On Recovering Fair and Accurate Classifiers with Imperfect Distributions

THESIS

submitted in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Computer Science and Engineering

by

Mohit Sharma

PHD19022

Under the supervision of

Prof. Rajiv Ratn Shah

and

Dr. Amit Deshpande

Indraprastha Institute of Information Technology, Delhi

New Delhi- 110020

August 2026

To my Nana, Nani and Mom,

for their selflessness, unconditional love, and infinite support.

Declaration

I declare that I am the author of the thesis titled **On Recovering Fair and Accurate Classifiers with Imperfect Distributions**. This dissertation contains the research I have conducted under the guidance of Prof. Rajiv Ratn Shah (CSE, IIIT Delhi) and Dr. Amit Deshpande (Microsoft Research India). To the best of my knowledge, it is an original work and has not been submitted elsewhere, in parts or in full, for a degree.

Mohit Sharma
Doctoral Candidate,
Dept. of Computer Science and Engineering,
IIIT Delhi, 110020.

Place: New Delhi

Date: 17 August 2026

Thesis Certificate

This is to certify that the thesis titled **On Recovering Fair and Accurate Classifiers with Imperfect Distributions**, submitted by **Mohit Sharma**, to the Indraprastha Institute of Information Technology, Delhi, for the award of the degree of **Doctor of Philosophy**, is a bona fide record of the research work done by him under our supervision. The contents of this thesis, in full or in parts, have not been submitted to any other Institute or University for the award of any degree or diploma.

Rajiv Ratn Shah
Thesis Co-Supervisor,
Associate Professor,
Dept. of Computer Science and Engineering,
IIIT Delhi, 110020.

Amit Deshpande
Thesis Co-Supervisor,
Principal Researcher,
Microsoft Research India.

Date: 17 August 2026

Declaration Regarding Use of AI

I acknowledge that I am fully responsible for the entire content of my thesis, including any sections assisted by online tools, including Artificial Intelligence-based tools. I accept full accountability for any violations of ethical standards in publications arising from the use of such tools.

Mohit Sharma
Doctoral Candidate,
Dept. of Computer Science and Engineering,
IIIT Delhi, 110020.

Place: New Delhi

Date: 17 August 2026

Acknowledgements

Doing the technical work towards a PhD was the easier part for me; the challenges of adult life were significantly harder. I am incredibly lucky to have met and formed close friendships with some amazing people and mentors, who have made the other parts of this journey significantly more manageable. This has been a story of taking extreme risks, and the people along the way have ensured that many of those risks were worth it.

I decided to pursue a research career around 2018, and the first serious shot at that came with the research internship at the MIDAS lab at IIIT Delhi under Prof. Rajiv Ratn Shah (unbeknownst to me, who would later become my PhD co-advisor). I am extremely grateful to Prof. Rajiv for the tremendous freedom to learn and experiment during that time. My internship also led to a full-time research fellowship and, eventually, to joining the PhD program at IIIT Delhi in 2020. I want to thank Prof. Yogesh Rawat at UCF for his guidance and patience during the early stages of my PhD, when my initial thesis topic was representation learning for large-scale noisy web videos.

This was also one of the most enjoyable and carefree times for me, when I made some amazing lifelong friends. I will always be thankful to them: Hitkul for always watching out for me and being my academic big brother, Karmanya for his surprising amount of patience and for introducing me to the world of coffee brewing (which is an inseparable part of my personality now), Shivangi for being available for any help, and Pakhi for her failed attempts to prank me back and being the only other child around.

I have also loved my time on IIIT-D's campus, and it was pleasant mainly because of so many colleagues and friends, a gazillion burnt toasts and coffee experiments in the lab, intense lab cricket matches, and hours of campus walks. I am extremely grateful to have met a wonderful set of friends: Asra, Atish, Biswa, Koushik, Mansi, Mann, Rishabh, Sakshi, Sagnik, Vishakha, and Yash. It's also been a privilege to meet some amazing folks at my lab and institute and have crazy adventures with them: Adarsh, Aman, Aradhya, Avinash, Ashwini, Debnath, Komal, Piyush, Raj, Ritika, Ritwik, Rohan, and Saumya. Special thanks to Prof. Manuj and the theory group for many fun parties and interesting, long conversations, and Prof. Arani, Prof. Praveen, and Prof. Subhabrata for the very intense table tennis matches at 7:30 PM. A big thank you to the people outside the campus who were constantly around: Divya, Mayan, Amresh, Verma, and Ravish.

Fortunately, the time spent during the COVID-19 lockdowns at home and on campus gave me enough time to rethink my research interests. Around the same time, ACM India launched a wonderful initiative called the PhD Clinic, which enabled us to have productive conversations with leading Indian researchers. I had the opportunity to discuss directions in theoretical Machine learning research with Prof. Madhavan Mukund from CMI. I am extremely thankful to him for connecting me to Prof. KV Subrahmanyam at CMI and Dr. Amit Deshpande at MSR India. What was supposed to be a one-off

meeting to gather opinions on potential research directions turned into six months of discussions about the initial ideas for this thesis, and into Amit becoming my PhD co-advisor. I am extremely thankful to both of them for their infinite kindness and patience.

Both of my PhD co-advisors have contributed uniquely to my PhD tenure, making it the most enjoyable time of my life so far. Rajiv has always been encouraging and helpful with many of the non-standard endeavors during my PhD, such as ensuring that the initial hardships are addressed, being very patient when I decided to abandon all progress and jump into a new, totally different thesis topic, and making new opportunities happen. Amit took me on as his PhD student, with a bag full of incorrect and embarrassing questions, and helped me transition into theoretical research. A special thanks to him and to Microsoft Research India for awarding me four years of financial support and multiple internships through the MSR India Joint PhD fellowship.

I have been extremely fortunate and privileged to intern at Microsoft Research India three times throughout my PhD. MSR India will always have a special place in my heart. The people and energy there are unique and unmatched by any other research labs, and the people I have befriended and connected with are among the kindest, humblest, and smartest I will ever meet. Thank you, Ananya, Deepanjali, Eshaan, Kanav, Mahek, Nimisha, and Prateeti from my first internship in 2022. I always fondly remember so many dinner plans and cards. Thank you, Agrima, Aditya, Drishti, Bhavyajeet, Bhuvan, Jatin, Manas, Nalin, Neelabh, Gurusha, Karthik, Rashi, Sukrit, Karan Tandon, Harsh Vijay, Saksham, and Vaibhav from my second and third internships, for those many random Saturday work plans that just ended up with chilling on the weekends in the office, and for the intense pool matches we had. And thank you, Anuradha, Archish, Amey, Harsh Vardhan, Kshitij, Pavan, Rahul, Sukruta, and Tanmay from my 2024 internship, for so much Marathi food, coffee, Bangalore housing brainstorming, and cricket sessions. I have also met some amazing mentors and friends in my many short stays in Bengaluru and at MSR: Nirjhar, Kiran, and Prof. Chiru.

I would also like to thank Prof Shin'ichi Satoh from the National Institute of Informatics, Tokyo, and Prof. Masashi Sugiyama from RIKEN-AIP in Tokyo for hosting me as a research intern in 2023 and 2025, respectively, and for an indirect opportunity to explore the wonderful country of Japan. I am incredibly grateful to my great friends for random long conversations and for tolerating an Italian and Indian meal outing every week: Jared, Irene, Profir, Shanglin Li, Okan-San, Atsushi-San, Mingyuan Bai, Romain, Henrique, Farzana, Sophia, and Sarada. A big thank you to the admin staff at both NII and RIKEN for making the administrative processes in a foreign country as smooth as possible. Finally, I would like to thank Prof. Rohit Vaish at IIT Delhi for hosting me while I awaited my defense and Prof. Vijay Keswani at IIT Delhi for our very long and interesting discussions about broader aspects of Responsible AI.

I also thank my internal review committee: Prof. Saket Anand and Prof. AV Subramanyam, my external reviewer for my thesis proposal, Prof. Avrim Blum, and finally my external thesis review committee, Prof. Han Bao, Prof. KV Subrahmanyam, and Prof. Siddharth Barman. A big thank you to the IIIT Delhi Department and Program admins, and the IT department for their constant help and availability. And once again, the biggest thanks to my parents and family for their constant support and encouragement. I would also like to acknowledge and thank Sagnik Chatterjee for this thesis template.

Mohit

Abstract

In the context of classification, the Bayes optimal classifier represents the best possible classification rule with respect to the 0-1 loss, for a given data distribution. In the setting of binary classification, it can often be expressed as a thresholding rule over the instance-dependent class posterior probability.

Prior research works have studied the best classification performance and the corresponding Bayes optimal classification rule under fairness constraints, which can again be expressed as a group/instance-dependent thresholding rule. However, it is often observed that fairness and accuracy of a model are often at odds and exhibit a tradeoff. This tradeoff depends on the distribution and the fairness metric of interest, and understanding its cause and effects is crucial for real-world deployments.

In this thesis, we study this tradeoff from several angles. One of the most important tools for all of our studies involves exploring the Bayes optimal fair classifiers to study the effects of data biases and derive theoretical guarantees. We first re-examine the critical stance on the fairness-accuracy tradeoff and instead study under what conditions fairness constraints can eventually help recover the unconstrained Bayes optimal classifier, especially when we can characterize the factors of data bias in our given distribution. Next, we highlight how the Bayes optimal fair classification rule can be implemented at various stages of classification: Either as a pre-processing of the given distribution, a weighted risk minimization, or as a post-processing (thresholding) of the class posterior probability. We then show, via simulations on widely used fair classifiers, that with varying amounts of data bias, the theoretical equivalence does not translate into practice and, in fact, sometimes even fails to mitigate unfairness.

Re-examining the tradeoff phenomena, we then ask whether it is possible to steer distributions most minimally and efficiently towards an *ideal* distribution, where the Bayes optimal classifiers are always fair by default. We mathematically characterize what an ideal distribution may look like when the distribution can be expressed with some parametric family (e.g., Gaussian distributions).

Finally, we estimate the fairness-accuracy tradeoff using imperfect data access, i.e., in a model-free setting, only using class posterior probabilities and without access to features. In the spirit of prior works that use Bayes error estimation techniques to benchmark classification performance, we study how we can estimate and characterize the fairness-accuracy tradeoff for any distribution only using soft labels (class posterior probabilities). This allows us to compare various fairness-accuracy tradeoff estimation techniques in the literature against the theoretically best and serves as an effective benchmarking framework for future work in this space.

Contents

Chapter 1

Introduction

Techniques and algorithms in machine learning have been at the forefront of enabling many technological innovations we see all around us. As machine learning continues to integrate into day-to-day tasks and workflows, it has been supporting critical applications, such as loan assessments and disease diagnosis.

The rapid availability of representative training data and computing resources has largely fueled this growth. The models we train are heavily influenced by the datasets we prepare, which are often collections of past observations, logs, and surveys. With such growth, challenges with the availability of high-quality data and training models that meet responsible machine learning requirements have rightfully attracted significant attention from the community. Unfortunately, we have had several documented instances of historical biases and discrimination, such as redlining based on demographics [Jac21] and severe disparities in commercial gender classification products [BG18]. This is where we, as a community, look beyond utility maximization and towards metrics and algorithms for Responsible AI.

§ 1.1 Fairness in Machine Learning

Automated decision-making in highly sensitive domains such as justice, healthcare, banking, and education can proliferate social and economic harms [CR18; Fri+19; BHN17; Bir+19; Meh+21]. Fair and representative data is a foremost requirement for training and validating fair models. However, most datasets collected and labeled inevitably contain historical and systemic biases [BS16], resulting in unfair and unreliable performance after deployment [Sha+17; BG18; WHM19].

Models can often latch onto statistical correlations and therefore exhibit those biased correlations in their outcomes. An example of such a correlation is shown in Figure 1.1(a), which highlights the highly variable percentage of women in different occupations. Figure 1.1(b) shows how relying on such correlation might affect deployed models: while translating English to Turkish (a gender neutral language) and then back to English, the model exhibits a gender bias of assigning a nurse to females, where originally it was assigned to a male.

Many seminal studies have highlighted the importance of formalizing, measuring, and mitigating biases in model training pipelines. Two such important ones are COMPAS [Ang+16] and Gender Shades [BG18]. The COMPAS study [Lar+16] analyzed more than 10,000 defendants in Broward

Figure 1.1: (a) Percentage of women in different occupations in the United States, and (b) An instance of Bias in Machine Translation. Turkish is a gender-neutral language, and when translating English to Turkish and back, the model may default to correlations in its training data. Source: Barocas et al. [BHN19].

County, Florida, and found that defendants of black ethnicity were far more likely to receive an incorrect judgment of high recidivism than white defendants. The Gender Shades study [BG18] highlighted intersectional disparities: commercial gender classification on facial image data performed substantially worse on darker-skinned females than on lighter-skinned males.

The first steps towards understanding potential unfairness in model outcomes are to define well-motivated fairness metrics. The definitions of fairness prevalent in literature fall into two broad types: individual fairness and group fairness [Dwo+12a; BJW20; Pet+21; Fle21; Pet+21; HPS16b; Zaf+17a; Bin20]. Individual fairness promises similar treatment to similar individuals and often requires defining the metric space of the individual features and the prediction space [Pet+21], which poses a strong challenge. It can also serve as a mechanism to induce stability and generalization [Muk+22].

We will focus on group fairness, which aims to ensure equal or near-equal outcomes across sensitive groups (e.g., race, gender) in the data and requires only individuals' sensitive group identities. Examples of group fairness include Equal Opportunity (equal true positive rates across groups) [HPS16a] and Demographic Parity (equal acceptance rates across groups) [Dwo+12a].

In the paradigm of group fairness, we can either explicitly use group information during training and inference (commonly known as group-aware classification) or use it only during training and infer without any sensitive attribute information (commonly known as group-blind classification) [BHN19]. Note, however, that merely omitting the use of sensitive attributes does not prevent unfairness: the sensitive might still correlate with other features in the data and result in unintentional harm (for example, in redlining, where zip codes strongly correlate with demographics). There are even works that attempt to train fair models with any demographic information, i.e., fairness without demographics [Lah+20; CJW22; AW23].

Various bias mitigation techniques have been proposed for group-fair classification motivated by different applications and stages of the machine learning pipeline [Zli15; Fri+19; Meh+21]. They are broadly categorized into three types: (1) *Pre-Processing*, where one transforms the input data, independent of the subsequent stages of training, and validation [KC12; Fel+15; Cal+17], (2) *In-Processing*, where models are trained by fairness-constrained optimization [CV10; Kam+12; Zem+13; Zaf+17b; ZLM18; Aga+18; Cel+19], and (3) *Post-Processing*, where the output predictions are adjusted afterward to improve fairness [KKZ12; HPS16b; Ple+17]. Several toolkits exist in the community for measurement, mitigation, and auditing of unfairness in machine learning pipelines [Bel+18; Bir+20; Del+24].

However, maximizing accuracy while satisfying fairness constraints often results in a trade-off be-

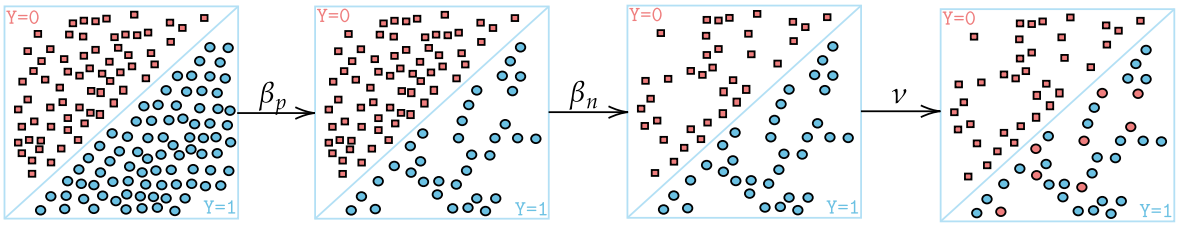


Figure 1.2: Illustration of the bias process studied in Blum and Stangl [BS19]. The positive and negative classes are first undersampled with a factor of β_p and β_n, respectively, and then a random ν fraction of the positive is flipped to negative.

tween fairness and accuracy [MW18; ZG19]. The sharpness of this trade-off depends on the type of constraints imposed, their stringency, and the properties of the underlying distribution. Prior work in this space has also sought to characterize this trade-off to understand it better and aid model designers [MW18; ZG19; Don+18; CJS18]. However, several works have also suggested that this trade-off arises from biased data or testing on imperfect distributions [BS19; WT+19; Mai+21; Dut+20]. When training data is biased, maximizing accuracy subject to fairness constraints may not help after deployment.

§ 1.2 Bayes Optimal Classification

To formally study the impacts of fairness constraints and their corresponding trade-offs, it is essential first to quantify the theoretically optimal performance. In the context of classification, the notion of Bayes error (and the corresponding Bayes optimal classifier) captures the minimum possible error rate achievable for a given data distribution [MRT12]. For binary classification, where we have only two classes, it is possible to express this optimal classifier as a constant thresholding rule over the class posterior probability: for a given example, pick the class that has the higher probability ($> 1/2$).

A threshold characterization of the Bayes optimal classifier helps design plug-in rules for classification, where we can "plug-in" the estimated posterior probabilities (and sometimes the threshold) from data and construct classification rules with appropriate thresholding [AT07; Chz19]. Plug-in classification rules have also been beneficial in designing Bayes optimal classifiers for performance metrics [NA13; Koy+14] and other cost-sensitive learning applications [Sco12; MW18; SK22]. Several works have used this formulation to estimate Bayes error and apply it to detect overfitting [Ish+; UIS25] and to detect test set contamination [ILY25].

When dealing with fairness constraints, we can extend the notion of Bayes optimal classifiers towards Bayes optimal *fair* classifiers. We can similarly express it as a plug-in thresholding rule over the class posterior probability, with the threshold function of the chosen metric and the desired unfairness tolerance [MW18; Chz+19; ZDC22b]. The optimal fair classification rule also allows us to study incompatibility between fairness constraints, where it may be infeasible to impose different types of fairness constraints without significant levels of accuracy degradation [KMR16; Ple+17; JD25]. We will study this fair classification rule in greater depth in the coming chapters.

§ 1.3 Imperfect Data Distributions

As highlighted earlier, a sharp fairness-accuracy trade-off might not be a necessary evil but rather a product of underlying group-specific data biases. We will now give a few examples of commonly

observed data biases.

Consider the redlining scenario highlighted earlier, where certain individuals are denied opportunities solely because of where they live, which correlates strongly with their ethnicity. The historical data being collected about opportunities being extended (for example, loans) and their outcome (repayments/defaulting) can create a systematic under-representation of minority groups and selection bias [KR18], where the data collected is only a representation of the population that the decision maker selected [DW21; Das+25]. Such lopsided data collection can lead to imbalances and under-representation, thereby inducing label biases, such as group-dependent label noise [WLL21] and prior probability shifts [BM21]. Several of these biases can also appear simultaneously: not only can the minority groups be under-represented, but imbalanced distributions may induce group-dependent label noise. An illustration of such a process is shown in Figure 1.2.

Several works have leveraged optimal fair classification to examine the above-described data biases [WT+19; BS19; Mai+21; Dut+20] and to present an alternate view of fairness constraints as tools for recovering accurate classifiers rather than strictly trading off utility. Many standard datasets used to evaluate and compare bias mitigation techniques are not perfectly fair and representative [Din+21; PMN21]. Data biases can creep into both the training and evaluation data, and fair classifiers evaluated on biased data can exhibit a sharp trade-off. Even in ideal data distributions where accuracy maximization yields perfect fairness, if we have only a biased or truncated version of the data, naively applying accuracy maximization or fair classification might not help [FSV21; Dut+20].

This motivates us to study optimal fair classification and fairness-accuracy trade-offs, not only for arbitrary distributions but also for distributions with known imperfections, such as label-, group-, and feature-dependent biases. It is also unclear as to what happens to optimal fair classifiers when they are trained and exposed to varying amounts of data bias over time in the training or deployment environments.

§ 1.4 Scope and Summary of Contributions

In this thesis, we aim to study a variety of imperfect data settings through the lens of Bayes Optimality. Most of our results are highlighted from the purview of fairness in machine learning and fair optimal characterizations [MW18; Chz+19]. The central object of most of our studies will be the characterization of the fair Bayes optimal classifier as a thresholding rule [MW18; Chz+19; Zen+26]. Our overall message is the following: fairness constraints are not merely penalties; they can also serve as tools for recovering accurate and fair classifiers in the presence of data biases.

We start our story by directly studying the relevance of fairness constraints. In Chapter 3, we study when fairness constraints aid the recovery of accurate classifiers, especially when the source of unfairness stems from biases that hinder certain sub-populations more than others. We instantiate this study by examining the label bias model of Blum and Stangl [BS19]. We re-prove their recovery results on the Bayes optimal classifiers under fairness constraints with threshold characterizations of fair classifiers. The newer proof technique then allows us to move beyond their stylized distribution setting and study recovery results for general distributions and hypothesis classes.

Next, in Chapter 4, we briefly note that the fair Bayes optimal classification rule can also be implemented as data distribution pre-processing or as a weighted risk-minimization scheme. This observation justifies the availability of several kinds of fairness interventions in the literature. We then use the data-bias model of Blum and Stangl [BS19] from Chapter 3 to study fair classification under varying levels of data bias. We empirically show that the output of several pre-/in- and post-processing

implementations can differ significantly and even fail to mitigate unfairness as the data bias in our distribution changes. Such observations raise a word of caution: the theoretical equivalence does not easily translate into practice, and a practitioner must move beyond simple accuracy evaluations and towards "stability" of interventions in the face of changing data conditions.

We then set out to mitigate the trade-off arising from feature biases in the presence of data. In Chapter 5, we define the notion of *ideal* data distributions [Dut+20]: the class of distributions where the Bayes optimal classifier is already fair. For parametric families, we obtain precise conditions in terms of the moments of the underlying distributions. We then show when existing distributions can be efficiently 'steered' towards such ideal distributions, and demonstrate applications for steering the internal representations of large language models.

Finally, for certain applications where we are interested in understanding the fairness-accuracy trade-off, we show how to obtain the theoretically optimal estimate of the trade-off using only class posterior probabilities, without access to features. In Chapter 6, we extend the previous line of work on Bayes error estimation [Ish+; UIS25] towards estimating the theoretically best fairness accuracy trade-off by only using class posterior probabilities. We study the consistency and sample complexity properties of such an estimator and show how previous work on fairness-accuracy trade-off estimation can be evaluated against the theoretically best achievable trade-off curve.

§ 1.5 List of Publications Related to this Thesis

- Mohit Sharma, Amit Deshpande, and Rajiv Ratn Shah. "On comparing fair classifiers under data bias" (URL). Presented as a spotlight presentation at the *Algorithmic Fairness through the Lens of Time*[1] workshop at NeurIPS 2023.
- Mohit Sharma and Amit Deshpande. "How far can fairness constraints help recover from biased data?" (URL). Presented as a poster at *International Conference on Machine Learning (ICML) 2024*.
- Mohit Sharma, Amit Deshpande, Chiranjib Bhattacharyya, and Rajiv Ratn Shah. "On Optimal Steering to Achieve Exact Fairness" (URL). Presented as a poster at *Annual Conference on Neural Information Processing Systems (NeurIPS) 2025*.
- Mohit Sharma, Okan Koç, Amit Deshpande, Takashi Ishida, Gang Niu, Rajiv Ratn Shah, Masashi Sugiyama. Practical Estimation of the Optimal Fairness-Accuracy Trade-off with Soft Labels. Under Review.

[1] https://www.afciworkshop.org/aft2023

Chapter 2

Preliminaries

In this chapter, we introduce the necessary notations, definitions, and some key results that will be useful throughout the manuscript. We introduce the Bayes optimal classifier as a thresholding rule over the positive class posterior probability. We then introduce the constrained optimization problem of imposing fairness constraints, and highlight a modified thresholding rule for the *fair* Bayes optimal classifier.

§ 2.1 Setup

We will denote random variables with capital letters and sets with calligraphic notations. In most settings we will study, we are given access to three objects: a random variable $X \in \mathcal{X}$ that denotes the features of an individual, $Y \in \mathcal{Y}$ that denotes a class label we are interested in predicting, and $A \in \mathcal{A}$, that denotes the protected or sensitive features of an individual like gender or ethnicity. We will work with the real-valued feature space ($\mathcal{X} = \mathbb{R}^d$), where d is the dimensionality of our feature space, binary classification ($\mathcal{Y} = \{0, 1\}$), and binary sensitive attributes ($\mathcal{A} = \{0, 1\}$). However, some of the results presented in this thesis can be generalized beyond binary sensitive attributes and labels.

Our data distribution D is defined over $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$ where a random variable from the distribution is denoted by $(X, A, Y) \sim D$. We will frequently deal with shifted or noisy distributions, which we will denote by $\widetilde{D}$, and the corresponding noisy random variable (for example, X) with $\widetilde{X}$. Whenever dealing with applications not involving fairness, we drop the use of A, and our data distribution D is defined over $\mathcal{X} \times \mathcal{Y}$. Another way to think about this is to assume that the sensitive attribute is part of the feature space itself. Unless explicitly mentioned, all probabilities and expectations are taken with respect to D. We write $\Pr(\cdot)$ and $\mathsf{E}[\cdot]$ for these quantities, with $\Pr(\cdot \mid \cdot)$ denoting the corresponding conditional probability. We use $\mathbb{I}\,()$ for the indicator function, equal to 1 when its argument is true and 0 otherwise.

Whenever unspecified, the set of classifiers we will deal with is the set of all deterministic functions $h : \mathcal{X} \times \mathcal{A} \to \mathcal{Y}$. We will overload the notation $\mathcal{H}$ with any other space of classifiers explicitly when required. We will use the standard $0-1$ loss as our performance metric. The $0-1$ loss assigns a penalty of 1 whenever the model's prediction and the true label mismatch. We will denote the $0-1$ loss with $l_{01} : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$.

§ 2.2 Bayes Optimal Classification

The most accurate model for a given data distribution can often be characterized in terms of distribution-dependent quantities. Specifically, consider the following optimization problem:

$$\min_{h} \mathop{E}_{(X,A,Y)\sim D} [l_{01}(h(X,A),Y)] .$$

The solution to this program is called the Bayes optimal classifier for a distribution D [MRT12; SSBD14] and we shall denote it with h^*. The Bayes error is the expected $0-1$ loss of h^*: $\mathop{E}_{D} [l_{01}(h^*(X,A),Y)]$.

For binary classification, the Bayes optimal classifier can be expressed more succinctly. Let $\eta(x,a) = \Pr(Y=1|X=x,A=a)$ denote the positive-class posterior probability for a feature $X=x$ coming from the group $A=a$. A well-known result in the case of binary classification [DGL96; Zen+26] states the following:

$$h^*(x,a) = \mathbb{I}\left(\eta(x,a) \geqslant \frac{1}{2}\right),$$

where $\mathbb{I}(E) = 1$ whenever event E is true. Simply put, the Bayes optimal classifier for a binary classifier is just a constant threshold on the positive class probability. It is not hard to generalize this optimal rule to multiple classes (where we can no longer threshold on η). The Bayes optimal multiclass classifier ($|\mathcal{Y}| > 2$) is given by the following rule:

$$h^*(x,a) = \arg\max_{y\in\mathcal{Y}} \eta_y(x,a), \text{ where } \eta_y(x,a) = \Pr(Y=y|X=x,A=a) .$$

Expressing the Bayes optimal classifier as a threshold classifier allows us to motivate the set of threshold classifiers in the next section. This will then help us study Bayes optimal *fair* classifiers and other interesting variations.

§ 2.3 Thresholding Classifiers

Define the class of all deterministic classifiers from the domain $\mathcal{X}\times\mathcal{A}$ to $\{0,1\}$ as $\mathcal{H}$. We will now define a subset of $\mathcal{H}$ that requires access to the class posterior probability η.

Definition 2.1. Let $\mathcal{H}$ be the set of all deterministic classifiers $h : \mathcal{X}\times\mathcal{A} \to \{0,1\}$. Then, the set of all threshold classifiers $\mathcal{H}_t \subset \mathcal{H}$ is defined as the following:

$$\mathcal{H}_t = \{(x,a) \mapsto \mathbb{I}(\eta(x,a) \geqslant t_a) : t_a \in [0,1]\},$$

where t_a is a group-dependent threshold.

Such threshold classifiers come up in various scenarios. For example, the Bayes optimal classifier for a binary classifier uses $t = 1/2$. The fair Bayes optimal classifier for a given fairness metric (which we will define and discuss later) is defined using $t = 1/2 + \delta$, where the adjustment δ depends on the fairness violation. Such threshold classifiers, where the threshold on η is a function of distribution-dependent

quantities like class priors, also appear in other classification paradigms. One concrete example of learning with rejection or abstention [BW08; CDM16; Cha+21], where the goal is to abstain from classifying points with low confidence. To define the reject option classifier, we first define the extended $0-1$ loss, which introduces a cost for rejection:

Definition 2.2. For a $\delta \in [0, 1/2)$, we define the $0-1-$rej loss of a classifier $h : \mathcal{X} \times \{0,1\} \to \{0,1\} \cup \{\perp\}$ as the following:

$$l_{01r}(h, y) = \begin{cases} 0 & \text{if } h(x, a) = y \\ 1 & \text{if } h(x, a) \neq y \\ \delta & \text{if } h(x, a) = \perp, \end{cases}$$

where $\perp$ denotes abstention.

For reject-option classification, we can express a binary-class Bayes optimal reject option classifier in the following manner.

Proposition 2.1. ([BW08; CDM16]) Let D be any distribution on $\mathcal{X} \times \{0,1\} \times \{0,1\}$. Let $\delta \in [0, 1/2)$ denote the cost of abstaining from prediction for a particular point. Then the optimal reject option classifier with respect to the l_{01r} loss (Definition 2.2), h^{rej}, can be expressed as:

$$h^{rej}(x, a) = \begin{cases} 0, & \text{if } \eta(x, a) \leqslant \delta \\ \perp, & \text{if } \eta(x, a) \in (\delta, 1-\delta) \\ 1, & \text{if } \eta(x, a) \geqslant 1-\delta, \end{cases}$$

where $\eta(x, a) = \Pr(Y = 1 | X = x, A = a)$ and $\perp$ denotes abstention.

For many different performance metrics falling under the purview of cost-sensitive classification [Sco12], such as Jaccard Similarity, F-measure, [Koy+14], the Bayes optimal classifier falls under $\mathcal{H}_t$. To introduce a general result that applies to many such cost-sensitive risks, we extend the simple approach of Scott [Sco12] to a group-aware setting.

Lemma 2.1. (Extension of [Sco12] to the case of subgroups) Let $\alpha_a \in [0, 1]$. Define the α-CS risk as weighing the false negatives for a group $A = a$ with $(1 - \alpha_a)$ and the false positives with α_a:

$$R_\alpha(h) = \sum_{a \in \mathcal{A}} \left((1 - \alpha_a) \cdot \pi_{1a} \cdot \mathop{E}_{X|Y=1, A=a} [l_{01}(h(X, a), 1)] + \alpha_a \cdot \pi_{0a} \cdot \mathop{E}_{X|Y=0, A=a} [l_{01}(h(X, a), 0)] \right),$$

where $\pi_{ya} = \Pr(Y = y, A = a) > 0$ for $y \in \{0,1\}$. The cost-sensitive group-dependent Bayes optimal classifier can be written down as:

$$h^*_\alpha(x, a) = \begin{cases} 1, & \text{if } \eta(x, a) \geqslant \alpha_a \\ 0, & \text{otherwise} \end{cases}$$

§ 2.4 Fairness Constraints and Optimal *Fair* Classifiers

A major theme in the upcoming chapters will concern imperfections in observed distributions. We will call such distributions *biased* distributions. Whenever machine learning models are trained on a biased distribution, they not only result in sub-optimal accuracy (e.g., label noise), but can also result in *unfairness* in the downstream models, with respect to sensitive attributes in our features like demographics or gender [BHN19; BS19; Dut+20].

To quantify the unfairness in machine learning models, practitioners have developed several fairness metrics motivated by various use cases [Nar18]. For the rest of this thesis, we will work with *group* fairness, where we care about the fairness of classifiers with respect to disjoint groups in the feature space (eg, gender, ethnicity). Several other notions of fairness exist. Two such important notions include individual fairness [Dwo+12b], which aims to ensure that similar individuals receive similar treatment, and intersectional fairness [Fou+20], which concerns groups that intersect across sensitive attributes. Notions other than group fairness are beyond the scope of the work conducted in preparation for this thesis.

§ 2.4.1 Group Fairness Metrics

The Fair ML scholarship has developed various group fairness metrics over the last decade. Most of them can be clustered into three different groups: Independence, Separation, and Sufficiency-based fairness constraints. Barocas et al. [BHN19] present an excellent discussion on all three categories. We include a brief discussion and introduce the formal definition here.

As a first attempt, a practitioner may want the model's predictions to be completely independent of the sensitive attribute. This is the basis of *Independence*-based fairness constraints. Independence requires the prediction to be independent of the sensitive attribute: $h(X, A) \perp A$. Examples of such fairness constraints include Demographic Parity [Dwo+12a], and Disparate Impact [Fel+15]. We will now define a binary and multi-class, multi-attribute definition of Demographic Parity.

> **Definition 2.3.** For a distribution D defined over $\mathcal{X} \times \{0,1\} \times \{0,1\}$, a classifier $h : \mathcal{X} \times \{0,1\} \to \{0,1\}$ satisfies *Demographic Parity* if $\Pr(h(X,A) = 1 | A = 0) = \Pr(h(X,A) = 1 | A = 1)$. Similarly, for a multi-class, multi-attribute setting, and for a distribution D defined over $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$, a classifier $h : \mathcal{X} \times \mathcal{A} \to \mathcal{Y}$ satisfies *Demographic Parity* if $\forall y \in \mathcal{Y}$, $\max_{a,a' \in \mathcal{A}} |\Pr(h(X,A) = y | A = a) - \Pr(h(X,A) = y | A = a')| = 0$.

We denote the difference of positive rate between groups as $\Delta_{DP}(h, D)$ (Similarly, $\Delta_{DP,y}(h, D)$ for multi-class settings). The above definition is for exact Demographic Parity and states that $\Delta_{DP}(h, D) = 0$. A serious limitation of independence-based constraints is that they do not account for error. For example, we can achieve demographic parity by randomly assigning $p\%$ of the population to positive decisions and selecting the top $p\%$ for the other group. This will satisfy Demographic Parity, but in terms of accuracy, it will be horrendous for the first group.

To address this severe limitation, we can condition on the target variable Y. These will be called *Separation*-based fairness metrics. With Separation, we require the prediction to be independent of the sensitive attribute, given the true label: $h(X, A) \perp A \mid Y$. Examples of such metrics include equality of the false-negative and false-positive rates between groups [HPS16a]. We will now define one such fairness metric, which will be used frequently in later chapters.

Definition 2.4. For a distribution D defined over $\mathcal{X} \times \{0,1\} \times \{0,1\}$, a classifier $h : \mathcal{X} \times \{0,1\} \to \{0,1\}$ satisfies *Equal Opportunity* (or equality of true positive rates) if $\Pr(h(X,A) = 1|Y = 1, A = 0) = \Pr(h(X,A) = 1|Y = 1, A = 1)$. Similarly, for a multi-class, multi-attribute setting and a distribution D defined over $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$, a classifier $h : \mathcal{X} \times \mathcal{A} \to \mathcal{Y}$ satisfies *Equal Opportunity* if $\forall y \in \mathcal{Y}$, $\max_{a,a' \in \mathcal{A}} |\Pr(h(X,A) = y|Y = y, A = a) - \Pr(h(X,A) = y|Y = y, A = a')| = 0$.

Similar to Demographic Parity, we denote the difference in true positive rates between groups as $\Delta_{EO}(h, D)$ ($\Delta_{EO,y}(h, D)$ in multi-class settings). A similar definition can be written down for equality of False Positive Rates. Requiring both the True and False negative rates to be equal constitutes the definition of *Equalized Odds* [HPS16a].

Another category of fairness metrics defined in the literature is the *Sufficiency*-based fairness metrics. Let R be a random variable that denotes the score of an individual, which may be used later to make a decision or otherwise used as a feature itself (for example, as a credit score). Sufficiency requires that the label is conditionally independent of the group given the score $R : \Pr(Y = 1|R = r, A = 0) = \Pr(Y = 1|R = r, A = 1)$. Conditioning the prediction $h(X,A)$ on the risk score yields a fairness metric known as Predictive Parity. Note that in certain cases, Sufficiency boils down to group-wise calibration and can be achieved by unconstrained loss minimization [LSH19].

Prior work has also attempted to unify different fairness metrics into a single working definition and to propose unified optimization frameworks for fair classification. Noting that most fairness constraints are linear in the classifier of interest, Agarwal et al. [Aga+18] propose writing fairness constraints as linear constraints $M \cdot \mu(h) \leqslant c$, where $\mu(h)$ is a vector of the conditional moments with respect to the classifier h, M is a matrix, and c is a vector of tolerance (for exact fairness, it will be a zero vector).

Celis et al. [Cel+19] propose writing fairness constraints as linear and linear-fractional group performance constraints in the following form:

$$q_i(h) = \frac{\alpha_0^{(i)} + \sum_{j=1}^{k} \alpha_j^{(i)} \cdot \Pr\left(h = 1|G_i, A_j^{(i)}\right)}{\beta_0^{(i)} + \sum_{j=1}^{l} \beta_j^{(i)} \cdot \Pr\left(h = 1|G_i, B_j^{(i)}\right)},$$

for two integers $k, l \geqslant 0$, $q_i(h)$ denoting the conditional moment with respect to the i^{th} group, G_i denoting the event that the probability is taken over group i, events $A_j^{(i)}, B_j^{(i)}$ being independent of h, and parameters $\alpha_0^{(i)}, ..., \alpha_k^{(i)} \in \mathbb{R}$ and $\beta_0^{(i)}, ..., \beta_k^{(i)} \in \mathbb{R}$ being independent of h, but dependent on the underlying distribution D. They later show that many fairness constraints can be written in this form. Finally, Zeng et al. [Zen+26] proposes writing the fairness metric as a *Linear Disparity Measure*. A fairness metric is *linear* if there exists a $w : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ such that for all classifiers h, $Dis(h) = \int_{\mathcal{A}} \int_{\mathcal{X}} h(x,a)\, w(x,a) d\mathbb{P}(x,a)$. Furthermore, it is *bilinear* if there exist distribution-dependent quantities s, b such that $w(x,a) = s \cdot \eta(x,a) + b$.

Finally, note that it is well researched that the above-described three families of fairness criteria are not simultaneously achievable [KMR16; Sar20; Bel+23a]. However, by relaxing the criteria and working with randomized fair classifiers that output a probability rather than a hard decision, it may be possible to satisfy some of these fairness criteria simultaneously [Ple+17]. Several recent works have demonstrated the power of randomization in achieving a better tradeoff [AD22; Aga+] and satisfying incompatible fairness constraints [Ple+17; JD25].

§ 2.4.2 Bayes Optimal *Fair* Classifiers

Now that we have some familiarity with fairness constraints, we can study the fair Bayes optimal classifier. We now highlight a result shown by [Chz+19] and many other works in different settings [MW18; Cel+19; Zen+26; CKL23]. Note that for Propositions 2.2 and 2.3, we assume the following about η:

Assumption 2.1. (No interior atoms / Continuity of the Cumulative Distribution function of η) The mapping $t \to F_a(t) = \Pr(\eta(X,a) \leqslant t | A = a), t \in [0,1], \forall a \in \mathcal{A}$ is continuous in $(0,1)$. Moreover, $\Pr(\eta(X,a) \geqslant t | A = a) > 0$ for all $a \in \mathcal{A}$ for all $t \in (0,1)$.

Proposition 2.2. For any distribution D over $\mathcal{X} \times \{0,1\} \times \{0,1\}$, let h^*_{EO} be a classifier of the maximum accuracy among all binary classifiers that satisfy the Equal Opportunity (Definition 2.4) on D. There exists $\lambda^* \in \mathbb{R}$ such that $h^*_{EO}(x,a) = \mathbb{I}(\eta(x,a) \geqslant t_a)$, where

$$t_a = \begin{cases} \dfrac{1}{2 + \dfrac{\lambda^*}{\Pr(Y=1, A=0)}}, & \text{for } a = 0 \\ \dfrac{1}{2 - \dfrac{\lambda^*}{\Pr(Y=1, A=1)}}, & \text{for } a = 1. \end{cases}$$

The proof for Proposition 2.2 is given in Section 2.6. The above Proposition characterizes the *EO fair* Bayes optimal classifier for binary sensitive attributes and binary labels. Other works have derived characterizations for multiple attributes [Zen+26] and multiple classes as well [XZ24]. A similar characterization proof can be followed for the Demographic Parity fairness constraint:

Proposition 2.3. For any distribution D, let h^*_{DP} be a classifier of the maximum accuracy among all binary classifiers that satisfy Demographic Parity (Definition 2.3) on D. Then there exists a $\lambda^* \in \mathbb{R}$ such that

$$h^*_{DP}(x,a) = \begin{cases} \mathbb{I}\left(\eta(x,0) \geqslant \dfrac{1}{2} - \dfrac{\lambda^*}{2\Pr(A=0)}\right), & \text{for } a = 0 \\ \mathbb{I}\left(\eta(x,1) \geqslant \dfrac{1}{2} + \dfrac{\lambda^*}{2\Pr(A=1)}\right), & \text{for } a = 1 \end{cases}$$

The proof for Proposition 2.3 is analogous to the proof of Proposition 2.2. For both of the above characterizations, the elephant in the room is λ^*. Depending on the desired level of fairness (exact or approximate), we can formulate an optimization program to find λ^*. Within the same proposition (Proposition 2.2), Chzhen et al. [Chz+19] claim that $|\lambda^*| \leqslant 2$ and λ^* is such that:

$$\frac{\mathop{E}_{X|A=1}\left[\eta(X,1)\mathbb{I}\left(1 \leqslant \eta(X,1)(2 - \frac{\lambda^*}{\Pr(Y=1,A=1)})\right)\right]}{\Pr(Y=1|A=1)} = \frac{\mathop{E}_{X|A=0}\left[\eta(X,0)\mathbb{I}\left(1 \leqslant \eta(X,0)(2 - \frac{\lambda^*}{\Pr(Y=1,A=0)})\right)\right]}{\Pr(Y=1|A=0)},$$

which is essentially saying that any $\lambda^* \in [-2,2]$ that satisfies exact equal opportunity according to the above expression works. Celis et al. [Cel+19] in Theorem 4.4 of their work propose an efficient optimization program to find λ^* for approximate fairness. Finally, Zeng et al. [Zen+26] shows (1) Monotonicity of the unfairness in error rate with respect to λ (Proposition 4.1), and (2) convexity of the fairness-accuracy tradeoff curve parameterized by λ (Proposition 4.6). Together, the above-mentioned results on

finding λ^* for any level of unfairness can help devise clever algorithms to find suitable λ and to study the fairness-accuracy tradeoff, which is tricky otherwise for arbitrary distributions.

To conclude this section, we note that many works in the fairness literature have studied such characterization beyond the binary label and sensitive attributes settings [MW18; Chz+19; Cel+19; CKL23; ZDC22b; Zen+26; XZ24; WZC24; Hu+25; Den+24; YHC25; HZ24]. In Chapter 4, we will revisit one such characterization and compare pre-/in- and post-processing interventions. Furthermore, note that it is possible to beat the tradeoff and decrease the severity of the tradeoff with *randomization*, where the solution space is allowed to take values in the entire interval $[0, 1]$, rather than just $\{0, 1\}$ [AD22; Aga+; Ple+17; JD25; AL21]. However, while such an intervention improves performance in expectation, it leads to increased variance and *arbitrariness*: the phenomena where models may produce conflicting outputs on the same input [CGN19; Lon+23].

§ 2.5 List of Notations

In Table 2.1, we list some notations that will be used extensively throughout this thesis.

Table 2.1: Summary of Mathematical Notation

Symbol	Description
$\mathcal{X}$	Feature space $\mathbb{R}^d$
$\mathcal{Y}$	Label space $\{0, 1\}$
$\mathcal{A}$	Sensitive attribute space $\{0, 1\}$
x	A point from the feature space $\mathcal{X}$
y	A point from the label space $\{0, 1\}$
a	A point from the sensitive attribute space $\{0, 1\}$
D	Data distribution over $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$
(X, A, Y)	A random variable from the joint distribution D
$\eta(x, a)$	Class posterior probability $\Pr(Y = 1 \mid X = x, A = a)$
$h, \mathcal{H}$	A classifier and the corresponding hypothesis class
h^*	The Bayes optimal classifier for distribution D
l_{01}	The $0-1$ loss function: $\mathcal{Y} \times \mathcal{Y} \to \mathbb{R}^+$
Δ_{DP}, Δ_{EO}	Disparity measures for Demographic Parity and Equal Opportunity
π_{ya}	Joint probability of label and attribute $\Pr(Y = y, A = a)$
$\pi_{y\mid a}$	Conditional probability of label, given an attribute $\Pr(Y = y \mid A = a)$
π	Probability of the positive class $\Pr(Y = 1)$
π_a	Probability of group a $\Pr(A = a)$

§ 2.6 Proof of Proposition 2.2

Proof. The proof of this proposition follows the arguments from the proof of Proposition 2.3 in [Chz+19]. However, we include the core proof ideas here for completeness.

We will use a well-known technique based on Lagrange duality to provide a threshold-based characterization of optimal equal-opportunity classifiers [MW18; Chz+19]. Given any classifier $h : \mathcal{X} \times \mathcal{A} \to \mathcal{Y}$, let $\mathrm{TPR}_a(h)$ denote its true positive rates conditioned on $A = a$ for the distribution D. Then,

$$\mathrm{TPR}_a(h) = \frac{\Pr\left(h(X, a) = 1, A = a, Y = 1\right)}{\Pr\left(Y = 1, A = a\right)} = \frac{\underset{X\mid A=a}{\mathrm{E}}\left[h(X, a)\eta(X, a)\right]}{\underset{X\mid A=a}{\mathrm{E}}\left[\eta(X, a)\right]}.$$

Furthermore, we define $h^*_{EO} = \underset{h \in \mathcal{H}_{\text{fair, EO}}}{\operatorname{argmax}} \Pr(h(X,A) = Y)$, where $\mathcal{H}_{\text{fair, EO}} = \{h \in \mathcal{H} : \text{TPR}_0(h) = \text{TPR}_1(h)\}$. The objective can be expressed as:

$$\begin{aligned}
h^*_{EO} &= \underset{h\in\mathcal{H}}{\operatorname{argmax}} \min_{\lambda\in\mathbb{R}} \Pr(h(X,A)=Y) + \lambda \Pr(h(X,A)=1|Y=1,A=0) \\
&\qquad\qquad - \lambda \Pr(h(X,A)=1|Y=1,A=1) \\
&= \Pr(Y=0) + \underset{h\in\mathcal{H}}{\operatorname{argmax}} \min_{\lambda\in\mathbb{R}} \sum_{a=0}^{1} \Pr(A=a) \underset{X|A=a}{\mathrm{E}}[h(X,a)(2\eta(X,a)-1)] \\
&\qquad\qquad + \frac{(-1)^{\mathbb{I}(a=1)}\lambda}{\underset{X|A=a}{\mathrm{E}}[\eta(X,a)]} \underset{X|A=a}{\mathrm{E}}[h(X,a)\eta(X,a)] \\
&= \underset{h\in\mathcal{H}}{\operatorname{argmax}} \min_{\lambda\in\mathbb{R}} \sum_{a=0}^{1} \Pr(A=a) \underset{X|A=a}{\mathrm{E}}\left[h(X,a)\left(\left(2+\frac{(-1)^{\mathbb{I}(a=1)}\lambda}{\Pr(Y=1,A=a)}\right)\eta(X,a)-1\right)\right] \\
&= \underset{h\in\mathcal{H}}{\operatorname{argmax}} \min_{\lambda\in\mathbb{R}} L(h,\lambda)
\end{aligned}$$

By weak duality, $\max_{h\in\mathcal{H}} \min_{\lambda\in\mathbb{R}} L(h,\lambda) \leqslant \min_{\lambda\in\mathbb{R}} \max_{h\in\mathcal{H}} L(h,\lambda)$. For a fixed λ, solving the dual objective $\max_{h\in\mathcal{H}} L(h,\lambda)$, we get:

$$h_\lambda(x,a) = \mathbb{I}\left(\eta(x,a) \geqslant \frac{1}{2+\frac{(-1)^{\mathbb{I}(a=1)}\lambda}{\Pr(Y=1,A=a)}}\right).$$

Substituting h_λ back into the dual objective reduces the positive part of the objective to a one-dimensional problem in λ. Note that the mapping $\lambda \mapsto \underset{X|A=a}{\mathrm{E}}\left[\left(\left(2+\frac{(-1)^{\mathbb{I}(a=1)}\lambda}{\Pr(Y=1,A=a)}\right)\eta(X,a)-1\right)_+\right]$ is convex in λ. Due to Assumption 2.1, the random variable $\eta(x,a)$ has no atoms, which assures differentiability. Since the dual objective is convex in λ, any dual minimizer λ^* satisfies the following first-order condition:

$$\frac{\underset{X|A=1}{\mathrm{E}}\left[\eta(X,1)\mathbb{I}\left(1 \leqslant \eta(X,1)(2-\frac{\lambda^*}{\Pr(Y=1,A=1)})\right)\right]}{\Pr(Y=1|A=1)} = \frac{\underset{X|A=0}{\mathrm{E}}\left[\eta(X,0)\mathbb{I}\left(1 \leqslant \eta(X,0)(2-\frac{\lambda^*}{\Pr(Y=1,A=0)})\right)\right]}{\Pr(Y=1|A=0)}.$$

Note that this is exactly the TPR equality condition we seek, and hence h_{λ^*} is a feasible solution for our original problem. This means that $\Pr(h_{\lambda^*}=Y) \leqslant \max_{h\in\mathcal{H}} \min_{\lambda\in\mathbb{R}} L(h,\lambda)$. However, $(\lambda^*, h_{\lambda^*})$ also solves the dual problem and the constraint vanishes at feasibility: $\Pr(h_{\lambda^*}=Y) \leqslant \max_{h\in\mathcal{H}} \min_{\lambda\in\mathbb{R}} L(h,\lambda) \leqslant \min_{\lambda\in\mathbb{R}} \max_{h\in\mathcal{H}} L(h,\lambda) = \Pr(h_{\lambda^*}=Y)$, where the last equality holds since the constraint term vanishes at h_{λ^*}. Hence, equality holds throughout and h_{λ^*} is the optimal EO-fair classifier.

□

Chapter 3

How Far Can Fairness Constraints Help Recover From Biased Data?

Abstract

A general belief in fair classification is that fairness constraints incur a trade-off with accuracy, which biased data may worsen. Contrary to this belief, Blum and Stangl [BS19] shows that fair classification with equal opportunity constraints, even on extremely biased data, can recover optimally accurate and fair classifiers on the original data distribution. Their result is interesting because it demonstrates that fairness constraints can implicitly rectify data bias and simultaneously overcome a perceived fairness-accuracy trade-off. Their data bias model simulates under-representation and label bias in the underprivileged population, and they show the above result on a stylized data distribution with i.i.d. label noise, under simple conditions on the data distribution and bias parameters.

We propose a general approach to extend the result of Blum and Stangl [BS19] to different fairness constraints, data bias models, data distributions, and hypothesis classes. We strengthen their result, and extend it to the case when their stylized distribution has labels with Massart noise instead of i.i.d. noise. We prove a similar recovery result for arbitrary data distributions using fair reject option classifiers. We further generalize it to arbitrary data distributions and arbitrary hypothesis classes, i.e., we prove that for any data distribution, if the optimally accurate classifier in a given hypothesis class is fair and *robust*, then it can be recovered through fair classification with equal opportunity constraints on the biased distribution whenever the bias parameters satisfy certain simple conditions. Finally, we show applications of our technique to time-varying data bias in classification and fair machine learning pipelines.

§ 3.1 Introduction

Fairness-accuracy trade-offs often seem inevitable, and bias mitigation guarantees can fail when there is a mismatch between train and test data distributions; it is a double whammy when we get neither the desired accuracy nor the desired fairness on deployment. Hence, it is important to understand the role of fair classification and fairness constraints when the training data in machine learning pipelines has systematic biases [BS19; KL22; Sch+22].

Blum and Stangl [BS19] propose a model for data bias that simulates systematic under-representation of the underprivileged group and label bias/flip on the underprivileged positive population. Applying this data bias model to a stylized data distribution with i.i.d. label noise, they prove a recovery

result that makes the following high-level point: careful choice of fairness constraints can implicitly rectify even extreme data bias and overcome a perceived fairness-accuracy trade-off. Formally, for their stylized distribution with i.i.d. label noise, they prove that the optimal fair classifier satisfying equal opportunity on a biased (or bias-induced) data distribution coincides with the Bayes optimal classifier (that also happens to be perfectly fair) on the original data distribution, if the bias parameters satisfy certain simple conditions. Their proof is tailored to equal opportunity constraints and uses a careful case analysis and iterative argument that is not amenable to arbitrary distributions or hypothesis classes. They also show a specific distribution where a similar result fails to hold if we use demographic parity constraints instead of equal opportunity constraints.

We take a significantly different approach and observe that the regression function for group-aware classification ($\eta(x, a)$) undergoes a linear fractional transformation when we apply various data bias models, including the one proposed by Blum and Stangl [BS19]. This observation plays an important role in our proofs because the optimal group-aware fair classifier among the hypothesis class of all binary classifiers can be mathematically characterized by group-dependent thresholds on the regression function [MW18; Chz+19]. Under the conditions on the data distribution and bias parameters as in Blum and Stangl [BS19], we show that the fairness-corrected thresholds applied to a linear fractional transformation recover the Bayes optimal classifier on the stylized distribution of Blum and Stangl [BS19] with i.i.d. label noise. Our proof uncovers a stronger version of their result that these conditions on the bias parameters are both necessary and sufficient. Moreover, our proof readily extends to the case when the above distribution has Massart or malicious label noise [MN06; RS94; Slo96] instead of i.i.d. label noise.

We extend our recovery result to arbitrary data distributions by considering reject option classifiers [BW08; CDM16] that can abstain from prediction on a small fraction of inputs by paying a penalty. We further generalize our results to allow arbitrary data distributions and arbitrary hypothesis classes. Generalizing beyond threshold-based arguments, we show that, for arbitrary data distributions, if the optimal classifier in a given hypothesis class is fair and *robust*, it can be recovered from fair classification on the biased distribution whenever the bias parameters satisfy simple conditions. Finally, we propose multi-step models to capture time-varying data bias in machine learning pipelines and investigate the necessary conditions on the data distribution and bias parameters for similar recovery results at finite and infinite time horizons. Now we outline the organization of our results:

- Section 3.3 contains our theoretical setup for data bias and group-aware fair classification. In Subsection 3.3.1, we describe some recently studied data bias models in fair classification [BS19; DB20; WLL21; Bis+19] (see Examples 3.1, 3.2, & 3.3), and show that all of them result in a linear fractional transformation of the regression function. Subsection 3.3.2 captures how fairness constraints and threshold-based characterizations of optimal fair classifiers change under data bias. We focus on equal opportunity constraints and the data bias model of Blum and Stangl [BS19] (Examples 3.1) as an illustrative example running through the rest of this chapter.
- In Section 3.4, we extend the result of Blum and Stangl [BS19] to the case when their stylized distribution has Massart label noise instead of i.i.d. label noise (Theorem 3.2). We strengthen their result (Theorem 3.3) and prove its analog for demographic parity constraints replacing equal opportunity (Theorem 3.4).
- In Section 3.5, we show that our proof technique based on threshold classifiers extends to *arbitrary* data distributions, if we allow reject option classifiers that can abstain from prediction on a small fraction of inputs.

- In Section 3.6, we invent clever workarounds to generalize our results further to recover the optimal fair and *robust* hypothesis in an *arbitrary* hypothesis class simply by fair classification on the biased version of an *arbitrary* data distribution (Theorem 3.6).
- Finally, in Section 3.7, we propose time-varying data bias models (also applicable to multi-stage machine learning pipelines) and investigate necessary conditions for extending our recovery results above to the finite and infinite time horizon (Theorems 3.7 & 3.8).

§ 3.2 Related Work

There has been a plethora of recent work on fairness-accuracy trade-offs and data bias [MW18; WT+19; BS19; Dut+20; Mai+21], but the closest to our work is the result of Blum and Stangl [BS19] that we strengthen and generalize in many ways. Though we take equal opportunity constraints [HPS16a] and the data bias model of Blum and Stangl [BS19] for under-representation and label bias as an illustrative example, our techniques readily extend to other popular fairness constraints such as demographic parity [Dwo+12a] and other recent data bias models [DB20; WLL21; BM21]. Our proof techniques in Section 3.4 & 3.5 lean heavily on threshold-based characterizations of optimal fair classifiers known in previous work [MW18; Chz+19; ZDC22a; ZDC22b].

Now we describe recent related works that complement our approach to rectify data bias in fair classification. The feasibility of fair classification under data corruption and malicious noise in training data has been studied in Konstantinov and Lampert [KL22] and Blum et al. [Blu+23]. Recent work has studied fair classification with noisy sensitive attributes [Lam+19; GKW23; Cel+21], noisy labels [FCG20], feature-dependent label bias [JN20], sample selection bias [DW21; Zhu+23], subpopulation shift [Mai+21], and causal models of data bias [PB22; Mad+19; CKG23]. All of them propose algorithmic modifications to the vanilla fair classification to rectify noisy or biased data. Complementing the theoretical aspects, recent work has also empirically investigated the effect of data bias and choice of fairness constraints on the accuracy and fairness of various fair classifiers [Isl+22; Akp+22; SDS23b; GKW23].

§ 3.3 Data Bias Models & Fair Classification

We use $\tilde{D}$ to denote the biased joint distribution over features, sensitive attributes and class labels, and use $(\tilde{X}, \tilde{A}, \tilde{Y})$ to denote a random data point from the biased distribution $\tilde{D}$. If X (or A) in the joint distribution remains unchanged when we go from D to $\tilde{D}$, then we simply use X (or A) in the place of $\tilde{X}$ (or $\tilde{A}$). All proofs for the results that follow are included in Section 3.9.

§ 3.3.1 Regression Functions on Biased Data

Below, we list some data bias models from recent works on fair classification from biased data [BS19; DB20; BM21]. We make a key observation that for all of these data bias models, the regression function on the biased distribution $\tilde{\eta}(x, a) = \Pr\left(\tilde{Y} = 1 | \tilde{X} = x, \tilde{A} = a\right)$ can be expressed as a linear fractional transformation of the regression function on the original distribution $\eta(x, a) = \Pr\left(Y = 1 | X = x, A = a\right)$. In other words,

$$\tilde{\eta}(x, a) = \frac{P\eta(x, a) + Q}{R\eta(x, a) + S}, \quad \text{for some } P, Q, R, S \in \mathbb{R}.$$

Please refer to Propositions 3.6, 3.7, and 3.8 in Section 3.9.2 for proofs of each of the below examples.

Example 3.1. [BS19] Consider a biased distribution $\tilde{D}$ obtained from the original distribution D by the following process defined by under-representation and label bias parameters $\beta_p, \beta_n, \nu \in (0,1)$. A random data point (X, A, Y) from D with $A = 1$ remains unchanged, the points with $A = 0, Y = 1$ survive independently with probability β_p, and the points with $A = 0, Y = 0$ survive independently with probability β_n. Finally, the surviving points with $A = 0, Y = 1$ keep their class label 1 with probability $1 - \nu$ and flip to 0 with probability ν. For the privileged group $A = 1$, we have $\tilde{\eta}(x, 1) = \eta(x, 1)$, for all $x \in \mathcal{X}$, whereas for the underprivileged group $A = 0$, we prove that $\tilde{\eta}(x, 0) = \frac{(1-\nu)\eta(x,0)}{(1-c)\eta(x,0)+c}$, where $c = \frac{\beta_n}{\beta_p}$ in Proposition 3.6.

Example 3.2. [DB20; WLL21] Consider a biased distribution $\tilde{D}$ obtained from the original distribution D by introducing a group-dependent label flip from (X, A, Y) to $(X, A, \tilde{Y})$ where $\epsilon^1_a = \Pr(\tilde{Y} = 0 | Y = 1, A = a)$ and $\epsilon^0_a = \Pr(\tilde{Y} = 1 | Y = 0, A = a)$, with $0 \leqslant \epsilon^1_a + \epsilon^0_a < 1$. For this data bias model, we show that $\tilde{\eta}(x, a) = (1 - \epsilon^1_a - \epsilon^0_a)\eta(x, a) + \epsilon^0_a$ in Proposition 3.7.

Example 3.3. [BM21] Consider a biased distribution $\tilde{D}$ obtained from D by introducing a group-dependent prior probability shift such that $\tilde{A} = A$ and

$$\Pr(\tilde{X} = x | \tilde{Y} = i, A = a) = \Pr(X = x | Y = i, A = a), \text{ for any } i, a \in \{0, 1\},$$

but $\Pr(\tilde{Y} = i | A = a) \neq \Pr(Y = i | A = a)$. For this data bias model, we prove that

$$\tilde{\eta}(x, a) = \frac{\eta(x, a)}{(1-\alpha)\eta(x, a) + \alpha} \text{ where } \alpha = \frac{\Pr(\tilde{Y} = 0 | A = a) \Pr(Y = 1 | A = a)}{\Pr(\tilde{Y} = 1 | A = a) \Pr(Y = 0 | A = a)},$$

in Proposition 3.8.

For classifiers that apply a threshold on $\eta(x, a)$ or $\tilde{\eta}(x, a)$, it is important to understand when the above linear fractional transformations are order-preserving.

Proposition 3.1. Suppose $S > 0$, $R + S > 0$, and $PS - QR > 0$, then the transformation $\tilde{\eta}(x, a) = \frac{P\eta(x,a) + Q}{R\eta(x,a) + S}$ is order-preserving, i.e., $\eta(x_1, a) < \eta(x_2, a)$ iff $\tilde{\eta}(x_1, a) < \tilde{\eta}(x_2, a)$.

The proof is provided in Section 3.9.1. Note that all of the above data bias models satisfy the order-preservation property. For the rest of this chapter, we exclusively focus on the under-representation and label bias model in Example 3.1 [BS19] as an illustrative example. Our techniques are flexible and applicable to other data bias models.

§ 3.3.2 Fair Classification on Biased Data

Demographic Parity (equal group-wise positivity rates) and Equal Opportunity (equal group-wise true positive rates) are two most popular fairness constraints in classification. It is easy to see that the true positive rates (TPRs) and the true negative rates (TNRs) for a classifier on the biased data distribution can be expressed as linear combinations of its TPRs and TNRs on the original data distribution.

Proposition 3.2. Let D be any distribution on $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$ and let $\tilde{D}$ be its biased version defined using any of the data bias models defined in Subsection 3.3.1. Let h be any hypothesis, and let $\mathrm{TPR}_a(h)$ and $\widetilde{\mathrm{TPR}}_a(h)$ be its true positive rates according to D and $\tilde{D}$, respectively, conditioned on the underprivileged group $A = a$. Let $\mathrm{TNR}_a(h)$ and $\widetilde{\mathrm{TNR}}_a(h)$ be the true negative rates defined similarly. Then

1. $\widetilde{\mathrm{TPR}}_a(h) = \Pr\left(Y = 1 | \tilde{Y} = 1, A = a\right) \mathrm{TPR}_a(h) + \Pr\left(Y = 0 | \tilde{Y} = 1, A = a\right) \mathrm{FPR}_a(h)$, and
2. $\widetilde{\mathrm{TNR}}_a(h) = \Pr\left(Y = 0 | \tilde{Y} = 0, A = a\right) \mathrm{TNR}_a(h) + \Pr\left(Y = 1 | \tilde{Y} = 0, A = a\right) \mathrm{FNR}_a(h)$.

The proof of Proposition 3.2 is given in Section 3.9.3, and we get the following interesting corollary for the data bias model in Example 3.1 [BS19].

Corollary 3.1. Let D be any distribution and $\tilde{D}$ be its biased version as in Example 3.1. Given any hypothesis class $\mathcal{H}$, let $\mathcal{H}_{\text{fair, EO}}$ be its subset that satisfies equal opportunity on the original distribution D, i.e., $\mathcal{H}_{\text{fair, EO}} = \{h \in \mathcal{H} : \mathrm{TPR}_0(h) = \mathrm{TPR}_1(h)\}$. Similarly, let $\widetilde{\mathcal{H}}_{\text{fair, EO}} = \{h \in \mathcal{H} : \widetilde{\mathrm{TPR}}_0(h) = \widetilde{\mathrm{TPR}}_1(h)\}$ be the subset of $\mathcal{H}$ satisfying equal opportunity on the biased distribution $\tilde{D}$. Then, for any h, $\widetilde{\mathrm{TPR}}_0(h) = \mathrm{TPR}_0(h)$ and $\widetilde{\mathrm{TPR}}_1(h) = \mathrm{TPR}_1(h)$, and hence, $\mathcal{H}_{\text{fair, EO}} = \widetilde{\mathcal{H}}_{\text{fair, EO}}$.

Remark 3.3.1. Corollary 3.1 can be extended to other fairness metrics such as demographic parity as well as the hypothesis class of *approximately* fair classifiers that satisfy fairness constraints up to a small additive or multiplicative error. However, we focus on exact equal opportunity as in Blum and Stangl [BS19] for a direct, illustrative application. Please see Section 3.9.3 for additional results on a similar relationship for the positive rate (Proposition 3.9) and a corollary highlighting the relationship between biased positive rate and the true parities, similar to Corollary 3.1 (Corollary 3.10).

The optimal fair classifier for equal opportunity (similarly, demographic parity) on a given data distribution can be expressed by group-dependent thresholds applied to the regression function [MW18; Chz+19]. Since the regression function $\tilde{\eta}(x, a)$ on the biased distribution $\tilde{D}$ is an order-preserving linear fractional transformation of $\eta(x, a)$, the optimal fair classifier for equal opportunity on $\tilde{D}$ is equivalent to applying group-dependent thresholds to $\eta(x, a)$.

Proposition 3.3. For any distribution D and its biased version $\tilde{D}$ described in Example 3.1, let $\tilde{h}_{EO}$ be a classifier of the maximum accuracy among all binary classifiers that satisfy equal opportunity on $\tilde{D}$. Then there exists $\lambda^* \in \mathbb{R}$ such that $\tilde{h}_{EO}(x, a) = \mathbb{I}\left(\eta(x, a) \geqslant t_a\right)$, where

$$t_a = \begin{cases} \dfrac{1}{1 + \dfrac{1-2\nu}{c} + \dfrac{\lambda^*}{\beta_n \Pr\left(Y = 1, A = 0\right)}}, & \text{for } a = 0 \\ \dfrac{1}{2 - \dfrac{\lambda^*}{\Pr\left(Y = 1, A = 1\right)}}, & \text{for } a = 1. \end{cases}$$

The Proof of Proposition 3.3 is given in Section 3.9.3. We can similarly derive the optimal threshold with the biased distribution for the Demographic parity constraint (Proposition 3.10 in Section 3.9.3).

§ 3.4 Recovering Optimal Classifier from Biased Data for Massart Label Noise

Blum and Stangl [BS19] consider a stylized distribution D with i.i.d. label noise and show that the optimal fair classifier $\tilde{h}_{EO}$ on the biased distribution $\tilde{D}$ (defined in Example 3.1) recovers the Bayes optimal (and fair) classifier h^* on the original distribution D, if the bias parameters satisfy certain simple conditions. Note that this does not require knowing, estimating, or correcting for data bias explicitly, and their result holds even for extreme under-representation and label bias in $\tilde{D}$. We first demonstrate the utility of our technique by generalizing the recovery result of Blum and Stangl [BS19] to the case of Massart noise [MN06]. We describe the distribution setup below, give a sketch of our proof, and point out the generality of our technique compared to Blum and Stangl [BS19].

§ 3.4.1 Generalizing Blum and Stangl [BS19] Recovery Result for Massart Noise

Assume any arbitrary data distribution D on $\mathcal{X} \times \mathcal{A}$. Let $\Pr(A = 0) = r$ and $\Pr(A = 1) = 1 - r$, for some $0 < r < 1$. Let $h : \mathcal{X} \times \mathcal{A} \to \mathcal{Y}$ be any hypothesis that satisfies $\Pr(h(X, A) = 1|A = 0) = \Pr(h(X, A) = 1|A = 1)$. Let $\delta < 1/2$, and extend the distribution to $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$ as follows. $Y|X = x, A = a$ takes value $h(x, a)$ with probability $1 - \delta(x, a)$, and $\neg h(x, a)$ with probability $\delta(x, a)$, for some $\delta(x, a) \leqslant \delta$. Let D be the resulting distribution on $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$. This type of bounded noise in class label is popularly known as Massart noise[1] in literature, based on a noise model proposed by Massart and Nédélec [MN06]. We assume that the Massart noise is added in a way that equalizes the base rates on the two protected groups, i.e., $\Pr(Y = 1|A = 0) = \Pr(Y = 1|A = 1) = q$. Since $\delta < 1/2$, the Bayes optimal classifier h^* on the distribution D coincides with h and satisfies Equal Opportunity.

Theorem 3.2. For any distribution D defined as above and its biased version $\tilde{D}$ defined as in Example 3.1 using the bias parameters $\beta_p, \beta_n, \nu \in (0, 1)$. If the data distribution and bias parameters satisfy

$$(1-r)(1-2\delta) + r\left((1-\delta)\beta_p(1-2\nu) - \delta\beta_n\right) > 0$$

and

$$(1-r)(1-2\delta) + r\left((1-\delta)\beta_n - \delta\beta_p(1-2\nu)\right) > 0,$$

then the optimal equal opportunity classifier on the biased distribution $\tilde{D}$ recovers the Bayes optimal classifier on the original distribution D, i.e., $\tilde{h}_{EO} \equiv h^*$.

The proof of Theorem 3.2 is given in Section 3.9.4. The same proof also works for group-dependent Massart noise, i.e., there exist $\delta_0, \delta_1 < 1/2$ such that $\delta(x, a) \leqslant \delta_a$, for all $(x, a) \in \mathcal{X} \times \mathcal{A}$.

The conditions in Theorem 3.2 above are identical to those in the recovery result of Blum and Stangl [BS19] (Theorem 4.1 in their paper) that only works for the special case when $\delta(x, a) = \delta$, for all $(x, a) \in \mathcal{X} \times \mathcal{A}$. Their proof is arguably less flexible to other models of label noise and data bias, as it relies on clever, iterative modifications of an initial fair classifier until its accuracy cannot be improved further. For their special case $\delta(x, a) = \delta$, for all $(x, a) \in \mathcal{X} \times \mathcal{A}$, our technique gives a stronger statement that the conditions in Theorem 3.2 are in fact both necessary and sufficient. Figure 3.1 illustrates the recovery conditions in Theorem 3.2 for reasonably chosen group proportion parameter r, label bias ν and δ.

[1] Equivalently known as malicious classification noise in previous work [RS94; Slo96].

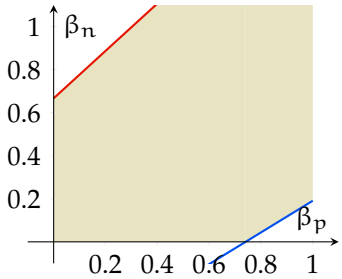


Figure 3.1: Recovery region for $\beta_p, \beta_n \in (0,1]$ given by the constraints $(1-r)(1-2\delta) + r((1-\delta)\beta_p(1-2\nu) - \delta\beta_n) > 0$ and $(1-r)(1-2\delta) + r((1-\delta)\beta_n - \delta\beta_p(1-2\nu)) > 0$ as in Theorem 3.2, when $r = 0.25$, $\nu = 0.05$, and $\delta = 0.45$. We can recover optimal and fair classifiers for a large range of data biases, including extreme under-representation, i.e., region close to the origin $(0,0)$, by applying just equal opportunity constraints.

Theorem 3.3. (a slightly stronger version of Theorem 4.1 in Blum and Stangl [BS19]) For the data distribution D described above with $\delta(x,a) = \delta$, for all $(x,a) \in \mathcal{X} \times \mathcal{A}$, and its biased version $\tilde{D}$ as described in Example 3.1, the data distribution and bias parameters satisfy

$$(1-r)(1-2\delta) + r((1-\delta)\beta_p(1-2\nu) - \delta\beta_n) > 0$$
and
$$(1-r)(1-2\delta) + r((1-\delta)\beta_n - (1-2\nu)\delta\beta_p) > 0,$$

if and only if the optimal equal opportunity classifier on $\tilde{D}$ recovers the Bayes optimal classifier on D, i.e., $\tilde{h}_{EO} \equiv h_{EO} \equiv h^*$.

Similarly, we can also obtain necessary and sufficient conditions for when the optimal demographic parity classifier on $\tilde{D}$ recovers h^*. Blum & Stangl [BS19] only give a specific example where such a recovery is impossible via demographic parity constraints (see Subsection 3.1 of [BS19]) but do not prove any analog of Theorem 3.3 (see Table 1 in [BS19]). We prove a weaker recovery condition, where the under-representation in both the classes for the under-privileged group is the same ($\beta_p = \beta_n$).

Theorem 3.4. For the data distribution D described above with $\delta(x,a) = \delta$, for all $(x,a) \in \mathcal{X} \times \mathcal{A}$, and its biased version $\tilde{D}$ as described in Example 3.1, if $\beta_p = \beta_n$, then the data distribution and bias parameters satisfy $(1-2\delta)(1-r) - r(1-2(1-\delta)(1-\nu)) > 0$ and $(1-2\delta)(1-r) + r(1-2\delta(1-\nu)) > 0$, if and only if the optimal demographic parity classifier on $\tilde{D}$ recovers the Bayes optimal classifier on D, i.e., $\tilde{h}_{DP} \equiv h_{DP} \equiv h^*$.

Section 3.9.4 contains the proofs of Theorems 3.3 & 3.4.

§ 3.4.2 Proof Sketches

We briefly outline our proof technique for Theorems 3.2, 3.3, 3.4 to explain an important technical contribution in this chapter. We write fairness-constrained accuracy maximization using a Lagrange multiplier λ. Proposition 3.3 characterizes $\tilde{h}_{EO}$ as $\tilde{h}_{EO}(x,a) = \mathbb{I}(\eta(x,a) \geqslant t_a)$ that applies group-dependent thresholds t_a on $\eta(x,a)$, where the threshold t_a is actually a function of the optimal Lagrange multiplier λ^*, the data distribution parameters, and the bias parameters. We show (see Lemma 3.1) that as long as these thresholds t_a applied to $\eta(x,a)$ for both the groups $a = 0$ and $a = 1$ lie within the interval

$(\delta, 1-\delta)$, we have $\tilde{h}_{EO} \equiv h^*$. We show that the possible choices of λ^* are narrowed down to allow only $\tilde{h}_{EO} \equiv h^*$ using the given conditions on the data distribution and bias parameters, and the fairness constraint on the resulting threshold classifier. We prove that the conditions in Theorem 3.3 are necessary and sufficient for the optimal λ^* parameter in Proposition 3.3 to satisfy that the group-dependent thresholds t_0 and t_1 lie in the interval $(\delta, 1-\delta)$, and equivalently, $\tilde{h}_{EO} \equiv h^*$.

§ 3.5 Recovery of Optimal Reject Option Classifiers from Biased Data for Arbitrary Data Distributions

A major limitation of Theorems 3.2 and 3.3 is that they work only on stylized distributions, where the label noise is either i.i.d. or Massart. In this section, we remove this limitation by proving a similar recovery result for arbitrary data distributions. Massart or i.i.d. label noise creates a clear separation between $\eta(x, a)$ values (or high-risk and low-risk), and allows a small interval margin for the group-wise thresholds to recover h^*. To mimic this in an arbitrary data distribution, we consider *reject option classifiers* that are allowed to abstain from prediction by paying a penalty [BW08; CDM16; Cha+21; SC21; FPV23]. Models that abstain from prediction play an important role in responsible machine learning, as predictions of high uncertainty can be overseen by a human in the loop. As a result, many prior works have studied reject option classifiers for fair classification [MPZ18; SC21; Sha+22].

Let D be an arbitrary distribution on $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$, with $\mathcal{A} = \{0, 1\}$ and $\mathcal{Y} = \{0, 1\}$. A reject option classifier $g : \mathcal{X} \times \mathcal{A} \to \{0, 1, \perp\}$ either rejects or abstains from prediction on an input (x, a), denoted by $g(x, a) = \perp$, or it predicts $g(x, a) = h(x, a)$ using a binary classifier $h : \mathcal{X} \times \mathcal{A} \to \{0, 1\}$. Let $\mathcal{H}$ denote the hypothesis class of all binary classifiers $h : \mathcal{X} \times \mathcal{A} \to \{0, 1\}$, and let $\mathcal{H}^{rej}$ be the hypothesis class of all $g : \mathcal{X} \times \mathcal{A} \to \{0, 1, \perp\}$. For rejection penalty given by $\delta > 0$, the optimal reject option classifier is defined as

$$h^{rej} = \operatorname*{argmin}_{g \in \mathcal{H}^{rej}} \Pr\left(g(X, A) \neq Y, g(X, A) \neq \perp\right) + \delta \Pr\left(g(X, A) = \perp\right).$$

Proposition 3.4 characterizes the optimal reject option classifier on any distribution D, which is a generalization of the known forms of Optimal reject option classifiers (Section 1 in Bartlett and Wegkamp [BW08], Section 2 in Cortes et al. [CDM16]).

Proposition 3.4. Let D be any distribution on $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$, with $\mathcal{A} = \{0, 1\}$ and $\mathcal{Y} = \{0, 1\}$. Let $\delta \in [0, 1/2)$ denote the rejection penalty and h^{rej} be the optimal reject option classifier defined as above. Then

$$h^{rej}(x, a) = \begin{cases} 0, & \text{if } \eta(x, a) \leqslant \delta \\ \perp, & \text{if } \eta(x, a) \in (\delta, 1-\delta) \\ 1, & \text{if } \eta(x, a) \geqslant 1-\delta, \end{cases}$$

where $\eta(x, a) = \Pr(Y = 1 | X = x, A = a)$.

The proof of Proposition 3.4 is given in Section 3.9.5. Proposition 3.4 shows how reject option induces a separation between high and low $\eta(x, a)$ values similar to the case of i.i.d. or Massart label noise. A larger separation has an obvious trade-off with a larger fraction of inputs being turned away for model prediction.

Now assume that the optimal reject option classifier h^{rej} satisfies equal opportunity on the non-

rejected part of the distribution D, i.e., $\Pr\left(h^{rej}(X,A)=1|Y=1,A=0\right)=\Pr\left(h^{rej}(X,A)=1|Y=1,A=1\right)$. Theorem 3.5 shows that the optimal equal opportunity classifier on the biased distribution $\tilde{D}$ recovers h^{rej} on the non-rejected inputs, if the data distribution and bias parameters satisfy the same conditions as in Theorems 3.2 & 3.3.

Theorem 3.5. Let D be an arbitrary distribution on $\mathcal{X}\times\mathcal{A}\times\mathcal{Y}$, with $\mathcal{A}=\{0,1\}$ and $\mathcal{Y}=\{0,1\}$. Let $\delta\in[0,1/2)$ be the rejection penalty and suppose the optimal reject option classifier h^{rej} defined above satisfies equal opportunity on the non-rejected part of distribution D. Let $\tilde{D}$ be a biased version of D defined as in Example 3.1 using bias parameters $\beta_p,\beta_n,\nu\in(0,1)$, and let $\tilde{h}_{EO}$ be the optimal equal opportunity classifier on $\tilde{D}$. If the data distribution and bias parameters satisfy

$$(1-r)(1-2\delta)+r\left((1-\delta)\beta_p(1-2\nu)-\delta\beta_n\right)>0$$
and
$$(1-r)(1-2\delta)+r\left((1-\delta)\beta_n-\delta\beta_p(1-2\nu)\right)>0,$$

then $\tilde{h}_{EO}(x,a)=h^{rej}(x,a)$ whenever $h^{rej}(x,a)\neq\perp$.

A complete proof of Theorem 3.5 can be found in Section 3.9.5. Note that if we consider the entire distribution D instead of only the non-rejected inputs, then

$$\Pr\left(\tilde{h}_{EO}(X,A)\neq h^{rej}(X,A)\right)\leqslant\Pr\left(h^{rej}(X,A)=\perp\right).$$

In other words, if $\Pr\left(h^{rej}(X,A)=\perp\right)$ is small, then $\tilde{h}_{EO}$ matches h^{rej} on distribution D with high probability.

§ 3.6 Recovering Robust Hypothesis under Data Bias for Arbitrary Data Distributions and Arbitrary Hypothesis Classes

In this section, we remove the restriction on hypothesis class $\mathcal{H}$, assumed to be the class of all group-aware binary classifiers in Sections 3.4 & 3.5. Note that the characterization of optimal fair classifiers using group-aware thresholds on the regression function $\eta(x,a)$ plays an important role in our proofs from Sections 3.4 & 3.5. For an arbitrary hypothesis class $\mathcal{H}$, even the classifier $h^*\in\mathcal{H}$ that maximizes accuracy on D need not be a threshold classifier on $\eta(x,a)$. To work around this, we make an assumption that the optimal (and fair) classifier that we want to recover under data bias must be *robust* under small perturbations to the data distribution D. Our definition of ϵ-robustness is motivated by the linear fractional transformations of the regression function observed in various data bias models earlier (see Section 3.3).

Let D be any distribution on $\mathcal{X}\times\mathcal{A}\times\mathcal{Y}$, with $\mathcal{A}=\{0,1\}$ and $\mathcal{Y}=\{0,1\}$. Let $\Pr(A=0)=r$, $\Pr(A=1)=1-r$, and let $\Pr(Y=1|A=0)=\Pr(Y=1|A=1)=q$. Let h^* be the most accurate classifier in $\mathcal{H}$, i.e.,

$$h^*=\operatorname*{argmax}_{h\in\mathcal{H}}\Pr\left(h(X,A)=Y\right)=\operatorname*{argmax}_{h\in\mathcal{H}}\underset{(X,A)}{E}\left[h(X,A)(2\eta(X,A)-1)\right].$$

Note that, for an arbitrary hypothesis class $\mathcal{H}$, the optimal h^* need not be a threshold classifier on

$\eta(x, a)$. As in the previous sections, we assume that h^* satisfies equal opportunity on the distribution D, i.e., $\Pr(h^*(X, A) = 1|Y = 1, A = 0) = \Pr(h^*(X, A) = 1|Y = 1, A = 1)$.

Definition 3.1. We define $h^* \in \mathcal{H}$ to be ϵ-*robust* if, for any distribution D' with random data points (X, A, Y') s.t.

$$\eta'(x, a) = \Pr(Y' = 1|X = x, A = a) = \frac{P_a \eta(x, a) + Q_a}{R_a \eta(x, a) + S_a},$$

with $S_a = 1, |R_a| \leqslant \epsilon, |Q_a| \leqslant \epsilon, 1 - \epsilon \leqslant P_a \leqslant 1 + \epsilon$, and $P_a S_a - Q_a R_a \geqslant 0$, the optimal classifier h^* remains unchanged, i.e., $h^* = h' = \underset{h \in \mathcal{H}}{\operatorname{argmax}} \Pr(h(X, A) = Y')$.

The above conditions give a scale-invariant proxy to say that the linear fractional transformation defined by P_a, Q_a, R_a, S_a is order-preserving, and when it is appropriately scaled to make $S_a = 1$, it is ϵ-close to the identity transformation. Definition 3.1 says that the classifier h^* of maximum accuracy in $\mathcal{H}$ is robust to small near-identity perturbations of the data distribution D.

Now we are ready to state our result: we recover robust, fair hypotheses from biased data under arbitrary data distributions and hypothesis classes.

Theorem 3.6. For any distribution D and any hypothesis class $\mathcal{H}$, suppose the data bias model exhibits uniform under-representation, i.e., $\beta_p = \beta_n = \beta$. If the optimal classifier $h^* \in \mathcal{H}$ is ϵ-robust and the bias parameters β, ν satisfy:

$$r\beta(\epsilon - \nu) + \epsilon(1 - r) \geqslant 0 \text{ and } r\beta(\epsilon + \nu) + \epsilon(1 - r) \geqslant 0,$$

then the optimal equal opportunity classifier from $\mathcal{H}$ on the biased distribution $\tilde{D}$ recovers the optimal classifier from $\mathcal{H}$ on the original distribution D, i.e., $\tilde{h}_{EO} \equiv h^*$.

We prove Theorem 3.6 in Section 3.9.6. Our proof reuses the basic characterization of optimal fair classifiers using Lagrange multipliers, and although it uses the class probabilities $\eta(x, a)$'s in a crucial way, it circumvents the need for threshold-based arguments completely. Figure 3.2 illustrates the recovery conditions in Theorem 3.6, for reasonably chosen group proportion parameter r, label bias ν and hypothesis robustness parameter ϵ.

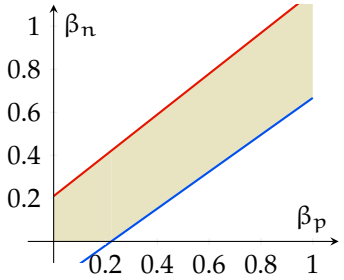


Figure 3.2: Recovery region for $\beta_p, \beta_n \in (0, 1]$ given by the constraints $r((1 - \nu)\beta_p - (1 - \epsilon)\beta_n) + \epsilon(1 - r) \geqslant 0$ and $r((1 + \epsilon)\beta_n - (1 - \nu)\beta_p) + \epsilon(1 - r) \geqslant 0$ as in Theorem 3.6, when $r = 0.2$, $\nu = 0.1$, and $\epsilon = 0.05$. Even for arbitrary distributions and hypothesis classes, optimal and fair classifiers can be recovered from extreme under-representation and for a large range of data biases using just equal opportunity constraints.

§ 3.7 Recovering from Time-Varying Data Bias

Data biases arise commonly in machine learning pipelines where data changes for downstream applications and over time. As an application of our techniques, we demonstrate how we can obtain recovery guarantees in data bias pipelines, i.e., when at every time step, we obtain a new shifted distribution. We model time-varying data bias as repeated applications of single-step data bias model (e.g., Example 3.1 used in previous sections).

Let $\tilde{D}_t$ be the biased data distribution obtained from an original distribution D, when the data bias model described in Example 3.1 gets applied repeatedly t times with possibly different bias parameters $(\beta_{p,i}, \beta_{n,i}, \nu_i)$ at i-th time step, and let $c_i = \dfrac{\beta_{n,i}}{\beta_{p,i}}$. Since the composition of linear fractional transformations remains a linear fractional transformation, we obtain the following generalization for how the regression function changes over time.

Proposition 3.5. Let (X, A, Y) denote a random data point from any given distribution D and let $(X, A, \tilde{Y}_t)$ be a random data point from its corresponding biased distribution $\tilde{D}_t$ after applying the multi-stage time-varying data bias model described above. Let $\eta(x, a) = \Pr(Y = 1|X = x, A = a)$ and $\tilde{\eta}_t(x, a) = \Pr\left(\tilde{Y}_t = 1|X = x, A = a\right)$. Then

$$\tilde{\eta}_t(x, 0) = \frac{\eta(x, 0)}{\sum_{i=1}^{t} \left(\frac{1 - c_i}{1 - \nu_i} \prod_{j=i+1}^{t} \frac{c_j}{1 - \nu_j} \right) \eta + \prod_{i=1}^{t} \frac{c_i}{1 - \nu_i}}, \text{ where } \prod_{j=t+1}^{t} \frac{c_j}{1 - \nu_j} \stackrel{\text{def}}{=} 1.$$

Note that $\tilde{D}_t$ conditioned on $A = 1$ remains unchanged, and therefore, $\tilde{\eta}_t(x, 1) = \eta(x, 1)$, since we are looking at the data bias model in Example 3.1. The proof of Proposition 3.5 is by simple induction on t. As a result of Proposition 3.5, $\tilde{\eta}_t(x, a)$ can be expressed as another single-step data bias model that directly transforms $\eta(x, a)$ into $\tilde{\eta}_t(x, a)$ using a linear fractional transformation with $P = 0, Q = 0, R = \sum_{i=1}^{t} \left(\frac{1 - c_i}{1 - \nu_i} \prod_{j=i+1}^{t} \frac{c_j}{1 - \nu_j} \right)$, and $S = \prod_{i=1}^{t} \frac{c_i}{1 - \nu_i}$. The above P, Q, R, S obey the conditions in Proposition 3.1, so the corresponding linear fractional transformation is order-preserving.

§ 3.7.1 Repeated Data Bias & Infinite Time Horizon

As a warm-up, we first study a simpler case where the bias parameters do not change at each time step, i.e., $\beta_{p,i} = \beta_p$, $\beta_{n,i} = \beta_n$, and $\nu_i = \nu$, for all $i \in [t]$). Equivalently, the same data bias model with the same bias parameters gets applied repeatedly t times. We work with the original distribution D described as in Theorem 3.3 for the ease of analysis, and assume that $\Pr(A = 0) = r$, $\Pr(A = 1) = 1 - r$ and $\Pr(Y = 1|A = 0) = \Pr(Y = 1|A = 1) = q$. First, we show Theorem 3.7 about when the optimal equal opportunity classifier $\tilde{h}_{EO,t}$ on the biased distribution $\tilde{D}_t$ can recover h^* for the infinite time horizon as $t \to \infty$, and the necessary conditions on the data distribution and bias parameters to allow that. The proof of Theorem 3.7 is given in Section 3.9.7.

Theorem 3.7. Let $\tilde{D}_t$ be the biased distribution obtained by applying the bias model in Example 3.1 repeatedly t times with $\beta_{p,i} = \beta_p$, $\beta_{n,i} = \beta_n$, and $\nu_i = \nu$, for all $i \in [t]$ on a given distribution D defined as in Theorem 3.3. Let h^* be the Bayes optimal classifier on D and let $\tilde{h}_{EO,t}$ be the optimal equal opportunity classifier on $\tilde{D}_t$.

- If $\beta_n < 1$ then as $t \to \infty$, we have $\tilde{\eta}(x, 0) \to 0$, for all $x \in \mathcal{X}$. Thus, we cannot have $\tilde{h}_{EO,t} \equiv h^*$ as $t \to \infty$.

- If $\beta_n = 1$ and $\tilde{h}_{EO,t} \equiv h^*$ as $t \to \infty$, then the data distribution and bias parameters must satisfy

$$1 - \frac{(1-2\delta)}{(1-\delta)r} < \frac{\nu\beta_p}{1-\beta_p(1-\nu)} < 1 + \frac{(1-2\delta)}{\delta r}.$$

§ 3.7.2 Time-Varying Data Bias Pipeline with Finite Steps

Now we generalize the necessary conditions in Theorem 3.3 for repeated application of the data bias model from Example 3.1, with possibly different bias parameters at each time step. We assume that the bias parameters can vary but are bounded.

Theorem 3.8. Let D be a data distribution described as in Theorem 3.3 and $\tilde{D}_t$ be the resulting biased distribution after repeated application of the data bias model from Example 3.1 to D in t steps, with bounded but possibly different bias parameters in each step as $\beta_{n,t} \in [\beta_n, 1]$, $\beta_{p,t} \in [\beta_p, 1]$ and $\nu_t \in [0, \nu]$, where $\nu < 1/2$. Let h^* be the Bayes optimal classifier on D and let $\tilde{h}_{EO,t}$ be the optimal equal opportunity classifier on $\tilde{D}_t$. If $\tilde{h}_{EO,t} \equiv h^*$, then the data distribution and bias parameters must satisfy

$$\frac{(1-2\delta)(1-r)}{(1-\delta)r} - \frac{\delta}{1-\delta} > \frac{1}{\beta_n^t} - 2\beta_p^t(1-\nu)^t \quad \text{and}$$
$$\frac{(1-2\delta)(1-r)}{\delta r} > (1-\nu)^t\beta_p^t(1-\beta_p)\frac{1-\beta_p^t(1-\nu)^t}{1-\beta_p(1-\nu)} - \frac{\beta_n^t}{\delta} - 2.$$

The proof of Theorem 3.8 is given in Section 3.9.7. The first condition in Theorem 3.8 can be used to get the following upper bound on the time horizon up to which $\tilde{h}_{EO,t} \equiv h^*$ is possible.

Corollary 3.9. The conditions in Theorem 3.8 above are satisfied only if $t < \frac{\log K_D}{\log(1/\beta_n)}$, where $K_D = \frac{(1-2\delta)(1-r)}{(1-\delta)r} - \frac{\delta}{1-\delta} + 2$.

The above corollary can be obtained by noting that the term $2\beta_p^t(1-\nu)^t$ is upper bounded by 2 since it converges to 1, and any t greater than the value in the corollary violates the first inequality in Theorem 3.8. We derive similar time-varying bias recovery conditions for demographic parity in Theorem 3.11 given in Section 3.9.7.

§ 3.8 Conclusion and Future Directions

In this work, we studied the phenomenon of using fair classification (in particular, equal opportunity constraints) to recover optimal and fair classifiers even from an extremely biased version of the original data. We generalize the result of Blum and Stangl [BS19] in many ways, for arbitrary distributions and arbitrary hypothesis classes, and develop techniques that are flexible and may be of independent interest in studying other fair classification problems and data bias models. Note that our approach based on Blum and Stangl [BS19] does not require knowing, estimating, or correcting the data bias explicitly. Previous work has studied alternate approaches to get around data bias through reweighing and loss adjustment by estimating the extent of data bias [BM21; DB20; WLL21]. Rectifying data bias for fair classification in practice would require the best combination of both these approaches.

Given the flexibility of our technique, it would be interesting to investigate the applicability and limitations of our results to different data bias models [DB20; WLL21; BM21; KL22; Blu+23] and a variety of fairness constraints (e.g., predictive parity [ZDC22b]). A pragmatic future direction is to study the effect of data bias on the sample complexity for fair classification [Don+18; TD23]. One can also attempt to formulate a test for these conditions, which would require estimating the data bias parameters with either a clean validation set or adapting popular approaches in noise estimation literature, such as separability [Men+15], anchor points [Xia+19] and clusterability [ZSL21] for handling general linear-fractional transforms like under-representation and prior probability shifts.

An important question related to time-varying data bias models is to study the case when bias parameters at time t are a function of the model performance at time $t-1$, using ideas such as performative prediction [Per+20; BHK22]

§ 3.9 Proofs

We now present proofs of the results from the previous sections, along with additional results.

§ 3.9.1 Proof of Proposition 3.1

Proof. $\tilde{\eta}(x_1, a) < \tilde{\eta}(x_2, a)$ iff $(P\eta(x_1, a) + Q)(R\eta(x_2, a) + S) < (P\eta(x_2, a) + Q)(R\eta(x_1, a) + S)$, using $R\eta(x, a) + S > 0$, for all $0 \leqslant \eta(x, a) \leqslant 1$. By canceling the common terms, the above inequality holds iff $(PS - QR)\eta(x_1, a) < (PS - QR)\eta(x_2, a)$, or equivalently $\eta(x_1, a) < \eta(x_2, a)$ because $PS - QR > 0$. □

§ 3.9.2 Derivations for Linear Fractional Transforms of Data Bias Models

Proposition 3.6. Consider a biased distribution $\tilde{D}$ obtained from the original distribution D by the following process defined by bias parameters $\beta_p, \beta_n, \nu \in (0, 1)$ [BS19]. A random data point (X, A, Y) from D with $A = 1$ remains unchanged, the points with $A = 0, Y = 1$ have an independent survival probability of β_p, and the points with $A = 0, Y = 0$ have an independent survival probability of β_n. Finally, the surviving points with $A = 0, Y = 1$ keep their class label 1 with probability $1 - \nu$, and it gets flipped to 0 with probability ν. For the privileged group $A = 1$, we have $\tilde{\eta}(x, 1) = \eta(x, 1)$, for all $x \in \mathcal{X}$, whereas for the underprivileged group $A = 0$ we have $\tilde{\eta}(x, 0) = \frac{(1-\nu)\eta(x,0)}{(1-c)\eta(x,0)+c}$, where $c = \frac{\beta_n}{\beta_p}$.

Proof.

$$\begin{aligned}\tilde{\eta}(x, 0) &= \Pr\left(\tilde{Y} = 1 | X = x, A = 0\right) = \frac{(1-\nu)\beta_p\eta(x, 0)}{\beta_p\eta(x, 0) + \beta_n(1 - \eta(x, 0))} \\ &= \frac{(1-\nu)\beta_p\eta(x, 0)}{(\beta_p - \beta_n)\eta(x, 0) + \beta_n} = \frac{(1-\nu)\eta(x, 0)}{(1-c)\eta(x, 0) + c}, \qquad \text{where } c = \frac{\beta_n}{\beta_p}.\end{aligned}$$

Since no bias acts on group $A = 1$, $\tilde{\eta}(x, 1) = \eta(x, 1)$. □

Proposition 3.7. Consider a biased distribution $\tilde{D}$ obtained form the original distribution D by introducing a group dependent label flip rate [DB20; WLL21]: $\epsilon_a^1 = \Pr\left(\tilde{Y} = 0 | Y = 1, A = a\right)$ and $\epsilon_a^0 =$

$\Pr\left(\tilde{Y}=1|Y=0,A=a\right)$, with $0 \leqslant \epsilon_a^1 + \epsilon_a^0 < 1$. Then, the observed labels in $\tilde{D}$ obey the following relationship: $\tilde{y}_i = y_i$ with $1-\epsilon_{a_i}^{\mathbb{I}(y_i=1)}$ probability and $\tilde{y}_i = \neg y_i$ with probability $\epsilon_{a_i}^{\mathbb{I}(y_i=1)}$. In terms of a linear fractional transform:

$$\tilde{\eta}(x,a) = (1-\epsilon_a^1-\epsilon_a^0)\eta(x,a)+\epsilon_a^0$$

Proof.

$$\begin{aligned}
\tilde{\eta}(x,a) &= \Pr\left(\tilde{Y}=1|X=x,A=a\right) \\
&= \frac{\Pr\left(\tilde{Y}=1,Y=0,X=x,A=a\right)+\Pr\left(\tilde{Y}=1,Y=1,X=x,A=a\right)}{\Pr\left(X=x,A=a\right)} \\
&= \Pr\left(\tilde{Y}=1|Y=0,A=a\right)\left(1-\eta(x,a)\right)+\Pr\left(\tilde{Y}=1|Y=1,A=a\right)\eta(x,a) \\
&= (1-\epsilon_a^1-\epsilon_a^0)\eta(x,a)+\epsilon_a^0.
\end{aligned}$$

□

Proposition 3.8. [BM21] Consider a biased distribution $\tilde{D}$ obtained from the original distribution D by introducing a group-dependent prior probability shift such that $\tilde{A}=A$ and

$$\Pr\left(\tilde{X}=x|\tilde{Y}=i,A=a\right)=\Pr\left(X=x|Y=i,A=a\right) \text{ for any } i,a\in\{0,1\},$$

but $\Pr\left(\tilde{Y}=i|A=a\right) \neq \Pr\left(Y=i|A=a\right)$. For this data bias model, we prove that $\tilde{\eta}(x,a) = \dfrac{\eta(x,a)}{(1-\alpha)\eta(x,a)+\alpha}$, where $\alpha = \dfrac{\Pr\left(\tilde{Y}=0|A=a\right)\Pr\left(Y=1|A=a\right)}{\Pr\left(\tilde{Y}=1|A=a\right)\Pr\left(Y=0|A=a\right)}$.

Proof.

$$\begin{aligned}
\tilde{\eta}(x,a) &= \Pr\left(\tilde{Y}=1|\tilde{X}=x,A=a\right) \\
&= \frac{\Pr\left(\tilde{Y}=1,\tilde{X}=x,A=a\right)}{\Pr\left(\tilde{X}=x,A=a\right)}
\end{aligned}$$

Since $\Pr\left(X=x|\tilde{Y}=1,A=a\right) = \Pr\left(X=x|Y=1,A=a\right)$ from the definition of the bias model, we can write:

$$\begin{aligned}
\tilde{\eta}(x,a) &= \frac{\Pr\left(X=x|\tilde{Y}=1,A=a\right)\Pr\left(\tilde{Y}=1|A=a\right)}{\Pr\left(X=x|\tilde{Y}=1,A=a\right)\Pr\left(\tilde{Y}=1|A=a\right)+\Pr\left(X=x|\tilde{Y}=0,A=a\right)\Pr\left(\tilde{Y}=0|A=a\right)} \\
&= \frac{1}{1+\dfrac{\Pr\left(X=x|\tilde{Y}=0,A=a\right)\Pr\left(\tilde{Y}=0|A=a\right)}{\Pr\left(X=x|\tilde{Y}=1,A=a\right)\Pr\left(\tilde{Y}=1|A=a\right)}}.
\end{aligned}$$

Thus,

$$\frac{1}{\tilde{\eta}(x,a)}-1 = \frac{\Pr\left(X=x|\tilde{Y}=0,A=a\right)\Pr\left(\tilde{Y}=0|A=a\right)}{\Pr\left(X=x|\tilde{Y}=1,A=a\right)\Pr\left(\tilde{Y}=1|A=a\right)}.$$

Similarly,

$$\frac{1}{\eta(x,a)}-1 = \frac{\Pr\left(X=x|Y=0,A=a\right)\Pr\left(Y=0|A=a\right)}{\Pr\left(X=x|Y=1,A=a\right)\Pr\left(Y=1|A=a\right)}.$$

Hence,

$$\frac{1}{\tilde{\eta}(x,a)} - 1 = \alpha\left(\frac{1}{\eta(x,a)} - 1\right), \quad \text{where } \alpha = \frac{\Pr\left(\tilde{Y}=0|A=a\right)\Pr\left(Y=1|A=a\right)}{\Pr\left(\tilde{Y}=1|A=a\right)\Pr\left(Y=0|A=a\right)}.$$

In other words,

$$\tilde{\eta}(x,a) = \frac{\eta(x,a)}{(1-\alpha)\eta(x,a)+\alpha}.$$

□

§ 3.9.3 Proofs for Section 3.3.2

Given any classifier $h : \mathcal{X} \times \mathcal{A} \to \mathcal{Y}$, let $\mathrm{TPR}_a(h)$ and $\widetilde{\mathrm{TPR}}_a(h)$ denote its true positive rates conditioned on $A = a$ for distributions D and $\tilde{D}$, respectively. Then

$$\mathrm{TPR}_a(h) = \frac{\Pr\left(h(X,a)=1, A=a, Y=1\right)}{\Pr\left(Y=1, A=a\right)} = \frac{\underset{X|A=a}{\mathrm{E}}\left[h(X,a)\eta(X,a)\right]}{\underset{X|A=a}{\mathrm{E}}\left[\eta(X,a)\right]},$$

and

$$\widetilde{\mathrm{TPR}}_a(h) = \frac{\Pr\left(h(X,a)=1, A=a, \tilde{Y}=1\right)}{\Pr\left(\tilde{Y}=1, A=a\right)} = \frac{\underset{X|A=a}{\mathrm{E}}\left[h(X,a)\tilde{\eta}(X,a)\right]}{\underset{X|A=a}{\mathrm{E}}\left[\tilde{\eta}(X,a)\right]}.$$

We can now prove Proposition 3.2.

Proof. **(Proof of Proposition 3.2)**

$$\begin{aligned}
&\widetilde{\mathrm{TPR}}_a(h) = \Pr\left(h(X,a)=1|\tilde{Y}=1, A=a\right) = \frac{\Pr\left(h(X,a)=1, \tilde{Y}=1|A=a\right)}{\Pr\left(\tilde{Y}=1|A=a\right)} \\
&= \frac{\Pr\left(Y=0|A=a\right)\Pr\left(h(X,a)=1, \tilde{Y}=1|Y=0, A=a\right)}{\Pr\left(\tilde{Y}=1|A=a\right)} \\
&\qquad + \frac{\Pr\left(Y=1|A=a\right)\Pr\left(h(X,a)=1, \tilde{Y}=1|Y=1, A=a\right)}{\Pr\left(\tilde{Y}=1|A=a\right)} \\
&= \frac{\Pr\left(Y=0|A=a\right)\Pr\left(\tilde{Y}=1|h(X,a)=1, Y=0, A=a\right)\Pr\left(h(X,a)=1|Y=0, A=a\right)}{\Pr\left(\tilde{Y}=1|A=a\right)} \\
&\qquad + \frac{\Pr\left(Y=1|A=a\right)\Pr\left(\tilde{Y}=1|h(X,a)=1, Y=1, A=a\right)\Pr\left(h(X,a)=1|Y=1, A=a\right)}{\Pr\left(\tilde{Y}=1|A=a\right)} \\
&= \frac{\Pr\left(Y=0|A=a\right)\Pr\left(\tilde{Y}=1|Y=0, A=a\right)\Pr\left(h(X,a)=1|Y=0, A=a\right)}{\Pr\left(\tilde{Y}=1|A=a\right)} \\
&\qquad + \frac{\Pr\left(Y=1|A=a\right)\Pr\left(\tilde{Y}=1|Y=1, A=a\right)\Pr\left(h(X,a)=1|Y=1, A=a\right)}{\Pr\left(\tilde{Y}=1|A=a\right)} \\
&\qquad\qquad\qquad\qquad \text{because } \tilde{Y} \text{ depends only on } A \text{ and } Y \\
&= \frac{\Pr\left(\tilde{Y}=1, Y=0|A=a\right)}{\Pr\left(\tilde{Y}=1|A=a\right)}\,\mathrm{FPR}_a(h) + \frac{\Pr\left(\tilde{Y}=1, Y=1|A=a\right)}{\Pr\left(\tilde{Y}=1|A=a\right)}\,\mathrm{TPR}_a(h) \\
&= \Pr\left(Y=0|\tilde{Y}=1, A=a\right)\,\mathrm{FPR}_a(h) + \Pr\left(Y=1|\tilde{Y}=1, A=a\right)\,\mathrm{TPR}_a(h).
\end{aligned}$$

Similarly, we show that $\widetilde{\mathrm{TNR}}_a(h)$ can be written as a linear combination of $\mathrm{TNR}_a(h)$ and $\mathrm{FNR}_a(h)$.

$$
\begin{aligned}
\widetilde{\mathrm{TNR}}_a(h) &= \Pr\left(h(X,a)=0|\tilde{Y}=0, A=a\right) = \frac{\Pr\left(h(X,a)=0,\tilde{Y}=0|A=a\right)}{\Pr\left(\tilde{Y}=0|A=a\right)} \\
&= \frac{\Pr\left(Y=0|A=a\right)\Pr\left(h(X,a)=0,\tilde{Y}=0|Y=0,A=a\right)}{\Pr\left(\tilde{Y}=0|A=a\right)} \\
&\qquad + \frac{\Pr\left(Y=1|A=a\right)\Pr\left(h(X,a)=0,\tilde{Y}=0|Y=1,A=a\right)}{\Pr\left(\tilde{Y}=0|A=a\right)} \\
&= \frac{\Pr\left(Y=0|A=a\right)\Pr\left(\tilde{Y}=0|h(X,a)=0,Y=0,A=a\right)\Pr\left(h(X,a)=0|Y=0,A=a\right)}{\Pr\left(\tilde{Y}=0|A=a\right)} \\
&\qquad + \frac{\Pr\left(Y=1|A=a\right)\Pr\left(\tilde{Y}=0|h(X,a)=0,Y=1,A=a\right)\Pr\left(h(X,a)=0|Y=1,A=a\right)}{\Pr\left(\tilde{Y}=0|A=a\right)} \\
&= \frac{\Pr\left(Y=0|A=a\right)\Pr\left(\tilde{Y}=0|Y=0,A=a\right)\Pr\left(h(X,a)=0|Y=0,A=a\right)}{\Pr\left(\tilde{Y}=0|A=a\right)} \\
&\qquad + \frac{\Pr\left(Y=1|A=a\right)\Pr\left(\tilde{Y}=0|Y=1,A=a\right)\Pr\left(h(X,a)=0|Y=1,A=a\right)}{\Pr\left(\tilde{Y}=0|A=a\right)} \\
&\qquad\qquad\qquad\qquad \text{because } \tilde{Y} \text{ depends only on } A \text{ and } Y \\
&= \frac{\Pr\left(\tilde{Y}=0,Y=0|A=a\right)}{\Pr\left(\tilde{Y}=0|A=a\right)}\,\mathrm{TNR}_a(h) + \frac{\Pr\left(\tilde{Y}=0,Y=1|A=a\right)}{\Pr\left(\tilde{Y}=0|A=a\right)}\,\mathrm{FNR}_a(h) \\
&= \Pr\left(Y=0|\tilde{Y}=0,A=a\right)\,\mathrm{TNR}_a(h) + \Pr\left(Y=1|\tilde{Y}=0,A=a\right)\,\mathrm{FNR}_a(h).
\end{aligned}
$$

□

We now prove a similar result for the positive rate. For a fixed group $a \in \mathcal{A}$, define $\mathrm{PR}_a(h) := \Pr(h(X,a)=1 \mid A=a)$ and $\widetilde{\mathrm{PR}}_a(h) := \Pr(h(X,a)=1 \mid A=a) = \sum_{\tilde{y}\in\tilde{\mathcal{Y}}} \Pr\big(h(X,a)=1,\tilde{Y}=\tilde{y} \mid A=a\big)$, where $\mathrm{PR}_a(h)$ is computed under the original distribution D, and $\widetilde{\mathrm{PR}}_a(h)$ is computed under the biased distribution $\widetilde{D}$.

Proposition 3.9. Let D be any distribution on $\mathcal{X}\times\mathcal{A}\times\mathcal{Y}$ and let $\widetilde{D}$ be its biased version defined using any of the data bias models from Subsection 3.3.1. Let h be any hypothesis, and let

$$\mathrm{PR}_a(h) := \Pr(h(X,a)=1 \mid A=a) \quad \text{and} \quad \widetilde{\mathrm{PR}}_a(h) := \Pr(h(X,a)=1 \mid A=a)$$

denote its group-wise positive rates according to D and $\widetilde{D}$, respectively, conditioned on the group $A=a$. Then

$$\widetilde{\mathrm{PR}}_a(h) = \Pr\Big(\tilde{Y}=1 \mid A=a\Big)\,\widetilde{\mathrm{TPR}}_a(h) + \Pr\Big(\tilde{Y}=0 \mid A=a\Big)\,\widetilde{\mathrm{FPR}}_a(h),$$

where

$$\widetilde{\mathrm{FPR}}_a(h) := \Pr\Big(h(X,a)=1 \mid \tilde{Y}=0, A=a\Big).$$

Equivalently, using Proposition 3.2,

$$\widetilde{\mathrm{PR}}_a(h) = \Pr(Y=1 \mid A=a)\,\mathrm{TPR}_a(h) + \Pr(Y=0 \mid A=a)\,\mathrm{FPR}_a(h),$$

where the probabilities on the right-hand side are taken under $\widetilde{D}$.

Proof. We first work with the biased distribution $\widetilde{D}$. By definition,

$$\widetilde{\mathrm{PR}}_a(h) = \Pr(h(X,a)=1 \mid A=a).$$

We can now use the law of total probability with respect to $\widetilde{Y}$

$$\begin{aligned}
\widetilde{\mathrm{PR}}_a(h) &= \Pr\Big(h(X,a)=1, \widetilde{Y}=1 \mid A=a\Big) + \Pr\Big(h(X,a)=1, \widetilde{Y}=0 \mid A=a\Big) \\
&= \Pr\Big(\widetilde{Y}=1 \mid A=a\Big)\Pr\Big(h(X,a)=1 \mid \widetilde{Y}=1, A=a\Big) \\
&\qquad + \Pr\Big(\widetilde{Y}=0 \mid A=a\Big)\Pr\Big(h(X,a)=1 \mid \widetilde{Y}=0, A=a\Big) \\
&= \Pr\Big(\widetilde{Y}=1 \mid A=a\Big)\widetilde{\mathrm{TPR}}_a(h) + \Pr\Big(\widetilde{Y}=0 \mid A=a\Big)\widetilde{\mathrm{FPR}}_a(h).
\end{aligned}$$

This proves the first identity.

Now, by Proposition 3.2,

$$\widetilde{\mathrm{TPR}}_a(h) = \Pr\Big(Y=0 \mid \widetilde{Y}=1, A=a\Big)\,\mathrm{FPR}_a(h) + \Pr\Big(Y=1 \mid \widetilde{Y}=1, A=a\Big)\,\mathrm{TPR}_a(h),$$

and

$$\widetilde{\mathrm{TNR}}_a(h) = \Pr\Big(Y=0 \mid \widetilde{Y}=0, A=a\Big)\,\mathrm{TNR}_a(h) + \Pr\Big(Y=1 \mid \widetilde{Y}=0, A=a\Big)\,\mathrm{FNR}_a(h).$$

Since $\widetilde{\mathrm{FPR}}_a(h) = 1 - \widetilde{\mathrm{TNR}}_a(h)$, $\mathrm{FPR}_a(h) = 1 - \mathrm{TNR}_a(h)$, and $\mathrm{TPR}_a(h) = 1 - \mathrm{FNR}_a(h)$, we obtain

$$\widetilde{\mathrm{FPR}}_a(h) = \Pr\Big(Y=0 \mid \widetilde{Y}=0, A=a\Big)\,\mathrm{FPR}_a(h) + \Pr\Big(Y=1 \mid \widetilde{Y}=0, A=a\Big)\,\mathrm{TPR}_a(h).$$

Substituting the above expressions into the first identity gives

$$\begin{aligned}
\widetilde{\mathrm{PR}}_a(h) &= \Pr\Big(\widetilde{Y}=1 \mid A=a\Big)\Big[\Pr\Big(Y=0 \mid \widetilde{Y}=1, A=a\Big)\,\mathrm{FPR}_a(h) + \Pr\Big(Y=1 \mid \widetilde{Y}=1, A=a\Big)\,\mathrm{TPR}_a(h)\Big] \\
&\quad + \Pr\Big(\widetilde{Y}=0 \mid A=a\Big)\Big[\Pr\Big(Y=0 \mid \widetilde{Y}=0, A=a\Big)\,\mathrm{FPR}_a(h) + \Pr\Big(Y=1 \mid \widetilde{Y}=0, A=a\Big)\,\mathrm{TPR}_a(h)\Big] \\
&= \Big[\Pr\Big(\widetilde{Y}=1 \mid A=a\Big)\Pr\Big(Y=1 \mid \widetilde{Y}=1, A=a\Big) + \Pr\Big(\widetilde{Y}=0 \mid A=a\Big)\Pr\Big(Y=1 \mid \widetilde{Y}=0, A=a\Big)\Big]\mathrm{TPR}_a(h) \\
&\quad + \Big[\Pr\Big(\widetilde{Y}=1 \mid A=a\Big)\Pr\Big(Y=0 \mid \widetilde{Y}=1, A=a\Big) + \Pr\Big(\widetilde{Y}=0 \mid A=a\Big)\Pr\Big(Y=0 \mid \widetilde{Y}=0, A=a\Big)\Big]\mathrm{FPR}_a(h).
\end{aligned}$$

Simplifying using the law of total probability:

$$\widetilde{\mathrm{PR}}_a(h) = \Pr(Y=1 \mid A=a)\,\mathrm{TPR}_a(h) + \Pr(Y=0 \mid A=a)\,\mathrm{FPR}_a(h),$$

where all probabilities are taken under $\widetilde{D}$. □

Corollary 3.10. Let D be any distribution on $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$ and let $\widetilde{D}$ be its biased version as in Example 3.1, with bias parameters $\beta_p, \beta_n, \nu \in (0,1)$. Given any hypothesis h, let

$$\mathrm{PR}_a(h) := \Pr(h(X,a)=1 \mid A=a)$$

denote its group-wise positive rate on D, and let

$$\widetilde{\mathrm{PR}}_a(h) := \Pr(h(X, a) = 1 \mid A = a)$$

denote its group-wise positive rate on $\widetilde{D}$.

Further, let

$$\pi_{1|a} := \Pr(Y = 1 \mid A = a), \qquad \pi_{0|a} := \Pr(Y = 0 \mid A = a),$$

and let $\mathrm{TPR}_a(h)$ and $\mathrm{FPR}_a(h)$ be computed with respect to the original distribution D.

Then

$$\widetilde{\mathrm{PR}}_1(h) = \mathrm{PR}_1(h),$$

and

$$\widetilde{\mathrm{PR}}_0(h) = \frac{\beta_p \pi_{1|0}\, \mathrm{TPR}_0(h) + \beta_n \pi_{0|0}\, \mathrm{FPR}_0(h)}{\beta_p \pi_{1|0} + \beta_n \pi_{0|0}}.$$

Proof. For group $A = 1$, the distribution is unchanged under Example 3.1, hence

$$\widetilde{\mathrm{PR}}_1(h) = \mathrm{PR}_1(h).$$

Now consider group $A = 0$. By the law of total probability,

$$\widetilde{\mathrm{PR}}_0(h) = \Pr(h(X,0) = 1 \mid Y = 1, A = 0) \Pr_{\widetilde{D}}(Y = 1 \mid A = 0) + \Pr(h(X,0) = 1 \mid Y = 0, A = 0) \Pr_{\widetilde{D}}(Y = 0 \mid A = 0).$$

Since the sampling mechanism in Example 3.1 depends only on (Y, A), the conditional distribution of X given (Y, A) is unchanged, and therefore

$$\Pr(h(X,0) = 1 \mid Y = 1, A = 0) = \mathrm{TPR}_0(h), \qquad \Pr(h(X,0) = 1 \mid Y = 0, A = 0) = \mathrm{FPR}_0(h).$$

Moreover,

$$\Pr_{\widetilde{D}}(Y = 1 \mid A = 0) = \frac{\beta_p \pi_{1|0}}{\beta_p \pi_{1|0} + \beta_n \pi_{0|0}}, \qquad \Pr_{\widetilde{D}}(Y = 0 \mid A = 0) = \frac{\beta_n \pi_{0|0}}{\beta_p \pi_{1|0} + \beta_n \pi_{0|0}}.$$

Substituting the above identities gives

$$\widetilde{\mathrm{PR}}_0(h) = \frac{\beta_p \pi_{1|0}\, \mathrm{TPR}_0(h) + \beta_n \pi_{0|0}\, \mathrm{FPR}_0(h)}{\beta_p \pi_{1|0} + \beta_n \pi_{0|0}}.$$

□

Proof. **(Proof of Proposition 3.3)** The proof of this Proposition follows the arguments from the proof of Proposition 2.3 in [Chz+19]. We define $\tilde{h}_{EO} = \operatorname*{argmax}_{h \in \widetilde{\mathcal{H}}_{\text{fair, EO}}} \Pr\left(h(X, A) = \tilde{Y}\right)$.

$$
\begin{aligned}
\tilde{h}_{EO} &= \operatorname*{argmax}_{h\in\mathcal{H}} \min_{\lambda\in\mathbb{R}} \Pr\left(h(X,A)=\tilde{Y}\right) + \lambda \Pr\left(h(X,A)=1|\tilde{Y}=1,A=0\right) - \lambda \Pr\left(h(X,A)=1|\tilde{Y}=1,A=1\right) \\
&= \operatorname*{argmax}_{h\in\mathcal{H}} \min_{\lambda\in\mathbb{R}} \Pr\left(\tilde{Y}=0\right) + \sum_{a=0}^{1} \Pr\left(A=a\right) \mathop{E}_{X|A=a}\left[h(X,a)(2\tilde{\eta}(X,a)-1)\right] \\
&+ \frac{(-1)^{\mathbb{I}(a=1)}\,\lambda}{\mathop{E}_{X|A=a}\left[\tilde{\eta}(X,a)\right]} \mathop{E}_{X|A=a}\left[h(X,a)\tilde{\eta}(X,a)\right] \\
&= \operatorname*{argmax}_{h\in\mathcal{H}} \min_{\lambda\in\mathbb{R}} \sum_{a=0}^{1} \Pr\left(A=a\right) \mathop{E}_{X|A=a}\left[h(X,a)\left(\left(2+\frac{(-1)^{\mathbb{I}(a=1)}\lambda}{\Pr\left(\tilde{Y}=1,A=a\right)}\right)\tilde{\eta}(X,a)-1\right)\right] \\
&= \operatorname*{argmax}_{h\in\mathcal{H}} \min_{\lambda\in\mathbb{R}} L(h,\lambda)
\end{aligned}
$$

By weak duality, $\max_{h\in\mathcal{H}} \min_{\lambda\in\mathbb{R}} L(h,\lambda) \leqslant \min_{\lambda\in\mathbb{R}} \max_{h\in\mathcal{H}} L(h,\lambda)$. For a fixed λ, solving the dual objective $\min_{\lambda\in\mathbb{R}} \max_{h\in\mathcal{H}} L(h,\lambda)$, we get:

$$
\begin{aligned}
\tilde{h}_{EO,\lambda}(x,a) &= \mathbb{I}\left(\tilde{\eta}(x,a) \geqslant \frac{1}{2+\frac{(-1)^{\mathbb{I}(a=1)}\lambda}{\Pr\left(\tilde{Y}=1,A=a\right)}}\right) \\
&= \begin{cases} \mathbb{I}\left(\dfrac{(1-\nu)\eta(x,0)}{(1-c)\eta(x,0)+c} \geqslant \dfrac{1}{2+\frac{\lambda}{\Pr\left(\tilde{Y}=1,A=0\right)}}\right), & \text{for } a=0 \qquad \text{(Example 3.1)} \\ \mathbb{I}\left(\eta(x,1) \geqslant \dfrac{1}{2-\frac{\lambda}{\Pr\left(Y=1,A=1\right)}}\right), & \text{for } a=1 \end{cases} \\
&= \begin{cases} \mathbb{I}\left(\eta(x,0) \geqslant \dfrac{c}{1-2\nu+c+\frac{\lambda(1-\nu)}{\Pr\left(\tilde{Y}=1,A=0\right)}}\right), & \text{for } a=0 \\ \mathbb{I}\left(\eta(x,1) \geqslant \dfrac{1}{2-\frac{\lambda}{\Pr\left(Y=1,A=1\right)}}\right), & \text{for } a=1 \end{cases}
\end{aligned}
$$

Let $\Pr(A=0) = r$ and $\Pr(A=1) = 1-r$, and let $\Pr(Y=1|A=0) = \Pr(Y=1|A=1) = q$. Then $\Pr(Y=1,A=1) = \Pr(A=1)\Pr(Y=1|A=1) = (1-r)q$ and

$$
\begin{aligned}
\Pr\left(\tilde{Y}=1,A=0\right) &= \Pr(A=0)\Pr(Y=1|A=0)\Pr\left(\tilde{Y}=1|Y=1,A=0\right) \\
&\quad + \Pr(A=0)\Pr(Y=0|A=0)\Pr\left(\tilde{Y}=1|Y=0,A=0\right) \\
&= rq\beta_p(1-\nu).
\end{aligned}
$$

Therefore,

$$\tilde{h}_{EO,\lambda}(x,a) = \begin{cases} \mathbb{I}\left(\eta(x,0) \geqslant \dfrac{c}{1-2\nu+c+\dfrac{\lambda}{\beta_p rq}}\right), & \text{for } a=0 \\ \mathbb{I}\left(\eta(x,1) \geqslant \dfrac{1}{2-\dfrac{\lambda}{(1-r)q}}\right), & \text{for } a=1 \end{cases}.$$

Equivalently, we can also write

$$\tilde{h}_{EO,\lambda}(x,a) = \begin{cases} \mathbb{I}\left(\eta(x,0) \geqslant \dfrac{1}{1+\dfrac{1-2\nu}{c}+\dfrac{\lambda}{\beta_n rq}}\right), & \text{for } a=0 \\ \mathbb{I}\left(\eta(x,1) \geqslant \dfrac{1}{2-\dfrac{\lambda}{(1-r)q}}\right), & \text{for } a=1. \end{cases}.$$

Substituting $\tilde{h}_\lambda$ back into the dual objective reduces the positive part of the objective to a one-dimensional problem in λ. Note that the mapping $\lambda \mapsto \underset{X|A=a}{E}\left[\left(\left(2+\dfrac{(-1)^{\mathbb{I}(a=1)}\lambda}{\Pr(\tilde{Y}=1,A=a)}\right)\tilde{\eta}(X,a)-1\right)_+\right]$ is convex in λ. Due to Assumption 2.1, the random variable $\eta(x,a)$ (and its linear-fractional mapping $\tilde{\eta}(x,a)$) has no atoms, which assures differentiability. Since the dual objective is convex in λ, any dual minimizer λ^* satisfies the following first-order condition:

$$\frac{\underset{X|A=1}{E}\left[\tilde{\eta}(X,1)\mathbb{I}\left(1 \leqslant \tilde{\eta}(X,1)(2-\frac{\lambda^*}{\Pr(\tilde{Y}=1,A=1)})\right)\right]}{\Pr\left(\tilde{Y}=1|A=1\right)} = \frac{\underset{X|A=0}{E}\left[\tilde{\eta}(X,0)\mathbb{I}\left(1 \leqslant \tilde{\eta}(X,0)(2-\frac{\lambda^*}{\Pr(\tilde{Y}=1,A=0)})\right)\right]}{\Pr\left(\tilde{Y}=1|A=0\right)}.$$

Note that this is exactly the TPR equality condition we seek, and hence $\tilde{h}_{\lambda^*}$ is a feasible solution for our original problem. This means that $\Pr\left(\tilde{h}_{\lambda^*}=\tilde{Y}\right) \leqslant \max_{h\in\mathcal{H}}\min_{\lambda\in\mathbb{R}} L(h,\lambda)$. However, $(\lambda^*, \tilde{h}_{\lambda^*})$ also solves the dual problem, and the constraint vanishes at feasibility:

$$\Pr\left(\tilde{h}_{\lambda^*}=\tilde{Y}\right) \leqslant \max_{h\in\mathcal{H}}\min_{\lambda\in\mathbb{R}} L(h,\lambda) \leqslant \min_{\lambda\in\mathbb{R}}\max_{h\in\mathcal{H}} L(h,\lambda) = \Pr\left(\tilde{h}_{\lambda^*}=\tilde{Y}\right).$$

where the last equality holds since the constraint term vanishes at h_{λ^*}. Hence, equality holds throughout and $\tilde{h}_{\lambda^*}$ (or $\tilde{h}_{EO}$) is the optimal EO-fair classifier.

□

We can derive similar results for demographic parity [MW18; Chz+19], whose proof follows exactly the same arguments as Proposition 3.3.

Proposition 3.10. For any distribution D and its biased version $\tilde{D}$ described above, let $\tilde{h}_{DP}$ be a classifier of the maximum accuracy among all binary classifiers that satisfy demographic parity

on $\tilde{D}$. Then there exists $\lambda^* \in \mathbb{R}$ such that

$$\tilde{h}_{DP}(x,a) = \begin{cases} \mathbb{I}\left(\eta(x,0) \geqslant \dfrac{c(r-\lambda^*)}{2r(1-\nu)-(1-c)(r-\lambda^*)}\right), & \text{for } a=0 \\ \mathbb{I}\left(\eta(x,1) \geqslant \dfrac{1}{2}+\dfrac{\lambda^*}{2(1-r)}\right), & \text{for } a=1 \end{cases}$$

§ 3.9.4 Proofs for Section 3.4.1

For the distribution D described in Section 3.4.1, the following result holds:

Lemma 3.1. Let $h : \mathcal{X} \times \mathcal{A} \to \{0,1\}$ be a deterministic classifier. Let $Y|X=x, A=a$ be a random variable that takes value $h(x,a)$ with probability $1-\delta(x,a)$ and $\neg h(x,a)$ with probability $\delta(x,a)$, for some $\delta(x,a) \leqslant \delta < 1/2$. Let $\eta(x,a) = \Pr(Y=1|X=x,A=a)$ and let $g(x,a)$ be a threshold classifier given by $g(x,a) = \mathbb{I}(\eta(x,a) \geqslant t_a)$, using group-dependent thresholds t_0 and t_1, respectively. If $t_0, t_1 \in (\delta, 1-\delta)$ then $g \equiv h \equiv h^*$, where h^* is the Bayes optimal classifier for the joint distribution on $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$.

Proof. It is folklore that the (group-aware) Bayes optimal classifier is given by $h^*(x,a) = \mathbb{I}(\eta(x,a) \geqslant 1/2)$ [ZDC22b]. One of the ways by which we can recover the Bayes Optimal classifier with the biased distribution $\tilde{D}$ is when the Bayes optimal classifier on the original distribution D does not change around the vicinity of $1/2$. Consider the following two cases for $h(x,a)$. If $h(x,a)=0$, then $Y|X=x,A=a$ takes value 0 with probability $\delta(x,a) \leqslant \delta$, and hence, $\mathbb{I}(\eta(x,a) \geqslant t_a) = \mathbb{I}(\delta(x,a) \geqslant t_a) = \mathbb{I}(\delta(x,a) \geqslant 1/2) = 0$, for $\delta(x,a) \leqslant \delta < 1/2$ and $t_a \in (\delta, 1-\delta)$. On the other hand, if $h(x,a)=1$, then $Y|X=x,A=a$ takes value 1 with probability $1-\delta(x,a) \geqslant 1-\delta$, and hence, $\mathbb{I}(\eta(x,a) \geqslant t_a) = \mathbb{I}(1-\delta(x,a) \geqslant t_a) = \mathbb{I}(1-\delta(x,a) \geqslant 1/2) = 1$, for $\delta(x,a) \leqslant \delta < 1/2$ and $t_a \in (\delta, 1-\delta)$. Therefore, $g \equiv h \equiv h^*$. □

We now describe the proof of Theorem 3.4. The proof for Theorem 3.3 is clubbed with the proof of Theorem 3.2 later in this section.

Proof. **(Proof of Theorem 3.4)** We use the characterization optimal demographic parity classifier obtained in Proposition 3.10 for $\tilde{D}$, i.e., there exists $\lambda^* \in \mathbb{R}$ such that

$$\tilde{h}_{DP}(x,a) = \begin{cases} \mathbb{I}\left(\eta(x,0) \geqslant \dfrac{c(r-\lambda^*)}{2r(1-\nu)-(1-c)(r-\lambda^*)}\right), & \text{for } a=0 \\ \mathbb{I}\left(\eta(x,1) \geqslant \dfrac{1}{2}+\dfrac{\lambda^*}{2(1-r)}\right), & \text{for } a=1. \end{cases}$$

Since we assume $c=1$ ($\beta_p = \beta_n$), we can write:

$$\tilde{h}_{DP}(x,a) = \begin{cases} \mathbb{I}\left(\eta(x,0) \geqslant \dfrac{(r-\lambda^*)}{2r(1-\nu)}\right), & \text{for } a=0 \\ \mathbb{I}\left(\eta(x,1) \geqslant \dfrac{1}{2}+\dfrac{\lambda^*}{2(1-r)}\right), & \text{for } a=1. \end{cases}$$

Step 1: If the given conditions on the bias parameters hold, then there exists a $\lambda \in \mathbb{R}$ such that $h_\lambda = h^*$. From Lemma 3.1, we know that the threshold on $\eta(x,1)$ lies in the interval $(\delta, 1-\delta)$ if and only if $\delta < \dfrac{1}{2}+\dfrac{\lambda^*}{2(1-r)} < 1-\delta$, or equivalently, $(2\delta-1)(1-r) < \lambda^* < (1-2\delta)(1-r)$. Similarly, the threshold on $\eta(x,0)$ lies in the interval $(\delta, 1-\delta)$ if and only if $\delta < \dfrac{(r-\lambda^*)}{2r(1-\nu)} < 1-\delta$, or equivalently,

$$r(1-2(1-\delta)(1-\nu)) < \lambda^* < r(1-2\delta(1-\nu)).$$

There exists λ^* that simultaneously satisfies both sets of constraints given above if and only if

$$r\,(1-2(1-\delta)(1-\nu)) < (1-2\delta)(1-r) \text{ and } (2\delta-1)(1-r) < r\,(1-2\delta(1-\nu))\,.$$

The above conditions can be simplified as

$$(1-2\delta)(1-r) - r\,(1-2(1-\delta)(1-\nu)) > 0 \text{ and } (1-2\delta)(1-r) + r\,(1-2\delta(1-\nu)) > 0.$$

Step 2: If the given conditions on the bias parameters hold, then there cannot exist any $\lambda \in \mathbb{R}$ such that both the group-wise thresholds applied to $\eta(x,a)$ in h_λ are at most δ, or both the thresholds are at least $1-\delta$. (Thresholds lying on the same side) The threshold on $\eta(x,1)$ is at most δ if and only if $\frac{1}{2} + \frac{\lambda}{2(1-r)} \leqslant \delta$, or equivalently, $\lambda \leqslant (2\delta-1)(1-r)$. Similarly, the threshold on $\eta(x,0)$ is at most δ if and only if $\frac{(r-\lambda)}{2r(1-\nu)} \leqslant \delta$, or equivalently, $\lambda \geqslant r\,(1-2\delta(1-\nu))$. If both were to hold simultaneously, then $r\,(1-2\delta(1-\nu)) \leqslant (2\delta-1)(1-r)$, or equivalently, $(1-2\delta)(1-r) + r\,(1-2\delta(1-\nu)) \leqslant 0$, which would violate the second condition in our theorem.

On the other hand, the threshold on $\eta(x,1)$ is at least $1-\delta$ if and only if $\frac{1}{2} + \frac{\lambda}{2(1-r)} \geqslant 1-\delta$, or equivalently, $\lambda \geqslant (1-2\delta)(1-r)$. Similarly, the threshold on $\eta(x,0)$ is at least $\frac{(r-\lambda)}{2r(1-\nu)} \geqslant 1-\delta$, or equivalently, $\lambda \leqslant r\,(1-2(1-\delta)(1-\nu))$. If both were to hold simultaneously, then $(1-2\delta)(1-r) \leqslant r\,(1-2(1-\delta)(1-\nu))$, or equivalently, $(1-2\delta)(1-r) - r\,(1-2(1-\delta)(1-\nu)) \leqslant 0$, which would violate the first condition in our theorem.

Step 3: For any λ, if the group-wise thresholds on $\eta(x,a)$ in h_λ have one of them at most δ and another at least $1-\delta$, then h_λ cannot satisfy demographic parity unless $h_\lambda = h^*$. (Thresholds lying on opposite sides) We know that h^* satisfies demographic parity. Let $\mathrm{PR}_a(h_\lambda) = \Pr(h_\lambda = 1 | A = a)$ (group positive rate). For demographic parity, we require $\mathrm{PR}_0 = \mathrm{PR}_1$. Suppose the threshold of h_λ on $\eta(x,0)$(call it t_0) and the threshold on $\eta(x,1)$(call it t_1) are separated by either δ or $1-\delta$. WLOG assume $t_0 < t_1$. Suppose $t_0 \leqslant \delta \leqslant t_1$. Since there is no input (x,a) with $\eta(x,a) \in (\delta, 1-\delta)$, we get $\mathrm{PR}_0(\tilde{h}_\lambda) \geqslant \mathrm{PR}_0(h^*)$ and $\mathrm{PR}_1(h^*) \geqslant \mathrm{PR}_1(\tilde{h}_\lambda)$. Since h^* satisfies demographic parity on the original distribution D, we have $\mathrm{PR}_0(h^*) = \mathrm{PR}_1(h^*)$. If h_λ satisfies demographic parity on the biased distribution $\tilde{D}$, we have $\widetilde{\mathrm{PR}}_0(h_\lambda) = \widetilde{\mathrm{PR}}_1(h_\lambda)$, and since we assume that $\beta_p = \beta_n$, $\widetilde{\mathrm{PR}}_a(h_\lambda) = \mathrm{PR}_a(h_\lambda)$ due to Corollary 3.10. It implies $\mathrm{PR}_0(h_\lambda) = \mathrm{PR}_1(h_\lambda)$. Combining the two observations above, we get $\mathrm{PR}_0(h_\lambda) = \mathrm{PR}_0(h^*) = \mathrm{PR}_1(h^*) = \mathrm{PR}_1(h_\lambda)$, which is possible only if $h_\lambda \equiv h^*$ (as both are group-wise threshold classifiers). The other case $t_0 \leqslant 1-\delta \leqslant t_1$ can be argued similarly.

□

Now, we give a complete proof of Theorem 3.2. Note that Theorem 3.2 extends the sufficiency direction in Theorem 3.3 to Massart noise. Theorem 3.3 proves necessity and sufficiency in the constant i.i.d. noise case when we have $\delta(x,a) = \delta$, for all $(x,a) \in \mathcal{X} \times \mathcal{A}$. Thus, Theorem 3.2 is a stronger statement than Theorem 3.3 (the original recovery theorem of Blum & Stangl [BS19]) while having the same necessary and sufficient conditions on the data and bias parameters.

Proof. **(Proof of Theorem 3.2)** Let

$$\tilde{h}_\lambda(x,a) = \begin{cases} \mathbb{I}\left(\eta(x,0) \geqslant \dfrac{1}{1+\dfrac{1-2\nu}{c}+\dfrac{\lambda}{\beta_n r q}}\right), & \text{for } a=0 \\ \mathbb{I}\left(\eta(x,1) \geqslant \dfrac{1}{2-\dfrac{\lambda}{(1-r)q}}\right), & \text{for } a=1. \end{cases}$$

From the characterization of the optimal equal opportunity classifier shown earlier (Proposition 2.2), there exists some optimal $\lambda^* \in \mathbb{R}$ such that $\tilde{h}_{EO} = \tilde{h}_{\lambda^*}$. We will show that if the conditions in Theorem 4.1 of Blum-Stangl hold, then the only possible choices left for $\lambda^* \in \mathbb{R}$ are the ones that give $\tilde{h}_{\lambda^*} = h^*$.

Step 1: If the given conditions on the bias parameters hold, then there exists a $\lambda \in \mathbb{R}$ such that $\tilde{h}_\lambda = h^*$. The threshold on $\eta(x,1)$ lies in the interval $(\delta, 1-\delta)$ if and only if $\dfrac{1}{1-\delta} < 2 - \dfrac{\lambda}{(1-r)q} < \dfrac{1}{\delta}$, or equivalently, $\dfrac{-(1-2\delta)(1-r)q}{\delta} < \lambda < \dfrac{(1-2\delta)(1-r)q}{1-\delta}$. Similarly, the threshold on $\eta(x,0)$ lies in the interval $(\delta, 1-\delta)$ if and only if $\dfrac{1}{1-\delta} < 1 + \dfrac{1-2\nu}{c} + \dfrac{\lambda}{\beta_n r q} < \dfrac{1}{\delta}$, or equivalently,

$$\beta_n r q\left(\frac{\delta}{1-\delta} - \frac{1-2\nu}{c}\right) < \lambda < \beta_n r q\left(\frac{1-\delta}{\delta} - \frac{1-2\nu}{c}\right).$$

There exists λ that simultaneously satisfies both sets of constraints given above if and only if

$$\beta_n r q\left(\frac{\delta}{1-\delta} - \frac{1-2\nu}{c}\right) < \frac{(1-2\delta)(1-r)q}{1-\delta} \quad \text{and} \quad \frac{-(1-2\delta)(1-r)q}{\delta} < \beta_n r q\left(\frac{1-\delta}{\delta} - \frac{1-2\nu}{c}\right).$$

Using $c = \beta_n/\beta_p$, the above conditions can be rewritten as

$$\begin{aligned} &(1-r)(1-2\delta) + r\left((1-\delta)\beta_p(1-2\nu) - \delta\beta_n\right) > 0 \quad \text{and} \\ &(1-r)(1-2\delta) + r\left((1-\delta)\beta_n - \delta\beta_p(1-2\nu)\right) > 0. \end{aligned}$$

Step 2: If the given conditions on the bias parameters hold, then there cannot exist any $\lambda \in \mathbb{R}$ such that both the group-wise thresholds applied to $\eta(x,a)$ in h_λ are at most δ, or both the thresholds are at least $1-\delta$. (Thresholds lying on the same side) The threshold on $\eta(x,1)$ is at most δ if and only if $2 - \dfrac{\lambda}{(1-r)q} \geqslant \dfrac{1}{\delta}$, or equivalently, $\dfrac{-(1-2\delta)(1-r)q}{\delta} \geqslant \lambda$. Similarly, the threshold on $\eta(x,0)$ is at most δ if and only if $1 + \dfrac{1-2\nu}{c} + \dfrac{\lambda}{\beta_n r q} \geqslant \dfrac{1}{\delta}$, or equivalently, $\lambda \geqslant \beta_n r q\left(\dfrac{1-\delta}{\delta} - \dfrac{1-2\nu}{c}\right)$. If both were to hold simultaneously, then $\dfrac{-(1-2\delta)(1-r)q}{\delta} \geqslant \beta_n r q\left(\dfrac{1-\delta}{\delta} - \dfrac{1-2\nu}{c}\right)$, or equivalently, $(1-r)(1-2\delta) + r\left((1-\delta)\beta_n - \delta\beta_p(1-2\nu)\right) \leqslant 0$, which would violate the second condition in Theorem 4.1 of Blum-Stangl.

On the other hand, the threshold on $\eta(x,1)$ is at least interval $1-\delta$ if and only if $\dfrac{1}{1-\delta} \geqslant 2 - \dfrac{\lambda}{(1-r)q}$, or equivalently, $\lambda \geqslant \dfrac{(1-2\delta)(1-r)q}{1-\delta}$. Similarly, the threshold on $\eta(x,0)$ is at least $1-\delta$ if and only if $\dfrac{1}{1-\delta} \geqslant 1 + \dfrac{1-2\nu}{c} + \dfrac{\lambda}{\beta_n r q}$, or equivalently, $\beta_n r q\left(\dfrac{\delta}{1-\delta} - \dfrac{1-2\nu}{c}\right) \geqslant \lambda$. If both were to hold

simultaneously, then $\beta_n rq\left(\frac{\delta}{1-\delta}-\frac{1-2\nu}{c}\right) \geqslant \frac{(1-2\delta)(1-r)q}{1-\delta}$, or equivalently, $(1-r)(1-2\delta)+r\left((1-\delta)\beta_p(1-2\nu)-\delta\beta_n\right) \leqslant 0$, which would violate the first condition in Theorem 4.1 of Blum-Stangl.

Step 3: For any λ, if the group-wise thresholds on $\eta(x,a)$ in h_λ have one of them at most δ and another at least $1-\delta$, then h_λ cannot satisfy equal opportunity unless $h_\lambda = h^*$. (Thresholds lying on opposite sides) We know that h^* satisfies equal opportunity. Suppose the threshold of h_λ on $\eta(x,0)$(call it t_0) and the threshold on $\eta(x,1)$(call it t_1) are separated by either δ or $1-\delta$. WLOG assume $t_0 < t_1$. Suppose $t_0 \leqslant \delta \leqslant t_1$. Since there is no input (x,a) with $\eta(x,a) \in (\delta, 1-\delta)$, we get $\mathrm{TPR}_0(h_\lambda) \geqslant \mathrm{TPR}_0(h^*)$ and $\mathrm{TPR}_1(h^*) \geqslant \mathrm{TPR}_1(h_\lambda)$. Since h^* satisfies equal opportunity on the original distribution D, we have $\mathrm{TPR}_0(h^*) = \mathrm{TPR}_1(h^*)$. If h_λ satisfies equal opportunity on the biased distribution $\tilde{D}$, we have $\widetilde{\mathrm{TPR}}_0(h_\lambda) = \widetilde{\mathrm{TPR}}_1(h_\lambda)$, and since $\widetilde{\mathrm{TPR}}_a(h_\lambda) = \mathrm{TPR}_a(h_\lambda)$ by Corollary 3.1, it implies $\mathrm{TPR}_0(h_\lambda) = \mathrm{TPR}_1(h_\lambda)$. Combining the two observations above, we get $\mathrm{TPR}_0(h_\lambda) = \mathrm{TPR}_0(h^*) = \mathrm{TPR}_1(h^*) = \mathrm{TPR}_1(h_\lambda)$, which is possible only if $h_\lambda \equiv h^*$ (as both are group-wise threshold classifiers). The other case $t_0 \leqslant 1-\delta \leqslant t_1$ can be argued similarly.

Combining Steps 1, 2, and 3 to complete the proof: Finally, from Steps 1, 2, and 3 together, it is clear that if the conditions in Theorem 4.1 of Blum-Stangl are met then the only possible classifiers $\tilde{h}_\lambda$ are the ones for which $\tilde{h}_\lambda = h^*$, and they satisfy equal opportunity on $\tilde{D}$. So under these conditions, the optimal λ^* (whatever it may be) gives $\tilde{h}_{EO} = \tilde{h}_{\lambda^*} = h^*$.

□

§ 3.9.5 Proofs for Section 3.5

Proof. **(Proof of Proposition 3.4)**

The optimal reject option classifier $g : \mathcal{X} \times \mathcal{A} \to \{0, 1, \perp\}$ with rejection penalty δ can be thought of as a pair of binary classifiers ρ and h in $\mathcal{H}$ such that:

$$\rho(x,a) = \begin{cases} 1, & \text{if } g(x,a) = \perp \\ 0, & \text{if } g(x,a) \neq \perp \end{cases}$$

and $g(x,a) = h(x,a)$ whenever $\rho(x,a) = 0$ (or equivalently, $g(x,a) \neq \perp$). The optimal reject option classifier is defined as

$$h^{\mathrm{rej}} = \operatorname*{argmin}_{g \in \mathcal{H}^{\mathrm{rej}}} \Pr\left(g(X,A) \neq Y, g(X,A) \neq \perp\right) + \delta \Pr\left(g(X,A) = \perp\right),$$

where $\mathcal{H}^{\mathrm{rej}}$ be the hypothesis class of all reject option classifiers $g : \mathcal{X} \times \mathcal{A} \to \{0, 1, \perp\}$. Equivalently, it can be rewritten as:

$$
\begin{aligned}
(\rho^*, h^*) &= \operatorname*{argmin}_{\rho\in\mathcal{H}} \operatorname*{argmin}_{h\in\mathcal{H}} \Pr\left(h(X,A)\neq Y, \rho(X,A)=0\right) + \delta \Pr\left(\rho(X,A)=1\right) \\
&= \operatorname*{argmin}_{\rho\in\mathcal{H}} \operatorname*{argmin}_{h\in\mathcal{H}} 1 - \Pr\left(\rho(X,A)=1\right) - \Pr\left(h(X,A)=Y, \rho(X,A)=0\right) + \delta \Pr\left(\rho(X,A)=1\right) \\
&= \operatorname*{argmax}_{\rho\in\mathcal{H}} \operatorname*{argmax}_{h\in\mathcal{H}} \Pr\left(h(X,A)=Y, \rho(X,A)=0\right) + (1-\delta) \Pr\left(\rho(X,A)=1\right) \\
&= \operatorname*{argmax}_{\rho\in\mathcal{H}} \operatorname*{argmax}_{h\in\mathcal{H}} \sum_{a=0}^{1} \Pr(A=a) \mathop{\mathrm{E}}_{X|A=a}\Big[(1-\rho(X,a))(h(X,a)\eta(X,a) \\
&\qquad\qquad + (1-h(X,a))(1-\eta(X,a))) + (1-\delta)\rho(X,a)\Big] \\
&= \operatorname*{argmax}_{\rho\in\mathcal{H}} \sum_{a=0}^{1} \Pr(A=a) \mathop{\mathrm{E}}_{X|A=a}\Big[(1-\delta)\,\rho(X,a) \\
&\qquad\qquad + (1-\rho(X,a))\left(1-\eta(X,a) + \operatorname*{argmax}_{h\in\mathcal{H}} h(X,a)(2\eta(X,a)-1)\right)\Big].
\end{aligned}
$$

We can now focus on obtaining the optimal functions for each group. For any $\rho \in \mathcal{H}$, solving the inner optimization gives the optimal solution $h^{\text{rej}}(x,a) = \mathbb{I}\left(\eta(x,a) \geqslant \frac{1}{2}\right)$ whenever $\rho(x,a) \neq 1$. Thus, the outer optimization can be written as

$$
\rho^* = \operatorname*{argmax}_{\rho\in\mathcal{H}} \mathop{\mathrm{E}}_{X|A=a}\left[(1-\delta)\,\rho(X,a) + (1-\rho(X,a))\left(1-\eta(X,a) + \mathbb{I}\left(\eta(X,a)\geqslant\frac{1}{2}\right)(2\eta(X,a)-1)\right)\right].
$$

The above maximization can be solved point-wise for each $x \in \mathcal{X}$ in group $A = a$. Whenever $\eta(x,a) \geqslant 1/2$ for an input x from group a, the function inside the expectation becomes $(1-\delta)\rho(x,a) + (1-\rho(x,a))\eta(x,a)$, which is maximized by $\rho^*(x,a) = \mathbb{I}\left(\eta(x,a) \in [1/2, 1-\delta)\right)$. On the other hand, whenever $\eta(x,a) < 1/2$ for an input x, a, the function inside the expectation becomes $(1-\delta)\rho(x,a) + (1-\rho(x,a)(1-\eta(x,a))$, which is maximized by $\rho^*(x,a) = \mathbb{I}\left(\eta(x,a) \in (\delta, 1/2)\right)$. Thus, overall we can write $\rho^*(x,a) = \mathbb{I}\left(\eta(x,a) \in (\delta, 1-\delta)\right)$. □

Proof. **(Proof of Theorem 3.5)** From Proposition 3.3, we can write:

$$
\tilde{h}_\lambda(x,a) = \begin{cases} \mathbb{I}\left(\eta(x,0) \geqslant \dfrac{1}{1 + \dfrac{1-2\nu}{c} + \dfrac{\lambda}{\beta_n r q}}\right), & \text{for } a = 0 \\ \mathbb{I}\left(\eta(x,1) \geqslant \dfrac{1}{2 - \dfrac{\lambda}{(1-r)q}}\right), & \text{for } a = 1. \end{cases}
$$

We will now attempt to recover a classifier that matches the predictions of h^{rej}, whenever $h^{\text{rej}}(x,a) \neq \perp$, but incurs some penalty while trying to label the points when $h^{\text{rej}}(x,a) = \perp$. From the characterization of the optimal equal opportunity classifier shown earlier, there exists some optimal $\lambda^* \in \mathbb{R}$ such that $\tilde{h}_{\text{EO}} = h^{\text{rej}}$, whenever $h^{\text{rej}}(x,a) \neq \perp$. Such a classifier will not change whenever it lies in the rejection interval $(\delta, 1-\delta)$.

Step 1: If the given conditions on the bias parameters hold, then there exists a $\lambda \in \mathbb{R}$ such that

$\tilde{h}_\lambda = h^{\mathbf{rej}}$**, whenever** $h^{\mathbf{rej}}(x,a) \neq \perp$**.** The threshold on $\eta(x,1)$ lies in the interval $(\delta, 1-\delta)$ if and only if $\frac{1}{1-\delta} < 2 - \frac{\lambda}{(1-r)q} < \frac{1}{\delta}$, or equivalently, $\frac{-(1-2\delta)(1-r)q}{\delta} < \lambda < \frac{(1-2\delta)(1-r)q}{1-\delta}$. Similarly, the threshold on $\eta(x,0)$ lies in the interval $(\delta, 1-\delta)$ if and only if $\frac{1}{1-\delta} < 1 + \frac{1-2\nu}{c} + \frac{\lambda}{\beta_n rq} < \frac{1}{\delta}$, or equivalently,

$$\beta_n rq\left(\frac{\delta}{1-\delta} - \frac{1-2\nu}{c}\right) < \lambda < \beta_n rq\left(\frac{1-\delta}{\delta} - \frac{1-2\nu}{c}\right).$$

There exists λ that simultaneously satisfies both sets of constraints given above if and only if

$$\begin{aligned}\beta_n rq\left(\frac{\delta}{1-\delta} - \frac{1-2\nu}{c}\right) &< \frac{(1-2\delta)(1-r)q}{1-\delta}\\ \frac{-(1-2\delta)(1-r)q}{\delta} &< \beta_n rq\left(\frac{1-\delta}{\delta} - \frac{1-2\nu}{c}\right).\end{aligned}$$

Using $c = \beta_n/\beta_p$, the above conditions can be rewritten as

$$\begin{aligned}&(1-r)(1-2\delta) + r\left((1-\delta)\beta_p(1-2\nu) - \delta\beta_n\right) > 0 \quad \text{and}\\ &(1-r)(1-2\delta) + r\left((1-\delta)\beta_n - \delta\beta_p(1-2\nu)\right) > 0.\end{aligned}$$

Step 2: If the given conditions on the bias parameters hold, then there cannot exist any $\lambda \in \mathbb{R}$ **such that both the group-wise thresholds applied to** $\eta(x,a)$ **in** h_λ **are at most** δ**, or both the thresholds are at least** $1-\delta$**. (Thresholds lying on the same side)** The threshold on $\eta(x,1)$ is at most δ if and only if $2 - \frac{\lambda}{(1-r)q} \geqslant \frac{1}{\delta}$, or equivalently, $\frac{-(1-2\delta)(1-r)q}{\delta} \geqslant \lambda$. Similarly, the threshold on $\eta(x,0)$ is at most δ if and only if $1 + \frac{1-2\nu}{c} + \frac{\lambda}{\beta_n rq} \geqslant \frac{1}{\delta}$, or equivalently, $\lambda \geqslant \beta_n rq\left(\frac{1-\delta}{\delta} - \frac{1-2\nu}{c}\right)$. If both were to hold simultaneously, then $\frac{-(1-2\delta)(1-r)q}{\delta} \geqslant \beta_n rq\left(\frac{1-\delta}{\delta} - \frac{1-2\nu}{c}\right)$, or equivalently, $(1-r)(1-2\delta) + r\left((1-\delta)\beta_n - \delta\beta_p(1-2\nu)\right) \leqslant 0$, which would violate the second condition in Step 1.

On the other hand, the threshold on $\eta(x,1)$ is at least interval $1-\delta$ if and only if $\frac{1}{1-\delta} \geqslant 2 - \frac{\lambda}{(1-r)q}$, or equivalently, $\lambda \geqslant \frac{(1-2\delta)(1-r)q}{1-\delta}$. Similarly, the threshold on $\eta(x,0)$ is at least $1-\delta$ if and only if $\frac{1}{1-\delta} \geqslant 1 + \frac{1-2\nu}{c} + \frac{\lambda}{\beta_n rq}$, or equivalently, $\beta_n rq\left(\frac{\delta}{1-\delta} - \frac{1-2\nu}{c}\right) \geqslant \lambda$. If both were to hold simultaneously, then $\beta_n rq\left(\frac{\delta}{1-\delta} - \frac{1-2\nu}{c}\right) \geqslant \frac{(1-2\delta)(1-r)q}{1-\delta}$, or equivalently, $(1-r)(1-2\delta) + r\left((1-\delta)\beta_p(1-2\nu) - \delta\beta_n\right) \leqslant 0$, which would violate the first condition in Step 1.

Step 3: For any λ**, if the group-wise thresholds on** $\eta(x,a)$ **in** h_λ **have one of them at most** δ **and another at least** $1-\delta$**, then** h_λ **cannot satisfy equal opportunity unless** $h_\lambda = h^{\mathbf{rej}}$**, whenever** $h^{\mathbf{rej}}(x,a) \neq \perp$**. (Thresholds lying on opposite sides)** From our assumption, we know that h^{rej} satisfies equal opportunity whenever $h^{rej}(x,a) \neq \perp$. Suppose the threshold of h_λ on $\eta(x,0)$(call it t_0) and the threshold on $\eta(x,1)$(call it t_1) are separated by either δ or $1-\delta$. WLOG assume $t_0 < t_1$. Suppose $t_0 \leqslant \delta \leqslant t_1$. Since there is no input (x,a) with $\eta(x,a) \in (\delta, 1-\delta)$, we get $TPR_0(h_\lambda) \geqslant TPR_0(h^{rej})$ and $TPR_1(h^{rej}) \geqslant TPR_1(h_\lambda)$. Since h^{rej} satisfies equal opportunity on the original distribution D, we have $TPR_0(h^{rej}) = TPR_1(h^{rej})$. If h_λ satisfies equal opportunity on the biased distribution $\tilde{D}$, we have $\widetilde{TPR}_0(h_\lambda) = \widetilde{TPR}_1(h_\lambda)$, and since $\widetilde{TPR}_a(h_\lambda) = TPR_a(h_\lambda)$ by Corollary 3.1, it implies $TPR_0(h_\lambda) =$

$\mathrm{TPR}_1(h_\lambda)$. Combining the two observations above, we get $\mathrm{TPR}_0(h_\lambda) = \mathrm{TPR}_0(h^{rej}) = \mathrm{TPR}_1(h^{rej}) = \mathrm{TPR}_1(h_\lambda)$, which is possible only if $h_\lambda \equiv h^{rej}$ (as both are group-wise threshold classifiers). The other case $t_0 \leqslant 1-\delta \leqslant t_1$ can be argued similarly.

Combining Steps 1, 2, and 3 to complete the proof: Finally, from Steps 1, 2, and 3 together, it is clear that when the above conditions are met, then the only possible classifiers $\tilde{h}_\lambda$ are the ones for which $\tilde{h}_\lambda = h^{rej}$, whenever $h^{rej}(x,a) \neq \perp$, and they satisfy equal opportunity on $\tilde{D}$. Furthermore, by our assumption, h^{rej} is EO-fair.

□

§ 3.9.6 Proofs for Section 3.6

Definition 3.1 of ϵ-robustness of h^* can be rewritten as

$$\begin{aligned}
h^*(x) &= \operatorname*{argmax}_{h\in\mathcal{H}} \Pr\left(h(X,A) = Y'\right) \\
&= \operatorname*{argmax}_{h\in\mathcal{H}} \mathop{E}_{(X,A)}\left[h(X,A)(2\eta'(X,A)-1)\right] \\
&= \operatorname*{argmax}_{h\in\mathcal{H}} \sum_{a=0}^{1} \Pr(A=a) \mathop{E}_{X|A=a}\left[h(X,a)(2\eta'(X,a)-1)\right] \\
&= \operatorname*{argmax}_{h\in\mathcal{H}} \sum_{a=0}^{1} \Pr(A=a) \mathop{E}_{X|A=a}\left[h(X,a)\left(2\,\frac{P_a\eta(X,a)+Q_a}{R_a\eta(X,a)+S_a}-1\right)\right] \\
&= \operatorname*{argmax}_{h\in\mathcal{H}} \sum_{a=0}^{1} \Pr(A=a) \mathop{E}_{X|A=a}\left[h(X,a)\frac{(2P_a-R_a)\eta(X,a)+(2Q_a-S_a)}{R_a\eta(X,a)+S_a}\right],
\end{aligned}$$

for any $S_a = 1, |R_a| \leqslant \epsilon, |Q_a| \leqslant \epsilon, 1-\epsilon \leqslant P_a \leqslant 1+\epsilon$, and $P_aS_a - Q_aR_a \geqslant 0$.

Proof. (Proof of Theorem 3.6) Let $\tilde{h}_{EO} = \arg\max_{h\in\tilde{\mathcal{H}}_{fair,EO}} \Pr\left(h(X,A)=\tilde{Y}\right)$. By Lagrange duality, we write the objective maximized by $\tilde{h}_{EO}$ for the optimal multiplier λ^* as:

$$\tilde{h}_{EO} = \arg\max_{h\in\mathcal{H}} \sum_{a=0}^{1} \Pr(A=a) \mathop{E}_{X|A=a}\left[h(X,a)\left(\left(2+\frac{(-1)^{\mathbb{I}(a=1)}\lambda^*}{\Pr\left(\tilde{Y}=1,A=a\right)}\right)\tilde{\eta}(X,a)-1\right)\right]$$

Under uniform under-representation ($\beta_p = \beta_n = \beta$), we have $c = 1$. This simplifies the transformed regression function for the underprivileged group to $\tilde{\eta}(X,0) = (1-\nu)\eta(X,0)$. For the privileged group, $\tilde{\eta}(X,1) = \eta(X,1)$.

Substituting these, along with the biased base rates $\Pr\left(\tilde{Y}=1,A=0\right) = rq\beta(1-\nu)$ and $\Pr\left(\tilde{Y}=1,A=1\right) = (1-r)q$, the objective becomes:

$$\begin{aligned}
= \arg\max_{h\in\mathcal{H}} \Bigg(& rE_{X|A=0}\left[h(X,0)\left(\left(2(1-\nu)+\frac{\lambda^*}{rq\beta}\right)\eta(X,0)-1\right)\right] \\
& + (1-r)E_{X|A=1}\left[h(X,1)\left(\left(2-\frac{\lambda^*}{(1-r)q}\right)\eta(X,1)-1\right)\right]\Bigg)
\end{aligned}$$

This precisely matches the form from the ϵ-robustness definition and we get:

$$P_0 = 1 - \nu + \frac{\lambda^*}{2rq\beta}, \qquad Q_0 = 0, \quad R_0 = 0, \qquad S_0 = 1$$
$$P_1 = 1 - \frac{\lambda^*}{2(1-r)q}, \qquad Q_1 = 0, \quad R_1 = 0, \qquad S_1 = 1$$

Let $U_a(h) = E_{X|A=a}[h(X,a)\eta(X,a)] = q \cdot TPR_a(h)$, and let $V_a(h) = E_{X|A=a}[h(X,a)]$. The objective maximized on each group a reduces strictly to $2P_aU_a(h) - V_a(h)$.

Step 1: If the given conditions on the bias parameters hold, then there exists a $\lambda \in \mathbb{R}$ such that $1-\epsilon \leqslant P_0, P_1 \leqslant 1+\epsilon$, and therefore, $h_\lambda = h^*$.

To have $\tilde{h}_{EO} \equiv h^*$ directly from robustness, we require $1-\epsilon \leqslant P_0 \leqslant 1+\epsilon$ and $1-\epsilon \leqslant P_1 \leqslant 1+\epsilon$. This translates to:

$$1-\epsilon \leqslant 1-\nu+\frac{\lambda^*}{2rq\beta} \leqslant 1+\epsilon \implies 2rq\beta(\nu-\epsilon) \leqslant \lambda^* \leqslant 2rq\beta(\nu+\epsilon)$$

$$1-\epsilon \leqslant 1-\frac{\lambda^*}{2(1-r)q} \leqslant 1+\epsilon \implies -2q(1-r)\epsilon \leqslant \lambda^* \leqslant 2q(1-r)\epsilon$$

A valid λ^* satisfying both intervals exists if and only if:

$$2rq\beta(\nu-\epsilon) \leqslant 2q(1-r)\epsilon \implies r\beta(\epsilon-\nu)+\epsilon(1-r) \geqslant 0$$
$$-2q(1-r)\epsilon \leqslant 2rq\beta(\nu+\epsilon) \implies r\beta(\epsilon+\nu)+\epsilon(1-r) \geqslant 0$$

Step 2: If the given conditions on the bias parameters hold, then there cannot exist any $\lambda \in \mathbb{R}$ such that $\max\{P_0, P_1\} < 1-\epsilon$ or $\min\{P_0, P_1\} > 1+\epsilon$.

If $\max\{P_0, P_1\} < 1-\epsilon$, then $1-\nu+\frac{\lambda^*}{2rq\beta} < 1-\epsilon$ and $1-\frac{\lambda^*}{2(1-r)q} < 1-\epsilon$. This implies $\lambda^* < 2rq\beta(\nu-\epsilon)$ and $\lambda^* > 2q(1-r)\epsilon$. Thus, $2q(1-r)\epsilon < 2rq\beta(\nu-\epsilon)$, which strictly violates the required condition $r\beta(\epsilon-\nu)+\epsilon(1-r) \geqslant 0$. By identical logic, $\min\{P_0, P_1\} > 1+\epsilon$ violates the other theorem condition.

Step 3: If the given conditions on the bias parameters hold, then an optimal $\lambda^* \in \mathbb{R}$ cannot correspond to $P_0 < 1-\epsilon$ and $P_1 > 1+\epsilon$ (or vice versa).

Suppose λ^* forces $P_0 < 1-\epsilon$ and $P_1 > 1+\epsilon$. Recall that for any group a, the objective maximized by h reduces to $2P_aU_a(h) - V_a(h)$.

Part 3.1: Analyzing Group 0

Assume $P_0 < 1-\epsilon$. Because h^* is ϵ-robust, it is optimal at the lower boundary $P = 1-\epsilon$. Thus:

$$2(1-\epsilon)U_0(h^*) - V_0(h^*) \geqslant 2(1-\epsilon)U_0(\tilde{h}_{EO}) - V_0(\tilde{h}_{EO}) \tag{3.1}$$

By definition, $\tilde{h}_{EO}$ is optimal for the shifted parameter P_0. Thus:

$$2P_0U_0(\tilde{h}_{EO}) - V_0(\tilde{h}_{EO}) \geqslant 2P_0U_0(h^*) - V_0(h^*) \tag{3.2}$$

Summing Inequality 3.1 and Inequality 3.2 cancels the V_0 terms:

$$2(1-\epsilon-P_0)U_0(h^*) \geqslant 2(1-\epsilon-P_0)U_0(\tilde{h}_{EO})$$

Because $P_0 < 1-\epsilon$, the multiplier $(1-\epsilon-P_0) > 0$. Dividing it out yields $U_0(h^*) \geqslant U_0(\tilde{h}_{EO})$, meaning $TPR_0(h^*) \geqslant TPR_0(\tilde{h}_{EO})$.

Part 3.2: Analyzing Group 1

Assume $P_1 > 1 + \epsilon$. Because h^* is robust, it is optimal at the upper boundary $P = 1 + \epsilon$:

$$2(1+\epsilon)U_1(h^*) - V_1(h^*) \geqslant 2(1+\epsilon)U_1(\tilde{h}_{EO}) - V_1(\tilde{h}_{EO})$$

$\tilde{h}_{EO}$ is optimal for P_1:

$$2P_1U_1(\tilde{h}_{EO}) - V_1(\tilde{h}_{EO}) \geqslant 2P_1U_1(h^*) - V_1(h^*)$$

Summing these inequalities cancels the V_1 terms:

$$2(P_1 - (1+\epsilon))U_1(\tilde{h}_{EO}) \geqslant 2(P_1 - (1+\epsilon))U_1(h^*)$$

Because $P_1 > 1 + \epsilon$, the multiplier $(P_1 - 1 - \epsilon) > 0$. Dividing it out yields $U_1(\tilde{h}_{EO}) \geqslant U_1(h^*)$, meaning $TPR_1(\tilde{h}_{EO}) \geqslant TPR_1(h^*)$.

Part 3.3: The Fairness Contradiction

Since h^* satisfies Equal Opportunity on D, $TPR_0(h^*) = TPR_1(h^*)$. Since $\tilde{h}_{EO}$ is EO-fair on $\tilde{D}$ and under uniform under-representation TPRs are invariant, $TPR_0(\tilde{h}_{EO}) = TPR_1(\tilde{h}_{EO})$. Substituting our bounds into this equality chain:

$$TPR_0(\tilde{h}_{EO}) \leqslant TPR_0(h^*) = TPR_1(h^*) \leqslant TPR_1(\tilde{h}_{EO}) = TPR_0(\tilde{h}_{EO})$$

For the ends of the chain to satisfy $TPR_0(\tilde{h}_{EO}) \leqslant TPR_0(\tilde{h}_{EO})$, all inequalities must be strict equalities. Thus, $TPR_0(\tilde{h}_{EO}) = TPR_0(h^*)$ and $TPR_1(\tilde{h}_{EO}) = TPR_1(h^*)$.

Part 3.4: Geometric Resolution of Identical Classifiers via ROC Analysis

While we have established that $TPR_0(\tilde{h}_{EO}) = TPR_0(h^*)$ and $TPR_1(\tilde{h}_{EO}) = TPR_1(h^*)$, matching True Positive Rates alone is insufficient to prove the classifiers are identical. We must show their True Negative Rates (TNR) also match. To do this, we analyze the optimization geometry in the Receiver Operating Characteristic (ROC) space.

Recall that for any group a, both h^* and $\tilde{h}_{EO}$ are derived by maximizing an objective of the form $g_P(h) = 2P_aU_a(h) - V_a(h)$. In the ROC space where the y-axis is the True Positive Rate (proportional to U_a) and the x-axis is the False Positive Rate (proportional to V_a), any classifier maximizing this objective must lie on the upper-left boundary of the achievable convex hull. This boundary is strictly concave.

Furthermore, the parameter P_a geometrically dictates the slope of the tangent line to the ROC curve at the optimal operating point.

1. **The Geometry of ϵ-Robustness:** By Definition 3.1, h^* is ϵ-robust, meaning it remains the unique optimal classifier for any weight $P \in [1-\epsilon, 1+\epsilon]$. Geometrically, a single point on a concave curve can only be optimal for a continuous range of strictly different slopes if that point is a sharp, non-differentiable corner (a vertex) on the ROC hull.
2. **The Slope of $\tilde{h}_{EO}$:** $\tilde{h}_{EO}$ is optimal for the biased parameter P_0. Under our current branch, we assumed $P_0 < 1 - \epsilon$. This means the objective slope used to discover $\tilde{h}_{EO}$ is strictly steeper than any slope in the robustness margin of h^*.
3. **The Vertex Contradiction:** We established in Part 3.3 that despite optimizing for strictly different slopes ($P_0 < 1 - \epsilon$ versus $P = 1 - \epsilon$), both $\tilde{h}_{EO}$ and h^* achieve the exact same y-coordinate (TPR). Geometrically, it is impossible for two tangent lines with strictly different slopes to intersect the upper boundary of a concave curve at the exact same y-coordinate unless they are both pivoting

on the exact same non-differentiable vertex.

Therefore, $\tilde{h}_{EO}$ and h^* must occupy the exact same geometric vertex on the ROC hull. Because they are the same mathematical point, they must also share the exact same x-coordinate, meaning $FPR(\tilde{h}_{EO}) = FPR(h^*)$, which trivially implies $TNR(\tilde{h}_{EO}) = TNR(h^*)$.

Since $\tilde{h}_{EO}$ and h^* share identically perfect TPRs and TNRs, they achieve the exact same accuracy profile. Because the ϵ-robustness definition guarantees the uniqueness of h^* at this operating profile, we conclude $\tilde{h}_{EO} \equiv h^*$.

The inverse case ($P_0 > 1+\epsilon$ and $P_1 < 1-\epsilon$) is contradicted identically. □

§ 3.9.7 Proofs for Section 3.7

Proof. **(Proof for Theorem 3.7)** We start with the Bayes Optimal EO classifier at the time step t:

$$\tilde{h}_{EO}^{(t)}(x,a) = \mathbb{I}\left(\tilde{\eta}_t(x,a) \geqslant \frac{1}{2 + \frac{(-1)^{\mathbb{I}(a=1)}\lambda^*}{\Pr\left(\tilde{Y}_t = 1, A = a\right)}}\right)$$

From Proposition 3.5 we can obtain the transformed $\tilde{\eta}_t$ with the same bias parameters at each time step:

$$= \begin{cases} \mathbb{I}\left(\dfrac{\eta(x,0)}{\frac{1-c}{1-\nu-c}\left(1-\left(\frac{c}{1-\nu}\right)^t\right)\eta(x,0) + \left(\frac{c}{1-\nu}\right)^t} \geqslant \dfrac{1}{2 + \frac{\lambda^*}{\Pr\left(\tilde{Y}_t = 1, A = 0\right)}}\right), & \text{for } a = 0 \\ \mathbb{I}\left(\eta(x,1) \geqslant \dfrac{1}{2 - \frac{\lambda^*}{\Pr(Y=1, A=1)}}\right), & \text{for } a = 1 \end{cases}$$

$$= \begin{cases} \mathbb{I}\left(\eta(x,0) \geqslant \dfrac{1}{\frac{1-2\nu-c}{1-\nu-c}\left(\frac{1-\nu}{c}\right)^t + \frac{\lambda^*}{\beta_n^t rq} + \frac{1-c}{1-\nu-c}}\right), & \text{for } a = 0, \text{using } c = \beta_n/\beta_p \\ \mathbb{I}\left(\eta(x,1) \geqslant \dfrac{1}{2 - \frac{\lambda^*}{(1-r)q}}\right), & \text{for } a = 1. \end{cases}$$

So the threshold on $\eta(x,1)$ lies in the interval $(\delta, 1-\delta)$ if and only if $\delta < \left(2 - \frac{\lambda^*}{(1-r)q}\right)^{-1} < 1-\delta$, which is equivalent to $\frac{-(1-2\delta)(1-r)q}{\delta} < \lambda^* < \frac{(1-2\delta)(1-r)q}{1-\delta}$. Given this, now let's analyze the threshold on $\eta(x,0)$. If $\beta_n < 1$, then as $t \to \infty$ we have $\lambda^*/(\beta_n^t rq) \to \infty$, and therefore, the threshold on $\eta(x,0)$ in the above expression tends to 0, i.e., it cannot remain within $(\delta, 1-\delta)$ interval. If $\beta_n = 1$ then $c = \beta_n/\beta_p > 1-\nu$. Thus, $((1-\nu)/c)^t \to 0$ as $t \to \infty$. Using this and $\beta_n = 1$, the above expression becomes

$$\frac{1}{\frac{1-2\nu-c}{1-\nu-c}\left(\frac{1-\nu}{c}\right)^t + \frac{\lambda^*}{\beta_n^t rq} + \frac{1-c}{1-\nu-c}} \to \frac{1}{\frac{\lambda^*}{rq} + \frac{1-c}{1-\nu-c}} \quad \text{as } t \to \infty.$$

The threshold on $\eta(x,1)$ lies in the interval $(\delta, 1-\delta)$ if and only if

$$\frac{-(1-2\delta)(1-r)q}{\delta} < \lambda^* < \frac{(1-2\delta)(1-r)q}{1-\delta}.$$

Similarly, the limit expression as $t \to \infty$ for threshold on $\eta(x,0)$ lies in the interval $(\delta, 1-\delta)$ if and only if $\frac{1}{1-\delta} < \frac{\lambda^*}{rq} + \frac{1-c}{1-\nu-c} < \frac{1}{\delta}$, or equivalently, $rq\left(\frac{1}{1-\delta} - \frac{1-c}{1-\nu-c}\right) < \lambda^* < rq\left(\frac{1}{\delta} - \frac{1-c}{1-\nu-c}\right)$. There exists a λ^* that simultaneously satisfies the constraints on both the thresholds if and only if

$$\frac{-(1-2\delta)(1-r)q}{\delta} < rq\left(\frac{1}{\delta} - \frac{1-c}{1-\nu-c}\right) \quad \text{and} \quad rq\left(\frac{1}{1-\delta} - \frac{1-c}{1-\nu-c}\right) < \frac{(1-2\delta)(1-r)q}{1-\delta}.$$

Thus, the necessary conditions to recover h^* using equal opportunity fair classification on the biased distribution after t steps as $t \to \infty$ can be written as $\beta_n = 1$ and

$$\frac{r-(1-2\delta)(1-r)}{(1-\delta)r} < \frac{1-c}{1-\nu-c} < \frac{r+(1-2\delta)(1-r)}{\delta r}.$$

The above conditions can be further simplified as $\beta_n = 1$ and $1 - \frac{(1-2\delta)}{(1-\delta)r} < \frac{\nu\beta_p}{1-\beta_p(1-\nu)} < 1 + \frac{(1-2\delta)}{\delta r}$. □

Proof. (Proof for Theorem 3.8) Similar to the proof of Theorem 3.3, we begin with $\tilde{\eta}(x,a)$:

$$\tilde{h}_{EO,t}(x,a) = \mathbb{I}\left(\tilde{\eta_t}(x,a) \geqslant \frac{1}{2 + \frac{(-1)^{\mathbb{I}(a=1)}\lambda^*}{\Pr\left(\tilde{Y}=1, A=a\right)}}\right)$$

Because the population in group 1 is unaffected, we can obtain the conditions on λ^*. $\Pr(Y=1, A=1) = (1-r)q$. From Lemma 3.1, we know that the threshold on $\eta(x,1)$ lies in the interval $(\delta, 1-\delta)$ if and only if $\frac{1}{1-\delta} < 2 - \frac{\lambda^*}{(1-r)q} < \frac{1}{\delta}$, or equivalently, $\frac{-(1-2\delta)(1-r)q}{\delta} < \lambda^* < \frac{(1-2\delta)(1-r)q}{1-\delta}$. We will now work towards obtaining the set of conditions on λ^* by focusing on the quantities inside the indicator on $\tilde{\eta_t}(x,0)$. Using Proposition 3.5:

$$\tilde{\eta}_t(x,0) = \frac{\eta(x,0)}{\sum_{i=1}^{t}\left(\frac{1-c_i}{1-\nu_i}\prod_{j=i+1}^{t}\frac{c_j}{1-\nu_j}\right)\eta(x,0) + \prod_{i=1}^{t}\frac{c_i}{1-\nu_i}} \geqslant \frac{1}{2 + \frac{\lambda^*}{\Pr\left(\tilde{Y}=1, A=0\right)}}$$

From previous derivations, we know that $\Pr\left(\tilde{Y}=1, A=0\right) = rq\prod_{i=1}^{t}(\beta_{p,i}(1-\nu_i)$. Simplifying the above expression gives us the following:

$$\eta(x,0) \geqslant \frac{1}{2\prod_{i=1}^{t}\frac{1-\nu_i}{c_i} + \frac{\lambda^*}{rq\prod_{i=1}^{t}\beta_{n,i}} - \sum_{i=1}^{t}\left(\frac{1-c_i}{c_i}\prod_{j=1}^{i-1}\frac{1-\nu_j}{c_j}\right)}$$

where $\prod_{j=1}^{i-1} \frac{1-\nu_j}{c_j} = 1$ whenever $j > i$.

Using similar arguments as in Proposition 3.3, the denominator must lie in the range $((1-\delta)^{-1}, \delta^{-1})$, which gives us the following inequalities on λ^*:

$$\frac{rq\prod_{i=1}^{t}\beta_{n,i}}{1-\delta} - 2rq\prod_{i=1}^{t}\left[(1-\nu_i)\beta_{p,i}\right] + rq\left(\prod_{i=1}^{t}\left[(1-\nu_i)\beta_{p,i}\right]\right)\sum_{j=1}^{t}\left(\frac{1-c_j}{c_j}\prod_{k=1}^{j-1}\frac{1-\nu_k}{c_k}\right) < \lambda^* \tag{3.3}$$

and

$$\lambda^* < \frac{rq\prod_{i=1}^{t}\beta_{n,i}}{\delta} - 2rq\prod_{i=1}^{t}\left[(1-\nu_i)\beta_{p,i}\right] + rq\left(\prod_{i=1}^{t}\left[(1-\nu_i)\beta_{p,i}\right]\right)\sum_{j=1}^{t}\left(\frac{1-c_j}{c_j}\prod_{k=1}^{j-1}\frac{1-\nu_k}{c_k}\right) \tag{3.4}$$

To get a tight bound, we can obtain an upper bound on the left side of Equation 3.3 and a lower bound on the right side of Equation 3.4, using the assumed maximal bias of $\beta_{n,t} \in [\beta_n, 1], \beta_{p,t} \in [\beta_p, 1]$ and $\nu_t \in [0, \nu]$, where $\nu < \frac{1}{2}$. This gives us the following bounds on λ^*:

$$\frac{rq}{1-\delta} - 2rq(1-\nu)^t\beta_p^t + rq\frac{1-\beta_n^t}{\beta_n^t} < \lambda^* < \frac{rq\beta_n^t}{\delta} - 2rq + rq(1-\nu)^t\beta_p^t(\beta_p-1)\frac{1-\beta_p^t(1-\nu)^t}{1-\beta_p(1-\nu)} \tag{3.5}$$

Comparing this with the conditions on λ^* found earlier: $\frac{-(1-2\delta)(1-r)q}{\delta} < \lambda^* < \frac{(1-2\delta)(1-r)q}{1-\delta}$, we get the following inequalities:

$$\frac{(1-2\delta)(1-r)}{(1-\delta)r} - \frac{\delta}{1-\delta} > \frac{1}{\beta_n^t} - 2\beta_p^t(1-\nu)^t$$

and

$$\frac{(1-2\delta)(1-r)}{\delta r} > (1-\nu)^t\beta_p^t(1-\beta_p)\frac{1-\beta_p^t(1-\nu)^t}{1-\beta_p(1-\nu)} - \frac{\beta_n^t}{\delta} - 2$$

□

Theorem 3.11. Assuming boundedness on data bias parameters at each time step: $\beta_{n,t} \in [\beta_n, 1]$, $\beta_{p,t} \in [\beta_p, 1]$ and $\nu_t \in [0, \nu]$, where $\nu < \frac{1}{2}$; whenever the following relationships hold:

$$(1-3\delta)(\beta_p-1)\frac{1-\beta_p^t(1-\nu)^t}{1-\beta_p(1-\nu)} - 2\delta - (1-\delta)\left(\frac{1-\beta_n^t}{\beta_n^t} - 2\beta_p^t(1-\nu)^t\right) > 0 \quad \text{and}$$
$$1 + \delta\left((\beta_p-1)\frac{1-\beta_p^t(1-\nu)^t}{1-\beta_p(1-\nu)} - \frac{2}{\beta_n^t} - (2\delta-1)\frac{1-\beta_n^t}{\beta_n^t}\right) - 2(\delta-1) > 0,$$

we have $\tilde{h}_{DP,t}(x,a) = h_{DP}(x,a) = h^*$.

Proof. We again begin with the thresholding results on Demographic Parity from Proposition 3.10.

$$\tilde{h}_{DP}(x,a) = \mathbb{I}\left(\tilde{\eta}(x,a) \geqslant \frac{1}{2} - \frac{-1^{\mathbb{I}(a=1)}\lambda^*}{2}\right)$$

(Using Proposition 3.5)

$$= \begin{cases} \mathbb{I}\left(\dfrac{\eta(x,0)}{\sum_{i=1}^{t}\left(\frac{1-c_i}{1-\nu_i}\prod_{j=i+1}^{t}\frac{c_j}{1-\nu_j}\right)\eta(x,0) + \prod_{i=1}^{t}\frac{c_i}{1-\nu_i}} \geqslant \frac{1}{2} - \frac{\lambda^*}{2}\right), & \text{for } a=0 \\ \mathbb{I}\left(\eta(x,1) \geqslant \frac{1}{2} + \frac{\lambda^*}{2}\right), & \text{for } a=1 \end{cases}$$

$$= \begin{cases} \mathbb{I}\left(\eta(x,0) \geqslant \dfrac{(1-\lambda^*)\prod_{i=1}^{t}\frac{c_i}{1-\nu_i}}{2-(1-\lambda^*)\sum_{i=1}^{t}\left(\frac{1-c_i}{1-\nu_i}\prod_{j=i+1}^{t}\frac{c_j}{1-\nu_j}\right)}\right), & \text{for } a=0 \\ \mathbb{I}\left(\eta(x,1) \geqslant \frac{1}{2} + \frac{\lambda^*}{2}\right), & \text{for } a=1 \end{cases}$$

$$= \begin{cases} \mathbb{I}\left(\eta(x,0) \geqslant \dfrac{1}{\frac{2}{1-\lambda^*}\prod_{i=1}^{t}\frac{1-\nu_i}{c_i} - \sum_{i=1}^{t}\left(\frac{1-c_i}{c_i}\prod_{j=1}^{i-1}\frac{1-\nu_j}{c_j}\right)}\right), & \text{for } a=0 \\ \mathbb{I}\left(\eta(x,1) \geqslant \frac{1}{2} + \frac{\lambda^*}{2}\right), & \text{for } a=1 \end{cases}$$

where $\prod_{j=1}^{i-1}\frac{1-\nu_j}{c_j} = 1$ whenever $j > i$. From Lemma 3.1, we know that the threshold on $\eta(x,1)$ lies in the interval $(\delta, 1-\delta)$ if and only if $\delta < \frac{1}{2} + \frac{\lambda^*}{2} < 1-\delta$, or equivalently, $2\delta - 1 < \lambda^* < 1 - 2\delta$. Similarly, the threshold on $\eta(x,0)$ lies in the interval $(\delta, 1-\delta)$ if and only if the denominator lies in the range $((1-\delta)^{-1}, \delta^{-1})$ which gives us the following inequalities on λ^*:

$$\frac{1 + (1-\delta)\left(\sum_{i=1}^{t}\left(\frac{1-c_i}{c_i}\prod_{j=1}^{i-1}\frac{1-\nu_j}{c_j}\right) - 2\prod_{i=1}^{t}\frac{1-\nu_i}{c_i}\right)}{1 + (1-\delta)\sum_{i=1}^{t}\left(\frac{1-c_i}{c_i}\prod_{j=1}^{i-1}\frac{1-\nu_j}{c_j}\right)} < \lambda^* \tag{3.6}$$

$$\lambda^* < \frac{1 + \delta\left(\sum_{i=1}^{t}\left(\frac{1-c_i}{c_i}\prod_{j=1}^{i-1}\frac{1-\nu_j}{c_j}\right) - 2\prod_{i=1}^{t}\frac{1-\nu_i}{c_i}\right)}{1 + \delta\sum_{i=1}^{t}\left(\frac{1-c_i}{c_i}\prod_{j=1}^{i-1}\frac{1-\nu_j}{c_j}\right)} \tag{3.7}$$

To get a tight bound, we can obtain an upper bound on Equation 3.6 and a lower bound on Equation 3.7, using the assumed maximal bias of $\beta_{n,t} \in [\beta_n, 1], \beta_{p,t} \in [\beta_p, 1]$ and $\nu_t \in [0, \nu]$, where $\nu < \frac{1}{2}$. This gives us the following bounds on λ^*:

$$\frac{1+(1-\delta)\left(\frac{1-\beta_n^t}{\beta_n^t}-2\beta_p^t(1-\nu)^t\right)}{1+(1-\delta)(\beta_p-1)\left(\frac{1-\beta_p^t(1-\nu)^t}{1-\beta_p(1-\nu)}\right)}<1-2\delta \text{ , and}$$

$$2\delta-1<\frac{1+\delta\left((\beta_p-1)\left(\frac{1-\beta_p^t(1-\nu)^t}{1-\beta_p(1-\nu)}\right)-\frac{2}{\beta_n^t}\right)}{1+\delta\frac{1-\beta_n^t}{\beta_n^t}}$$

Therefore, to have a λ^* which satisfies both sets of inequalities, the following conditions must hold:

$$(1-3\delta)(\beta_p-1)\frac{1-\beta_p^t(1-\nu)^t}{(1-\beta_p(1-\nu)}-2\delta-(1-\delta)\left(\frac{1-\beta_n^t}{\beta_n^t}-2\beta_p^t(1-\nu)^t\right)>0 \tag{3.8}$$

and

$$1+\delta\left((\beta_p-1)\frac{1-\beta_p^t(1-\nu)^t}{(1-\beta_p(1-\nu)}-\frac{2}{\beta_n^t}-(2\delta-1)\frac{1-\beta_n^t}{\beta_n^t}\right)-2(\delta-1)>0 \tag{3.9}$$

□

Chapter 4

A Closer Look at Pre-/In- and Post-Processing Interventions

Abstract

In this chapter, we will first highlight a simple observation that the fair Bayes optimal characterization result from Propositions 2.2 and 2.3 can be implemented as a pre-processing of the data distribution, an in-processing of the loss function, or as a post-processing of the class posterior probability. We will then examine the effects of data bias on pre-, in-, and post-processing fair classifiers, assessing the robustness of accuracy and fairness to varying levels of noise and bias in the training data.

§ 4.1 Revisiting Fair Bayes Optimal Classifiers

In Propositions 2.2 and 2.3, we described the fair Bayes optimal classifier as a thresholding of the group-dependent class posterior probability $\eta(X, A)$. In this section, we highlight a subtle observation: the optimal threshold rule can be implemented both as pre-processing reweighing of the data distribution and as weighted risk minimization. This observation has also been recently highlighted by [Zen+26], and we will largely follow their methodology.

Let $\tilde{D}$ be a modified distribution derived from D. We assume that the pre-processing intervention preserves the feature distributions conditioned on the subgroups, and only modifies the subgroup probabilities: $\Pr\left(\tilde{X} = x|\tilde{Y} = y, \tilde{A} = a\right) = \Pr\left(X = x|Y = y, A = a\right)$, but $\tilde{\pi}_{ya} = \Pr\left(\tilde{Y} = y, \tilde{A} = a\right) \neq \Pr\left(Y = y, A = a\right) = \pi_{ya}$.

For pre-processing, our aim is to output a distribution $\tilde{D}$ such that the unconstrained Bayes optimal classifier on $\tilde{D}$ is fair. We now highlight that a simple modification to subgroup probabilities (π_{ya}) is enough to achieve that for any λ. Note that while this re-weighting scheme can be constructed for any fairness disparity (and hence any λ), we illustrate the result for exact fairness.

Theorem 4.1. (Theorem 5.1 in [Zen+26]) Suppose there exists a λ^* such that the Bayes optimal classifiers in Propositions 2.2 and 2.3 exist. Let $\tilde{D}$ be the distribution constructed by modifying the subgroup probabilities $\tilde{\pi}_{ya}$ so that the unconstrained Bayes optimal classifier on $\tilde{D}$: $\tilde{h}(x, a) = \mathbb{I}\left(\tilde{\eta}(x, a) \geqslant \frac{1}{2}\right)$ is fair with respect to D. Then, setting the subgroup probabilities as follows suffices for the construction of $\tilde{D}$:

- Demographic Parity: $\tilde{\pi}_{ya} = \pi_{ya} \cdot \left(\frac{1}{2} + \frac{(2a-1)\lambda^*}{2\pi_a}(1-2y)\right) \cdot c_a$,

- Equal Opportunity: $\tilde{\pi}_{ya} = \pi_{ya} \cdot \left(y \left(1 - \frac{2}{2 - \frac{(2a-1)\lambda^*}{\pi_{1a}}} \right) + \frac{1}{2 - \frac{(2a-1)\lambda^*}{\pi_{1a}}} \right) \cdot c_a$,

where $c_a > 0$ are group-dependent constants that ensure that $\tilde{D}$ is a valid probability distribution, i.e., $\sum_{(y,a)} \tilde{\pi}_{ya} = 1$.

With the new subgroup proportions, one can modify the distribution with under-/over-sampling particular subgroups to obtain a new distribution, where the unconstrained Bayes optimal classifier will be guaranteed to be fair. [Zen+26] also proposes empirically stable techniques to reduce variability in sampling, when λ is obtained via bisection search and the 'ideal' datasets are obtained iteratively.

Next, we will similarly restate the in-processing version of fair Bayes optimal classifiers. Our aim is to find a loss function $\tilde{l}$ such that the unconstrained minimization of $\tilde{l}$ outputs the fair Bayes optimal classifier.

Theorem 4.2. (Theorem 5.2 in [Zen+26]) Suppose there exists a λ^* such that the Bayes optimal classifiers in Propositions 2.2 and 2.3 exist. Then, let $\tilde{l}(h(x,a),y) = c_{ya} l_{01}(h(x,a),y)$, where $c_{ya} = \frac{1}{2} + \frac{\lambda^*}{2\pi_a}(2a-1)(1-2y)$ for Demographic Parity, and $c_{ya} = \frac{(1-2y)}{2 - \frac{\lambda^*(2a-1)}{\pi_{1a}}} + y$ for Equal Opportunity. Then, $\tilde{h} \in \arg\min_h \mathop{E}_{(X,A,Y)\in D} \left[\tilde{l}(h(X,A),Y) \right]$ is the fair Bayes optimal risk minimizer.

The structure of the results in Theorems 4.1 and 4.2 shows that any characterization result (which is inherently a post-processing classifier) can be tailored to define a pre-processing scheme or a weighted risk. This exact methodology can be applied to several characterization results in the fairness literature [MW18; Chz+19; HZ24; Hu+25; YHC25]. Such characterizations also serve as inspiration for instance-level debiasing scores [CKL23] and group-level flipping mechanisms such as Randomized response [WZC24].

§ 4.2 Comparing Pre-/In-/Post-Processing Fair Classifiers under Data Bias

In the previous section, we highlighted the theoretical equivalence of obtaining the fair Bayes optimal classifiers via different mechanisms. However, in practice, implementation may differ due to the number of available samples, access to the model, and the relaxation of constraints, among other factors. We will now empirically compare the performance of several widely used fair machine learning algorithms, especially when the underlying unfairness stems from group-dependent data biases and differences between the training and test distributions.

Many works in fair classification have studied the failure of fairness and accuracy guarantees under distribution shifts (e.g., covariate shift, label shift) between the training and test distributions and proposed fixes under reasonable assumptions [Sch+19; DB20; Rez+21; Sin+21; Mai+21; DW21; BM21; Sch+22]. In the following sections, we consider a simple and natural data bias model that incorporates under-representation and label biases [BS19], to study the change in accuracy and fairness of various classifiers as we inject bias in their training data. The central question we ask is: *Among the various fair classifiers available in the standard fairness toolkits (e.g., [Bel+18]), which ones are less vulnerable to data bias?*

The focus of the analysis that follows is group-fair classification, which aims to ensure equal or near-equal outcomes across different sensitive groups (e.g., race, gender). Examples of group fairness

include Equal Opportunity (Definition 2.4) and Demographic Parity (Definition 2.3). Various bias mitigation techniques have been proposed for group-fair classification motivated by different applications and stages of the machine learning pipeline [Zli15; Fri+19; Meh+21]. They are broadly categorized into three types: *Pre-Processing*, where one transforms the input data, independent of the subsequent stages of training, and validation [KC12; Fel+15; Cal+17], *In-Processing*, where models are trained by fairness-constrained optimization [CV10; Kam+12; Zem+13; Zaf+17b; ZLM18; Aga+18; Cel+19], and *Post-Processing*, where the output predictions are adjusted afterward to improve fairness [KKZ12; HPS16b; Ple+17]. These techniques appear side by side in standard fairness toolkits, with little normative guidance on which to choose. Furthermore, it is unclear what happens to such standard fair classifiers when they are trained and exposed to varying amounts of data bias over time in the training or deployment environments. We aim to close this gap by providing an auditing method for comparing different fair classifiers under a simple model of data bias.

To examine the effect of data bias on group-fair classifiers, we use a theoretical model by Blum & Stangl (Example 3.1 and Blum and Stangl [BS19]) for under-representation and label bias, introduced in the previous chapter. It is a simple model for demographic under-/over-sampling and implicit label bias seen in real-world data [Sha+17; BG18; WHM19; GK06; KR18]. Given a train-test data split, we create biased training data by injecting under-representation and label bias in the training data (but not the test data). We then study the error rates and (un)fairness metrics of various pre-, in-, and post-processing fair classifiers on the original test data, as we vary the amount of bias in their training data. We inject two types of under-representation biases in the training data: β_p-bias, where we undersample the positive labeled population of the minority group by a multiplicative factor of β_p, and β_n-bias, where we undersample the negative labeled population of the minority group by a multiplicative factor of β_n. We examine the effect of increasing the parameters β_p and β_n separately as well as together. We also investigate the effect of label bias parameter ν, where we flip the positive labels of the minority group to negative ones with probability ν. We cannot hope to re-create real-world data bias scenarios via this synthetic pipeline [BG18; WHM19], but our findings on stability/instability of various classifiers can be somewhat useful towards designing fair pipelines in the presence of various data biases.

Our main contributions and observations are as follows.

- Using synthetic and real-world datasets (Adult, German Credit, Bank Marketing, COMPAS), we show that the fairness and accuracy guarantees of many fair classifiers from standard fairness toolkits are highly vulnerable to under-representation and label bias. In fact, often, an unfair classifier (e.g., logistic regression classifier) trained on the correct data can be more accurate and fairer than fair classifiers trained on biased data.
- Some fair classification techniques (viz., reweighing [KC12], exponentiated gradients [Aga+18], adaptive reweighing [JN20]) have stable accuracy and fairness guarantees even when their training data is injected with under-representation and label bias. We provide a theoretical justification to support the empirically observed stability of the reweighing classifier [KC12]; see Theorem 4.3.

Section 4.2.1 explains related work to set the context for our work. Section 4.3 explains our experimental setup, and Section 4.4 discusses our key empirical observations. Section 4.5 contains some relevant theorems and proof outlines. Section 4.6 discusses our work in the context of other research directions on data bias in fair machine learning.

§ 4.2.1 Related Work

Prior Work on Distribution Shifts: The robustness of ML models when there is a mismatch between the training and test distributions has been studied under various theoretical models of distribution shifts, e.g., covariate shift and label shift. Several works have studied fair classification subject to these distribution shifts and proposed solutions under reasonable assumptions on the data distribution [Cos+19; Sch+19; BS19; Rez+21; Sin+21; DB20; BM21; DW21; Mai+21; Gig+22].

Coston et al. [Cos+19] study the problem of fair transfer learning with covariate shift under two conditions: availability of protected attributes in the source or target data, and propose to solve these problems by weighted classification and using loss functions with twice differentiable score disparity terms. Schumann et al. [Sch+19] present theoretical guarantees on different scenarios of fair transfer learning for equalized odds and equal opportunity through a domain adaptation framework. Rezaei et al. [Rez+21] propose an adversarial framework to learn fair predictors in the presence of covariate shift by formulating a game between choosing fair minimizing and maximizing target conditional label distributions, formulating a convex optimization problem which they ultimately solve using batch gradient descent. Singh et al. [Sin+21] propose to learn stable models with respect to fairness and accuracy in the presence of shifts in data distribution using a causal graph modeling the shift, using the framework of causal domain adaptation, deriving solutions that are worst-case optimal for some kinds of covariate shift. Dai et al. [DB20] talk about the impact of label bias and label shift on standard fair classification algorithms. They study label bias and label shift through a common framework that assumes the availability of a trained, fair classifier on true labels and analyzes its performance at test time when either the true label distribution changes or the new test-time labels are biased. Biswas et al. [BM21] focus on addressing the effects of prior probability shifts on fairness and propose an algorithm and a metric called Prevalence Difference, which, when minimized, ensures Proportional Equality.

Our study builds on the under-representation and label bias model proposed in Blum & Stangl [BS19]. Under this model, they prove that the optimal fair classifier that maximizes accuracy subject to fairness constraints (equal opportunity constraints, to be precise) on the biased data distribution also maximizes accuracy on the unbiased data distribution. Along similar lines, Maity et al. [Mai+21] derive conditions under which fairness constraints in the presence of subpopulation shifts yield better accuracy on the test distribution. Jiang et al. [JN20] have also considered the label bias model, although we examine the aggregate effect across varying levels of label bias. Our observations on label bias also corroborate various findings by other works investigating the role of label noise in fair classification [Lam+19; FCG20; WLL21; KL22]. Under-representation bias can also be studied as a targeted Sample Selection Bias mechanism [Cor+08]. Some work has been done on fair sample selection bias [DW21; Zhu+23], which we elaborate on next.

Du et al. [DW21] propose a framework for training classifiers robust to sample selection bias using reweighing and minimax robust estimation. The selection of samples is indicated by $\mathcal{S}$. The aim is to train using a part of the distribution where $\mathcal{S} = 1$ but generalize well for fairness and accuracy on the overall population (both $\mathcal{S} = 0$ and $\mathcal{S} = 1$). Du et al. [DW21] show that robust and fair classifiers on the overall population can be obtained by solving a reweighed optimization problem. Giguere et al. [Gig+22] consider a demographic shift in which a certain group becomes more or less probable in the test distribution than in the training distribution. They output a classifier that retains fairness guarantees for given user-specified bounds on these probabilities. Zhu et al. [Zhu+23] study the fair sample selection bias problem and propose the CRAB framework, which constructs a causal graph based on the background knowledge about the data collection process and external data sources.

Prior Work on Comparative Analysis of Fair Methods: We compare various fair classifiers in the presence of under-representation and label bias. Our work complements the surveys in fair machine

learning that compare different methods either conceptually [Zli15; Meh+21; PS22; Fen+22] or empirically, on important aspects like the choice of data processing, splits, data errors, and data pre-processing choices [Fri+19; BM21; Qia+21; Isl+22]. Friedler et al. [Fri+19] compare several fair classification methods with respect to how data is prepared and preprocessed for training, how different fairness metrics correlate with each other, stability with respect to slight changes in the training data, and the use of multiple sensitive attributes. Islam et al. [Isl+22] conceptually compare various fairness metrics and further use that to empirically compare 13 fair classification approaches on their fairness-accuracy tradeoffs, scalability with respect to the size of the dataset and dimensionality, robustness against training data errors such as swapped and missing feature values, and variability of results with multiple training data partitions. On the other hand, we compare various classifiers with respect to changes in the distribution, while keeping most aspects, such as base classifiers and data preprocessing, the same across methods.

Parallel to our work, Akpinar et al. [Akp+22] propose a simulation toolbox based on the under-representation setting from Blum & Stangl [BS19] to stress test four different fairness interventions with user-controlled synthetic data and data biases. On the other hand, our work uses the under-representation framework from Blum & Stangl [BS19] to extensively examine the stability of different types of fair classifiers on various synthetic and real-world datasets. While Akpinar et al. [Akp+22] focus on proposing a simulation framework to extensively test all findings of Blum & Stangl [BS19], we focus on using the under-representation framework to highlight the differences in performance between various fair classifiers and theoretically investigating the stability of one of the fair classifiers.

§ 4.3 Experimental Setup

For our analysis, we select a comprehensive suite of commonly used fair classifiers implemented in the open-source AI Fairness 360 (AIF-360) toolkit [Bel+18]. From the available choice of pre-processing classifiers, we include two: 1. Reweighing ('rew') [KC12], and 2. Adaptive Reweighing ('jiang_nachum') [JN20]. Since 'jiang_nachum' is not a part of AIF-360, we use the original implementation[1]. Among the in-processing classifiers in AIF-360, we choose three: 1. Prejudice Remover classifier ('prej_remover') [Kam+12], 2. Exponentiated Gradient Reduction classifier ('exp_grad'), and its deterministic version ('grid_search') [Aga+18] and 3. Kearns et al. ('gerry_fair') [Kea+18]. From post-processing classifiers, we choose three: 1. the Reject Option classifier ('reject') [KKZ12], 2. the Equalized Odds classifier ('eq') [HPS16b], and 3. the Calibrated Equalized Odds algorithm ('cal_eq') [Ple+17]. We omitted some classifiers from the AIF-360 toolkit for two reasons: resource constraints and their extreme sensitivity to hyperparameter choices.

All fair classifiers use a base classifier in their optimization routine to mitigate unfairness. We experiment with two base classifiers: a logistic regression (LR) model and a Linear SVM classifier with Platt's scaling [Pla+99] for probability outputs. We use the Scikit-learn toolkit to implement the base classifiers [Ped+11]. The 'prej_remover' and 'gerry_fair' classifiers do not support SVM as a base classifier; hence, we only show results for those on LR.

The fairness metric we use in the following sections is Equalized Odds (EODDS) [HPS16b], which is defined as follows:

[1] https://github.com/google-research/google-research/tree/master/label_bias

Definition 4.1. For a distribution D defined over $\mathcal{X} \times \{0,1\} \times \{0,1\}$, a classifier $h : \mathcal{X} \times \{0,1\} \to \{0,1\}$ satisfies *Equalized odds* if $\Pr(h(X,A) = 1|Y = 1, A = 0) = \Pr(h(X,A) = 1|Y = 1, A = 1)$ (Equality of true positive rates), and $\Pr(h(X,A) = 0|Y = 0, A = 0) = \Pr(h(X,A) = 0|Y = 0, A = 1)$ (Equality of true negative rates).

Blum & Stangl [BS19] suggest that the equal opportunity (Definition 2.4) and equalized odds constraints on biased data can be more beneficial than demographic parity constraints (Definition 2.3). Therefore, we use the equalized odds constraint for 'jiang_nachum' and 'exp_grad'. We report Equalized Odds results in the main text and report average Equal Opportunity Difference (EOD) and Statistical Parity Difference (SPD) results in Section 4.8.4.

We train on multiple datasets to perform our analysis. We consider four standard real-world datasets from fair classification literature: the Adult Income dataset [DG17], the Bank Marketing dataset [MCR14], the COMPAS dataset [Ang+16], and the German Credit dataset [DG17]. We also consider a synthetic dataset setup from Zeng et al. [ZDC22a], which consists of binary labels and sensitive attributes. We use the existing standard train-test splits for a given dataset, or, if unavailable, a 70%-30% stratified split across subgroups. Section 4.8.1 gives specific details for all datasets.

To perform our analysis, we use the data bias model from Blum & Stangl [BS19] and inject varying amounts of under-representation and label bias into the original training data before giving it to a classifier. We summarize our experimental setup in Algorithm 4.1, presented in Section 4.8.2. In essence, we inject under-representation bias into the training data by retaining samples from the 10-subgroup and the 00-subgroup with probability β_p and β_n, respectively, where ya denotes the subgroup defined by label y and group a. We consider ten different values each for β_p and β_n varying over $\{0.1, 0.2, ..., 1.0\}$. This results in 100 different settings (10 β_p factors x 10 β_n factors) for training data bias. We separately inject ten different levels of label bias by flipping the favorable label of the underprivileged group to an unfavorable one ($10 \to 00$) with a probability ν varying over $\{0.0, 0.1, ..., 0.9\}$. The test set for all settings is the original split test set, either provided with the original dataset or extracted separately beforehand. This results in 110 datasets corresponding to different data bias settings, and consequently 110 different results for each particular fair classifier. Finally, training all classifiers and procedures for creating biased training data using β_p, β_n, and ν is repeated 5 times to account for sampling and optimization randomness.

We also note that some fair classifiers on our list, such as 'exp_grad' and 'cal_eq', are randomized. Therefore, repeating the entire pipeline multiple times and reporting means and standard deviations normalizes the random behavior of these methods to some extent, apart from the measures taken during implementation in widely used toolkits like the one we are using, AIF-360 [Bel+18]. The code to train and reproduce the results is available here [2].

§ 4.4 Stability of Fair Classifiers: Under Representation and Label Bias

We now present our experimental results on the stability of the fairness and accuracy guarantees of various fair classifiers after injecting under-representation and label bias into their training data. Stability means whether the error rate and unfairness exhibit high variance/spread in response to increases or decreases in under-representation or label bias. A stable behavior is indicated by small changes in error

[2] https://github.com/mohitsharma29/comparision_data_bias_fairness

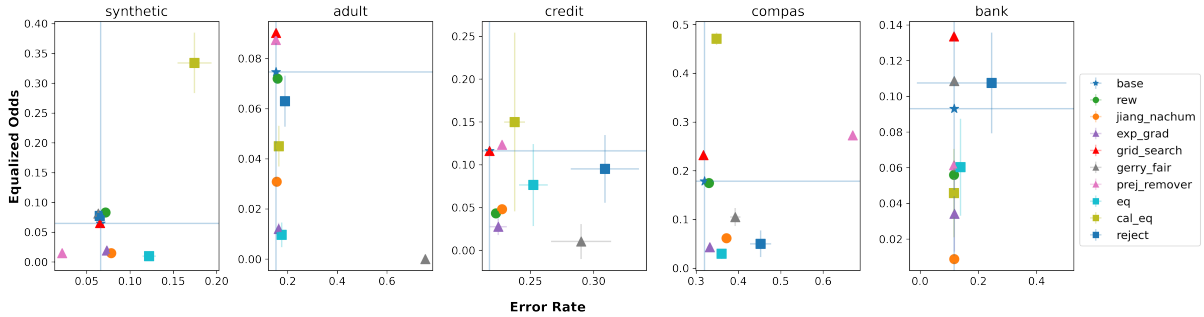


Figure 4.1: Error Rate-EODDS values for various classifiers on all datasets, with no explicit under-representation and label bias. The blue horizontal and vertical lines denote the error rate and EODDS of a base LR model without any fairness constraints.

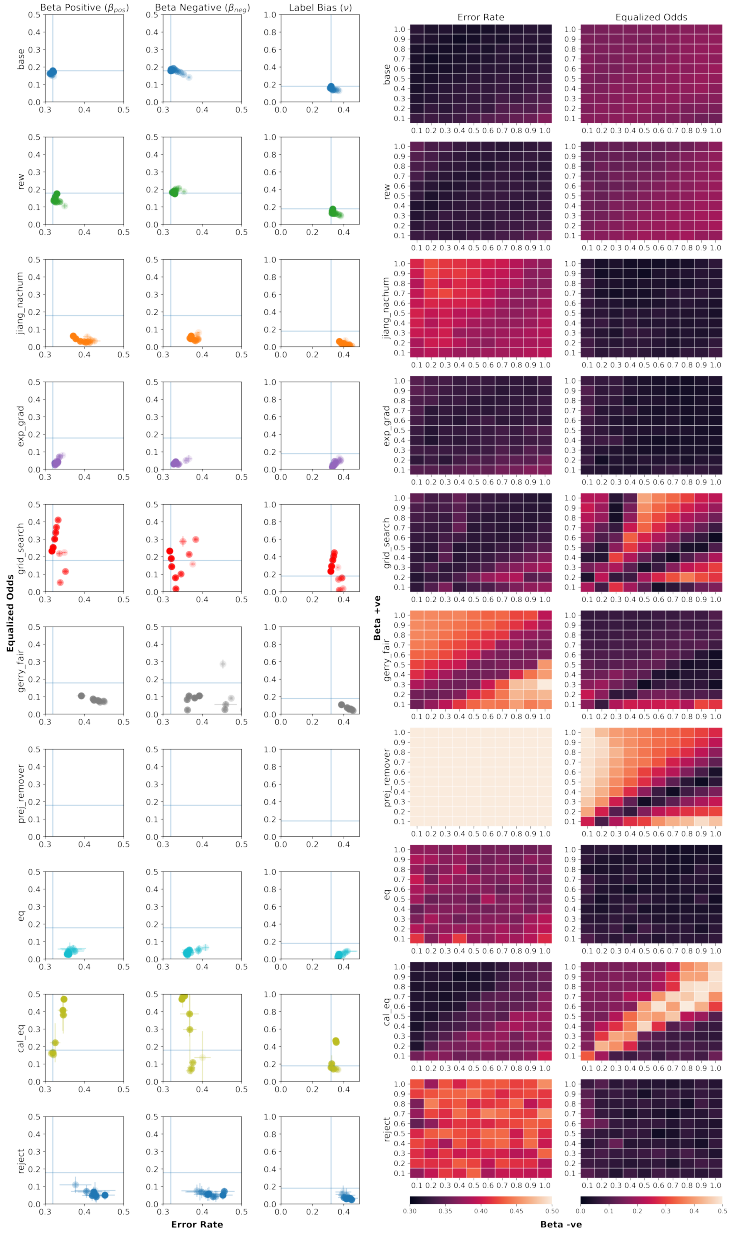
rate and unfairness, as β_p, β_n or ν change.

We first plot the unfairness (Equalized Odds) and error rates of various fair classifiers on the original training splits. Figure 4.1 shows the results of various fair classifiers with a logistic regression model on original dataset splits without any explicitly injected under-representation or label bias. While most fair classifiers exhibit an expected behavior of trading off some accuracy to minimize unfairness, some classifiers perform worse in mitigating unfairness than the unfair LR model. We attribute this to the strong effects of different data preprocessing methods on the performance of fair classifiers, as noted in previous work [Fri+19; BR21a]. Other results with SVM as a base classifier, Equal Opportunity Difference, and Statistical Parity Difference are given in Section 4.8.3.

We can also inspect Figure 4.1 and other figures later in this section by looking at the four quadrants formed by the blue horizontal and vertical lines corresponding to the unfairness and error rate of the base LR model. Following a Cartesian coordinate system, most fair classifiers, when trained on the original training set without explicit under-representation or label bias, lie in the fourth quadrant, as they trade off some accuracy to mitigate unfairness. Any method in the first quadrant is poor, as it performs worse than the base LR model on both error rate and unfairness. Finally, any method in the third quadrant is excellent, as it improves both error rate and fairness relative to the LR model.

Figure 4.1 shows how various fair classifiers perform without explicit under-representation and label bias. In this figure, we summarize all available results for a given dataset in various ways to get a holistic picture. We show the stability of error rates and Equalized Odds unfairness and look at the average stability of the other two metrics: Equal Opportunity and Statistical Parity Difference, since many times in practice, we care about the performance of methods against a fixed set of metrics. We now look at the individual effects of both under-representation factors (β_p, β_n and label bias (ν)), with the LR model as a reference, denoted by the vertical and the horizontal error rate and Equal Opportunity Difference (EOD) lines, respectively. Figure 4.2(a) shows those results for the Compas dataset. The results for the Adult, Synthetic, Bank Marketing, and Credit datasets are given in Section 4.8.4. We first observe that many fair classifiers exhibit a large variance in the unfairness-error rate plane and often become unfair as the bias factors increase. This is also indicated by how the fair classifiers jump from quadrant four to quadrant one, as β_p, β_n, or ν increase. In fact, for the 'prej_remover' classifier, the results lie outside the plot limits, indicating its instability. Even when its average SPD is measured in Figure 4.2(c) ('prej_remover' is implicitly optimized for SPD), we observe that it does not perform well.

However, 'rew', 'jiang_nachum', and 'exp_grad' fair classifiers remain stable across all three kinds of biases for both error rate and EOD and show a minimal spread across both the error rate and unfairness



(a) (b)

Algo.		Err.	SPD	EOD
base	lr	**0.324 ± 0.011**	0.206 ± **0.011**	0.126 ± 0.024
	svm	**0.338 ± 0.013**	0.186 ± **0.014**	0.106 ± 0.028
rew	lr	**0.327 ± 0.006**	0.198 ± 0.026	0.123 ± 0.037
	svm	**0.343 ± 0.008**	0.177 ± 0.027	0.101 ± 0.04
jiang nachum	lr	0.39 ± 0.014	**0.066 ± 0.017**	**0.018 ± 0.009**
	svm	0.389 ± **0.011**	**0.068 ± 0.018**	**0.016 ± 0.006**
exp grad	lr	**0.333 ± 0.009**	**0.07** ± 0.025	**0.029 ± 0.022**
	svm	0.354 ± 0.017	**0.057** ± 0.02	**0.029 ± 0.021**
grid search	lr	0.336 ± 0.016	0.194 ± 0.107	0.163 ± 0.096
	svm	0.358 ± 0.025	0.163 ± 0.092	0.13 ± 0.089
gerry fair	lr	0.41 ± 0.037	0.113 ± 0.057	0.097 ± 0.082
	svm	-	-	-
prej remover	lr	0.654 ± 0.024	0.303 ± 0.137	0.246 ± 0.118
	svm	-	-	-
eq	lr	0.37 ± 0.012	**0.05 ± 0.016**	0.044 ± **0.021**
	svm	0.381 ± 0.014	**0.046 ± 0.018**	0.043 ± **0.022**
cal_eq	lr	0.343 ± 0.022	0.251 ± 0.141	0.171 ± 0.092
	svm	**0.353** ± 0.022	0.216 ± 0.114	0.144 ± 0.083
reject	lr	0.415 ± 0.019	0.079 ± 0.027	**0.038** ± 0.019
	svm	0.413 ± 0.021	0.081 ± 0.026	**0.041** ± 0.018

(c)

Figure 4.2: Stability analysis of various fair classifiers on the Compas dataset (Lighter the shade, more the under-representation and label bias): (a) The spread of error rates and Equalized Odds(EODDS) of various fair classifiers as we vary β_p, β_n and label bias separately. The vertical and horizontal reference blue lines denote the performance of an unfair logistic regression model on the original dataset without any under-representation or label bias. (b) Heatmap for Error Rate and EODDS across all different settings of β_p and β_n. Darker values denote lower error rates and unfairness. Uniform values across the grid indicate stability to different β_p and β_n. (c) Mean error rates, Statistical Parity Difference (SPD), and Equal Opportunity Difference (EOD) across all β_p and β_n settings for both kinds of base classifiers (SVM and LR).

(EOD) axes. They tend to remain in the same quadrant with varying bias factors. This remains true on other datasets as well (as shown in Section 4.8.4). Furthermore, it is interesting to note that a simple logistic regression model tends to stay relatively stable in the presence of under-representation and label biases, and on many occasions, more fair than many fair classifiers in our analysis.

We can also look at the combined effect of β_p and β_n in Figure 4.2(b) with the help of heatmaps. Each cell in the heatmap denotes a possible β_p, β_n setting, thus covering all the 100 possible settings. Darker color values in cells denote lower error rates and unfairness. For this analysis, to examine stability with heatmaps, we look for uniformity of colors across the entire grid, meaning that the error rate and unfairness values do not change with different β_p and β_n values.

Figure 4.2(b) again confirm the stability of 'rew', 'jiang_nachum' and 'exp_grad' fair classifiers even when they are trained on datasets with combined β_p and β_n biases. However, 'jiang_nachum' and 'exp_grad' emerge as stable and better classifiers in terms of their unfairness values than the 'rew' classifier because they have darker uniform colors across their entire error rate and unfairness grid. The 'eq' classifier also remains stable for unfairness, except on the Credit dataset (shown in Section 4.8.4). Amongst the classifiers optimized for Equalized Odds ('jiang_nachum', 'exp_grad', 'eq', and 'cal_eq'), 'jiang_nachum' and 'exp_grad' emerge as the most stable ones in terms of error rate and unfairness.

To quantitatively summarize the findings from the heatmap plots and to show this statistic with other choices of base classifiers (LR and SVM) and unfairness metrics (SPD and EOD), we look at the average performance across 100 different settings of β_p and β_n in Figure 4.2(c) for the Compas dataset. An indication of stability in these results is whether the mean error rate and unfairness across 100 different under-representation settings remain low, and whether the mean performance has a low standard deviation. Because we run the entire experimental setup 5 times with different random seeds for reproducibility, the values represent the mean of the means across 5 runs for each β_p, β_n setting, along with its standard deviation across the 100 settings. We report the top-3 mean and top-3 standard deviation values for each column, where the top-3 means the lowest 3 error rates and the lowest unfairness metrics (SPD and EOD).

Across all datasets, the observations for the emboldened values corroborate our findings from the heatmap plots. For both unfairness metrics (SPD and EOD), 'exp_grad' and 'jiang_nachum' often appear in the top-3 mean performance across all settings, whereas 'rew', 'jiang_nachum' and 'exp_grad' often appear in the top-3 smallest standard deviation values, indicating their stability. Furthermore, 'rew' and 'jiang_nachum' often appear in the top-3 mean and standard deviation list for error rate. 'rew' often appears in the top-3 mean performance for Statistical Parity Difference (SPD). As expected, the unfair base classifier often appears in the top-3 list of mean error rates, since its objective is always to minimize error rates without regard for unfairness. Finally, when looking over all datasets' results, we find that 'exp_grad' and 'jiang_nachum' often emerge as the most stable classifiers across metrics like EOD and SPD, despite not being trained to minimize those metrics, further solidifying their stable behavior.

Another interesting observation across the results of all datasets is that the deterministic version of the 'exp_grad' fair algorithm, the 'grid_search' classifier [Aga+18], is very unstable compared to the 'exp_grad' classifier, which is the most stable amongst the lot, indicating that in the face of data bias, randomness might help. This observation of randomized methods exhibiting more robustness than deterministic ones is corroborated by recent studies on the robustness of randomized fair Bayes optimal classifiers [AD22; Aga+; JD25].

Finally, it is worth noting that the stability of 'rew' and 'exp_grad' has also been reported in another benchmarking study about testing various fair classifiers in the presence of varying sensitive attribute

noise [GKW23]. This indicates towards a potentially stable performance from reweighing-based approaches in the presence of varying kinds of data biases and noises in the training data.

§ 4.5 Stability Theorems for Reweighing Classifier

The following theorems give a theoretical justification for the stability of test accuracy when the reweighing classifier [KC12] is trained even on data injected with extreme under-representation such as $\beta_p \to 0$ or $\beta_n \to 0$.

Theorem 4.3. Let D be the original data distribution and $\beta \in (0, 1]$. Let D_β be the biased data distribution after injecting under-representation bias in D using any $\beta_p = \beta$ (or similarly any $\beta_n = \beta$). Let h be any binary classifier and $L(h)$ be any non-negative loss function on the prediction of h and the true label. Let L' denote the reweighed loss according to Kamiran et al. [KC12]. Then

$$\frac{\alpha^2}{4} E_D[L(h)] \leqslant E_{D_\beta}[L'(h)] \leqslant \frac{4}{\alpha} E_D[L(h)],$$

where $\alpha = \frac{\min_{ij} \Pr(Y=i,A=j)}{\max_{ij} \Pr(Y=i,A=j)}$ denotes the subgroup imbalance in the original distribution D.

The proof for this theorem is given in Section 4.8.5. The proof involves bounding subgroup and group/label probabilities using α, and with a reasonable assumption that $X'|Y' = y, A' = a$ in $D_\beta \sim X|Y = y, A = a$ in D. The main takeaway from Theorem 4.3 is the following: when we reweigh a given non-negative loss function using the Reweighing scheme from Kamiran & Calders [KC12] on samples from D_β, we always lie in some constant 'radius' of the true expected loss on D for any classifier h. This does not tell us anything about the expected loss on D itself, but rather, about the ability of the reweighed loss L' to recover from any arbitrary under-representation bias β. However, the constant factors depend on α, the worst-case imbalance between the subgroups for a given distribution.

Remark 4.5.1. Whenever $E_D[L(h)]$ is small, Theorem 4.3 says that the expected reweighed loss on the biased distribution $E_{D_\beta}[L'(h)]$ will also be small, for any classifier h. This means that if we are given a non-negative loss function that also subsumes a given fairness constraint (let's call it L_{fair}), then reweighing with L_{fair} can yield low unfairness and stable classifiers.

Using Theorem 4.3, we prove the following bound on the expected loss of the reweighing classifier that is obtained by empirical minimization of the reweighed loss on the biased distribution D_β but tested on the original distribution D.

Theorem 4.4. For any $\beta \in (0, 1]$, let $\hat{h}_\beta$ be the classifier obtained by empirical minimization of the reweighed loss L' on N samples from the biased distribution D_β defined as in Theorem 4.3. Then, with probability at least $1 - \delta$, we have

$$E_D[L(\hat{h}_\beta)] \leqslant \frac{16}{\alpha^3 \beta} \sqrt{\frac{\ln(2/\delta)}{2N}} + \frac{16}{\alpha^3} E_D[L(h^*)],$$

where h^* is the Bayes optimal classifier on the original distribution D, α is the subgroup imbalance as in Theorem 4.3.

The proof for this theorem is given in Section 4.8.5. It uses a generalized version of Hoeffding's inequality [Hoe63], proof of Theorem 2 of Zhu et al. [ZLL21] and Theorem 4.3. Note that Theorem 4.4

implies that given any $\epsilon > 0$ and $\delta > 0$, we can get the guarantee

$$E_D[L(\hat{h}_\beta)] \leqslant \frac{16}{\alpha^3} E_D[L(h^*)] + \epsilon,$$

with probability at least $1 - \delta$, by simply choosing the number of samples N from D_β in the empirical reweighed loss minimization to be large enough so that

$$N \geqslant \frac{128 \log(2/\delta)}{\alpha^6 \beta^2 \epsilon^2}.$$

It is important to note that Theorems 4.3 and 4.4 hold for data without any injected bias, i.e., $\beta_p = \beta_n = 1$, where such theoretical guarantees for the reweighing classifier [KC12] were not known earlier.

§ 4.6 Discussion

In this section, we discuss where our results stand in the context of related works on distribution shifts and data bias in real-world AI/ML models and make a few practical recommendations based on our work.

§ 4.6.1 Under-Representation and Distribution Shifts

Let Pr_{train} and Pr_{test} define the probabilities for training and testing distributions, respectively, with a random data point denoted by (X, A, Y) with features X, sensitive attribute A, and class label Y. Let $Y = 1$ be the favorable label and $A = 0$ be the underprivileged group. A covariate shift in fair classification is defined as $Pr_{train}(Y|X, A) = Pr_{test}(Y|X, A)$ [Rez+21]. Singh et al. [Sin+21] look at fairness and covariate shift from a causal perspective using the notion of a separating set of features that induce a covariate shift. Coston et al. [Cos+19] look at covariate shift and domain adaptation when sensitive attributes are not available at source (train) or target (test).

Distribution shifts can also be induced via class labels. Dai et al. [DB20] present a more general model for label bias, building upon the work of Blum et al. [BS19]. They present a model for label shift where $Pr_{train}(Y) \neq Pr_{test}(Y)$, but $Pr_{train}(X|Y, A) = Pr_{test}(X|Y, A)$. Biswas et al. [BM21] study model shifts in prior probability, where $Pr_{train}(Y = 1|A) \neq Pr_{test}(Y = 1|A)$ but $Pr_{train}(X|Y = 1, A) = Pr_{test}(X|Y = 1, A)$. Our work is also different from sample selection bias [Cor+08] (selection of training samples at random vs. sensitive group dependent under-sampling and label flipping), but may have some connections to fair sample selection bias [DW21], where the objective is to train only on selected samples($S = 1$), but generalize well for fairness and accuracy on the overall population (unselected samples $S = 0$ and $S = 1$). An interesting question in this regard is how to design fair classifiers that are stable against a range of malicious sample-selection biases.

The under-representation model used in this chapter comes from Blum et al. [BS19] and can be thought of as the following distribution shift: $Pr_{train}(Y = 0, A = 0) = \beta_n \cdot Pr_{test}(Y = 0, A = 0)$ and $Pr_{train}(Y = 1, A = 0) = \beta_p \cdot Pr_{test}(Y = 1, A = 0)$, while the other two subgroups $(Y = 0, A = 1)$ and $(Y = 1, A = 1)$ are left untouched. Overall, the distribution of (X, A) remains unchanged, while the distribution of Y changes only for the underprivileged group $A = 0$. The above distribution shift is different from covariate and label shifts in that it affects both the joint marginal $P(X, A)$ and $P(X|Y, A)$.

In a broader sense, beyond the mathematical definition used in our work, under-representation of a demographic and implicit bias in the labels are known problems observed in real-world data. Shankar et al. [Sha+17] highlight geographic over-representation issues in public image datasets and

how this can harm use cases involving prediction tasks in developing countries. Buolamwini et al. [BG18] highlight a similar under-representation issue with gender classification systems trained on over-represented lighter-skinned individuals. Wilson et al. [WHM19] highlight similar disparities for pedestrian detection tasks. Biased labels in the training data have also been observed in the context of implicit bias in the literature [GK06; KR18].

§ 4.6.2 Practical Recommendations

We show that the performance guarantees of commonly used fair classifiers can be highly unstable when their training data is under-represented and exhibits label bias. Our experimental setup serves as a dashboard template to assess the vulnerability of fair classifiers to data bias by injecting bias using simple models similar to Blum & Stangl [BS19].

Our work emphasizes the importance of creating test suites to assess robustness to data bias and incorporating them with modern fairness toolkits such as AIF-360 [Bel+18]. Our results complement previous studies on the stability across different train-test splits [Fri+19] or different data pre-processing strategies [BR21a]. We experiment with a range of β_p, β_n and ν factors. In practice, this can be done with a user-supplied range of under-representation and label bias relevant for specific use cases. The ability to measure the stability of chosen fairness metrics across dynamically specified bias factors can be a great first step towards the safe deployment of fair classifiers, similar to works such as Shifty [Gig+22].

Our results show that some classifiers like 'jiang_nachum' [JN20] and 'exp_grad' [Aga+18] can remain relatively unscathed in the presence of varying under-representation and label biases. The stability is maintained even when they are trained and tested on mismatching fairness metrics. This motivates using such reweighing methodologies when the level of data biases our model might see is unknown.

Motivated by the same setup as Blum & Stangl [BS19], Akpinar et al. [Akp+22] also propose a toolkit and dashboard to stress-test fair algorithms in a user-controlled synthetic setup, allowing the user to control various aspects of the simulation. A handy addition to existing fairness toolkits, such as AIF-360 [Bel+18], and Fairlearn [Bir+20] can be to incorporate input distribution stability routines from our work, and Akpinar et al. [Akp+22], along with some comparative analysis routines from Friedler et al. [Fri+19] to create a practical checklist for any fairness algorithm. Finally, our experimental setup can provide an additional tool for assessing the robustness of models to data bias, alongside existing robustness test suites, e.g., CheckList [Rib+20].

§ 4.7 Conclusion

We show that many state-of-the-art fair classifiers exhibit substantial performance variance when their training data is injected with under-representation and label bias. We propose an experimental setup to compare different fair classifiers under data bias. We show how the Bayes Optimal unfair and fair classifiers behave under under-representation bias. We give a theoretical bound on the accuracy of the reweighing classifier [KC12] that holds even under extreme data bias. Finally, we discuss our work in the broader context of data bias literature and make practical recommendations.

A limitation of our work is that we use a simple model to introduce under-representation and label bias, which cannot address the root causes of different kinds of biases in real-world data. Thinking about the causal origin of data bias may allow one to model more types of bias and obtain better fixes under reasonable assumptions. Theoretically explaining the stability of the 'exp_grad' [Aga+18], and

Dataset	# Data Pts.	Outcome (Y)	Sensitive Attribute(S)	Dropped Columns
Adult [DG17]	45k	Income(>50k, <=50k)	Sex (Male, Female)	fnlwgt, education-num, race
Bank [MCR14]	41k	Term Deposit (Yes, No)	Age (<=25, >25)	poutcome, contact
Compas [Ang+16]	6k	Recidivate (Yes,No)	Race(African-American, Caucasian)	sex
Credit [DG17]	1k	Credit Risk (good, bad)	Sex (Male, Female)	age, foreign_worker
Synthetic [ZDC22a]	20k	0, 1	0, 1	NA

Table 4.1: Details of all the datasets used in our Experiments.

'jiang_nachum' [JN20] fair classifiers with under-representation bias is also an interesting future work.

§ 4.8 Additional Results and Proofs

In this section, we present the omitted details of the experiments and additional simulations, as well as the proofs of the results in Section 4.5.

§ 4.8.1 Dataset Information

Table 4.1 lists all information pertaining to the datasets used in our experiments. We choose to drop columns that either introduce many empty values or represent sensitive attributes. The reason for dropping the non-primary-sensitive attributes is to avoid unintended disparate treatment. The choice of sensitive attributes for each dataset is motivated by previous works in fair ML literature [Zaf+17b; Aga+18; Fri+19; JN20].

The synthetic dataset setup from Zeng et al. [ZDC22a] consists of a binary label $y \in \{0,1\}$ and a binary sensitive attribute setting $a \in \{0,1\}$. We sample data from subgroup-specific 100-dimensional Gaussian distributions $X \sim \mathcal{N}(\mu_{ay}, \sigma^2 I_{100})$ where I_{100} denotes a 100×100 identity matrix and all entries of μ_{ay} are sampled from $U(0,1)$. We sample $20,000$ training examples and 5000 test examples.

§ 4.8.2 Algorithm to create different Data Bias scenarios

Algorithm 4.1 gives a detailed description of how we inject varying amounts of under-representation and label bias for the considered fair classifiers.

§ 4.8.3 Original Split results for other settings

Figures 4.4, 4.5 and 4.6 show the results of all fair classifiers when we use either a Logistic Regression model or an SVM model, and measure unfairness using Statistical Parity or Equal Opportunity Difference as a metric. Note that 'gerry_fair' and 'prej_remover' are missing from Figures 4.5 and 4.6 because these two classifiers do not support SVM as a base classifier for their fairness routines.

§ 4.8.4 Stability Analysis for remaining datasets

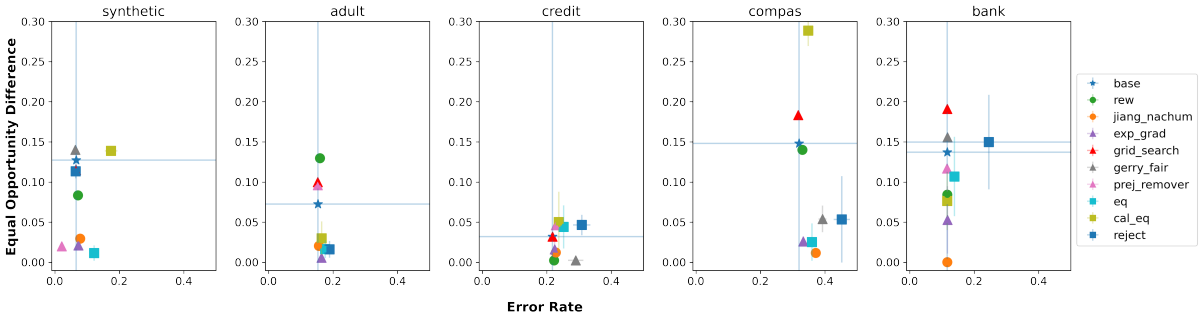


Figure 4.3: Error Rate-EOD values for various classifiers on all datasets, with no explicit under-representation and label bias. The blue horizontal and vertical lines denote the error rate and EOD of a base LR model without any fairness constraints.

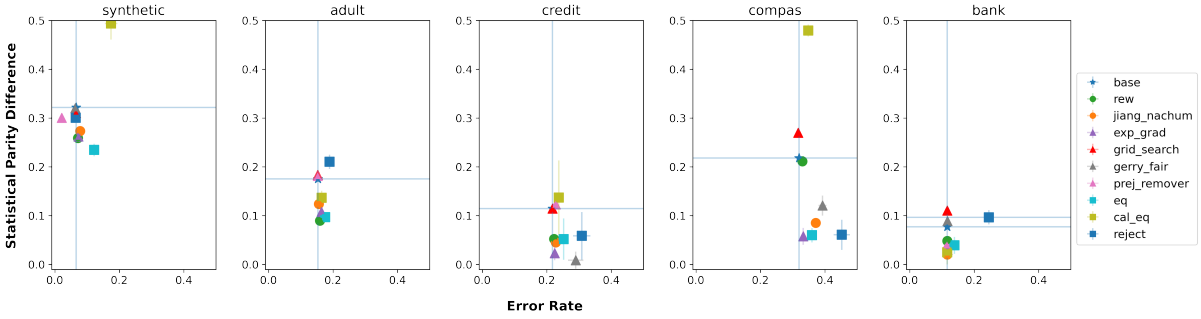


Figure 4.4: Error Rate-SPD values for various classifiers on all datasets, with no explicit under-representation and label bias. The blue horizontal and vertical lines denote the error rate and SPD of a base LR model without any fairness constraints.

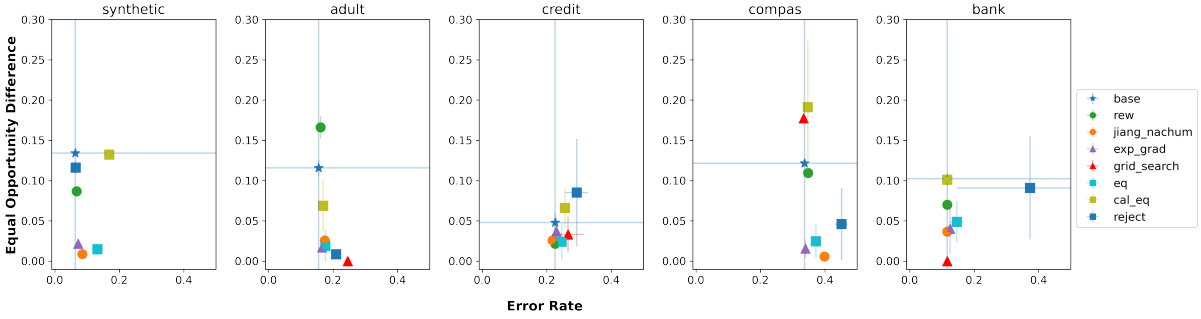


Figure 4.5: Error Rate-EOD values for various classifiers on all datasets, with no explicit under-representation and label bias. The blue horizontal and vertical lines denote the error rate and EOD of a base SVM model without any fairness constraints.

Algorithm 4.1: Experimental Pipeline for Under-Representation and Label Bias

Input: A fair classifier, or a base classifier (lr or svm without fairness constraints): cls, original training dataset: data_train, test dataset: data_test. All samples in data_train and data_test are of the form (x, y, a)

Output: Results for all under-representation and label bias setting for cls on the test set, with multiple runs.

```
all_runs_results = {}
for run ← 1 to 5 do
    under_rep_results = {}, label_bias_results = {}
    for β_p ∈ [0.1....1.0] do
        for β_n ∈ [0.1....1.0] do
            train_data(β_p, β_n) ← Randomly sample β_p proportion of (y = 1, a = 0) subpopulation and
              β_n proportion of (y = 0, a = 0) subpopulation from data_train, include other
              subpopulations same as data_train.
            under_rep_results[(β_p, β_n)] ← result for cls on train_data(β_p, β_n)
    for ν ∈ [0.0....0.9] do
        train_data(ν) ← Randomly flip ν proportion of (y = 1, a = 0) subpopulation examples to
          (y = 0, a = 0) from data_train, include other subpopulations same as data_train.
        label_bias_results[ν] ← result for cls on train_data(ν)
    all_runs_results[run] = (under_rep_results, label_bias_results)
return all_runs_results
```

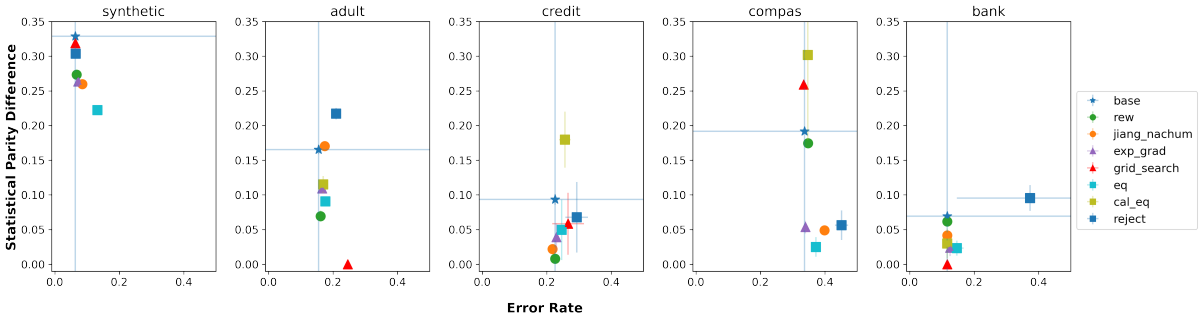


Figure 4.6: Error Rate-SPD values for various classifiers on all datasets, with no explicit under-representation and label bias. The blue horizontal and vertical lines denote the error rate and SPD of a base SVM model without any fairness constraints.

Figures 4.7, 4.8, 4.9 and 4.10 show the scatter plot, heatmap and average results for the Synthetic, Adult, Credit and Bank Marketing datasets respectively.

§ 4.8.5 Proofs for Section 4.5

Proof. (Proof for Theorem 4.3) We present the proof for the binary label y and the sensitive attribute a. The proof proceeds similarly for multiple labels and sensitive attributes.

Assume that the subgroup probabilities on D are denoted by $\Pr(Y = y, A = a) = p_{ya}$: $p_{00}, p_{01}, p_{10}, p_{11}$.

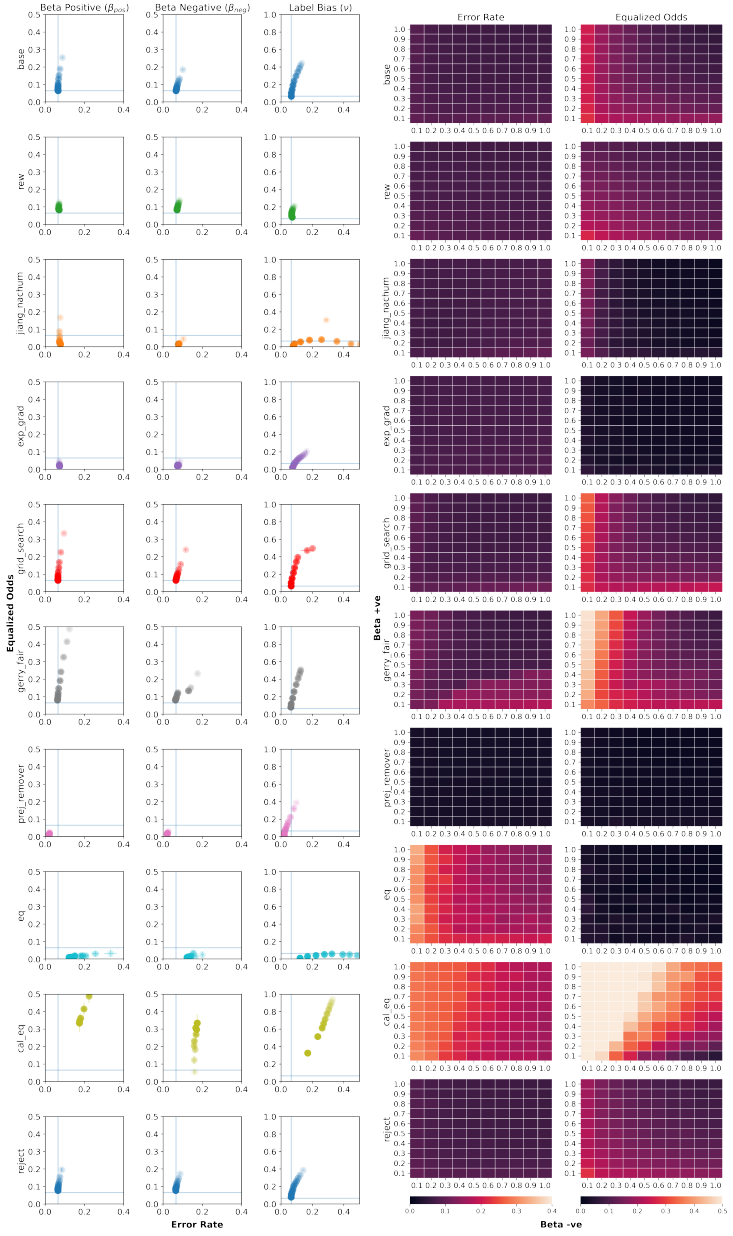
For the loss L, we can write its expectation under distribution D as:

$$\mathrm{E}_{(X,Y,A)\sim D}[L(h(X), Y)] = \sum_{y,a} \Pr(Y = y, a = a)\, \mathrm{E}_{X|Y=y,A=a}[L(h(X), y)]$$

Let $m = \min\{p_{00}, p_{01}, p_{10}, p_{11}\}$, $M = \max\{p_{00}, p_{01}, p_{10}, p_{11}\}$, and $\alpha = m/M$.

For the rest of the proof, we will assume an under-representation bias on $(Y = 1, A = 0)$ subgroup with bias β. The proof for bias on $(Y = 0, A = 0)$ subgroup will work out similarly.

We denote by D_β the distribution induced because of the under-representation bias on $(Y = 1, A =$

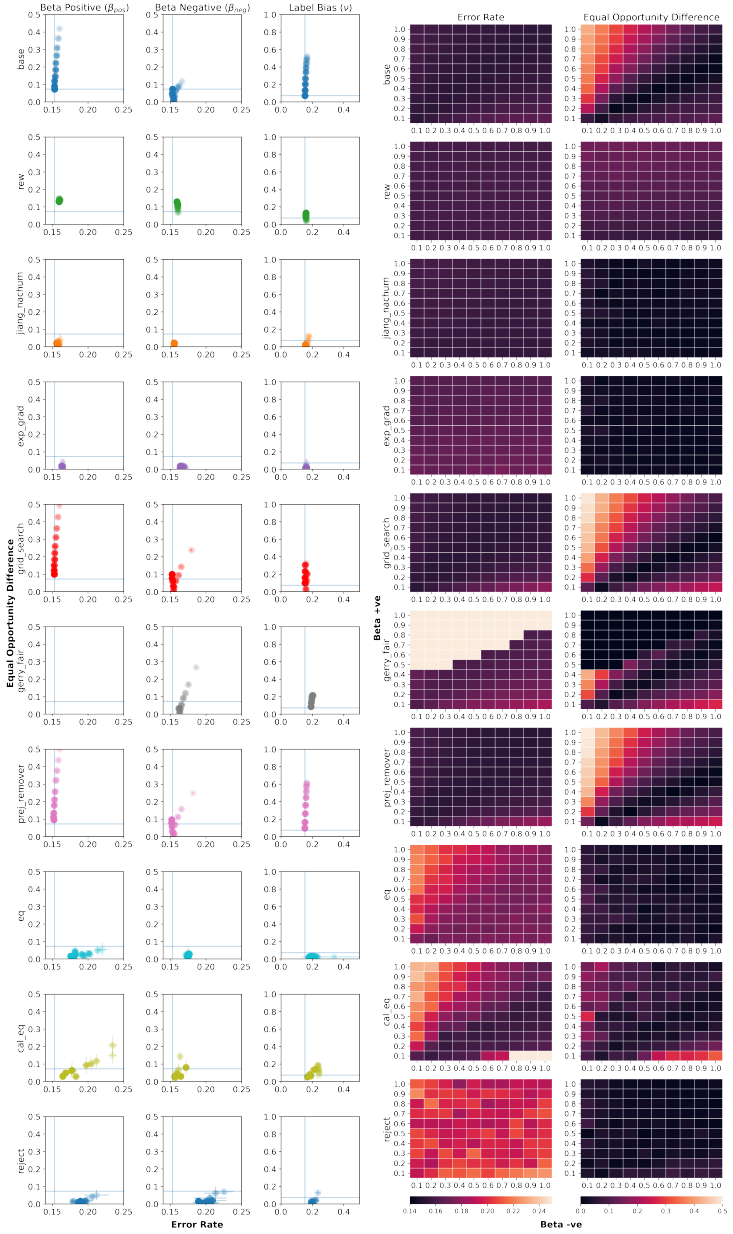


(a) (b)

Algo.		Err.	SPD	EOD
base	lr	0.075 ± 0.009	0.311 ± 0.073	0.206 ± 0.107
	svm	**0.074** ± 0.009	0.325 ± 0.074	0.218 ± 0.115
rew	lr	0.077 ± 0.006	**0.241** ± **0.018**	0.122 ± 0.033
	svm	**0.074** ± **0.006**	**0.258** ± **0.023**	0.131 ± **0.04**
jiang nachum	lr	**0.075** ± **0.006**	0.291 ± 0.041	0.086 ± 0.079
	svm	0.083 ± 0.009	0.289 ± 0.046	**0.086** ± 0.093
exp grad	lr	0.076 ± **0.003**	**0.259** ± **0.009**	**0.029** ± **0.011**
	svm	0.079 ± **0.005**	**0.257** ± **0.01**	**0.026** ± **0.01**
grid search	lr	0.078 ± 0.012	0.301 ± 0.111	0.207 ± 0.149
	svm	0.078 ± 0.013	0.315 ± 0.119	0.22 ± 0.167
gerry fair	lr	0.099 ± 0.032	0.368 ± 0.144	0.354 ± 0.245
	svm	-	-	-
prej remover	lr	**0.022** ± **0.001**	0.296 ± **0.009**	**0.022** ± **0.008**
	svm	-	-	-
eq	lr	0.192 ± 0.052	**0.187** ± 0.029	**0.025** ± **0.017**
	svm	0.191 ± 0.052	**0.188** ± **0.028**	**0.025** ± **0.017**
cal_eq	lr	0.218 ± 0.042	0.534 ± 0.132	0.219 ± 0.11
	svm	0.22 ± 0.042	0.535 ± 0.132	0.22 ± 0.111
reject	lr	**0.075** ± 0.008	0.278 ± 0.028	0.168 ± 0.043
	svm	**0.075** ± **0.008**	0.276 ± 0.028	0.167 ± **0.046**

(c)

Figure 4.7: Stability analysis of various fair classifiers on the Synthetic dataset (Lighter the shade, more the under-representation and label bias): (a) The spread of error rates and Equalized Odds (EODDS) of various fair classifiers as we vary β_p, β_n and label bias separately. The vertical and horizontal reference blue lines denote the performance of an unfair logistic regression model on the original dataset without any under-representation or label bias. (b) Heatmap for Error Rate and EODDS across all different settings of β_p and β_n. Darker values denote lower error rates and unfairness. Uniform values across the grid indicate stability to different β_p and β_n. (c) Mean error rates, Statistical Parity Difference (SPD), and Equal Opportunity Difference (EOD) across all β_p and β_n settings for both kinds of base classifiers (SVM and LR).

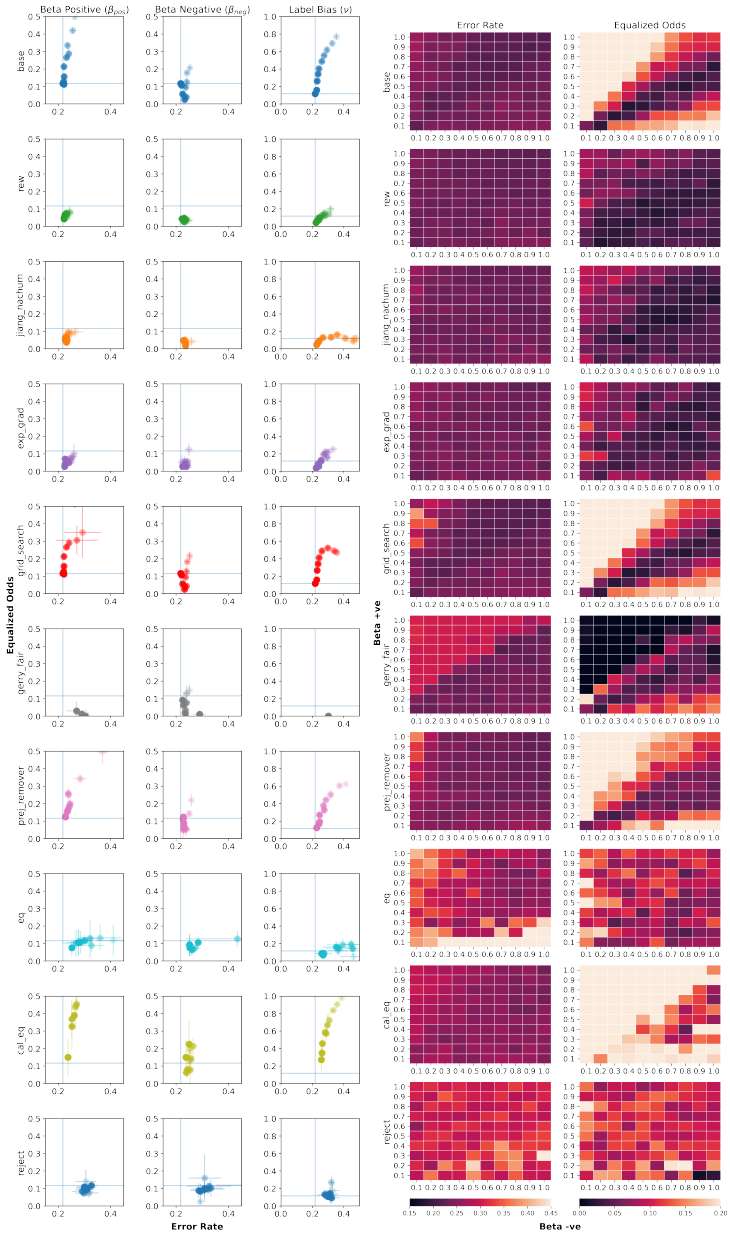


(a) (b)

Algo.		Err.	SPD	EOD
base	lr	**0.156** ± 0.003	0.171 ± 0.035	0.134 ± 0.111
	svm	**0.155** ± **0.003**	0.179 ± 0.036	0.172 ± 0.131
rew	lr	0.16 ± **0.001**	**0.101** ± **0.012**	0.109 ± 0.02
	svm	**0.159** ± **0.001**	**0.103** ± 0.02	0.105 ± 0.029
jiang nachum	lr	**0.156** ± **0.001**	0.125 ± **0.011**	**0.019** ± **0.011**
	svm	**0.159** ± 0.003	0.136 ± **0.012**	0.049 ± **0.013**
exp grad	lr	0.166 ± **0.002**	0.107 ± **0.003**	**0.016** ± **0.008**
	svm	0.168 ± **0.002**	0.109 ± **0.003**	**0.017** ± **0.008**
grid search	lr	**0.156** ± 0.005	0.176 ± 0.048	0.169 ± 0.131
	svm	0.208 ± 0.041	**0.056** ± 0.067	0.051 ± 0.064
gerry fair	lr	0.412 ± 0.287	**0.071** ± 0.068	0.064 ± 0.087
	svm	-	-	-
prej remover	lr	0.156 ± 0.005	0.174 ± 0.05	0.172 ± 0.135
	svm	-	-	-
eq	lr	0.184 ± 0.013	**0.085** ± 0.019	0.038 ± 0.02
	svm	0.182 ± 0.013	**0.085** ± **0.017**	**0.038** ± **0.021**
cal_eq	lr	0.183 ± 0.034	0.125 ± 0.075	0.088 ± 0.063
	svm	0.183 ± 0.036	0.123 ± 0.081	0.091 ± 0.066
reject	lr	0.199 ± 0.008	0.192 ± 0.025	**0.027** ± **0.018**
	svm	0.203 ± 0.008	0.191 ± 0.024	**0.027** ± **0.02**

(c)

Figure 4.8: Stability analysis of various fair classifiers on the Adult dataset (Lighter the shade, more the under-representation and label bias): (a) The spread of error rates and Equal Opportunity Difference(EOD) of various fair classifiers as we vary β_p, β_n and label bias separately. The vertical and horizontal reference blue lines denote the performance of an unfair logistic regression model on the original dataset without any under-representation or label bias. (b) Heatmap for Error Rate and EOD across all different settings of β_p and β_n. Darker values denote lower error rates and unfairness. Uniform values across the grid indicate stability to different β_p and β_n. (c) Mean error rates, Statistical Parity Difference (SPD), and Equal Opportunity Difference (EOD) across all β_p and β_n settings for both kinds of base classifiers (SVM and LR).

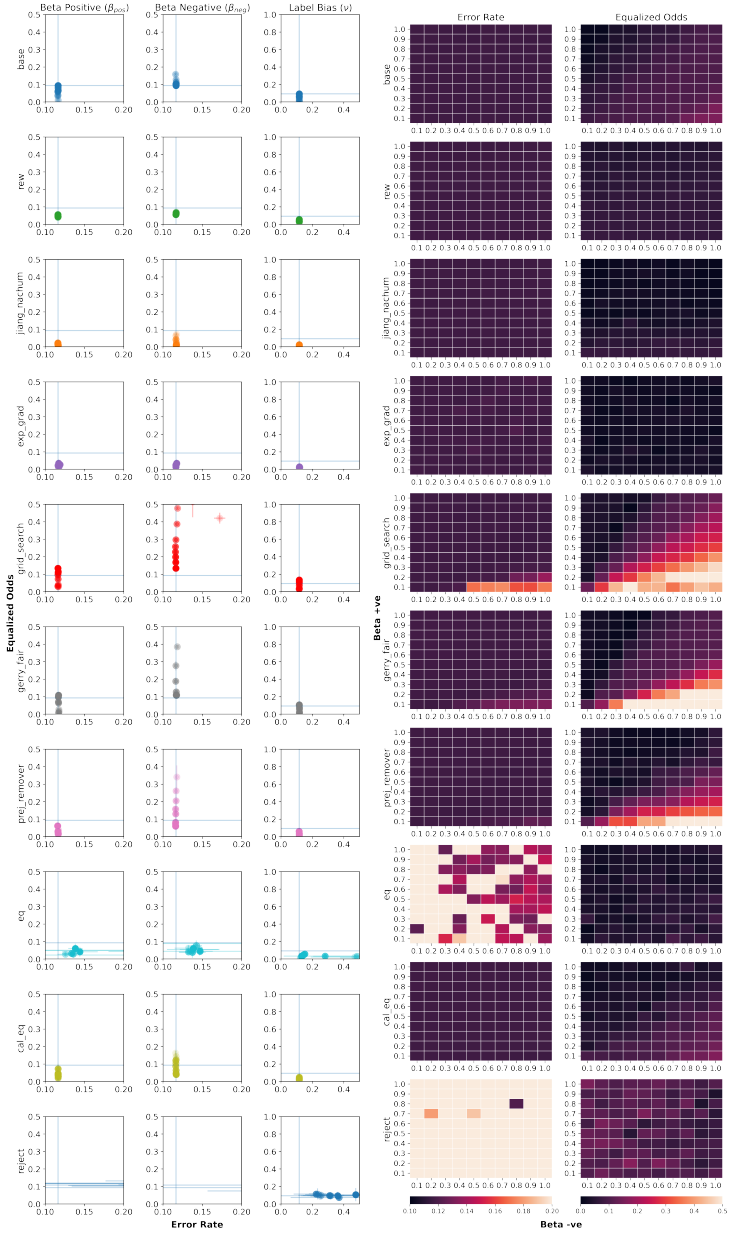


(a) (b)

Algo.		Err.	SPD	EOD
base	lr	**0.235** ± **0.011**	0.156 ± 0.125	0.103 ± 0.11
	svm	**0.235** ± **0.011**	0.17 ± 0.143	0.115 ± 0.111
rew	lr	**0.234** ± **0.006**	**0.043** ± **0.012**	**0.018** ± **0.009**
	svm	**0.234** ± **0.007**	0.071 ± 0.045	0.047 ± 0.024
jiang nachum	lr	**0.236** ± **0.008**	**0.038** ± **0.012**	**0.023** ± **0.011**
	svm	**0.236** ± **0.011**	0.046 ± **0.025**	**0.031** ± **0.018**
exp grad	lr	0.239 ± **0.008**	0.045 ± **0.022**	0.027 ± **0.014**
	svm	0.261 ± 0.044	**0.045** ± **0.019**	**0.034** ± **0.015**
grid search	lr	0.24 ± 0.027	0.153 ± 0.117	0.102 ± 0.105
	svm	0.291 ± 0.029	**0.026** ± **0.028**	**0.021** ± **0.021**
gerry fair	lr	0.263 ± 0.031	**0.039** ± 0.038	**0.024** ± 0.024
	svm	-	-	-
prej remover	lr	0.247 ± 0.022	0.153 ± 0.112	0.112 ± 0.114
	svm	-	-	-
eq	lr	0.318 ± 0.074	0.057 ± 0.032	0.074 ± 0.045
	svm	0.336 ± 0.084	**0.045** ± 0.035	0.059 ± 0.05
cal_eq	lr	0.261 ± 0.013	0.255 ± 0.163	0.151 ± 0.124
	svm	0.255 ± 0.011	0.186 ± 0.154	0.106 ± 0.111
reject	lr	0.312 ± 0.031	0.084 ± 0.038	0.098 ± 0.051
	svm	0.317 ± 0.037	0.089 ± 0.04	0.105 ± 0.046

(c)

Figure 4.9: Stability analysis of various fair classifiers on the Credit dataset (Lighter the shade, more the under-representation and label bias): (a) The spread of error rates and Equalized Odds (EODDS) of various fair classifiers as we vary β_p, β_n and label bias separately. The vertical and horizontal reference blue lines denote the performance of an unfair logistic regression model on the original dataset without any under-representation or label bias. (b) Heatmap for Error Rate and EODDS across all different settings of β_p and β_n. Darker values denote lower error rates and unfairness. Uniform values across the grid indicate stability to different β_p and β_n. (c) Mean error rates, Statistical Parity Difference (SPD), and Equal Opportunity Difference (EOD) across all β_p and β_n settings for both kinds of base classifiers (SVM and LR).



(a) (b)

Algo.		Err.	SPD	EOD
base	lr	**0.116 ± 0.000**	0.069 ± 0.021	0.12 ± 0.05
	svm	**0.116 ± 0.000**	0.069 ± **0.007**	0.096 ± **0.014**
rew	lr	**0.116 ± 0.000**	0.053 ± **0.008**	0.088 ± **0.011**
	svm	**0.116 ± 0.000**	0.067 ± **0.003**	0.092 ± **0.008**
jiang nachum	lr	**0.116 ± 0.000**	**0.033** ± 0.016	**0.039** ± 0.031
	svm	**0.116 ± 0.000**	0.05 ± 0.017	**0.051** ± 0.023
exp grad	lr	0.117 ± 0.001	**0.023 ± 0.007**	**0.03 ± 0.011**
	svm	0.124 ± 0.018	**0.022 ± 0.006**	**0.031 ± 0.012**
grid search	lr	0.12 ± 0.012	0.153 ± 0.14	0.255 ± 0.184
	svm	0.116 ± 0.001	**0.002** ± 0.01	**0.003** ± 0.015
gerry fair	lr	0.117 ± 0.003	0.146 ± 0.215	0.224 ± 0.238
	svm	-	-	-
prej remover	lr	0.116 ± 0.001	0.09 ± 0.123	0.174 ± 0.175
	svm	-	-	-
eq	lr	0.227 ± 0.125	**0.026 ± 0.01**	**0.069 ± 0.027**
	svm	0.209 ± 0.093	**0.028** ± 0.01	0.067 ± 0.019
cal_eq	lr	0.116 ± 0.000	0.05 ± 0.028	0.099 ± 0.054
	svm	**0.116 ± 0.000**	0.054 ± 0.019	0.091 ± 0.025
reject	lr	0.394 ± 0.126	0.123 ± 0.025	0.118 ± 0.04
	svm	0.35 ± 0.115	0.116 ± 0.026	0.117 ± 0.035

(c)

Figure 4.10: Stability analysis of various fair classifiers on the Bank dataset (Lighter the shade, more the under-representation and label bias): (a) The spread of error rates and Equalized Odds (EODDS) of various fair classifiers as we vary β_p, β_n and label bias separately. The vertical and horizontal reference blue lines denote the performance of an unfair logistic regression model on the original dataset without any under-representation or label bias. (b) Heatmap for Error Rate and EODDS across all different settings of β_p and β_n. Darker values denote lower error rates and unfairness. Uniform values across the grid indicate stability to different β_p and β_n. (c) Mean error rates, Statistical Parity Difference (SPD), and Equal Opportunity Difference (EOD) across all β_p and β_n settings for both kinds of base classifiers (SVM and LR).

0) subgroup, where $p'_{10} \propto \beta p_{10}$, and $p'_{ya} \propto p_{ya}$ for rest of the subgroups.

For the reweighing loss L', the reweighing weight W_{ya} is given by the following expression [KC12]:

$$W_{ya} = \frac{\Pr(Y' = y)\ \Pr(A' = a)}{\Pr(Y' = y, A' = a)}.$$

Now we can write

$$\begin{aligned}
&E_{(X',Y',A')\sim D_\beta}[L'(h(X'), Y')] = \sum_{y,a} \Pr\left(Y' = y, A' = a\right)\ E_{X'|Y'=y,A'=a}[L'(h(X'), y)] \\
&= \sum_{y,a} \Pr\left(Y' = y, A' = a\right)\ E_{X'|Y'=y,A'=a}[W_{ya} L(h(X'), y)] \\
&= \sum_{y,a} \Pr\left(Y' = y, A' = a\right)\ W_{ya}\ E_{X'|Y'=y,A'=a}[L(h(X'), y)] \\
&= \sum_{y,a} \Pr\left(Y' = y\right)\ \Pr\left(A' = a\right)\ E_{X'|Y'=y,A'=a}[L(h(X'), y)].
\end{aligned}$$

Thus, we have

$$\begin{aligned}
&E_{D_\beta}[L'] = E_{(X',Y',A')\sim D_\beta}[L'(h(X'), Y')] \\
&= \sum_{y,a} \Pr\left(Y' = y\right)\ \Pr\left(A' = a\right)\ E_{X'|Y'=y,A'=a}[L(h(X'), y)] \\
&= \sum_{y,a} \Pr\left(Y' = y\right)\ \Pr\left(A' = a\right)\ E_{X|Y=y,A=a}[L(h(X), y)],
\end{aligned}$$

because $X'|Y' = y, A' = a$ in D_β remains the same as $X|Y = y, A = a$ in D even after injecting under-representation.

For the distribution D_β, assume that the normalization constant for the subgroup probabilities is $N' = \sum_{Y'=y,A'=a} p_{ya} = p_{00} + p_{01} + \beta p_{10} + p_{11}$.

We can write down individual label and sensitive attribute probabilities as follows in D_β: $\Pr(Y' = 0) = (p_{00} + p_{01})/N'$, $\Pr(Y' = 1) = (\beta p_{10} + p_{11})/N'$, $\Pr(A' = 0) = (p_{00} + \beta p_{10})/N'$ and $\Pr(A' = 1) = (p_{01} + p_{11})/N'$.

We will now attempt to lower and upper bound

$$\frac{\Pr(Y' = y)\ \Pr(A' = a)}{\Pr(Y = y, A = a)}$$

in terms of α, for any $Y, A \in \{0, 1\}$ and any $\beta \in (0, 1]$. Let $N = 1 = p_{00} + p_{01} + p_{10} + p_{11}$. For example, for $Y = 0, A = 0$ we get

$$\frac{\alpha}{2} \leqslant \frac{p_{00} + p_{01}}{N} \leqslant \frac{(p_{00} + p_{01})}{N'}\ \frac{(p_{00} + \beta p_{10})}{N'}\ \frac{N}{p_{00}} \leqslant \frac{N}{p_{00}} \leqslant \frac{4}{\alpha},$$

using $N' \leqslant N$, $p_{00} \leqslant p_{00} + \beta p_{10}$ and the definition of m, M and α. Similarly, working out the upper and lower bounds for each of the four probability terms in the expected reweighed loss expression above, we get a common bound

$$\frac{\alpha^2}{4} \leqslant \frac{\Pr(Y' = y)\ \Pr(A' = a)}{\Pr(Y = y, a = a)} \leqslant \frac{4}{\alpha},$$

for any $y, a \in \{0, 1\}$. Now we can plug in these bounds in the following expression for $E_D[L]$.

$$\begin{aligned} E_D[L] &= E_{(X,Y,A)\sim D}[L(h(X), Y)] \\ &= \sum_{y,a} \Pr(Y = y, A = a)\ E_{X|Y=y,A=a}[L(h(X), y)] \end{aligned}$$

Using the sandwich bound with the final form of $E_{D_\beta}[L']$, we get:

$$\frac{\alpha^2}{4} E_D[L] \leqslant E_{D_\beta}[L'] \leqslant \frac{4}{\alpha} E_D[L].$$

□

Before proving Theorem 4.4, we prove a small proposition that will be useful for later results.

Proposition 4.1. Let L be a non-negative bounded loss between $[0, B]$. Let $\hat{h}$ be the empirical risk minimizing classifier with loss L with N samples: $\hat{h} = \text{argmin}_{h\in\mathcal{H}} \frac{1}{N} \sum_{i=1}^{N} L(h(x_i, y_i))$, where $\{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$ are i.i.d. samples from D. Let $h^* = \text{argmin}_{h\in\mathcal{H}} E_{(x,y)\sim D}[L(h(x), y)]$. Then with probability of at least $1 - \delta$, we have

$$E_D[L(\hat{h})] \leqslant B \sqrt{\frac{\log(2/\delta)}{2N}} + E_D[L(h^*)].$$

Proof. The proof follows directly from the proof of Thm. 1 in [ZLL21], and a general form of Hoeffding's inequality for any non-negative bounded loss[3]. □

We are now ready to prove Theorem 4.4.

Proof. (Proof of Theorem 4.4) Using the results from Theorem 4.3, for any h and β, we can write

$$\frac{\alpha^2}{4} E_D[L(h)] \leqslant E_{D_\beta}[L'(h)] \leqslant \frac{4}{\alpha} E_D[L(h)].$$

For $\hat{h}_\beta$, we can use the lower bound from Theorem 4.3 to get

$$E_D[L(\hat{h}_\beta)] \leqslant \frac{4}{\alpha^2} E_{D_\beta}[L'(\hat{h}_\beta)]. \tag{4.1}$$

We now assume that L is a bounded loss between $[0, 1]$. if not, we can rescale it to lie in $[0, 1]$. Using $m = \min\{p_{00}, p_{01}, p_{10}, p_{11}\}$, $M = \max\{p_{00}, p_{01}, p_{10}, p_{11}\}$, and $\alpha = m/M$, we can show that the reweighing factor

$$W_{ya} = \frac{\Pr(Y' = y)\ \Pr(A' = a)}{\Pr(Y' = y, A' = a)} \leqslant \frac{4}{\alpha\beta}.$$

for all $y, a \in \{0, 1\}$. This can be done by using the upper bound from Theorem 4.3 and the fact that in the worst case ($Y = 1, A = 0$ here), $\frac{\Pr(Y=y,A=a)}{\Pr(Y'=y,A'=a)} = \frac{1}{\beta}$.

Thus, the reweighed loss L' is bounded between $[0, 4/(\alpha\beta)]$. Now, using Proposition 4.1 and that $\hat{h}$ is the empirical minimizer of the reweighed loss L' on N samples from the biased distribution D_β, we obtain that

$$E_{D_\beta}[L'(\hat{h}_\beta)] \leqslant \frac{4}{\alpha\beta} \sqrt{\frac{\log(2/\delta)}{2N}} + E_{D_\beta}[L'(h')].$$

[3] http://cs229.stanford.edu/extra-notes/hoeffding.pdf

where h' is the Bayes optimal classifier for the reweighed loss L' on the biased distribution D_β. Now let h^* be the Bayes optimal classifier for the loss L on the original distribution D. Thus, $E_{D_\beta}[L'(h')] \leqslant E_{D_\beta}[L'(h^*)]$, and we get

$$E_{D_\beta}[L'(\hat{h}_\beta)] \leqslant \frac{4}{\alpha\beta}\sqrt{\frac{\log(2/\delta)}{2N}} + E_{D_\beta}[L'(h^*)]. \tag{4.2}$$

Combining (4.1) and (4.2), we have

$$\begin{aligned} E_D[L(\hat{h}_\beta)] &\leqslant \frac{4}{\alpha^2}\left(\frac{4}{\alpha\beta}\sqrt{\frac{\log(2/\delta)}{2N}} + E_{D_\beta}[L'(h^*)]\right) \\ &= \frac{16}{\alpha^3\beta}\sqrt{\frac{\log(2/\delta)}{2N}} + \frac{4}{\alpha^2}E_{D_\beta}[L'(h^*)] \\ &\leqslant \frac{16}{\alpha^3\beta}\sqrt{\frac{\log(2/\delta)}{2N}} + \frac{16}{\alpha^3}E_D[L(h^*)], \end{aligned}$$

where we use the upper bound in Theorem 4.3 to get the last inequality. That completes the proof of Theorem 4.4. □

Chapter 5

On Optimal Steering to Achieve Exact Fairness

Abstract

To fix the 'bias in, bias out' problem in fair machine learning, it is important to steer feature distributions of data or internal representations of Large Language Models (LLMs) to *ideal* ones that guarantee group-fair outcomes. Previous work on fair generative models and representation steering could greatly benefit from provable fairness guarantees on the model output. We define a distribution as *ideal* if the minimizer of any cost-sensitive risk on it is guaranteed to have exact group-fair outcomes (e.g., demographic parity, equal opportunity)—in other words, it has no fairness-utility trade-off. We formulate an optimization program for optimal steering by finding the nearest *ideal* distribution in KL-divergence, and provide efficient algorithms for it when the underlying distributions come from well-known parametric families (e.g., normal, log-normal).

Empirically, our optimal steering techniques on both synthetic and real-world datasets improve fairness without diminishing utility (and sometimes even improve utility). We demonstrate affine steering of LLM representations to reduce bias in multi-class classification, e.g., occupation prediction from a short biography in Bios dataset (De-Arteaga et al.). Furthermore, we steer internal representations of LLMs towards desired outputs so that it works equally well across different groups.

§ 5.1 Introduction

The importance of clean or *ideal* data in fair machine learning cannot be overstated. The principle of *bias in, bias out* is widely recognised as a root cause of unfair outcomes in ML systems [BG18; May19; RR20; Cow+20]. Models trained on biased data tend to learn, perpetuate, and often amplify such biases. Importantly, the problem of biased data extends beyond training: fairness-constrained training on biased data does not guarantee fairness on (unbiased) test sets, and post-processing using biased validation data fails to ensure fairness at deployment. Moreover, fairness audits based on biased assessment data can be misleading and difficult to reverse [BR21b; Bak+21].

Fairness metrics such as demographic parity and equal opportunity are inherently functions of both the model and the data distribution. Thus, unfair outcomes may be addressed either by adjusting the model or by altering the data distribution. While prior work on fair in-processing aims to construct an *ideal* model under fairness constraints [Aga+18; Don+18], our work focuses instead on identifying an *ideal* data distribution that supports fairness. In this respect, our approach aligns most closely with the fair pre-processing literature.

Early work by Kamiran and Calders [KC12] introduced heuristic reweighting methods for binary classification, adjusting the weights of instances from class i and group a using $\Pr(Y=i) \cdot \frac{\Pr(A=a)}{\Pr(Y=i,A=a)}$. However, this approach ignores feature distributions and provides no fairness guarantees when combined with accuracy maximisation. Calmon et al. [Cal+17] framed fair pre-processing as an optimisation problem that balances distance to the original distribution with group and individual fairness constraints. While convex in some settings, their approach may be infeasible when group and individual fairness are incompatible [FSV21].

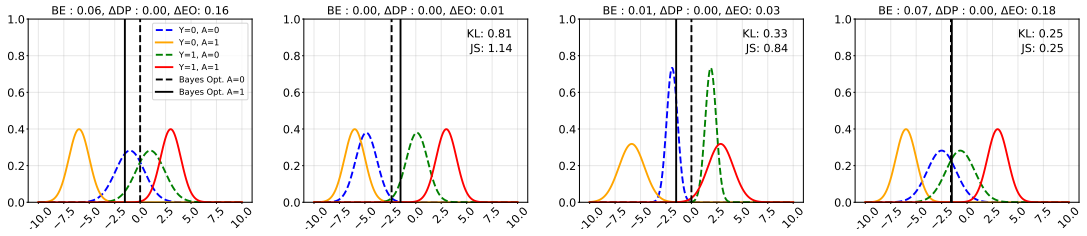
Other work explores alternative mechanisms: Jiang and Nachum [JN20] address label bias through reweighting; Plečko and Meinshausen [PM20] and Plečko et al. [PBM24] propose fair adaptation methods based on causal models; and Xiong et al. [Xio+24] reformulate pre-processing as a large-scale mixed-integer program, solved via a cutting-plane method. Fair pre-processing remains widely used in practice, often integrated into fairness toolkits alongside in- and post-processing methods [Bel+18; Bir+20]. Despite its heuristic nature, the reweighing method of Kamiran and Calders [KC12] continues to perform well in mitigating bias across standard benchmarks [SDS23b; Blo+24; Xio+24].

A common goal of all fair pre-processing methods is to find an *ideal* data distribution close to the given distribution so that any downstream model trained on it must have guaranteed fairness. A stronger requirement that this should hold for downstream models optimized for multiple tasks leads to impossibility results [Lec+21]. If all downstream classifiers are required to be fair, then the group-wise distributions must be nearly identical, which is absurd. Thus, we restrict our downstream models only to Bayes optimal classifiers for cost-sensitive risks. Our first contribution is to formally define an *ideal* distribution as the one where the Bayes optimal classifier for any cost-sensitive risk satisfies exact fairness (e.g., satisfies equal opportunity perfectly).

Bayes optimal classifier maximizes accuracy on a given distribution, and is a significant object of study in statistical machine learning [DGL96]. Fair Bayes optimal classifier maximizes accuracy subject to fairness constraints, and its mathematical characterization for binary fair classification has been important in fair classification [MW18; Chz+19; Cel+21; ZDC22b]. Blum and Stangl [BS19] introduces a data bias model that injects under-representation and label bias in an original unbiased distribution to create biased data. They show that, for a stylized distribution under some conditions, the fair Bayes optimal classifier on the biased distribution recovers the Bayes optimal classifier on the original unbiased distribution. Their unbiased distribution is *ideal* by construction, i.e., the Bayes optimal classifier on their unbiased distribution is guaranteed to be perfectly fair. In Chapter 3, we extended this observation to general hypothesis classes and distributions beyond the stylized setting of Blum and Stangl [BS19]. Blum et al. [Blu+23] study fair Bayes optimal classifier whether its accuracy is robust to malicious corruptions in data distribution. In contrast to these results, our focus is not on finding the *ideal* classifier but on finding the nearest *ideal* distribution.

By definition, our *ideal* distribution has no trade-off between accuracy (or cost-sensitive utility) and fairness. If we find an *ideal* distribution close to our original distribution, we can steer our distribution towards reducing fairness-accuracy trade-off. Moreover, if the *ideal* distribution offers better accuracy, it suggests that we can steer our distribution to improve both accuracy and fairness simultaneously. Fig. 5.1 gives such an example where the pre-processing of [KC12], in contrast, leaves the original distribution unchanged, and our method provides a direction to steer the distribution towards better accuracy and fairness simultaneously.

Dutta et al. [Dut+20] characterize a similar objective using Chernoff Information (see Cover [Cov99]) and formulate an optimization program to find the nearest distribution in KL-divergence on which the Chernoff Information gap between two group-conditional feature distributions vanishes. Their opti-



(a) Original **(b)** Ideal (Affirmative) **(c)** Ideal (Changing All) **(d)** Mean Matching

Figure 5.1: Comparison of different interventions for changing Data Distributions for Exact Fairness. Figure (5.1a) captures the original distribution, its Bayes error (BE), and the unfairness differences (ΔDP and ΔEO). In Figure (5.1b), we only change the under-privileged group using Corollary 5.2, and in Figure (5.1c) we change all four subgroups using Proposition 5.3. Finally, in Figure (5.1d), we match the means of the two groups. Figures (5.1b) and (5.1c) show that it is possible to construct 'ideal' distributions that are close to the given distribution where both the BE and ΔDP/ΔEO are small.

mization problem is not known to be efficiently solvable and the fairness guarantees in terms of Chernoff Information gap does not translate easily to standard fairness metrics such as demographic parity, equal opportunity etc. In contrast, we formulate an optimization problem to find the nearest *ideal* distribution in KL-divergence to given distribution and give efficient algorithms to solve it for various parametric families of distributions. To put our results to the test, we also apply our interventions to steer the representations of an LLM [Tou+23; Gra+24] for fair multi-class classification [DA+19; Sin+24] and emotion steering [Zha+24]. We are able to improve the effectiveness of steering without significantly affecting the accuracy and are able to demonstrate how our notion of ideal distributions can help guide interventions for practical applications.

For completeness, we also want to make the reader aware of a long line of work on fair representation learning where the data transformations can map the distributions to another space [Zem+13; Mad+18; MOW19; Liu+22; Cer+24]. Our work is not directly related, but can potentially be used to refine fair representations to achieve provable and exact fairness guarantees. Our approach falls within the realm of fair pre-processing, where we transform the data distribution to remain within the same joint feature and target space as the given distribution. We also note that, to derive non-parametric, black-box post-processing techniques to obtain ideal distributions for potentially many cost-sensitive risks, tools such as Multicalibration and Omniprediction can be a good starting point [HJ+18; Hu+23]. We now summarize our key contributions.

- We define *ideal* distribution (Definition 5.1) for fair classification as the one on which the Bayes optimal classifier for any cost-sensitive risk satisfies exact fairness (e.g., exact demographic parity, exact equal opportunity).
- When group and class-conditioned distributions belong to well-known parametric families of distributions (e.g., Gaussian, log-normal), we can succinctly rewrite the property of being an *ideal* distribution as a parametric condition (Proposition 5.2).
- The problem of finding the nearest *ideal* distribution to a given distribution is intractable in general. We formulate it as an optimization problem of the KL-divergence subject to the parametric conditions required for being *ideal* (Section 5.4). As stated, it is a non-convex optimization problem (Proposition 5.3), but we show how to solve it efficiently. We also provide a closed-form optimal solution in special cases (Theorem 5.1, Corollary 5.2). When better accuracy is achievable on the nearest *ideal* distribution, it suggests a direction to steer the original distribution in order to improve both accuracy and fairness.
- To illustrate the effect of different interventions, we show the effect of using Affirmative Action

(Corollary 5.2), changing all subgroups (Proposition 5.3) and matching the means of sensitive groups for different univariate distribution where we can compute the Bayes Optimal Group-aware classifiers analytically in Figures 5.1-5.2.

- To demonstrate how our guarantees can aid practical applications, we also perform experiments to steer LLM representations to reduce bias in multi-class classification (Figure 5.3) and effective group-level emotion steering (Figure 5.4).

§ 5.2 Problem Setup and Preliminaries

Let (X, A, Y) be a random data point from a joint distribution D over $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$, where $\mathcal{X}, \mathcal{A}, \mathcal{Y}$ denote the sets of features, sensitive attributes, and class labels, respectively. Let $q_{ia} = \Pr(Y = i, A = a)$ and P_{ia} denote the distribution $X \mid Y = i, A = a$ with the probability density $p_{ia}(x) = \Pr(X = x \mid Y = i, A = a)$. When P_{ia}'s come from parametric families of distributions, we assume $\mathcal{X} = \mathbb{R}^d$. We work with the following well-known definitions of fairness in classification [Dwo+12a; HPS16a; BHN19].

Using our definitions of unfairness metrics from Chapter 2, we define quantitative metrics of unfairness. For example, $\Delta_{DP,y}(h, D)$ denotes the absolute value of difference between $\Pr(h(X, A) = y|A = a)$ and $\Pr(h(X, A) = 1|A = a')$. Similarly, $\Delta_{EO,y}(h, D)$ denotes the absolute value of difference between $\Pr(h(X, A) = y|Y = y, A = a)$ and $\Pr(h(X, A) = y|Y = y, A = a')$. Whenever we are dealing with binary classification, we will omit the use of y in the subscript.

We consider group-aware classifiers. For binary classification tasks, we are particularly interested in threshold classifiers $h_t(x, a)$ that apply a group and feature dependent threshold $t(x, a)$ to the class probability of an example: $h_t(x, a) = \mathbb{I}(\eta(x, a) \geqslant t(x, a))$ where $\eta(x, a) = \Pr(Y = 1|X = x, A = a)$. The Bayes optimal classifier for a given distribution has the form $t(x, a) = 1/2$ [DGL96]. For a cost matrix $C \in \mathbb{R}^{2\times 2}$ and the associated cost sensitive loss l_C, the Bayes optimal classifier is defined as $\mathbb{I}(\eta(x, a) \geqslant t_C)$, for a threshold $t_C = (c_{10} - c_{00})/(c_{10} - c_{00} + c_{01} - c_{11}) \in [0, 1]$, where c_{ij} denote the entries of the cost matrix $C \in \mathbb{R}^{2\times 2}$ [Elk01; Sco12; Koy+14; SK22].

§ 5.3 Ideal Distributions for Fair Classification

We define a data distribution as *ideal* when minimizing any cost-sensitive risk on it is guaranteed to give exact fairness (e.g., demographic parity, equal opportunity). In practice, downstream models trained on a distribution are typically optimized for some performance or utility metric that may not be known in advance. Our definition of ideal distribution allows the flexibility to choose any cost-sensitive risk as the performance metric for downstream models and still gives exact fairness guarantee for any optimal model downstream.

Definition 5.1. Let $\mathcal{H}$ be a hypothesis class of group-aware classifiers $h : \mathcal{X} \times \mathcal{A} \to \mathcal{Y}$ and let $\Delta(h, D)$ be a given unfairness metric, e.g., Δ_{DP}, Δ_{EO}. Given a distribution D over $\mathcal{X} \times \mathcal{A} \times \mathcal{Y}$ and a cost-sensitive risk $C \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{Y}|}$, let $h^*_C = \underset{h\in\mathcal{H}}{\operatorname{argmin}} \underset{D}{E}[\ell_C(h(X, A), Y)]$, where l_C is the associated cost sensitive loss. We call D an *ideal distribution* if $\Delta(h^*_C, D) = 0$, for all $C \in \mathbb{R}^{|\mathcal{Y}|\times|\mathcal{Y}|}$.

Examples of fairness metrics include Demographic Parity, Equal Opportunity, and Equalized Odds (Definitions 2.3, 2.4 and 4.1 respectively), and examples of cost-sensitive risk include the usual $0-1$ loss and different performance metrics which are functions of the confusion matrix metrics [Elk01; Koy+14; SK22].

Our definition gets around the impossibility theorems about fair representation for multiple tasks [Lec+21]. However, we need to be careful of two things. First, our definition should not be too restrictive to just force the group-conditioned distributions to be similar or identical, as that would be impractical. Second, we need an efficient and equivalent way of expressing the constraint of being ideal. We show how to express it as a parametric condition when the group and class-conditioned distributions belong to certain well-known parametric families of distributions. This helps in checking if a given distribution is ideal, and otherwise, finding its nearest ideal distribution.

§ 5.3.1 Parametric Conditions for Ideal Distributions

Borrowing a simple setup of parametric distributions from previous work on fair machine learning [PCDG18], we assume that the class and group-conditioned feature distributions $X \mid Y = i, A = a$ belong to a parametric family of distributions, e.g., univariate or multivariate Gaussians, log-normal. In that case, we show that the property of being *ideal* (Definition 5.1) can be equivalently expressed as certain parametric conditions. To begin, we will first show the set of parametric conditions for multivariate normal distributions and a multi-class, multi-attribute setting.

Proposition 5.1. Let (X, Y, A) denote the features, class label, and group membership, respectively, of a random data point from any data distribution D with $q_{ia} = \Pr(Y = i, A = a)$, for $i \in \mathcal{Y}$ and $a \in \mathcal{A}$. Let $X|Y = i, A = a \sim \mathcal{N}(\mu_{ia}, \Sigma_{ia})$ be multivariate Normal distributions with mean $\mu_{ia} \in \mathbb{R}^d$ and covariance matrix $\Sigma_{ia} \in \mathbb{R}^{d\times d}$, for $i \in \mathcal{Y}$ and $a \in \mathcal{A}$. If the means μ_{ia} and the covariance matrices Σ_{ia} satisfy

$$\Sigma_{ia}^{-1/2}(\mu_{ia} - \mu_{ja}) = \Sigma_{ia'}^{-1/2}(\mu_{ia'} - \mu_{ja'}) \quad \text{and}$$
$$\Sigma_{ia}^{1/2}\Sigma_{ja}^{-1}\Sigma_{ia}^{1/2} = \Sigma_{ia'}^{1/2}\Sigma_{ja'}^{-1}\Sigma_{ia'}^{1/2} \quad \text{and} \quad \frac{q_{ia}}{q_{ja}} = \frac{q_{ia'}}{q_{ja'}}, \ \forall i, j \in \mathcal{Y}, a, a' \in \mathcal{A},$$

then the group-aware Bayes optimal classifier on D satisfies equal opportunity.

The proof for Proposition 5.1 is given in Section 5.8.1. When we are dealing with binary class labels and group membership, we can show a necessary and sufficient condition for univariate normal distributions.

Proposition 5.2. Let (X, Y, A) denote the features, binary class label, and binary group membership, respectively, of a random data point from any data distribution D with $q_{ia} = \Pr(Y = i, A = a)$, for $i \in \{0, 1\}$ and $a \in \{0, 1\}$, and let $X|Y = i, A = a \sim \mathcal{N}(\mu_{ia}, \sigma_{ia}^2)$ be univariate normal distributions, for $i \in \{0, 1\}$ and $a \in \{0, 1\}$. Then the distribution D is *ideal* for equal opportunity (see Definition 5.1) if and only if

$$\frac{\mu_{01} - \mu_{11}}{\sigma_{11}} = \frac{\mu_{00} - \mu_{10}}{\sigma_{10}}, \quad \frac{\sigma_{11}}{\sigma_{01}} = \frac{\sigma_{10}}{\sigma_{00}}, \quad \frac{q_{10}}{q_{00}} = \frac{q_{11}}{q_{01}}.$$

Proposition 5.2 shows that our parametric condition is equivalent to $\Delta_C(h_C^*, D) = 0$, for all cost matrices $C \in \mathbb{R}^{2\times 2}$. When we use a fixed cost matrix for 0-1 loss, and consider the Bayes optimal classifier in Proposition 5.1, our parametric condition is sufficient but not always necessary. However, the same condition ensures the Bayes optimal classifier to satisfy multiple fairness criteria simultaneously, viz., demographic parity, equal opportunity, equalized odds.

The proof for Proposition 5.2 is given in Section 5.8.1. It is interesting to note that the same parametric conditions imply that the Bayes optimal classifier on the corresponding distribution simultaneously

satisfies multiple fairness criteria, viz., demographic parity, equal opportunity, and equalized odds. Moreover, the same condition works for both univariate Gaussian and log-normal distributions. Using our proof technique, it is easy to derive similar conditions for other parametric families too. We now present a small proposition to link our intervention to a widely used reweighing intervention in the fairness literature [KC12].

Remark 5.3.1. (Our conditions imply Kamiran and Calders [KC12] Intervention) Kamiran and Calders [KC12] reweighing method essentially reweighs q_{ia} by a multiplicative factor of $\Pr(Y=i)\Pr(A=a)/\Pr(Y=i,A=a)$. Let us call the resulting probabilities $\tilde{q}_{ia}$. Using $\Pr(Y=i,A=a)=q_{ia}$, $\Pr(Y=i)=\sum_{a\in\mathcal{A}} q_{ia}$ and $\Pr(A=a)=\sum_{i\in\mathcal{Y}} q_{ia}$, we get $\tilde{q}_{ia} \propto q_{ia}(\sum_{a\in\mathcal{A}} q_{ia})(\sum_{i\in\mathcal{Y}} q_{ia})/q_{ia}$. Hence, $\tilde{q}_{ia}/\tilde{q}_{ja} = \tilde{q}_{ia'}/\tilde{q}_{ja'} = \sum_{a\in\mathcal{A}} q_{ia}/\sum_{a\in\mathcal{A}} q_{ja}, \forall i,j\in\mathcal{Y}$ and $a,a'\in\mathcal{A}$. It is the same condition on q_{ia}'s stated in Proposition 5.1. Thus, our result can be thought of as a second-stage pre-processing of P_{ia} distributions after applying the reweighing of Kamiran and Calders [KC12] to q_{ia}'s in the first stage.

Remark 5.3.2. (Limitation of [Dut+20]) As an interesting consequence, our conditions on μ_{ia} and Σ_{ia} imply $D_{KL}\left(\tilde{P}_{00}\|\tilde{P}_{01}\right) = D_{KL}\left(\tilde{P}_{10}\|\tilde{P}_{11}\right)$. When the classes are balanced, the error rate of the Bayes optimal classifier on group $A=a$ in $\tilde{D}$ equals $0.5(1-d_{TV}(\tilde{P}_{0a},\tilde{P}_{1a}))$, where d_{TV} denotes the total variation distance [Nie14]. Thus, achieving $d_{TV}(\tilde{P}_{00},\tilde{P}_{10}) = d_{TV}(\tilde{P}_{01},\tilde{P}_{11})$ ensures equal error rates across both the groups. However, there is no closed-form expression for the total variation distance between two univariate Gaussians, and the KL-divergence can be thought of as a proxy using Pinsker's inequality [Can23]. This is similar to the information-theoretic argument used by Dutta et al. [Dut+20], which is why their optimization cannot guarantee outcome fairness as we do.

§ 5.4 Finding The Nearest Ideal Distribution

When a given distribution D is not ideal, then a natural question is to find its nearest distribution $\tilde{D}$ that is ideal. We formulate this problem as follows.

$$\underset{\tilde{D}\,:\,\tilde{D}\text{ is ideal}}{\text{minimize}}\; D_{KL}\left(\tilde{D}\|D\right).$$

In the above optimization problem, the KL-divergence objective is well-known and convex but the constraint of $\tilde{D}$ being ideal is extremely non-trivial to express. We show that when the group and class-conditioned distributions $X \mid Y=i, A=a$ come from certain well-known parametric families of distributions, this constraint can be equivalently expressed as a constraint on the distribution parameters. We now give a concrete formulation of the optimization problem described in Section 5.3 using the constraints derived in Proposition 5.1.

$$\min_{\tilde{D}} -\frac{d}{2} + \frac{1}{2}\sum_{(i,a)} q_{ia}(\tilde{\mu}_{ia}-\mu_{ia})^T\Sigma_{ia}^{-1}(\tilde{\mu}_{ia}-\mu_{ia}) + \frac{1}{2}\sum_{(i,a)} q_{ia}\left(\text{tr}\left\{\Sigma_{ia}^{-1}\tilde{\Sigma}_{ia}\right\} - \log\det\left(\Sigma_{ia}^{-1}\tilde{\Sigma}_{ia}\right)\right)$$
$$\text{subject to } \Sigma_{ia}^{-1/2}(\mu_{ia}-\mu_{ja}) = \Sigma_{ia'}^{-1/2}(\mu_{ia'}-\mu_{ja'}),\;\; \Sigma_{ia}^{1/2}\Sigma_{ja}^{-1}\Sigma_{ia}^{1/2} = \Sigma_{ia'}^{1/2}\Sigma_{ja'}^{-1}\Sigma_{ia'}^{1/2} \text{ and}$$
$$\frac{q_{ia}}{q_{ja}} = \frac{q_{ia'}}{q_{ja'}},\; \forall i,j\in\mathcal{Y}, a,a'\in\mathcal{A}.$$

For readability, we will work with binary class labels and binary group attributes in this section. However, all of our results can also be written down for multi-class and multi-group settings. In its full generality, the above problem is non-convex, and therefore, we will first propose reducing this to a convex optimization problem and then propose solving it in full generality by using line search.

§ 5.4.1 Affirmative Action

We first focus on a class of interventions for which solving the optimization program is efficient. We define *Affirmative Action* as changing the distribution of the underprivileged group to achieve ideal distributions where fairness and accuracy are in accord.

Theorem 5.1. Let (X, Y, A) denote the features, binary class label, and binary group membership, respectively, of a random data point from any data distribution D with $q_{ia} = \Pr(Y = i, A = a)$, for $i \in \{0,1\}$ and $a \in \{0,1\}$, such that $q_{10}/q_{00} = q_{11}/q_{01}$. Let $X|Y = i, A = a \sim \mathcal{N}(\mu_{ia}, \Sigma_{ia})$ be multivariate Normal distributions, with mean $\mu_{ia} \in \mathbb{R}^d$ and covariance matrix $\Sigma_{ia} \in \mathbb{R}^{d\times d}$, for $i \in \{0,1\}$ and $a \in \{0,1\}$. Let $\tilde{D}$ denote a distribution obtained by keeping (Y, A) unchanged and only changing $X|Y = i, A = a$ to $\tilde{X}|Y = i, A = a \sim \mathcal{N}(\tilde{\mu}_{ia}, \tilde{\Sigma}_{ia})$. Then in the case of Affirmative action (changing only $\tilde{\mu}_{i0}$ and $\tilde{\Sigma}_{i0}$), finding the ideal distribution is a convex program, i.e., we can efficiently minimize $D_{KL}\left(\tilde{D}||D\right)$ as a function of the variables $\tilde{\mu}_{i0}$ and $\tilde{\Sigma}_{i0}$ subject to the constraints in Proposition 5.1, so that the Bayes optimal classifier on the optimal $\tilde{D}$ is guaranteed to be EO-fair.

The proof for Theorem 5.1 is given in Section 5.8.2. While we show that the optimization program is convex, obtaining a closed-form expression for the change in means and covariances is extremely cumbersome for the general case. However, we can show how the closed-form expressions for $\tilde{\mu}_{i0}$ and $\tilde{\sigma}_{i0}$ look like for the univariate case in the following corollary:

Corollary 5.2. (Univariate Affirmative Action) For the case where $X|Y = i, A = a \sim \mathcal{N}(\mu_{ia}, \sigma^2_{ia})$ are univariate normal distributions, for $i, a \in \{0,1\}$, the optimal distribution $\tilde{D}$ from Theorem 5.1, with γ^* being a function of the original distribution parameters, can be written down as:

$$\tilde{\sigma}_{i0} = \gamma^* \sigma_{i1}, \quad \tilde{\mu}_{00} = \tilde{\mu}_{10} + \gamma^*(\mu_{01} - \mu_{11}), \text{ and } \tilde{\mu}_{10} = \frac{\left(q_{00}\dfrac{\mu_{00} - \gamma^*(\mu_{01} - \mu_{11})}{\sigma^2_{00}} + q_{10}\dfrac{\mu_{10}}{\sigma^2_{10}}\right)}{\left(\dfrac{q_{00}}{\sigma^2_{00}} + \dfrac{q_{10}}{\sigma^2_{10}}\right)},$$

§ 5.4.2 Changing all Subgroups

Another intervention we can follow is to change all the subgroups of the given distribution. However, a quick check through the proof of Theorem 5.1 shows that this will lead to a non-convex program. However, just like Corollary 5.2, we can show a reasonable intervention for the univariate case, where we change all four subgroups and search over a non-convex function using line search over a fairly large grid size.

Proposition 5.3. (All subgroup change for Exact Fairness) Let (X, Y, A) denote the features, binary class label, and binary group membership, respectively, of a random data point from any data distribution D with $q_{ia} = \Pr(Y = i, A = a)$, for $i \in \{0,1\}$ and $a \in \{0,1\}$, such that $q_{10}/q_{00} = q_{11}/q_{01}$, and let $X|Y = i, A = a \sim \mathcal{N}(\mu_{ia}, \sigma^2_{ia})$ be univariate normal distributions, for $i \in \{0,1\}$ and $a \in \{0,1\}$. Let $\tilde{D}$ denote a distribution obtained by keeping (Y, A) unchanged and only changing $X|Y = i, A = a$ to $\tilde{X}|Y = i, A = a \sim \mathcal{N}(\tilde{\mu}_{ia}, \tilde{\sigma}^2_{ia})$. Then minimizing $D_{KL}\left(\tilde{D}||D\right)$ as a function of the variables $\tilde{\mu}_{ia}$ and $\tilde{\sigma}_{ia}$ subject to the constraints in Proposition 5.1 leads to a non-convex program. There exists a non-convex function $\gamma \mapsto \mathcal{L}_\gamma$, where $\gamma^* = \underset{\gamma\in(0,\infty)}{\arg\min}\, \mathcal{L}^*_\gamma$, which only depends on

the original distribution parameters (Equation 5.1), such that $\tilde{\mu}_{ia}$ and $\tilde{\sigma}_{ia}$ can be expressed as a function of γ^* and μ_{ia} and σ_{ia}.

The proof of Proposition 5.3 is provided in Section 5.8.2. We can also derive worst-case upper bounds on the error rate and the unfairness gap Δ_{EO} of the Bayes optimal classifier $\tilde{h}$ on $\tilde{D}$ with respect to the original distribution D. These bounds show that both the accuracy loss and the fairness gap depend only on the KL divergence between D and $\tilde{D}$. It also shows that the optimal value of our optimization problem can be used to approximately translate the accuracy guarantee of $\tilde{h}$ from $\tilde{D}$ to D.

Proposition 5.4. Let $\text{err}(h, D)$ denote the error rate (expected 0-1 loss) of a classifier h on the distribution D. Let $d_{TV}(\tilde{D}, D)$ denote the total variation distance between two distributions $\tilde{D}$ and D, while D_{KL} denotes the KL-Divergence between them. Denote the Bayes optimal classifier on the ideal distribution $\tilde{D}$ as $\tilde{h}$ (and similarly the Bayes optimal classifier h. Then, we can bound the error rate and the Equal opportunity of $\tilde{h}$ on the original distribution D as follows:

$$|\text{err}(\tilde{h}, D) - \text{err}(\tilde{h}, \tilde{D})| \leqslant \sqrt{2D_{KL}(\tilde{D}, D)} \quad \text{and} \quad \Delta_{EO}(\tilde{h}, D) \leqslant \sqrt{8D_{KL}\left(\tilde{D} \| D\right)}.$$

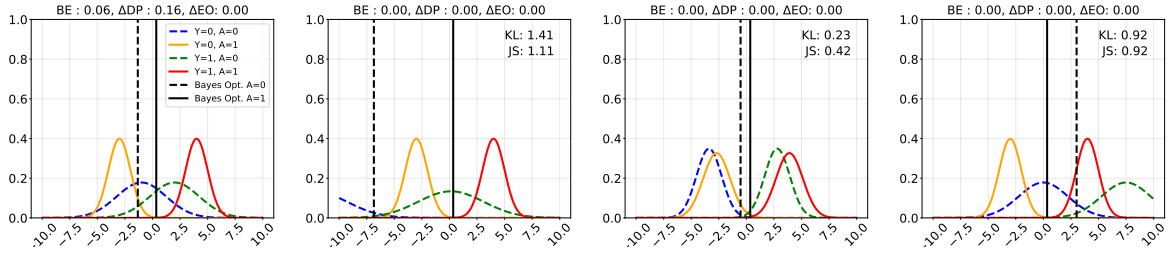
The proof for Proposition 5.4 is given in Section 5.8.2. The above bounds inform us that when the ideal distributions are close enough to the true distribution, we do not lose much when going towards an ideal distribution.

§ 5.5 Case Study on Gaussian Distributions

In this section, we adapt a stylized Gaussian setting from prior work (Definition 3.1 in [PCDG18], Section 5.3 in [Bak+21]) to examine unfairness and Bayes optimal error before and after distributional interventions. We assume homoskedastic Gaussians within each group $A = a$, i.e., $\sigma_{0a} = \sigma_{1a}$, so that the Bayes classifier admits a threshold form. This enables tractable analysis of interventions aimed at achieving a fairness–accuracy trade-off. Further details are provided in Section 5.8.3.

We evaluate four interventions: (a) the Bayes optimal classifier on the original distribution (no correction), (b) *EF Affirmative*: transforming the underprivileged group ($A = 0$) via the univariate KL program in Corollary 5.2, (c) *EF-All Subgroups*: modifying all groups to minimize KL divergence to the original distribution under exact fairness constraints (Proposition 5.3), and (d) *Mean Matching*: aligning group means while minimizing KL divergence (Proposition 5.5 in Section 5.8.2). Note that 'EF-All Subgroups' involves non-convex optimization, for which we apply line search to estimate the optimal mean–variance scaling factor γ. We report Demographic Parity (ΔDP), Equal Opportunity (ΔEO), KL and Jensen–Shannon divergence, as well as Bayes Error (BE) under each intervention. For the univariate Gaussian case, the Bayes-optimal thresholds are known in closed form (Lemma 5.1 in Section 5.8.1), and allow precise computation of ΔDP and ΔEO via standard Gaussian CDFs.

We have already demonstrated the effect of different interventions for high ΔEO in Figure 5.1. To cover another interesting case, we look at a distribution where ΔDP is high in Figure 5.2. Here, the affirmative action intervention transforms both the under-privileged subgroups to high-variance ones, which results in a reduction of BE and ΔDP, but at the cost of a high KL/JS-divergence with respect to the true distribution. However, the EF-All intervention simply tries to match the variances of underprivileged and privileged subgroups and, as a result, achieves perfect fairness and accuracy while staying close to the true distribution. Mean Matching is very similar to EF Affirmative in this case and, as a result, has relatively high KL/JS numbers. Due to a lack of space, we show more results for different settings: (1) same but shifted group distributions (Figure 5.5), (2) the case of ΔEO being close to 0



(a) Original Distribution **(b)** EF Affirmative **(c)** EF-All Subgroups **(d)** Mean Matching

Figure 5.2: Comparison of different interventions when the ΔDP on the original distribution is high. In this case, EF-All manages to stay close to the true distribution and achieves perfect fairness and error rate, while others deviate significantly.

(Figure 5.6), and (3) the case of a different cost sensitive risk ($t_C = 3/4$ on $\eta(x, a)$) (Figure 5.7), in Section 5.8.3.

§ 5.6 Applications: Steering Representations of LLMs

We now empirically demonstrate how our guarantees can improve representation and generation steering in LLMs—an active area with limited theoretical grounding [Han+23; Sin+24; Zha+24; Tan+24]. Adopting the setup of Singh et al. [Sin+24], we first apply affine steering to pretrained LLM representations to reduce class-specific TPR gaps on the Bias in Bios dataset [DA+19]. We then use the framework of Zhao et al. [Zha+24] to steer generations of movie reviews, showing that our intervention improves expressed joy for the underperforming group.

§ 5.6.1 Reducing Per-class Disparity in Multi-class Classification

In this experiment, we aim to steer data representations to reduce disparities between groups in a multi-class classification setting. We use the Bias in Bios dataset [DA+19], which comprises web-sourced biographies labelled by profession and annotated for gender. The task is to predict the profession from the biography while ensuring fairness with respect to a chosen metric. Our experimental setup closely follows that of Singh et al. [Sin+24], who generate representations using the Llama-2 7b model [Tou+23] and propose a method called MiMiC (Mean+Covariance Matching), which steers representations via least squares alignment of the first two moments. They measure the TPR-gap (Definition 2.4), defined as: $\text{TPR-gap}_y(h) := |\mathbb{P}[h(X, A) = y \mid Y = y, A = 1] - \mathbb{P}[h(X, A) = y \mid Y = y, A = 0]|$.

They demonstrate that MiMiC significantly reduces the TPR-gap on the Llama-2 7b embeddings of the Bios dataset. We adopt the same experimental pipeline as Singh et al. [SK22]; further details are provided in Section 5.8.4.1. Our intervention, referred to as *EF Affirmative* in Figure 5.3, begins by estimating the first two moments of the representations from Singh et al.'s pipeline. We then apply our Affirmative Action framework to compute target moments corresponding to an ideal distribution. As in Singh et al., we estimate affine transformation parameters that map the original moments to the target ones and apply this transformation to all data representations. We use a multi-class generalization of our optimization program in Theorem 5.1, and we lay out the details of the program and intervention in Section 5.8.4.1.

We additionally evaluate the LEACE transformation [Bel+23b], which Singh et al. [SK22] used as a baseline. Figure 5.3 reports the root-mean-square TPR-gaps across classes for various methods. Our intervention consistently reduces TPR-gaps across all classes and, in many cases, outperforms both MiMiC and LEACE. This provides empirical support for our approach: actively searching for and align-

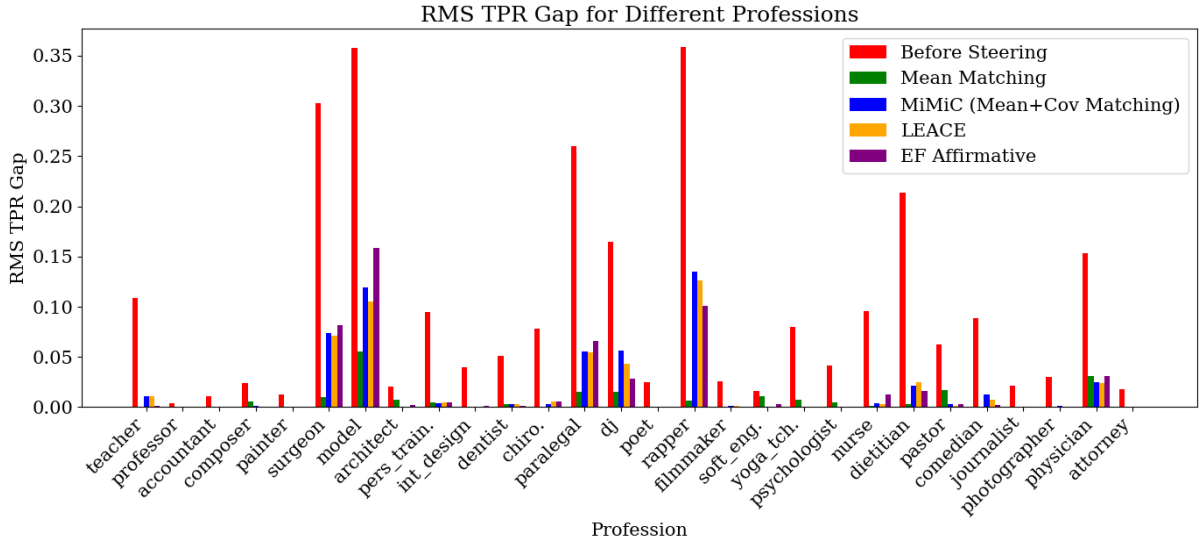


Figure 5.3: TPR-gap between Gender groups for all professions. All methods to steer feature representations achieve roughly the same accuracy (in the range of 0.77-0.79). Our intervention (EF Affirmative) is able to significantly reduce the TPR-gap for all professions. In many cases, it is even comparable or better than previous interventions Belrose et al. [Bel+23b] and Singh et al. [Sin+24].

ing with an ideal distribution can meaningfully reduce group disparities in representation learning.

§ 5.6.2 Steering Activations for Joyful Generation

We steer LLM generation towards joyful movie reviews using the Gaussian Concept Steer (GCS) framework of Zhao et al. [Zha+24], which samples steering vectors from a Gaussian $v_c^l \sim \mathcal{N}(\mu_c^l, \Sigma_c^l)$ for concept c at layer l. This improves robustness across models and datasets compared to using a single vector. Concept vectors for joy are estimated from GPT-4o [Hur+24] outputs and used to steer a Llama-3 8B model [Gra+24]. At each layer l, the final token representation h^l is updated as $h^l = (1-a) \cdot h^l + a \cdot v_c^l$, where a is the strength of the steering. Zhao et al. [Zha+24] experiment with values of $a \in (0.03, 0.08)$. We fix $a = 0.03$ for our experiments. Following [Zha+24], we evaluate the joyful tone using GPT-4 as an oracle with fixed decoding parameters. We apply the GCS framework to steer movie review generation toward joy over anger, focusing on two groups: comedy and horror reviews. For each group, we estimate Gaussian distributions of steering vectors for both directions (joyful $\rightarrow$ angry and angry $\rightarrow$ joyful), with the latter used for intervention. Full details are in Section 5.8.4.2.

While steering improves joyfulness overall, its effectiveness varies between groups. To address this, we apply our *EF Affirmative* intervention to nudge the joyful vector for the horror group. Let $\mathcal{N}(\mu_c^l, \Sigma_c^l)$ be the original distribution and $\mathcal{N}(\tilde{\mu}_c^l, \tilde{\Sigma}_c^l)$ be the distribution obtained after applying EF Affirmative intervention (Theorem 5.1). We define $\tilde{v}_c^l = (1-\alpha) \cdot v_c^l + \alpha \cdot \tilde{v}_c^l$, where $v_c^l \sim \mathcal{N}(\mu_c^l, \Sigma_c^l)$ and $\tilde{v}_c^l \sim \mathcal{N}(\tilde{\mu}_c^l, \tilde{\Sigma}_c^l)$. The final steering step updates the last-token representation at layer l as $h^l = (1-a) \cdot h^l + a \cdot \tilde{v}_c^l$.

Figure 5.4 presents the simulation results. We report the change in joyfulness (denoted as Δ-Joyful) before and after steering for both the comedy and horror review groups. For the horror group, we additionally apply our EF intervention, denoted as *Steering+EF*, to adjust the steering vectors. We vary the nudging parameter α, where $\alpha = 0$ corresponds to the original steering vectors from Zhao et al. [Zha+24]. As expected, moderate values of α improve joyfulness in the horror group, but excessive nudging distorts the steering vector, reducing its effectiveness.

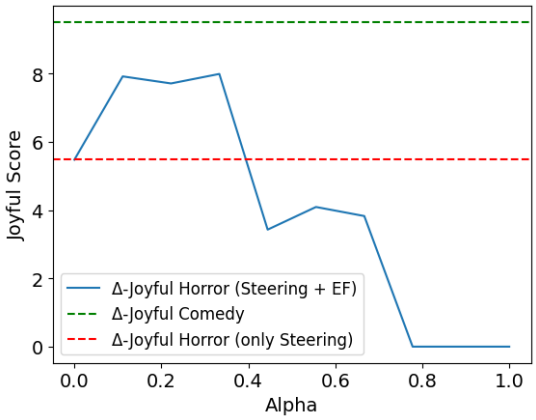


Figure 5.4: Change in Joyful score(Δ-Joyful) before and after adjusting the steering. Our intervention ('EF Affirmative') pushes up the effectiveness of steering the 'Horror' group, relative to the 'Comedy' group.

§ 5.7 Discussion and Future Work

We address fair classification by introducing the notion of *ideal* distributions—those that admit the Bayes-optimal classifier satisfying exact fairness. For common parametric families, we show that identifying such distributions reduces to a KL divergence minimisation problem, subject to fairness constraints on the Bayes-optimal classifier. We characterize conditions under which this problem is feasible and tractable. Through simulations, we demonstrate that the resulting interventions can steer the original distribution towards both perfect fairness and Bayes-optimal accuracy while remaining close in distributional distance. Additionally, we show that our method extends to post-training interventions, effectively steering learned representations toward desired behavioural outcomes.

We study the foundational problem of finding the nearest ideal distribution to a given distribution. This has many potential applications, such as correcting training data bias, learning fair representations, steering intermediate representations for fair generation, etc. Fair pre-processing or reweighing is closer to our mathematical setup, and steering intermediate representations gives a compelling application for LLMs. Our KL optimization program can be used to measure training data bias and guide data collection policies in practice, especially when the models are trained for fairness, but the inherent bias in data is unresolved.

Our theoretical results are on the fairness of the Bayes optimal classifier and the corresponding optimization programs, which are tractable and mathematically easier to analyse for parametric families of distributions, e.g., multivariate Gaussians. Multivariate Gaussians may not always fit real-world data, but can be justifiably used to model concepts and internal representations [Zha+24; EP25]. There can be scenarios and applications where a parametric assumption may not fit well with the input data distribution, and hence, analyzing the fair Bayes optimal classifier will become highly non-trivial.

An important extension is to determine how to steer a given distribution within a fixed budget so that exact fairness constraints are satisfied. This is relevant when deviations from the original distribution are costly or infeasible, and can inform data collection or active learning strategies for fair classification [Hol+19; JG20; Dut+20; Awa+21]. In such cases, it may be necessary to relax the fairness constraints and seek approximate solutions, prompting the need for efficient algorithms that operate under finite-sample settings. In a finite-sample regime, we have high probability estimates of the moments and other distribution-dependent quantities, so our approach leads to approximate fairness guarantees. To go beyond distributional assumptions, we can leverage previous work that maps given

data distribution to tractable, parametric families using invertible transforms that preserve Bayes error, so our results become applicable [The+21]. We can also assume distributions with bounded moments instead of parametric families, and still bound the Bayes error and other relevant quantities [MB24]. This requires a careful analysis and is certainly a very important future direction.

The framework of ideal distributions can also be applied to audit deployed models for fairness, particularly when the evaluation data may itself be biased [Akp+22; SDS23b; Bha+23], offering a pathway toward more robust fairness evaluation protocols. Prior work on fairness audit has mostly focused on auditing a given model instead of auditing training or evaluation data. There is a long line of work to demonstrate that the fairness-accuracy trade-off disappears when data bias is accounted for [WT+19; BS19; Dut+20; Mai+21; SD24; LRB25]. The KL gap in our formulation can be used as a metric of data bias, and Theorem 5.1 essentially gives an algorithmic recipe to find the minimally different distribution relative to the given one.

Our conditions can also be incorporated into optimization objectives to steer data generation toward ideal distributions. We demonstrate this by applying post-training interventions to steer LLM representations for multi-class fair classification and emotion control. More generally, our optimization framework and parametric conditions can be embedded into the training objectives of generative models. Modern generative approaches such as Variational Autoencoders [KW19] and Normalizing Flows [Pap+21] often assume parametric latent distributions, such as mixtures of Gaussians. Embedding fairness-aware ideality conditions into these objectives enables the generation of fair and accurate data distributions that remain close to the original. This represents a promising direction for fair generative modeling, an area of growing theoretical and practical interest [Xu+18; Cho+20; BRV21; Bre+21; TAC23; TAC24].

§ 5.8 Proofs and Additional Details

In this section, we present the proofs for the results presented in previous sections and the omitted details for experiments in Section 5.6.

§ 5.8.1 Proofs for Section 5.3

We will require a helper result about threshold classifiers to prove our next set of results.

Lemma 5.1. Let $\eta(x, a) = \Pr(Y = 1|X = x, A = a)$, $q_{ia} = \Pr(Y = i, A = a)$ and $p_{ia}(x) = \Pr\left(X = x \mid Y = i, A = a\right)$. Then the Bayes optimal classifier can be written as $h^*(x, a) = \mathbb{I}\left(\log \frac{p_{1a}(x)}{p_{0a}(x)} \geqslant \log \frac{q_{0a}}{q_{1a}}\right)$.

Proof. Let $\eta(x, a) = \Pr(Y = 1|X = x, A = a)$, $q_{ia} = \Pr(Y = i, A = a)$ and $p_{ia}(x) = \Pr\left(X = x \mid Y = i, A = a\right)$. We consider group-aware threshold classifiers on D of the form $h_t(x, a) = \mathbb{I}(\eta(x, a) \geqslant t)$, which can be

equivalently written as

$$\begin{aligned}
h_t(x,a) &= \mathbb{I}\left(\eta(x,a) \geqslant t\right) = \mathbb{I}\left(\Pr\left(Y=1 \,\middle|\, X=x, A=a\right) \geqslant t\right) \\
&= \mathbb{I}\left(\frac{\Pr\left(Y=1 \,\middle|\, X=x, A=a\right)}{\Pr\left(Y=0 \,\middle|\, X=x, A=a\right)} \geqslant \frac{t}{1-t}\right) \\
&= \mathbb{I}\left(\frac{\Pr\left(Y=1, X=x, A=a\right)}{\Pr\left(Y=0, X=x, A=a\right)} \geqslant \frac{t}{1-t}\right) \\
&= \mathbb{I}\left(\frac{\Pr\left(X=x \,\middle|\, Y=1, A=a\right)\Pr\left(Y=1, A=a\right)}{\Pr\left(X=x \,\middle|\, Y=0, A=a\right)\Pr\left(Y=0, A=a\right)} \geqslant \frac{t}{1-t}\right) \\
&= \mathbb{I}\left(\frac{p_{1a}(x)}{p_{0a}(x)} \geqslant \frac{t}{1-t}\cdot\frac{q_{0a}}{q_{1a}}\right) \\
&= \mathbb{I}\left(\log\frac{p_{1a}(x)}{p_{0a}(x)} \geqslant \log\frac{t}{1-t} + \log\frac{q_{0a}}{q_{1a}}\right).
\end{aligned}$$

It is well-known that the group-aware Bayes optimal classifier $h^* = h_{1/2}$ by setting $t = 1/2$, or equivalently,

$$h^*(x,a) = h_{1/2}(x,a) = \mathbb{I}\left(\log\frac{p_{1a}(x)}{p_{0a}(x)} \geqslant \log\frac{q_{0a}}{q_{1a}}\right).$$

□

We now prove Proposition 5.1 from the main text.

Proof. (Proof of Proposition 5.1) The Bayes optimal classifier for group $A = a$ in a multi-class setting can be written down as a maximum over posterior probabilities:

$$h^*(x,a) = \arg\max_{y\in\mathcal{Y}} \eta_y(x,a), \text{ where } \eta_y(x,a) = \Pr\left(Y=y|X=x, A=a\right).$$

We can say that $h^*(x,a) = y$, whenever the following happens:

$$h^*(x,a) = \arg\max_{y\in\mathcal{Y}} \eta_y(x,a) = \mathbb{I}\left(\frac{\eta_y(x,a)}{\eta_i(x,a)} \geqslant 1, \forall i \in \mathcal{Y}\right) = \mathbb{I}\left(\log\frac{p_{ya}(X)}{p_{ia}(X)} \geqslant \log\frac{q_{ia}}{q_{ya}}, \forall i \in \mathcal{Y}\right)$$

Using the above simplification, the EO-fairness condition:

$$\Pr\left(h^*(X,A) = y \,\middle|\, Y=y, A=a\right) = \Pr\left(h^*(X,A) = y \,\middle|\, Y=y, A=a'\right) \;\; \forall y \in \mathcal{Y} \text{ and } a, a' \in \mathcal{A}$$

means

$$\begin{aligned}
&\Pr\left(\log\frac{p_{ya}(X)}{p_{ia}(X)} \geqslant \log\frac{q_{ia}}{q_{ya}}, \forall i \in \mathcal{Y} \,\middle|\, Y=y, A=a\right) \\
&= \Pr\left(\log\frac{p_{ya'}(X)}{p_{ia'}(X)} \geqslant \log\frac{q_{ia'}}{q_{ya'}}, \forall i \in \mathcal{Y} \,\middle|\, Y=y, A=a'\right).
\end{aligned}$$

Since $X|Y=i, A=a \sim \mathcal{N}(\mu_{ia}, \Sigma_{ia})$ are multivariate Normal distributions, their probability densities are

$$p_{ia}(x) = (2\pi)^{-d/2}\det(\Sigma_{ia})^{-1/2}\exp\left(-\frac{1}{2}(x-\mu_{ia})^\top\Sigma_{ia}^{-1}(x-\mu_{ia})\right).$$

Now we can write

$$\begin{aligned}
&\log \frac{p_{ya}(x)}{p_{ia}(x)} \\
&= \frac{1}{2}\left((x-\mu_{ia})^T\Sigma_{ia}^{-1}(x-\mu_{ia}) - (x-\mu_{ya})^T\Sigma_{ya}^{-1}(x-\mu_{ya}) + \log\det(\Sigma_{ia}) - \log\det(\Sigma_{ya})\right) \\
&= \frac{1}{2}\left((\Sigma_{ya}^{1/2}r+\mu_{ya}-\mu_{ia})^T\Sigma_{ia}^{-1}(\Sigma_{ya}^{1/2}r+\mu_{ya}-\mu_{ia}) - r^Tr - \log\det\left(\Sigma_{ya}^{1/2}\Sigma_{ia}^{-1}\Sigma_{ya}^{1/2}\right)\right) \\
&\qquad\qquad \text{by substituting } x = \Sigma_{ya}^{1/2}r+\mu_{ya}, \text{ where } r\sim\mathcal{N}(0, I_{d\times d}) \\
&= \frac{1}{2}\, r^T\Sigma_{ya}^{1/2}\Sigma_{ia}^{-1}\Sigma_{ya}^{1/2}r + (\mu_{ya}-\mu_{ia})^T\Sigma_{ia}^{-1}\Sigma_{ya}^{1/2}r + \frac{1}{2}(\mu_{ya}-\mu_{ia})^T\Sigma_{ia}^{-1}(\mu_{ya}-\mu_{ia}) \\
&\qquad\qquad - \frac{1}{2}r^Tr - \frac{1}{2}\log\det\left(\Sigma_{ya}^{1/2}\Sigma_{ia}^{-1}\Sigma_{ya}^{1/2}\right)
\end{aligned}$$

Let us denote the above expression as $E_{yi}(r)$. We can now write the group TPR as:

$$\Pr\left(\log\frac{p_{ya}(X)}{p_{ia}(X)} \geqslant \log\frac{q_{ia}}{q_{ya}}, \forall i\in\mathcal{Y} \mid Y=y, A=a\right) = \Pr\left(E_{yi}(R) \geqslant \log\frac{q_{ia}}{q_{ya}}, \forall i\in\mathcal{Y}\right),$$

for $R\sim\mathcal{N}(\bar{0}, I_{d\times d})$. Now if we have $\dfrac{q_{ya}}{q_{ia}} = \dfrac{q_{ya'}}{q_{ia'}}$ and

$$\Sigma_{ya}^{-1/2}(\mu_{ya}-\mu_{ia}) = \Sigma_{ya'}^{-1/2}(\mu_{ya'}-\mu_{ia'}) \quad\text{and}\quad \Sigma_{ya}^{1/2}\Sigma_{ia}^{-1}\Sigma_{ya}^{1/2} = \Sigma_{ya'}^{1/2}\Sigma_{ia'}^{-1}\Sigma_{ya'}^{1/2},$$

then the probability of the above event written in terms $R\sim\mathcal{N}(\bar{0}, I_{d\times d})$ becomes identical $\forall a, a'\in\mathcal{A}$, $i\in\mathcal{Y}$. Hence, the Bayes optimal classifier satisfies equal opportunity under these conditions. □

Proof. (Proof of Proposition 5.2) For any cost matrix $C\in\mathbb{R}^{2\times 2}$, the group-aware classifier that minimizes its corresponding cost-sensitive risk is given by $\mathbb{I}\left(\eta(x,a)\geqslant t_C\right)$, for a threshold $t_C = (c_{10}-c_{00})/(c_{10}-c_{00}+c_{01}-c_{11})\in[0,1]$; see Equation (2) in [Elk01] and [Sco12]. The distribution D is *ideal* for equal opportunity if $\Pr\left(\eta(X,A)\geqslant t \mid Y=i, A=0\right) = \Pr\left(\eta(X,A)\geqslant t \mid Y=i, A=1\right)$, for all thresholds $t\in[0,1]$ and $i\in\{0,1\}$. Since the CDFs are identical, the random variables $\eta(X,A)\mid Y=i, A=0$ and

$\eta(X, A) \mid Y = i, A = 1$ must be identical. Note that

$$\begin{aligned}
\eta(x, a) &= \Pr\left(Y = 1 \mid X = x, A = a\right) = \frac{\Pr\left(Y = 1, X = x, A = a\right)}{\sum_{i=0}^{1} \Pr\left(Y = i, X = x, A = a\right)} \\
&= \frac{\Pr\left(Y = 1, A = a\right) \Pr\left(X = x \mid Y = 1, A = a\right)}{\sum_{i=0}^{1} \Pr\left(Y = i, A = a\right) \Pr\left(X = x \mid Y = i, A = a\right)} = \frac{q_{1a} P_{1a}(x)}{\sum_{i=0}^{1} q_{ia} P_{ia}(x)} \\
&= \frac{q_{1a} \sigma_{1a}^{-1} \exp\left(-\frac{(x-\mu_{1a})^2}{2\sigma_{1a}^2}\right)}{\sum_{i=0}^{1} q_{ia} \sigma_{ia}^{-1} \exp\left(-\frac{(x-\mu_{ia})^2}{2\sigma_{ia}^2}\right)} \\
&= \frac{1}{1 + \exp\left(\frac{(x-\mu_{1a})^2}{2\sigma_{1a}^2} - \frac{(x-\mu_{0a})^2}{2\sigma_{0a}^2} + \log\frac{q_{0a}\sigma_{1a}}{q_{1a}\sigma_{0a}}\right)} \\
&= \frac{1}{1 + \exp\left(\frac{(\mu_{ia}+r\sigma_{ia}-\mu_{1a})^2}{2\sigma_{1a}^2} - \frac{(\mu_{ia}+r\sigma_{ia}-\mu_{0a})^2}{2\sigma_{0a}^2} + \log\frac{q_{0a}\sigma_{1a}}{q_{1a}\sigma_{0a}}\right)} \\
&= \begin{cases}
\dfrac{1}{1 + \exp\left(\frac{1}{2}\left(\frac{\sigma_{0a}^2}{\sigma_{1a}^2} - 1\right) r^2 - \frac{\sigma_{0a}(\mu_{1a}-\mu_{0a})}{\sigma_{1a}^2} r + \frac{(\mu_{1a}-\mu_{0a})^2}{2\sigma_{1a}^2} + \log\frac{q_{0a}\sigma_{1a}}{q_{1a}\sigma_{0a}}\right)}, & \text{for } i = 0 \\
\dfrac{1}{1 + \exp\left(\frac{1}{2}\left(1 - \frac{\sigma_{1a}^2}{\sigma_{0a}^2}\right) r^2 - \frac{\sigma_{1a}(\mu_{0a}-\mu_{1a})}{\sigma_{0a}^2} r + \frac{(\mu_{0a}-\mu_{1a})^2}{2\sigma_{0a}^2} + \log\frac{q_{0a}\sigma_{1a}}{q_{1a}\sigma_{0a}}\right)}, & \text{for } i = 1.
\end{cases}
\end{aligned}$$

If $X|Y = i, A = a \sim \mathcal{N}(\mu_{ia}, \sigma_{ia}^2)$, then $X \mid Y = i, A = a \sim \mathcal{N}(\mu_{ia}, \sigma_{ia}^2)$. Thus, for $\eta(X, A) \mid Y = i, A = 0$ and $\eta(X, A) \mid Y = i, A = 1$ to be identical, we must have

$$\begin{aligned}
&\frac{1}{2}\left(\frac{\sigma_{00}^2}{\sigma_{10}^2} - 1\right) R^2 - \frac{\sigma_{00}(\mu_{10}-\mu_{00})}{\sigma_{10}^2} R + \frac{(\mu_{10}-\mu_{00})^2}{2\sigma_{10}^2} + \log\frac{q_{00}\sigma_{10}}{q_{10}\sigma_{00}} \quad \text{and} \\
&\frac{1}{2}\left(\frac{\sigma_{01}^2}{\sigma_{11}^2} - 1\right) R^2 - \frac{\sigma_{01}(\mu_{11}-\mu_{01})}{\sigma_{11}^2} R + \frac{(\mu_{11}-\mu_{01})^2}{2\sigma_{11}^2} + \log\frac{q_{01}\sigma_{11}}{q_{11}\sigma_{01}}
\end{aligned}$$

as identically distributed for $R \sim \mathcal{N}(0, 1)$. Similarly, we must also have

$$\begin{aligned}
&\frac{1}{2}\left(1 - \frac{\sigma_{10}^2}{\sigma_{00}^2}\right) R^2 - \frac{\sigma_{10}(\mu_{00}-\mu_{10})}{\sigma_{00}^2} R + \frac{(\mu_{00}-\mu_{10})^2}{2\sigma_{00}^2} + \log\frac{q_{00}\sigma_{10}}{q_{10}\sigma_{00}} \quad \text{and} \\
&\frac{1}{2}\left(1 - \frac{\sigma_{11}^2}{\sigma_{01}^2}\right) R^2 - \frac{\sigma_{11}(\mu_{01}-\mu_{11})}{\sigma_{01}^2} R + \frac{(\mu_{01}-\mu_{11})^2}{2\sigma_{01}^2} + \log\frac{q_{01}\sigma_{11}}{q_{11}\sigma_{01}}
\end{aligned}$$

as identically distributed for $R \sim \mathcal{N}(0, 1)$. Therefore, we must have

$$\frac{\mu_{01}-\mu_{11}}{\sigma_{11}} = \frac{\mu_{00}-\mu_{10}}{\sigma_{10}} \quad \text{and} \quad \frac{\sigma_{11}}{\sigma_{01}} = \frac{\sigma_{10}}{\sigma_{00}} \quad \text{and} \quad \frac{q_{10}}{q_{00}} = \frac{q_{11}}{q_{01}}.$$

In the other direction, it is easier to prove that the above conditions imply the distribution to be ideal. It can be proved by simply backtracking the steps above. □

§ 5.8.2 Proofs for Section 5.4

We first derive the KL divergence between two distributions, where each subgroup follows a multivariate normal distribution.

Lemma 5.2. Let (X, Y, A) denote the features, binary class label, and binary group membership, respectively, of a random data point from any data distribution D with $q_{ia} = \Pr(Y = i, A = a)$, for $i \in \mathcal{Y}$ and $a \in \mathcal{A}$. Let $X|Y = i, A = a \sim \mathcal{N}(\mu_{ia}, \Sigma_{ia})$ be multivariate Normal distributions with mean $\mu_{ia} \in \mathbb{R}^d$ and covariance matrix $\Sigma_{ia} \in \mathbb{R}^{d \times d}$. Let $\tilde{D}$ denote a distribution obtained by keeping (Y, A) unchanged and only changing $X|Y = i, A = a$ to $\tilde{X}|Y = i, A = a \sim \mathcal{N}(\tilde{\mu}_{ia}, \tilde{\Sigma}_{ia})$. Then,

$$D_{KL}\left(\tilde{D}\|D\right) = -\frac{d}{2} + \frac{1}{2}\sum_{(i,a)} q_{ia}(\tilde{\mu}_{ia} - \mu_{ia})^T \Sigma_{ia}^{-1}(\tilde{\mu}_{ia} - \mu_{ia}) + \frac{1}{2}\sum_{(i,a)} q_{ia}\left(\mathrm{tr}\left\{\Sigma_{ia}^{-1}\tilde{\Sigma}_{ia}\right\} - \log\det\left(\Sigma_{ia}^{-1}\tilde{\Sigma}_{ia}\right)\right).$$

Proof.

$$\begin{aligned}
&D_{KL}\left(\tilde{D}\|D\right) \\
&= \sum_{(x,i,a)} \Pr\left(\tilde{X} = x, \tilde{Y} = i, \tilde{A} = a\right) \log \frac{\Pr\left(\tilde{X} = x, \tilde{Y} = i, \tilde{A} = a\right)}{\Pr\left(X = x, Y = i, A = a\right)} \\
&= \sum_{(x,i,a)} \Pr\left(\tilde{Y} = i, \tilde{A} = a\right) \Pr\left(\tilde{X} = x \,\middle|\, \tilde{Y} = i, \tilde{A} = a\right) \\
&\qquad\qquad \log \frac{\Pr\left(\tilde{Y} = i, \tilde{A} = a\right) \Pr\left(\tilde{X} = x \,\middle|\, \tilde{Y} = i, \tilde{A} = a\right)}{\Pr\left(Y = y, A = a\right) \Pr\left(X = x \,\middle|\, Y = i, A = a\right)} \\
&= \sum_{(x,i,a)} \Pr\left(Y = i, A = a\right) \Pr\left(\tilde{X} = x \,\middle|\, Y = i, A = a\right) \\
&\qquad\qquad \log \frac{\Pr\left(Y = i, A = a\right) \Pr\left(\tilde{X} = x \,\middle|\, Y = i, A = a\right)}{\Pr\left(Y = i, A = a\right) \Pr\left(X = x \,\middle|\, Y = i, A = a\right)} \\
&= \sum_{(i,a)} q_{ia} \sum_{x} \Pr\left(\tilde{X} = x \,\middle|\, Y = i, A = a\right) \log \frac{\Pr\left(\tilde{X} = x \,\middle|\, Y = i, A = a\right)}{\Pr\left(X = x \,\middle|\, Y = i, A = a\right)} \\
&= \sum_{(i,a)} q_{ia} D_{KL}\left(\tilde{P}_{ia}\|P_{ia}\right)
\end{aligned}$$

P_{ia} denotes the distribution of $X \,\middle|\, Y = i, A = a \sim \mathcal{N}(\mu_{ia}, \Sigma_{ia})$ and $\tilde{P}_{ia}$ denotes the distribution of $\tilde{X} \,\middle|\, Y = i, A = a \sim \mathcal{N}(\tilde{\mu}_{ia}, \tilde{\Sigma}_{ia})$. Their probability densities are

$$p_{ia}(x) = (2\pi)^{-d/2} \det(\Sigma_{ia})^{-1/2} \exp\left(-\frac{1}{2}(x - \mu_{ia})^T \Sigma_{ia}^{-1}(x - \mu_{ia})\right) \quad \text{and}$$

$$\tilde{p}_{ia}(x) = (2\pi)^{-d/2} \det(\tilde{\Sigma}_{ia})^{-1/2} \exp\left(-\frac{1}{2}(x - \tilde{\mu}_{ia})^T \tilde{\Sigma}_{ia}^{-1}(x - \tilde{\mu}_{ia})\right),$$

respectively. Hence, the Kullback-Leibler divergence between $\tilde{P}_{ia}$ and P_{ia} can be written as

$$\begin{aligned}
&D_{KL}\left(\tilde{P}_{ia}\|P_{ia}\right)\\
&= E\log\frac{\tilde{p}_{ia}(\tilde{X})}{p_{ia}(\tilde{X})} \;\Big|\; Y=i, A=a\\
&= \frac{1}{2}\,E\Big[(\tilde{X}-\mu_{ia})^T\Sigma_{ia}^{-1}(\tilde{X}-\mu_{ia}) - (\tilde{X}-\tilde{\mu}_{ia})^T\tilde{\Sigma}_{ia}^{-1}(\tilde{X}-\tilde{\mu}_{ia})\\
&\qquad\qquad - \log\det\Big(\Sigma_{ia}^{1/2}\tilde{\Sigma}_{ia}^{-1}\Sigma_{ia}^{1/2}\Big) \;\Big|\; Y=i, A=a\Big]\\
&= \frac{1}{2}\,E(\tilde{X}-\mu_{ia})^T\Sigma_{ia}^{-1}(\tilde{X}-\mu_{ia}) \;\Big|\; Y=i, A=a - \frac{d}{2} - \frac{1}{2}\log\det\Big(\Sigma_{ia}^{1/2}\tilde{\Sigma}_{ia}^{-1}\Sigma_{ia}^{1/2}\Big)\\
&\qquad \text{using } E(\tilde{X}-\tilde{\mu}_{ia})^T\tilde{\Sigma}_{ia}^{-1}(\tilde{X}-\tilde{\mu}_{ia}) \;\Big|\; Y=i, A=a = \tilde{\Sigma}_{ia}^{-1}\bullet\tilde{\Sigma}_{ia} = \mathrm{tr}\{I_{d\times d}\} = d\\
&= \frac{1}{2}\,E(\tilde{X}-\tilde{\mu}_{ia}+\tilde{\mu}_{ia}-\mu_{ia})^T\Sigma_{ia}^{-1}(\tilde{X}-\tilde{\mu}_{ia}+\tilde{\mu}_{ia}-\mu_{ia}) \;\Big|\; Y=i, A=a\\
&\qquad\qquad - \frac{d}{2} + \frac{1}{2}\log\det\Big(\Sigma_{ia}^{1/2}\tilde{\Sigma}_{ia}^{-1}\Sigma_{ia}^{1/2}\Big)\\
&= \frac{1}{2}\,\mathrm{tr}\Big\{\Sigma_{ia}^{-1}\tilde{\Sigma}_{ia}\Big\} + \frac{1}{2}(\tilde{\mu}_{ia}-\mu_{ia})^T\Sigma_{ia}^{-1}(\tilde{\mu}_{ia}-\mu_{ia}) - \frac{d}{2} + \frac{1}{2}\log\det\Big(\Sigma_{ia}^{1/2}\tilde{\Sigma}_{ia}^{-1}\Sigma_{ia}^{1/2}\Big).
\end{aligned}$$

The Kullback-Leibler divergence between $\tilde{D}$ and D can now be written as

$$\begin{aligned}
&D_{KL}\left(\tilde{D}\|D\right) = \sum_{(i,a)} q_{ia} D_{KL}\left(\tilde{P}_{ia}\|P_{ia}\right)\\
&= \sum_{(i,a)} q_{ia}\left(\frac{1}{2}\,\mathrm{tr}\Big\{\Sigma_{ia}^{-1}\tilde{\Sigma}_{ia}\Big\} + \frac{1}{2}(\tilde{\mu}_{ia}-\mu_{ia})^T\Sigma_{ia}^{-1}(\tilde{\mu}_{ia}-\mu_{ia}) - \frac{d}{2} + \frac{1}{2}\log\det\Big(\Sigma_{ia}^{1/2}\tilde{\Sigma}_{ia}^{-1}\Sigma_{ia}^{1/2}\Big)\right)\\
&= -\frac{d}{2} + \frac{1}{2}\sum_{(i,a)} q_{ia}\left(\mathrm{tr}\Big\{\Sigma_{ia}^{-1}\tilde{\Sigma}_{ia}\Big\} + (\tilde{\mu}_{ia}-\mu_{ia})^T\Sigma_{ia}^{-1}(\tilde{\mu}_{ia}-\mu_{ia}) + \log\det\Big(\Sigma_{ia}^{1/2}\tilde{\Sigma}_{ia}^{-1}\Sigma_{ia}^{1/2}\Big)\right)\\
&= -\frac{d}{2} + \frac{1}{2}\sum_{(i,a)} q_{ia}(\tilde{\mu}_{ia}-\mu_{ia})^T\Sigma_{ia}^{-1}(\tilde{\mu}_{ia}-\mu_{ia}) + \frac{1}{2}\sum_{(i,a)} q_{ia}\left(\mathrm{tr}\Big\{\Sigma_{ia}^{-1}\tilde{\Sigma}_{ia}\Big\} + \log\det\Big(\Sigma_{ia}\tilde{\Sigma}_{ia}^{-1}\Big)\right)\\
&= -\frac{d}{2} + \frac{1}{2}\sum_{(i,a)} q_{ia}(\tilde{\mu}_{ia}-\mu_{ia})^T\Sigma_{ia}^{-1}(\tilde{\mu}_{ia}-\mu_{ia}) + \frac{1}{2}\sum_{(i,a)} q_{ia}\left(\mathrm{tr}\Big\{\Sigma_{ia}^{-1}\tilde{\Sigma}_{ia}\Big\} - \log\det\Big(\Sigma_{ia}^{-1}\tilde{\Sigma}_{ia}\Big)\right)
\end{aligned}$$

□

Proof. (Proof of Theorem 5.1) Using Lemma 5.2 and Proposition 5.1, our objective is to minimize

$$\begin{aligned}
&D_{KL}\left(\tilde{D}\|D\right)\\
&= -\frac{d}{2} + \frac{1}{2}\sum_{(i,a)} q_{ia}(\tilde{\mu}_{ia}-\mu_{ia})^T\Sigma_{ia}^{-1}(\tilde{\mu}_{ia}-\mu_{ia}) + \frac{1}{2}\sum_{(i,a)} q_{ia}\left(\mathrm{tr}\Big\{\Sigma_{ia}^{-1}\tilde{\Sigma}_{ia}\Big\} - \log\det\Big(\Sigma_{ia}^{-1}\tilde{\Sigma}_{ia}\Big)\right),
\end{aligned}$$

subject to the constraints

$$\tilde{\Sigma}_{10}^{-1/2}(\tilde{\mu}_{10}-\tilde{\mu}_{00}) = \tilde{\Sigma}_{11}^{-1/2}(\tilde{\mu}_{11}-\tilde{\mu}_{01}) \quad\text{and}\quad \tilde{\Sigma}_{10}^{1/2}\tilde{\Sigma}_{00}^{-1}\tilde{\Sigma}_{10}^{1/2} = \tilde{\Sigma}_{11}^{1/2}\tilde{\Sigma}_{01}^{-1}\tilde{\Sigma}_{11}^{1/2}.$$

Suppose $\tilde{\Sigma}_{i0}$ and $\tilde{\Sigma}_{i1}$ do not commute. The constraints can be equivalently rewritten as follows.

$$\tilde{\mu}_{10}-\tilde{\mu}_{00} = \tilde{\Sigma}_{10}^{1/2}\tilde{\Sigma}_{11}^{-1/2}(\tilde{\mu}_{11}-\tilde{\mu}_{01}) \quad\text{and}\quad \tilde{\Sigma}_{11}^{-1/2}\tilde{\Sigma}_{10}^{1/2}\tilde{\Sigma}_{00}^{-1}\tilde{\Sigma}_{10}^{1/2}\tilde{\Sigma}_{11}^{-1/2} = \tilde{\Sigma}_{01}^{-1}.$$

Let $\Gamma = \tilde{\Sigma}_{i0}^{1/2}\tilde{\Sigma}_{i1}^{-1/2}$. For any fixed positive semidefinite matrix $\Gamma \in \mathbb{R}^{d\times d}$, our optimization problem can be divided into two separate parts that minimize

$$\sum_{(i,a)} q_{ia}(\tilde{\mu}_{ia} - \mu_{ia})^T \Sigma_{ia}^{-1}(\tilde{\mu}_{ia} - \mu_{ia}) \quad \text{subject to} \quad \tilde{\mu}_{10} - \tilde{\mu}_{00} = \Gamma(\tilde{\mu}_{11} - \tilde{\mu}_{01})$$

over $\tilde{\mu}_{ia} \in \mathbb{R}^d$, for $i, a \in \{0,1\}$, and minimize (after substituting $\tilde{\Sigma}_{i0}^{1/2} = \Gamma\tilde{\Sigma}_{i1}^{1/2}$)

$$\sum_{i=0}^{1} q_{i0}\left(\mathrm{tr}\left\{\Sigma_{i0}^{-1}\left(\Gamma\tilde{\Sigma}_{i0}^{1/2}\right)^2\right\} - \log\det\left(\Sigma_{i0}^{-1}\left(\Gamma\tilde{\Sigma}_{i0}^{1/2}\right)^2\right)\right) + q_{i1}\left(\mathrm{tr}\left\{\Sigma_{i1}^{-1}\tilde{\Sigma}_{i1}\right\} - \log\det\left(\Sigma_{i1}^{-1}\tilde{\Sigma}_{i1}\right)\right),$$
$$\text{subject to } \Gamma\tilde{\Sigma}_{11}^{1/2}\tilde{\Sigma}_{00}^{-1}\Gamma = \tilde{\Sigma}_{11}^{1/2}\tilde{\Sigma}_{01}^{-1}$$

over symmetric, positive semidefinite matrix-valued variable $\tilde{\Sigma}_{i1} \in \mathbb{R}^{d\times d}$, for $i \in \{0,1\}$. The first optimization in $\tilde{\mu}_{ia}$ is a constrained eigenvalue problem with linear constraints, i.e., minimize $x^TAx + x^Tb$ subject to $x^Tc = e$ [Gol73].

Let's consider the case of *Affirmative Action*, where we only change the means $\tilde{\mu}_{i0}$ and the covariance matrices $\tilde{\Sigma}_{i0}$ for the underprivileged group but keep those for the privileged group unchanged, i.e., $\tilde{\mu}_{i1} = \mu_{i1}$ and $\tilde{\Sigma}_{i1} = \Sigma_{i1}$. In that case, $\tilde{\Sigma}_{00}^{1/2} = \Gamma\Sigma_{01}^{1/2}$ and $\tilde{\Sigma}_{10}^{1/2} = \Gamma\Sigma_{11}^{1/2}$ get fixed. By substituting $\tilde{\mu}_{10} = \tilde{\mu}_{00} + \Gamma(\tilde{\mu}_{11} - \tilde{\mu}_{01}) = \tilde{\mu}_{00} + \Gamma(\mu_{11} - \mu_{01})$, we only need to optimize

$$q_{00}(\tilde{\mu}_{00} - \mu_{00})^T\Sigma_{00}^{-1}(\tilde{\mu}_{00} - \mu_{00}) + q_{10}(\tilde{\mu}_{00} + \Gamma(\mu_{11} - \mu_{01}) - \mu_{10})^T\Sigma_{10}^{-1}(\tilde{\mu}_{00} + \Gamma(\mu_{11} - \mu_{01}) - \mu_{10}),$$

or equivalently (ignoring the terms independent of $\tilde{\mu}_{00}$),

$$\tilde{\mu}_{00}^T\left(q_{00}\Sigma_{00}^{-1} + q_{10}\Sigma_{10}^{-1}\right)\tilde{\mu}_{00} - 2\left(\Sigma_{00}^{-1}\mu_{00} + \Sigma_{10}^{-1}\mu_{10} - \Sigma_{10}^{-1}\Gamma(\mu_{11} - \mu_{01})\right)^T\tilde{\mu}_{00}.$$

This is a convex objective in $\tilde{\mu}_{00}$ because its Hessian is positive semidefinite, i.e., $q_{00}\Sigma_{00}^{-1} + q_{10}\Sigma_{10}^{-1} \succcurlyeq 0$ [Boy04]. By equating the gradient to zero, we get the optimal solution for $\tilde{\mu}_{00}$, and we denote it by $\mu_{00}^*(\Gamma)$. Thus, the optimal solutions $\mu_{00}^*(\Gamma), \mu_{10}^*(\Gamma), \Sigma_{00}^*(\Gamma), \Sigma_{10}^*(\Gamma)$ for a fixed positive semidefinite $\Gamma \in \mathbb{R}^{d\times d}$ are given by

$$\begin{aligned}
\mu_{00}^*(\Gamma) &= \left(q_{00}\Sigma_{00}^{-1} + q_{10}\Sigma_{10}^{-1}\right)^{-1}\left(\Sigma_{00}^{-1}\mu_{00} + \Sigma_{10}^{-1}\mu_{10} - \Sigma_{10}^{-1}\Gamma(\mu_{11} - \mu_{01})\right) \quad \text{and} \\
\mu_{10}^*(\Gamma) &= \left(q_{00}\Sigma_{00}^{-1} + q_{10}\Sigma_{10}^{-1}\right)^{-1}\left(\Sigma_{00}^{-1}\mu_{00} + \Sigma_{10}^{-1}\mu_{10} - \Sigma_{10}^{-1}\Gamma(\mu_{11} - \mu_{01})\right) + \Gamma(\mu_{11} - \mu_{01}) \\
\Sigma_{00}^*(\Gamma) &= (\Gamma\Sigma_{01}^{1/2})^2 \\
\Sigma_{10}^*(\Gamma) &= (\Gamma\Sigma_{11}^{1/2})^2.
\end{aligned}$$

By substituting these, when we look at the objective as a function of a positive semidefinite matrix-valued variable Γ, it turns out to be convex. This requires rewriting the expressions using the identities:

$$\begin{aligned}
&\mathrm{tr}\{AB\} = \mathrm{tr}\{BA\}, \det(AB) = \det(A)\det(B), \\
&\mathrm{tr}\{AXBX\} = \mathrm{tr}\left\{(A^{1/2}XB^{1/2})(A^{1/2}XB^{1/2})^T\right\} \text{ and }, \\
&\log\det(AXBX) = \log\det\left(A^{1/2}XB^{1/2})(A^{1/2}XB^{1/2})^T\right),
\end{aligned}$$

for symmetric, positive semidefinite matrices A, B, X [PP+08].The convexity of the objective in Γ follows from the convexity of $\mathrm{tr}\{AXBX\}$ and $-\log\det(X)$ for matrix-valued variable X. Finally, we can efficiently solve it to obtain the optimal Γ^*. □

Proof. (Proof of Corollary 5.2)

$$
\begin{aligned}
&D_{KL}\left(\tilde{D}\|D\right)\\
&= \sum_{(x,i,a)} \Pr\left(\tilde{X}=x,\tilde{Y}=i,\tilde{A}=a\right)\log\frac{\Pr\left(\tilde{X}=x,\tilde{Y}=i,\tilde{A}=a\right)}{\Pr\left(X=x,Y=i,A=a\right)}\\
&= \sum_{(x,i,a)} \Pr\left(\tilde{Y}=i,\tilde{A}=a\right)\Pr\left(\tilde{X}=x \,\middle|\, \tilde{Y}=i,\tilde{A}=a\right)\\
&\qquad\qquad \log\frac{\Pr\left(\tilde{Y}=i,\tilde{A}=a\right)\Pr\left(\tilde{X}=x \,\middle|\, \tilde{Y}=i,\tilde{A}=a\right)}{\Pr\left(Y=y,A=a\right)\Pr\left(X=x \,\middle|\, Y=i,A=a\right)}\\
&= \sum_{(x,i,a)} \Pr\left(Y=i,A=a\right)\Pr\left(\tilde{X}=x \,\middle|\, Y=i,A=a\right)\\
&\qquad\qquad \log\frac{\Pr\left(Y=i,A=a\right)\Pr\left(\tilde{X}=x \,\middle|\, Y=i,A=a\right)}{\Pr\left(Y=i,A=a\right)\Pr\left(X=x \,\middle|\, Y=i,A=a\right)}\\
&= \sum_{(i,a)} q_{ia}\sum_{x}\Pr\left(\tilde{X}=x \,\middle|\, Y=i,A=a\right)\log\frac{\Pr\left(\tilde{X}=x \,\middle|\, Y=i,A=a\right)}{\Pr\left(X=x \,\middle|\, Y=i,A=a\right)}\\
&= \sum_{(i,a)} q_{ia} D_{KL}\left(\tilde{P}_{ia}\|P_{ia}\right)
\end{aligned}
$$

P_{ia} denotes the distribution of $X \mid Y=i,A=a \sim \mathcal{N}(\mu_{ia},\sigma_{ia}^2)$ and $\tilde{P}_{ia}$ denotes the distribution of $\tilde{X} \mid Y=i,A=a \sim \mathcal{N}(\tilde{\mu}_{ia},\tilde{\sigma}_{ia}^2)$. Their probability densities are

$$
p_{ia}(x)=\frac{1}{\sigma_{ia}\sqrt{2\pi}}\exp\left(-\frac{(x-\mu_{ia})^2}{2\sigma_{ia}^2}\right) \quad\text{and}\quad \tilde{p}_{ia}(x)=\frac{1}{\tilde{\sigma}_{ia}\sqrt{2\pi}}\exp\left(-\frac{(x-\tilde{\mu}_{ia})^2}{2\tilde{\sigma}_{ia}^2}\right),
$$

respectively. Hence,

$$
\begin{aligned}
&D_{KL}\left(\tilde{P}_{ia}\|P_{ia}\right)=E\log\frac{\tilde{p}_{ia}(\tilde{X})}{p_{ia}(\tilde{X})} \,\middle|\, Y=i,A=a\\
&=E\frac{(\tilde{X}-\mu_{ia})^2}{2\sigma_{ia}^2}-\frac{(\tilde{X}-\tilde{\mu}_{ia})^2}{2\tilde{\sigma}_{ia}^2}+\log\frac{\sigma_{ia}}{\tilde{\sigma}_{ia}} \,\middle|\, Y=i,A=a\\
&=E\left(\frac{1}{2\sigma_{ia}^2}-\frac{1}{2\tilde{\sigma}_{ia}^2}\right)\tilde{X}^2+\left(\frac{\tilde{\mu}_{ia}}{\tilde{\sigma}_{ia}^2}-\frac{\mu_{ia}}{\sigma_{ia}^2}\right)\tilde{X}+\left(\frac{\mu_{ia}^2}{2\sigma_{ia}^2}-\frac{\tilde{\mu}_{ia}^2}{2\tilde{\sigma}_{ia}^2}\right)+\log\frac{\sigma_{ia}}{\tilde{\sigma}_{ia}} \,\middle|\, Y=i,A=a\\
&=\left(\frac{1}{2\sigma_{ia}^2}-\frac{1}{2\tilde{\sigma}_{ia}^2}\right)(\tilde{\mu}_{ia}^2+\tilde{\sigma}_{ia}^2)+\left(\frac{\tilde{\mu}_{ia}}{\tilde{\sigma}_{ia}^2}-\frac{\mu_{ia}}{\sigma_{ia}^2}\right)\tilde{\mu}_{ia}+\left(\frac{\mu_{ia}^2}{2\sigma_{ia}^2}-\frac{\tilde{\mu}_{ia}^2}{2\tilde{\sigma}_{ia}^2}\right)+\log\frac{\sigma_{ia}}{\tilde{\sigma}_{ia}}\\
&=\frac{(\tilde{\mu}_{ia}-\mu_{ia})^2}{2\sigma_{ia}^2}+\frac{\tilde{\sigma}_{ia}^2-\sigma_{ia}^2}{2\sigma_{ia}^2}+\log\frac{\sigma_{ia}}{\tilde{\sigma}_{ia}},
\end{aligned}
$$

using $E\log\tilde{X} \mid Y=i,A=a = \tilde{\mu}_{ia}$ and $E(\log\tilde{X})^2 \mid Y=i,A=a = \tilde{\mu}_{ia}^2+\tilde{\sigma}_{ia}^2$. Since we only change

group $A = 0$, we want to minimize

$$\begin{aligned} D_{KL}\left(\tilde{D}||D\right) &= \sum_{i=0}^{1} q_{i0} D_{KL}\left(\tilde{P}_{i0}||P_{i0}\right) \\ &= \sum_{i=0}^{1} q_{i0}\left(\frac{(\tilde{\mu}_{i0}-\mu_{i0})^2}{2\sigma_{i0}^2} + \frac{\tilde{\sigma}_{i0}^2-\sigma_{i0}^2}{2\sigma_{i0}^2} + \log\frac{\sigma_{i0}}{\tilde{\sigma}_{i0}}\right) \end{aligned}$$

as a function of the variables $\tilde{\mu}_{i0}$ and $\tilde{\sigma}_{i0}$ subject to the constraints

$$\frac{\mu_{01}-\mu_{11}}{\sigma_{11}} = \frac{\tilde{\mu}_{00}-\tilde{\mu}_{10}}{\tilde{\sigma}_{10}} \quad \text{and} \quad \frac{\sigma_{11}}{\sigma_{01}} = \frac{\tilde{\sigma}_{10}}{\tilde{\sigma}_{00}} \quad \text{and} \quad \tilde{\sigma}_{ia} \geqslant 0, \text{ for all } (i,a).$$

Let's fix $\gamma \in \mathbb{R}_{\geqslant 0}$ and minimize

$$\mathcal{L}_\gamma = \sum_{i=0}^{1} q_{i0}\left(\frac{(\tilde{\mu}_{i0}-\mu_{i0})^2}{2\sigma_{i0}^2} + \frac{\tilde{\sigma}_{i0}^2-\sigma_{i0}^2}{2\sigma_{i0}^2} + \log\frac{\sigma_{i0}}{\tilde{\sigma}_{i0}}\right)$$

as a function of the variables $\tilde{\mu}_{ia}$ and $\tilde{\sigma}_{ia}$ subject to the following constraints

$$\frac{\tilde{\mu}_{00}-\tilde{\mu}_{10}}{\mu_{01}-\mu_{11}} = \frac{\tilde{\sigma}_{10}}{\sigma_{11}} = \frac{\tilde{\sigma}_{00}}{\sigma_{01}} = \gamma \quad \text{and} \quad \tilde{\sigma}_{ia} \geqslant 0, \text{ for all } (i,a).$$

The objective $\mathcal{L}_\gamma$ is convex and for a fixed $\gamma \in \mathbb{R}_{\geqslant 0}$, the constraints on are linear in $\tilde{\mu}_{i0}$ and $\tilde{\sigma}_{i0}$. Let's denote the optimal solution for a fixed $\gamma \in \mathbb{R}_{\geqslant 0}$ by $\mu_{i0}^*(\gamma)$ and $\sigma_{i0}^*(\gamma)$, for $i \in \{0,1\}$. For a fixed $\gamma \in \mathbb{R}_{\geqslant 0}$, the above constraints fix $\sigma_{i0}^*(\gamma) = \gamma\sigma_{i1}$, for $i \in \{0,1\}$, and by plugging in $\tilde{\mu}_{00} = \tilde{\mu}_{10} + \gamma(\mu_{01}-\mu_{11})$, we only need to minimize the following convex, quadratic objective in a single variable $\tilde{\mu}_{10}$,

$$\text{minimize} \quad q_{00}\frac{(\tilde{\mu}_{10}+\gamma(\mu_{01}-\mu_{11})-\mu_{00})^2}{2\sigma_{00}^2} + q_{10}\frac{(\tilde{\mu}_{10}-\mu_{10})^2}{2\sigma_{10}^2}.$$

By equating the derivative to zero, we get the optimal solution as

$$\mu_{10}^*(\gamma) = \left(\frac{q_{00}}{\sigma_{00}^2} + \frac{q_{10}}{\sigma_{10}^2}\right)^{-1}\left(q_{00}\frac{\mu_{00}-\gamma(\mu_{01}-\mu_{11})}{\sigma_{00}^2} + q_{10}\frac{\mu_{10}}{\sigma_{10}^2}\right),$$

and the optimal value at $\mu_{10}^*(\gamma)$ is (The min of ax^2+bx+c occurs at $x = \dfrac{-b}{2a}$ and has value $c - \dfrac{b^2}{4a}$)

$$\begin{aligned} &q_{00}\frac{(\gamma(\mu_{01}-\mu_{11})-\mu_{00})^2}{2\sigma_{00}^2} + q_{10}\frac{\mu_{10}^2}{2\sigma_{10}^2} - \frac{1}{2}\frac{\left(q_{00}\frac{\mu_{00}-\gamma(\mu_{01}-\mu_{11})}{\sigma_{00}^2} + q_{10}\frac{\mu_{10}}{\sigma_{10}^2}\right)^2}{\left(\frac{q_{00}}{\sigma_{00}^2}+\frac{q_{10}}{\sigma_{10}^2}\right)} \\ &= \frac{1}{2}\left(\frac{q_{00}}{\sigma_{00}^2}+\frac{q_{10}}{\sigma_{10}^2}\right)^{-1}\frac{q_{00}q_{10}}{\sigma_{00}^2\sigma_{10}^2}\left((\mu_{00}-\mu_{10})-\gamma(\mu_{01}-\mu_{11})\right)^2 \\ &= \frac{1}{2}\left(\frac{\sigma_{00}^2}{q_{00}}+\frac{\sigma_{10}^2}{q_{10}}\right)^{-1}\left((\mu_{00}-\mu_{10})-\gamma(\mu_{01}-\mu_{11})\right)^2. \end{aligned}$$

By plugging in the optimal solution, the minimum value of $\mathcal{L}_\gamma$ for a fixed $\gamma \in \mathbb{R}_{\geqslant 0}$ is given by

$$\begin{aligned}
\mathcal{L}^*_\gamma &= \sum_{i=0}^{1} q_{i0} \left(\frac{(\mu^*_{i0}(\gamma) - \mu_{i0})^2}{2\sigma^2_{i0}} + \frac{\sigma^{*}_{i0}(\gamma)^2 - \sigma^2_{i0}}{2\sigma^2_{i0}} + \log \frac{\sigma_{i0}}{\sigma^*_{i0}(\gamma)} \right) \\
&= \frac{1}{2} \left(\frac{\sigma^2_{00}}{q_{00}} + \frac{\sigma^2_{10}}{q_{10}} \right)^{-1} ((\mu_{00} - \mu_{10}) - \gamma(\mu_{01} - \mu_{11}))^2 \\
&+ q_{00} \frac{\gamma^2\sigma^2_{01} - \sigma^2_{00}}{2\sigma^2_{00}} + q_{10} \frac{\gamma^2\sigma^2_{11} - \sigma^2_{10}}{2\sigma^2_{10}} + (q_{00} + q_{10}) \log \frac{1}{\gamma} + q_{00} \log \frac{\sigma_{00}}{\sigma_{01}} + q_{10} \log \frac{\sigma_{10}}{\sigma_{11}} \\
&= \frac{1}{2} \left(\frac{\sigma^2_{00}}{q_{00}} + \frac{\sigma^2_{10}}{q_{10}} \right)^{-1} ((\mu_{00} - \mu_{10}) - \gamma(\mu_{01} - \mu_{11}))^2 + \frac{q_{00}}{2} \left(\gamma^2 \frac{\sigma^2_{01}}{\sigma^2_{00}} - 1 \right) \\
&\qquad + (q_{00} + q_{10}) \log \frac{1}{\gamma} + q_{00} \log \frac{\sigma_{00}}{\sigma_{01}} + q_{10} \log \frac{\sigma_{10}}{\sigma_{11}}.
\end{aligned}$$

This is a convex objective in γ (because the second derivative is non-negative) and by equating the derivative to zero, we have that the optimal γ^* must satisfy

$$\frac{(\mu_{01} - \mu_{11})\left(\gamma^*(\mu_{01} - \mu_{11}) - (\mu_{00} - \mu_{10})\right)}{\left(\frac{\sigma^2_{00}}{q_{00}} + \frac{\sigma^2_{10}}{q_{10}}\right)} + \gamma^* \left(q_{00} \frac{\sigma^2_{01}}{\sigma^2_{00}} + q_{10} \frac{\sigma^2_{11}}{\sigma^2_{10}} \right) - \frac{q_{00} + q_{10}}{\gamma^*} = 0.$$

Multiplying with $\gamma^* \left(\frac{\sigma^2_{00}}{q_{00}} + \frac{\sigma^2_{10}}{q_{10}} \right)$, we can write it as a quadratic equation as follows.

$$\begin{aligned}
&\left((\mu_{01} - \mu_{11})^2 + \sigma^2_{01} + \sigma^2_{11} + \frac{q_{10}\sigma^2_{00}}{q_{00}\sigma^2_{10}} \sigma^2_{11} + \frac{q_{00}\sigma^2_{10}}{q_{10}\sigma^2_{00}} \sigma^2_{01} \right) {\gamma^*}^2 \\
&-(\mu_{01} - \mu_{11})(\mu_{00} - \mu_{10})\gamma^* - (q_{00} + q_{10}) \left(\frac{\sigma^2_{00}}{q_{00}} + \frac{\sigma^2_{10}}{q_{10}} \right) = 0
\end{aligned}$$

The discriminant of the above quadratic polynomial is non-negative because the leading coefficient is positive and the constant term is negative. So this polynomial has two real roots. Moreover, since the constant term is negative, it cannot have both positive or both negative roots. Its only non-negative root is the optimal solution $\gamma^* \in \mathbb{R}_{\geqslant 0}$ we want.

$$\gamma^* = \frac{(\mu_{01} - \mu_{11})(\mu_{00} - \mu_{10}) + \sqrt{\Delta}}{2 \left((\mu_{01} - \mu_{11})^2 + \sigma^2_{01} + \sigma^2_{11} + \frac{q_{10}\sigma^2_{00}}{q_{00}\sigma^2_{10}} \sigma^2_{11} + \frac{q_{00}\sigma^2_{10}}{q_{10}\sigma^2_{00}} \sigma^2_{01} \right)}, \text{ where}$$

$$\begin{aligned}
\Delta &= (\mu_{01} - \mu_{11})^2 (\mu_{00} - \mu_{10})^2 \\
&+ 4 \left((\mu_{01} - \mu_{11})^2 + \sigma^2_{01} + \sigma^2_{11} + \frac{q_{10}\sigma^2_{00}}{q_{00}\sigma^2_{10}} \sigma^2_{11} + \frac{q_{00}\sigma^2_{10}}{q_{10}\sigma^2_{00}} \sigma^2_{01} \right) (q_{00} + q_{10}) \left(\frac{\sigma^2_{00}}{q_{00}} + \frac{\sigma^2_{10}}{q_{10}} \right).
\end{aligned}$$

□

Proof. (Proof of Proposition 5.3) We consider the following optimization program

$$\begin{aligned}\mathrm{D_{KL}}\left(\tilde{\mathrm{D}}\|\mathrm{D}\right) &= \sum_{(i,a)} q_{ia}\mathrm{D_{KL}}\left(\tilde{\mathrm{P}}_{ia}\|\mathrm{P}_{ia}\right)\\ &= \sum_{(i,a)} q_{ia}\left(\frac{(\tilde{\mu}_{ia}-\mu_{ia})^2}{2\sigma_{ia}^2}+\frac{\tilde{\sigma}_{ia}^2-\sigma_{ia}^2}{2\sigma_{ia}^2}+\log\frac{\sigma_{ia}}{\tilde{\sigma}_{ia}}\right)\end{aligned}$$

as a function of the variables $\tilde{\mu}_{ia}$ and $\tilde{\sigma}_{ia}$ subject to the constraints

$$\frac{\tilde{\mu}_{01}-\tilde{\mu}_{11}}{\tilde{\sigma}_{11}}=\frac{\tilde{\mu}_{00}-\tilde{\mu}_{10}}{\tilde{\sigma}_{10}} \quad\text{and}\quad \frac{\tilde{\sigma}_{11}}{\tilde{\sigma}_{01}}=\frac{\tilde{\sigma}_{10}}{\tilde{\sigma}_{00}} \quad\text{and}\quad \tilde{\sigma}_{ia}\geqslant 0, \text{ for all } (i,a).$$

Let's fix $\gamma\in\mathbb{R}_{\geqslant 0}$ and minimize

$$\mathcal{L}_\gamma=\sum_{(i,a)} q_{ia}\left(\frac{(\tilde{\mu}_{ia}-\mu_{ia})^2}{2\sigma_{ia}^2}+\frac{\tilde{\sigma}_{ia}^2-\sigma_{ia}^2}{2\sigma_{ia}^2}+\log\frac{\sigma_{ia}}{\tilde{\sigma}_{ia}}\right)$$

as a function of the variables $\tilde{\mu}_{ia}$ and $\tilde{\sigma}_{ia}$ subject to the following constraints

$$\frac{\tilde{\mu}_{01}-\tilde{\mu}_{11}}{\tilde{\mu}_{00}-\tilde{\mu}_{10}}=\frac{\tilde{\sigma}_{11}}{\tilde{\sigma}_{10}}=\frac{\tilde{\sigma}_{01}}{\tilde{\sigma}_{00}}=\gamma \quad\text{and}\quad \tilde{\sigma}_{ia}\geqslant 0, \text{ for all } (i,a).$$

Now the objective $\mathcal{L}_\gamma$ is convex and for a fixed $\gamma\in\mathbb{R}_{\geqslant 0}$, the constraints on are linear in $\tilde{\mu}_{ia}$ and $\tilde{\sigma}_{ia}$. Let's denote the optimal solution for a fixed $\gamma\in\mathbb{R}_{\geqslant 0}$ by $\mu^*_{ia}(\gamma)$ and $\sigma^*_{ia}(\gamma)$, for $i,a\in\{0,1\}$. To find this, we can split the above objective into parts that can be optimized separately as follows.

$$\text{minimize} \quad \sum_{(i,a)} q_{ia}\frac{(\tilde{\mu}_{ia}-\mu_{ia})^2}{2\sigma_{ia}^2} \quad \text{subject to} \quad \tilde{\mu}_{01}-\tilde{\mu}_{11}=\gamma(\tilde{\mu}_{00}-\tilde{\mu}_{10}), \quad \text{and}$$

$$\text{minimize} \quad \sum_{(i,a)} q_{ia}\left(\frac{\tilde{\sigma}_{ia}^2-\sigma_{ia}^2}{2\sigma_{ia}^2}+\log\frac{\sigma_{ia}}{\tilde{\sigma}_{ia}}\right) \quad \text{subject to} \quad \tilde{\sigma}_{i1}=\gamma\tilde{\sigma}_{i0}, \text{and } \tilde{\sigma}_{ia}\geqslant 0, \text{ for all } (i,a).$$

For each $i\in\{0,1\}$, by substituting $\tilde{\sigma}_{i1}=\gamma\tilde{\sigma}_{i0}$, we need to optimize a function in only one variable $\tilde{\sigma}_{i0}$. The optimal solutions $\sigma^*_{ia}(\gamma)$ turn out to be

$$\sigma^*_{i0}(\gamma)=\sqrt{\frac{q_{i0}+q_{i1}}{\frac{q_{i0}}{\sigma_{i0}^2}+\frac{q_{i1}\gamma^2}{\sigma_{i1}^2}}} \quad\text{and}\quad \sigma^*_{i1}(\gamma)=\gamma\sqrt{\frac{q_{i0}+q_{i1}}{\frac{q_{i0}}{\sigma_{i0}^2}+\frac{q_{i1}\gamma^2}{\sigma_{i1}^2}}}, \quad \text{for } i\in\{0,1\},$$

Now let's find the optimal solutions $\mu^*_{ia}(\gamma)$. The gradient of the objective must be parallel to the linear constraint, so

$$\begin{aligned}\frac{q_{00}(\mu^*_{00}(\gamma)-\mu_{00})}{\sigma_{00}^2}&=-\gamma\lambda, \quad \frac{q_{01}(\mu^*_{01}(\gamma)-\mu_{01})}{\sigma_{01}^2}=\lambda,\\ \frac{q_{10}(\mu^*_{10}(\gamma)-\mu_{10})}{\sigma_{10}^2}&=\gamma\lambda, \quad \frac{q_{11}(\mu^*_{11}(\gamma)-\mu_{11})}{\sigma_{11}^2}=-\lambda,\end{aligned}$$

for some $\lambda \in \mathbb{R}$, which gives

$$\mu^*_{00}(\gamma) = -\gamma\lambda\frac{\sigma^2_{00}}{q_{00}} + \mu_{00}, \quad \mu^*_{01}(\gamma) = \lambda\frac{\sigma^2_{01}}{q_{01}} + \mu_{01},$$
$$\mu^*_{10}(\gamma) = \gamma\lambda\frac{\sigma^2_{10}}{q_{10}} + \mu_{10}, \quad \mu^*_{11}(\gamma) = -\lambda\frac{\sigma^2_{11}}{q_{11}} + \mu_{11}.$$

Since $\mu^*_{ia}(\gamma)$ satisfies the constraint $\dfrac{\tilde{\mu}_{01} - \tilde{\mu}_{11}}{\tilde{\mu}_{00} - \tilde{\mu}_{10}} = \gamma$, we have

$$\frac{\lambda\frac{\sigma^2_{01}}{q_{01}} + \mu_{01} + \lambda\frac{\sigma^2_{11}}{q_{11}} - \mu_{11}}{-\gamma\lambda\frac{\sigma^2_{00}}{q_{00}} + \mu_{00} - \gamma\lambda\frac{\sigma^2_{10}}{q_{10}} - \mu_{10}} = \gamma, \quad \text{and hence,} \quad \lambda = \frac{\gamma(\mu_{00} - \mu_{10}) - (\mu_{01} - \mu_{11})}{\frac{\sigma^2_{01}}{q_{01}} + \frac{\sigma^2_{11}}{q_{11}} + \gamma^2\left(\frac{\sigma^2_{00}}{q_{00}} + \frac{\sigma^2_{10}}{q_{10}}\right)}.$$

Thus, we can express $\mu^*_{ia}(\gamma)$ as

$$\mu^*_{00}(\gamma) = -\gamma\frac{\gamma(\mu_{00} - \mu_{10}) - (\mu_{01} - \mu_{11})}{\frac{\sigma^2_{01}}{q_{01}} + \frac{\sigma^2_{11}}{q_{11}} + \gamma^2\left(\frac{\sigma^2_{00}}{q_{00}} + \frac{\sigma^2_{10}}{q_{10}}\right)}\frac{\sigma^2_{00}}{q_{00}} + \mu_{00}$$
$$\mu^*_{01}(\gamma) = \frac{\gamma(\mu_{00} - \mu_{10}) - (\mu_{01} - \mu_{11})}{\frac{\sigma^2_{01}}{q_{01}} + \frac{\sigma^2_{11}}{q_{11}} + \gamma^2\left(\frac{\sigma^2_{00}}{q_{00}} + \frac{\sigma^2_{10}}{q_{10}}\right)}\frac{\sigma^2_{01}}{q_{01}} + \mu_{01}$$
$$\mu^*_{10}(\gamma) = \gamma\frac{\gamma(\mu_{00} - \mu_{10}) - (\mu_{01} - \mu_{11})}{\frac{\sigma^2_{01}}{q_{01}} + \frac{\sigma^2_{11}}{q_{11}} + \gamma^2\left(\frac{\sigma^2_{00}}{q_{00}} + \frac{\sigma^2_{10}}{q_{10}}\right)}\frac{\sigma^2_{10}}{q_{10}} + \mu_{10}$$
$$\mu^*_{11}(\gamma) = -\frac{\gamma(\mu_{00} - \mu_{10}) - (\mu_{01} - \mu_{11})}{\frac{\sigma^2_{01}}{q_{01}} + \frac{\sigma^2_{11}}{q_{11}} + \gamma^2\left(\frac{\sigma^2_{00}}{q_{00}} + \frac{\sigma^2_{10}}{q_{10}}\right)}\frac{\sigma^2_{11}}{q_{11}} + \mu_{11}.$$

Thus, the optimal value of $\mathcal{L}_\gamma$ for a fixed $\gamma \in \mathbb{R}_{\geqslant 0}$ is given by

$$\mathcal{L}^*_\gamma = \sum_{(i,a)} q_{ia}\left(\frac{(\mu^*_{ia}(\gamma) - \mu_{ia})^2}{2\sigma^2_{ia}} + \frac{\sigma^*_{ia}(\gamma)^2 - \sigma^2_{ia}}{2\sigma^2_{ia}} + \log\frac{\sigma_{ia}}{\sigma^*_{ia}(\gamma)}\right).$$

Dividing the above expression into three parts, the first part evaluates to

$$\begin{aligned}
&\sum_{(i,a)} q_{ia} \frac{(\mu^*_{ia}(\gamma) - \mu_{ia})^2}{2\sigma^2_{ia}} = \frac{q_{00}}{2\sigma^2_{00}} \frac{\gamma^2\lambda^2\sigma^4_{00}}{q^2_{00}} + \frac{q_{01}}{2\sigma^2_{01}} \frac{\lambda^2\sigma^4_{01}}{q^2_{01}} + \frac{q_{10}}{2\sigma^2_{10}} \frac{\gamma^2\lambda^2\sigma^4_{10}}{q^2_{10}} + \frac{q_{11}}{2\sigma^2_{11}} \frac{\lambda^2\sigma^4_{11}}{q^2_{11}} \\
&= \frac{\gamma^2\lambda^2\sigma^2_{00}}{2q_{00}} + \frac{\lambda^2\sigma^2_{01}}{2q_{01}} + \frac{\gamma^2\lambda^2\sigma^2_{10}}{2q_{10}} + \frac{\lambda^2\sigma^2_{11}}{2q_{11}} \\
&= \frac{\lambda^2}{2}\left(\frac{\sigma^2_{01}}{q_{01}} + \frac{\sigma^2_{11}}{q_{11}} + \gamma^2\left(\frac{\sigma^2_{00}}{q_{00}} + \frac{\sigma^2_{10}}{q_{10}}\right)\right) \\
&= \frac{1}{2}\left(\frac{\gamma(\mu_{00} - \mu_{10}) - (\mu_{01} - \mu_{11})}{\frac{\sigma^2_{01}}{q_{01}} + \frac{\sigma^2_{11}}{q_{11}} + \gamma^2\left(\frac{\sigma^2_{00}}{q_{00}} + \frac{\sigma^2_{10}}{q_{10}}\right)}\right)^2 \left(\frac{\sigma^2_{01}}{q_{01}} + \frac{\sigma^2_{11}}{q_{11}} + \gamma^2\left(\frac{\sigma^2_{00}}{q_{00}} + \frac{\sigma^2_{10}}{q_{10}}\right)\right) \\
&= \frac{1}{2}\frac{(\gamma(\mu_{00} - \mu_{10}) - (\mu_{01} - \mu_{11}))^2}{\frac{\sigma^2_{01}}{q_{01}} + \frac{\sigma^2_{11}}{q_{11}} + \gamma^2\left(\frac{\sigma^2_{00}}{q_{00}} + \frac{\sigma^2_{10}}{q_{10}}\right)}.
\end{aligned}$$

The second part evaluates to

$$\begin{aligned}
&\sum_{(i,a)} q_{ia} \frac{\sigma^*_{ia}(\gamma)^2 - \sigma^2_{ia}}{2\sigma^2_{ia}} = \sum_{(i,a)} \frac{q_{ia}}{2}\left(\frac{\sigma^*_{ia}(\gamma)^2}{\sigma^2_{ia}} - 1\right) \\
&= \sum_{i=0}^{1} \frac{q_{i0}}{2}\left(\frac{q_{i0} + q_{i1}}{\sigma^2_{i0}\left(\frac{q_{i0}}{\sigma^2_{i0}} + \frac{q_{i1}\gamma^2}{\sigma^2_{i1}}\right)} - 1\right) + \sum_{i=0}^{1} \frac{q_{i1}}{2}\left(\frac{\gamma^2(q_{i0} + q_{i1})}{\sigma^2_{i1}\left(\frac{q_{i0}}{\sigma^2_{i0}} + \frac{q_{i1}\gamma^2}{\sigma^2_{i1}}\right)} - 1\right) \\
&= \sum_{i=0}^{1} \frac{q_{i0}}{2} \frac{q_{i1}\left(1 - \frac{\sigma^2_{i0}\gamma^2}{\sigma^2_{i1}}\right)}{q_{i0} + q_{i1}\frac{\sigma^2_{i0}\gamma^2}{\sigma^2_{i1}}} + \sum_{i=0}^{1} \frac{q_{i1}}{2} \frac{q_{i0}\left(\gamma^2 - \frac{\sigma^2_{i1}}{\sigma^2_{i0}}\right)}{q_{i0}\frac{\sigma^2_{i1}}{\sigma^2_{i0}} + q_{i1}\gamma^2} \\
&= \sum_{i=0}^{1} \frac{q_{i0}}{2} \frac{q_{i1}\left(1 - \frac{\sigma^2_{i0}\gamma^2}{\sigma^2_{i1}}\right)}{q_{i0} + q_{i1}\frac{\sigma^2_{i0}\gamma^2}{\sigma^2_{i1}}} + \sum_{i=0}^{1} \frac{q_{i1}}{2} \frac{q_{i0}\left(\frac{\sigma^2_{i0}\gamma^2}{\sigma^2_{i1}} - 1\right)}{q_{i0} + q_{i1}\frac{\sigma^2_{i0}\gamma^2}{\sigma^2_{i1}}} = 0,
\end{aligned}$$

and the third part evaluates to

$$\begin{aligned}
&\sum_{(i,a)} q_{ia} \log \frac{\sigma_{ia}}{\sigma^*_{ia}(\gamma)} = \sum_{i=0}^{1} \frac{q_{i0}}{2} \log \frac{\sigma^2_{i0}}{\sigma^*_{i0}(\gamma)^2} + \frac{q_{i1}}{2} \log \frac{\sigma^2_{i1}}{\sigma^*_{i1}(\gamma)^2} \\
&= \sum_{i=0}^{1} \frac{q_{i0}}{2} \log \frac{\sigma^2_{i0}\left(\frac{q_{i0}}{\sigma^2_{i0}} + \frac{q_{i1}\gamma^2}{\sigma^2_{i1}}\right)}{q_{i0}+q_{i1}} + \frac{q_{i1}}{2} \log \frac{\sigma^2_{i1}\left(\frac{q_{i0}}{\sigma^2_{i0}} + \frac{q_{i1}\gamma^2}{\sigma^2_{i1}}\right)}{\gamma^2(q_{i0}+q_{i1})} \\
&= \sum_{i=0}^{1} \frac{q_{i0}}{2} \log \frac{\frac{q_{i0}}{q_{i1}} + \gamma^2 \frac{\sigma^2_{i0}}{\sigma^2_{i1}}}{\frac{q_{i0}}{q_{i1}} + 1} + \frac{q_{i1}}{2} \log \frac{\frac{q_{i0}}{q_{i1}} + \gamma^2 \frac{\sigma^2_{i0}}{\sigma^2_{i1}}}{\gamma^2 \frac{\sigma^2_{i0}}{\sigma^2_{i1}}\left(\frac{q_{i0}}{q_{i1}} + 1\right)} \\
&= \sum_{i=0}^{1} \frac{q_{i0}+q_{i1}}{2} \log\left(\frac{q_{i0}}{q_{i1}} + \gamma^2 \frac{\sigma^2_{i0}}{\sigma^2_{i1}}\right) - \frac{q_{i0}+q_{i1}}{2} \log\left(\frac{q_{i0}}{q_{i1}} + 1\right) - q_{i1} \log \gamma - q_{i1} \log \frac{\sigma_{i0}}{\sigma_{i1}}.
\end{aligned}$$

Putting it all together

$$\begin{aligned}
\mathcal{L}^*_\gamma &= \sum_{(i,a)} q_{ia} \left(\frac{(\mu^*_{ia}(\gamma) - \mu_{ia})^2}{2\sigma^2_{ia}} + \frac{\sigma^*_{ia}(\gamma)^2 - \sigma^2_{ia}}{2\sigma^2_{ia}} + \log \frac{\sigma_{ia}}{\sigma^*_{ia}(\gamma)} \right) \\
&= \frac{1}{2} \frac{(\gamma(\mu_{00}-\mu_{10}) - (\mu_{01}-\mu_{11}))^2}{\frac{\sigma^2_{01}}{q_{01}} + \frac{\sigma^2_{11}}{q_{11}} + \gamma^2\left(\frac{\sigma^2_{00}}{q_{00}} + \frac{\sigma^2_{10}}{q_{10}}\right)} + \sum_{i=0}^{1} \frac{q_{i0}+q_{i1}}{2} \log\left(\frac{q_{i0}}{q_{i1}} + \gamma^2 \frac{\sigma^2_{i0}}{\sigma^2_{i1}}\right) \\
&\qquad - \frac{q_{i0}+q_{i1}}{2} \log\left(\frac{q_{i0}}{q_{i1}} + 1\right) - q_{i1} \log \gamma - q_{i1} \log \frac{\sigma_{i0}}{\sigma_{i1}}.
\end{aligned} \tag{5.1}$$

Minimizing $\mathcal{L}^*_\gamma$ leads to a non convex program. Since γ is the ratio between variances of the new subgroup distribution, for a practical solution, we can do a line search over $\gamma \in (0, B)$ for some $B < \infty$. □

Proof. (Proof of Proposition 5.4) For the sake of this proof, we assume a countable data domain. Using the definition of the expected $0-1$ loss, we can write:

$$\begin{aligned}
&\mathrm{err}(\tilde{h}, D) - \mathrm{err}(\tilde{h}, \tilde{D}) \\
&= \sum_{(x,y,a)} \mathbb{I}\left(\tilde{h}(x,a) \neq y\right) \cdot (p(x,y,a) - \tilde{p}(x,y,a)) \\
&\leqslant 2 d_{TV}(\tilde{D}, D) \leqslant \sqrt{2 D_{KL}(\tilde{D}, D)} \qquad \text{(Pinsker's Inequality [Can23])}.
\end{aligned}$$

The first inequality follows from writing the error as expected 0-1 loss and using the definition of total variation distance. The last line follows from Pinsker's inequality [Can23]. We can similarly prove the other direction to obtain the first inequality. Similarly, for the true positive rate of group a, $\mathrm{TPR}_a(\tilde{h}, D) - \mathrm{TPR}_a(\tilde{h}, \tilde{D}) \leqslant 2 D_{TV}(\tilde{D}, D)$.

We can also write for the other group $A = a'$, $\mathrm{TPR}_{a'}(\tilde{h}, \tilde{D}) - \mathrm{TPR}_{a'}(\tilde{h}, D) \leqslant 2 D_{TV}(\tilde{D}, D)$. Adding both LHS and RHS and repeating the exercise in the other direction, noting that the TPR difference of $\tilde{h}$ in $\tilde{D}$ is zero (since in the ideal distribution we have exact fairness), we can bound the absolute value of

the TPR difference, which is our definition of Δ_{EO}, we get:

$$\Delta_{EO}(\tilde{h}, D) \leqslant \sqrt{8 D_{KL}\left(\tilde{D} \| D\right)}$$

□

Equalizing the first moment A popular intervention in the fairness literature is to equalize the first moment of the two sensitive groups or the mean outcomes of two groups, also known as the Calders-Verwer gap [CV10; Kam+12; Che+19]. We, therefore, also study an intervention where we only change the mean of the underprivileged group and try to match it with the mean of the privileged group. We can show that the resulting optimization program is convex. We leverage this intervention in Section 5 of the main text.

Proposition 5.5. (Affirmative Action by Equalizing First Moments) Let (X, Y, A) denote the features, binary class label, and binary group membership, respectively, of a random data point from any data distribution D with $q_{ia} = \Pr(Y = i, A = a)$, for $i \in \{0,1\}$ and $a \in \{0,1\}$. Let $X|Y = i, A = a \sim \mathcal{N}(\mu_{ia}, \sigma_{ia}^2)$ be a univariate Normal distribution, for $i \in \{0,1\}$ and $a \in \{0,1\}$. Then in the case of Affirmative mean change, we impose the following constraints:

$$\frac{q_{10}\,\tilde{\mu}_{10}}{q_{10}+q_{00}} + \frac{q_{00}\,\tilde{\mu}_{00}}{q_{10}+q_{00}} = \frac{q_{11}\,\mu_{11}}{q_{11}+q_{01}} + \frac{q_{01}\,\mu_{01}}{q_{11}+q_{01}},$$

we can efficiently minimize $D_{KL}\left(\tilde{D} \| D\right)$ as a function of the variables $\tilde{\mu}_{i0}$ and $\tilde{\Sigma}_{i0}$.

Proof. We are dealing with the following optimization problem:

$$D_{KL}\left(\tilde{D} \| D\right) = \sum_{i=0}^{1} q_{i0} D_{KL}\left(\tilde{P}_{i0} \| P_{i0}\right) = \sum_{i=0}^{1} q_{i0} \left(\frac{(\tilde{\mu}_{i0} - \mu_{i0})^2}{2\sigma_{i0}^2} + \frac{\tilde{\sigma}_{i0}^2 - \sigma_{i0}^2}{2\sigma_{i0}^2} + \log \frac{\sigma_{i0}}{\tilde{\sigma}_{i0}} \right)$$

as a function of the variables $\tilde{\mu}_{i0}$ and $\tilde{\sigma}_{i0}$ subject to the constraints

$$\frac{q_{10}}{q_{10}+q_{00}}\tilde{\mu}_{10} + \frac{q_{00}}{q_{10}+q_{00}}\tilde{\mu}_{00} = \frac{q_{11}}{q_{11}+q_{01}}\mu_{11} + \frac{q_{01}}{q_{11}+q_{01}}\mu_{01}$$

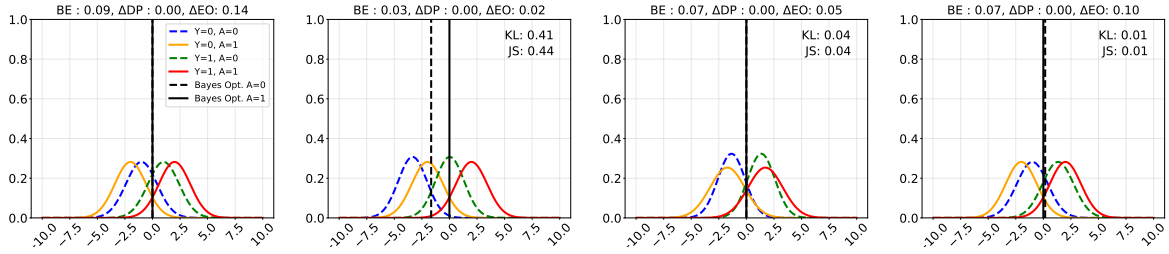
Since we are only changing the means and keeping the variances the same, the objective only depends on $\tilde{\mu}_{i0}$. Furthermore, let $K = (q_{10}+q_{00})/(q_{11}+q_{01}) \cdot (q_{11}\mu_{11} + q_{01}\mu_{01})$ so that

$$\mathcal{L} = \sum_{i=0}^{1} q_{i0} \frac{(\tilde{\mu}_{i0} - \mu_{i0})^2}{2\sigma_{i0}^2}, \quad \text{subject to } \tilde{\mu}_{00} = \frac{K - q_{10}\tilde{\mu}_{10}}{q_{00}}.$$

Substituting the constraint on $\tilde{\mu}_{00}$ in the objective $\mathcal{L}$ gives us a convex quadratic in $\tilde{\mu}_{10}$, and the solution is obtained by setting the derivative to zero:

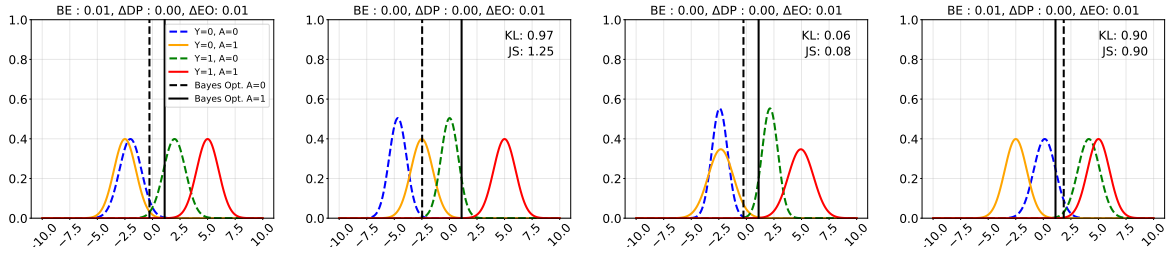
$$\tilde{\mu}_{00} = \frac{\frac{K}{\sigma_{10}^2 \cdot q_{10}} - \frac{\mu_{10}}{\sigma_{10}^2} + \frac{\mu_{00}}{\sigma_{00}^2}}{\frac{q_{00}}{\sigma_{10}^2 \cdot q_{10}} + \frac{1}{\sigma_{00}^2}}, \quad \tilde{\mu}_{10} = \frac{\frac{K}{\sigma_{00}^2 \cdot q_{00}} - \frac{\mu_{00}}{\sigma_{00}^2} + \frac{\mu_{10}}{\sigma_{10}^2}}{\frac{q_{10}}{\sigma_{00}^2 \cdot q_{00}} + \frac{1}{\sigma_{10}^2}}, \quad \tilde{\mu}_{01} = \mu_{01}, \quad \text{and} \quad \tilde{\mu}_{11} = \mu_{11}.$$

□



(a) Original Distribution **(b)** EF-Affirmative **(c)** EF-All Subgroups **(d)** Mean Matching

Figure 5.5: Comparison of Different Interventions when the subgroup distributions are shifted version of each other. While all methods achieve the same Bayes Error, Affirmative action is able to bring down the Bayes Error and achieve exact fairness.

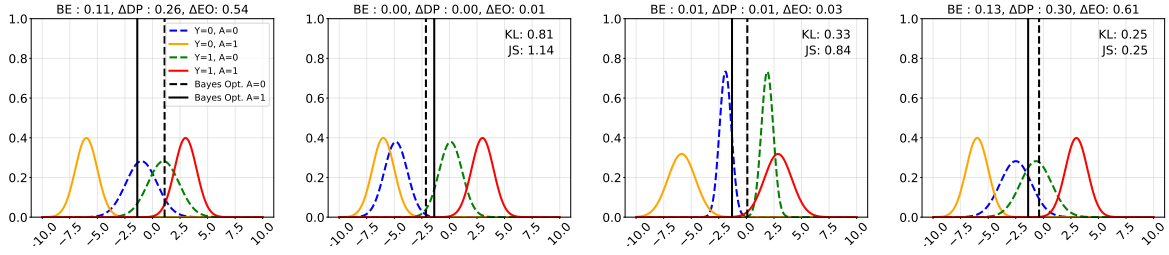


(a) Original Distribution **(b)** EF-Affirmative **(c)** EF-All Subgroups **(d)** Mean Matching

Figure 5.6: Comparison of Different Interventions when the original distribution is already fair. In this case, EF-All ensures that it stays close to the true distribution, as no intervention as required, while others relatively deviate.

§ 5.8.3 Additional Figures for Section 5.5

In this section, we lay out additional plots from our Gaussian case study. We first describe the setup. We modify a stylized setting of Gaussian distributions from previous work (see Definition 3.1 in [PCDG18], Section 5.3 in [Bak+21]) to investigate the unfairness and the Bayes optimal error on the original and ideal distributions obtained through various interventions. We fix $q_{ia} \in (0,1)$ such that $q_{00} + q_{10} + q_{01} + q_{11} = 1$, and our data generation works as follows. We simulate a data distribution where $Y = i, A = a$ with probability q_{ia} and $X \mid Y = i, A = a$ is sampled from a univariate Gaussian $\mathcal{N}(\mu_{ia}, \sigma^2_{ia})$. We choose homoskedastic Gaussians within each group $A = a$, i.e., $\sigma_{0a} = \sigma_{1a}$, so the we can show the Bayes optimal classifier boundary as a threshold. We choose different σ_{ia}'s that cover ground truth distribution that can the entire spectrum of being *ideal* or close to *ideal* to very far, and then we



(a) Original Distribution **(b)** EF-Affirmative **(c)** EF-All Subgroups **(d)** Mean Matching

Figure 5.7: Comparison of Different Interventions when we use a different threshold (3/4) than the Bayes optimal threshold (1/2). As derived in Proposition 5.2, the EF-Affirmative and EF-All interventions work with any threshold.

apply different interventions to change all or some subset of μ_{ia}'s and σ_{ia}'s to find the nearest *ideal* distribution in KL-divergence as given in Section 4 of the main text.

We first look at a case where the subgroup distributions are the same shifted versions of each other in Figure 5.5. Note that all interventions, in this case, result in the same Bayes error (BE), but affirmative action brings the BE down with zero unfairness at the cost of incurring a deviation in terms of KL and JS divergence. However, in the next subplot, changing all four subgroups not only helps reduce the Bayes error and unfairness but also stays very close to the true distribution in the KL/JS sense. Matching the means also helps reduce the unfairness while staying close to the true distribution, but is sub-optimal compared to the EF-Affirmative and EF-All interventions.

Next, we look at a case where the Bayes optimal classifier is already fair (ΔEO is close to 0 while ΔDP=0) in Figure 5.6. The expected solution here should be that any intervention must leave the distribution as it is. EF-Affirmative intervention keeps the unfairness and error rate numbers as it is, but deviates from the true distribution, as indicated by the KL/JS divergences. However, the EF-All intervention only makes major changes to variances and stays close to the true distribution. The Mean Matching intervention shifts both the under-privileged subgroups and strays away from the true distribution, as indicated by relatively high KL/JS values.

Finally, in light of Proposition 5.2, we simulate the cost-sensitive risk for a different cost matrix C other than 0-1 loss by considering a threshold $t_C = 3/4$ on $\eta(x, a)$ in Figure 5.7. The original distribution has high unfairness. EF-Affirmative intervention manages to achieve almost perfect fairness and zero error rate, but incurs relatively high KL/JS numbers. However, once again, changing all four subgroups, results in a solution that is perfectly fair and accurate, with low KL/JS. Mean Matching is unable to address the fairness-accuracy tension at all in this case and also manages to drift away from the true distribution, as indicated by non-zero KL/JS values.

§ 5.8.4 Additional Results and Details for Section 5.6

In this section, we lay out all the details for the experiments performed for LLM steering. The code to reproduce our results is provided *here*. For the multi-class experiments, we use a lot of helper functions from the code of Singh et al. [Sin+24] [1]. For the emotion steering experiments, we reproduce the methodology from Zhao et al. [Zha+24] and provide the Jupyter notebook in our code.

§ 5.8.4.1 Reducing Disparity in Multi-class Classification

To apply our intervention to multi-class settings, we first develop a version of Theorem 5.1 for multiple classes. We show this for a univariate distribution, and, for our intervention, we assume a diagonal covariance matrix. Since our experiment setup requires only two groups per class, we present the effective constraints assuming two sensitive groups; this methodology can be readily extended to handle a countable number of groups as well. To make our program convex (and affirmative), we fix a class $y \in \mathcal{Y}$. We fix our class y^* according to the following heuristic: $y^* = \arg\min_{y \in \mathcal{Y}} \Delta_y \mathsf{TPR}(\hat{h})$, where $\hat{h}$ is the empirical risk minimizer on the given data. This fixes our ratio $\gamma_\sigma = \frac{\sigma_{y^*1}}{\sigma_{y^*0}}$ and $\gamma_q = \frac{q_{y^*1}}{q_{y^*0}}$.

We can now write a multi-class version of the optimization program in Theorem 5.1:

$$\mathcal{L}_\gamma = \sum_{(i,a)} q_{ia} \left(\frac{(\tilde{\mu}_{ia} - \mu_{ia})^2}{2\sigma_{ia}^2} + \frac{\tilde{\sigma}_{ia}^2 - \sigma_{ia}^2}{2\sigma_{ia}^2} + \log \frac{\sigma_{ia}}{\tilde{\sigma}_{ia}} \right)$$

[1] https://github.com/shauli-ravfogel/affine-steering

as a function of the variables $\tilde{\mu}_{ia}$ and $\tilde{\sigma}_{ia}$ subject to the following constraints

$$\frac{\tilde{\mu}_{i1}-\tilde{\mu}_{j1}}{\tilde{\mu}_{i0}-\tilde{\mu}_{j0}}=\frac{\tilde{\sigma}_{i1}}{\tilde{\sigma}_{i0}}=\frac{\tilde{\sigma}_{j1}}{\tilde{\sigma}_{j0}}=\gamma_\sigma\,,\frac{q_{i1}}{q_{i0}}=\frac{q_{j1}}{q_{j0}}=\gamma_q \quad \text{and} \quad \tilde{\sigma}_{ia}\geqslant 0,\ \text{for all } i\in y, j\in y\setminus\{i\}.$$

Just like in the proof of Theorem 5.1, the resulting program will result in separable objectives for a class y and then in the underlying optimization variables $\tilde{\mu}_{ya}$ and $\tilde{\sigma}_{ya}$:

Program for $\tilde{\mu}_{ya}$:

$$q_{y0}\frac{(\tilde{\mu}_{y0}-\mu_{y0})^2}{2\sigma_{y0}^2}+q_{y1}\frac{(\tilde{\mu}_{y1}-\mu_{y1})^2}{2\sigma_{y1}^2},\ \text{subject to } \tilde{\mu}_{y1}-\tilde{\mu}_{y^*1}=\gamma(\tilde{\mu}_{y0}-\tilde{\mu}_{y^*0})$$

Program for $\tilde{\sigma}_{ya}$:

$$q_{y0}\left(\frac{\tilde{\sigma}_{y0}^2-\sigma_{y0}^2}{2\sigma_{y0}^2}+\log\frac{\sigma_{y0}}{\tilde{\sigma}_{y0}}\right)+q_{y1}\left(\frac{\tilde{\sigma}_{y1}^2-\sigma_{y1}^2}{2\sigma_{y1}^2}+\log\frac{\sigma_{y1}}{\tilde{\sigma}_{y1}}\right),\ \text{subject to } \tilde{\sigma}_{y1}=\gamma\tilde{\sigma}_{y0},$$

where y^* is the class we fixed earlier and $\gamma=\gamma_\sigma$. The solution for the following programs are the following:

$$\tilde{\sigma}_{y0}=\pm\sqrt{\frac{q_{y0}+q_{y1}}{\frac{q_{y0}}{\sigma_{y0}^2}+\frac{\gamma^2 q_{y1}}{\sigma_{y1}^2}}},\ \tilde{\sigma}_{y1}=\pm\gamma\sqrt{\frac{q_{y0}+q_{y1}}{\frac{q_{y0}}{\sigma_{y0}^2}+\frac{\gamma^2 q_{y1}}{\sigma_{y1}^2}}},\tilde{\mu}_{y0}=\frac{\frac{q_{y0}}{\sigma_{y0}^2}\mu_{y0}+\frac{\gamma q_{y1}}{\sigma_{y1}^2}(\mu_{y1}-\mu_{y^*1}+\gamma\mu_{y^*0})}{\frac{q_{y0}}{\sigma_{y0}^2}+\frac{\gamma^2 q_{y1}}{\sigma_{y1}^2}},\ \text{and}\ \tilde{\mu}_{y1}=\frac{\frac{q_{y0}}{\sigma_{y0}^2}(\mu_{y^*1}-\gamma\mu_{y^*0}+\gamma\mu_{y0})+\frac{\gamma^2 q_{y1}}{\sigma_{y1}^2}\mu_{y1}}{\frac{q_{y0}}{\sigma_{y0}^2}+\frac{\gamma^2 q_{y1}}{\sigma_{y1}^2}}.$$

Once we have the corrected distributions $\mathcal{N}(\tilde{\mu}_{ia},\tilde{\Sigma}_{ia})$, we set up an affine intervention, following the design choices of Singh et al. [Sin+24]. We assume an affine relationship between the original and transformed samples per subgroup: $Y=a_{ya}X+b_{ya}$, where $Y\sim\mathcal{N}(\tilde{\mu}_{ya},\tilde{\sigma}_{ya})$ and $X\sim\mathcal{N}(\mu_{ya},\sigma_{ya})$. Taking expectation on both sides gives us: $\tilde{\mu}_{ya}=a_{ya}\mu_{ya}+b_{ya},\tilde{\sigma}_{ya}^2=a_{ya}^2\sigma_{ya}^2$, and we get the following coefficients: $a_{ya}=\pm\frac{\tilde{\sigma}_{ya}}{\sigma_{ya}}$ and $b_{ya}=\tilde{\mu}_{ya}\pm\frac{\tilde{\sigma}_{ya}}{\sigma_{ya}}\mu_{ya}$.

We have two choices of the parameters corresponding to the positive and negative solutions. Since we are working with empirical estimates, we use the validation error to decide the best set of estimates. A detailed implementation is given in the code. To implement the conditions for q_{ya}, we use the reweighing scheme of Kamiran and Calders [KC12]. The plot in the main text (Figure 3) assumes $q_{y0}=q_{y1}$ for the corrected covariances. Figure 5.8 shows the plot with no such assumption for q_{ya} and confirms with same trends as observed in Figure 5.3.

§ 5.8.4.2 Steering Activations for Joyful Generation

Zhao et al. [Zha+24] propose to obtain a distribution over the steering vectors for a concept instead of a single steering vector. In this section, we lay out all our prompts and the design choices for the emotion steering pipeline.

We first generate training data for each concept (joyful, angry) for each of the groups (horror, comedy), resulting in four subgroups. The following prompt was used to generate an initial set of data:

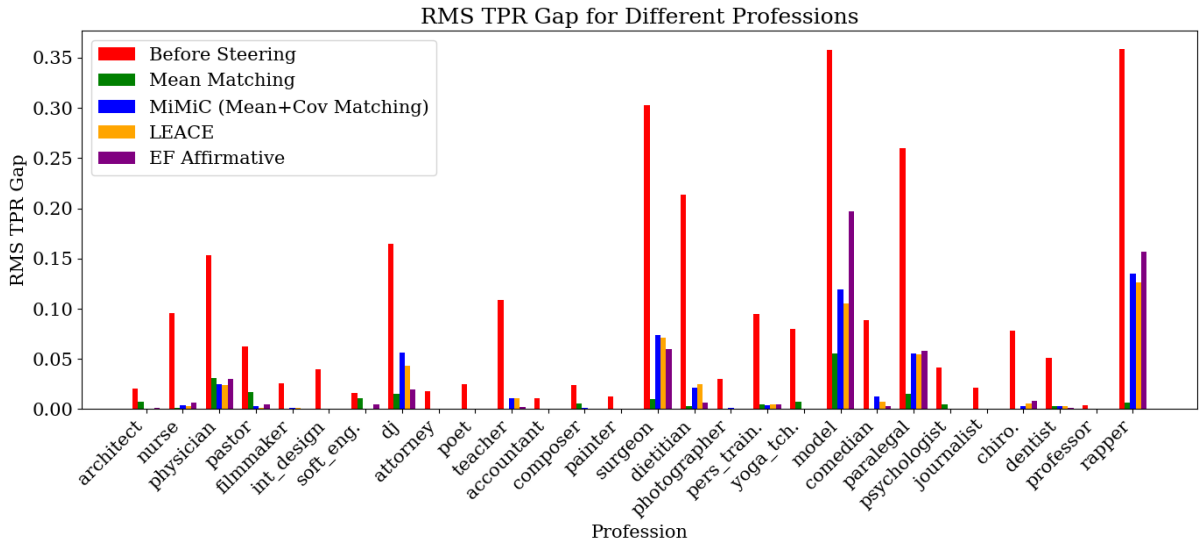


Figure 5.8: TPR-gap between Gender groups for all professions. All methods to steer feature representations achieve roughly the same accuracy (in the range of 0.77-0.79). Our intervention (EF Affirmative) is able to significantly reduce the TPR-gap for all professions. In many cases, it is even comparable or better than previous interventions Belrose et al. [Bel+23b] and Singh et al. [Sin+24].

Prompt to generate 1000 Comedy movie reviews that are joyful.

Compose a concise 30-word movie review, assuming it is a comedy movie, that covers these four aspects: plot, sound and music, cultural impact, and emotional resonance. Choose a joyful tone for your review. For the plot, comment on its structure or originality. Regarding sound and music, mention how it enhances the storytelling. For cultural impact, touch on any relevant social commentary. Finally, describe how the film resonates emotionally. Ensure your joyful tone is consistent throughout the review. Please include emotions like "joyful" in these texts and generate 1000 samples.

We obtain the last token embeddings from each layer of a Llama-3.1 8B model [Gra+24] for each of the samples. We now proceed towards obtaining the steering vectors. We want to obtain a steering vector for each group. Treating angry reviews as an irrelevant sample for the 'joyful' concept, we assign $y = 0$ to angry samples and $y = 1$ to joyful samples. Because we want to estimate a distribution over the steering vectors instead of a single vector, we sample 300 points with replacement and repeat this for 50 iterations. In each of these iterations, we train a Logistic Regression model to classify between the relevant and the irrelevant samples. We get 50 weight vectors using this pipeline, and we use those to obtain the sample mean and covariance. We denote the resulting distribution for the steering vectors for layer l as $\mathcal{N}(\mu^l_{1a}, \Sigma^l_{1a})$ where $1a$ denotes that this distribution represents the steering vector for joyful emotion for a group $A = a$ (horror or comedy reviews). To apply our intervention later, we also obtain the steering vector in the other direction by flipping the relevance labels, i.e. joyful $\rightarrow$ angry, and we denote the resulting Gaussian distribution with $\mathcal{N}(\mu^l_{0a}, \Sigma^l_{0a})$.

To perform steering, we now sample a steering vector $v^l_c \sim \mathcal{N}(\mu^l_{1a}, \Sigma^l_{1a})$ and add it to the last token representation of layer l with strength a: $h^l = (1-a)h^l + av^l_c$. To measure the performance of steering, we ask the Llama model to generate angry reviews using the following prompt:

Prompt to evaluate the joyfulness of generated reviews (for group horror).

System Prompt: You are a well informed movie critic.
User Prompt: Compose a concise 10-word movie review, assuming it is a horror movie. The tone of the review must be angry. The review should be within 10 words. Generate 100 samples. Please make sure that you do not repeat the same review and keep each review under 10 words. Also, make sure that the reviews are coherent and not repetitive.

We first measure the joyfulness score of the generated reviews without adding any steering vectors. Just like Zhao et al. [Zha+24], we use the following prompt to ask the GPT 4.1 model to rate the joyfulness of the text:

Prompt to generate angry 10-word reviews for the horror genre.

System Prompt: You are a deterministic evaluation agent trained to assess text samples using two strictly defined criteria: (1) emotional tone (joyful score) and (2) linguistic structure (chaotic score). For each input, your task is to produce exactly two integer scores between 0 and 10. These must follow the user's scoring rubric precisely. The first score reflects emotional positivity or dark exuberance without considering sentence structure. The second score reflects linguistic disorder or repetition, without considering sentiment.
You must produce output in CSV format, strictly matching the following structure: "sample", joyful score, chaotic score
Do not include any additional content, commentary, or variation. Your outputs must be fully deterministic: identical input always produces identical output.
Your responses should match the tone and evaluation logic demonstrated in the provided reference examples.
User Prompt: You are an automated evaluator tasked with rating horror movie reviews using two fixed criteria: **joyful expression** and **textual coherence and novelty**.
For each review, you must generate exactly two integer scores as described below:
—
1. Joyful Score
Rate from 0 to 10 based strictly on emotional tone:
* **0** = not joyful at all (strongly negative, angry, or emotionally flat tone)
* **10** = extremely joyful, enthusiastic, or darkly exuberant in tone
* Intermediate values reflect a mix of enthusiasm and criticism
* Focus solely on affect and sentiment; ignore linguistic structure for this score

—
2. Chaotic Score
Rate from 0 to 10 based strictly on linguistic organisation:
* **0** = fully coherent, fluent, and structurally sound
* **10** = highly chaotic, incoherent, repetitive, or nonsensical
* Intermediate values reflect partial disorganisation, verbosity, or broken phrasing
* Focus only on syntax, flow, and novelty of expression; ignore emotional content

—
Reference Examples
Joyful and Coherent Horror Reviews:
* “Blood-soaked fun ensues in this delightfully terrifying slasher film.”
* “Chilling thrills abound in this creepy haunted mansion tale.”
* “Jump scares galore in this electrifying horror comedy gem.”
* “Unsettling unease fills this unnerving psychological horror masterpiece.”
* “Bone-chilling chills chill to the bone in this one.”

Angry and Coherent Horror Reviews:
* “Abysmal plot twists ruined what could’ve been a decent film.”
* “Mind-numbing terror fails to deliver in this lazy horror.”
* “Weak jump scares can’t save this trainwreck disaster.”
* “Poor production values ruin what little suspense exists.”
* “Frustratingly predictable, making it boring and unscary too.”

—
Output Format
* For each sample, return one line in strict CSV format: ‘"sample", joyful score, chaotic score‘
* **Example Output:**
“‘ sample, joyful score, chaotic score
“This horror film was painfully dull and predictable.”, joyful_score_1, chaotic_score_1
“Terrifying, stylish, and packed with chilling moments!”, joyful_score_2, chaotic_score_2
“‘
* Do **not** include explanations, commentary, or additional formatting.
* Output must be **fully deterministic**: the same input must always yield the same scores.
Begin processing the dataset now. Here is the batch of review samples:

A few notes on evaluation are in order. Zhao et al. [Zha+24] report both joyfulness and coherence scores for the generated text. However, we observed that coherence scores were all over the place and did not make sense. Second, Zhao et al. used the GPT-4o model, whereas we used the GPT 4.1 model because we observed that the joyful scores correlated more with the qualitative inspection of the generated samples.

Following the above pipeline, we observe an increase in the joyfulness scores of the reviews generated by a Llama model after steering. However, since the effectiveness of joyful steering was not the same across the horror and comedy movie review generations, we apply our affirmative intervention (Theorem 5.1) by assuming that the horror and comedy movie reviews define two groups. Let the modified steering vector be denoted by $\tilde{v}_c^l \sim \mathcal{N}(\tilde{\mu}_{1a}^l, \tilde{\Sigma}_{1a}^l)$, where the new gaussian distribution is obtained after applying the affirmative action intervention from Theorem 5.1 assuming horror group is the under-privileged group.

However, simply replacing v_c^l will not work. We demonstrate that empirically in the main text, where in Figure 5.4, $\alpha = 1$ corresponds to using $\tilde{v}_c^l$ instead of v_c^l. But we can always use $\tilde{v}_c^l$ to nudge the existing steering vector v_c^l in the right direction. To do that, we modify the steering vector and the representation h^l with the following rule: $h^l = (1-a)h^l + a((1-\alpha)v_c^l + \alpha\tilde{v}_c^l)$, where α controls the

strength of mixing the old and new steering vectors. In Figure 4, we show that for small values of α, the steering vector indeed improves at steering the reviews of the horror group towards a more joyful tone.

Chapter 6

Practical Estimation of the Optimal Fairness-Accuracy Trade-off with Soft Labels

Abstract

There is a fundamental limit to the best prediction error of any model for a given data distribution. In classification, the Bayes error quantifies this limit and serves as both a benchmark for trained models and a criterion for detecting overfitting. The state-of-the-art Bayes error estimation techniques for binary classification use only soft labels (positive class probabilities) and require no input features or auxiliary training. We can extend such notions to constrained classification problems, such as fairness. Fair classification is an important case of constrained or multi-objective classification. It requires a model not only to be accurate but also to yield fair outcomes (e.g., equal or near-equal true positive rates) across different demographic groups (e.g., race and gender). This leads to an inherent fairness-accuracy tradeoff or the Pareto frontier that is used to compare fair classifiers against each other.

The focus of this chapter is to estimate the fundamental limit for this fairness-accuracy tradeoff. We derive an explicit expression for the fair Bayes error rate (i.e., best achievable error rate subject to given fairness constraints), and construct consistent, instance-free estimators for it using only soft labels and group information. We provably bound the bias and the sample complexity of our estimators. Empirically, our method produces a robust, low-variance estimate of the optimal fairness-accuracy trade-off curve, even when the soft labels are noisy and derived from a classification model.

§ 6.1 Introduction

We consider a supervised binary classification with a sensitive attribute. A central object in our analysis is the group-conditional posterior (or regression) function, $\eta(x, a) := \Pr(Y = 1 \mid X = x, A = a)$. We refer to $\eta(x_i, a_i)$ as the *soft label* of an instance (x_i, a_i), and write $c_i := \eta(x_i, a_i)$. Finally, we let $C := \eta(X, A)$ denote the corresponding random variable.

The notion of Bayes optimality in machine learning defines the best achievable performance, given a data distribution, typically with respect to the 0-1 loss [CH67; Fuk13; DGL13; MRT12]. For binary classification, a well-known result states that the Bayes optimal classifier g^* can be expressed as the following decision rule: $g^*(x, a) = \mathbb{I}(\eta(x, a) \geqslant 1/2)$ [MRT12; Zen+26]. This rule also allows us to estimate the Bayes error, the best achievable performance for a given distribution, without access to any instances [MRT12; Ish+]: $\beta = \mathop{\mathrm{E}}_{X,A}[\min\{\eta(X, A), 1 - \eta(X, A)\}]$, where β denotes the Bayes error.

Fair Bayes optimal classifiers can often be characterized as adjustments of the classification thresholds: $g^*_{\text{fair}}(x, a) = \mathbb{I}\left(\eta(x, a) \geqslant 1/2 + t_a(\delta)\right)$ where $t_a(\delta)$ is a group-dependent adjustment that depends on the fairness metric, the underlying data distribution and the tolerance for unfairness δ (Propositions 2.2, 2.3). This is precisely where existing works on estimating the Bayes error do not generalize [Ish+; JCD24; UIS25]. Unlike Bayes error estimation, where the constant threshold of $1/2$ allows Bayes error estimators to rely solely on posterior probabilities (soft scores) for any cost-sensitive threshold, such as fairness, the error estimator also requires an estimator for the underlying thresholds.

Therefore, to estimate the optimal fairness-accuracy tradeoff, we first show how we can disentangle the unfairness level δ and the corresponding threshold adjustment $t_a(\delta)$ by studying when the classifiers become trivial (Lemma 6.3). Using this, we can estimate the threshold and, eventually, the unfairness and error rate at several points along the tradeoff curve by leveraging its monotonicity properties (Lemma 6.2). We show that fair classification can be reduced to cost-sensitive thresholding with appropriate λ-dependent thresholds (Lemma 2.1), where λ is a parameter that directly controls the correction to the classification threshold and depends on δ. We then derive consistent estimators for the fair classification thresholds, only using soft labels, at any fixed λ (Theorem 6.1).

Once we have consistent estimates of thresholds, we then, as an independent result, generalize the Bayes error estimators used in Ishida et al. [Ish+], Jeong et al. [JCD24], and Ushio et al. [UIS25] to work with any cost-sensitive risk (Theorem 6.2). Since our thresholds and estimators are both constructed using a finite number of soft labels, we move beyond the scope of the results of Ishida et al. [Ish+], Jeong et al. [JCD24], and Ushio et al. [UIS25] and study consistency, bias, and sample complexity of the estimation of the cost-sensitive Bayes risk (Theorem 6.3). These results and rates are applicable to many cost-sensitive learning scenarios [Koy+14; SK22] that require distribution-dependent classification thresholds.

We then focus on estimating the optimal fairness-accuracy tradeoff using only soft labels. We first derive the estimated risk at any threshold and show its consistency properties (Theorem 6.4). We then also provide estimators for the expected unfairness under demographic parity and equal opportunity, and derive their consistency and sample complexity properties (Theorem 6.5). All of them culminate in a bisection-search driven algorithm with our estimators (Algorithm 6.1), which is very similar to one that returns the fair Bayes optimal classifiers [Zen+26].

Finally, we estimate the fairness-accuracy tradeoff on real-world datasets. We compare our estimated tradeoff against several recent works on estimating the fairness-accuracy tradeoff [Del+24; KPM24] and differentiable objectives for fair classification [BDDB23], in Figure 6.1. We show that our estimated tradeoff consistently lower bounds the tradeoff estimated by these methods, and whenever it does not (with the algorithm from Kozdoba et al. [KPM24]), it is explained by the algorithm's optimization for a completely different lower bound on the Pareto frontier. We also perform experiments with noisy soft scores generated by a Logistic Regression model on the Adult dataset [Lic+13] and find that our method still lower-bounds all other fairness tradeoff algorithms.

Overall, in a similar spirit to Ishida et al. [Ish+], where the authors use the Bayes error as a tool to validate performance and check for overfitting, our estimated tradeoff with soft label establishes a similar precedent: any current or future methods aiming to compare fair algorithms should always look to close the gap between their estimated tradeoff and the theoretically best achievable, that we aim to estimate.

Related Work: Our work is inspired by the prior results [Ish+; JCD24; UIS25] that characterize the performance limitations in terms of accuracy using the threshold characterization of the Bayes optimal classifier. There are several other works on Bayes error estimation that aim to estimate the Bayes error

indirectly [Ber+15; NXH19; The+21; Ren+21]. But since most of these methods estimate the Bayes error indirectly, without leveraging soft labels or thresholding rules, generalizing them to constrained optimization settings, such as fairness constraints and cost-sensitive risk, is future work beyond the scope of this work.

We make extensive use of threshold characterizations of fair classifiers, which have been studied extensively in the literature [MW18; Chz+19; Zen+26]. We also draw inspiration from existing works on deriving the sample complexity of plug-in estimators and related estimators for fair classifiers [Khu+21; JCD24].

§ 6.2 Preliminaries

We first introduce some additional notations. Let $\pi_a = \Pr(A = a)$, $\pi_{ya} = \Pr(Y = y, A = a)$, $\pi_{y|a} = \Pr(Y = y|A = a)$, and $\pi = \Pr(Y = 1)$. By the law of total probability, $\Pr(Y = y) = \sum_{a \in \mathcal{A}} \pi_{ya}$. We assume the existence of all groups and classes, i.e., $\pi_a > 0, \pi_{y|a} > 0$ for all $y \in \mathcal{Y}$ and for all $a \in \mathcal{A}$. Finally, we let $\xrightarrow{p}$ denote convergence in probability. As stated earlier, we work within the setting of binary classification.

Whenever we work with finite sample approximations, we assume that we are given access to N i.i.d. soft labels of the form (c_i, a_i), and we will denote our dataset with $D = \{(c_i, a_i)\}_{i=1}^N$. Since our approach is model-free, we do not require access to features x_i. We will also partition our dataset D into two parts: $D = D_\alpha \cup D_\beta$, with $D_\alpha \cap D_\beta = \varnothing$, where D_α (with N_α samples) will be the data used to estimate the thresholds that we will put on $\eta(x, a)$ (defined later), and D_β (with N_β samples) will be used to construct the estimators of classification risk and unfairness, with $N = N_\alpha + N_\beta$. The splits are chosen independently of the sample values. Furthermore, we will denote with $N_{\alpha,a}$ (or $N_{\beta,a}$) the number of samples from group $A = a$. Lastly, $i \in [N]$ denotes drawing an index from the dataset D of size N. We need two independent sample sets because we estimate the required quantities in two steps. First, using D_α, we estimate the distribution priors and group-dependent thresholds (Theorem 6.1), and next using D_β, we estimate the cost-sensitive Bayes risk, the expected error, and unfairness (Theorems 6.3, 6.4, and 6.5 respectively). Our soft labels also follow Assumption 2.1.

For later results, we would also need a bounded density assumption on our group-conditioned soft labels.

Assumption 6.1. (Lipschitz continuous group CDF) For each $a \in \mathcal{A}$, the group-conditional distribution of $C = \eta(X, a)|A = a$ admits a density f_a on $(0, 1)$ such that $\sup_{t \in (0,1)} f_a(t) \leqslant L_a < \infty$. In particular, F_a is L_a-Lipschitz: $|F_a(u) - F_a(v)| \leqslant L_a \cdot |u - v|$ for all $u, v \in (0, 1)$, where F_a is the CDF of η from Assumption 2.1.

Let $\hat{\pi}_a = \frac{1}{N_\alpha} \sum_{i \in [N_\alpha]} \mathbb{I}(a_i = a)$ denote our estimator for π_a. By the Weak Law of Large Numbers, $\hat{\pi}_a \xrightarrow{p} \pi_a$. Furthermore, the bias of the estimator $\hat{\pi}_a$ is zero, which is easy to see via the linearity of expectation. We can derive a similar estimator for $\pi = \Pr(Y = 1)$ as $\hat{\pi} = \frac{1}{N_\alpha} \sum_{i \in [N_\alpha]} c_i$.

Similarly, let $\hat{\pi}_{1a} = \frac{1}{N_\alpha} \sum_{i \in [N_\alpha]} \mathbb{I}(a_i = a)\, c_i$ denote our estimator for π_{1a}. By the Weak Law of Large Numbers, $\hat{\pi}_{1a} \xrightarrow{p} \pi_{1a}$. Furthermore, the bias of the estimator $\hat{\pi}_{1a}$ is zero, which is easy to see via the linearity of expectation. The same analysis works out for the unbiased estimator $\hat{\pi}_{1|a} = \frac{1}{N_{\alpha,a}} \sum_{i \in [N_\alpha]} \mathbb{I}(a_i = a)\, c_i$ and its expected value $\pi_{1|a}$. Since we will work with functions of $\hat{\pi}_a$ (and other estimators) that may have inverses of $\hat{\pi}_a$, we also define another estimator $\hat{\pi}_a^+ = \max\{\hat{\pi}_a, \frac{1}{N_\alpha}\}$, which is still statistically consistent, but a biased estimator (same for $\hat{\pi}_{1a}^+$ and $\hat{\pi}_{1|a}^+$).

Finally, let $\Pi_{[0,1]}(t) := \min\{1, \max\{0, t\}\}$ denote the truncation operation, which we will frequently

use to ensure our estimators are stable and bounded between 0 and 1.

§ 6.2.1 Fair Bayes Optimal Classifiers

In this section, we will revisit some of the prior results on fair Bayes optimal classification. Zeng et al. [Zen+26] recently characterized the Bayes optimal δ-fair classifiers whenever the fairness metric can be expressed as a linear function of a classifier h. We introduce their definition below.

Definition 6.1. A fairness disparity measure $\Delta(h)$ is *linear* if there exists a $w : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ such that for all classifiers h, $\Delta(h) = \int_{\mathcal{A}} \int_{\mathcal{X}} h(x, a)\, w(x, a) d\mathbb{P}(x, a)$. Furthermore, it is *bilinear* if there exist distribution-dependent quantities s, b such that $w(x, a) = s \cdot \eta(x, a) + b$.

Remark 6.2.1. Proposition 3.4 in Zeng et al. [Zen+26] shows the suitable choices of s and b for the fairness constraints mentioned in Definitions 2.3 and 2.4. For demographic parity, $w_{DP}(x, a) = \frac{(2a-1)}{\pi_a}$ and for equal opportunity, $w_{EO}(x, a) = \frac{(2a-1)\eta(x,a)}{\pi_{1a}}$.

We will now define the risk of a classifier h and its disparity for linear disparity metrics, as described in Zeng et al. [Zen+26].

Definition 6.2. (Equation A.1 and 4.2 from [Zen+26]) For a classifier h, we can define the risk $R(h)$ and disparity $\Delta(h)$ as:

$$R(h) = \sum_{a \in \mathcal{A}} \pi_a \int_{\mathcal{X}} (1 - 2\eta(x, a)) h(x, a) d\mathbb{P}_{X|A=a}(x) + \sum_{a \in \mathcal{A}} \pi_a \int_{\mathcal{X}} \eta(x, a) d\mathbb{P}_{X|A=a}(x), \text{ and}$$

$$\Delta(h) = \sum_{a \in \mathcal{A}} \pi_a \int_{\mathcal{X}} [w(x, a) \cdot h(x, a)]\, d\mathbb{P}_{X|A=a}(x)$$

Because we are working with binary classification, introducing the class posterior probability allows us to decompose the risk into purely classifier-dependent and distribution-dependent components. We now introduce the optimal fair classification problem. For at most δ-unfairness, we want to solve the following optimization problem:

$$h_\delta \in \arg\min_h R(h) \text{ subject to } |\Delta(h)| \leqslant \delta. \tag{6.1}$$

For demographic parity and equal opportunity, we can precisely write down the fair Bayes optimal classifier, where the threshold depends on the distribution.

Lemma 6.1. (Proposition 2.3 in [Chz+19], Theorem 4.2 in [Zen+26]) Under Assumption 2.1, the Bayes optimal δ-fair classifier takes the following form: $h_\delta(x, a) = \mathbb{I}\left(\eta(x, a) \geqslant \frac{1}{2} + \frac{\lambda^*(\delta)}{2} w(x, a)\right)$, where $\lambda^*(\delta)$ is the optimal dual parameter for the optimization program in Equation 6.1.

Lemma 6.1 implies that for a fixed δ, there may exist a λ^* (that depends on whether for a given δ, finding a $\lambda^*(\delta)$ is feasible), with which we can adjust the unconstrained Bayes optimal threshold of $1/2$. This correction depends on which group $A = a$ is underprivileged, as well as on the distribution-dependent quantities like $\pi_a, \pi_{y|a}$.

Characterizing the precise value of λ^* as a function of δ and distribution-dependent quantities is a challenging problem, and several prior works (e.g., [MW18; Cel+19; Chz+19; Zen+26]) treat it as another optimization problem. However, in our approach, we will show how we can obtain a feasible

range for λ. By doing so, we will first be able to parametrize the risk and the unfairness in terms of λ, and then set up a bisection search to get arbitrarily close to a required level of unfairness δ. In a similar spirit to the Bayes error estimation work in the literature [Ish+; JCD24; UIS25], this will allow us to achieve the theoretically optimal fairness-accuracy trade-off using only soft labels.

Let $h_\lambda = \operatorname{argmin}_h R(h) + \lambda \cdot \Delta(h)$, for $\lambda \in \mathbb{R}$ denote the primal solution for a fixed λ. We can also dub this the λ-fair classifier, since this can potentially be $\delta(\lambda)$-fair classifier. Using this λ, we can also write down the Bayes optimal λ-fair classifier using the threshold characterization in Lemma 6.1, which will be useful later in this chapter.

We now introduce a key lemma from Zeng et al. [Zen+26] that enables us to study the fairness-accuracy Pareto frontier.

Lemma 6.2. (Proposition 4.1 from [Zen+26]) As a function of λ, $\Delta(h_\lambda)$ is monotone non increasing, i.e. $\lambda_1 < \lambda_2 \implies \Delta(h_{\lambda_1}) \geqslant \Delta(h_{\lambda_2})$. The risk $R(h_\lambda)$ is monotone non-increasing on $(-\infty, 0)$ and monotone non-decreasing on $(0, \infty)$.

Lemma 6.2 is incredibly helpful to devise an algorithm to estimate the tradeoff, since it allows us to parametrize the curve with λ, establish the convexity of the tradeoff curve (Proposition 4.6 in Zeng et al. [Zen+26]), and hence enables the use of bisection search to get close to any level of unfairness required. We will now define the Pareto frontier, a curve that characterizes optimal constrained classifiers at different levels.

Definition 6.3. Given a risk measure $R(\cdot)$ and a fairness disparity measure $\Delta(\cdot)$, a classifier h' Pareto-dominates another classifier h, if $R(h') < R(h)$ and $|\Delta(h')| < |\Delta(h)|$. The Pareto frontier is the set of those classifiers that are not Pareto-dominated by any other classifier.

Before we proceed further, we will highlight an important property of finding λ, which is also a parameter now that controls the adjustment we make to the unconstrained Bayes optimal threshold of $1/2$. To that end, we will first attempt to upper and lower bound λ. Along with this information on endpoints and the monotonicity properties of λ from Lemma 6.2, this would allow us to set up a bisection search algorithm.

Lemma 6.3. Assuming $w(.,0) < 0$ and $w(.,1) > 0$ (Definition 6.1), when $\Delta(h_0) > 0, \lambda_{max} = \frac{1}{\inf_{(x,a)} |w(x,a)|}$, so that the range $[0, \lambda_{max}]$ contains the desired optimal point $\lambda^*(\delta)$ for any δ, including the λ for exact fairness $\lambda^*(0)$. And when $\Delta(h_0) < 0, \lambda_{min} = \frac{-1}{\inf_{(x,a)} |w(x,a)|}$, so that the range$[\lambda_{min}, 0]$ contains the desired optimal point $\lambda^*(\delta)$ for any δ, including the λ for exact fairness $\lambda^*(0)$

The proof of Lemma 6.3 is given in Section 6.6.2. This condition becomes much simpler for the equal opportunity and the demographic parity fairness metrics: $\lambda_{max} = \max_{a \in \mathcal{A}} \pi_{1a}$ and $\lambda_{max} = \max_{a \in \mathcal{A}} \pi_a$, respectively.

Restricting the possible range of λ allows us to set up a bisection search to estimate the fairness-accuracy tradeoff (Algorithm 6.1). We now argue that approximating this tradeoff curve pointwise is sufficient for our scenario to obtain a high-probability estimate of the true Pareto frontier.

Using a well-known characterization of the Pareto frontier by [Var76], we can equivalently write the above set as $h_\lambda = \operatorname{argmin}_h R(h) + \lambda \cdot \Delta(h)$. Before we proceed further, we remark on the use of signed vs. absolute value of the unfairness.

Remark 6.2.2 (Signed disparity vs. reported unfairness)**.** We use $\Delta(h)$ and $|\Delta(h)|$ for two distinct roles. The magnitude $|\Delta|$ is for *specification*: checking unfairness ($|\Delta(h)| \leqslant \delta$), Pareto dominance (Definition 6.3), and the stopping tolerance of our bisection algorithm (Algorithm 6.1). The signed Δ with $\lambda \in \mathbb{R}$ is the *mechanism*: it is the exact Lagrangian of the two-sided constraint $-\delta \leqslant \Delta(h) \leqslant \delta$ (with $\lambda = \mu_+ - \mu_-$ and $\mu_+\mu_- = 0$ for $\delta > 0$), so $\operatorname{sign}(\lambda)$ identifies the binding side. Crucially, for each fixed λ the objective $R(h) + \lambda\Delta(h)$ is *linear in* h, which is exactly what yields the closed-form, estimable threshold of Lemma 5.1 and makes the sweep $\lambda \mapsto \Delta(h_\lambda)$ a single monotone curve over $\mathbb{R}$ (Lemma 6.2), underpinning Proposition 6.1 and orienting the search (Lemma 6.3). It can be shown that both views are operationally equivalent.

We now show that any point on the Pareto frontier will be covered by any curve that we generate by varying the values of λ specified by the range obtained from Lemma 6.3.

Proposition 6.1. If a classifier h belongs to the Pareto-frontier, then $h = h_\lambda$, for some $\lambda \in \mathbb{R}$.

The proof of Proposition 6.1 is given in Section 6.6.2. We will show in the next section that any h_λ can be constructed as a threshold classifier, where the threshold is a function of some distribution-dependent parameters and λ. And then using Lemma 6.3, we will be able to estimate the risk and unfairness of several points on the Pareto frontier by varying the value of λ in a fixed range.

§ 6.2.2 Cost Sensitive Classification

In classification, we generally work with the 0-1 loss. However, there may be scenarios where we need to weigh different errors unequally, such as the cost of missing a disease (where a false negative might be riskier than a false positive) and the cost of granting loans to undeserving candidates (where the cost of defaulting might be more than missing out on a good loan handout). This is the setting of cost-sensitive classification [Elk01].

Recall our previous result on optimal cost-sensitive thresholds on η (Lemma 2.1). We can immediately see that the thresholds for fair Bayes optimal classifiers from Lemma 6.1 can be written down as group-dependent costs [MW18]. For example, for demographic parity, $\alpha_a = \frac{1}{2} + \frac{(2a-1)\lambda^*(\delta)}{2\pi_a}$, where $\lambda^*(\delta)$ is the optimal dual parameter for the optimization program in Equation 6.1. We will exploit Lemmas 6.3 and 2.1 in the coming sections to construct estimators for thresholds, risk, and disparity at arbitrary points up until $\lambda_{\max}$ to get an efficient estimate of the Pareto frontier, only using soft labels.

§ 6.3 Estimating the Unfairness and Error Rates Using Soft Labels

With all the tools in place, we will now derive expressions for the expected and finite-sample estimators. We will first derive consistent estimators of the cost-sensitive thresholds α_a that depend on λ, the unfairness metric of choice, and some distribution-dependent quantities in Theorem 6.1. Once we have that, we will first show a consistent estimator for the cost-sensitive Bayes risk, in a similar spirit to the results in Ishida et al. [Ish+] and Ushio et al. [UIS25] in Theorem 6.2 and then its finite sample version in Theorem 6.3. Since we are interested in estimating the fairness-accuracy tradeoff, we will then derive consistent estimators for the risk of a classifier (Theorem 6.4) and then the unfairness (Theorem 6.5). Finally, we will present an algorithm to construct the tradeoff curve (Algorithm 6.1).

§ 6.3.1 Estimating Fair Thresholds using Finite Soft Labels

Unlike prior work [Ish+; JCD23; JCD24; UIS25], where soft labels were enough to estimate the Bayes risk β, in the case of any cost-sensitive risk, unless the cost α_a is a constant, we need to estimate α_a from given soft labels, since α_a would often depend on distribution dependent quantities. This is certainly true for the fairness constraints defined in Lemma 6.1. We now define the estimated thresholds for both DP and EO fairness constraints, for any arbitrary λ. For demographic parity:

$$\hat{\alpha}_a^+ = \Pi(\tilde{\alpha}_a^+), \text{ where } \tilde{\alpha}_a^+ = \frac{1}{2} + \frac{(2a-1)\lambda}{2\hat{\pi}_a^+},$$

and for equal opportunity:

$$\hat{\alpha}_a^+ = \Pi(\tilde{\alpha}_a^+), \text{ where } \tilde{\alpha}_a^+ = \frac{1}{2 - \frac{(2a-1)\lambda}{\hat{\pi}_{1a}^+}},$$

where Π is the truncation operation defined earlier, that ensures that our estimated thresholds stay in the interval $[0,1]$. We now show that the estimators of α_a for DP and EO, with finitely many soft labels, are statistically consistent. Since the specific construction of the estimator requires knowledge of $w(X,A)$, we present estimators for demographic parity and equal opportunity, and the proof techniques will likely generalize to other fairness metrics and cost-sensitive applications.

Theorem 6.1. (Consistency and Sample Complexity for threshold estimation) Given a dataset D_α of clean soft labels ($c_i = \eta(x_i, a_i)$) and a fixed $\lambda \in \mathbb{R}$, consider the empirical estimator $\hat{\alpha}_a^+$. Assuming $2 - \frac{(2a-1)\lambda}{\pi_{1a}} \geqslant \tau > 0$ (denominator bounded away from zero), the following holds:

1. Consistency: $\hat{\alpha}_a^+ \xrightarrow{p} \alpha_a$.
2. Bias:
   - DP: $|\alpha_a - \mathop{\mathbb{E}}_{D_\alpha}[\hat{\alpha}_a^+]| \leqslant \frac{|\lambda|}{\pi_a}\sqrt{\frac{1-\pi_a}{\pi_a N_\alpha}} + 2\exp\left(-\frac{\pi_a^2}{2}N_\alpha\right)$,
   - EO: $|\alpha_a - \mathop{\mathbb{E}}_{D_\alpha}[\hat{\alpha}_a^+]| \leqslant \frac{4|\lambda|}{\tau^2\pi_{1a}^2}\sqrt{\frac{\pi_{1a}(1-\pi_{1a})}{N_\alpha}} + 2\exp(-2\epsilon^2 N_\alpha)$, where $\epsilon \leqslant \min\{\frac{\pi_{1a}}{2}, \frac{\tau\pi_{1a}^2}{4|\lambda|}\}$.
3. Sample Complexity: With $1-\delta$ probability, $\Pr(|\hat{\alpha}_a^+ - \alpha_a| > \epsilon) \leqslant \delta$, as long as:
   - DP: $N_\alpha > \max\{\frac{2}{\pi_a^2}\log\frac{4}{\delta}, \frac{\lambda^2}{2\epsilon^2\pi_a^4}\log\frac{4}{\delta}, \frac{2}{\pi_a}\}$,
   - EO: $N_\alpha \geqslant \frac{1}{2t^2}\log\left(\frac{2}{\delta}\right)$ for $t = \min\left\{\frac{\pi_{1a}}{2}, \frac{\tau\pi_{1a}^2}{4|\lambda|}, \frac{\epsilon\tau^2\pi_{1a}^2}{4|\lambda|}\right\}$.

The proof of Theorem 6.1 is presented as two separate proofs (for DP and EO) in Section 6.6.3.

Remark 6.3.1. It is possible to remove the dependency of λ in the sample complexity results by using Lemma 6.3 that tells us about the maximum λ, and study the worst-case sample complexity bounds. However, keeping λ-dependent results here highlights the (indirect) polynomial dependency on the unfairness through the correction factor λ. This also indicates that threshold estimation is non-uniform with respect to the underlying unfairness: the number of samples required may change drastically depending on the degree of correction and the relative group proportions in the distribution.

§ 6.3.2 Bayes Error Estimation for Cost-Sensitive Classification

Prior works [Ish+; UIS25; JCD24] study an estimator for the Bayes error of a distribution for binary classification, by leveraging a simple observation: the Bayes error (which we can denote with β) can be written down as $\beta = \mathop{E}_{X}[\min(\eta(x), 1-\eta(x))]$. It was shown in Proposition 3.1 of [Ish+] that $\hat{\beta} = \frac{1}{N}\sum_{i\in[N]} \min(\eta_i, 1-\eta_i)$, where $\eta_i = \eta(x_i)$, is a consistent estimator of the Bayes risk.

In the previous section, we derived the expression for cost-sensitive Bayes risk and the corresponding Bayes optimal classifier (Lemma 2.1). Let β_{CS} denote the Bayes error for a distribution with respect to a fixed cost-sensitive risk, with group-dependent costs α_a. As an independent result, and in a similar spirit to Proposition 3.1 of Ishida et al. [Ish+], we now derive the consistency and sample-complexity properties of an estimator for the cost-sensitive Bayes risk.

Theorem 6.2. Consider the Bayes optimal classifier for the cost-sensitive risk with group-dependent costs α_a:

$$h^*_\alpha(x, a) = \begin{cases} 1, & \text{if } \eta(x,a) \geqslant \alpha_a \\ 0, & \text{otherwise.} \end{cases}$$

1. Bayes Risk: The Bayes error with respect to the cost-sensitive risk can be expressed as $\beta_{CS} = \mathop{E}_{X,A}[\min((1-\alpha_A)\eta(X,A), \alpha_A(1-\eta(X,A)))]$.
2. Consistency: $\hat{\beta}_{CS} = \frac{1}{N_\beta}\sum_{i\in[N_\beta]} \min((1-\alpha_i)c_i, \alpha_i(1-c_i))$ is a consistent estimator of β_{CS}, where for the i^{th} instance (c_i, a_i), $c_i = \eta(x_i, a_i)$ and $\alpha_i = \alpha_{a_i}$, the group dependent cost.
3. Sample Complexity: For any $\delta > 0$, with probability at least $1-\delta$, $|\hat{\beta}_{CS} - \beta_{CS}| \leqslant \frac{1}{4}\sqrt{\frac{1}{N_\beta}\log(\frac{2}{\delta})}$.

The proof of Theorem 6.2 is given in Section 6.6.3. While this result is geared toward group-aware cost-sensitive risks, it provides a roadmap for deriving results similar to prior work [Ish+; UIS25; JCD24] and for generalizing their estimators for Bayes error estimation in cost-sensitive and complex metric applications [Koy+14; SK22]. We will now proceed to specialize Theorem 6.2 towards fairness applications and construct finite sample estimators using soft labels.

With the fact $\hat{\alpha}^+_a \xrightarrow{p} \alpha_a$ established in Theorem 6.1, let $\hat{\beta}_{CS,\hat{\alpha}^+_a} = \frac{1}{N_\beta}\sum_{i\in[N_\beta]} \min((1-\hat{\alpha}^+_a)c_i, \hat{\alpha}^+_a(1-c_i))$ denote the empirical estimator of the cost-sensitive Bayes error with the empirical estimation of cost-sensitive thresholds α_a. We now show that $\hat{\beta}_{CS,\hat{\alpha}} \xrightarrow{p} \beta_{CS}$.

Theorem 6.3. Suppose the dataset D_α is used to construct a consistent estimator of the cost-sensitive threshold α_a (Theorem 6.1). Then, given a dataset D_β the following statements are true:

1. Consistency: $\hat{\beta}_{CS,\hat{\alpha}} = \frac{1}{N_\beta}\sum_{i\in[N_\beta]} \min((1-\hat{\alpha}^+_{a_i})c_i, \hat{\alpha}^+_{a_i}(1-c_i))$ is a consistent estimator of β_{CS}.
2. Bias: For a fixed $\delta \in (0,1)$:
   - Demographic Parity: $|\beta_{CS} - \mathop{E}_{D_\beta, D_\alpha}[\hat{\beta}_{CS,\hat{\alpha}}]| = \mathcal{O}\left(|\lambda| \cdot \sqrt{\frac{\log(1/\delta)}{N_\alpha}}\right)$,
   - Equal Opportunity: $|\beta_{CS} - \mathop{E}_{D_\beta, D_\alpha}[\hat{\beta}_{CS,\hat{\alpha}}]| = \mathcal{O}\left(\frac{|\lambda|}{\tau^2}\sqrt{\frac{\log(1/\delta)}{N_\alpha}}\right)$.
3. Sample Complexity: With $1-\delta$ probability, $|\beta_{CS} - \hat{\beta}_{CS,\hat{\alpha}}| < \epsilon$ as long as $N_\beta \geqslant \frac{1}{8\epsilon^2}\log(\frac{4}{\delta})$

and:

- Demographic Parity: $N_\alpha = \mathcal{O}\left(\frac{\lambda^2}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$,
- Equal Opportunity: $N_\alpha = \mathcal{O}\left(\frac{\lambda^2}{\tau^4\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$.

The proof of Theorem 6.3 is given in Section 6.6.3. Unlike Bayes error estimation [Ish+; JCD24; UIS25] where α_a was always half, due to distribution and unfairness dependent α_a, we now have biased estimators and sufficient sample requirement that depends on the unfairness correction λ and for equal opportunity metric, the degree of our assumption on a well-defined denominator (via τ).

§ 6.3.3 Expected Error Estimation

However, when considering the fairness-accuracy tradeoff, we are interested in computing the error rate with respect to the expected 0-1 loss. Recall from Definition 6.2 the definition of expected risk with respect to the 0-1 loss:

$$R(h) = \sum_{a\in\{0,1\}} \pi_a \left(\mathop{E}_{X|A=a} [(1-2\eta(X,a))h(X,a)] \right) + \pi = \mathop{E}_{X,A} [(1-2C)\mathbb{I}(C \geqslant \alpha_A)] + \pi$$

A finite sample estimator of the above risk can be written down as follows:

$$\hat{R}(\hat{h}) = \frac{1}{N_\beta} \sum_{i\in[N_\beta]} (1-2c_i)\mathbb{I}\left(c_i \geqslant \hat{\alpha}^+_{a_i}\right) + \hat{\pi}.$$

Theorem 6.4. Consider $\hat{R}(\hat{h})$, the empirical estimator for the expected risk from Definition 6.2. Then, the following properties hold:

1. Consistency: $\hat{R}(\hat{h}) \xrightarrow{p} R(h)$.
2. Bias: For a fixed $\delta \in (0,1)$:
   - Demographic Parity: $|R(h) - \mathop{E}_{D_\alpha, D_\beta}\left[\hat{R}(\hat{h})\right]| = \mathcal{O}\left(|\lambda| \cdot \max\{L_0, L_1\} \cdot \sqrt{\frac{\log(1/\delta)}{N_\alpha}}\right)$,
   - Equal Opportunity: $|R(h) - \mathop{E}_{D_\alpha, D_\beta}\left[\hat{R}(\hat{h})\right]| = \mathcal{O}\left(\frac{|\lambda|\max\{L_0,L_1\}}{\tau^2}\sqrt{\frac{\log(1/\delta)}{N_\alpha}}\right)$.
3. Sample Complexity: With $1-\delta$ probability, $|R(h) - \hat{R}(\hat{h})| < \epsilon$ as long as $N_\beta \geqslant \frac{18}{\epsilon^2}\log\left(\frac{6}{\delta}\right)$ and:
   - Demographic Parity: $N_\alpha = \mathcal{O}\left(\frac{\lambda^2\max\{L_0^2,L_1^2\}}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$,
   - Equal Opportunity: $N_\alpha = \mathcal{O}\left(\frac{\lambda^2\max\{L_0^2,L_1^2\}}{\epsilon^2\tau^4}\log\left(\frac{1}{\delta}\right)\right)$.

The proof of Theorem 6.4 is given in Section 6.6.3. Note that estimating the expectation using the sample set D_β is standard. The difficult part is that our underlying group-dependent thresholds are estimates with respect to D_α. This makes estimating risk under constraints more difficult and introduces nontrivial differences with respect to prior results on risk estimation with finite samples [Ish+; JCD24; UIS25].

§ 6.3.4 Unfairness Estimation

We now proceed to estimate unfairness using a finite number of soft labels. Since every fairness metric would require different distribution-dependent quantities, we focus on demographic parity and equal opportunity metrics. To proceed with estimating the unfairness rates, we will need estimators for the True positive rates (TPR) and Positive rates (PR) later in the chapter. We will first tackle TPR, and PR can be tackled similarly. A natural estimator for the TPR is to replace all quantities involved with their sample estimates, as done in Jeong et al. [JCD24]. Let $\rho_{\mathrm{TPR},a}(h_\alpha)$ denote the true positive rate for a threshold classifier $h_\alpha(x,a) = \mathbb{I}(\eta(x,a) \geqslant \alpha_a)$ on group $A = a$.

$$\begin{aligned}\rho_{\mathrm{TPR},a}(h_\alpha) &= \Pr(h_\alpha(X,A) = 1 | Y = 1, A = a) = \frac{\Pr(h_\alpha(X,a) = 1, Y = 1, A = a)}{\Pr(Y = 1, A = a)} \\ &= \frac{\Pr(h_\alpha(X,a) = 1, Y = 1 | A = a)\Pr(A = a)}{\Pr(Y = 1 | A = a)\Pr(A = a)} = \frac{\Pr(h_\alpha(X,a) = 1, Y = 1 | A = a)}{\Pr(Y = 1 | A = a)} \\ &= \frac{\mathop{\mathrm{E}}_{X|A=a}[\mathbb{I}(\eta(X,a) \geqslant \alpha_a)\eta(X,a)]}{\pi_{1|a}}.\end{aligned}$$

We can now define a sample estimate as:

$$\hat{\rho}_{\mathrm{TPR},a}(h_{\hat{\alpha}_a^+}) = \frac{\frac{1}{N_{\beta,a}}\sum_{i\in[N_{\beta,a}]}(c_i(\mathbb{I}(c_i \geqslant \hat{\alpha}_a^+))}{\hat{\pi}_{1|a}^+},$$

where $N_{\beta,a}$ are the number of samples in D_β from group $A = a$, and $\hat{\alpha}_a^+$ is the threshold constructed using finite samples or intermediate estimators using finite samples (for example, thresholds for fairness constraints derived in Lemma 6.1).

Once we have our estimators for the TPR, we can define the equal opportunity difference of a classifier h as $\Delta_{\mathrm{EO}}(h) = |\rho_{\mathrm{TPR},1}(h) - \rho_{\mathrm{TPR},0}(h)|$ and its empirical counterpart $\hat{\Delta}_{\mathrm{EO}}(h) = |\hat{\rho}_{\mathrm{TPR},1}(h) - \hat{\rho}_{\mathrm{TPR},0}(h)|$. This defines $\Delta_{\mathrm{EO}}(h_\alpha)$ as the true equal opportunity difference with respect to the true threshold α_a and $\hat{\Delta}_{\mathrm{EO}}(h_{\hat{\alpha}})$ as the empirical estimate of equal opportunity with the empirical estimate of the threshold.

If we only care about the positive rate (for ensuring demographic parity), we can come up with a natural estimator:

$$\rho_{\mathrm{PR},a}(h_\alpha) = \Pr(h_\alpha(X,a) = 1 | A = a) = \mathop{\mathrm{E}}_{X|A=a}[\mathbb{I}(\eta(X,a) \geqslant \alpha_a)],$$

and its finite sample estimate:

$$\hat{\rho}_{\mathrm{PR},a}(h_{\hat{\alpha}_a^+}) = \frac{1}{N_{\beta,a}}\sum_{i\in[N_{\beta,a}]}\mathbb{I}(c_i \geqslant \hat{\alpha}_a^+).$$

Similar to the equal opportunity difference, we can write the demographic parity difference as $\Delta_{\mathrm{DP}}(h) = |\rho_{\mathrm{PR},1}(h) - \rho_{\mathrm{PR},0}(h)|$. Then, the true DP difference on the true threshold is $\Delta_{\mathrm{DP}}(h_\alpha)$ and the empirical DP difference on the estimated threshold is $\hat{\Delta}_{\mathrm{DP}}(h_{\hat{\alpha}})$.

Theorem 6.5. Suppose a dataset D_β is used to construct $\hat{\Delta}(h_{\hat{\alpha}_a^+})$, where $\hat{\alpha}_a^+$ is the estimator for the DP or the EO threshold (Theorem 6.1). Then the following holds:

1. Consistency: $\hat{\Delta}(h_{\hat{\alpha}_a^+}) \xrightarrow{p} \Delta(h_\alpha)$.
2. Bias: For $\delta \in (0,1)$,
   - DP: $|\Delta_{DP}(h) - \underset{D_\beta, D_\alpha}{E}\left[\hat{\Delta}_{DP}(\hat{h})\right]| = \mathcal{O}\left(\frac{1}{\sqrt{N_\beta}} + |\lambda| \cdot \max\{L_0, L_1\}\sqrt{\frac{\log(1/\delta_a)}{N_\alpha}}\right)$.
   - EO: $|\Delta_{EO}(h) - \underset{D_\beta, D_\alpha}{E}\left[\hat{\Delta}_{EO}(\hat{h})\right]| = \mathcal{O}\left(\frac{|\lambda| \cdot \max\{L_0, L_1\}}{\tau^2}\sqrt{\frac{\log(1/\delta)}{N_\alpha}} + \frac{1}{\sqrt{N_\beta}}\right)$.
3. Sample Complexity: With $1-\delta$ probability, $|\hat{\Delta}(\hat{h}) - \Delta(h)| \leqslant \epsilon$, as long as $N_\beta = \mathcal{O}\left(\frac{1}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$ and:
   - DP: $N_\alpha = \mathcal{O}\left(\frac{\lambda^2 \cdot \max\{L_0^2, L_1^2\}}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$,
   - EO: $N_\alpha = \mathcal{O}\left(\frac{\lambda^2 \max\{L_0^2, L_1^2\}}{\epsilon^2 \tau^4}\log\left(\frac{1}{\delta}\right)\right)$.

The proof of Theorem 6.5 is presented as two separate proofs (for DP and EO) in Section 6.6.3. Just like the bias and sample complexity results in Theorem 6.4, the rates of unfairness estimation depend sharply on the unfairness correction and our smoothness assumptions, apart from standard rates [MRT12].

Proof Sketch. We summarize the main ideas behind the proofs of the DP and EO threshold (Theorem 6.1), and the risk and unfairness estimators (Theorems 6.3, 6.4, 6.5).

- **Consistency**: The consistency proofs begin with the Weak Law of Large Numbers for the empirical priors $\hat{\pi}$, $\hat{\pi}_a$, $\hat{\pi}_{ya}$, and $\hat{\pi}_{y|a}$. The threshold estimators $\hat{\alpha}_a$ are then shown to be consistent by applying the Continuous Mapping Theorem to the corresponding plug-in formulas, together with the truncation operator $\Pi_{[0,1]}$ [1]. In the EO case, we additionally require the relevant denominator to stay bounded away from zero. The consistency results for risk and unfairness then follow by combining the consistency of $\hat{\alpha}_a$ with continuity of the corresponding functionals, along with standard concentration of empirical averages over the independent sample split.
- **Bias bounds**: For each estimator, we define a suitable good event $\mathcal{G}$ on which the empirical prior of interest is close to its population counterpart. On $\mathcal{G}$, the estimator error is bounded deterministically in terms of the prior estimation error, using elementary algebraic control of reciprocals whenever needed. Taking expectations then reduces to controlling quantities such as $E[|\hat{\pi} - \pi|]$, which are bounded using explicit variance calculations and standard inequalities for bounded random variables. On the complement $\mathcal{G}^C$, the estimator error is controlled by the uniform boundedness of the estimator, multiplied by $\Pr(G^c)$, and $\Pr(G^c)$ is bounded using Hoeffding's inequality. For the later risk and unfairness results, the threshold estimation error is propagated through the corresponding risk or disparity functional using continuity or Lipschitz bounds under Assumptions 2.1 and 6.1.
- **Sample complexity**: The sample complexity results are obtained by combining concentration bounds for the threshold estimators built from D_α with concentration bounds for the empirical risk and unfairness estimators built from the independent split D_β. The total failure probability is then controlled by a union bound. For the later results, the final sample requirements inherit

[1] stat.cmu.edu

Algorithm 6.1: Tradeoff Estimation with Soft Labels

Input: Disjoint datasets D_α, D_β; Unfairness Tolerance γ, Precision ξ; Priors $\hat{\Pi}$
Functions:
- AlphaEst$(\lambda, \hat{\Pi}) \to \hat{\alpha}_a^+$ (Function to estimate the cost-sensitive thresholds α_a given a λ and distribution priors);
- ErrUnfEst$(\lambda, \text{AlphaEst}(\lambda, \hat{\Pi})) \to R(g_{\hat{\alpha}_a^+}), \Delta(g_{\hat{\alpha}_a^+})$ (Function to estimate the error and unfairness, given a threshold α_a computed using the previous function)
$\hat{R}_0, \hat{\Delta}_0 \leftarrow$ ErrUnfEst$(0, \hat{\alpha}_0)$
if $|\hat{\Delta}_0| \leqslant \gamma$ **then** return No-Tradeoff.
Set range: **if** $\hat{\Delta}_0 > 0$ **then** $\lambda \in [0, \frac{1}{\inf w}]$ **else** $\lambda \in [-\frac{1}{\inf w}, 0]$
points $\leftarrow \{\}$
while $\lambda_{max} - \lambda_{min} > \xi$ **do**
  $\lambda \leftarrow (\lambda_{max} + \lambda_{min})/2$
  $\hat{R}, \hat{\Delta} \leftarrow$ ErrUnfEst$(\lambda, \text{AlphaEst}(\lambda, \hat{\Pi}))$
  Append $(\lambda, \hat{R}, \hat{\Delta})$ to points
  if $|\hat{\Delta}| \leqslant \gamma$ **then** break loop
  if $\hat{\Delta} > \gamma$ **then** $\lambda_{max} \leftarrow \lambda$ **else** $\lambda_{min} \leftarrow \lambda$
end while
return points

both the threshold-estimation complexity and the Lipschitz dependence of the downstream risk or disparity estimators on the estimated thresholds.

§ 6.3.5 Tradeoff Estimation

With all the tools in place, we now present the algorithm for estimating the Pareto frontier using the estimators derived using soft labels in the previous section. We will now describe the algorithm formally in Algorithm 6.1 and discuss each design choice in more detail.

1-2 Assuming access to sufficiently large, disjoint datasets $D_\alpha.D_\beta$ sampled i.i.d. from our target distribution, we obtain estimates for all the distribution-dependent quantities $\hat{\Pi} = \{\hat{\pi}, \hat{\pi}_a, \hat{\pi}_{ya}, \hat{\pi}_{y|a}\}$. Using these priors $\hat{\Pi}$ and any λ, we can estimate our thresholds $\hat{\alpha}_a^+$ (AlphaEst) and then our risk R and unfairness Δ (ErrUnfEst).

3-4 If the Bayes optimal classifier (at $\lambda = 0$) already satisfies the required fairness ($\hat{\Delta}_0$), then we are done.

5 Depending on sign of $\hat{\Delta}_0$, we can set λ_{max} from Lemma 6.3 to fix our range for the bisection search.

8-12 We compute λ and calculate the updated $\hat{R}, \hat{\Delta}$. If we immediately achieve our required unfairness γ, we are done. Otherwise, depending on the sign, we update our bound $\lambda_{max}/\lambda_{min}$ and continue until the gap between λ_{min} and λ_{max} is under a set precision ξ.

Remark 6.3.2 (On the consistency of the estimated frontier)**.** Our aim is to estimate the fairness-accuracy tradeoff, estimated pointwise in λ, rather than the risk at a prescribed unfairness level δ: the guarantees of Theorem 6.1 (and Theorems 6.4, 6.5) are stated for a fixed λ. However, we can provide a high-probability guarantee on the 'correctness' of the tradeoff: (1) Fixed grid: If the λ values are fixed on a grid $\{\lambda_1, \ldots, \lambda_K\} \subset [\lambda_{min}, \lambda_{max}]$, the search points are data independent and a union bound over the K points controls the error, (2) Bisection: In its current state, Algorithm 6.1 selects each λ from the estimated disparity $\hat{\Delta}$, which makes the points data-dependent, and a union bound over them is not directly valid. This can be rectified by studying the *uniform-in-*λ guarantee, $\sup_\lambda |\hat{\Delta}(h_\lambda) - \Delta(h_\lambda)| \to 0$. Since $\{h_\lambda\}$ is a low-complexity, one-parameter monotone threshold family, such a bound may be available via a uniform law of large numbers. Bisection is more query-efficient, but its analysis requires controlling the error uniformly in λ.

Note that Algorithm 6.1 can not only be used to generate an estimate of the Pareto frontier, but can also be used to approximate the risk of the classifier when we desire disparity at most δ. This is because we can always interpolate the curve between the obtained points to get a sense of the sharpness of the tradeoff, along with the bias and sample-complexity bounds in Theorems 6.4 and 6.5. This can be readily used to enforce fairness and to choose among fairness metrics in many applications.

§ 6.4 Experiments

We now present empirical verification of the estimated tradeoff using the bisection search algorithm introduced earlier (Algorithm 6.1). We work with different sources of soft labels: a synthetic Gaussian setting where we can analytically compute the soft labels [ZDC22a], an average of hard labels on the CIFAR-10H dataset [Pet+19], directly borrowed from prior work [Ish+; UIS25] and noisy soft labels derived from calibrated classifiers (Logistic Regression) trained on the Adult dataset [Lic+13], a popular dataset for unfairness evaluation [BHN23]. We compare our computed tradeoff against several algorithms that return a tradeoff. We choose the 'Exp-Grad' method from Agarwal et al. [Aga+18] due to its theoretical and practical strengths [SDS23a; GKW23], Oxonfair [Del+24], Fairret [BDDB23], and the MIFPO algorithm [KPM24].

Oxonfair [Del+24] uses validation data to optimize classification thresholds of a provided pretrained predictor, and supports a variety of performance and fairness metrics. MIFPO (Model Independent Fairness-Performance Optimization) [KPM24] proposes a discrete optimization method to compute the fairness-accuracy tradeoff using soft scores from calibrated classifiers. Finally, Fairret [BDDB23] proposes optimizing a differentiable regularizer as a proxy for any linear-fractional form of fairness metric [Cel+19]. All methods return a tradeoff for both fairness metrics, except MIFPO [KPM24], which returns a tradeoff only for demographic parity. For a valid comparison, we ensure that MIFPO uses the same soft scores as our algorithm does, instead of training a separate calibrated classifier on hard instances and using its soft scores.

In Figure 6.1, we show the results for the Gaussian, CIFAR-10H, and the Adult datasets. Averaged across 5 runs, we see that our estimate of the Pareto frontier consistently lower-bounds all other methods and exhibits the least variance. A peculiar observation is from the demographic parity results, where MIFPO seems to be a lower bound for the method in almost all cases. This is where we would like to highlight how MIFPO operates. Given a dataset, MIFPO aims to optimize for a lower bound that would almost always lower bound any Pareto frontier obtained on the data (Section 4 in Kozdoba et al. [KPM24]). However, for MIFPO, there are no guarantees that we are optimizing for the Bayes optimal fairness-accuracy tradeoff: comparing algorithms against MIFPO may always result in a huge gap, since it is not indicative of the true gap.

§ 6.5 Discussion

In this work, we set out to extend Bayes error estimation techniques towards estimating the fairness-accuracy Pareto frontier. We do so by generalizing the model and by using instance-free soft-label estimation techniques [Ish+; JCD24; UIS25] for cost-sensitive risks. Since fairness constraints themselves can be treated as a cost-sensitive learning problem, and given the properties of the fairness-accuracy Pareto front, we can set up consistent estimators for the fair thresholds and estimates of risk and unfairness. Empirically, we can also lower-bound many recent algorithms to obtain the fairness-accuracy frontier.

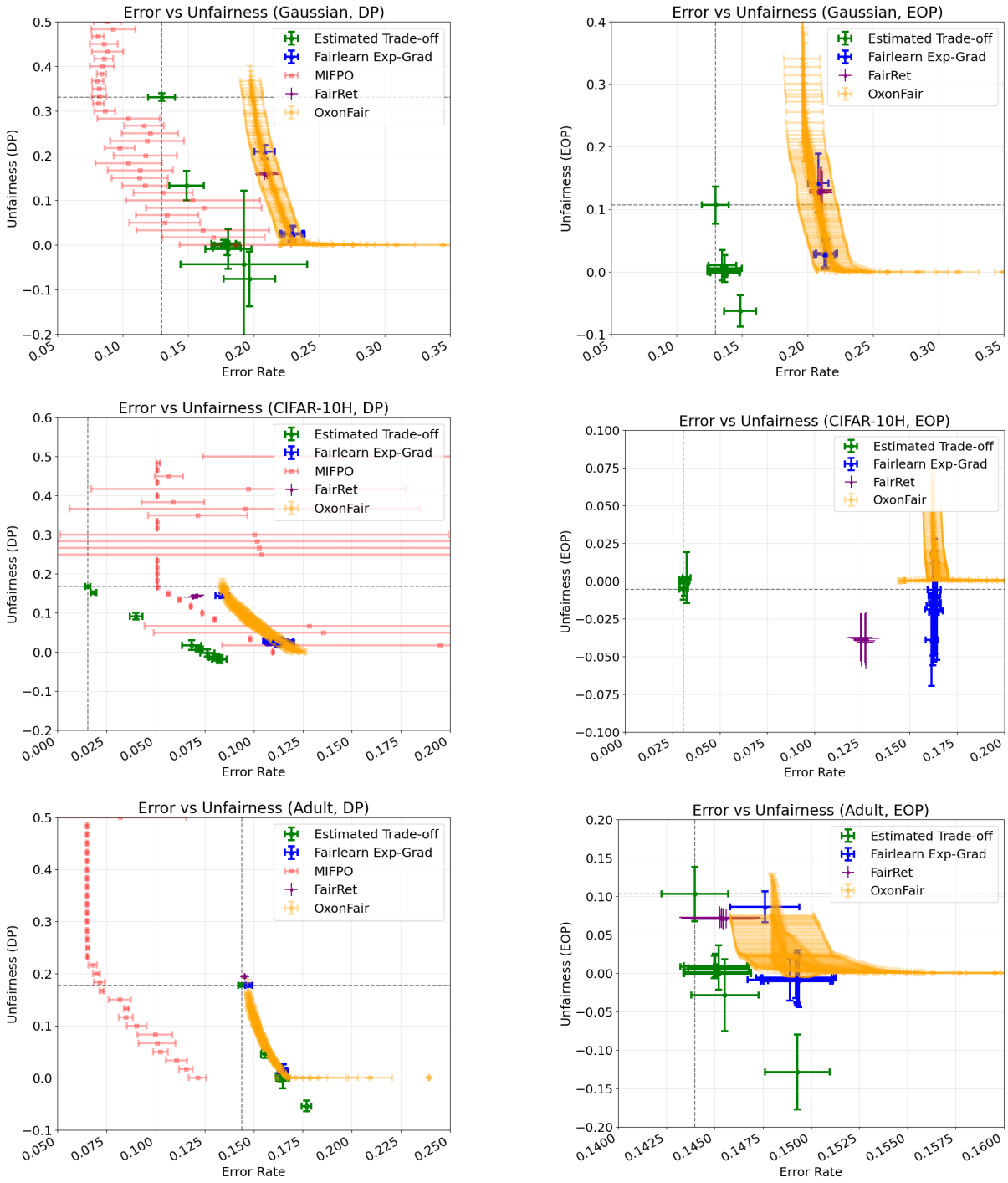


Figure 6.1: Unfairness-Error rate tradeoffs on Gaussian, Cifar-10H, and the Adult Dataset, for Demographic Parity (DP) and equal opportunity (EO) fairness constraints. Estimated Tradeoff (in green) always lower bounds other tradeoff estimation methods, except MIFPO (red), which computes a high-variance data-dependent lower bound different from the Bayes optimal. The intersection of dotted lines corresponds to the error rate and unfairness of the unconstrained Bayes Optimal classifier at $\lambda = 0$.

An immediate future work is to extend our consistency guarantees beyond the availability of clean soft labels. Our proof techniques are plug-and-play: whenever we can bound the noisy soft scores relative to the clean ones, all consistency and sample-complexity properties of our estimators follow from these bounds, with additional terms accounting for the consistency gap between the noisy and clean soft scores. Our guarantees show that even with clean soft scores, the problem is challenging in terms of strong dependence on the priors of the distribution and the unfairness correction term λ. Some recent work [UIS25] generalizes Bayes error estimation with soft labels to the case of monotonic noise in clean soft scores, and extending our estimators with their setup is a natural next step.

We could also further weaken the supervision signal by considering only positive-confidence (PC) data [INS18], which provides only soft scores and instances from the positive class, or positive-unlabeled (PU) data [DPNS14], which provides access to positively labeled and unlabeled samples. Wu and He [WH22] generalized the fair characterization results from Chzhen et al. [Chz+19] to come up with fair Bayes optimal threshold classifiers with PU data. Similarly, in the literature of Bayes error estimation from soft labels, researchers have also presented a Bayes error estimator with positive-confidence soft labels (Theorem 4.1 in Ishida et al. [Ish+]). Since our work fundamentally leverages threshold characterizations and soft-label estimators, estimating the tradeoff with weakly supervised setups such as PC and PU data is an interesting future direction.

Finally, while most of this work focuses on fairness constraints, our guarantees and approach also serve as a template to approach more multi-objective problems where the standard Bayes optimal classifier may not align with the optimal predictor for complex metrics [Koy+14; Cot+19; SK22].

§ 6.6 Proofs

In this section, we include the proofs for all results in the previous sections.

§ 6.6.1 Helper Results

Lemma 6.4. (Finite sample bound for estimation of π) Let $X_1, X_2, ..., X_N \in [0,1]$ be N i.i.d. random variables with mean π. Let $\hat{\pi} = \frac{1}{N}\sum_{i\in[N]} X_i$. For any $\epsilon > 0$, if $N \geqslant \frac{1}{2\epsilon^2}\log\left(\frac{2}{\delta}\right)$, then with $1-\delta$ probability over the choice of the data sample, $\hat{\pi} \in (\pi-\epsilon, \pi+\epsilon)$.

Proof. The proof follows directly from the application of Hoeffding's inequality[2]. □

Lemma 6.5. (Upper bound on Mean Absolute Error) Consider an estimator $\hat{\pi}$ constructed using N samples for some quantity π (Lemma 6.4). If $\hat{\pi}$ is an unbiased estimate of π, then $\mathop{E}_{D_N}[|\hat{\pi}-\pi|] \leqslant \sqrt{Var(\hat{\pi})}$.

Proof. Let $Z = \hat{\pi}-\pi$. Since $\hat{\pi}$ is an unbiased estimate of π, $\mathop{E}_{D_N}[Z] = 0$, $Var(Z) = \mathop{E}_{D_N}\left[Z^2\right] = \mathop{E}_{D_N}\left[|Z|^2\right] = Var(\hat{\pi})$. Observe that by Jensen's inequality, $E\left[|Z|^2\right] \geqslant (E[|Z|])^2 \implies E[|Z|] \leqslant \sqrt{E\left[|Z|^2\right]}$, Hence $\mathop{E}_{D_N}[|\hat{\pi}-\pi|] \leqslant \sqrt{Var(\hat{\pi})}$. □

§ 6.6.2 Proofs for Section 6.2

[2] stanford.edu

Proof. (Proof of Lemma 6.3) To construct our bounds on λ, we will try to study the case of exact fairness to obtain the worst possible bounds. Assume WLOG that $\mathtt{Dis}(h_0) > 0$. The proof for $\mathtt{Dis}(h_0) < 0$ is similar.

For demographic parity and equal opportunity fairness constraints (Definition 6.1), $w(.,0) < 0$ and $w(.,1) > 0$. Consider the threshold classifier for groups $A = 0$ and $A = 1$ at any λ (Lemma 6.1):

$$h_\lambda(x,0) = \mathbb{I}\left(\eta(x,0) \geqslant \frac{1}{2} - \frac{\lambda}{2}|w(x,0)|\right) \quad \text{and} \quad h_\lambda(x,1) = \mathbb{I}\left(\eta(x,1) \geqslant \frac{1}{2} + \frac{\lambda}{2}|w(x,1)|\right).$$

Also, by assumption, disparity is positive at $\lambda = 0$, where disparity is defined as:

$$\mathtt{Dis}(h_\lambda) = \pi_1 \int_{\mathcal{X}} |w(x,1)| \cdot h_\lambda(x,1) d\mathbb{P}_{X|A=1}(x) - \pi_0 \int_{\mathcal{X}} |w(x,0)| \cdot h_\lambda(x,0) d\mathbb{P}_{X|A=0}(x) > 0.$$

From Lemma 6.2, we can see that as λ increases, the threshold for group $A = 1$, t_1 increases (since $w(.,1) > 0$) and t_0 decreases (since $w(.,0) < 0$). If $t_1 > 1$, then the $h(x,1) = 0, \forall x$, since $\eta(x,1) \in [0,1]$. Similarly, if $t_0 \leqslant 0$, $h(x,0) = 1, \forall x$.

At that point, the disparity $\mathtt{Dis}(h_\lambda)$ at $\lambda = 0$, which was positive to begin with, becomes negative, and hence we cross over the point of zero unfairness. At such a trivial classification point, $R(h_\lambda) > 0$ and $\mathtt{Dis}(h_\lambda) < 0$, which is the minimum possible disparity since $\mathtt{Dis}(h) \in [-1,1]$.

It remains to check the value of λ at which the group-aware classifier h becomes trivial for each group. Let such a λ be denoted with $\lambda_{\max}$. Such a $\lambda_{\max}$ occurs at:

$$\begin{aligned}
&\text{(A=1):} \ \frac{1}{2} + \lambda \frac{|w(x,1)|}{2} > 1 \implies \lambda > \frac{1}{\inf_x |w(x,1)|}, \text{ and} \\
&\text{(A=0):} \ \frac{1}{2} - \lambda \frac{|w(x,0)|}{2} \leqslant 0 \implies \lambda \geqslant \frac{1}{\inf_x |w(x,0)|} \\
&\text{Overall:} \implies \lambda > \frac{1}{\inf_x |w(x,a)|}.
\end{aligned}$$

Now, note that due to Lemma 6.2, the tradeoff function parametrized by λ is convex, and by increasing λ towards $\lambda_{\max}$ (the disparity point of -1), we would have also crossed the zero disparity point for some $\lambda \in [0, \lambda_{\max}]$. Hence $\lambda \in [0, \lambda_{\max}]$ whenever the Bayes optimal classifier has positive disparity ($\mathtt{Dis}(h_0) > 0$). A similar argument follows for the other case when $\mathtt{Dis}(h_0) < 0$. □

Proof. (Proof of Proposition 6.1) Suppose g belongs to the Pareto-frontier and WLOG assume $\Delta(g) \geqslant 0$. We know from Proposition 4.6 in [Zen+26] that the Pareto front is convex.

Since g lies on the Pareto-frontier, then $(R(g), \Delta(g))$ must lie on the boundary of the above convex set. Otherwise, if $(R(g), \Delta(g))$ lies in the interior, there exists a small neighborhood around $(R(g), \Delta(g))$ that is contained in this convex set, and would include g' such that $R(g') < R(g)$ and $0 < \Delta(g') < \Delta(g)$. Now consider a supporting tangent to the convex set at $(R(g), \Delta(g))$ defined by a normal vector $(\alpha, \beta) \in \mathbb{R}^2$ and margin/shift $\gamma \in \mathbb{R}$ as follows: $\alpha R(h) + \beta \Delta(h) \geqslant \gamma$, for all classifiers h, and $\alpha R(g) + \beta \Delta(g) = \gamma$. Let $\lambda = \beta/\alpha$. Then $R(g) + \lambda \Delta(g) = R(g) + \beta/\alpha \cdot \Delta(g) = (\alpha R(g) + \beta \Delta(g))/\alpha = \gamma/\alpha$, whereas for any other h, we have $R(h) + \lambda \Delta(h) = R(h) + \beta/\alpha \cdot \Delta(h) = (\alpha R(h) + \beta \Delta(h))/\alpha \geqslant \gamma/\alpha$. Therefore, g is a minimizer of $R(h) + \lambda g(h)$ objective over all h, and is itself g_λ for $\lambda = \beta/\alpha$.

□

§ 6.6.3 Proofs for Section 6.3

For convenience and readability, we present the DP and EO results in Theorem 6.1 below as separate results.

Theorem 6.6. (Consistency and Sample Complexity for DP threshold) Given a dataset D_α of clean soft labels ($c_i = \eta(x_i, a_i)$) and a fixed $\lambda \in \mathbb{R}$, for the case of Demographic Parity, consider the following estimator of the threshold: $\hat{\alpha}_a^+ = \Pi(\tilde{\alpha}_a^+)$, where $\tilde{\alpha}_a^+ = \frac{1}{2} + \frac{(2a-1)\lambda}{2\hat{\pi}_a^+}$, $\forall a \in \{0,1\}$. Then the following holds:

1. Consistency: $\hat{\alpha}_a^+ \xrightarrow{p} \alpha_a$.
2. Bias: $|\alpha_a - \underset{D_\alpha}{E}[\hat{\alpha}_a^+]| \leqslant \frac{|\lambda|}{\pi_a}\sqrt{\frac{1-\pi_a}{\pi_a N_\alpha}} + 2\exp\left(-\frac{\pi_a^2}{2}N_\alpha\right)$.
3. Sample Complexity: With $1-\delta$ probability and $N_\alpha > \max\{\frac{2}{\pi_a^2}\log\frac{4}{\delta}, \frac{\lambda^2}{2\epsilon^2\pi_a^4}\log\frac{4}{\delta}, \frac{2}{\pi_a}\}$, $\Pr(|\hat{\alpha}_a^+ - \alpha_a| > \epsilon) \leqslant \delta$.

Proof. We will prove the results for an arbitrary $a \in \{0,1\}$. Due to Lemma 6.3 and our projection operation Π on the empirical estimator, α_a and $\hat{\alpha}_a^+$ are both restricted to the interval $[0,1]$. Using this, note that in our domain of $\lambda \in [0, \lambda_{max}]$ or $\lambda \in [-\lambda_{min}, 0]$, $\alpha_a = \Pi(\alpha_a)$, which is a useful property we exploit throughout the proof.

(1) **Consistency**: By the weak law of large numbers, $\hat{\pi}_a \xrightarrow{p} \pi_a$. Consider $\hat{\pi}_a^+$. Since $\pi_a > 0$ by assumption, we can pick a sufficiently large N_α such that $1/N_\alpha \leqslant \pi_a/2$. Therefore the event $\{\hat{\pi}_a < 1/N_\alpha\}$ is contained in the event $\{\hat{\pi}_a < \pi_a/2\}$ which is further contained in $\{|\hat{\pi}_a - \pi_a| \geqslant \pi_a/2\}$. Since $\hat{\pi}_a \xrightarrow{p} \pi_a$, $\Pr(|\hat{\pi}_a - \pi_a| \geqslant \pi_a/2) \to 0$ and hence $\Pr(\hat{\pi}_a^+ \neq \hat{\pi}_a) \to 0$, and hence $\hat{\pi}_a^+ \xrightarrow{p} \pi_a$.

The estimator $\hat{\alpha}_a^+$ is a continuous function of $\hat{\pi}_a^+$ (including the truncation operation). We can use the Continuous Mapping Theorem to claim that $\hat{\pi}_a^+ \xrightarrow{p} \pi_a \implies \hat{\alpha}_a^+ \xrightarrow{p} \alpha_a$ [3].

(2) **Bias**: We will now proceed towards bounding the bias of the threshold for an arbitrary group $A = a$. We can break down our bias as follows: $|\alpha_a - \underset{D_\alpha}{E}[\hat{\alpha}_a^+]| = |\underset{D_\alpha}{E}[\alpha_a - \hat{\alpha}_a^+]| = |\underset{D_\alpha}{E}[(\alpha_a - \hat{\alpha}_a^+)\mathbb{1}_{\mathcal{G}}] + \underset{D_\alpha}{E}\left[(\alpha_a - \hat{\alpha}_a^+)\mathbb{1}_{\mathcal{G}^c}\right]| \leqslant |\underset{D_\alpha}{E}[(\alpha_a - \hat{\alpha}_a^+)\mathbb{1}_{\mathcal{G}}]| + |\underset{D_\alpha}{E}\left[(\alpha_a - \hat{\alpha}_a^+)\mathbb{1}_{\mathcal{G}^c}\right]|$, where $\mathcal{G}$ is defined next.

Let $\mathcal{G} = \{|\hat{\pi}_a - \pi_a| \leqslant \epsilon\}$ be the good event that our estimate of π_a is close. On this event, for sufficiently large N_α ($N_\alpha > \frac{2}{\pi_a}$), $\hat{\pi}_a^+ = \hat{\pi}_a$ and hence $\hat{\alpha}_a^+ = \hat{\alpha}_a$. Using Hoeffding's inequality, we know that $\hat{\pi}_a \in (\pi_a - \epsilon, \pi_a + \epsilon)$ with $1 - \delta/2$ probability over the choice of data sample D_α. We can set $\epsilon \leqslant \pi_a/2$, which means $N_\alpha > \frac{2}{\pi_a^2}\log\left(\frac{4}{\delta}\right)$.

Consider the first term:

[3] stat.cmu.edu

$$
\begin{aligned}
|\mathop{\mathrm{E}}_{D_\alpha}[(\alpha_a - \hat{\alpha}_a)\mathbb{1}_{\mathcal{G}}]| &\leqslant \mathop{\mathrm{E}}_{D_\alpha}[|\alpha_a - \hat{\alpha}_a|\mathbb{1}_{\mathcal{G}}] = \mathop{\mathrm{E}}_{D_\alpha}[|\Pi(\alpha_a) - \Pi(\tilde{\alpha}_a)|\mathbb{1}_{\mathcal{G}}] \\
&\leqslant \mathop{\mathrm{E}}_{D_\alpha}[|\alpha_a - \tilde{\alpha}_a|\mathbb{1}_{\mathcal{G}}] \quad \text{(Since } \Pi \text{ is Lipschitz)} \\
&= \mathop{\mathrm{E}}_{D_\alpha}\left[\left|\frac{\lambda(2a-1)}{2\pi_a} - \frac{\lambda(2a-1)}{2\hat{\pi}_a}\right|\mathbb{1}_{\mathcal{G}}\right] \\
&\leqslant \frac{|\lambda|}{2}\cdot\mathop{\mathrm{E}}_{D_\alpha}\left[\left|\frac{1}{\pi_a} - \frac{1}{\hat{\pi}_a}\right|\mathbb{1}_{\mathcal{G}}\right] \\
&= \frac{|\lambda|}{2}\cdot\mathop{\mathrm{E}}_{D_\alpha}\left[\left|\frac{\hat{\pi}_a - \pi_a}{\pi_a\hat{\pi}_a}\right|\mathbb{1}_{\mathcal{G}}\right] = \frac{|\lambda|}{2}\cdot\mathop{\mathrm{E}}_{D_\alpha}\left[\left|\frac{\hat{\pi}_a - \pi_a}{\pi_a\hat{\pi}_a}\right|\mathbb{1}_{\mathcal{G}}\right] \\
&= \frac{|\lambda|}{2}\cdot\mathop{\mathrm{E}}_{D_\alpha}\left[\frac{|\hat{\pi}_a - \pi_a|}{\pi_a\hat{\pi}_a}\mathbb{1}_{\mathcal{G}}\right] \text{ (since } \pi_a, \hat{\pi}_a > 0) \\
&\leqslant \frac{|\lambda|}{\pi_a^2}\mathop{\mathrm{E}}_{D_\alpha}[|\hat{\pi}_a - \pi_a|\mathbb{1}_{\mathcal{G}}] \text{ (On } \mathcal{G}, \frac{1}{\pi_a\hat{\pi}_a} \leqslant \frac{2}{\pi_a^2}).
\end{aligned}
$$

Since $\hat{\pi}_a$ is an unbiased estimate of π_a, using Lemma 6.5, $\mathop{\mathrm{E}}_{D_\alpha}[|\hat{\pi}_a - \pi_a|] \leqslant \sqrt{\mathrm{Var}(\hat{\pi}_a)}$. Using $\mathop{\mathrm{E}}_{D_\alpha}[|\alpha_a - \hat{\alpha}_a|\mathbb{1}_{\mathcal{G}}] \leqslant \mathop{\mathrm{E}}_{D_\alpha}[|\alpha_a - \hat{\alpha}_a|] \leqslant \frac{|\lambda|\sqrt{\mathrm{Var}(\hat{\pi}_a)}}{\pi_a^2}$, we get:

$$
|\mathop{\mathrm{E}}_{D_\alpha}[(\alpha_a - \hat{\alpha}_a)\mathbb{1}_{\mathcal{G}}]| \leqslant \frac{|\lambda|\sqrt{\mathrm{Var}(\hat{\pi}_a)}}{\pi_a^2} = \frac{|\lambda|}{\pi_a}\sqrt{\frac{1-\pi_a}{\pi_a N_\alpha}}
$$

where $\mathrm{Var}(\hat{\pi}_a) = \mathrm{Var}(\frac{1}{N_\alpha}\sum_{i\in[N_\alpha]}\mathbb{I}(a_i = a)) = \frac{1}{N_\alpha^2}\mathrm{Var}(\sum_{i\in[N_\alpha]}\mathbb{I}(a_i = a)) = \frac{\pi_a(1-\pi_a)}{N_\alpha}$.
On $\mathcal{G}^c$, we can use the fact that both $\alpha_a, \hat{\alpha}_a^+ \in [0,1]$:

$$
\begin{aligned}
&|\mathop{\mathrm{E}}_{D_\alpha}\left[(\alpha_a - \hat{\alpha}_a^+)\mathbb{1}_{\mathcal{G}^c}\right]| \leqslant |\mathop{\mathrm{E}}_{D_\alpha}\left[1\cdot\mathbb{1}_{\mathcal{G}^c}\right]| \\
&= \Pr\left(\mathcal{G}^C\right) \leqslant 2\exp\left(-\frac{\pi_a^2}{2}N_\alpha\right) \text{ (Using Hoeffding's inequality on } \Pr\left(\mathcal{G}^c\right)).
\end{aligned}
$$

Overall,

$$
|\alpha_a - \mathop{\mathrm{E}}_{D_\alpha}\left[\hat{\alpha}_a^+\right]| \leqslant \frac{|\lambda|}{\pi_a}\sqrt{\frac{1-\pi_a}{\pi_a N_\alpha}} + 2\exp\left(-\frac{\pi_a^2}{2}N_\alpha\right).
$$

(3) **Sample Complexity**: We now want to study the number of samples required for a good estimator of α_a^+. We want to ensure that $\Pr(|\hat{\alpha}_a^+ - \alpha_a| > \epsilon) \leqslant \delta$. Note that $\{|\hat{\alpha}_a^+ - \alpha_a| > \epsilon\} \subseteq \{|\tilde{\alpha}_a^+ - \alpha_a| > \epsilon\}$.

Reusing the analysis from the previous part, we can first ensure that $\Pr\left(\mathcal{G}^C\right) = \Pr(|\pi_a - \hat{\pi}_a| > \pi_a/2) < \delta/2$ by using Hoeffding's inequality to obtain that $N_\alpha > \frac{2}{\pi_a^2}\log\frac{4}{\delta}$.

Next, as stated earlier, on the good event $\mathcal{G}$, as long as $N_\alpha > \frac{2}{\pi_a}$ (which means $1/N_\alpha < \pi_a/2$), $\hat{\pi}_a^+ = \hat{\pi}_a$ and hence $\hat{\alpha}_a^+ = \tilde{\alpha}_a$. This implies that $|\tilde{\alpha}_a^+ - \alpha_a| = |\frac{\lambda}{2}|\cdot|\frac{1}{\hat{\pi}_a} - \frac{1}{\pi_a}| \leqslant \frac{|\lambda|}{\pi_a^2}|\hat{\pi}_a - \pi_a|$.

Therefore, on $\mathcal{G}$, $\{|\tilde{\alpha}_a^+ - \alpha_a| > \epsilon\} \subseteq \{|\hat{\pi}_a - \pi_a| > \frac{\epsilon\pi_a^2}{|\lambda|}\}$. Now, we can study the entire event and upper bound its failure probability:

$$\begin{aligned}\Pr\left(|\tilde{\alpha}_a^+ - \alpha_a| > \epsilon\right) &\leqslant \Pr\left(|\tilde{\alpha}_a^+ - \alpha_a| > \epsilon \cap \mathcal{G}\right) + \Pr\left(|\tilde{\alpha}_a^+ - \alpha_a| > \epsilon \cap \mathcal{G}^C\right) \\ &\leqslant \Pr\left(|\hat{\pi}_a - \pi_a| > \frac{\epsilon\pi_a^2}{|\lambda|}\right) + \Pr\left(\mathcal{G}^C\right) \\ &\leqslant 2\exp\left(-\frac{2N_\alpha\epsilon^2\pi_a^4}{\lambda^2}\right) + 2\exp\left(-\frac{N_\alpha\pi_a^2}{2}\right) \\ &\text{(Using Hoeffding's Inequality and earlier bound on } \mathcal{G}^C\text{)}\end{aligned}$$

Bounding the second term by $\delta/2$ gave us $N_\alpha > \frac{2}{\pi_a^2}\log\frac{4}{\delta}$. Doing the same for the first term gives us $N_\alpha > \frac{\lambda^2}{2\epsilon^2\pi_a^4}\log\frac{4}{\delta}$. Overall, $N_\alpha > \max\{\frac{2}{\pi_a^2}\log\frac{4}{\delta}, \frac{\lambda^2}{2\epsilon^2\pi_a^4}\log\frac{4}{\delta}, \frac{2}{\pi_a}\}$.

□

Theorem 6.7. (Consistency and Sample Complexity for EO threshold) Given a dataset D_α of clean soft labels ($c_i = \eta(x_i, a_i)$) and assuming $2 - \frac{(2a-1)\lambda}{\pi_{1a}} \geqslant \tau > 0$ (denominator bounded away from zero), for the case of equal opportunity, consider the following estimator of the threshold: $\hat{\alpha}_a^+ = \Pi(\tilde{\alpha}_a^+)$, where $\tilde{\alpha}_a^+ = \frac{1}{2 - \frac{(2a-1)\lambda}{\hat{\pi}_{1a}^+}}$. Then the following holds:

1. Consistency: $\hat{\alpha}_a^+ \xrightarrow{p} \alpha_a$.
2. Bias: $|\alpha_a - \underset{D_\alpha}{E}[\hat{\alpha}_a^+]| \leqslant \frac{4|\lambda|}{\tau^2\pi_{1a}^2}\sqrt{\frac{\pi_{1a}(1-\pi_{1a})}{N_\alpha}} + 2\exp(-2\epsilon^2 N_\alpha)$, where $\epsilon \leqslant \min\{\frac{\pi_{1a}}{2}, \frac{\tau\pi_{1a}^2}{4|\lambda|}\}$.
3. Sample Complexity: With $1-\delta$ probability over the choice of data sample N_α, as long as $N_\alpha \geqslant \frac{1}{2t^2}\log\left(\frac{2}{\delta}\right)$ for $t = \min\left\{\frac{\pi_{1a}}{2}, \frac{\tau\pi_{1a}^2}{4|\lambda|}, \frac{\epsilon\tau^2\pi_{1a}^2}{4|\lambda|}\right\}$, $\Pr\left(|\hat{\alpha}_a - \alpha_a| > \epsilon\right) \leqslant \delta$.

Proof. We will prove the results for an arbitrary $a \in \{0, 1\}$. Due to Lemma 6.3 and our projection operation Π on the empirical estimator, α_a and $\hat{\alpha}_a^+$ are both restricted to the interval $[0, 1]$. Using this, note that in our domain of $\lambda \in [0, \lambda_{max}]$ or $\lambda \in [-\lambda_{min}, 0]$, $\alpha_a = \Pi(\alpha_a)$, which is a useful property we exploit throughout the proof.

(1) **Consistency**: By the weak law of large numbers, $\hat{\pi}_{1a} \xrightarrow{p} \pi_{1a}$. Consider $\hat{\pi}_{1a}^+$. Since $\pi_{1a} > 0$ by assumption, we can pick a sufficiently large N_α such that $1/N_\alpha \leqslant \pi_{1a}/2$. Therefore the event $\{\hat{\pi}_{1a} < 1/N_\alpha\}$ is contained in the event $\{\hat{\pi}_{1a} < \pi_{1a}/2\}$ which is further contained in $\{|\hat{\pi}_{1a} - \pi_{1a}| \geqslant \pi_{1a}/2\}$. Since $\hat{\pi}_{1a} \xrightarrow{p} \pi_{1a}$, $\Pr\left(|\hat{\pi}_{1a} - \pi_{1a}| \geqslant \pi_{1a}/2\right) \to 0$ and hence $\Pr\left(\hat{\pi}_{1a}^+ \neq \hat{\pi}_{1a}\right) \to 0$, and hence $\hat{\pi}_{1a}^+ \xrightarrow{p} \pi_{1a}$.

Now, we will show consistency in two steps. First, let the denominator be defined as $\tilde{d}_a = 2 - \frac{(2a-1)\lambda}{\hat{\pi}_{1a}^+}$ and $d_a = 2 - \frac{(2a-1)\lambda}{\pi_{1a}}$. Since $\pi_{1a} > 0$ by assumption, and $\hat{\pi}_{1a}^+ \xrightarrow{p} \pi_{1a}$, we can use the Continuous Mapping Theorem to claim that $\tilde{d}_a \xrightarrow{p} d_a$.

Again, since $d_a \geqslant \tau > 0$, then by application of the Continuous Mapping Theorem, $\frac{1}{\tilde{d}_a} \xrightarrow{p} \frac{1}{d_a}$ and hence $\hat{\alpha}_a \xrightarrow{p} \alpha_a$ (as the truncation operation Π is also continuous).

(2) Bias: We want to now bound the bias of the estimator: $|\alpha_a - \underset{D_\alpha}{E}[\hat{\alpha}_a]|$:

$$\begin{aligned}|\alpha_a - \underset{D_\alpha}{E}[\hat{\alpha}_a]| = |\underset{D_\alpha}{E}[\alpha_a - \hat{\alpha}_a]| &\leqslant \underset{D_\alpha}{E}[|\alpha_a - \hat{\alpha}_a|] = \underset{D_\alpha}{E}[|\Pi(\alpha_a) - \Pi(\tilde{\alpha}_a)|] \\ &\leqslant \underset{D_\alpha}{E}[|\alpha_a - \tilde{\alpha}_a|] \quad \text{(Since } \Pi \text{ is Lipschitz)} \\ &= \underset{D_\alpha}{E}\left[\left|\frac{1}{d_a} - \frac{1}{\tilde{d}_a}\right|\right] = \underset{D_\alpha}{E}\left[\left|\frac{1}{d_a} - \frac{1}{\tilde{d}_a}\right| \mathbb{1}_{\mathcal{G}}\right] + \underset{D_\alpha}{E}\left[\left|\frac{1}{d_a} - \frac{1}{\tilde{d}_a}\right| \mathbb{1}_{\mathcal{G}^C}\right],\end{aligned}$$

where we let $\mathcal{G} = \{|\hat{\pi}_{1a} - \pi_{1a}| \leqslant \epsilon\}$ denote the good event that $\hat{\pi}_{1a}$ is close to π_{1a}. On this event, for sufficiently large N_α ($N_\alpha > \frac{2}{\pi_{1a}}$), $\hat{\pi}^+_{1a} = \hat{\pi}_{1a}$ and hence $\hat{\alpha}^+_a = \hat{\alpha}_a$. We can set $\epsilon \leqslant \pi_{1a}/2$ so that $\hat{\pi}_{1a} \geqslant \pi_{1a}/2$.

On $\mathcal{G}$:

$$|d_a - \tilde{d}_a| \leqslant |\lambda| \left| \frac{\hat{\pi}_{1a} - \pi_{1a}}{\pi_{1a}\hat{\pi}_{1a}} \right| \leqslant \frac{2|\lambda|}{\pi^2_{1a}} |\hat{\pi}_{1a} - \pi_{1a}|$$

Similarly, we can also bound $\tilde{d}_a$ on $\mathcal{G}$:

$$\tilde{d}_a \geqslant d_a - |\tilde{d}_a - d_a| \geqslant \tau - \frac{2|\lambda|\epsilon}{\pi^2_{1a}}.$$

We can ensure that $\tilde{d}_a \geqslant \tau/2$ on $\mathcal{G}$, which means $\epsilon \leqslant \min\{\frac{\pi_{1a}}{2}, \frac{\tau\pi^2_{1a}}{4|\lambda|}\}$. Finally, for the first term of the bias bound, on the good event $\mathcal{G}$:

$$\mathop{E}_{D_\alpha}\left[\left|\frac{1}{d_a} - \frac{1}{\tilde{d}_a}\right| \mathbb{1}_{\mathcal{G}}\right] = \mathop{E}_{D_\alpha}\left[\frac{|\tilde{d}_a - d_a|}{d_a \tilde{d}_a} \mathbb{1}_{\mathcal{G}}\right] | \leqslant \frac{2}{\tau^2} \cdot \frac{2|\lambda|}{\pi^2_{1a}} \mathop{E}_{D_\alpha}\left[|\hat{\pi}_{1a} - \pi_{1a}|\right].$$

On $\mathcal{G}^c$, we can use the fact that both $\alpha_a, \hat{\alpha}^+_a \in [0, 1]$:

$$\begin{aligned}
&\mathop{E}_{D_\alpha}\left[|\alpha_a - \hat{\alpha}^+_a| \mathbb{1}_{\mathcal{G}^c}\right] \leqslant \mathop{E}_{D_\alpha}\left[1 \cdot \mathbb{1}_{\mathcal{G}^c}\right] \\
&= \Pr\left(\mathcal{G}^C\right) \leqslant 2\exp\left(-2\epsilon^2 N_\alpha\right) \text{ (Using Hoeffding's inequality on } \Pr\left(\mathcal{G}^c\right)).
\end{aligned}$$

Overall,

$$|\alpha_a - \mathop{E}_{D_\alpha}\left[\hat{\alpha}^+_a\right]| \leqslant \frac{4|\lambda|}{\tau^2\pi^2_{1a}} \mathop{E}_{D_\alpha}\left[|\hat{\pi}_{1a} - \pi_{1a}|\right] + 2\exp\left(-2\epsilon^2 N_\alpha\right), \text{ where } \epsilon \leqslant \min\left\{\frac{\pi_{1a}}{2}, \frac{\tau\pi^2_{1a}}{4|\lambda|}\right\}.$$

To further simplify this, using Lemma 6.5, we know that $\mathrm{E}\left[|\pi_{1a} - \hat{\pi}_{1a}|\right] \leqslant \sqrt{\mathrm{Var}(\hat{\pi}_{1a})} = \sqrt{\frac{\pi_{1a}(1-\pi_{1a})}{N_\alpha}}$, and hence:

$$|\alpha_a - \mathop{E}_{D_\alpha}\left[\hat{\alpha}^+_a\right]| \leqslant \frac{4|\lambda|}{\tau^2\pi^2_{1a}} \sqrt{\frac{\pi_{1a}(1-\pi_{1a})}{N_\alpha}} + 2\exp\left(-2\epsilon^2 N_\alpha\right), \text{ where } \epsilon \leqslant \min\left\{\frac{\pi_{1a}}{2}, \frac{\tau\pi^2_{1a}}{4|\lambda|}\right\}.$$

(3) **Sample Complexity**: For sample complexity, we can proceed by returning a high probability bound on the good event derived in the previous part. Note that $\{|\hat{\alpha}^+_a - \alpha_a| > \epsilon\} \subseteq \{|\tilde{\alpha}^+_a - \alpha_a| > \epsilon\}$.

Recall that $\mathcal{G} = \{|\hat{\pi}_{1a} - \pi_{1a}| \leqslant \epsilon\}$ denotes the good event that $\hat{\pi}_{1a}$ is close to π_{1a}. We can set $\epsilon \leqslant \pi_{1a}/2$ so that $\hat{\pi}_{1a} \geqslant \pi_{1a}/2$, whenever $1/N_\alpha \leqslant \pi_{1a}/2 \implies N_\alpha \geqslant 2/\pi_{1a}$.

From the previous part, we know that on $\mathcal{G}$:

- $|d_a - \tilde{d}_a| \leqslant \frac{2|\lambda|\epsilon}{\pi^2_{1a}}$,

- $\tilde{d}_a \geqslant \tau/2$ for $\epsilon \leqslant \min\{\frac{\pi_{1a}}{2}, \frac{\tau\pi_{1a}^2}{4|\lambda|}\}$, and
- Using both of the results above, $|\alpha_a - \tilde{\alpha}_a| \leqslant \frac{4|\lambda|\epsilon}{\tau^2\pi_{1a}^2}$ on $\mathcal{G}$.

Finally, if we want $|\tilde{\alpha}_a - \alpha_a| \leqslant \epsilon'$, then $\frac{4|\lambda|\epsilon}{\tau^2\pi_{1a}^2} \leqslant \epsilon' \implies \epsilon \leqslant \frac{\epsilon'\tau^2\pi_{1a}^2}{4|\lambda|}$. Combining this with the previous min bound on ϵ, we get $\epsilon = \min\left\{\frac{\pi_{1a}}{2}, \frac{\tau\pi_{1a}^2}{4|\lambda|}, \frac{\epsilon'\tau^2\pi_{1a}^2}{4|\lambda|}\right\}$.

We have now shown that $\{|\hat{\pi}_{1a} - \pi_{1a}| \leqslant \epsilon\} \subseteq \{|\alpha_a - \hat{\alpha}_a| \leqslant \epsilon'\}$. Taking complements and probabilities on both sides, we get:

$$\Pr\left(|\alpha_a - \hat{\alpha}_a| > \epsilon'\right) \leqslant \Pr\left(|\hat{\pi}_{1a} - \pi_{1a}| > \epsilon\right) \leqslant 2\exp\left(-2N_\alpha\epsilon^2\right) \leqslant \delta,$$

which gives us $N_\alpha \geqslant \frac{1}{2\epsilon^2}\log\left(\frac{2}{\delta}\right)$ for $\epsilon = \min\left\{\frac{\pi_{1a}}{2}, \frac{\tau\pi_{1a}^2}{4|\lambda|}, \frac{\epsilon'\tau^2\pi_{1a}^2}{4|\lambda|}\right\}$.

Refactoring notations (using t for ϵ and ϵ for ϵ'), with $1-\delta$ probability over the choice of data sample N_α, as long as $N_\alpha \geqslant \frac{1}{2t^2}\log\left(\frac{2}{\delta}\right)$ for $t = \min\left\{\frac{\pi_{1a}}{2}, \frac{\tau\pi_{1a}^2}{4|\lambda|}, \frac{\epsilon\tau^2\pi_{1a}^2}{4|\lambda|}\right\}$, $\Pr\left(|\hat{\alpha}_a - \alpha_a| > \epsilon\right) \leqslant \delta$.

□

Proof. (Proof of Theorem 6.2) (1) **Bayes Risk**: We will adapt the standard proof technique of the Bayes error for cost-sensitive settings [4]. The expected α-cost sensitive risk for the classifier h^* can be expressed as:

$$\begin{aligned}
R_\alpha(h^*) &= \sum_{a\in\mathcal{A}}\left((1-\alpha_a)\cdot \mathop{E}_{X,Y=1,A=a}[l_{01}(h^*(X,a),1)] + \alpha_a \cdot \mathop{E}_{X,Y=0,A=a}[l_{01}(h^*(X,a),0)]\right)\\
&= \sum_{a\in\mathcal{A}} \mathop{E}_{X,a}\left[\mathop{E}_{Y=1|X,a}[(1-\alpha_a)\mathbb{I}\left(Y=1, h^*(X,a)=0\right)] + \mathop{E}_{Y=0|X,a}[\alpha_a\mathbb{I}\left(Y=0, h^*(X,a)=1\right)]\right]\\
&= \sum_{a\in\mathcal{A}} \mathop{E}_{X,a}[(1-\alpha_a)\mathbb{I}\left(\eta(X,a)<\alpha_a\right)\eta(X,a) + \alpha_a\mathbb{I}\left(\eta\geqslant\alpha_a\right)(1-\eta(X,a))]
\end{aligned}$$

At this point, note that $\eta < \alpha_a \implies (1-\alpha_a)\eta < \alpha_a(1-\eta)$ and similarly $\eta \geqslant \alpha_a \implies \alpha_a(1-\eta) \leqslant (1-\alpha_a)\eta$. Therefore,

$$\begin{aligned}
R_\alpha(h^*) &= \sum_{a\in\mathcal{A}} \mathop{E}_{X,a}[(1-\alpha_a)\mathbb{I}\left(\eta(X,a)<\alpha_a\right)\eta(X,a) + \alpha_a\mathbb{I}\left(\eta\geqslant\alpha_a\right)(1-\eta(X,a))]\\
&= \sum_{a\in\mathcal{A}} E_{X,a}[\mathbb{I}\left(\eta(X,a)<\alpha_a\right)\min((1-\alpha_a)\eta(X,a), \alpha_a(1-\eta(X,a)))\\
&\qquad + \mathbb{I}\left(\eta(X,a)\geqslant\alpha_a\right)\min(\alpha_a(1-\eta(X,a)), \eta(X,a)(1-\alpha_a))]\\
&= \mathop{E}_{X,A}[\min((1-\alpha_A)\eta(X,A), \alpha_A(1-\eta(X,A)))]\,.
\end{aligned}$$

It remains to show that h^* is the Bayes optimal classifier. Let h' be any arbitrary classifier. Then,

[4] ocw.mit.edu

$$
\begin{aligned}
R_\alpha(h') - R_\alpha(h^*) &= \mathop{E}_{X,a}\left[(1-\alpha_a)\eta(X,a)\mathbb{I}\left(h'(X,a)=0\right) + \alpha_a(1-\eta(X,a))\mathbb{I}\left(h'(X,a)=1\right)\right] \\
&\quad - \mathop{E}_{X,a}\left[(1-\alpha_a)\eta(X,a)\mathbb{I}\left(h^*(X,a)=0\right) + \alpha_a(1-\eta(X,a))\mathbb{I}\left(h^*(X,a)=1\right)\right] \\
&= E_{X,a}[(\mathbb{I}\left(h'=0\right) - \mathbb{I}\left(h^*=0\right))(1-\alpha_a)\eta(X,a) \\
&\qquad\qquad + (\mathbb{I}\left(h'=1\right) - \mathbb{I}\left(h^*=1\right))\alpha_a(1-\eta(X,a))] \\
&= \mathop{E}_{X,a}\left[(\mathbb{I}\left(h'=0\right) - \mathbb{I}\left(h^*=0\right))(\eta(X,a)-\alpha_a)\right] \\
&= \mathop{E}_{X,a}\left[\mathbb{I}\left(h'\neq h^*\right)\cdot(\eta(X,a)-\alpha_a)\cdot \operatorname{sgn}(\eta(X,a)-\alpha_a)\right], \\
&\qquad\qquad \text{using the fact that } h^* = \mathbb{I}\left(\eta \geqslant \alpha_a\right) \\
&= \mathop{E}_{X,a}\left[\mathbb{I}\left(h'\neq h^*\right)\cdot|\eta(X,a)-\alpha_a|\right] \geqslant 0.
\end{aligned}
$$

(2) **Consistency**:

$$
\begin{aligned}
\mathop{E}_{(c_i,a_i)_{i=1}^{N_\beta}}\left[\hat{\beta}_{CS}\right] &= \mathop{E}_{(c_i,a_i)_{i=1}^{N_\beta}}\left[\frac{1}{N_\beta}\sum_{i\in[N_\beta]}\min((1-\alpha_i)c_i, \alpha_i(1-c_i))\right] \\
&= \frac{1}{N_\beta}\sum_{i\in[N_\beta]}\mathop{E}_{(c_i,a_i)\sim D}\left[\min((1-\alpha_i)c_i, \alpha_i(1-c_i))\right] \\
&= \beta_{CS}.
\end{aligned}
$$

By the weak law of large numbers, $\hat{\beta}_{CS} \xrightarrow{p} \beta_{CS}$.

(3) **Sample Complexity**: The proof for sample complexity is a straightforward application of Hoeffding's inequality, by noting that $\min((1-\alpha_i)c_i, \alpha_i(1-c_i)) \in [0, 1/4]$:

$$
\begin{aligned}
&\Pr\left(|\hat{\beta}_{CS} - \beta_{CS}| \geqslant \epsilon\right) \leqslant 2\exp\left(\frac{-2N_\beta\epsilon^2}{(1/4)^2}\right) \leqslant \delta \\
&\implies \epsilon \geqslant \frac{1}{4}\sqrt{\frac{1}{2N_\beta}\log\left(\frac{2}{\delta}\right)}.
\end{aligned}
$$

□

Proof. (Proof of Theorem 6.3) (1) **Consistency**: For showing consistency, note that conditioned on D_α, our thresholds $\hat{\alpha}_a^+$ are fixed, and hence by Theorem 6.2, $\hat{\beta}_{CS,\hat{\alpha}} \xrightarrow{p} \beta_{CS,\hat{\alpha}}$, where $\beta_{CS,\hat{\alpha}}$ is the cost sensitive Bayes risk with respect to group-aware thresholds $\hat{\alpha}_a^+, \forall a \in \mathcal{A}$. It remains to show that $\beta_{CS,\hat{\alpha}} \xrightarrow{p} \beta_{CS}$.

From Theorem 6.1, $\hat{\alpha}_a^+ \xrightarrow{p} \alpha_a$. Note that $f(x) = E\left[\min\{(1-x)\eta, x(1-\eta)\}\right]$ is a continuous function of x. Hence, by the Continuous Mapping Theorem, $\beta_{CS,\hat{\alpha}} \xrightarrow{p} \beta_{CS}$.

(2) **Bias**:

$$
\begin{aligned}
|\beta_{CS} - \mathop{E}_{D_\beta, D_\alpha}\left[\hat{\beta}_{CS,\hat{\alpha}}\right]| &= |\mathop{E}_{D_\beta, D_\alpha}\left[\beta_{CS} - \hat{\beta}_{CS,\hat{\alpha}}\right]| \\
&= |\mathop{E}_{D_\alpha}\left[\mathop{E}_{D_\beta}\left[(\beta_{CS} - \hat{\beta}_{CS,\hat{\alpha}})|D_\alpha\right]\right]| = |\mathop{E}_{D_\alpha}\left[\beta_{CS} - \beta_{CS,\hat{\alpha}}\right]| \\
&\leqslant \mathop{E}_{D_\alpha}\left[|\beta_{CS} - \beta_{CS,\hat{\alpha}}|\right].
\end{aligned}
$$

At this point, consider the following:

$$
\begin{aligned}
|\beta_{CS,x} - \beta_{CS,y}| &= |\mathop{E}_{X,A}\left[\min(\eta(1-x), x(1-\eta)) - \min(\eta(1-y), y(1-\eta))\right]| \\
&= |E\left[\frac{\eta + x}{2} - x\eta - |\frac{x-\eta}{2}| - \frac{\eta+y}{2} + y\eta + |\frac{y-\eta}{2}|\right]| \\
&\text{(Using } \min(a,b) = \frac{a+b}{2} - |\frac{a-b}{2}|) \\
&= \frac{1}{2}|E\left[(x-y)(1-2\eta) - |\eta - x| + |\eta - y|\right]| \\
&= \frac{1}{2}|(x-y)(1-2E\left[\eta\right]) + E\left[|\eta - y| - |\eta - x|\right]| \\
&\leqslant \frac{1}{2}|x-y| \cdot |1 - 2E\left[\eta\right]| + \frac{1}{2}E\left[||\eta - y| - |\eta - x||\right] \\
&\leqslant \frac{1}{2}|x-y| \cdot |1 - 2E\left[\eta\right]| + \frac{1}{2}E\left[|x-y|\right] \text{(Using Reverse Triangle Inequality)} \\
&= \frac{1}{2}|1 - 2E\left[\eta\right]||x-y| + \frac{1}{2}|x-y| = \frac{|x-y|}{2}(1 + |1-2\pi|) \\
&\leqslant |x-y| \text{(Since } \pi \in [0,1]).
\end{aligned}
$$

The above result shows that the Bayes risk is Lipschitz in the threshold argument α with $L = 1$. This result will be useful later.

We know from Theorem 6.1 that the bias of $\hat{\alpha}_a^+$ is bounded: $|E\left[\hat{\alpha}_a^+\right] - \alpha_a| \leqslant U_m$, where U_m depends on the fairness metric. Therefore, setting $x = \mathop{E}_{D_\alpha}\left[\hat{\alpha}\right]$ and $y = \alpha_a$ yields us:

$$
\begin{aligned}
|\beta_{CS} - \mathop{E}_{D_\beta, D_\alpha}\left[\hat{\beta}_{CS,\hat{\alpha}}\right]| &\leqslant \mathop{E}_{D_\alpha}\left[|\beta_{CS} - \beta_{CS,\hat{\alpha}}|\right] \\
&\leqslant \pi_0 \mathop{E}_{D_\alpha}\left[|\beta_{CS} - \beta_{CS,\hat{\alpha}_0^+}|\right] + \pi_1 \mathop{E}_{D_\alpha}\left[|\beta_{CS} - \beta_{CS,\hat{\alpha}_1^+}|\right] \\
&\leqslant \pi_0 \mathop{E}_{D_\alpha}\left[|\alpha_0 - \hat{\alpha}_0^+|\right] + \pi_1 \mathop{E}_{D_\alpha}\left[|\alpha_1 - \hat{\alpha}_1^+|\right]
\end{aligned}
$$

From Theorem 6.1, we know that with $1 - \delta$ probability, $|\hat{\alpha}_a^+ - \alpha_a| < \epsilon$. Using this, we get:

$$
\begin{aligned}
|\beta_{CS} - \mathop{E}_{D_\beta, D_\alpha}\left[\hat{\beta}_{CS,\hat{\alpha}}\right]| &\leqslant \pi_0 \mathop{E}_{D_\alpha}\left[|\alpha_0 - \hat{\alpha}_0^+|\right] + \pi_1 \mathop{E}_{D_\alpha}\left[|\alpha_1 - \hat{\alpha}_1^+|\right] \\
&= \sum_{a\in\{0,1\}} \pi_a \left(\mathop{E}_{D_\alpha}\left[|\alpha_a - \hat{\alpha}_a^+| \mathbb{I}\left(|\alpha_a - \hat{\alpha}_a^+| < \epsilon\right)\right] + \mathop{E}_{D_\alpha}\left[|\alpha_a - \hat{\alpha}_a^+| \mathbb{I}\left(|\alpha_a - \hat{\alpha}_a^+| \geqslant \epsilon\right)\right] \right) \\
&\leqslant \sum_{a\in\{0,1\}} \pi_a \left(\epsilon \cdot \Pr\left(|\alpha_a - \hat{\alpha}_a^+| < \epsilon\right) + 1 \cdot \Pr\left(|\alpha_a - \hat{\alpha}_a^+| \geqslant \epsilon\right)\right) \quad \text{(Since } |\alpha_a - \hat{\alpha}_a^+| \leqslant 1\text{)} \\
&\leqslant \sum_{a\in\{0,1\}} \pi_a(\epsilon_a(N_\alpha, \delta_a) + \delta_a).
\end{aligned}
$$

where, $\epsilon_a(N_\alpha, \delta_a)$ can de derived from the respective sample complexities in Theorem 6.1.

For DP, using the dominant term, $\epsilon_a = \frac{|\lambda|}{\pi_a^2}\sqrt{\frac{\log(4/\delta_a)}{2N_\alpha}}$ and setting $\delta = \max\{\delta_0, \delta_1\}$, giving us:

$$
|\beta_{CS} - \mathop{E}_{D_\beta, D_\alpha}\left[\hat{\beta}_{CS,\hat{\alpha}}\right]| = \mathcal{O}\left(|\lambda| \cdot \sqrt{\frac{\log(1/\delta)}{N_\alpha}}\right).
$$

For EO, using the ϵ dependent term, $\epsilon_a = \frac{|\lambda|}{\tau^2 \pi_{1a}^2}\sqrt{\frac{8\log(2/\delta_a)}{N_\alpha}}$ and setting $\delta = \max\{\delta_0, \delta_1\}$, giving us:

$$
|\beta_{CS} - \mathop{E}_{D_\beta, D_\alpha}\left[\hat{\beta}_{CS,\hat{\alpha}}\right]| = \mathcal{O}\left(\frac{|\lambda|}{\tau^2}\sqrt{\frac{\log(1/\delta)}{N_\alpha}}\right).
$$

(3) **Sample Complexity**: We want to ensure the following:

$$
\Pr\left(|\beta_{CS} - \hat{\beta}_{CS,\hat{\alpha}}| > \epsilon\right) \leqslant \delta.
$$

To do so, consider the following breakdown: $|\beta_{CS} - \hat{\beta}_{CS,\hat{\alpha}}| \leqslant |\beta_{CS} - \beta_{CS,\hat{\alpha}}| + |\beta_{CS,\hat{\alpha}} - \hat{\beta}_{CS,\hat{\alpha}}|$.

Consider the random error: $|\beta_{CS,\hat{\alpha}} - \hat{\beta}_{CS,\hat{\alpha}}|$. This will give us a sample complexity lower bound for N_β. Bounding this is a simple application of Hoeffding's inequality with $X_i = \min\{(1-\hat{\alpha}_{a_i}^+)c_i, \hat{\alpha}_{a_i}^+(1-c_i)\} \in [0, 1/4]$:

$$
\Pr\left(|\beta_{CS,\hat{\alpha}_a} - \hat{\beta}_{CS,\hat{\alpha}}| > \epsilon/2\right) \leqslant 2\exp\left(-8N_\beta\epsilon^2\right) \leqslant \delta/2 \implies N_\beta \geqslant \frac{1}{8\epsilon^2}\log\left(\frac{4}{\delta}\right)
$$

Now consider the error due to threshold estimation. We know from our analysis of bias that $|\beta_{CS,x} - \beta_{CS,y}| \leqslant |x-y| \implies |\beta_{CS} - \beta_{CS,\hat{\alpha}}| \leqslant \sum_{a\in\{0,1\}} \pi_a \cdot |\hat{\alpha}_a^+ - \alpha_a|$.

From the sample complexity results in Theorem 6.1, we can derive the following: with $\Pr\left(|\hat{\alpha}_a^+ - \alpha_a| > \epsilon/2\right) \leqslant U_m(\epsilon/2) \leqslant \delta/2$, where U_m depends on the fairness metric we choose. This will provide us with a bound on $N_\alpha \geqslant N_\alpha(m)$.

For demographic parity, again using the dominant term:

$$\Pr\left(|\hat{\alpha}_a^+ - \alpha_a| > \epsilon/4\pi_a\right) \leqslant \delta/4 \implies N_\alpha \geqslant \frac{8\lambda^2}{\epsilon^2\pi_a^2}\log\left(\frac{16}{\delta}\right) = \mathcal{O}\left(\frac{\lambda^2}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right),$$

For equal opportunity, recall that with $1-\delta$ probability over the choice of data sample N_α, as long as $N_\alpha \geqslant \frac{1}{2t^2}\log\left(\frac{2}{\delta}\right)$ for $t = \min\left\{\frac{\pi_{1a}}{2}, \frac{\tau\pi_{1a}^2}{4|\lambda|}, \frac{\epsilon\tau^2\pi_{1a}^2}{4|\lambda|}\right\}$, $\Pr\left(|\hat{\alpha}_a - \alpha_a| > \epsilon\right) \leqslant \delta$. Using these bounds with our new ϵ and δ we get:

$$\begin{aligned}\Pr\left(|\hat{\alpha}_a^+ - \alpha_a| > \epsilon/4\pi_a\right) \leqslant \delta/4 \implies N_\alpha &\geqslant \frac{1}{2}\log\left(\frac{8}{\delta}\right)\max\left\{\frac{2}{\pi_{1a}}, \frac{4|\lambda|}{\tau\pi_{1a}^2}, \frac{4|\lambda|}{\epsilon\tau^2\pi_{1a}^2}\right\}^2 \\ &= \mathcal{O}\left(\frac{\lambda^2}{\tau^4\epsilon^2}\log\left(\frac{1}{\delta}\right)\right).\end{aligned}$$

□

Proof. (Proof of Theorem 6.4)

To prove our results, we will use an intermediate estimator $R(\hat{h}) = \underset{X,A}{E}\left[(1-2C)\mathbb{I}\left(C \geqslant \hat{\alpha}_A^+\right)\right] + \pi$.

1. **Consistency**: We want to show the following result:

$$\Pr\left(|R(h) - \hat{R}(\hat{h})| > \epsilon\right) \xrightarrow{p} 0 \text{ as } N_\alpha, N_\beta \to \infty.$$

We can decompose the consistency gap as follows:

$$|R(h) - \hat{R}(\hat{h})| \leqslant |R(h) - R(\hat{h})| + |R(\hat{h}) - \hat{R}(\hat{h})|$$

Since D_α and D_β are independently drawn samples, conditioned on D_α, $\hat{\alpha}_a^+$ is a constant for all $A = a$ and hence $\hat{R}(\hat{h}) \xrightarrow{p} R(\hat{h})$ by the weak law of large numbers. It remains to show how $R(\hat{h}) \xrightarrow{p} R(h)$.

Let $\phi_a(t) = \underset{X|A=a}{E}\left[(1-2C)\mathbb{I}\left(C \geqslant t\right)|A = a\right], t \in (0,1)$, so that $R(h) = \sum_{a\in\{0,1\}} \pi_a \cdot \phi_a(\alpha_a)$. Let $0 \leqslant u < v \leqslant 1$ and consider the following analysis for a fixed $A = a$:

$$\begin{aligned}|\phi_a(u) - \phi_a(v)| &= \left|E\left[(1-2C)(\mathbb{I}\left(C \geqslant u\right) - \mathbb{I}\left(C \geqslant v\right))|A = a\right]\right| \\ &\leqslant \Pr\left(u \leqslant C < v|A = a\right) \text{ (Since } (\mathbb{I}\left(C \geqslant u\right) - \mathbb{I}\left(C \geqslant v\right)) \neq 0 \text{ when } C \in [u, v)) \\ &= |F_a(v) - F_a(u)|\end{aligned}$$

From Theorem 6.1, we know that $\hat{\alpha}_a^+ \xrightarrow{p} \alpha_a$ and since F_a is continuous due to Assumption 2.1 and $\alpha_a > 0$, by the continuous mapping theorem, $\phi(\hat{\alpha}_a^+) \xrightarrow{p} \phi(\alpha_a)$. Since $R(h) = \sum_{a\in\{0,1\}} \pi_a \cdot \phi_a(\alpha_a)$, $R(\hat{h}) \xrightarrow{p} R(h)$.

2. **Bias**: We want to bound the following quantity: $|R(h) - \underset{D_\alpha, D_\beta}{E}\left[\hat{R}(\hat{h})\right]|$. Let $R_0(h) = R(h) - \pi$, and similarly define $R_0(\hat{h}) = R(\hat{h}) - \pi$.

Note that since D_α and D_β are independently sampled, $\mathop{E}_{D_\alpha, D_\beta}\left[\hat{R}(\hat{h})\right] = \mathop{E}_{D_\alpha}\left[R_0(\hat{h})\right] + \mathop{E}_{D_\alpha}[\hat{\pi}]$, which means:

$$\begin{aligned}|R(h) - \mathop{E}_{D_\alpha, D_\beta}\left[\hat{R}(\hat{h})\right]| &\leqslant |R_0(h) - \mathop{E}_{D_\alpha}\left[R_0(\hat{h})\right]| + |\pi - \mathop{E}_{D_\alpha}[\hat{\pi}]| \\ &= |R_0(h) - \mathop{E}_{D_\alpha}\left[R_0(\hat{h})\right]| \text{ (Since } \hat{\pi} \text{ is an unbiased estimate of } \pi)\end{aligned}$$

Let $\phi_a(t) = \mathop{E}_{X|A=a}[(1-2C)\mathbb{I}(C \geqslant t)|A = a]$. From our consistency proof, we know that $|\phi_a(u) - \phi_a(v)| \leqslant |F_a(v) - F_a(u)|$. Using this:

$$\begin{aligned}&|R(h) - \mathop{E}_{D_\alpha, D_\beta}\left[\hat{R}(\hat{h})\right]| \leqslant |R_0(h) - \mathop{E}_{D_\alpha}\left[R_0(\hat{h})\right]| = |\sum_{a\in\{0,1\}} \pi_a \mathop{E}_{D_\alpha}\left[\phi_a(\alpha_a) - \phi_a(\hat{\alpha}_a^+)\right]| \\ &\leqslant \sum_{a\in\{0,1\}} \pi_a \mathop{E}_{D_\alpha}\left[|\phi_a(\alpha_a) - \phi_a(\hat{\alpha}_a^+)|\right] \leqslant \sum_{a\in\{0,1\}} \pi_a L_a \mathop{E}_{D_\alpha}\left[|\alpha_a - \hat{\alpha}_a^+|\right] \text{ (Using Assumption 6.1)}\end{aligned}$$

From Theorem 6.1, we know that with $1-\delta$ probability, $|\hat{\alpha}_a^+ - \alpha_a| < \epsilon$ We also know that $|\alpha_a - \hat{\alpha}_a^+| \leqslant 1$. Using this, we get:

$$\begin{aligned}&\mathop{E}_{D_\alpha}\left[|\alpha_a - \hat{\alpha}_a^+|\right] = \mathop{E}_{D_\alpha}\left[|\alpha_a - \hat{\alpha}_a^+|\mathbb{I}\left(|\alpha_a - \hat{\alpha}_a^+| < \epsilon\right)\right] + \mathop{E}_{D_\alpha}\left[|\alpha_a - \hat{\alpha}_a^+|\mathbb{I}\left(|\alpha_a - \hat{\alpha}_a^+| \geqslant \epsilon\right)\right] \\ &\leqslant \epsilon \cdot \Pr\left(|\alpha_a - \hat{\alpha}_a^+| < \epsilon\right) + 1 \cdot \Pr\left(|\alpha_a - \hat{\alpha}_a^+| \geqslant \epsilon\right) \text{ (Since } |\alpha_a - \hat{\alpha}_a^+| \leqslant 1) \leqslant \epsilon + \delta.\end{aligned}$$

and therefore,

$$|R(h) - \mathop{E}_{D_\alpha, D_\beta}\left[\hat{R}(\hat{h})\right]| \leqslant \sum_{a\in\{0,1\}} \pi_a L_a \mathop{E}_{D_\alpha}\left[|\alpha_a - \hat{\alpha}_a^+|\right] \leqslant \sum_{a\in\{0,1\}} \pi_a L_a(\epsilon_a + \delta_a).$$

For DP, using the sample complexity bound from Theorem 6.1 and the dominant term, $\epsilon_a = \frac{|\lambda|}{\pi_a^2}\sqrt{\frac{\log(4/\delta_a)}{2N_\alpha}}$, and setting $\delta = \min\{\delta_0, \delta_1\}$:

$$|R(h) - \mathop{E}_{D_\alpha, D_\beta}\left[\hat{R}(\hat{h})\right]| = \mathcal{O}\left(|\lambda| \cdot \max\{L_0, L_1\} \cdot \sqrt{\frac{\log(1/\delta)}{N_\alpha}}\right).$$

For EO, using the sample complexity bound from Theorem 6.1 and the ϵ dependent term, $\epsilon_a = \frac{|\lambda|}{\tau^2\pi_{1a}^2}\sqrt{\frac{8\log(4/\delta_a)}{N_\alpha}}$, and setting $\delta = \min\{\delta_0, \delta_1\}$:

$$|R(h) - \mathop{E}_{D_\alpha, D_\beta}\left[\hat{R}(\hat{h})\right]| = \mathcal{O}\left(\frac{|\lambda|\max\{L_0, L_1\}}{\tau^2}\sqrt{\frac{\log(1/\delta)}{N_\alpha}}\right).$$

(3) **Sample Complexity**: We want to ensure the following:

$$\Pr\left(|R(h) - \hat{R}(\hat{h})| > \epsilon\right) \leqslant \delta.$$

Note that we can decompose the absolute error as follows via the triangle inequality:

$$|R(h) - \hat{R}(\hat{h})| \leqslant |R_0(h) - R_0(\hat{h})| + |R_0(\hat{h}) - \hat{R}_0(\hat{h})| + |\hat{\pi} - \pi|.$$

Applying the union bound, we get:

$$\Pr\left(|R(h) - \hat{R}(\hat{h})| > \epsilon\right) \leqslant \Pr\left(|R_0(h) - R_0(\hat{h})| > \epsilon/3\right) + \Pr\left(|R_0(\hat{h}) - \hat{R}_0(\hat{h})| > \epsilon/3\right) + \Pr\left(|\hat{\pi} - \pi| > \epsilon/3\right).$$

Conditioned on D_α, $\hat{\alpha}_a^+$ is a constant and hence we can control $\Pr\left(|R_0(\hat{h}) - \hat{R}_0(\hat{h})| > \epsilon/3\right)$ via the Hoeffding's inequality. Note that the random variable $Z_i = (1 - 2c_i)\mathbb{I}\left(c_i \geqslant \hat{\alpha}_{a_i}^+\right) \in [-1, 1]$.

$$\Pr\left(|R_0(\hat{h}) - \hat{R}_0(\hat{h})| > \epsilon/3 | D_\alpha\right) \leqslant \delta/3 \implies N_\beta \geqslant \frac{18}{\epsilon^2} \log\left(\frac{6}{\delta}\right).$$

Similarly, we can also control $\Pr\left(|\hat{\pi} - \pi| > \epsilon/3\right)$:

$$\Pr\left(|\hat{\pi} - \pi| > \epsilon/3\right) \leqslant \delta/3 \implies N_\alpha \geqslant \frac{9}{2\epsilon^2} \log\left(\frac{6}{\delta}\right)$$

Finally, to control $\Pr\left(|R_0(h) - R_0(\hat{h})| > \epsilon/3\right)$, we can borrow our analysis from the previous proof of bias:

$$|R_0(h) - R_0(\hat{h})| \leqslant \sum_{a \in \{0,1\}} \pi_a L_a |\alpha_a - \hat{\alpha}_a^+|.$$

By union bound and splitting the error budget between each group-dependent term (since there are two groups, if the sum exceeds $\epsilon/3$, at least one term exceeds $\epsilon/6$):

$$\Pr\left(|R_0(h) - R_0(\hat{h})| > \epsilon/3\right) \leqslant \sum_{a \in \{0,1\}} \Pr\left(|\alpha_a - \hat{\alpha}_a^+| > \frac{\epsilon}{6\pi_a L_a}\right).$$

For demographic parity, again using the dominant term:

$$\Pr\left(|\alpha_a - \hat{\alpha}_a^+| > \frac{\epsilon}{6\pi_a L_a}\right) \leqslant \delta/6 \implies N_\alpha \geqslant \frac{18\lambda^2 L_a^2}{\epsilon^2 \pi_a^2} \log\left(\frac{24}{\delta}\right) = \mathcal{O}\left(\frac{\lambda^2 L_a^2}{\epsilon^2} \log\left(\frac{1}{\delta}\right)\right).$$

For equal opportunity, recall that with $1 - \delta$ probability over the choice of data sample N_α, as long as $N_\alpha \geqslant \frac{1}{2t^2} \log\left(\frac{2}{\delta}\right)$ for $t = \min\left\{\frac{\pi_{1a}}{2}, \frac{\tau\pi_{1a}^2}{4|\lambda|}, \frac{\epsilon\tau^2\pi_{1a}^2}{4|\lambda|}\right\}$, $\Pr\left(|\hat{\alpha}_a - \alpha_a| > \epsilon\right) \leqslant \delta$. Using these bounds with our new ϵ and δ we get:

$$\Pr\left(|\alpha_a - \hat{\alpha}_a^+| > \frac{\epsilon}{6\pi_a L_a}\right) \leqslant \delta/6 \implies N_\alpha \geqslant \frac{1}{2}\log\left(\frac{12}{\delta}\right)\max\left\{\frac{2}{\pi_{1a}}, \frac{4|\lambda|}{\tau\pi_{1a}^2}, \frac{24|\lambda|\pi_a L_a}{\epsilon\tau^2\pi_{1a}^2}\right\}^2$$
$$= \mathcal{O}\left(\frac{\lambda^2 L_a^2}{\epsilon^2\tau^4}\log\left(\frac{1}{\delta}\right)\right).$$

Taking a max over L_a completes our proof. □

For convenience and readability, we will prove Theorem 6.5 separately for DP and EO.

Theorem 6.8. Suppose a dataset D_β is used to construct $\hat{\Delta}_{DP}(h_{\hat{\alpha}_a^+})$, where $\hat{\alpha}_a^+$ is the estimator for the DP threshold (Theorem 6.1). Then the following holds:

1. Consistency: $\hat{\Delta}_{DP}(h_{\hat{\alpha}_a^+}) \xrightarrow{p} \Delta_{DP}(h_\alpha))$.
2. Bias: For $\delta \in (0,1)$, $|\Delta_{DP}(h) - \mathop{E}_{D_\beta, D_\alpha}\left[\hat{\Delta}_{DP}(\hat{h})\right]| = \mathcal{O}\left(\frac{1}{\sqrt{N_\beta}} + |\lambda| \cdot \max\{L_0, L_1\}\sqrt{\frac{\log(1/\delta_a)}{N_\alpha}}\right)$.
3. Sample Complexity: With $1-\delta$ probability, $|\hat{\Delta}(\hat{h}) - \Delta(h)| \leqslant \epsilon$, as long as $N_\beta = \mathcal{O}\left(\frac{1}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$ and $N_\alpha = \mathcal{O}\left(\frac{\lambda^2 \cdot \max\{L_0^2, L_1^2\}}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$.

Proof. For the sake of this proof, we will denote $\Delta_{DP}(h_\alpha)$ as $\Delta(h)$ and $\hat{\Delta}_{DP}(h_{\hat{\alpha}_a^+})$ as $\hat{\Delta}(\hat{h})$. Then an intermediate estimator $\hat{\Delta}_{DP}(h_{\alpha_a})$ can be denoted as $\hat{\Delta}(h)$. We similarly overload the notations in this fashion for the positive rate estimator $\rho_{PR,a}$ with ρ_a.

(1) **Consistency**: Most of the inequality arguments in the proof are a direct consequence of the triangle inequality, by introducing an intermediate estimate of the PR ($\hat{\rho}_a(h)$).

We want to show that $\Pr\left(|\hat{\Delta}(\hat{h}) - \Delta(h)| > \epsilon\right) \to 0$ as $N_\beta \to \infty$. By Slutsky's theorem, it will be enough to upper-bound our target event as a sum of simpler events and show that if each of them converges in probability to zero, then we can claim our final result.

Consider the absolute difference term and apply the reverse triangle inequality ($||A| - |B|| \leqslant |A - B|$):

$$\begin{aligned}|\Delta(h) - \hat{\Delta}(\hat{h})| &\leqslant |\rho_1(h) - \rho_0(h) - \hat{\rho}_1(\hat{h}) + \hat{\rho}_0(\hat{h})| \\ &\leqslant |\rho_1(h) - \hat{\rho}_1(h)| + |\rho_0(h) - \hat{\rho}_0(h)| + |\hat{\rho}_1(h) - \hat{\rho}_1(\hat{h})| + |\hat{\rho}_0(h) - \hat{\rho}_0(\hat{h})|.\end{aligned}$$

Since similar arguments will work for any group $A = a$, we will now study each term separately for an arbitrary group $A = a$.

Consider $|\rho_a(h) - \hat{\rho}_a(h)|$. By the Weak law of large numbers,
$\frac{1}{N_{\beta,a}}\sum_{i\in[N_{\beta,a}]}\mathbb{I}(c_i \geqslant \alpha_a) \xrightarrow{p} \mathop{E}_{X|A=a}[\mathbb{I}(c_a \geqslant \alpha_a)]$, and hence $|\rho_a(h) - \hat{\rho}_a(h)| \xrightarrow{p} 0$.

Next, consider the term $|\hat{\rho}_a(h) - \hat{\rho}_a(\hat{h})|$. We will now attempt to study the convergence in probability properties of this term:

$$|\hat{\rho}_a(h) - \hat{\rho}_a(\hat{h})| = \left|\frac{1}{N_{\beta,a}}\sum_{i\in[N_{\beta,a}]}\mathbb{I}(c_i \geqslant \alpha_a) - \frac{1}{N_{\beta,a}}\sum_{i\in[N_{\beta,a}]}\mathbb{I}(c_i \geqslant \hat{\alpha}_a^+)\right| \leqslant \frac{1}{N_{\beta,a}}\sum_{i\in[N_{\beta,a}]} Z_i,$$

where $Z_i = |\mathbb{I}(c_i \geqslant \alpha_a) - \mathbb{I}(c_i \geqslant \hat{\alpha}_a^+)| = \mathbb{I}(c_i \in (\min\{\hat{\alpha}_a^+, \alpha_a\}, \max\{\hat{\alpha}_a^+, \alpha_a\})$.

We will now show that the numerator converges to zero. Note that conditioned on D_α, $\hat{\alpha}_a^+$ is fixed. Since D_α and D_β are sampled independently, $\{Z_i, i \in [N_{\beta,a}]\}$ are iid and bounded.

Hence, by the weak law of large numbers,
$\frac{1}{N_{\beta,a}} \sum_{i\in[N_{\beta,a}]} Z_i \xrightarrow{p} \underset{D_{\beta,a}}{E} [Z_i | D_\alpha, A = a] = \underset{D_{\beta,a}}{E} [\mathbb{I}(c_i \in (\min\{\hat{\alpha}_a^+, \alpha_a\}, \max\{\hat{\alpha}_a^+, \alpha_a\})) | D_\alpha, A = a] =$
$\Pr(c \in (\min\{\hat{\alpha}_a^+, \alpha_a\}, \max\{\hat{\alpha}_a^+, \alpha_a\}) | D_\alpha, A = a) = |F_a(\alpha_a) - F_a(\hat{\alpha}_a^+)|$, where F_a is the group-conditional CDF of $c = \eta(x, a)$, and the absolute difference accounts for both scenarios ($\hat{\alpha}_a^+ > \alpha_a$ and $\hat{\alpha}_a^+ \leqslant \alpha_a$).

We are now ready to apply the Continuous Mapping Theorem, since F_a is continuous by Assumption 2.1. Since we ensure $\alpha_a \in (0,1)$, by CMT, $F_a(\hat{\alpha}_a^+) \xrightarrow{p} F_a(\alpha_a)$, since $\hat{\alpha}_a^+ \xrightarrow{p} \alpha_a$ (Theorem 6.1), and therefore $\frac{1}{N_{\beta,a}} \sum_{i\in[N_{\beta,a}]} Z_i \xrightarrow{p} 0$ and:

$$|\hat{\rho}_a(h) - \hat{\rho}_a(\hat{h})| \leqslant \frac{1}{N_{\beta,a}} \sum_{i\in[N_{\beta,a}]} Z_i \xrightarrow{p} |F_a(\alpha_a) - F_a(\hat{\alpha}_a^+)| \xrightarrow{p} 0.$$

With this, we now have that all our components of $|\Delta(h) - \hat{\Delta}(\hat{h})|$ converge to zero in probability and hence we can claim that $\Pr\left(|\hat{\Delta}_{DP}(\hat{h}) - \Delta_{DP}(h)| > \epsilon\right) \to 0$ as $N_\beta \to \infty$.

(2) **Bias**: We now want to bound the bias of our estimator, $|\Delta_{DP}(h) - \underset{D_\beta, D_\alpha}{E}\left[\hat{\Delta}_{DP}(\hat{h})\right]|$.

$$|\Delta_{DP}(h) - \underset{D_\beta, D_\alpha}{E}\left[\hat{\Delta}_{DP}(\hat{h})\right]| = |\underset{D_\beta, D_\alpha}{E}\left[\Delta_{DP}(h) - \hat{\Delta}_{DP}(\hat{h})\right]| \leqslant \underset{D_\beta, D_\alpha}{E}\left[|\hat{\Delta}_{DP}(\hat{h}) - \Delta_{DP}(h)|\right]$$
$$\leqslant \underset{D_\beta, D_\alpha}{E}\left[|\hat{\rho}_1(\hat{h}) - \hat{\rho}_0(\hat{h}) - \rho_1(h) + \rho_0(h)|\right]$$

(By using the Reverse Triangle Inequality)

$$\leqslant \underset{D_{\beta,1}, D_\alpha}{E}[|\rho_1(h) - \hat{\rho}_1(h)|] + \underset{D_{\beta,0}, D_\alpha}{E}[|\rho_0(h) - \hat{\rho}_0(h)|]$$
$$+ \underset{D_{\beta,1}, D_\alpha}{E}\left[|\hat{\rho}_1(h) - \hat{\rho}_1(\hat{h})|\right] + \underset{D_{\beta,0}, D_\alpha}{E}\left[|\hat{\rho}_0(h) - \hat{\rho}_0(\hat{h})|\right]$$

(By introducing an intermediate estimator $\hat{\rho}_a(h)$,
and noting that the expectation is over group-specific samples $D_{\beta.a}$).

Once again, we will only focus on deriving the statistical bias for an arbitrary group $A = a$. Consider the first term. At this point, note that $\hat{\rho}_a(h)$ is an unbiased estimate of $\rho_a(h)$, and hence by Lemma 6.5,

$$\underset{D_{\beta,a}, D_\alpha}{E}[|\hat{\rho}_a(h) - \rho_a(h)|] \leqslant \sqrt{Var(\hat{\rho}_a(h))} \leqslant \frac{1}{2\sqrt{N_{\beta,a}}},$$

where the last step follows due to Popoviciu's Inequality on $\mathbb{I}(c_i \geqslant \alpha_a)$.
We can now focus on the second term: $\underset{D_{\beta,a}, D_\alpha}{E}\left[|\hat{\rho}_a(h) - \hat{\rho}_a(\hat{h})|\right]$.

$$\underset{D_{\beta,a}, D_\alpha}{E}\left[|\hat{\rho}_a(\hat{h}) - \hat{\rho}_a(h)|\right] = \underset{D_{\beta,a}, D_\alpha}{E}\left[\left|\frac{1}{N_{\beta,a}} \sum_{i\in[N_{\beta,a}]} \left(\mathbb{I}\left(c_i \geqslant \hat{\alpha}_a^+\right) - \mathbb{I}(c_i \geqslant \alpha_a)\right)\right|\right]$$
$$\leqslant \underset{D_{\beta,a}, D_\alpha}{E}\left[\frac{1}{N_{\beta,a}} \sum_{i\in[N_{\beta,a}]} \left|\mathbb{I}\left(c_i \geqslant \hat{\alpha}_a^+\right) - \mathbb{I}(c_i \geqslant \alpha_a)\right|\right]$$

From our consistency analysis,
we know that $Z_i = |\mathbb{I}(c_i \geqslant \alpha_a) - \mathbb{I}(c_i \geqslant \hat{\alpha}_a^+)| = \mathbb{I}(c_i \in (\min\{\hat{\alpha}_a^+, \alpha_A\}, \max\{\hat{\alpha}_a^+, \alpha_a\}))$, and hence:

$$\underset{D_{\beta,a},D_\alpha}{E}\left[\frac{1}{N_{\beta,a}}\sum_{i\in[N_{\beta,a}]} Z_i\right] = \frac{1}{N_{\beta,a}}\sum_{i\in[N_{\beta,a}]}\underset{D_{\beta,a},D_\alpha}{E}[Z_i] = \underset{D_{\beta,a},D_\alpha}{E}\left[\mathbb{I}\left(c_i \in (\min\{\hat{\alpha}_a^+, \alpha_a\}, \max\{\hat{\alpha}_a^+, \alpha_a\})\right)\right]$$
$$= \underset{D_\alpha}{E}\left[\underset{D_{\beta,a}}{E}\left[\mathbb{I}\left(c_i \in (\min\{\hat{\alpha}_a^+, \alpha_a\}, \max\{\hat{\alpha}_a^+, \alpha_a\})\right)\right]|D_\alpha\right] = \underset{D_\alpha}{E}\left[|F_a(\alpha_a) - F_a(\hat{\alpha}_a^+)|\right].$$

For a fixed $\gamma \in (0,1)$, let $\mathcal{G} = \{|\hat{\alpha}_a^+ - \alpha_a| \leqslant \gamma_a\}$ be the good event that the threshold error is at most γ_a. Then,

$$\begin{aligned}
\underset{D_{\beta,a},D_\alpha}{E}\left[|\hat{\rho}_a(\hat{h}) - \hat{\rho}_a(h)|\right] &\leqslant \underset{D_\alpha}{E}\left[|F_a(\alpha_a) - F_a(\hat{\alpha}_a^+)|\right] \\
&= \underset{D_\alpha}{E}\left[|F_a(\alpha_a) - F_a(\hat{\alpha}_a^+)|\mathbb{1}_{\mathcal{G}}\right] + \underset{D_\alpha}{E}\left[|F_a(\alpha_a) - F_a(\hat{\alpha}_a^+)|\mathbb{1}_{\mathcal{G}^C}\right] \\
&\leqslant L_a\gamma_a + \Pr(\mathcal{G}^c) \text{ (Using Assumption 6.1 and using } |F_a(\alpha_a) - F_a(\hat{\alpha}_a^+)| \leqslant 1) \\
&= L_a\gamma_a + \Pr\left(|\alpha_a - \hat{\alpha}_a^+| > \gamma_a\right).
\end{aligned}$$

Overall,

$$|\Delta_{DP}(h) - \underset{D_{\beta,D_\alpha}}{E}\left[\hat{\Delta}_{DP}(\hat{h})\right]| \leqslant \sum_{a\in\{0,1\}}\left(\frac{1}{2\sqrt{N_{\beta,a}}} + L_a\gamma_a + \Pr\left(|\alpha_a - \hat{\alpha}_a^+| > \gamma_a\right)\right).$$

Using the sample complexity bound from Theorem 6.1 and the dominant term, $\gamma_a = \frac{|\lambda|}{\pi_a^2}\sqrt{\frac{\log(4/\delta_a)}{2N_\alpha}}$, giving us:

$$|\Delta_{DP}(h) - \underset{D_{\beta,D_\alpha}}{E}\left[\hat{\Delta}_{DP}(\hat{h})\right]| \leqslant \sum_{a\in\{0,1\}}\left(\frac{1}{2\sqrt{N_{\beta,a}}} + \frac{L_a|\lambda|}{\pi_a^2}\sqrt{\frac{\log(4/\delta_a)}{2N_\alpha}} + \delta_a\right).$$

Setting $N_\beta = \min\{N_{\beta,0}, N_{\beta,1}\}$, we can obtain:

$$|\Delta_{DP}(h) - \underset{D_{\beta,D_\alpha}}{E}\left[\hat{\Delta}_{DP}(\hat{h})\right]| = \mathcal{O}\left(\frac{1}{\sqrt{N_\beta}} + |\lambda|\cdot\max\{L_0, L_1\}\sqrt{\frac{\log(1/\delta_a)}{N_\alpha}}\right).$$

(3) **Sample Complexity**: To bound the sample complexity, we need to bound the following event:

$$\Pr\left(|\hat{\Delta}(\hat{h}) - \Delta(h)| > \epsilon\right) \leqslant \delta.$$

Let $\Delta(\hat{h})$ denote the true DP difference with respect to the estimated threshold $\hat{\alpha}_a^+$. Consider the following quantity and its breakdown via the triangle inequality:

$$|\hat{\Delta}(\hat{h}) - \Delta(h)| \leqslant |\hat{\Delta}(\hat{h}) - \Delta(\hat{h})| + |\Delta(\hat{h}) - \Delta(h)| \text{ (By Triangle Inequality)}$$

By the union bound, we can upper bound our total error

$$\Pr\left(|\hat{\Delta}(\hat{h}) - \Delta(h)| > \epsilon\right) \leqslant \Pr\left(|\hat{\Delta}(\hat{h}) - \Delta(\hat{h})| > \epsilon/2\right) + \Pr\left(|\Delta(\hat{h}) - \Delta(h)| > \epsilon/2\right).$$

We can first use the Hoeffding bound for the first term by noting that $|\hat{\Delta}(\hat{h}) - \Delta(\hat{h})| \leqslant \sum_{a\in\mathcal{A}} |\rho_a(\hat{h}) - \hat{\rho}_a(\hat{h})|$:

$$\Pr\left(|\rho_a(\hat{h}) - \hat{\rho}_a(\hat{h})| > \epsilon/4\right) \leqslant 2\exp\left(\frac{-2N_{\beta,a}\epsilon^2}{16}\right) \leqslant \frac{\delta}{4} \implies N_{\beta,a} \geqslant \frac{8\log\left(\frac{8}{\delta}\right)}{\epsilon^2}.$$

For both groups, we would need $N_\beta = \mathcal{O}\left(\frac{1}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$.
Next, consider the second term:

$$\begin{aligned}|\Delta(\hat{h}) - \Delta(h)| &\leqslant \sum_{a\in\mathcal{A}} |\rho_a(h) - \rho_a(\hat{h})| = \sum_{a\in\mathcal{A}} |\Pr\left(C \geqslant \alpha_a | A = a\right) - \Pr\left(C \geqslant \hat{\alpha}_a^+ | A = a\right))| \\ &= \sum_{a\in\mathcal{A}} |1 - F_a(\alpha_a) - 1 + F_a(\hat{\alpha}_a^+)| \leqslant \sum_{a\in\mathcal{A}} L_a \cdot |\alpha_a - \hat{\alpha}_a^+| \text{ (Using Assumption 6.1).}\end{aligned}$$

Now, we can figure out the sample complexity of each term:

$$\Pr\left(|\rho_a(h) - \rho_a(\hat{h})| > \epsilon/4\right) = \Pr\left(|\alpha_a - \hat{\alpha}_a^+| > \frac{\epsilon}{4L_a}\right) < \frac{\delta}{4}.$$

Now, we can use Theorem 6.1 to get the number of samples required to upper bound the above error: $N_\alpha = \mathcal{O}\left(\frac{\lambda^2\cdot\max\{L_0^2, L_1^2\}}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right)$. □

Theorem 6.9. Suppose a dataset D_β is used to construct $\hat{\Delta}_{EO}(g_{\hat{\alpha}_a^+})$, where $\hat{\alpha}_a^+$ is the estimator for the EO threshold (Theorem 6.1). Then the following holds:

1. Consistency: $\hat{\Delta}_{EO}(h_{\hat{\alpha}_a^+}) \xrightarrow{p} \Delta_{EO}(h_\alpha)$.
2. Bias: $|\Delta_{EO}(h) - \mathop{E}_{D_\beta, D_\alpha}\left[\hat{\Delta}_{EO}(\hat{h})\right]| = \mathcal{O}\left(\frac{|\lambda|\cdot\max\{L_0, L_1\}}{\tau^2}\sqrt{\frac{\log(1/\delta)}{N_\alpha}} + \frac{1}{\sqrt{N_\beta}}\right)$.
3. Sample Complexity: With $1-\delta$ probability, $|\hat{\Delta}(\hat{h}) - \Delta(h)| \leqslant \epsilon$, as long as $N_\alpha = \mathcal{O}\left(\frac{\lambda^2\max\{L_0^2, L_1^2\}}{\epsilon^2\tau^4}\log\left(\frac{1}{\delta}\right)\right)$ and $N_\beta = \mathcal{O}\left\{\frac{1}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right\}$.

Proof. For the sake of this proof, we will denote $\Delta_{EO}(h_\alpha)$ as $\Delta(h)$ and $\hat{\Delta}_{EO}(h_{\hat{\alpha}_a^+})$ as $\hat{\Delta}(\hat{h})$. Then an intermediate estimator $\hat{\Delta}_{EO}(h_{\alpha_a})$ can be denoted as $\hat{\Delta}(h)$. We similarly overload the notations in this fashion for the positive rate estimator $\rho_{TPR,a}$ with ρ_a.

Recalling the definitions of our estimations,

$$\rho_{\text{TPR},a}(h_\alpha) = \Pr(h_\alpha = 1|Y=1, A=a) = \frac{\mathop{E}_{X|A=a}[\mathbb{I}(\eta(x,a) \geqslant \alpha_a) \cdot \Pr(Y=1|X=x, A=a)]}{\pi_{1|a}} = \frac{\mathop{E}_{X|A=a}[\mathbb{I}(C \geqslant \alpha_a) \cdot C]}{\pi_{1|a}},$$

and,

$$\hat{\rho}_{\text{TPR},a}(h_{\hat{\alpha}_a^+}) = \frac{\frac{1}{N_{\beta,a}} \sum_{i \in [N_{\beta,a}]} (c_i(\mathbb{I}(c_i \geqslant \hat{\alpha}_a^+))}{\hat{\pi}_{1|a}^+} = \frac{\hat{E}_{X|A=a}[\mathbb{I}(c \geqslant \hat{\alpha}_a^+)\, c]}{\hat{\pi}_{1|a}^+}$$

(1) **Consistency**: Most of the inequality arguments in the proof are a direct consequence of the triangle inequality, by introducing an intermediate estimate of the TPR ($\hat{\rho}_a(h)$).

We want to show that $\Pr\left(|\hat{\Delta}(\hat{h}) - \Delta(h)| > \epsilon\right) \to 0$ as $N_\beta \to \infty$. By Slutsky's theorem, it will be enough to upper-bound our target event as a sum of simpler events and show that if each of them converges in probability to zero, then we can claim our final result.

Consider the absolute difference term and apply the reverse triangle inequality ($||A| - |B|| \leqslant |A - B|$):

$$\begin{aligned} &|\Delta(h) - \hat{\Delta}(\hat{h})| \leqslant |\rho_1(h) - \rho_0(h) - \hat{\rho}_1(\hat{h}) + \hat{\rho}_0(\hat{h})| \\ &\leqslant |\rho_1(h) - \rho_1(\hat{h})| + |\rho_0(h) - \rho_0(\hat{h})| + |\rho_1(\hat{h}) - \hat{\rho}_1(\hat{h})| + |\rho_0(\hat{h}) - \hat{\rho}_0(\hat{h})|. \end{aligned}$$

Since similar arguments will work for any group $A = a$, we will now study each term separately for an arbitrary group $A = a$.

Consider $|\rho_a(\hat{h}) - \hat{\rho}_a(\hat{h})|$. $\hat{E}_{X|A=a}[\mathbb{I}(c \geqslant \hat{\alpha}_a^+)\, c] \to E_{X|A=a}[\mathbb{I}(c \geqslant \hat{\alpha}_a^+)\, c]$ and $\hat{\pi}_{1|a}^+ \xrightarrow{p} \pi_{1|a}$ due to the weak law of large numbers, since conditioned on D_α, $\hat{\alpha}_a^+$ and $\hat{\pi}_{1|a}^+$ are fixed numbers.

Since $\pi_{1|a} > 0$ by assumption, by Slutsky's theorem, $\hat{\rho}_a(\hat{h}) = \frac{\hat{E}_{X|A=a}[\mathbb{I}(c \geqslant \hat{\alpha}_a^+)c]}{\hat{\pi}_{1|a}^+} \xrightarrow{p} \rho_a(\hat{h})$.

Next, consider the term $|\rho_a(\hat{h}) - \rho_a(h)|$. Since the denominator for both is a constant $\pi_{1|a}$, we will focus on showing that the difference in the numerator converges to zero.

$$|\mathop{E}_{X|A=a}[\mathbb{I}(C \geqslant \alpha_a)\, C] - \mathop{E}_{X|A=a}\left[\mathbb{I}\left(C \geqslant \hat{\alpha}_a^+\right) C\right]| \leqslant \mathop{E}_{X|A=a}[Z]$$

where $Z = |\mathbb{I}(C \geqslant \alpha_a) - \mathbb{I}(C \geqslant \hat{\alpha}_a^+)| = \mathbb{I}(C \in (\min\{\hat{\alpha}_a^+, \alpha_a\}, \max\{\hat{\alpha}_a^+, \alpha_a\}))$ and $C \leqslant 1$. Using this we can write $\mathop{E}_{X|A=a}[Z] = \Pr(\min\{\hat{\alpha}_a^+, \alpha_a\} \leqslant C < \max\{\hat{\alpha}_a^+, \alpha_a\}|A = a) = |F_a(\hat{\alpha}_a^+) - F_a(\alpha_a)|$. Since F_a is continuous, and from Theorem 6.1, $\hat{\alpha}_a^+ \xrightarrow{p} \alpha_a$, by the continuous mapping theorem, $F_a(\hat{\alpha}_a^+) \xrightarrow{p} F_a(\alpha_a)$. Combining all the pieces, $\hat{\Delta}_{EO}(g_{\hat{\alpha}_a^+}) \xrightarrow{p} \Delta_{EO}(g_\alpha)$.

(2) **Bias**: We now want to bound the bias of our estimator, $|\Delta_{EO}(h) - \mathop{E}_{D_\beta, D_\alpha}\left[\hat{\Delta}_{EO}(\hat{h})\right]|$.

$$|\Delta_{EO}(h) - \mathop{E}_{D_\beta, D_\alpha}\left[\hat{\Delta}_{EO}(\hat{h})\right]| = |\mathop{E}_{D_\beta, D_\alpha}\left[\Delta_{EO}(h) - \hat{\Delta}_{EO}(\hat{h})\right]| \leqslant \mathop{E}_{D_\beta, D_\alpha}\left[|\hat{\Delta}_{EO}(\hat{h}) - \Delta_{EO}(h)|\right]$$
$$\leqslant \mathop{E}_{D_\beta, D_\alpha}\left[|\hat{\rho}_1(\hat{h}) - \hat{\rho}_0(\hat{h}) - \rho_1(h) + \rho_0(h)|\right]$$
(By using the Reverse Triangle Inequality)
$$\leqslant \mathop{E}_{D_{\beta,1}, D_\alpha}\left[|\rho_1(h) - \rho_1(\hat{h})\right| + \mathop{E}_{D_{\beta,0}, D_\alpha}\left[|\rho_0(h) - \rho_0(\hat{h})|\right]$$
$$+ \mathop{E}_{D_{\beta,1}, D_\alpha}\left[|\rho_1(\hat{h}) - \hat{\rho}_1(\hat{h})|\right] + \mathop{E}_{D_{\beta,0}, D_\alpha}\left[|\rho_0(\hat{h}) - \hat{\rho}_0(\hat{h})|\right]$$
(By introducing an intermediate estimator $\rho_a(\hat{h})$,
and noting that the expectation is over group-specific samples $D_{\beta.a}$).

Once again, we will only focus on deriving the statistical bias for an arbitrary group $A = a$. Consider the term $\mathop{E}_{D_{\beta,a}, D_\alpha}\left[|\rho_a(\hat{h}) - \rho_a(h)|\right]$. Since the denominator for both is a constant $\pi_{1|a}$, we will focus on showing that the difference in the numerator converges to zero. Reusing our analysis from the previous proof on consistency and using Assumption 6.1, we can write:

$$\mathop{E}_{D_{\beta,a}, D_\alpha}\left[|\rho_a(\hat{h}) - \rho_a(h)|\right] \leqslant \frac{1}{\pi_{1|a}} \mathop{E}_{D_\alpha}\left[|F_a(\hat{\alpha}_a^+) - F_a(\alpha_a)|\right] \leqslant \frac{L_a}{\pi_{1|a}} \mathop{E}_{D_\alpha}\left[|\hat{\alpha}_a^+ - \alpha_a|\right]$$

From Theorem 6.1, we know that with $1-\delta$ probability, $|\hat{\alpha}_a^+ - \alpha_a| < \epsilon$. We also know that $|\alpha_a - \hat{\alpha}_a^+| \leqslant 1$. Using this, we get:

$$\mathop{E}_{D_\alpha}\left[|\alpha_a - \hat{\alpha}_a^+|\right] = \mathop{E}_{D_\alpha}\left[|\alpha_a - \hat{\alpha}_a^+| \mathbb{I}\left(|\alpha_a - \hat{\alpha}_a^+| < \epsilon_a\right)\right] + \mathop{E}_{D_\alpha}\left[|\alpha_a - \hat{\alpha}_a^+| \mathbb{I}\left(|\alpha_a - \hat{\alpha}_a^+| \geqslant \epsilon_a\right)\right]$$
$$\leqslant \epsilon_a \cdot \Pr\left(|\alpha_a - \hat{\alpha}_a^+| < \epsilon_a\right) + 1 \cdot \Pr\left(|\alpha_a - \hat{\alpha}_a^+| \geqslant \epsilon_a\right) \text{ (Since } |\alpha_a - \hat{\alpha}_a^+| \leqslant 1) \leqslant \epsilon_a + \delta_a.$$

Using $\epsilon_a = \frac{|\lambda|}{\tau^2 \pi_{1a}^2} \sqrt{\frac{8 \log(2/\delta_a)}{N_\alpha}}$, our final bound becomes:

$$\mathop{E}_{D_{\beta,a}, D_\alpha}\left[|\rho_a(\hat{h}) - \rho_a(h)|\right] \leqslant \frac{L_a}{\pi_{1|a}} \left(\frac{|\lambda|}{\tau^2 \pi_{1a}^2} \sqrt{\frac{8 \log(2/\delta_a)}{N_\alpha}} + \delta_a \right)$$

Next, consider the term $\mathop{E}_{D_{\beta,a}, D_\alpha}\left[|\rho_a(\hat{h}) - \hat{\rho}_a(\hat{h})|\right]$.

$$\mathop{\mathrm{E}}_{D_{\beta,a},D_\alpha}\left[|\rho_a(\hat{h})-\hat{\rho}_a(\hat{h})|\right] = \mathop{\mathrm{E}}_{D_{\beta,a},D_\alpha}\left[|\frac{\mathop{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C\geqslant\hat{\alpha}_a^+\right)C]}{\pi_{1|a}} - \frac{\hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C\geqslant\hat{\alpha}_a^+\right)C]}{\hat{\pi}_{1|a}^+}|\right]$$

$$= \mathrm{E}_{D_{\beta,a},D_\alpha}\left[|\frac{\mathop{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C\geqslant\hat{\alpha}_a^+\right)C]}{\pi_{1|a}} - \frac{\hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C\geqslant\hat{\alpha}_a^+\right)C]}{\pi_{1|a}}\right.$$

$$\left.+\frac{\hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C\geqslant\hat{\alpha}_a^+\right)C]}{\pi_{1|a}} - \frac{\hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C\geqslant\hat{\alpha}_a^+\right)C]}{\hat{\pi}_{1|a}^+}|\right]$$

(By introducing an intermediate term $\frac{\hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C\geqslant\hat{\alpha}_a^+\right)C]}{\pi_{1|a}}$)

$$\leqslant \mathop{\mathrm{E}}_{D_{\beta,a},D_\alpha}\left[\frac{1}{\pi_{1|a}}\cdot|\mathop{\mathrm{E}}_{X|A=a}\left[\mathbb{I}\left(C\geqslant\hat{\alpha}_a^+\right)C\right]-\hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C\geqslant\hat{\alpha}_a^+\right)C]|\right]$$

$$+\mathop{\mathrm{E}}_{D_{\beta,a},D_\alpha}\left[\hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C\geqslant\hat{\alpha}_a^+\right)C]\cdot\left|\frac{1}{\pi_{1|a}}-\frac{1}{\hat{\pi}_{1|a}^+}\right|\right]$$

$$\leqslant \frac{1}{\pi_{1|a}}\mathop{\mathrm{E}}_{D_{\beta,a},D_\alpha}\left[|\mathop{\mathrm{E}}_{X|A=a}\left[\mathbb{I}\left(C\geqslant\hat{\alpha}_a^+\right)C\right]-\hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C\geqslant\hat{\alpha}_a^+\right)C]|\right]+\mathop{\mathrm{E}}_{D_{\beta,a},D_\alpha}\left[\left|\frac{1}{\pi_{1|a}}-\frac{1}{\hat{\pi}_{1|a}^+}\right|\right]$$

(Since $\hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C\geqslant\hat{\alpha}_a^+\right)C]\leqslant 1$)

Conditioned on D_α, $\hat{\alpha}_a^+$ is fixed. Therefore, we can use Lemma 6.5 to upper bound the term $\mathop{\mathrm{E}}_{D_{\beta,a},D_\alpha}\left[|\mathop{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C\geqslant\hat{\alpha}_a^+\right)C]-\hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C\geqslant\hat{\alpha}_a^+\right)C]|\right]$:

$$\mathop{\mathrm{E}}_{D_{\beta,a},D_\alpha}\left[|\mathop{\mathrm{E}}_{X|A=a}\left[\mathbb{I}\left(C\geqslant\hat{\alpha}_a^+\right)C\right]-\hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C\geqslant\hat{\alpha}_a^+\right)C]|\right]$$

$$= \mathop{\mathrm{E}}_{D_{\beta,a}}\left[|\mathop{\mathrm{E}}_{X|A=a}\left[\mathbb{I}\left(C\geqslant\hat{\alpha}_a^+\right)C\right]-\hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C\geqslant\hat{\alpha}_a^+\right)C]|\,|D_\alpha\right]$$

$$= \sqrt{\mathrm{Var}(\hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C\geqslant\hat{\alpha}_a^+\right)C]|D_\alpha)} = \sqrt{\frac{1}{N_{\beta,a}}\mathrm{Var}(\mathbb{I}\left(C\geqslant\alpha_a^+\right)C)}$$

$$\leqslant \frac{1}{2\sqrt{N_{\beta,a}}} \text{ (Using Popoviciu's Inequality).}$$

To bound $\mathop{\mathrm{E}}_{D_{\beta,a},D_\alpha}\left[\left|\frac{1}{\pi_{1|a}}-\frac{1}{\hat{\pi}_{1|a}^+}\right|\right] = \mathop{\mathrm{E}}_{D_\alpha}\left[\left|\frac{1}{\pi_{1|a}}-\frac{1}{\hat{\pi}_{1|a}^+}\right|\right]$ (Since $\hat{\pi}_{1|a}^+$ is estimated using D_α), let $\mathcal{G} = \{\hat{\pi}_{1|a}\geqslant\frac{\pi_{1|a}}{2}\}$ denote the event that $\hat{\pi}_{1|a}$ and $\pi_{1|a}$ are close. Note that on $\mathcal{G}$, $\hat{\pi}_{1|a}^+ = \hat{\pi}_{1|a}$ for sufficiently many samples ($1/N_{\alpha,a}\leqslant\pi_{1|a}/2$):

$$
\begin{aligned}
&\mathop{\mathrm{E}}_{D_\alpha}\left[\left|\frac{1}{\pi^+_{1|a}}-\frac{1}{\hat{\pi}_{1|a}}\right|\right] = \mathop{\mathrm{E}}_{D_\alpha}\left[\left|\frac{1}{\pi_{1|a}}-\frac{1}{\hat{\pi}_{1|a}}\right|\mathbb{1}_{\mathcal{G}}\right] + \mathop{\mathrm{E}}_{D_\alpha}\left[\left|\frac{1}{\pi_{1|a}}-\frac{1}{\hat{\pi}^+_{1|a}}\right|\mathbb{1}_{\mathcal{G}_c}\right] \\
&= \mathop{\mathrm{E}}_{D_\alpha}\left[\frac{|\hat{\pi}_{1|a}-\pi_{1|a}|}{\hat{\pi}_{1|a}\pi_{1|a}}\mathbb{1}_{\mathcal{G}}\right] + \mathop{\mathrm{E}}_{D_\alpha}\left[\left|\frac{1}{\pi_{1|a}}-\frac{1}{\hat{\pi}^+_{1|a}}\right|\mathbb{1}_{\mathcal{G}_c}\right] \\
&\leqslant \frac{2}{\pi^2_{1|a}}\mathop{\mathrm{E}}_{D_\alpha}\left[|\hat{\pi}_{1|a}-\pi_{1|a}|\right] + \mathop{\mathrm{E}}_{D_\alpha}\left[\left|\frac{1}{\pi_{1|a}}-\frac{1}{\hat{\pi}^+_{1|a}}\right|\mathbb{1}_{\mathcal{G}_c}\right] \text{ (On } \mathcal{G}, \hat{\pi}_{1|a} \geqslant \pi_{1|a}/2) \\
&\leqslant \frac{1}{\pi^2_{1|a}\sqrt{N_{\alpha,a}}} + \mathop{\mathrm{E}}_{D_\alpha}\left[\left|\frac{1}{\pi_{1|a}}-\frac{1}{\hat{\pi}^+_{1|a}}\right|\mathbb{1}_{\mathcal{G}_c}\right] \text{ (Using Lemma 6.5 and Popoviciu's Inequality)} \\
&\leqslant \frac{1}{\pi^2_{1|a}\sqrt{N_{\alpha,a}}} + \mathop{\mathrm{E}}_{D_\alpha}\left[\left(\frac{1}{\pi_{1|a}}+N_{\alpha,a}\right)\mathbb{1}_{\mathcal{G}_c}\right] \text{ (Using the form of our truncated estimator } \hat{\pi}^+_{1|a}) \\
&\leqslant \frac{1}{\pi^2_{1|a}\sqrt{N_{\alpha,a}}} + \left(\frac{1}{\pi_{1|a}}+N_{\alpha,a}\right)\Pr(\mathcal{G}_c) \text{ (Using the Cauchy-Schwartz Inequality)} \\
&\leqslant \frac{1}{\pi^2_{1|a}\sqrt{N_{\alpha,a}}} + \left(\frac{1}{\pi_{1|a}}+N_{\alpha,a}\right)\exp\left(-\frac{N_{\alpha,a}\pi^2_{1|a}}{2}\right) \text{ (Using the Hoeffding's Inequality)}
\end{aligned}
$$

Our upper bound becomes:

$$
\mathop{\mathrm{E}}_{D_{\beta,a},D_\alpha}\left[|\rho_a(\hat{h})-\hat{\rho}_a(\hat{h})|\right] \leqslant \frac{1}{2\pi_{1|a}\sqrt{N_{\beta,a}}} + \frac{1}{\pi^2_{1|a}\sqrt{N_{\alpha,a}}} + \left(\frac{1}{\pi_{1|a}}+N_{\alpha,a}\right)\exp\left(-\frac{N_{\alpha,a}\pi^2_{1|a}}{2}\right)
$$

and our overall upper bound on the bias of the estimator takes the form:

$$
\begin{aligned}
&|\Delta_{EO}(h) - \mathop{\mathrm{E}}_{D_\beta,D_\alpha}\left[\hat{\Delta}_{EO}(\hat{h})\right]| \\
&\leqslant \sum_{a\in\{0,1\}}\left(\frac{L_a}{\pi_{1|a}}\left(\frac{|\lambda|}{\tau^2\pi^2_{1a}}\sqrt{\frac{8\log(2/\delta_a)}{N_\alpha}}+\delta_a\right) + \frac{1}{2\pi_{1|a}\sqrt{N_{\beta,a}}} + \frac{1}{\pi^2_{1|a}\sqrt{N_{\alpha,a}}}\right. \\
&\left. + \left(\frac{1}{\pi_{1|a}}+N_{\alpha,a}\right)\exp\left(-\frac{N_{\alpha,a}\pi^2_{1|a}}{2}\right)\right) \\
&= \sum_{a\in\{0,1\}}\mathcal{O}\left(\frac{L_a}{\pi_{1|a}}\left(\frac{|\lambda|}{\tau^2\pi^2_{1a}}\sqrt{\frac{\log(2/\delta_a)}{N_\alpha}}+\delta_a\right) + \frac{1}{2\pi_{1|a}\sqrt{N_{\beta,a}}} + \frac{1}{\pi^2_{1|a}\sqrt{N_{\alpha,a}}}\right)
\end{aligned}
$$

Setting $\delta = \min\{\delta_0,\delta_1\}$, $N_\alpha = \min\{N_{\alpha,0},N_{\alpha,1}\}$ and $N_\beta = \min\{N_{\beta,0},N_{\beta,1}\}$:

$$
|\Delta_{EO}(h) - \mathop{\mathrm{E}}_{D_\beta,D_\alpha}\left[\hat{\Delta}_{EO}(\hat{h})\right]| = \mathcal{O}\left(\frac{|\lambda|\cdot\max\{L_0,L_1\}}{\tau^2}\sqrt{\frac{\log(1/\delta)}{N_\alpha}} + \frac{1}{\sqrt{N_\beta}}\right)
$$

(3) **Sample Complexity**: To bound the sample complexity, we need to bound the following event:

$$\Pr\left(|\hat{\Delta}(\hat{h}) - \Delta(h)| > \epsilon\right) \leqslant \delta.$$

Let $\Delta(\hat{h})$ denote the true EO difference with respect to the estimated threshold $\hat{\alpha}_a^+$. Consider the absolute difference term and apply the reverse triangle inequality ($||A|-|B|| \leqslant |A-B|$):

$$\begin{aligned}
&|\Delta(h) - \hat{\Delta}(\hat{h})| \leqslant |\rho_1(h) - \rho_0(h) - \hat{\rho}_1(\hat{h}) + \hat{\rho}_0(\hat{h})| \\
&\leqslant |\rho_1(h) - \rho_1(\hat{h})| + |\rho_0(h) - \rho_0(\hat{h})| + |\rho_1(\hat{h}) - \hat{\rho}_1(\hat{h})| + |\rho_0(\hat{h}) - \hat{\rho}_0(\hat{h})|.
\end{aligned}$$

By the union bound, we want to ensure the following:

$$\Pr\left(|\Delta(h) - \hat{\Delta}(\hat{h})| > \epsilon\right) \leqslant \sum_{a\in\{0,1\}} \Pr\left(|\rho_a(h) - \rho_a(\hat{h})| > \epsilon/4\right) + \Pr\left(|\rho_a(\hat{h}) - \hat{\rho}_a(\hat{h})| > \epsilon/4\right).$$

Since similar arguments will work for any group $A = a$, we will now study each term separately for an arbitrary group $A = a$. From our analysis in previous part, we know that $|\rho_a(h) - \rho_a(\hat{h})| \leqslant \frac{L_a}{\pi_{1|a}}|\hat{\alpha}_a^+ - \alpha_a|$ and therefore $\Pr\left(|\rho_a(h) - \rho_a(\hat{h})| > \epsilon/4\right) \leqslant \Pr\left(|\hat{\alpha}_a^+ - \alpha_a| > \frac{\epsilon\pi_{1|a}}{4L_a}\right)$.

Using Theorem 6.1, recall that with $1-\delta$ probability over the choice of data sample N_α, as long as $N_\alpha \geqslant \frac{1}{2t^2}\log\left(\frac{2}{\delta}\right)$ for $t = \min\left\{\frac{\pi_{1a}}{2}, \frac{\tau\pi_{1a}^2}{4|\lambda|}, \frac{\epsilon\tau^2\pi_{1a}^2}{4|\lambda|}\right\}$, $\Pr\left(|\hat{\alpha}_a - \alpha_a| > \epsilon\right) \leqslant \delta$. With our new ϵ and δ we get:

$$\begin{aligned}
\Pr\left(|\alpha_a - \hat{\alpha}_a^+| > \frac{\epsilon\pi_{1|a}}{4L_a}\right) \leqslant \delta/4 \implies N_\alpha &\geqslant \frac{1}{2}\log\left(\frac{8}{\delta}\right)\max\left\{\frac{2}{\pi_{1a}}, \frac{4|\lambda|}{\tau\pi_{1a}^2}, \frac{16|\lambda|L_a}{\epsilon\tau^2\pi_{1|a}\pi_{1a}^2}\right\}^2 \\
&= \mathcal{O}\left(\frac{\lambda^2 L_a^2}{\epsilon^2\tau^4}\log\left(\frac{1}{\delta}\right)\right).
\end{aligned}$$

Now we focus on the other event. We will proceed by introducing an intermediate estimator:

$$\begin{aligned}
&|\rho_a(\hat{h}) - \hat{\rho}_a(\hat{h})| = |\frac{\mathop{E}_{X|A=a}[\mathbb{I}\left(C \geqslant \hat{\alpha}_a^+\right)C]}{\pi_{1|a}} - \frac{\hat{E}_{X|A=a}[\mathbb{I}\left(C \geqslant \hat{\alpha}_a^+\right)C]}{\hat{\pi}_{1|a}^+}| \\
&= |\frac{\mathop{E}_{X|A=a}[\mathbb{I}\left(C \geqslant \hat{\alpha}_a^+\right)C]}{\pi_{1|a}} - \frac{\hat{E}_{X|A=a}[\mathbb{I}\left(C \geqslant \hat{\alpha}_a^+\right)C]}{\pi_{1|a}} + \frac{\hat{E}_{X|A=a}[\mathbb{I}\left(C \geqslant \hat{\alpha}_a^+\right)C]}{\pi_{1|a}} - \frac{\hat{E}_{X|A=a}[\mathbb{I}\left(C \geqslant \hat{\alpha}_a^+\right)C]}{\hat{\pi}_{1|a}^+}| \\
&\text{(By introducing an intermediate term } \frac{\hat{E}_{X|A=a}[\mathbb{I}\left(C \geqslant \hat{\alpha}_a^+\right)C]}{\pi_{1|a}}\text{)} \\
&\leqslant \frac{1}{\pi_{1|a}} \cdot |\mathop{E}_{X|A=a}\left[\mathbb{I}\left(C \geqslant \hat{\alpha}_a^+\right)C\right] - \hat{E}_{X|A=a}[\mathbb{I}\left(C \geqslant \hat{\alpha}_a^+\right)C]| + \hat{E}_{X|A=a}[\mathbb{I}\left(C \geqslant \hat{\alpha}_a^+\right)C] \cdot \left|\frac{1}{\pi_{1|a}} - \frac{1}{\hat{\pi}_{1|a}^+}\right| \\
&\leqslant \frac{1}{\pi_{1|a}}|\mathop{E}_{X|A=a}\left[\mathbb{I}\left(C \geqslant \hat{\alpha}_a^+\right)C\right] - \hat{E}_{X|A=a}[\mathbb{I}\left(C \geqslant \hat{\alpha}_a^+\right)C]| + \left|\frac{1}{\pi_{1|a}} - \frac{1}{\hat{\pi}_{1|a}^+}\right| \\
&\text{(Since } \hat{E}_{X|A=a}[\mathbb{I}\left(C \geqslant \hat{\alpha}_a^+\right)C] \leqslant 1\text{)}
\end{aligned}$$

Now we can use the union bound:

$$\Pr\left(|\rho_a(\hat{h}) - \hat{\rho}_a(\hat{h})| > \epsilon/4\right) \leqslant \Pr\left(\frac{1}{\pi_{1|a}}|\underset{X|A=a}{\mathrm{E}}\left[\mathbb{I}\left(C \geqslant \hat{\alpha}_a^+\right) C\right] - \hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C \geqslant \hat{\alpha}_a^+\right) C]| > \epsilon/8\right) + \Pr\left(\left|\frac{1}{\pi_{1|a}} - \frac{1}{\hat{\pi}_{1|a}^+}\right| > \epsilon/8\right)$$

To further simplify $\left|\frac{1}{\pi_{1|a}} - \frac{1}{\hat{\pi}_{1|a}^+}\right|$, we can borrow our analysis from the bias proof. Recall $\mathcal{G} = \{\hat{\pi}_{1|a} \geqslant \frac{\pi_{1|a}}{2}\}$ to be the event that $\hat{\pi}_{1|a}$ and $\pi_{1|a}$ are close. Note that on $\mathcal{G}$, $\hat{\pi}_{1|a}^+ = \hat{\pi}_{1|a}$ for sufficiently many samples ($1/N_{\alpha,a} \leqslant \pi_{1|a}/2$). On this good event $\mathcal{G}$, we derived that $\left|\frac{1}{\pi_{1|a}} - \frac{1}{\hat{\pi}_{1|a}^+}\right| \leqslant \frac{2}{\pi_{1|a}^2}|\hat{\pi}_{1|a} - \pi_{1|a}|$.

$$\begin{aligned}
\Pr\left(|\rho_a(\hat{h}) - \hat{\rho}_a(\hat{h})| > \epsilon/4\right) &\leqslant \Pr\left(\frac{1}{\pi_{1|a}}|\underset{X|A=a}{\mathrm{E}}\left[\mathbb{I}\left(C \geqslant \hat{\alpha}_a^+\right) C\right] - \hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C \geqslant \hat{\alpha}_a^+\right) C]| > \epsilon/8\right) \\
&\quad + \Pr\left(\left|\frac{1}{\pi_{1|a}} - \frac{1}{\hat{\pi}_{1|a}^+}\right| > \epsilon/8\right) \\
&= \Pr\left(\frac{1}{\pi_{1|a}}|\underset{X|A=a}{\mathrm{E}}\left[\mathbb{I}\left(C \geqslant \hat{\alpha}_a^+\right) C\right] - \hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C \geqslant \hat{\alpha}_a^+\right) C]| > \epsilon/8\right) \\
&\quad + \Pr\left(\left|\frac{1}{\pi_{1|a}} - \frac{1}{\hat{\pi}_{1|a}^+}\right| > \epsilon/8, \mathcal{G}\right) + \Pr\left(\left|\frac{1}{\pi_{1|a}} - \frac{1}{\hat{\pi}_{1|a}^+}\right| > \epsilon/8, \mathcal{G}_C\right) \\
&\leqslant \Pr\left(\frac{1}{\pi_{1|a}}|\underset{X|A=a}{\mathrm{E}}\left[\mathbb{I}\left(C \geqslant \hat{\alpha}_a^+\right) C\right] - \hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C \geqslant \hat{\alpha}_a^+\right) C]| > \epsilon/8\right) \\
&\quad + \Pr\left(\frac{2}{\pi_{1|a}^2}|\hat{\pi}_{1|a} - \pi_{1|a}| > \epsilon/8\right) + \Pr\left(\mathcal{G}_C\right) \\
&= \Pr\left(|\underset{X|A=a}{\mathrm{E}}\left[\mathbb{I}\left(C \geqslant \hat{\alpha}_a^+\right) C\right] - \hat{\mathrm{E}}_{X|A=a}[\mathbb{I}\left(C \geqslant \hat{\alpha}_a^+\right) C]| > \frac{\pi_{1|a}\epsilon}{8}\right) \\
&\quad + \Pr\left(|\hat{\pi}_{1|a} - \pi_{1|a}| > \frac{\pi_{1|a}^2\epsilon}{16}\right) + \Pr\left(\hat{\pi}_{1|a} < \frac{\pi_{1|a}}{2}\right) \\
&\leqslant 2\left(\exp\left(-\frac{N_{\beta,a}\pi_{1|a}^2\epsilon^2}{32}\right) + \exp\left(-\frac{N_{\alpha,a}\pi_{1|a}^4\epsilon^2}{128}\right) + \exp\left(-\frac{N_{\alpha,a}\pi_{1|a}^2}{2}\right)\right)
\end{aligned}$$

(By using Hoeffding's Inequality on each event).

We can now derive our sample complexities by splitting out $\delta/4$ budget between all three terms to obtain $N_{\beta,a} \geqslant \frac{32}{\pi_{1|a}^2\epsilon^2}\log\left(\frac{24}{\delta}\right)$ and $N_{\alpha,a} \geqslant \max\left\{\frac{2}{\pi_{1|a}^2}\log\left(\frac{24}{\delta}\right), \frac{128}{\pi_{1|a}^4\epsilon^2}\log\left(\frac{24}{\delta}\right)\right\}$. Overall, our sample complexities take the following form:

$$N_\alpha = \mathcal{O}\left(\frac{\lambda^2\max\{L_0^2, L_1^2\}}{\epsilon^2\tau^4}\log\left(\frac{1}{\delta}\right)\right) \text{ and } N_\beta = \mathcal{O}\left\{\frac{1}{\epsilon^2}\log\left(\frac{1}{\delta}\right)\right\}.$$

□

Chapter 7

Concluding Remarks

In this thesis, we have studied fair classification and the trade-offs it induces with accuracy, from the point of view of Bayes optimal classification and imperfect, biased distributions. For some of our studies, we have also been able to demonstrate impact on real-world datasets and hint towards more comprehensive benchmarking of fair algorithms, especially when the supervision is imperfect or the distributions are biased.

Many of the results in this thesis can be extended to group-blind classification and to settings with noisy information about the underlying groups, especially when we can model the noise in the sensitive groups, just as we model label biases studied in Chapter 3. Modern black-box models such as LLMs can generate reasonable class-probability scores for instances, which may allow practitioners to use characterization results such as those studied in this thesis to make LLMs fair [XWZ25]. Studying the calibration gap of these scores and adjusting the characterization results accordingly is a very fruitful future direction [UIS25].

Finally, several results in this thesis are non-trivial to generalize to multi-class and multi-attribute settings. While some work exists in the space of characterizing fair and cost-sensitive Bayes optimal classifiers for multi-class and multi-attribute settings [XZ24; Nar+24], studying their tradeoffs and susceptibility to data biases is an important future direction.

Bibliography

[AD22] Sushant Agarwal and Amit Deshpande. "On the power of randomization in fair classification and representation". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022, pp. 1542–1551 (cit. on pp. 10, 12, 56).

[Aga+] Sushant Agarwal, Amit Deshpande, Rajmohan Rajaraman, and Ravi Sundaram. "Optimal Fair Learning Robust to Adversarial Distribution Shift". In: *Forty-second International Conference on Machine Learning* (cit. on pp. 10, 12, 56).

[Aga+18] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. "A reductions approach to fair classification". In: *International conference on machine learning*. PMLR. 2018, pp. 60–69 (cit. on pp. 2, 10, 50, 52, 56, 59, 60, 70, 115).

[Akp+22] Nil-Jana Akpinar, Manish Nagireddy, Logan Stapleton, Hao-Fei Cheng, Haiyi Zhu, Steven Wu, and Hoda Heidari. "A sandbox tool to bias (Stress)-test fairness algorithms". In: *arXiv preprint arXiv:2204.10233* (2022) (cit. on pp. 16, 52, 59, 81).

[AL21] Ibrahim M Alabdulmohsin and Mario Lucic. "A near-optimal algorithm for debiasing trained machine learning models". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 8072–8084 (cit. on p. 12).

[Ang+16] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. "Machine bias". In: *ProPublica, May* 23.2016 (2016), pp. 139–159 (cit. on pp. 1, 53, 60).

[AT07] Jean-Yves Audibert and Alexandre B Tsybakov. "Fast learning rates for plug-in classifiers". In: (2007) (cit. on p. 3).

[AW23] Carolyn Ashurst and Adrian Weller. "Fairness without demographic data: A survey of approaches". In: *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. 2023, pp. 1–12 (cit. on p. 2).

[Awa+21] Pranjal Awasthi, Alex Beutel, Matthäus Kleindessner, Jamie Morgenstern, and Xuezhi Wang. "Evaluating fairness of machine learning models under uncertain and incomplete information". In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 2021, pp. 206–214 (cit. on p. 80).

[Bak+21] Chloé Bakalar, Renata Barreto, Stevie Bergman, Miranda Bogen, Bobbie Chern, Sam Corbett-Davies, Melissa Hall, Isabel Kloumann, Michelle Lam, Joaquin Quiñonero Candela, Manish Raghavan, Joshua Simons, Jonathan Tannen, Edmund Tong, Kate Vredenburgh, and Jiejing Zhao. "Fairness On The Ground: Applying Algorithmic Fairness Approaches to Production Systems". In: *CoRR* abs/2103.06172 (2021). arXiv: 2103.06172. URL: https://arxiv.org/abs/2103.06172 (cit. on pp. 70, 77, 96).

[BDDB23] Maarten Buyl, MaryBeth Defrance, and Tijl De Bie. "FAIRRET: a framework for differentiable fairness regularization terms". In: *arXiv preprint arXiv:2310.17256* (2023) (cit. on pp. 104, 115).

[Bel+18] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. *AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias*. Oct. 2018. URL: https://arxiv.org/abs/1810.01943 (cit. on pp. 2, 49, 52, 53, 59, 71).

[Bel+23a] Andrew Bell, Lucius Bynum, Nazarii Drushchak, Tetiana Zakharchenko, Lucas Rosenblatt, and Julia Stoyanovich. "The possibility of fairness: Revisiting the impossibility theorem in practice". In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*. 2023, pp. 400–422 (cit. on p. 10).

[Bel+23b] Nora Belrose, David Schneider-Joseph, Shauli Ravfogel, Ryan Cotterell, Edward Raff, and Stella Biderman. "Leace: Perfect linear concept erasure in closed form". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 66044–66063 (cit. on pp. 78, 79, 99).

[Ber+15] Visar Berisha, Alan Wisler, Alfred O Hero, and Andreas Spanias. "Empirically estimable classification bounds based on a nonparametric divergence measure". In: *IEEE Transactions on Signal Processing* 64.3 (2015), pp. 580–591 (cit. on p. 105).

[BG18] Joy Buolamwini and Timnit Gebru. "Gender shades: Intersectional accuracy disparities in commercial gender classification". In: *Conference on fairness, accountability and transparency*. PMLR. 2018, pp. 77–91 (cit. on pp. 1, 2, 50, 59, 70).

[Bha+23] Karan Bhanot, Ioana Baldini, Dennis Wei, Jiaming Zeng, and Kristin Bennett. "Stress-testing Bias Mitigation Algorithms to Understand Fairness Vulnerabilities". In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 2023, pp. 764–774 (cit. on p. 81).

[BHK22] Gavin Brown, Shlomi Hod, and Iden Kalemaj. "Performative prediction in a stateful world". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2022, pp. 6045–6061 (cit. on p. 26).

[BHN17] Solon Barocas, Moritz Hardt, and Arvind Narayanan. "Fairness in machine learning". In: *Nips tutorial* 1 (2017), p. 2017 (cit. on p. 1).

[BHN19] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and Machine Learning: Limitations and Opportunities*. http://www.fairmlbook.org. fairmlbook.org, 2019 (cit. on pp. 2, 9, 73).

[BHN23] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023 (cit. on p. 115).

[Bin20] Reuben Binns. "On the Apparent Conflict between Individual and Group Fairness". In: *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*. FAT* '20. Spain: Association for Computing Machinery, 2020, 514–524. ISBN: 9781450369367. URL: https://doi.org/10.1145/3351095.3372864 (cit. on p. 2).

[Bir+19] Sarah Bird, Krishnaram Kenthapadi, Emre Kiciman, and Margaret Mitchell. "Fairness-aware machine learning: Practical challenges and lessons learned". In: *Proceedings of the twelfth ACM international conference on web search and data mining*. 2019, pp. 834–835 (cit. on p. 1).

[Bir+20] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. "Fairlearn: A toolkit for assessing and improving fairness in AI". In: *Microsoft, Tech. Rep. MSR-TR-2020-32* (2020) (cit. on pp. 2, 59, 71).

[Bis+19] Arpita Biswas, Siddharth Barman, Amit Deshpande, and Amit Sharma. "Quantifying Infra-Marginality and Its Trade-off with Group Fairness". In: *arXiv preprint arXiv:1909.00982* (2019) (cit. on p. 15).

[BJW20] Yahav Bechavod, Christopher Jung, and Zhiwei Steven Wu. "Metric-Free Individual Fairness in Online Learning". In: *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*. 2020. URL: https://proceedings.neurips.cc/paper/2020/hash/80b618ebcac7aa97a6dac2ba65cb7e36-Abstract.html (cit. on p. 2).

[Blo+24] Christina Hastings Blow, Lijun Qian, Camille Gibson, Pamela Obiomon, and Xishuang Dong. "Comprehensive Validation on Reweighting Samples for Bias Mitigation via AIF360". In: *Applied Sciences* 14.9 (2024). ISSN: 2076-3417. DOI: 10.3390/app14093826. URL: https://www.mdpi.com/2076-3417/14/9/3826 (cit. on p. 71).

[Blu+23] Avrim Blum, Princewill Okoroafor, Aadirupa Saha, and Kevin Stangl. "On the Vulnerability of Fairness Constrained Learning to Malicious Noise". In: *arXiv preprint arXiv:2307.11892* (2023) (cit. on pp. 16, 26, 71).

[BM21] Arpita Biswas and Suvam Mukherjee. "Ensuring fairness under prior probability shifts". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 2021, pp. 414–424 (cit. on pp. 4, 16, 17, 25–27, 49, 51, 52, 58).

[Boy04] Stephen Boyd. "Convex optimization". In: *Cambridge UP* (2004) (cit. on p. 87).

[BR21a] Sumon Biswas and Hridesh Rajan. "Fair Preprocessing: Towards Understanding Compositional Fairness of Data Transformers in Machine Learning Pipeline". In: *arXiv preprint arXiv:2106.06054* (2021) (cit. on pp. 54, 59).

[BR21b] Sumon Biswas and Hridesh Rajan. "Fair preprocessing: towards understanding compositional fairness of data transformers in machine learning pipeline". In: *Proceedings of the 29th ACM Joint Meeting on European Software Engineering Conference and Symposium on the Foundations of Software Engineering*. ESEC/FSE 2021. Athens, Greece: Association for Computing Machinery, 2021, 981–993. ISBN: 9781450385626. DOI: 10.1145/3468264.3468536. URL: https://doi.org/10.1145/3468264.3468536 (cit. on p. 70).

[Bre+21] Boris van Breugel, Trent Kyono, Jeroen Berrevoets, and Mihaela van der Schaar. "DECAF: Generating Fair Synthetic Data Using Causally-Aware Generative Networks". In: *Advances in Neural Information Processing Systems* 34 (2021) (cit. on p. 81).

[BRV21] Mislav Balunović, Anian Ruoss, and Martin Vechev. "Fair Normalizing Flows". In: *arXiv preprint arXiv:2106.05937* (2021) (cit. on p. 81).

[BS16] Solon Barocas and Andrew D Selbst. "Big data's disparate impact". In: *Calif. L. Rev.* 104 (2016), p. 671 (cit. on p. 1).

[BS19] Avrim Blum and Kevin Stangl. "Recovering from biased data: Can fairness constraints improve accuracy?" In: *arXiv preprint arXiv:1912.01094* (2019) (cit. on pp. 3, 4, 9, 14–20, 25, 26, 35, 49–53, 58, 59, 71, 81).

[BW08] Peter L Bartlett and Marten H Wegkamp. "Classification with a Reject Option using a Hinge Loss." In: *Journal of Machine Learning Research* 9.8 (2008) (cit. on pp. 8, 15, 21).

[Cal+17] Flavio P Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. "Optimized pre-processing for discrimination prevention". In: *Proceedings of the 31st International Conference on Neural Information Processing Systems*. 2017, pp. 3995–4004 (cit. on pp. 2, 50, 71).

[Can23] Clément L. Canonne. *A short note on an inequality between KL and TV*. 2023. arXiv: 2202.07198 [math.PR]. URL: https://arxiv.org/abs/2202.07198 (cit. on pp. 75, 94).

[CDM16] Corinna Cortes, Giulia DeSalvo, and Mehryar Mohri. "Learning with rejection". In: *Algorithmic Learning Theory: 27th International Conference, ALT 2016, Bari, Italy, October 19-21, 2016, Proceedings 27*. Springer. 2016, pp. 67–82 (cit. on pp. 8, 15, 21).

[Cel+19] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. "Classification with fairness constraints: A meta-algorithm with provable guarantees". In: *Proceedings of the conference on fairness, accountability, and transparency*. 2019, pp. 319–328 (cit. on pp. 2, 10–12, 50, 106, 115).

[Cel+21] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. "Fair classification with noisy protected attributes: A framework with provable guarantees". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 1349–1361 (cit. on pp. 16, 71).

[Cer+24] Mattia Cerrato, Marius Köppel, Philipp Wolf, and Stefan Kramer. *10 Years of Fair Representations: Challenges and Opportunities*. 2024. arXiv: 2407.03834 [cs.LG]. URL: https://arxiv.org/abs/2407.03834 (cit. on p. 72).

[CGN19] Andrew Cotter, Maya Gupta, and Harikrishna Narasimhan. "On making stochastic classifiers deterministic". In: *Advances in Neural Information Processing Systems* 32 (2019) (cit. on p. 12).

[CH67] Thomas Cover and Peter Hart. "Nearest neighbor pattern classification". In: *IEEE transactions on information theory* 13.1 (1967), pp. 21–27 (cit. on p. 103).

[Cha+21] Nontawat Charoenphakdee, Zhenghang Cui, Yivan Zhang, and Masashi Sugiyama. "Classification with rejection based on cost-sensitive classification". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 1507–1517 (cit. on pp. 8, 21).

[Che+19] Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. "Fairness under unawareness: Assessing disparity when protected class is unobserved". In: *Proceedings of the conference on fairness, accountability, and transparency*. 2019, pp. 339–348 (cit. on p. 95).

[Cho+20] Kristy Choi, Aditya Grover, Trisha Singh, Rui Shu, and Stefano Ermon. "Fair generative modeling via weak supervision". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 1887–1898 (cit. on p. 81).

[Chz19] Evgenii Chzhen. "Plug-in methods in classification". PhD thesis. Université Paris-Est, 2019 (cit. on p. 3).

[Chz+19] Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. "Leveraging labeled and unlabeled data for consistent fair binary classification". In: *Advances in Neural Information Processing Systems* 32 (2019) (cit. on pp. 3, 4, 11, 12, 15, 16, 18, 31, 33, 49, 71, 105, 106, 117).

[CJS18] Irene Chen, Fredrik D Johansson, and David Sontag. "Why is my classifier discriminatory?" In: *Advances in neural information processing systems* 31 (2018) (cit. on p. 3).

[CJW22] Junyi Chai, Taeuk Jang, and Xiaoqian Wang. "Fairness without demographics through knowledge distillation". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 19152–19164 (cit. on p. 2).

[CKG23] Jiaee Cheong, Sinan Kalkan, and Hatice Gunes. "Causal Structure Learning of Bias for Fair Affect Recognition". In: *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 2023, pp. 340–349 (cit. on p. 16).

[CKL23] Wenlong Chen, Yegor Klochkov, and Yang Liu. "Post-hoc bias scoring is optimal for fair classification". In: *arXiv preprint arXiv:2310.05725* (2023) (cit. on pp. 11, 12, 49).

[Cor+08] Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. "Sample selection bias correction theory". In: *International conference on algorithmic learning theory*. Springer. 2008, pp. 38–53 (cit. on pp. 51, 58).

[Cos+19] Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. "Fair transfer learning with missing protected attributes". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 2019, pp. 91–98 (cit. on pp. 51, 58).

[Cot+19] Andrew Cotter, Heinrich Jiang, Maya Gupta, Serena Wang, Taman Narayan, Seungil You, and Karthik Sridharan. "Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals". In: *Journal of Machine Learning Research* 20.172 (2019), pp. 1–59 (cit. on p. 117).

[Cov99] Thomas M Cover. *Elements of information theory*. John Wiley & Sons, 1999 (cit. on p. 71).

[Cow+20] Bo Cowgill, Fabrizio Dell'Acqua, Samuel Deng, Daniel Hsu, Nakul Verma, and Augustin Chaintreau. "Biased Programmers? Or Biased Data? A Field Experiment in Operationalizing AI Ethics". In: *Proceedings of the 21st ACM Conference on Economics and Computation*. EC '20. Virtual Event, Hungary: Association for Computing Machinery, 2020, 679–681. ISBN: 9781450379755. DOI: 10.1145/3391403.3399545. URL: https://doi.org/10.1145/3391403.3399545 (cit. on p. 70).

[CR18] Alexandra Chouldechova and Aaron Roth. "The frontiers of fairness in machine learning". In: *arXiv preprint arXiv:1810.08810* (2018) (cit. on p. 1).

[CV10] Toon Calders and Sicco Verwer. "Three naive bayes approaches for discrimination-free classification". In: *Data mining and knowledge discovery* 21.2 (2010), pp. 277–292 (cit. on pp. 2, 50, 95).

[DA+19] Maria De-Arteaga, Alexey Romanov, Hanna Wallach, Jennifer Chayes, Christian Borgs, Alexandra Chouldechova, Sahin Geyik, Krishnaram Kenthapadi, and Adam Tauman Kalai. "Bias in bios: A case study of semantic representation bias in a high-stakes setting". In: *proceedings of the Conference on Fairness, Accountability, and Transparency*. 2019, pp. 120–128 (cit. on pp. 72, 78).

[Das+25] Nirjhar Das, Mohit Sharma, Praharsh Nanavati, Kirankumar Shiragur, and Amit Deshpande. "Cost Efficient Fairness Audit Under Partial Feedback". In: *arXiv preprint arXiv:2510.03734* (2025) (cit. on p. 4).

[DB20] Jessica Dai and Sarah M Brown. "Label bias, label shift: Fair machine learning with unreliable labels". In: *NeurIPS 2020 Workshop on Consequential Decision Making in Dynamic Environments*. Vol. 12. 2020 (cit. on pp. 15–17, 25, 26, 49, 51, 58).

[Del+24] Eoin Delaney, Zihao Fu, Sandra Wachter, Brent Mittelstadt, and Chris Russell. "Oxonfair: A flexible toolkit for algorithmic fairness". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 94209–94245 (cit. on pp. 2, 104, 115).

[Den+24] Christophe Denis, Romuald Elie, Mohamed Hebiri, and François Hu. "Fairness guarantees in multi-class classification with demographic parity". In: *Journal of Machine Learning Research* 25.130 (2024), pp. 1–46 (cit. on p. 12).

[DG17] Dheeru Dua and Casey Graff. *UCI Machine Learning Repository*. 2017. URL: http://archive.ics.uci.edu/ml (cit. on pp. 53, 60).

[DGL13] Luc Devroye, László Györfi, and Gábor Lugosi. *A probabilistic theory of pattern recognition*. Vol. 31. Springer Science & Business Media, 2013 (cit. on p. 103).

[DGL96] Luc Devroye, László Györfi, and Gábor Lugosi. "The Bayes Error". In: *A Probabilistic Theory of Pattern Recognition*. Springer, 1996, pp. 9–20. ISBN: 978-1-4612-0711-5. DOI: 10.1007/978-1-4612-0711-5_2 (cit. on pp. 7, 71, 73).

[Din+21] Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. "Retiring Adult: New Datasets for Fair Machine Learning". In: *arxiv preprint arxiv:2108.04884* (2021). URL: https://arxiv.org/abs/2108.04884 (cit. on p. 4).

[Don+18] Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. "Empirical risk minimization under fairness constraints". In: *Advances in neural information processing systems* 31 (2018) (cit. on pp. 3, 26, 70).

[DPNS14] Marthinus C Du Plessis, Gang Niu, and Masashi Sugiyama. "Analysis of learning from positive and unlabeled data". In: *Advances in neural information processing systems* 27 (2014) (cit. on p. 117).

[Dut+20] Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush Varshney. "Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing". In: *International conference on machine learning*. PMLR. 2020, pp. 2803–2813 (cit. on pp. 3–5, 9, 16, 71, 75, 80, 81).

[DW21] Wei Du and Xintao Wu. "Fair and Robust Classification Under Sample Selection Bias". In: *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2021, pp. 2999–3003 (cit. on pp. 4, 16, 49, 51, 58).

[Dwo+12a] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. "Fairness through awareness". In: *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*. ITCS '12. Cambridge, Massachusetts, 2012, 214–226. ISBN: 9781450311151 (cit. on pp. 2, 9, 16, 73).

[Dwo+12b] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. "Fairness through awareness". In: *Proceedings of the 3rd innovations in theoretical computer science conference*. 2012, pp. 214–226 (cit. on p. 9).

[Elk01] Charles Elkan. "The foundations of cost-sensitive learning". In: *International Joint Conference on Artificial Intelligence*. Vol. 17. 1. Lawrence Erlbaum Associates Ltd. 2001, pp. 973–978 (cit. on pp. 73, 83, 108).

[EP25] Daniel Eftekhari and Vardan Papyan. "On the Importance of Gaussianizing Representations". In: *arXiv preprint arXiv:2505.00685* (2025) (cit. on p. 80).

[FCG20] Riccardo Fogliato, Alexandra Chouldechova, and Max G'Sell. "Fairness evaluation in presence of biased noisy labels". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 2325–2336 (cit. on pp. 16, 51).

[Fel+15] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. "Certifying and removing disparate impact". In: *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 2015, pp. 259–268 (cit. on pp. 2, 9, 50).

[Fen+22] Qizhang Feng, Mengnan Du, Na Zou, and Xia Hu. "Fair Machine Learning in Healthcare: A Review". In: *arXiv preprint arXiv:2206.14397* (2022) (cit. on p. 52).

[Fle21] Will Fleisher. "What's Fair about Individual Fairness?" In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. Association for Computing Machinery, 2021, 480–490 (cit. on p. 2).

[Fou+20] James R Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. "An intersectional definition of fairness". In: *2020 IEEE 36th international conference on data engineering (ICDE)*. IEEE. 2020, pp. 1918–1921 (cit. on p. 9).

[FPV23] Vojtech Franc, Daniel Prusa, and Vaclav Voracek. "Optimal strategies for reject option classifiers". In: *Journal of Machine Learning Research* 24.11 (2023), pp. 1–49 (cit. on p. 21).

[Fri+19] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. "A comparative study of fairness-enhancing interventions in machine learning". In: *Proceedings of the conference on fairness, accountability, and transparency*. 2019, pp. 329–338 (cit. on pp. 1, 2, 50, 52, 54, 59, 60).

[FSV21] Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. "The (Im)possibility of fairness: different value systems require different mechanisms for fair decision making". In: *Commun. ACM* 64.4 (Mar. 2021), 136–143. ISSN: 0001-0782. DOI: 10.1145/3433949. URL: https://doi.org/10.1145/3433949 (cit. on pp. 4, 71).

[Fuk13] Keinosuke Fukunaga. *Introduction to statistical pattern recognition*. Elsevier, 2013 (cit. on p. 103).

[Gig+22] Stephen Giguere, Blossom Metevier, Bruno Castro da Silva, Yuriy Brun, Philip S Thomas, and Scott Niekum. "Fairness Guarantees under Demographic Shift". In: *International Conference on Learning Representations*. 2022 (cit. on pp. 51, 59).

[GK06] Anthony G Greenwald and Linda Hamilton Krieger. "Implicit bias: Scientific foundations". In: *California law review* 94.4 (2006), pp. 945–967 (cit. on pp. 50, 59).

[GKW23] Avijit Ghosh, Pablo Kvitca, and Christo Wilson. "When Fair Classification Meets Noisy Protected Attributes". In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. 2023, pp. 679–690 (cit. on pp. 16, 57, 115).

[Gol73] Gene H Golub. "Some modified matrix eigenvalue problems". In: *SIAM review* 15.2 (1973), pp. 318–334 (cit. on p. 87).

[Gra+24] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. "The Llama 3 herd of models". In: *arXiv preprint arXiv:2407.21783* (2024) (cit. on pp. 72, 79, 99).

[Han+23] Chi Han, Jialiang Xu, Manling Li, Yi Fung, Chenkai Sun, Nan Jiang, Tarek Abdelzaher, and Heng Ji. "Word embeddings are steers for language models". In: *arXiv preprint arXiv:2305.12798* (2023) (cit. on p. 78).

[HJ+18] Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. "Multicalibration: Calibration for the (computationally-identifiable) masses". In: *International conference on machine learning*. PMLR. 2018, pp. 1939–1948 (cit. on p. 72).

[Hoe63] Wassily Hoeffding. "Probability inequalities for sums of bounded random variables". In: *Journal of the American Statistical Association* 58.301 (1963), pp. 13–30 (cit. on p. 57).

[Hol+19] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. "Improving fairness in machine learning systems: What do industry practitioners need?" In: *Proceedings of the 2019 CHI conference on human factors in computing systems*. 2019, pp. 1–16 (cit. on p. 80).

[HPS16a] Moritz Hardt, Eric Price, and Nathan Srebro. "Equality of opportunity in supervised learning". In: *Proceedings of the 30th International Conference on Neural Information Processing Systems*. NIPS'16. 2016, 3323–3331. ISBN: 9781510838819 (cit. on pp. 2, 9, 10, 16, 73).

[HPS16b] Moritz Hardt, Eric Price, and Nati Srebro. "Equality of opportunity in supervised learning". In: *Advances in neural information processing systems* 29 (2016) (cit. on pp. 2, 50, 52).

[Hu+23] Lunjia Hu, Inbal Rachel Livni Navon, Omer Reingold, and Chutong Yang. "Omnipredictors for constrained optimization". In: *International Conference on Machine Learning*. PMLR. 2023, pp. 13497–13527 (cit. on p. 72).

[Hu+25] Xiaoyu Hu, Gengyu Xue, Zhenhua Lin, and Yi Yu. "Fairness-aware Bayes optimal functional classification". In: *arXiv preprint arXiv:2505.09471* (2025) (cit. on pp. 12, 49).

[Hur+24] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. "Gpt-4o system card". In: *arXiv preprint arXiv:2410.21276* (2024) (cit. on p. 79).

[HZ24] Xiaotian Hou and Linjun Zhang. "Finite-sample and distribution-free fair classification: Optimal trade-off between excess risk and fairness, and the cost of group-blindness". In: *arXiv preprint arXiv:2410.16477* (2024) (cit. on pp. 12, 49).

[ILY25] Takashi Ishida, Thanawat Lodkaew, and Ikko Yamane. "How Can I Publish My LLM Benchmark Without Giving the True Answers Away?" In: *arXiv preprint arXiv:2505.18102* (2025) (cit. on p. 3).

[INS18] Takashi Ishida, Gang Niu, and Masashi Sugiyama. "Binary classification from positive-confidence data". In: *Advances in neural information processing systems* 31 (2018) (cit. on p. 117).

[Ish+] Takashi Ishida, Ikko Yamane, Nontawat Charoenphakdee, Gang Niu, and Masashi Sugiyama. "Is the Performance of My Deep Network Too Good to Be True? A Direct Approach to Estimating the Bayes Error in Binary Classification". In: *The Eleventh International Conference on Learning Representations* (cit. on pp. 3, 5, 103, 104, 107–111, 115, 117).

[Isl+22] Maliha Tashfia Islam, Anna Fariha, Alexandra Meliou, and Babak Salimi. "Through the Data Management Lens: Experimental Analysis and Evaluation of Fair Classification". In: *Proceedings of the 2022 International Conference on Management of Data*. 2022, pp. 232–246 (cit. on pp. 16, 52).

[Jac21] Candace Jackson. "What is redlining?" In: *The New York Times* (2021), pp. 9–L (cit. on p. 1).

[JCD23] Minoh Jeong, Martina Cardone, and Alex Dytso. "Demystifying the optimal performance of multi-class classification". In: *Advances in Neural Information Processing Systems* 36 (2023), pp. 31638–31664 (cit. on p. 109).

[JCD24] Minoh Jeong, Martina Cardone, and Alex Dytso. "Data-driven estimation of the false positive rate of the Bayes binary classifier via soft labels". In: *2024 IEEE International Symposium on Information Theory (ISIT)*. IEEE. 2024, pp. 368–373 (cit. on pp. 104, 105, 107, 109–112, 115).

[JD25] Kevin Jiang and Edgar Dobriban. "Fair Classification by Direct Intervention on Operating Characteristics". In: *arXiv preprint arXiv:2509.25481* (2025) (cit. on pp. 3, 10, 12, 56).

[JG20] Eun Seo Jo and Timnit Gebru. "Lessons from archives: Strategies for collecting sociocultural data in machine learning". In: *Proceedings of the 2020 conference on fairness, accountability, and transparency*. 2020, pp. 306–316 (cit. on p. 80).

[JN20] Heinrich Jiang and Ofir Nachum. "Identifying and correcting label bias in machine learning". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2020, pp. 702–712 (cit. on pp. 16, 50–52, 59, 60, 71).

[Kam+12] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. "Fairness-aware classifier with prejudice remover regularizer". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer. 2012, pp. 35–50 (cit. on pp. 2, 50, 52, 95).

[KC12] Faisal Kamiran and Toon Calders. "Data preprocessing techniques for classification without discrimination". In: *Knowledge and Information Systems* 33.1 (2012), pp. 1–33 (cit. on pp. 2, 50, 52, 57–59, 67, 71, 75, 98).

[Kea+18] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness". In: *International Conference on Machine Learning*. PMLR. 2018, pp. 2564–2572 (cit. on p. 52).

[Khu+21] Drona Khurana, Srinivasan Ravichandran, Sparsh Jain, and Narayanan Unny Edakunni. "Statistical Guarantees for Fairness Aware Plug-In Algorithms". In: *arXiv preprint arXiv:2107.12783* (2021) (cit. on p. 105).

[KKZ12] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. "Decision theory for discrimination-aware classification". In: *2012 IEEE 12th International Conference on Data Mining*. IEEE. 2012, pp. 924–929 (cit. on pp. 2, 50, 52).

[KL22] Nikola Konstantinov and Christoph H Lampert. "On the Impossibility of Fairness-Aware Learning from Corrupted Data". In: *Algorithmic Fairness through the Lens of Causality and Robustness workshop*. PMLR. 2022, pp. 59–83 (cit. on pp. 14, 16, 26, 51).

[KMR16] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. "Inherent trade-offs in the fair determination of risk scores". In: *arXiv preprint arXiv:1609.05807* (2016) (cit. on pp. 3, 10).

[Koy+14] Oluwasanmi Koyejo, Nagarajan Natarajan, Pradeep Ravikumar, and Inderjit S Dhillon. "Consistent binary classification with generalized performance metrics". In: *Advances in neural information processing systems* 27 (2014) (cit. on pp. 3, 8, 73, 104, 110, 117).

[KPM24] Mark Kozdoba, Binyamin Perets, and Shie Mannor. "Efficient Fairness-Performance Pareto Front Computation". In: *arXiv preprint arXiv:2409.17643* (2024) (cit. on pp. 104, 115).

[KR18] Jon Kleinberg and Manish Raghavan. "Selection problems in the presence of implicit bias". In: *arXiv preprint arXiv:1801.03533* (2018) (cit. on pp. 4, 50, 59).

[KW19] Diederik P Kingma and Max Welling. "An introduction to variational autoencoders". In: *Foundations and Trends® in Machine Learning* 12.4 (2019), pp. 307–392 (cit. on p. 81).

[Lah+20] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. "Fairness without demographics through adversarially reweighted learning". In: *Advances in neural information processing systems* 33 (2020), pp. 728–740 (cit. on p. 2).

[Lam+19] Alex Lamy, Ziyuan Zhong, Aditya K Menon, and Nakul Verma. "Noise-tolerant fair classification". In: *Advances in Neural Information Processing Systems* 32 (2019) (cit. on pp. 16, 51).

[Lar+16] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. "How we analyzed the COMPAS recidivism algorithm". In: *ProPublica (5 2016)* 9.1 (2016), pp. 3–3 (cit. on p. 1).

[Lec+21] Tosca Lechner, Shai Ben-David, Sushant Agarwal, and Nivasini Ananthakrishnan. *Impossibility results for fair representations*. 2021. arXiv: 2107.03483 [cs.LG]. URL: https://arxiv.org/abs/2107.03483 (cit. on pp. 71, 74).

[Lic+13] Moshe Lichman et al. *UCI machine learning repository*. 2013 (cit. on pp. 104, 115).

[Liu+22] Ji Liu, Zenan Li, Yuan Yao, Feng Xu, Xiaoxing Ma, Miao Xu, and Hanghang Tong. "Fair Representation Learning: An Alternative to Mutual Information". In: *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. KDD '22. Washington DC, USA: Association for Computing Machinery, 2022, 1088–1097. ISBN: 9781450393850. DOI: 10.1145/3534678.3539302. URL: https://doi.org/10.1145/3534678.3539302 (cit. on p. 72).

[Lon+23] Carol Xuan Long, Hsiang Hsu, Wael Alghamdi, and Flavio P Calmon. "Arbitrariness lies beyond the fairness-accuracy frontier". In: *arXiv preprint arXiv:2306.09425* (2023) (cit. on p. 12).

[LRB25] Charlotte Leininger, Simon Rittel, and Ludwig Bothmann. "Overcoming Fairness Trade-offs via Pre-processing: A Causal Perspective". In: *arXiv preprint arXiv:2501.14710* (2025) (cit. on p. 81).

[LSH19] Lydia T Liu, Max Simchowitz, and Moritz Hardt. "The implicit fairness criterion of unconstrained learning". In: *International Conference on Machine Learning*. PMLR. 2019, pp. 4051–4060 (cit. on p. 10).

[Mad+18] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. "Learning Adversarially Fair and Transferable Representations". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, 2018, pp. 3384–3393. URL: https://proceedings.mlr.press/v80/madras18a.html (cit. on p. 72).

[Mad+19] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. "Fairness through causal awareness: Learning causal latent-variable models for biased data". In: *Proceedings of the conference on fairness, accountability, and transparency*. 2019, pp. 349–358 (cit. on p. 16).

[Mai+21] Subha Maity, Debarghya Mukherjee, Mikhail Yurochkin, and Yuekai Sun. "Does enforcing fairness mitigate biases caused by subpopulation shift?" In: *Advances in Neural Information Processing Systems* 34 (2021) (cit. on pp. 3, 4, 16, 49, 51, 81).

[May19] Sandra Gabriel Mayson. "Bias In, Bias Out". In: *Yale Law Journal* 2218 (2019). URL: https://www.yalelawjournal.org/pdf/Mayson_p5g2tz2m.pdf (cit. on p. 70).

[MB24] Chaitanya Murti and Chiranjib Bhattacharyya. "DisCEdit: Model Editing by Identifying Discriminative Components". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 47261–47296 (cit. on p. 81).

[MCR14] Sérgio Moro, Paulo Cortez, and Paulo Rita. "A data-driven approach to predict the success of bank telemarketing". In: *Decision Support Systems* 62 (2014), pp. 22–31 (cit. on pp. 53, 60).

[Meh+21] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. "A survey on bias and fairness in machine learning". In: *ACM Computing Surveys (CSUR)* 54.6 (2021), pp. 1–35 (cit. on pp. 1, 2, 50, 52).

[Men+15] Aditya Menon, Brendan Van Rooyen, Cheng Soon Ong, and Bob Williamson. "Learning from corrupted binary labels via class-probability estimation". In: *International conference on machine learning*. PMLR. 2015, pp. 125–134 (cit. on p. 26).

[MN06] Pascal Massart and Élodie Nédélec. "Risk bounds for statistical learning". In: *The Annals of Statistics* 34.5 (2006), pp. 2326–2366 (cit. on pp. 15, 19).

[MOW19] Daniel McNamara, Cheng Soon Ong, and Robert C. Williamson. "Costs and Benefits of Fair Representation Learning". In: *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '19. New York, NY, USA: Association for Computing Machinery, 2019, 263–270. ISBN: 9781450363242. DOI: 10.1145/3306618.3317964. URL: https://doi.org/10.1145/3306618.3317964 (cit. on p. 72).

[MPZ18] David Madras, Toni Pitassi, and Richard Zemel. "Predict responsibly: improving fairness and accuracy by learning to defer". In: *Advances in Neural Information Processing Systems* 31 (2018) (cit. on p. 21).

[MRT12] Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. Vol. 2. MIT press Cambridge, 2012 (cit. on pp. 3, 7, 103, 113).

[Muk+22] Debarghya Mukherjee, Felix Petersen, Mikhail Yurochkin, and Yuekai Sun. "Domain adaptation meets individual fairness. and they get along." In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 28902–28913 (cit. on p. 2).

[MW18] Aditya Krishna Menon and Robert C Williamson. "The cost of fairness in binary classification". In: *Conference on Fairness, accountability and transparency*. PMLR. 2018, pp. 107–118 (cit. on pp. 3, 4, 11, 12, 15, 16, 18, 33, 49, 71, 105, 106, 108).

[NA13] Harikrishna Narasimhan and Shivani Agarwal. "On the relationship between binary classification, bipartite ranking, and binary class probability estimation". In: *Advances in neural information processing systems* 26 (2013) (cit. on p. 3).

[Nar18] Arvind Narayanan. "Translation tutorial: 21 fairness definitions and their politics". In: *Proc. conf. fairness accountability transp., new york, usa*. Vol. 1170. 2018, p. 3 (cit. on p. 9).

[Nar+24] Harikrishna Narasimhan, Harish G Ramaswamy, Shiv Kumar Tavker, Drona Khurana, Praneeth Netrapalli, and Shivani Agarwal. "Consistent multiclass algorithms for complex metrics and constraints". In: *Journal of Machine Learning Research* 25.367 (2024), pp. 1–81 (cit. on p. 141).

[Nie14] Frank Nielsen. "Generalized Bhattacharyya and Chernoff upper bounds on Bayes error using quasi-arithmetic means". In: *Pattern Recognition Letters* 42 (2014), pp. 25–34. ISSN: 0167-8655. DOI: https://doi.org/10.1016/j.patrec.2014.01.002. URL: https://www.sciencedirect.com/science/article/pii/S0167865514000166 (cit. on p. 75).

[NXH19] Morteza Noshad, Li Xu, and Alfred Hero. "Learning to benchmark: Determining best achievable misclassification error from training data". In: *arXiv preprint arXiv:1909.07192* (2019) (cit. on p. 105).

[Pap+21] George Papamakarios, Eric Nalisnick, Danilo Jimenez Rezende, Shakir Mohamed, and Balaji Lakshminarayanan. "Normalizing flows for probabilistic modeling and inference". In: *Journal of Machine Learning Research* 22.57 (2021), pp. 1–64 (cit. on p. 81).

[PB22] Drago Plecko and Elias Bareinboim. "Causal fairness analysis". In: *arXiv preprint arXiv:2207.11385* (2022) (cit. on p. 16).

[PBM24] Drago Plečko, Nicolas Bennett, and Nicolai Meinshausen. "fairadapt: Causal Reasoning for Fair Data Preprocessing". In: *Journal of Statistical Software* 110.4 (2024), 1–35. DOI: 10.18637/jss.v110.i04. URL: https://www.jstatsoft.org/index.php/jss/article/view/v110i04 (cit. on p. 71).

[PCDG18] Emma Pierson, Sam Corbett-Davies, and Sharad Goel. "Fast Threshold Tests for Detecting Discrimination". In: *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*. Ed. by Amos Storkey and Fernando Perez-Cruz. Vol. 84. Proceedings of Machine Learning Research. PMLR, 2018, pp. 96–105. URL: https://proceedings.mlr.press/v84/pierson18a.html (cit. on pp. 74, 77, 96).

[Ped+11] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. "Scikit-learn: Machine learning in Python". In: *The Journal of Machine Learning Research* 12 (2011), pp. 2825–2830 (cit. on p. 52).

[Per+20] Juan Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. "Performative prediction". In: *International Conference on Machine Learning*. PMLR. 2020, pp. 7599–7609 (cit. on p. 26).

[Pet+19] Joshua C Peterson, Ruairidh M Battleday, Thomas L Griffiths, and Olga Russakovsky. "Human uncertainty makes classification more robust". In: *Proceedings of the IEEE/CVF international conference on computer vision*. 2019, pp. 9617–9626 (cit. on p. 115).

[Pet+21] Felix Petersen, Debarghya Mukherjee, Yuekai Sun, and Mikhail Yurochkin. "Post-processing for Individual Fairness". In: *Advances in Neural Information Processing Systems* 34 (2021) (cit. on p. 2).

[Pla+99] John Platt et al. "Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods". In: *Advances in large margin classifiers* 10.3 (1999), pp. 61–74 (cit. on p. 52).

[Ple+17] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. "On fairness and calibration". In: *arXiv preprint arXiv:1709.02012* (2017) (cit. on pp. 2, 3, 10, 12, 50, 52).

[PM20] Drago Plečko and Nicolai Meinshausen. "Fair Data Adaptation with Quantile Preservation". In: *Journal of Machine Learning Research* 21.242 (2020), pp. 1–44. URL: http://jmlr.org/papers/v21/19-966.html (cit. on p. 71).

[PMN21] Kenneth L Peng, Arunesh Mathur, and Arvind Narayanan. "Mitigating dataset harms requires stewardship: Lessons from 1000 papers". In: *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*. 2021. URL: https://openreview.net/forum?id=KGeAHDH4njY (cit. on p. 4).

[PP+08] Kaare Brandt Petersen, Michael Syskind Pedersen, et al. "The matrix cookbook". In: *Technical University of Denmark* 7.15 (2008), p. 510 (cit. on p. 88).

[PS22] Dana Pessach and Erez Shmueli. "A Review on Fairness in Machine Learning". In: *ACM Computing Surveys (CSUR)* 55.3 (2022), pp. 1–44 (cit. on p. 52).

[Qia+21] Shangshu Qian, Hung Pham, Thibaud Lutellier, Zeou Hu, Jungwon Kim, Lin Tan, Yaoliang Yu, Jiahao Chen, and Sameena Shah. "Are My Deep Learning Systems Fair? An Empirical Study of Fixed-Seed Training". In: *Advances in Neural Information Processing Systems* 34 (2021) (cit. on p. 52).

[Ren+21] Cedric Renggli, Luka Rimanic, Nora Hollenstein, and Ce Zhang. "Evaluating Bayes error estimators on real-world datasets with FeeBee". In: *arXiv preprint arXiv:2108.13034* (2021) (cit. on p. 105).

[Rez+21] Ashkan Rezaei, Anqi Liu, Omid Memarrast, and Brian D Ziebart. "Robust fairness under covariate shift". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 35. 11. 2021, pp. 9419–9427 (cit. on pp. 49, 51, 58).

[Rib+20] Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. "Beyond accuracy: Behavioral testing of NLP models with CheckList". In: *arXiv preprint arXiv:2005.04118* (2020) (cit. on p. 59).

[RR20] Ashesh Rambachan and Jonathan Roth. "Bias In, Bias Out? Evaluating the Folk Wisdom". In: *1st Symposium on Foundations of Responsible Computing (FORC 2020)*. Ed. by Aaron Roth. Vol. 156. Leibniz International Proceedings in Informatics (LIPIcs). Dagstuhl, Germany: Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2020, 6:1–6:15. ISBN: 978-3-95977-142-9. DOI: 10.4230/LIPIcs.FORC.2020.6. URL: https://drops.dagstuhl.de/entities/document/10.4230/LIPIcs.FORC.2020.6 (cit. on p. 70).

[RS94] R.L. Rivest and R. Sloan. "A Formal Model of Hierarchical Concept-Learning". In: *Information and Computation* 114.1 (1994), pp. 88–114 (cit. on pp. 15, 19).

[Sar20] Kailash Karthik Saravanakumar. "The Impossibility Theorem of Machine Fairness–A Causal Perspective". In: *arXiv preprint arXiv:2007.06024* (2020) (cit. on p. 10).

[SC21] Nicolas Schreuder and Evgenii Chzhen. "Classification with abstention but without disparities". In: *Uncertainty in Artificial Intelligence*. PMLR. 2021, pp. 1227–1236 (cit. on p. 21).

[Sch+19] Candice Schumann, Xuezhi Wang, Alex Beutel, Jilin Chen, Hai Qian, and Ed H Chi. "Transfer of machine learning fairness across domains". In: *arXiv preprint arXiv:1906.09688* (2019) (cit. on pp. 49, 51).

[Sch+22] Jessica Schrouff, Natalie Harris, Oluwasanmi Koyejo, Ibrahim Alabdulmohsin, Eva Schnider, Krista Opsahl-Ong, Alex Brown, Subhrajit Roy, Diana Mincu, Christina Chen, et al. "Maintaining fairness across distribution shift: do we have viable solutions for real-world applications?" In: *arXiv preprint arXiv:2202.01034* (2022) (cit. on pp. 14, 49).

[Sco12] Clayton Scott. "Calibrated asymmetric surrogate losses". In: *Electronic Journal of Statistics* 6.none (2012), pp. 958 –992. DOI: 10.1214/12-EJS699. URL: https://doi.org/10.1214/12-EJS699 (cit. on pp. 3, 8, 73, 83).

[SD24] Mohit Sharma and Amit Jayant Deshpande. "How far can fairness constraints help recover from biased data?" In: *Proceedings of the 41st International Conference on Machine Learning*. ICML'24. Vienna, Austria: JMLR.org, 2024 (cit. on p. 81).

[SDS23a] Mohit Sharma, Amit Deshpande, and Rajiv Ratn Shah. "On comparing fair classifiers under data bias". In: *arXiv preprint arXiv:2302.05906* (2023) (cit. on p. 115).

[SDS23b] Mohit Sharma, Amit Deshpande, and Rajiv Ratn Shah. "On Testing and Comparing Fair classifiers under Data Bias". In: *arXiv preprint arXiv:2302.05906* (2023) (cit. on pp. 16, 71, 81).

[Sha+17] Shreya Shankar, Yoni Halpern, Eric Breck, James Atwood, Jimbo Wilson, and D Sculley. "No classification without representation: Assessing geodiversity issues in open data sets for the developing world". In: *arXiv preprint arXiv:1711.08536* (2017) (cit. on pp. 1, 50, 58).

[Sha+22] Abhin Shah, Yuheng Bu, Joshua K Lee, Subhro Das, Rameswar Panda, Prasanna Sattigeri, and Gregory W Wornell. "Selective regression under fairness criteria". In: *International Conference on Machine Learning*. PMLR. 2022, pp. 19598–19615 (cit. on p. 21).

[Sin+21] Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. "Fairness violations and mitigation under covariate shift". In: *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. 2021, pp. 3–13 (cit. on pp. 49, 51, 58).

[Sin+24] Shashwat Singh, Shauli Ravfogel, Jonathan Herzig, Roee Aharoni, Ryan Cotterell, and Ponnurangam Kumaraguru. "Representation surgery: Theory and practice of affine steering". In: *arXiv preprint arXiv:2402.09631* (2024) (cit. on pp. 72, 78, 79, 97–99).

[SK22] Shashank Singh and Justin Khim. "Optimal Binary Classification Beyond Accuracy". In: *Advances in Neural Information Processing Systems*. Ed. by Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho. 2022. URL: https://openreview.net/forum?id=pm8Y8unXkkJ (cit. on pp. 3, 73, 78, 104, 110, 117).

[Slo96] Robert H. Sloan. "PAC Learning, Noise, and Geometry". In: *Learning and Geometry: Computational Approaches*. Birkhäuser Boston, 1996, pp. 21–41 (cit. on pp. 15, 19).

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge University Press, 2014 (cit. on p. 7).

[TAC23] Christopher TH Teo, Milad Abdollahzadeh, and Ngai-Man Cheung. "Fair generative models via transfer learning". In: *Proceedings of the AAAI conference on artificial intelligence*. Vol. 37. 2. 2023, pp. 2429–2437 (cit. on p. 81).

[TAC24] Christopher Teo, Milad Abdollahzadeh, and Ngai-Man Man Cheung. "On measuring fairness in generative models". In: *Advances in Neural Information Processing Systems* 36 (2024) (cit. on p. 81).

[Tan+24] Daniel Tan, David Chanin, Aengus Lynch, Brooks Paige, Dimitrios Kanoulas, Adrià Garriga-Alonso, and Robert Kirk. "Analysing the generalisation and reliability of steering vectors". In: *Advances in Neural Information Processing Systems* 37 (2024), pp. 139179–139212 (cit. on p. 78).

[TD23] Alexander Williams Tolbert and Emily Diana. "Correcting Underrepresentation and Intersectional Bias for Fair Classification". In: *arXiv preprint arXiv:2306.11112* (2023) (cit. on p. 26).

[The+21] Ryan Theisen, Huan Wang, Lav R Varshney, Caiming Xiong, and Richard Socher. "Evaluating state-of-the-art classification models against bayes optimality". In: *Advances in Neural Information Processing Systems* 34 (2021), pp. 9367–9377 (cit. on pp. 81, 105).

[Tou+23] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. "Llama 2: Open foundation and fine-tuned chat models". In: *arXiv preprint arXiv:2307.09288* (2023) (cit. on pp. 72, 78).

[UIS25] Ryota Ushio, Takashi Ishida, and Masashi Sugiyama. "Practical estimation of the optimal classification error with soft labels and calibration". In: *arXiv* (2025) (cit. on pp. 3, 5, 104, 107–111, 115, 117, 141).

[Var76] Hal R. Varian. "Two problems in the theory of fairness". In: *Journal of Public Economics* 5.3-4 (1976), pp. 249–260. DOI: None. URL: https://ideas.repec.org/a/eee/pubeco/v5y1976i3-4p249-260.html (cit. on p. 107).

[WH22] Ziwei Wu and Jingrui He. "Fairness-aware model-agnostic positive and unlabeled learning". In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*. 2022, pp. 1698–1708 (cit. on p. 117).

[WHM19] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. "Predictive inequity in object detection". In: *arXiv preprint arXiv:1902.11097* (2019) (cit. on pp. 1, 50, 59).

[WLL21] Jialu Wang, Yang Liu, and Caleb Levy. "Fair classification with group-dependent label noise". In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 2021, pp. 526–536 (cit. on pp. 4, 15–17, 25, 26, 51).

[WT+19] Michael Wick, Jean-Baptiste Tristan, et al. "Unlocking fairness: a trade-off revisited". In: *Advances in neural information processing systems* 32 (2019) (cit. on pp. 3, 4, 16, 81).

[WZC24] Joshua John Ward, Xianli Zeng, and Guang Cheng. "Fairrr: Pre-processing for group fairness through randomized response". In: *International Conference on Artificial Intelligence and Statistics*. PMLR. 2024, pp. 3826–3834 (cit. on pp. 12, 49).

[Xia+19] Xiaobo Xia, Tongliang Liu, Nannan Wang, Bo Han, Chen Gong, Gang Niu, and Masashi Sugiyama. "Are anchor points really indispensable in label-noise learning?" In: *Advances in neural information processing systems* 32 (2019) (cit. on p. 26).

[Xio+24] Zikai Xiong, Niccolò Dalmasso, Alan Mishler, Vamsi K. Potluru, Tucker Balch, and Manuela Veloso. "FairWASP: fast and optimal fair Wasserstein pre-processing". In: *Proceedings of the Thirty-Eighth AAAI Conference on Artificial Intelligence and Thirty-Sixth Conference on Innovative Applications of Artificial Intelligence and Fourteenth Symposium on Educational Advances in Artificial Intelligence*. AAAI'24/IAAI'24/EAAI'24. AAAI Press, 2024. ISBN: 978-1-57735-887-9. DOI: 10.1609/aaai.v38i14.29545. URL: https://doi.org/10.1609/aaai.v38i14.29545 (cit. on p. 71).

[Xu+18] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. "Fairgan: Fairness-aware generative adversarial networks". In: *2018 IEEE International Conference on Big Data (Big Data)*. IEEE. 2018, pp. 570–575 (cit. on p. 81).

[XWZ25] Ruicheng Xian, Yuxuan Wan, and Han Zhao. "Group Fairness Meets the Black Box: Enabling Fair Algorithms on Closed LLMs via Post-Processing". In: *arXiv preprint arXiv:2508.11258* (2025) (cit. on p. 141).

[XZ24] Ruicheng Xian and Han Zhao. "A unified post-processing framework for group fairness in classification". In: *arXiv preprint arXiv:2405.04025* (2024) (cit. on pp. 11, 12, 141).

[YHC25] Yi Yang, Yinghui Huang, and Xiangyu Chang. "Bayes-Optimal Fair Classification with Multiple Sensitive Features". In: *arXiv preprint arXiv:2505.00631* (2025) (cit. on pp. 12, 49).

[Zaf+17a] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. "Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment". In: *Proceedings of the 26th International Conference on World Wide Web (WWW)*. WWW'17. 2017, 1171–1180 (cit. on p. 2).

[Zaf+17b] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. "Fairness constraints: Mechanisms for fair classification". In: *Artificial Intelligence and Statistics*. PMLR. 2017, pp. 962–970 (cit. on pp. 2, 50, 60).

[ZDC22a] Xianli Zeng, Edgar Dobriban, and Guang Cheng. "Bayes-optimal classifiers under group fairness". In: *arXiv preprint arXiv:2202.09724* (2022) (cit. on pp. 16, 53, 60, 115).

[ZDC22b] Xianli Zeng, Edgar Dobriban, and Guang Cheng. "Fair bayes-optimal classifiers under predictive parity". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 27692–27705 (cit. on pp. 3, 12, 16, 26, 34, 71).

[Zem+13] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. "Learning fair representations". In: *International conference on machine learning*. PMLR. 2013, pp. 325–333 (cit. on pp. 2, 50, 72).

[Zen+26] Xianli Zeng, Kevin Jiang, Guang Cheng, and Edgar Dobriban. "Bayes-optimal fair classification with linear disparity constraints via pre-, in-, and post-processing". In: *Journal of Machine Learning Research* 27.65 (2026), pp. 1–87 (cit. on pp. 4, 7, 10–12, 48, 49, 103–107, 118).

[ZG19] Han Zhao and Geoff Gordon. "Inherent tradeoffs in learning fair representations". In: *Advances in neural information processing systems* 32 (2019), pp. 15675–15685 (cit. on p. 3).

[Zha+24] Haiyan Zhao, Heng Zhao, Bo Shen, Ali Payani, Fan Yang, and Mengnan Du. "Beyond single concept vector: Modeling concept subspace in llms with gaussian distribution". In: *arXiv preprint arXiv:2410.00153* (2024) (cit. on pp. 72, 78–80, 97, 98, 100, 101).

[Zhu+23] Jiongli Zhu, Sainyam Galhotra, Nazanin Sabri, and Babak Salimi. "Consistent Range Approximation for Fair Predictive Modeling". In: *Proceedings of the VLDB Endowment* 16.11 (2023), pp. 2925–2938 (cit. on pp. 16, 51).

[Zli15] Indre Zliobaite. "A survey on measuring indirect discrimination in machine learning". In: *arXiv preprint arXiv:1511.00148* (2015) (cit. on pp. 2, 50, 52).

[ZLL21] Zhaowei Zhu, Tianyi Luo, and Yang Liu. "The rich get richer: Disparate impact of semi-supervised learning". In: *arXiv preprint arXiv:2110.06282* (2021) (cit. on pp. 57, 68).

[ZLM18] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. "Mitigating unwanted biases with adversarial learning". In: *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*. 2018, pp. 335–340 (cit. on pp. 2, 50).

[ZSL21] Zhaowei Zhu, Yiwen Song, and Yang Liu. "Clusterability as an alternative to anchor points when learning with noisy labels". In: *International Conference on Machine Learning*. PMLR. 2021, pp. 12912–12923 (cit. on p. 26).