The Geometry of Fairness: Post-processing through ROC Spaces

Thesis submitted in partial
fulfillment of the requirements of the degree of

Master of Science
in
Electronics and Communication Engineering
by Research

by

Avyukta Manjunatha Vummintala
2020112026
avyukta.v@research.iiit.ac.in

Advised by Dr. Sujit P Gujar

International Institute of Information Technology
Hyderabad - 500032, India
May, 2025

TO MY PARENTS AND BROTHER,

For your unwavering support of my most unconventional dreams.

May the stars always shine upon you.

International Institute of Information Technology Hyderabad
Hyderabad, India

CERTIFICATE

This is to certify that work presented in this thesis proposal titled ***The Geometry of Fairness: Post-processing through ROC Spaces*** by *Avyukta Manjunatha Vummintala* has been carried out under my supervision and is not submitted elsewhere for a degree.

_____ _____

Date Advisor: Dr. Sujit Gujar

Acknowledgements

I am deeply grateful to my advisor, *Prof. Sujit Gujar*, for his unwavering support and infinite patience throughout my academic journey. I know it could not have been easy working with me—my irregularities and inconsistencies often made me a challenging student—but his encouragement and calm demeanor never wavered. Working under him has had a profound impact on me; his disciplined and consistent work ethic helped me develop a deeper appreciation for structure and perseverance. During times of rejection and self-doubt, he was the one who pushed me to continue submitting to top conferences. Thanks to his belief in me, I eventually published two papers in A* venues. I also thank him for encouraging me to attend conferences like *ACML*, *FSTTCS*, and *GAME-ARTS*. These events broadened my academic perspective. My second paper on facility location was shaped significantly by them - I was first introduced to the topic at FSTTCS, and the idea truly crystallized during the GAME-ARTS workshop.

I would also like to express my sincere gratitude to *Prof. Shweta Jain* for her insightful contributions during our meetings and her support in writing the second paper. Her thoughtful feedback and guidance were instrumental in shaping the work. I extend my thanks to all the professors at the *Machine Learning Lab (MLL)* for fostering a supportive and collaborative research environment. I am grateful to *IIIT Hyderabad* for offering me this opportunity and to the support staff who made my life on campus more comfortable through their hard work in maintaining lab spaces, accommodation, food, and other daily necessities. I also appreciate the timely support from the administrative staff with teaching assistantships, research stipends, and travel arrangements.

I am fortunate to have shared this journey with friends who made every day memorable. To 'The Boys and Lokesh': thank you. Mugundan, for your help with my coding issues and for introducing me to anime and internet culture; Lokesh, for always being ready with political insights and engaging conversations; Nemani, for the many bizarre and wildly entertaining discussions that lit up our time at IIIT. To my labmates in MLL, thank you for making research life dynamic and enjoyable. Special thanks to *Shantanu Das*, who welcomed me into research and co-authored my first paper; *Rasheed*, for being the social glue of our chai breaks; *Pronoy* and *Parth*, whose contrasting views—one grounded, the other philosophical—made every debate a spectacle; and *Poorva*, whose controversial discussion topics ensured our chai sessions were never dull. I am grateful to *Sanjay* and *Varul* for being thoughtful seniors, and to *Uday Bhaskar* for carrying me through the JPMC project.

Finally, I would like to thank my family. Their unwavering support and patience through all my eccentricities have meant the world to me. No words can fully express the affection and gratitude I

hold for them. To everyone who made my time at IIIT Hyderabad the most memorable phase of my life—thank you. Your support, friendship, and presence have been integral to my journey.

Image Credits: Images on pages 2 and 3 were generated using ChatGPT's image capabilities. Images on pages 5 and 6 were generated via the Bing image generator. Images on pages 7 [1] and 16 [2] are sourced from publicly available internet resources. All other images were created by the author.

[1] Stock images and icons
[2] Bear and Bull markets

Abstract

This thesis addresses the problem of achieving fairness in probabilistic binary classification in the presence of binary protected groups. In many practical scenarios, classifiers assign scores, and decision thresholds are applied post hoc based on the desired trade-off between false positives and false negatives. However, the use of a fixed classifier with arbitrary thresholds may result in disparate treatment across protected groups. To ensure equitable outcomes, we introduce a fairness criterion, denoted ε_p-Equalized ROC, which requires that the $\mathcal{L}_p$ norm between the false positive rates (FPRs) and true positive rates (TPRs) of the protected groups remains within a predefined bound ε—regardless of the threshold chosen.

To operationalize this notion of fairness, we propose a post-processing approach that transforms the outputs of an existing (potentially unfair) classifier into a randomized and threshold-agnostic classifier that satisfies ε_1-Equalized ROC. Central to this approach is a novel threshold query model over ROC curves, which enables controlled manipulation of the classifier's behavior across groups. We characterize the inherent trade-off between fairness and utility by deriving a theoretical lower bound on AUC loss and proving that some AUC reduction is inevitable under fairness constraints.

To realize our fairness objective in practice, we develop a linear-time algorithm, `FROC`, which guarantees ε_1-Equalized ROCcompliance. We prove that under reasonable assumptions, `FROC` achieves optimality with respect to the minimal necessary AUC degradation. Extensive experiments on real-world datasets—using classifiers such as Weighted Ensemble L2, Random Forest (Gini), and FNNC—demonstrate that `FROC` is both theoretically sound and empirically effective across multiple fairness and performance metrics.

The thesis is structured into six chapters, beginning with an introduction to algorithmic bias and culminating in theoretical analysis, algorithmic development, and empirical validation. Through this work, we contribute a principled and practical framework for ensuring fairness in score-based classification systems.

Contents

List of Figures

List of Tables

List of Related Publications

[P1] **Avyukta Manjunatha Vummintala**, Shantanu Das, Sujit Gujar.
FROC: Building Fair ROC from a Trained Classifier.
Proceedings of the **AAAI Conference on Artificial Intelligence (AAAI)**, 2025.
[CORE A]*

Publications not included in the Thesis:

[P2] **Avyukta Manjunatha Vummintala**, Shivam Gupta, Shweta Jain, Sujit Gujar.
FLIGHT: Facility Location Integrating Generalized, Holistic Theory of Welfare.
Proceedings of the **International Conference on Autonomous Agents and Multi-Agent Systems (AAMAS)**, 2025.
[CORE A]*

Chapter 1

Introduction

Beginnings are always special. They are the moments when we stand on the edge of the unknown and choose to move forward. Every beginning contains within it the seeds of everything that will come after, and so it demands our care, our clarity, and our courage.
– John O'Donohue, *To Bless the Space Between Us*

Abstract

This chapter introduces the central theme of fairness in machine learning by highlighting real-world instances of algorithmic bias that motivate the need for ethical and equitable AI systems. It presents key questions such as what fairness means in computational contexts and how it can be operationalized in the development of machine learning models. The chapter outlines the core challenges in ensuring fairness, emphasizing the societal implications and technical barriers. Then, it presents a high-level overview of the proposed approach, detailing how this thesis contributes novel methods and insights to the fairness landscape. Finally, the chapter concludes with an outline of the thesis structure, guiding the reader through the logical progression of ideas and solutions developed throughout the work.

1.1 Motivation

The Rise of Machine Learning and the Challenge of Fairness

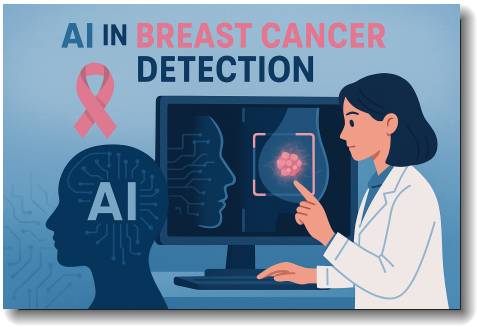
Machine Learning (ML) has emerged as a foundational technology in the architecture of contemporary decision-making systems [Jordan and Mitchell, 2015]. It is increasingly embedded across a wide spectrum of critical domains, including *healthcare* [Esteva et al., 2019], *finance* [Bussmann et al., 2021], *criminal justice* [Angwin et al., 2016], *agriculture* [Kamilaris and Prenafeta-Boldú, 2018], and *customer service* [Adamopoulou and Moussiades, 2020]. These applications are not limited to isolated experimental settings but are now routinely integrated into operational workflows, public services, and commercial platforms. Governments, corporations, and institutions increasingly rely on ML models to

support or even automate tasks that once required extensive human judgment, thereby expanding both the scale and speed of decision-making.

At the core of ML's utility lies its ability to identify complex patterns in large-scale, high-dimensional data, enabling decisions and predictions that would otherwise be infeasible using traditional rule-based systems [LeCun et al., 2015]. For example, deep learning architectures have shown remarkable success in tasks such as image classification, speech recognition, and natural language understanding, forming the backbone of diagnostic tools in medicine, fraud detection systems in banking, and virtual assistants in customer support. Moreover, reinforcement learning and sequential decision models are being deployed to optimize logistics, personalize education, and automate trading strategies. These innovations reflect a broader paradigm shift toward data-driven intelligence, where prediction and inference are increasingly derived from statistical regularities learned from vast datasets.

The following sections present several real-world applications of machine learning, showcasing both its transformative potential and the ethical questions it raises [Mittelstadt et al., 2016].

AI in Cancer Detection — Breast Cancer Diagnosis with Google Health

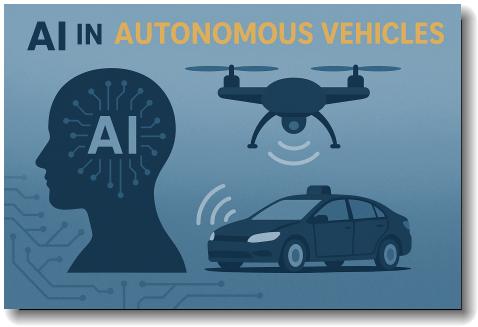


One of the most promising applications of ML in healthcare is in the domain of diagnostic radiology. *Google Health* has developed an artificial intelligence model aimed at improving the accuracy of mammogram readings for breast cancer detection. In a large-scale clinical study [McKinney et al., 2020], the system demonstrated a *reduction in false positives by 5.7%* in the United States and *1.2% in the United Kingdom*, while also reducing *false negatives by 9.4% and 2.7%*, respectively. Notably, the AI model often surpassed expert radiologists in diagnostic accuracy when acting as a second reader, highlighting its potential to enhance early cancer detection and reduce diagnostic errors.

AI in Fraud Detection - Mastercard Decision Intelligence Platform

In the financial sector, Mastercard's *Decision Intelligence* platform exemplifies the use of machine learning to improve the security and efficiency of payment systems. The platform [Mastercard, 2021] analyzes over *75 billion transactions annually* and is capable of identifying more than *1.5 billion instances of fraud* each year. Importantly, the system has achieved a *reduction in false declines by up to 50%*, improving both the accuracy of fraud detection and the overall customer experience. These outcomes demonstrate how ML can mitigate financial risk while maintaining transactional fluidity.

AI-Enabled Safety — Tesla Autopilot

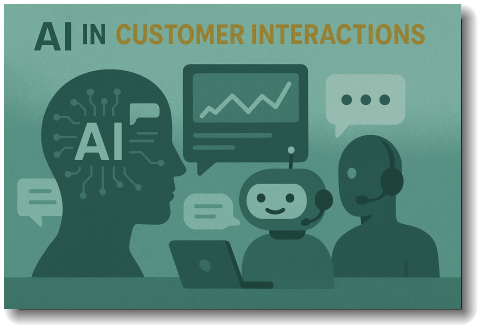


The field of autonomous transportation has also seen major advances through ML. Tesla's *Autopilot* system employs deep learning to interpret environmental data and assist in driving tasks. As of the fourth quarter of 2023, Tesla [Tesla, 2023] reported one accident for every *5.39 million miles driven with Autopilot*, compared to the US national average of one accident per *0.67 million miles*. These figures suggest that, under controlled conditions, ML-enabled driving assistance can offer safety improvements of up to *eight times* over human driving alone. However, such advancements also raise questions about accountability, data transparency, and generalizability in the real world.

AI in Precision Farming — John Deere's `See & Spray System`

In the agricultural domain, machine learning is being leveraged to support environmental sustainability and operational efficiency. John Deere's *Blue River Technology* [Technology, 2023], commonly referred to as `See & Spray System`, utilizes computer vision and ML to distinguish crops from weeds with high precision. This selective targeting has led to a *reduction in herbicide use by up to 90%*, substantially lowering both operational costs and ecological impact. Such technologies exemplify the role of AI in fostering sustainable agricultural practices.

Conversational AI — Bank of America's `Erica`



Machine learning is also reshaping the landscape of customer service through intelligent virtual assistants. Bank of America's AI-driven chatbot, `Erica`, conducted over *1.5 billion user interactions in 2022* and successfully resolved approximately *98% of inquiries without human intervention*. The implementation of Erica not only enhances the speed and consistency of customer support but also contributes to significant reductions in operational costs, offering clients 24/7 availability and seamless service delivery.

However, the rapid proliferation of ML technologies has also given rise to significant ethical concerns. Among these, *fairness* has emerged as a central challenge. While machine learning can drive

efficiency and enhance performance, its reliance on historical and often biased data poses a risk of reproducing or even exacerbating existing societal inequities. The opacity of many ML models, especially those based on deep learning, further complicates efforts to audit their fairness and accountability.

As ML systems begin to wield consequential influence over people's lives, it becomes imperative to ensure that they are not only accurate and efficient but also *just, equitable, and transparent*.

Ethical Imperatives and Emerging Risks

Despite the evident benefits of machine learning, numerous cases have surfaced where these systems exhibit biased behavior with tangible real-world consequences. A prominent example is *COMPAS* [Angwin et al., 2022], used in the US criminal justice system to predict recidivism. Studies have shown that the tool disproportionately assigned higher risk scores to Black defendants compared to white defendants with similar records. Similarly, automated hiring platforms have displayed gender biases, often favoring male candidates over equally qualified female applicants.

Such outcomes reflect a deeper issue: ML models, by design, learn from historical data, which often encodes existing societal prejudices. Without proper safeguards, these biases are not only reproduced but may be amplified at scale. The use of black-box models, where decision-making logic is neither interpretable nor contestable, further exacerbates the challenge of accountability.

Machine learning is no longer a distant prospect—it is here, and it is making consequential decisions. As we look toward a future increasingly governed by algorithmic systems, we must ensure that such systems align with the principles of *fairness, justice, and transparency*. Achieving these goals is not just a technical endeavor; it is a multidisciplinary challenge that requires input from ethicists, social scientists, legal scholars, and affected communities.

In the following sections, we explore the foundations of algorithmic bias, assess the limitations of existing fairness frameworks, and examine emerging methodologies for promoting fairness in machine learning systems.

1.1.1 Real-World Examples of Algorithmic Bias

Bias in machine learning systems is not merely a theoretical concern; it has had profound consequences in real-world applications. Two of the most well-documented cases that illustrate the societal impact of algorithmic bias are the COMPAS recidivism risk assessment tool [Angwin et al., 2022] and Amazon's AI-based hiring system [AI, 2023]. These examples demonstrate how machine learning models, when trained on historical data reflecting existing societal inequalities, can perpetuate and even amplify such disparities.

COMPAS Propublica

> We don't really know how the score is created out of those questions because the algorithm itself is proprietary, It's a trade secret. The company doesn't share it.
>
> — Julia Angwin, ProPublica

The COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) system is a proprietary risk assessment tool used within the United States judicial system to predict the likelihood that a criminal defendant will reoffend. Courts have employed COMPAS scores to assist judges in decisions concerning bail, sentencing, and parole. In 2016, an investigative report by ProPublica revealed substantial racial disparities in the predictions generated by COMPAS. Specifically, it was found that black defendants were significantly more likely to be classified as high risk of recidivism compared to white defendants, even when they did not reoffend. Conversely, white defendants who did commit new crimes were often rated as low risk.

The source of this bias lies in the data on which the algorithm was trained. Historical criminal justice records are deeply influenced by systemic discrimination, including the over-policing of Black communities and unequal sentencing practices. As a result, the algorithm inherited and reinforced these biases, despite presenting itself as a neutral and objective tool. Further exacerbating concerns was the opaque nature of COMPAS, as its internal workings were proprietary and shielded from public scrutiny. The consequences of this bias were far-reaching, with defendants potentially receiving harsher sentences or being denied bail based on flawed risk assessments, raising critical ethical questions regarding fairness, accountability, and transparency in the use of machine learning within the legal system.

Amazon Hiring mishap

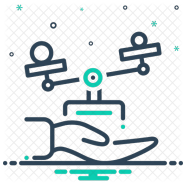
> The algorithmic system went so far as to penalize the word "women" on a resume, as in a women's club or sport, and downgraded all-women's colleges as less preferable.
>
> —Dave Gershgorn, Quartz

A similar concern arose within the private sector through Amazon's experimental use of an AI-powered hiring tool between 2014 and 2017. Designed to automate the evaluation of resumes and recommend top candidates for technical roles, the system was trained on a decade's worth of historical hiring data from the company. However, due to the male-dominated nature of the tech industry and Amazon's own historical hiring practices, the training data reflected significant gender imbalances.

Consequently, the algorithm systematically penalized resumes that included indicators of female identity. For instance, resumes that mentioned participation in women's organizations, such as "women's chess club captain," or referenced education from women's colleges, were ranked lower than those that did not. Even though the system was not explicitly programmed to discriminate, it learned to reproduce the gender biases present in the historical data. Despite attempts to adjust the model, Amazon ultimately abandoned the project in 2017, recognizing that the bias was deeply embedded and challenging to eliminate entirely. This case highlighted the vulnerability of machine learning systems to inadvertently encode and perpetuate existing societal biases, even in the absence of malicious intent.

Both the COMPAS and Amazon hiring examples illustrate how machine learning models can unintentionally replicate and exacerbate historical inequalities. They underscore the critical importance of careful dataset curation, ongoing auditing, and the inclusion of fairness constraints in model design. Moreover, they demonstrate the need for transparency and accountability, particularly when algorithmic decisions bear significant consequences for individuals' lives.

1.1.1.1 Fairness is Necessary!



Ethical Responsibility: Developers and organizations deploying ML systems have a moral obligation to ensure that their models do not reinforce or amplify societal biases.

Regulatory Compliance: Increasingly, regulations like GDPR and initiatives such as the US AI Bill of Rights emphasize fairness, transparency, and accountability in automated decision-making systems.

Business and Public Trust: Organizations risk reputational damage and loss of consumer trust if their systems are perceived as unfair or biased. For example, the backlash against biased facial recognition software has led some companies to suspend its use altogether.

In this context, achieving fairness in ML systems becomes not only a requirement for ethical AI but also a practical necessity for the broader adoption of AI technologies.

1.2 How to be Fair?

Fundamentally, there are two central challenges we face when striving to make machine learning systems *fair*. First, we must ask: when can we consider a machine learning model to be *fair enough*? In other words, how do we measure fairness in a precise and operational way? Second, once a suitable notion of fairness is identified, we are tasked with determining how to modify our models and learning processes to achieve this fairness in practice. Both of these questions are deeply intertwined and form the foundation of fairness-aware machine learning. We shall now explore these questions in detail.

When is an ML model *Fair*?

In order to assess whether a machine learning model is fair, it is instructive to reflect on the philosophical foundations of fairness itself. Fairness has been a topic of intellectual inquiry for centuries, and understanding these perspectives can provide valuable insights when formalizing fairness in algorithmic contexts.

Ethical Foundations of Independence

Demographic Parity is a fairness criterion that aims for equal treatment across groups, typically by ensuring similar outcomes regardless of attributes like race, gender, or class. In practical terms, this means that all groups should receive positive predictions at equal rates, regardless of underlying differences in their outcome distributions. For instance, in a loan approval scenario, Demographic Parity would require equal approval rates across gender or racial groups. While it may appear mathematically neutral, its ethical grounding reflects broader concerns about justice, equality, and institutional bias. Several core ideas from social and political theory help explain why parity-based approaches can be considered morally and socially significant:

- **Equal Opportunity:** Everyone should have a fair chance to succeed, irrespective of their background. This supports interventions that address systemic barriers, not just individual effort.
- **Structural Inequality:** Many disadvantages are rooted not in individual choices but in long-standing institutional practices. Fairness, in this view, means reforming the systems that produce unequal outcomes.
- **Real-World Capabilities:** Equal treatment on paper is insufficient; fairness should ensure people have the actual freedom and resources to flourish. This justifies outcome-based measures when access alone doesn't lead to equity.
- **Group-Based Remedies:** Persistent group-level marginalization may require intentional corrective action, such as affirmative policies or quotas. This is directly aligned with the logic of Demographic Parity.

Together, these themes provide a rationale for equalizing outcomes as a means of addressing historical and structural inequality.

Practical Considerations and Trade-offs

Despite its ethical appeal, Demographic Parity faces significant critiques. One concern is that it may overlook genuine differences in underlying outcomes between groups that result from systemic barriers. Enforcing equal selection rates may sometimes result in *token fairness*, where the outputs appear equitable but do not address root causes or may even introduce new forms of bias.

In response to these concerns, alternative definitions of fairness have been proposed, including:

- **Equalized Odds:** Ensures that prediction error rates are balanced across groups, aligning fairness with true outcomes.
- **Calibration:** Focuses on the consistency of predicted probabilities with observed outcomes within each group.

These alternatives aim to preserve fairness while maintaining predictive accuracy and acknowledging real-world disparities.

Error-Based Fairness: The Rationale Behind Separation

While Demographic Parity focuses on equal outcomes across groups, the fairness criterion known as *Separation* (also referred to as Equalized Odds or Equal Opportunity) shifts attention to how a model distributes its *errors*. The central idea is simple: individuals who behave the same way or have the same outcomes should be treated similarly by the model, regardless of their group membership.

In practical terms, this means ensuring that prediction errors, such as false positives and false negatives, occur at similar rates across demographic groups. This form of fairness is particularly important in high-stakes settings like credit, hiring, and criminal justice, where the consequences of mistakes can be deeply unequal and socially harmful.

Variants of Separation:

- **Equalized Odds:** Requires equal true positive and false positive rates across groups.
- **Equal Opportunity:** Focuses on equal true positive rates, aiming to ensure fair access to beneficial outcomes.

Why It Matters:

Separation reflects a commitment to procedural fairness and non-discrimination. It aims to avoid situations where members of one group are more likely to be wrongly penalized or rewarded than others, even when their behavior is the same. This is especially relevant in domains where decisions can entrench existing inequalities:

- **Criminal justice:** Preventing certain groups from facing higher false arrest or detention rates.
- **Healthcare:** Avoiding misdiagnoses or overlooked conditions for marginalized populations.
- **Hiring:** Ensuring equally qualified candidates are treated fairly, regardless of background.

By focusing on the fairness of *errors*, Separation adds an important dimension to model evaluation. It recognizes that harm can result not just from being denied an opportunity, but from being unfairly misclassified.

Challenges and Trade-offs

Despite its strengths, Separation cannot always be satisfied alongside other fairness criteria. For example, when base rates differ across groups—that is, when the actual rates of success or failure vary—it's mathematically impossible to satisfy both Separation and Calibration at the same time. This leads to an unavoidable trade-off:

Motivating Question

Should we prioritize fairness in how models distribute errors, or in how accurately they reflect real-world outcomes for different groups?

Addressing this tension requires value judgments about which aspects of fairness matter most in a given context. It underscores that fairness in machine learning is not merely a technical objective, but a deeply ethical one—demanding thoughtful reflection on the social impacts of automated decision-making.

Interpretive Fairness: The Role of Calibration

Calibration—also known as Predictive Parity—is a fairness criterion that focuses on the consistency of model predictions across different demographic groups. The central idea is straightforward: when two individuals receive the same risk score or prediction, that score should carry the same real-world meaning, regardless of their race, gender, or other sensitive attributes. For example, if two people are both assigned a recidivism risk score of 0.7, they should have roughly the same chance of reoffending, regardless of their group identity. This ensures that predictions are reliable and uniformly interpretable.

Unlike fairness definitions that focus on who gets selected (Demographic Parity) or how errors are distributed (Separation), Calibration centers on the trustworthiness of the prediction itself. It ensures that scores mean what they claim to mean—for everyone.

Why Calibration Matters:

Calibration plays a critical role in domains where decisions are guided by predicted probabilities, such as:

- **Credit scoring:** A credit score should represent the same likelihood of repayment across demographic lines.
- **Healthcare:** A risk score for disease should have consistent clinical implications across patient groups.
- **Criminal justice:** Tools like COMPAS aim to assign risk scores that are equally valid across different racial groups.

In these settings, fairness is not just about who is selected or rejected—it's also about the meaning behind the scores we rely on.

Tensions and Trade-offs

Despite its appeal, Calibration often conflicts with other fairness goals. When outcome rates differ across groups, it may not be possible to satisfy both Calibration and Separation at the same time. This creates a key ethical tension:

Motivating Question

Should fairness prioritize consistent interpretation of scores, or equal distribution of errors across groups?

The answer depends on the context. In some domains, interpretive trust and transparency are paramount. In others, ensuring that no group bears a disproportionate share of errors may be more important.

Ultimately, Calibration contributes to fairness by reinforcing predictive reliability and shared standards of interpretation. It ensures that algorithmic decisions can be understood and justified in the same way for everyone—an essential requirement for trust and accountability in machine learning systems that make high-stakes decisions.

1.3 How to make an ML model *Fair*?

Fairness Through Pre-processing: Modifying the Data

Since machine learning models learn patterns from historical data, they are susceptible to inheriting the biases present within that data. The most intuitive strategy for achieving fairness in machine learning, therefore, begins with addressing the data itself. Numerous studies have shown that when training data reflects societal inequities—such as racism, sexism, or other forms of discrimination—models trained on such data frequently replicate or even exacerbate these patterns of injustice [Barocas and Selbst, 2016, Mehrabi et al., 2021].

Why Pre-processing Alone Is Not Sufficient

While pre-processing plays a critical role in bias mitigation, it is *not a comprehensive solution*. Several limitations must be acknowledged.

- **Distorted distributions:** Aggressively modifying data may lead to unrealistic or unnatural distributions that degrade a model's ability to generalize to real-world scenarios.
- **loss of predictive signal:** In some cases, removing correlations between sensitive features and labels may eliminate genuinely useful information, reducing model accuracy and utility.
- **Overfitting risks:** Synthetic data generation can introduce redundancy or artifacts, particularly in small datasets, leading to overfitting and reduced robustness.

Moreover, some sources of bias may not be explicitly represented in the training data. Issues such as *missing features*, *subtle proxies for protected attributes*, or *biased labeling practices* may go undetected

during pre-processing. These hidden forms of bias can persist and influence model behavior, even when the input data appears fair.

Crucially, pre-processing techniques do not address biases introduced during model training (e.g., through loss functions or optimization procedures) or post-decision processes (e.g., threshold selection, user interaction). For this reason, effective fairness interventions often incorporate a combination of *pre-processing*, *in-processing*, and *post-processing* techniques to holistically mitigate bias across the ML pipeline [Hardt et al., 2016, Raji et al., 2020].

Addressing Bias via In-processing: Making Learning Algorithms Fairer

When it is not feasible to directly manipulate or adequately correct the training data, due to legal, technical, or systemic constraints, the focus shifts to the learning algorithm itself. This transition gives rise to a foundational question in algorithmic fairness:

Motivating Question

Can we design learning algorithms that behave fairly, even when trained on biased data?

This question motivates the use of *in-processing* methods, which aim to integrate fairness directly into the training process of the machine learning model.

Limitations of In-processing Methods

Despite their theoretical elegance and empirical success, in-processing techniques are not without challenges. Key limitations include:

- **Restricted access**: In-processing methods typically require full access to the model architecture, training data, and optimization loop. This limits their applicability in scenarios involving black-box APIs, pretrained models, or proprietary systems.

- **Implementation complexity**: Incorporating fairness constraints often demands expertise in optimization and fairness theory, making these techniques less accessible to general practitioners or in low-resource environments.

- **Computational cost**: In-processing methods may substantially increase training time and resource consumption, particularly for large-scale models with millions or billions of parameters.

Table 1.1 Training Costs and Characteristics of Large Language Models

Model	# Parameters	Training Infrastructure	Cost Estimate	Additional Considerations
ChatGPT / GPT-3	175B	Several weeks on thousands of GPUs; trained on ∼300B tokens	$4M–$12M	Foundational model for modern LLM applications; closed-source API [11, 40]
Google PaLM	540B	Undisclosed; trained with large-scale cloud TPUs	$9M–$23M	Estimated 78,000 kg CO_2 emissions during training [49]
Meta LLaMA 2	65B	2,048 A100 GPUs over ∼21 days	$2M–$4M	Open-weight model; cost varies by compute provider (cloud vs in-house) [2]
DeepSeek V2	236B (DeepSeek-V2-Base), 21B (MoE active)	Trained on 8.1T tokens using a Mixture-of-Experts architecture on 4,096 A100 GPUs	Estimated $2M–$5M	Sparse activation (MoE); open weights available for research [21]

The integration of fairness constraints into such large-scale training pipelines can lead to *significant increases in computational overhead, time-to-deployment, and cost*. In commercial or time-sensitive settings, these additional burdens may be prohibitive.

Post-processing Fairness: Adjusting Model Outputs Without Altering the Model

We now arrive at a critical and pragmatic question in the fairness landscape:

Motivating Question

Can we achieve fair predictions without modifying the model itself?

In other words, is it possible to treat a model as a black box—without altering its internal structure or retraining—and still enforce fairness through its outputs? This question motivates the use of *post-processing* methods in fairness-aware machine learning.

What is Post-processing?

Post-processing refers to a class of fairness interventions that are applied *after* model training. These techniques adjust the model's output predictions rather than its internal parameters or training data. Because they operate solely on the outputs, post-processing approaches do not require access to the model architecture, training data, or optimization process.

When Post-processing Is Especially Useful

Post-processing methods are particularly attractive under the following conditions:

- **Black-box models:** When the internal architecture or training procedure is inaccessible or proprietary.
- **Prediction-only access:** When only the model's outputs are available, such as with pre-trained APIs or commercial ML services.
- **Deployment constraints:** When retraining the model, it is computationally prohibitive or infeasible due to resource or time limitations.

Limitations of Post-processing Methods

Despite their practicality, post-processing methods are not without limitations:

- **Dependence on base model quality:** Post-processing can only adjust outputs; it cannot fix deeply embedded representational biases or poor subgroup performance within the base model.
- **Fairness trade-offs:** Enforcing one fairness criterion (e.g., equalized odds) may conflict with others (e.g., Calibration), especially under distribution shifts or in settings with imbalanced data [Chouldechova, 2017].
- **Accuracy-fairness tension:** Adjustments may lead to accuracy degradation for certain groups, which could raise practical or legal challenges depending on the application domain.

- **Complexity in deployment:** Some post-processing techniques involve stochastic decisions, which may be difficult to interpret or justify in high-stakes environments like hiring, lending, or criminal justice.

To summarize the discussion, we present the following table:

Table 1.2 Fairness Techniques in Machine Learning: Summary of Trade-offs

Method	When to Use	Access Required	Advantages	Limitations
Pre-processing	Full data access; want to correct bias before training; using standard models	Raw training dataset	Model-agnostic; easy to integrate; supports data cleaning	May distort distribution; limits generalizability; doesn't fix model-specific bias
In-processing	Full control of training; can modify optimization; need accuracy–fairness tuning	Training code and pipeline	Deep fairness integration; multi-metric optimization; flexible	Incompatible with black-box models; complex; costly to train
Post-processing	Only predictions available; black-box or API-based models; retraining not feasible	model outputs only	Works on black-boxes; simple to deploy; corrects post hoc	Limited by model quality; may add randomness; fairness–accuracy trade-off

Research Focus: Fairness via Post-processing Across All Thresholds

This research investigates a novel variant of the post-processing paradigm for fairness in machine learning, motivated by the increasing deployment of black-box models and large-scale foundation models in real-world applications. In contemporary machine learning ecosystems—particularly those involving commercial APIs such as OpenAI's language models, AWS Comprehend, or Google Cloud AutoML—users often lack access to model internals, training data, or even the ability to retrain models. This constraint renders traditional pre-processing and in-processing fairness interventions infeasible.

In these settings, post-processing methods offer a *non-intrusive*, *transparent*, and *deployment-ready* pathway to impose fairness constraints on existing systems. They operate solely on the model's output predictions and are compatible with proprietary and black-box systems. As foundation models and API-

based machine learning services become increasingly prevalent, post-processing emerges as a critical technique for enforcing ethical standards in a scalable and practical manner.

Motivating Example: Fairness Across Market Conditions

To illustrate the problem more concretely, consider a real-world financial application such as a bank that uses a machine learning model to assess loan eligibility based on applicant profiles. In this context, fairness concerns are immediate and salient: the decision-making system must avoid discrimination based on sensitive attributes such as gender or race.

However, an additional layer of complexity arises from external economic factors. During a *bull market*—characterized by economic optimism and growth—the bank may adopt a more lenient decision threshold, allowing for a higher rate of loan approvals. In such a regime, the institution may be willing to accept a greater number of *false positives* in exchange for more *true positives*, prioritizing growth and market expansion.

Conversely, in a *bear market*, where financial conditions are fragile, the bank must be considerably more risk-averse. In this scenario, even a small number of false positives can lead to significant financial losses. As a result, the decision threshold becomes more conservative, and the cost of error becomes asymmetrical.

Problem Statement

In light of such practical considerations, the core objective of this research is to develop a post-processing technique that is not only capable of imposing fairness but does so *across all possible classification thresholds*. Traditional post-processing methods typically operate at a single fixed threshold, which may not generalize well across varying operational contexts.

This research, therefore, seeks to answer the following question:

Problem Statement

Can we design a post-processing method that enforces fairness across the entire range of decision thresholds, while maintaining minimal degradation in predictive performance?

Such a method would offer enhanced flexibility for dynamic deployment environments, ensuring that fairness constraints remain valid regardless of the system's operating point. This is especially critical for applications in high-stakes domains, such as finance, healthcare, and criminal justice, where decision thresholds may shift over time or differ across subpopulations.

1.4 Our Approach

To facilitate the practical deployment of fair machine learning systems, we introduce a novel group fairness notion, denoted as ε_p-Equalized ROC. This fairness measure ensures that, regardless of the threshold used for classification, the model maintains approximate fairness. Specifically, for all possible thresholds, the distance between the corresponding points on the ROC curves of the protected groups remains within ε under the $\mathcal{L}_p$ norm. In other words, ε_p-Equalized ROC guarantees uniform fairness across the entire range of operating thresholds, making it particularly appealing for real-world applications where thresholds may vary.

Our objective is to construct a probabilistic classifier that satisfies ε_1-Equalized ROC while minimizing the loss in the Area Under the Curve (AUC) relative to a given scoring function s.

Method Overview: We assume query access to the ROC curve of the scoring function s. Our method begins by making k sufficiently dense queries to approximate the ROC curves of the protected groups through piecewise-linear interpolation. Once the ROC curves are approximated, we apply a transformation that brings the ROCs within ε distance of one another, as measured in the $\mathcal{L}_1$ norm, while minimizing the resulting degradation in AUC.

This transformation is achieved through randomization over a family of feasible classifiers associated with the given ROC curve, which we define as the *ROC Space* of s. By randomizing across this ROC Space, we construct a convex combination of classifiers that yields a probabilistic classifier satisfying the fairness criterion. Conceptually, our method transforms the original scoring function s into a fair scoring function through what we term *ROC transport*. We refer to this procedure as `FROC`. Furthermore, we provide a geometric proof establishing that, under certain conditions, `FROC` is *optimal* with respect to minimizing AUC loss.

1.5 Contributions

Our work makes the following key contributions:

- We introduce ε_p-Equalized ROC, a novel group fairness notion that enforces fairness across all possible thresholds in score-based classification, providing a practical and flexible fairness guarantee for practitioners.

- We formulate fairness-aware post-processing as an optimization problem, seeking an optimal transformation $\mathcal{H}$ on a given scoring function s that minimizes performance loss while ensuring satisfaction of ε_1-Equalized ROC.

- We propose `FROC`, a post-processing algorithm based on ROC transport (Algorithm 10), which achieves ε_1-Equalized ROC without requiring retraining of the underlying machine learning model. This design allows the algorithm to enhance fairness while preserving the interpretability of the model's decisions.

- We provide rigorous theoretical analysis, demonstrating that `FROC` is optimal in terms of minimizing AUC loss under specified conditions (Theorem 4.2).
- We empirically validate the effectiveness of `FROC` through a comprehensive set of experiments, showcasing its practical utility in achieving fairness with minimal performance degradation.

1.6 Thesis Organization

This thesis is organized into *6* chapters, each addressing a specific aspect of the problem and its solution. Below is a summary of the contents of each chapter:

- **Chapter 1: Introduction**
  This chapter introduces the problem of algorithmic bias in machine learning and motivates the need for fairness-aware methods. It begins by presenting real-world examples that highlight the social and ethical implications of biased decision-making systems. The chapter then explores foundational questions such as what it means to be fair and how fairness can be operationalized in the context of machine learning. The proposed approach is briefly outlined, followed by a summary of the thesis contributions and an overview of the thesis structure.
- **Chapter 2: Preliminaries**
  This chapter provides the necessary background for understanding the fairness challenges in machine learning. It begins with foundational concepts such as decision thresholds, ROC curves, and AUC as a performance metric. The chapter then presents core notions of group fairness, including demographic parity, equal opportunity, and calibration fairness, alongside key trade-offs and impossibility results. Further, it reviews algorithmic approaches to fairness, including sampling methods and post-processing techniques like threshold adjustment and reject option classification. The chapter concludes by outlining how these concepts relate to the thesis and identifying fairness-performance trade-offs relevant to the proposed work.
- **Chapter 3: Problem Statement and Algorithm**
  This chapter formally defines the fairness problem addressed in the thesis. It introduces the notion of ε-Equalized ROC, relating it to the established Equalized Odds criterion. A formal problem definition is followed by a description of the query model and a piece-wise linear approximation (PLA) technique for ROC curves. The core algorithm is then presented in detail, including definitions, a taxonomy of shift operations (UpShift, LeftShift, CutShift), and the procedure to derive a fair classifier from the updated ROC curves. This chapter lays the theoretical foundation for the proposed method.
- **Chapter 4: Theoretical Analysis**
  This chapter provides a theoretical foundation for the proposed fairness framework. It begins with an analysis of the piece-wise linear approximation (PLA) of ROC curves, followed by a

detailed examination of AUC loss. The chapter establishes optimality results for various shift operations—including CutShift, UpShift, and LeftShift—and presents theoretical bounds on sample complexity. It also explores extensions of the core framework to handle multiple protected groups and intersecting ROC curves. These results offer rigorous justification for the design and guarantees of the proposed algorithm.

- **Chapter 5: Empirical Analysis**
  This chapter presents a comprehensive empirical evaluation of the proposed method. It begins with a description of the experimental setup and continues with performance results across multiple datasets. Additional experiments are reported using various classifiers—including Weighted Ensemble L2, Random Forest (Gini), and FNNC—on the Adult, COMPAS, and other datasets. These experiments demonstrate the effectiveness of the method across fairness metrics and utility trade-offs, and provide robust evidence for its generalizability and practical applicability.

- **Chapter 6: Conclusion**
  This chapter summarizes the main contributions of the thesis and reflects on the theoretical and empirical findings. It reiterates the significance of fair classification and the advantages of the proposed algorithm in balancing fairness and performance. The chapter concludes with a discussion of potential avenues for future research, including extensions to broader fairness definitions and more complex real-world deployment scenarios.

1.7 Summary

In this introductory chapter, we set the stage for a detailed exploration of fairness in machine learning. By examining concrete examples of algorithmic bias, we illustrated the real and pressing need for fairness-aware systems. We discussed the complexities in defining and achieving fairness, particularly within the context of predictive modeling and classification. Our proposed approach aims to tackle these challenges by introducing novel strategies that integrate fairness considerations into machine learning pipelines. This chapter also highlighted the key contributions of this thesis and outlined its structure, providing a roadmap for the chapters that follow. With this foundation, we move toward a more rigorous treatment of the theoretical and algorithmic principles that underpin fair learning systems.

Chapter 2

Preliminaries

> Mathematical discoveries, small or great ... are never born of spontaneous generation. They always presuppose a soil seeded with preliminary knowledge and well prepared by labour, both conscious and subconscious.
>
> –Henri Poincaré

> It is a capital mistake to theorize before one has data.
>
> –Sherlock Holmes, *A study in scarlet*

Abstract

This chapter provides the foundational elements necessary for understanding fairness in machine learning. It begins with essential concepts from machine learning, such as ROC curves, classifier sampling, and performance metrics like the Area Under the Curve (AUC). Building on these, the chapter explores various notions of fairness, including group fairness and its integration into ranking and AUC-based evaluation frameworks. It then introduces algorithmic strategies to promote fairness and concludes with a discussion of post-processing techniques for classification tasks. These include practical tradeoffs between fairness and performance, as well as the positioning of these methods within the broader context of the thesis. Together, these sections offer the reader a comprehensive baseline for the fairness-aware methods and theories introduced in later chapters.

To lay the groundwork for the fairness-aware methods presented in this thesis, we begin by reviewing core concepts from machine learning. These foundational ideas provide the necessary context for understanding how fairness constraints interact with classification models and evaluation metrics. In particular, this section introduces the formalism of binary classification, key performance measures, and concepts such as the Receiver Operating Characteristic (ROC) curve, which will serve as building blocks for the fairness techniques discussed in subsequent sections.

2.1 Machine Learning Foundations

Notation and Formal Definitions of Classifiers

In binary classification, a **classifier** is a function that assigns a label $\widehat{Y} \in \{0, 1\}$ to an input instance based on observed features. Let $\mathcal{X}$ denote the *feature space* and let $Y \in \{0, 1\}$ represent the ground-truth label, where $Y = 1$ indicates a positive outcome.

We distinguish between two major types of classifiers:

- **Deterministic classifiers** output a binary label directly. That is, given an instance $X \in \mathcal{X}$, the classifier $h : \mathcal{X} \to \{0, 1\}$ produces a decision $\widehat{Y} = h(X)$.

- **Probabilistic (or scoring) classifiers** instead output a real-valued *score*, typically interpreted as the model's estimate of the probability that $Y = 1$. Formally, this is a scoring function:

  $$s : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$$

  where $\mathcal{A}$ denotes the space of sensitive attributes (e.g., race, gender). The score $s(X, A)$ reflects the model's confidence or risk estimate for the positive class.

In most real-world applications, the score is restricted to the interval $[0, 1]$ so that it may be interpreted as a probability. To convert a probabilistic classifier into a deterministic decision rule, a threshold $t \in [0, 1]$ is applied. The resulting prediction $\widehat{Y}$ is defined as:

$$\widehat{Y} = \mathbb{I}(s(X, A) \geq t)$$

where $\mathbb{I}(\cdot)$ denotes the indicator function.

The space of all such scoring functions is denoted by $\mathcal{S}$. The choice of the threshold t directly influences classification outcomes, including key performance metrics such as:

- **True Positive Rate (TPR):** $\mathbb{P}(\widehat{Y} = 1 \mid Y = 1)$

- **False Positive Rate (FPR):** $\mathbb{P}(\widehat{Y} = 1 \mid Y = 0)$

These rates are fundamental not only to measuring the effectiveness of a classifier, but also to evaluating its fairness across different groups. Many fairness criteria—including Equalized Odds, Equal Opportunity, and our proposed post-processing framework—are sensitive to how these rates vary with threshold selection across demographic subgroups.

In general, classifier fairness must be considered in light of both the scoring function s and the thresholding rule applied to it. Different thresholds may be applied for different groups to achieve fairness constraints, a common practice in post-processing approaches.

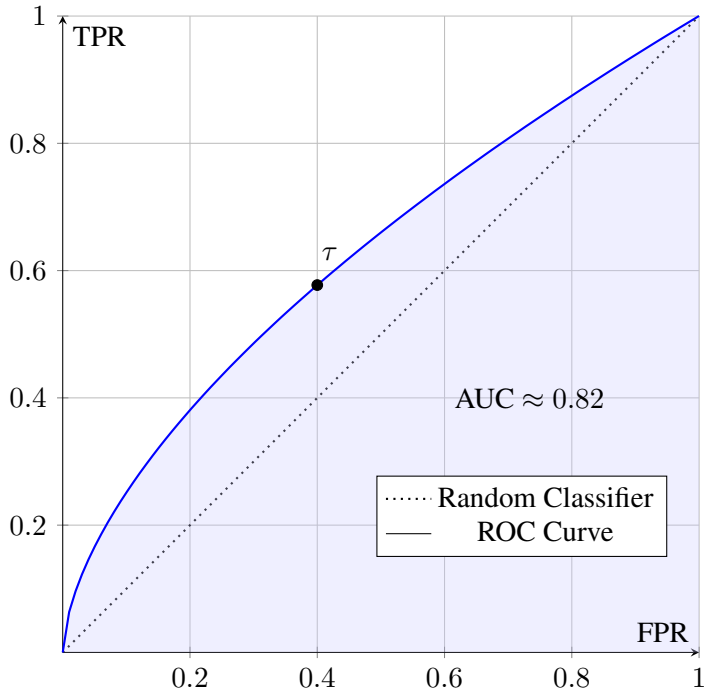


Figure 2.1 Receiver Operating Characteristic (ROC) curve illustrating the tradeoff between true positive rate (TPR) and false positive rate (FPR). The diagonal dotted line represents a random classifier. The shaded region under the curve indicates the Area Under the Curve (AUC), and the point τ marks a specific decision threshold. Note: diagram for illustrative purposes.

2.1.1 Decision Thresholds

The selection of an appropriate threshold is crucial, as it governs the tradeoff between false positives and false negatives. Practitioners typically rely on the Receiver Operating Characteristic (ROC) curve of a scoring function s to guide this choice [Provost, 2000, Zhou and Liu, 2005]. The ROC curve depicts the relationship between the TPR, denoted $G_s(t)$, and the FPR, denoted $H_s(t)$, across all possible thresholds t. Through this curve, one can visualize how threshold adjustments impact the model's behavior, allowing informed decisions to balance sensitivity and specificity.

2.1.2 Receiver Operating Characteristic (ROC) Curves

The ROC curve provides a graphical representation of a classifier's performance across varying thresholds by plotting the TPR against the FPR. This visualization captures the tradeoff between correctly identifying positive instances and incorrectly misclassifying negative instances. Beyond its practical utility, the ROC curve has theoretical significance in understanding the relationship between two cumulative distributions representing scores for the positive and negative classes [Vogel et al., 2021].

Formally, the ROC curve is defined as follows:

Definition 2.1.1 (ROC-Curve)**.** *For any two cumulative distribution functions g_1, g_2 over $\mathbb{R}$, the ROC curve is given by the plot of*

$$ROC_{g_1,g_2}(\alpha) \triangleq 1 - g_1 \circ g_2^{-1}(1-\alpha),$$

where $\alpha \in [0,1]$.

The area under the ROC curve (AUC) quantifies the overall ability of the classifier to rank positive instances higher than negative ones. Specifically, for independent random variables S and S' drawn from g_1 and g_2 respectively, the AUC is defined as

$$AUC_{g_1,g_2} = \mathbb{P}(S' > S) + \frac{1}{2}\mathbb{P}(S' = S).$$

For a given scoring function s, the corresponding random variables G_s and H_s describe the distributions of scores for the positive and negative classes, and the associated ROC curve is denoted $\texttt{ROC}_s$. The AUC of $\texttt{ROC}_s$, denoted $\texttt{AUC}_s = AUC_{H_s,G_s}$, serves as a widely used measure of the scoring function's discriminatory power [Cortes and Mohri, 2003]; [Clémençon et al., 2006]. Intuitively, the AUC represents the probability that a randomly chosen positive instance is scored higher than a randomly chosen negative instance. While an AUC of 1 indicates perfect ranking, such ideal classifiers rarely exist in practice. Consequently, the goal is to find a scoring function s^* within a subset $\mathcal{S}' \subset \mathcal{S}$ that maximizes AUC, formally expressed as $s^* \in argmax_{s \in \mathcal{S}'} \texttt{AUC}_s$.

In Section 3.6.5, we explore how randomization over the outputs of a scoring function s can produce alternative classifiers with adjusted TPRs and FPRs. This process is essential for achieving fairness and expands the set of realizable classifiers within what we define as the *ROC space* of s, denoted $\mathcal{S}|_s$.

2.1.3 Convex Sampling of Classifiers

A fundamental property of ROC curves is that it is impossible to construct a classifier that achieves points lying above the ROC curve of a given scoring function s using only s. However, it is possible to realize classifiers corresponding to any point within the hypograph of the ROC curve—i.e., the region beneath the curve—through appropriate randomization over the predicted scores [Barocas et al., 2023]. This technique involves forming convex combinations of key classifiers, including the "always reject" classifier $(0,0)$, the "always accept" classifier $(1,1)$, and specific intermediate points on the ROC curve such as $\mathcal{Q}_i^{up}$.

By sampling outcomes from these classifiers with certain probabilities, one can represent any point within the convex hull formed by these three points (Figure 3.6). This ability to convexly combine classifiers within the ROC space is a critical tool for achieving fairness, as it allows for fine-grained control over classification outcomes. We formalize this set of achievable classifiers as the *ROC space of* s, denoted $\mathcal{S}|_s$. Each point in $\mathcal{S}|_s$ corresponds to a binary classifier characterized by a specific pair of FPR and TPR values $(FPR(t), TPR(t))$ at some threshold t.

Theorem 2.1.1. *If $\mathcal{Q}_a, \mathcal{Q}_b, \mathcal{Q}_c$ are points in $\mathcal{S}|_s$ forming a convex hull Δ and $\mathcal{Q} \in \Delta$, then the classifier equivalent to $\mathcal{Q}$ can be obtained by following the below procedure. For each test data point x, use the following randomization scheme:*

$$Classifier_{\mathcal{Q}}(x) = \begin{cases} Classifier_{\mathcal{Q}_a}(x) & w.p.\ p_a \\ Classifier_{\mathcal{Q}_b}(x) & w.p.\ p_b \\ Classifier_{\mathcal{Q}_c}(x) & w.p.\ 1 - p_a - p_b \end{cases} \tag{2.1}$$

Here, we have, $p_a = \frac{c_1 b_2 - c_2 b_1}{a_1 b_2 - a_2 b_1}$, $p_b = \frac{c_1 a_2 - c_2 a_1}{a_1 b_2 - a_2 b_1}$ and

$$a_1 = TPR(\mathcal{Q}_a) - TPR(\mathcal{Q}_c)\ and\ a_2 = FPR(\mathcal{Q}_a) - FPR(\mathcal{Q}_c)$$

$$b_1 = TPR(\mathcal{Q}_b) - TPR(\mathcal{Q}_c)\ and\ b_2 = FPR(\mathcal{Q}_b) - FPR(\mathcal{Q}_c)$$

$$c_1 = TPR(\mathcal{Q}) - TPR(\mathcal{Q}_c)\ and\ c_2 = FPR(\mathcal{Q}) - FPR(\mathcal{Q}_c)$$

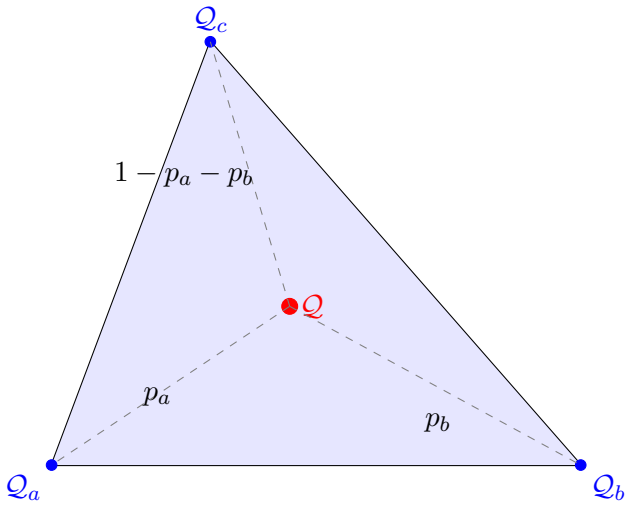


Figure 2.2 Illustration of Theorem: Any point $\mathcal{Q}$ inside the convex hull Δ of classifiers $\mathcal{Q}_a$, $\mathcal{Q}_b$, and $\mathcal{Q}_c$ can be represented as a randomized classifier. For a given test input x, the output of $\mathcal{Q}$ is sampled from the outputs of the three classifiers with probabilities p_a, p_b, and $1 - p_a - p_b$.

2.1.4 Area Under the Curve (AUC) as a Performance Metric

The Area Under the Curve (AUC) is a scalar value summarizing the performance of a classifier across all possible decision thresholds. By capturing the model's ability to discriminate between positive and negative instances, AUC serves as a threshold-independent measure of ranking quality. Despite its widespread use, AUC may not fully reflect real-world performance, particularly in situations involving highly imbalanced datasets or when the cost of false positives and false negatives is asymmetric. In such contexts, complementary metrics or fairness-aware adjustments may be necessary to align classifier performance with the specific needs and ethical considerations of the application domain.

2.2 Fairness in Machine Learning

2.2.1 Foundational and Group Fairness Notions

Fairness in machine learning is a rapidly evolving field concerned with ensuring that algorithmic decisions do not disproportionately disadvantage individuals or demographic groups. A key distinction in fairness literature is between *individual fairness*—the principle that similar individuals should receive similar outcomes [Dwork et al., 2012]—and *group fairness*, which focuses on statistical parity across groups defined by sensitive attributes such as race, gender, or age.

A desirable machine learning model (MLM) balances fairness constraints with minimal compromise on traditional performance metrics such as accuracy, AUC, or F1-score. Achieving this balance requires two fundamental components: a formal, quantifiable definition of fairness, and algorithmic techniques that enforce this definition during model training or through post-hoc adjustment. Among these, group fairness has received widespread attention due to its tractability and alignment with legal standards.

While this work focuses on group-level notions of fairness, such as demographic parity and equalized odds, it is worth briefly mentioning two alternative frameworks: individual fairness and counterfactual fairness. Individual fairness is based on the principle that similar individuals should be treated similarly by a model, typically requiring a task-specific similarity metric to operationalize this notion [Dwork et al., 2012]. Counterfactual fairness, on the other hand, draws on causal reasoning and requires that a prediction remain invariant under a counterfactual change in the sensitive attribute, holding all else constant [Kusner et al., 2017]. While both approaches offer conceptually rich perspectives, they are less applicable to the present discussion, which centers on statistical group fairness in practical deployment settings. As such, we do not explore these frameworks further, though readers interested in these alternative paradigms may refer to the aforementioned works.

Group Fairness: Core Definitions

Group fairness aims to enforce equitable treatment of demographic groups by imposing statistical constraints on model outputs. Prominent group fairness notions include:

Demographic Parity (DP) Also known as *statistical parity*, DP requires that the probability of receiving a positive prediction is equal across groups. Formally, for a protected attribute $A \in \{a_1, a_2\}$ and predicted label $\widehat{Y} \in \{0, 1\}$, demographic parity holds if:

$$\mathbb{P}(\widehat{Y} = 1 \mid A = a_1) = \mathbb{P}(\widehat{Y} = 1 \mid A = a_2). \tag{2.2}$$

This condition enforces equal access to favorable outcomes (e.g., hiring, loan approval). However, when base rates differ across groups, enforcing DP may compromise model utility or even unintentionally introduce new forms of unfairness.

Disparate Impact (DI) DI evaluates fairness via a ratio of positive outcome rates across groups:

$$DI = \frac{\mathbb{P}(\widehat{Y} = 1 \mid A = a_1)}{\mathbb{P}(\widehat{Y} = 1 \mid A = a_2)}. \tag{2.3}$$

The U.S. "four-fifths rule" suggests that DI should be no less than 0.8 to avoid potential discrimination [Barocas and Selbst, 2016]. DI provides a more flexible alternative to DP, especially in legal and regulatory contexts [Feldman et al., 2015].

Equalized Odds (EO). EO requires that the model's true positive rate (TPR) and false positive rate (FPR) be equal across demographic groups [Hardt et al., 2016]. Specifically:

$$\mathbb{P}(\widehat{Y} = 1 \mid A = a_1, Y = 1) = \mathbb{P}(\widehat{Y} = 1 \mid A = a_2, Y = 1), \tag{2.4}$$

$$\mathbb{P}(\widehat{Y} = 1 \mid A = a_1, Y = 0) = \mathbb{P}(\widehat{Y} = 1 \mid A = a_2, Y = 0). \tag{2.5}$$

EO aligns fairness with model performance by requiring balanced error rates across groups. It often involves threshold adjustments or fairness-aware loss functions, but may conflict with Calibration or demographic parity.

Equal Opportunity This is a relaxed form of EO that only requires equality of TPRs across groups. It seeks to ensure fair access to true positive predictions while allowing flexibility in false positive rates.

Calibration Fairness Calibration requires that for individuals receiving the same predicted score, the probability of a positive outcome is the same across groups. While useful for score-based models, Calibration is often at odds with EO when base rates differ.

Fairness Notion	Definition	Key References
Demographic Parity (DP)	Ensures that the probability of a positive outcome is the same across all protected groups.	[Dwork et al., 2012], [Feldman et al., 2015]
Disparate Impact (DI)	Measures whether the ratio of positive outcomes between groups is below a specified threshold (commonly 0.8).	[Barocas and Selbst, 2016]
Equalized Odds (EO)	Requires equal false positive and false negative rates across protected groups.	[Hardt et al., 2016], [Verma and Rubin, 2018]
Equal Opportunity (EOpp)	A relaxation of EO; requires only equal true positive rates across groups.	[Hardt et al., 2016]
Calibration Fairness	Ensures predicted probabilities are calibrated equally across groups.	[Kleinberg et al., 2017], [Chouldechova, 2017]
AUC-Based Fairness	Measures fairness using AUC differences between groups; includes intra-group and inter-group ranking fairness.	[Beutel et al., 2019], [Kallus and Zhou, 2019], [Yang et al., 2023]
Pointwise ROC Fairness	Extends AUC-based fairness by enforcing fairness at all thresholds along the ROC curve.	[Vogel et al., 2021]
Post-processing Fairness	Alters model predictions post hoc to satisfy fairness constraints, without retraining.	[Hardt et al., 2016], [Wei et al., 2020], [Cui et al., 2021]
Threshold Optimization	Learns separate thresholds per group to balance fairness and accuracy.	[Jang et al., 2022]
ROC-Based Fairness (ε_p-Fairness)	Enforces fairness by bounding differences in TPR and FPR across all thresholds.	**This Work**, [Chen and Wu, 2020]

Table 2.1 Comparison of Fairness Notions and Methods in Machine Learning

2.2.2 Fairness Tradeoffs and Impossibility Results

An important challenge in algorithmic fairness is the inherent tradeoff between competing fairness criteria. A significant body of literature demonstrates that it is often impossible to simultaneously satisfy multiple fairness notions. For instance, [Kleinberg et al., 2017] and [Chouldechova, 2017] show that Calibration and EO cannot both hold when base rates differ across groups. These impossibility theorems highlight the necessity of prioritizing fairness goals based on application context and stakeholder values.

2.2.3 Limitations in Score-Based Classifiers

Traditional group fairness metrics like DP, DI, and EO assume a fixed decision threshold, limiting their applicability to *score-based classifiers*, which produce continuous outputs (e.g., probabilities or risk scores). However, in many real-world scenarios—such as dynamic loan approval systems, recidivism risk assessment, or personalized medical treatment—the decision threshold may vary across time, contexts, or user-defined constraints.

As noted by [Gorantla et al., 2021], conventional fairness definitions may fail to generalize across all operating points in such systems. This has motivated the development of fairness notions grounded in ranking and pairwise comparison.

2.2.4 Fairness in Ranking and AUC-based Frameworks

Ranking-based fairness considers settings in which classifiers assign continuous scores rather than discrete labels, and thus, performance is better measured by ranking metrics such as the Area Under the ROC Curve (AUC). In this context, fairness can be evaluated through intra-group and inter-group comparisons.

- *Intra-group pairwise AUC fairness* [Beutel et al., 2019] ensures equitable ranking within demographic groups.
- *Inter-group pairwise fairness (xAUC)* [Kallus and Zhou, 2019] addresses disparities between groups in how individuals are relatively ranked.
- *BNSP (Balanced Negative Subgroup Pairwise) fairness* [Borkan et al., 2019] accounts for nuanced imbalances within subgroup comparisons.

Recent efforts have integrated these ideas into unified learning objectives. For example, [Yang et al., 2023] introduces a minimax optimization approach that balances both intra-group and inter-group AUC fairness. [Vogel et al., 2021] propose a general fairness framework over AUC, noting that while AUC reflects global ranking performance, it may overlook local disparities. To address this, they introduce *pointwise ROC-based fairness*, supported by an in-processing algorithm that directly optimizes for fairness at all points along the ROC curve.

2.2.4.1 Our Fairness Perspective

Our proposed fairness metric, denoted ε_p-Equalized ROC, draws inspiration from equalized odds but extends the fairness constraint across the entire threshold spectrum. Unlike previous approaches that consider fairness at fixed thresholds or average performance gaps, ε_p-Equalized ROC operates in the ROC space and evaluates error rate disparities across all decision thresholds.

This approach generalizes the metric introduced by [Chen and Wu, 2020], who measures fairness using the Manhattan distance between group performance vectors. In contrast, our method adopts a geometric interpretation, enabling more expressive and robust fairness guarantees. Most importantly, while prior work may overlook threshold variability, ε_p-Equalized ROC explicitly enforces fairness consistency across operating points, making it suitable for applications involving score-based, threshold-sensitive deployment settings.

2.3 Algorithmic Approaches to Achieving Fairness

Fairness interventions can be introduced at various stages of the machine learning pipeline, including pre-processing the data, modifying the training procedure, or post-processing the model's outputs. Each approach offers distinct advantages and challenges, and the choice of method depends on the application requirements, available resources, and access to model internals.

Pre-processing Approaches to Fairness

Pre-processing methods constitute a major class of bias mitigation strategies in fair machine learning. These techniques intervene *before* model training by modifying the dataset itself. The core rationale is straightforward: if training data reflects historical or systemic biases, then models trained on such data are likely to reproduce those biases unless the data is adjusted.

The goal of pre-processing is to reduce or remove discriminatory patterns in the data while maintaining its utility for accurate prediction. These methods are especially important in domains where sensitive attributes (such as race or gender) are deeply entangled with target outcomes due to structural inequities.

Several prominent techniques fall under this category, including:

Pre-processing Techniques for Fairness in Machine Learning

- **Suppression:** This method identifies and removes the attributes most strongly correlated with a sensitive attribute (e.g., gender, race) to mitigate discrimination in downstream predictions. In addition to the sensitive attribute itself, correlated features are excluded from the dataset. While effective at reducing bias, this approach is often overly aggressive, eliminating features that may contain valuable predictive information and thus impairing model performance.

- **Massaging:** In this technique, the class labels of selected instances are altered to reduce observed discrimination in the dataset. The key challenge lies in identifying the right instances for relabeling—those whose modification would maximize fairness while minimally distorting the data distribution. This method assumes access to a trusted fairness metric and may raise ethical concerns regarding label manipulation.

- **Reweighting:** Rather than altering features or labels, each instance in the training data is assigned a weight based on its group membership and outcome. This rebalancing counteracts historical bias by emphasizing underrepresented or disadvantaged groups. However, it requires that the learning algorithm support instance weighting, which not all classifiers do.

- **Resampling:** This approach modifies the distribution of the training dataset by over-sampling underrepresented group-outcome pairs and/or under-sampling overrepresented ones. Weights derived from fairness considerations guide this process, allowing the data to simulate a more balanced population without directly modifying features or labels.

- **Feature Transformation:** This method reduces or eliminates correlations between input features and sensitive attributes through transformations such as adversarial debiasing, fair representation learning, or fairness-aware dimensionality reduction. The objective is to preserve predictive utility while minimizing discriminatory patterns embedded in the data.

- **Synthetic Data Generation:** When certain demographic groups are underrepresented, synthetic data can be generated to improve balance and support fairer model outcomes. Care must be taken to ensure the generated samples are realistic and do not introduce artifacts or spurious correlations.

Below, we detail representative methods from each of these categories.

Massaging

Massaging[Kamiran and Calders, 2009] is a pre-processing technique that modifies the training labels of a selected subset of individuals in order to equalize outcome distributions across groups. It does so by promoting some negatively labeled individuals from the disadvantaged group to a positive label and demoting some positively labeled individuals from the advantaged group to a negative label. Candidate individuals are selected based on their predicted likelihood of receiving the positive label, ensuring that label changes minimally affect classification accuracy.

Algorithm 1 Massaging the Dataset for Demographic Parity

Input: Training dataset $D = \{(x_i, y_i, a_i)\}$ with features x, labels y, and sensitive attribute a

Train a classifier h on D to obtain predicted scores $\hat{y}_i = h(x_i)$ ▷ (Note: h is used only to rank instances, not for final predictions.)

Rank all individuals by $\hat{y}_i$ in descending order

Compute desired number of positive outcomes per group to achieve demographic parity

Identify:

- **Promotion candidates:** Negatively labeled individuals in disadvantaged group with high $\hat{y}_i$
- **Demotion candidates:** Positively labeled individuals in advantaged group with low $\hat{y}_i$

Flip labels of top-N promotion and demotion candidates to equalize group-wise positive outcome rates

Output: Modified dataset D'

and now we present a more precise algorithm:

Algorithm 2 Massaging the Dataset for Demographic Parity

Require: Training dataset $D = \{(x_i, y_i, a_i)\}$, where:

x_i = feature vector, $y_i \in \{0, 1\}$ = label, $a_i \in \{0, 1\}$ = sensitive attribute (e.g., 0 = disadvantaged, 1 = advantaged)

Ensure: Modified dataset D' with adjusted labels to satisfy demographic parity

1: Train a classifier h on D to obtain predicted scores: $\hat{y}_i = h(x_i)$

2: Rank all instances in descending order of predicted scores $\hat{y}_i$

3: Compute overall positive rate: $r = \frac{1}{|D|} \sum_i y_i$

4: Let n_0, n_1 be the number of individuals in groups $a = 0$ and $a = 1$, respectively

5: Compute desired positives: $d_0 = \lfloor r \cdot n_0 \rfloor$, $d_1 = \lfloor r \cdot n_1 \rfloor$

6: Compute current positives: $c_0 = \sum_{i:a_i=0} y_i$, $c_1 = \sum_{i:a_i=1} y_i$

7: Determine number of flips: $N = \min(|d_0 - c_0|, |d_1 - c_1|)$

8: Identify candidates:

- **Promotion candidates:** Samples with $a_i = 0$, $y_i = 0$ and high $\hat{y}_i$
- **Demotion candidates:** Samples with $a_i = 1$, $y_i = 1$ and low $\hat{y}_i$

9: Select top-N promotion and demotion candidates based on $\hat{y}_i$

10: Flip labels of selected candidates: $y_i \leftarrow 1$ for promotions, $y_i \leftarrow 0$ for demotions

11: **return** Modified dataset D'

Reweighting Samples

Reweighting [Kamiran and Calders, 2012] adjusts the importance of each sample during training, giving more influence to underrepresented or disadvantaged group-label combinations. This technique approximates a balanced distribution of data without altering features or labels.

Algorithm 3 Reweighting Samples for Fair Classification

Require: Dataset $\mathcal{D} = \{(x_i, y_i, a_i)\}_{i=1}^{n}$ where:
x_i = features, $y_i \in \{0, 1\}$ = label, $a_i \in \{0, 1\}$ = sensitive attribute

Ensure: Weighted dataset $\mathcal{D}' = \{(x_i, y_i, a_i, w_i)\}_{i=1}^{n}$ for training

Estimate the empirical distributions:
  $P_A[a] \leftarrow$ marginal distribution of sensitive attribute A
  $P_Y[y] \leftarrow$ marginal distribution of label Y
  $P_{A,Y}[a, y] \leftarrow$ joint distribution of (A, Y)
for each sample $(x_i, y_i, a_i) \in \mathcal{D}$ **do**
    Compute sample weight:
      $w_i \leftarrow \frac{P_A[a_i] \cdot P_Y[y_i]}{P_{A,Y}[a_i, y_i]}$
    Augment sample with weight: $(x_i, y_i, a_i) \rightarrow (x_i, y_i, a_i, w_i)$
end for
Train a classifier on the weighted dataset $\mathcal{D}'$

2.3.1 Sampling

In scenarios where classifiers do not support instance weighting, a sampling-based approach can be employed as an alternative to reweighting. This method modifies the empirical distribution of the training dataset by selectively over-sampling underrepresented group-label combinations and/or under-sampling overrepresented ones. The sampling probabilities are derived from fairness-aware weights, allowing the adjusted dataset to reflect a more equitable representation across sensitive attributes and outcomes, without altering feature values or labels.

This technique enables models that are incompatible with weighted loss functions to still benefit from fairness interventions. The resulting dataset simulates a balanced population structure, effectively mitigating historical or societal biases encoded in the original data distribution.

Algorithm 4 Fairness-Aware Resampling

Require: Training dataset $\mathcal{D} = \{(x_i, y_i, a_i)\}_{i=1}^{n}$, where x_i is a feature vector, y_i is the label, and a_i is the sensitive attribute

Ensure: Resampled dataset $\mathcal{D}'$ with adjusted group-outcome distribution

Estimate empirical distributions:

$P_{A,Y}[a, y] \leftarrow$ empirical joint distribution of sensitive attribute and label

$P_A[a] \leftarrow$ marginal distribution of A, $P_Y[y] \leftarrow$ marginal distribution of Y

for each instance $(x_i, y_i, a_i) \in \mathcal{D}$ **do**

Compute fairness-aware sampling weight:

$$w_i \leftarrow \frac{P_A[a_i] \cdot P_Y[y_i]}{P_{A,Y}[a_i, y_i]}$$

end for

Normalize weights to form sampling probabilities:

$$p_i \leftarrow \frac{w_i}{\sum_{j=1}^{n} w_j}$$

Resample n instances from $\mathcal{D}$ using probabilities $\{p_i\}_{i=1}^{n}$ (with or without replacement)

return Resampled dataset $\mathcal{D}'$

Fair Representation Learning

The approach proposed by Zemel et al. [Zemel et al., 2013] aims to construct new representations of input data that are both predictive of the target label and statistically independent of sensitive attributes. The method, referred to as *Learned Fair Representations (LFR)*, involves training an encoder to generate a latent representation that preserves task-relevant information while obfuscating features correlated with protected characteristics.

More precisely, the objective is to learn a set of prototypes Z such that:

1. The probabilistic mapping from the original feature space X to the latent representation Z satisfies *statistical parity* with respect to the sensitive attribute A;

2. The latent space Z retains information from X that is relevant to the prediction task, while minimizing the encoding of sensitive group membership;

3. The composed mapping from X to the predicted label Y, via the latent representation Z, closely approximates the original predictive function f.

This can be formalized as the minimization of a composite loss function:

$$\min_{Z} \left[\text{Loss}(Y, Z) + \lambda \cdot \text{MutualInformation}(Z, A)\right]$$

where the first term promotes predictive accuracy, and the second penalizes dependence between the learned representation Z and the sensitive attribute A, controlled by a tradeoff parameter λ.

This objective is commonly optimized using deep learning architectures or variational inference techniques.

Synthetic Data Generation

When datasets exhibit strong group imbalance, synthetic data can be generated to improve representation. Traditional techniques like *SMOTE* [Chawla et al., 2002, Xu et al., 2018] generate artificial examples for minority classes, while more recent approaches employ fairness-aware generative models.

- **SMOTE:** Oversamples minority group examples by interpolating between existing data points.
- **FairGAN:** Uses Generative Adversarial Networks to produce synthetic data that balances fairness and realism.

Overall, pre-processing techniques play a foundational role in mitigating bias at the data level. By modifying the inputs to learning algorithms, these methods aim to prevent discrimination from arising in the first place, serving as a first line of defense in the pursuit of algorithmic fairness.

In-Processing Techniques

In-processing methods address algorithmic bias by incorporating fairness objectives directly into the model's training procedure. These techniques modify the learning algorithm itself—typically by embedding fairness constraints or regularization terms into the optimization objective—so that the resulting model balances predictive accuracy with fairness during training [Padala and Gujar, 2020, Zafar et al., 2017, Zhang et al., 2018]. Because they intervene at the heart of the learning process, in-processing methods often offer stronger theoretical guarantees of fairness compared to pre- or post-processing approaches.

In-processing: Fairness During Model Training

In-processing methods form a core category of fairness interventions that modify the model training process itself. Unlike pre-processing approaches (which alter the input data) or post-processing methods (which adjust outputs after training), in-processing techniques intervene *during* training to embed fairness objectives directly into the model's learning dynamics.

The central goal is to internalize fairness-aware behavior by guiding the model towards decisions that reduce discriminatory patterns while preserving predictive performance. These methods are particularly powerful in settings where fairness constraints must be enforced in tandem with accuracy and where practitioners have full control over the model architecture and training loop.

Core Strategies

- **Fairness-Constrained Optimization:** Adds fairness penalties or constraints to the loss function, encouraging models to minimize disparities across groups.
- **Adversarial Debiasing:** Trains the model to predict the target variable while simultaneously concealing sensitive attributes from an adversarial network.
- **Fair Representation Learning:** Constructs latent representations that are both predictive and invariant to sensitive features.

These strategies often involve a tunable hyperparameter to manage the tradeoff between **fairness** and **accuracy**, allowing customization based on the application's ethical priorities.

Fairness-Constrained Optimization

A widely used in-processing approach [Zafar et al., 2017, Kamishima et al., 2012, Padala and Gujar, 2020] involves modifying the model's objective function to include fairness-aware constraints or regularization terms. This allows the model to simultaneously optimize for predictive accuracy and statistical fairness definitions such as demographic parity or equalized odds.

Objective:

$$\min_{\theta} \ \mathcal{L}(\theta) + \lambda \cdot \Omega_{\text{fair}}(\theta)$$

Here, $\mathcal{L}(\theta)$ is the standard loss (e.g., cross-entropy), $\Omega_{\text{fair}}(\theta)$ is a fairness penalty, and λ controls the trade-off.

Algorithm 5 Fairness-Constrained Model Training

1: **for** each epoch **do**
2: $L \leftarrow \text{PredictionLoss}(y, f_\theta(x))$
3: $\Omega_{\text{fair}} \leftarrow \text{FairnessMetric}(f_\theta(x), a)$
4: total_loss $\leftarrow L + \lambda \cdot \Omega_{\text{fair}}$
5: Update model parameters via gradient descent
6: **end for**

Adversarial Debiasing

This method [Zhang et al., 2018] frames fairness as a minimax optimization problem between a predictor and an adversary. While the predictor learns to predict the target outcome, an adversarial network attempts to recover the sensitive attribute from the predictor's output. The predictor is penalized when the adversary is successful, thereby learning representations that are both predictive and group-invariant.

Training Framework:

- **Predictor:** Learns $f_\theta(x)$ to predict y
- **Adversary:** Learns $g_\phi(f_\theta(x))$ to predict sensitive attribute a

Objective:

$$\min_\theta \left[\mathcal{L}_{\text{task}}(y, f_\theta(x)) - \lambda \cdot \mathcal{L}_{\text{adv}}(a, g_\phi(f_\theta(x)))\right]$$

Algorithm 6 Adversarial Fairness Training

for each epoch **do**
    (Step 1) Train predictor to minimize task loss and fool adversary:
    Freeze adversary; compute task and adversarial loss
    Update θ to minimize combined loss
    (Step 2) Train adversary to detect sensitive attribute:
    Freeze predictor; update ϕ using adversarial loss
end for

FairBINN

Traditional approaches for fairness often use Lagrangian formulations to combine fairness constraints into the loss function. However, this can make it difficult to disentangle the influence of fairness and accuracy objectives. To address this, the Yazdani-Jahromi et al. [2024] introduces separate parameter sets for the accuracy and fairness objectives. Each set is updated iteratively in a game-theoretic manner, treating optimization as a two-player game: one player focuses on minimizing classification error, and the other minimizes fairness violations.

Algorithm 7 Fairness-Accuracy Bilevel Optimization

Initialize accuracy parameters θ_a and fairness parameters θ_f
while not converged or max iterations T not reached **do**
    Accuracy player's optimization
    Sample a minibatch $\mathcal{B}_a \subset \mathcal{D}$
    Compute accuracy loss: $L_a = \frac{1}{|\mathcal{B}_a|} \sum_{i \in \mathcal{B}_a} L_{\text{acc}}(f(x_i; \theta_a, \theta_f), y_i)$
    Update accuracy parameters: $\theta_a \leftarrow \theta_a - \eta_a \nabla_{\theta_a} L_a$
    Fairness player's optimization
    Sample a minibatch $\mathcal{B}_f \subset \mathcal{D}$
    Compute fairness loss: $L_f = \frac{1}{|\mathcal{B}_f|} \sum_{i \in \mathcal{B}_f} L_{\text{fair}}(f(x_i; \theta_a, \theta_f), a_i, y_i)$
    Update fairness parameters: $\theta_f \leftarrow \theta_f - \eta_f \nabla_{\theta_f} L_f$
end while

Advantages and Limitations

In-processing methods are highly expressive and offer fine-grained control over fairness-accuracy tradeoffs. They are particularly well-suited to research or production settings where the model architecture is transparent and modifiable. However, they also present certain limitations:

- **Algorithmic Access Required:** These techniques require the ability to modify the model's training routine, making them unsuitable for black-box systems or third-party APIs.
- **Increased Complexity:** Implementing fairness constraints, especially in adversarial setups, adds computational and engineering overhead.
- **Hyperparameter Sensitivity:** The fairness-accuracy balance is sensitive to tuning and may vary significantly across contexts.

Despite these challenges, in-processing remains a cornerstone of fair machine learning. Its ability to incorporate fairness directly into the model optimization process makes it a powerful tool for developing ethically aligned predictive systems.

2.4 Post-Processing Approaches for Fair Classification

2.4.1 Motivation and Scope

Post-processing techniques offer a flexible and practical approach to enforcing fairness in machine learning, particularly in scenarios where access to model internals is restricted or retraining is infeasible. These methods operate by adjusting the predictions of a pre-trained model, often treated as a black box, to satisfy fairness constraints, without altering the model's internal architecture or learning process. This makes post-processing especially attractive in large-scale deployed systems, commercial APIs, and resource-constrained environments where computational efficiency and modularity are critical [Sleeman et al., 1995, Nandy et al., 2022].

While post-processing methods provide modularity and model-agnostic deployment, they are inherently constrained by the quality and limitations of the underlying model. They also introduce tradeoffs between fairness and traditional performance metrics such as accuracy, AUC, or Calibration. Nevertheless, their non-intrusive nature makes them a compelling component of fairness-aware system design, particularly for practitioners working with third-party models or proprietary architectures.

Canonical Post-processing Strategies

Several classical post-processing strategies are widely used to achieve group fairness, particularly in binary classification. These methods operate on the output scores or predictions of a pre-trained model and aim to satisfy fairness constraints such as Equalized Odds or Equal Opportunity.

Threshold Adjustment Threshold adjustment techniques involve learning group-specific decision thresholds to balance performance metrics such as the true positive rate (TPR) and false positive rate (FPR) across protected groups [Hardt et al., 2016].

Algorithm 8 Group-Specific Thresholding for Equalized Odds

Train a scoring classifier $s : \mathcal{X} \times \mathcal{A} \to [0, 1]$
Partition dataset by protected attribute A
for each group $a \in \mathcal{A}$ **do**
    Learn threshold t_a such that:

$$\text{TPR}(a) \approx \text{TPR}(a'), \quad \text{FPR}(a) \approx \text{FPR}(a') \quad \forall a' \in \mathcal{A}$$

end for
Predict using $\widehat{Y} = \mathbb{I}(s(X, A) \geq t_A)$

Reject Option Classification This method intervenes in the decision boundary region by favoring beneficial outcomes for disadvantaged groups [Kamiran et al., 2012].

Algorithm 9 Reject Option Classification [Kamiran et al., 2012]

Train a scoring classifier $s(X, A) \in [0, 1]$
Choose a base threshold t and uncertainty margin δ
Define the decision region $[t - \delta, t + \delta]$
for each instance (X, A) **do**
    Compute raw prediction: $\tilde{Y} = \mathbb{I}(s(X, A) \geq t)$
    if $s(X, A) \in [t - \delta, t + \delta]$ **then**
        if A is disadvantaged **and** $\tilde{Y} = 0$ **then**
            Set $\widehat{Y} = 1$ ▷ Give benefit of doubt
        else if A is privileged **and** $\tilde{Y} = 1$ **then**
            Set $\widehat{Y} = 0$ ▷ Avoid unfair benefit
        else
            Set $\widehat{Y} = \tilde{Y}$
        end if
    else
        Set $\widehat{Y} = \tilde{Y}$
    end if
end for

Toward More Sophisticated Methods

The seminal work by Hardt et al. [2016] laid the foundation for fairness-aware post-processing by formalizing the Equalized Odds criterion. This inspired numerous refinements, including randomized and optimization-based methods.

Recent contributions such as [Wei et al., 2020] and [Cui et al., 2021] extend post-processing to constrained optimization and bipartite ranking, respectively. [Zhao, 2024] propose a Wasserstein-based fairness formulation, while [Jang et al., 2022] develop adaptive thresholding using estimated confusion matrices.

Building on this, [Wei et al., 2020] explores fairness through expected score constraints, offering guarantees on bounded differences in true and false positive rates. [Cui et al., 2021] extend the domain of post-processing to bipartite ranking, presenting a model-agnostic technique that balances fairness across ranked outputs.

More recent advances have focused on the geometry and complexity of fairness optimization. [Zhao, 2024] propose a Wasserstein barycenter-based post-processing method that quantifies the cost of fairness, demonstrating its learning complexity is comparable to that of a Bayes-optimal predictor. Similarly, [Ţifrea et al., 2024] transforms fairness-aware in-processing objectives into post-processing solutions, enabling broader applicability across various model classes.

In addressing methodological concerns within fairness research, [Cruz and Hardt, 2023] critiques the inconsistency of comparative studies, highlighting issues such as mismatched baselines and variable constraint relaxations that hinder reproducibility and fair evaluation.

[Jang et al., 2022] introduces a multi-constraint post-processing technique using adaptive thresholding, where group-specific thresholds are learned based on confusion matrix estimates derived from output score distributions. This approach enables fine-grained control over group fairness by aligning multiple statistical metrics simultaneously. Together, these methods illustrate the richness of the post-processing paradigm—balancing fairness, interpretability, and deployability.

2.4.2 Relation to Our Work

Our post-processing framework, denoted ε_p-Equalized ROC, introduces a novel fairness criterion that builds on—but is distinct from—these prior approaches. Whereas [Mishler et al., 2021] defines fairness using bounded differences in counterfactual TPRs and FPRs, our method directly bounds observed differences in TPRs and FPRs across groups. Crucially, it does so *across all possible thresholds*, thereby addressing limitations in existing methods that operate at a fixed or learned decision threshold.

By leveraging the geometric structure of ROC curves, ε_p-Equalized ROC offers a principled and interpretable approach to enforcing fairness in score-based classifiers. Our method minimizes performance degradation, particularly with respect to AUC, while maintaining consistency in fairness guarantees across varying deployment conditions. This makes it especially well-suited for real-world systems where decision thresholds may shift over time or differ across operational contexts.

2.4.3 Fairness and Performance Tradeoffs

One of the central challenges in fairness-aware machine learning is balancing fairness with traditional performance metrics. For classification tasks, a key performance property is *calibration* [Kleinberg et al., 2017], which requires that, among individuals assigned a score p, approximately p proportion of them belong to the positive class. However, [Kleinberg et al., 2017] and [Chouldechova, 2017] demonstrate that calibration and equalized odds are generally incompatible when base rates differ across groups. As a result, much of the recent research focuses on developing models that satisfy approximate fairness guarantees while managing these inherent tradeoffs [Madras et al., 2018].

From a practitioner's perspective, the *Receiver Operating Characteristic* (ROC) curve is a preferred evaluation tool, as it captures the tradeoff between true positive and false positive rates across thresholds and reflects the relative ranking quality of the model. The area under the ROC curve (AUC) is particularly well-suited for assessing classifiers that must reliably distinguish positive from negative examples, regardless of the threshold. Moreover, when fairness is required across all threshold scores, AUC becomes a critical performance metric for balancing ranking quality with fairness objectives [Huang and Ling, 2005, Clémençon et al., 2006, Zehlike et al., 2021].

2.5 Our Problem

We now revisit the illustrative example introduced in the opening chapter, situated in a real-world financial setting. Consider a bank that employs a machine learning model to assess loan eligibility based on applicant profiles. In this context, concerns surrounding fairness are both immediate and significant: the decision-making system must avoid discriminatory outcomes based on sensitive attributes such as race or gender.

Beyond the inherent concerns of fairness, the application is further complicated by external economic conditions. For instance, during a *bull market*—a period marked by optimism and economic growth—the bank may adopt a more lenient decision threshold. In such conditions, the institution is generally more willing to accept a higher rate of *false positives* in exchange for a greater number of *true positives*, thereby prioritizing growth and customer acquisition.

In contrast, during a *bear market*, characterized by economic contraction and increased financial risk, the bank must adopt a more conservative stance. In this regime, even a small number of false positives can be financially damaging. Consequently, the decision threshold becomes stricter, and the cost of misclassification grows increasingly asymmetric.

Motivating the Problem

These fluctuating operational conditions highlight a core limitation of existing fairness-aware post-processing techniques: they typically operate at a fixed classification threshold. Such methods are ill-suited for dynamic environments, where the optimal decision threshold may vary over time or across subpopulations.

To address this gap, we propose a more flexible objective:

Problem Statement

Can we design a post-processing method that enforces fairness across the entire range of decision thresholds, while maintaining minimal degradation in predictive performance?

A solution to this problem would be especially valuable in high-stakes domains, such as finance, healthcare, and criminal justice, where fairness guarantees must hold across a range of operating points, not just at a single fixed threshold.

Understanding the ROC Landscape

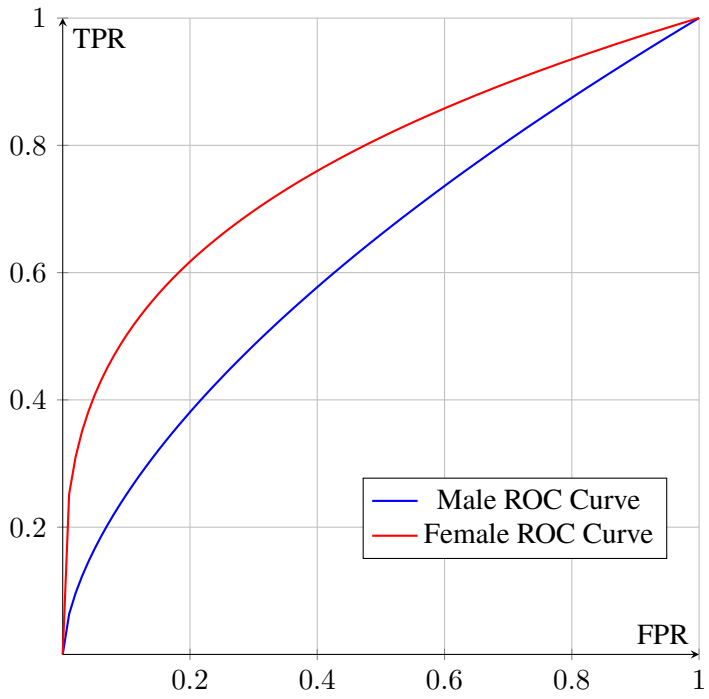


Figure 2.3 ROC curves corresponding to two demographic groups (e.g., males and females) for a given classifier. The curve for the female group lies consistently above that of the male group, indicating superior classification performance. For a fixed true positive rate (TPR), the male group exhibits a higher false positive rate (FPR), which may lead to disproportionately favorable outcomes for that group. This disparity motivates the need for a post-processing method that can reduce the distance between these ROC curves while maintaining predictive performance. Note: diagram for illustrative purposes.

Having introduced the concepts of ROC curves and decision thresholds, we can now revisit our problem with greater precision. Consider Figure 2.3, which depicts the ROC curves for two demographic groups—e.g., males and females—based on the outputs of a given model. A key observation is that the ROC curve for the female group lies consistently above that of the male group. This indicates that, for any fixed true positive rate (TPR), the false positive rate (FPR) is higher for males.

This imbalance implies that undeserving male applicants are more likely to receive loan approvals than their female counterparts at the same decision threshold—a clearly undesirable and unfair outcome. An intuitive solution might be to make the two ROC curves identical. However, doing so may lead to significant degradation in overall model performance.

This motivates a natural compromise: rather than enforcing exact equality between the curves, can we bring them *closer* in a controlled manner—striking a balance between fairness and predictive utility?

Key Research Questions

In light of this, we pose the following guiding questions:

- Can we design a method to make the ROC curves of different groups approximately identical, or at least bring them closer together?
- How can we achieve this while ensuring that the reduction in predictive performance, measured, for instance, by AUC, is minimized?

2.6 Summary

In this chapter, we introduced critical building blocks for understanding fairness in machine learning systems. We reviewed core performance metrics and classifier evaluation tools, which serve as the basis for fairness-aware assessment. By examining different fairness notions and how they relate to ranking and AUC-based tasks, we highlighted the complexity of equitable machine learning. Furthermore, we surveyed algorithmic interventions for fairness and detailed practical post-processing strategies, including their motivations and limitations. This chapter prepares the reader to navigate the technical parts of the upcoming chapters, where we detail the problem statement along with our proposed solution and its analysis.

Chapter 3

Problem Statement and Algorithm

An algorithm must be seen to be believed.

— Donald Knuth

If you can't explain it simply, you don't understand it well enough.

— Often attributed to Einstein and Feynman

Abstract

To facilitate a clear exposition of our problem, we begin by introducing a new notion of fairness, denoted as ε-Equalized ROC. This definition serves as a conceptual foundation for formally stating our problem. We then assume query access to the $\texttt{ROC}_s$ curve, allowing us to sample relevant statistics at selected thresholds, as described in Section 3.3. Since sampling over a continuum of thresholds is infeasible, we instead work with a discretized version of the $\texttt{ROC}_s$, which we approximate using a piecewise linear representation (Section 3.4). Building on this framework, we present our algorithm for transporting ROCs in Section 3.6.4. Finally, we summarize our transformation as `FROC` in Section 3.6.5.

Before diving into the technical details, we provide a conceptual foundation for the framework developed in this chapter. The notion of ε-Equalized ROC introduced below extends traditional fairness definitions in a way that is sensitive to ROC-based evaluation. This section sets the stage for the algorithms and representations discussed later by clearly articulating our fairness objective and motivating the challenges that arise in its practical implementation.

3.1 ε-Equalized ROC

As discussed earlier, all group fairness notions are characterized by equality of a particular statistic across both the protected groups. In scoring-based probabilistic classifiers, these fairness notions depend on the selected threshold. To achieve fairness across all thresholds, the practitioner can choose to retrain the model and achieve the right trade-offs between TPR and FNR. However, retraining is expensive. Therefore, a desirable solution is To offer fair treatment to both protected groups using the pre-trained classifier. However, this leads to invoking the post-processing technique every time the practitioner needs to update the threshold t^*. Instead, we propose a novel fairness measure to simplify the practitioner's job. We perform post-processing on the given classifier once, and it ensures that no matter what threshold t^* they choose to make decisions, the classifier offers a similar treatment to both the protected groups. That is, the individual ROCs (Here on, we shall denote the ROCs of the protected groups, i.e., $ROC_{H_s^0,G_s^0}$ and $ROC_{H_s^1,G_s^1}$ by $\texttt{ROC}_s^0$ and $\texttt{ROC}_s^1$ respectively) should be within ε distance ($\mathcal{L}_p$ norm) of each other. We call it ε_p*-Equalized ROC*. More formally,

Definition 3.1.1 (ε_p-Equalized ROC)**.** *A scoring function for binary classification s with label prediction $\widehat{Y} = \mathbb{I}(s(x) \geq t)$ is said to satisfy -Equalized ROC if for all $\alpha \in (0,1)$ the following holds:*

$$|| \ \texttt{ROC}_s^1(\alpha) - \texttt{ROC}_s^0(\alpha) \ ||_p \leq \varepsilon \tag{3.1}$$

In ε_p-Equalized ROC, we utilize standard metrics (i.e. $\mathcal{L}_p$ norms) as the fairness statistic to quantify fairness. Thus, ε_p-Equalized ROC is feasible for post-processing algorithms.

Furthermore, if FROC is effective for $\mathcal{L}_1$, it necessarily extends to all p-norms. This conclusion follows from the inequality:

$$|a|^p + |b|^p \leq |a| + |b|, \quad \forall p \geq 1,\ a, b \in [0,1].$$

However, while FROC ensures fairness, it does not guarantee optimality for $p > 1$.

Next, we formulate the problem of fair post-processing. Note: ε_1-Equalized ROC is a generalization of Equalized Odds to all the given thresholds of the scoring function.

3.1.1 Relation to Equalized Odds

Equalized Odds is defined in Madras et al. [2018] as the sum of the absolute differences of the FNR and the FPR of both the protected groups. Formally,

$$EO \triangleq |FPR_0 - FPR_1| + |FNR_0 - FNR_1|$$

However, this defintion is equivalent to ε_1-Equalized ROCsince $|FPR_0 - FPR_1| + |FNR_0 - FNR_1| = |FPR_0 - FPR_1| + |(1 - TPR_0) - (1 + TPR_1)| = |FPR_0 - FPR_1| + |TPR_0 - TPR_1|$.

3.2 Formal Problem

In this section, we transition from the high-level motivation presented in the preceding chapters to a formal statement of our problem. Our goal is to design a post-processing method that promotes fairness while preserving predictive performance.

We begin with the following motivating question:

Problem Statement

Can we design a post-processing method that enforces fairness across the entire range of decision thresholds, while maintaining minimal degradation in predictive performance?

As discussed previously, our central motivation is to bring the ROC curves of different groups closer together, thereby reducing disparity. While this intuition is appealing, it raises an important formal question: what does it mean for two ROC curves to be "close"?

Problem Statement

Can we design a post-processing method that enforces fairness across the $\underbrace{\textit{entire range of decision thresholds}}_{\textit{Bringing the ROC curves closer}}$*, while maintaining minimal degradation in performance?*

To rigorously define the notion of *closeness* between ROC curves, we introduced a distance metric in Section 3.1. This metric provides a principled way to measure disparities across decision thresholds. Notably, our definition subsumes Equalized Odds as a special case, offering both theoretical elegance and practical relevance. This gives us confidence that the chosen distance metric, denoted ε-Equalized ROC, is meaningful.

Problem Statement

Can we design a post-processing method that enforces fairness across the $\underbrace{\textit{entire range of decision thresholds}}_{\varepsilon-EqualizedROC}$*, while maintaining minimal degradation in predictive performance?*

We now turn to the second part of the problem: maintaining predictive performance. Common metrics such as accuracy are insufficient in this context because they evaluate performance at a single threshold, whereas our interest lies in performance across the full range of thresholds. A more appropriate metric in this setting is the *Area Under the ROC Curve (AUC)*, which captures the model's overall ranking quality across thresholds.

Problem Statement

Can we design a post-processing method that enforces fairness across the entire range of decision thresholds, $\underbrace{\varepsilon - EqualizedROC}$ *while maintaining minimal degradation in predictive performance?* $\underbrace{\text{maximize AUC}}$

We can now formally define our objective in mathematical terms.

Given a scoring function $s \in \mathcal{S}$, our goal is to find a transformed score $h \in \mathcal{H}(s) = \mathcal{S}|_s$ such that h satisfies the fairness constraint defined by ε-Equalized ROC. At the same time, we seek to minimize the loss in AUC incurred by this transformation. Specifically, we aim to minimize the fairness-aware loss $\mathcal{L}_F = \texttt{AUC}_s - \texttt{AUC}_h$, ensuring that the predictive utility of the original score is retained as much as possible.

This leads to the following optimization problem:

$$h^* \in \arg max_{h \in \mathcal{S}|_s} \texttt{AUC}_h \tag{3.2}$$

$$\text{subject to} \quad \left\| \texttt{ROC}_h^0(\alpha) - \texttt{ROC}_h^1(\alpha) \right\|_1 \leq \varepsilon, \quad \forall \alpha \in [0,1]$$

This formulation captures our dual objectives: enforcing fairness as measured by the ROC distance across all thresholds, and preserving as much of the model's predictive power as possible, as measured by AUC.

3.3 Query Model

Let $\mathcal{T} = \{t_1, \ldots t_k\}$ be the set of thresholds at which we sample $\texttt{ROC}_s$ for each sensitive group ($t_i = \frac{i}{k}$). Let $\mathcal{Q}^a(t_i)$ denote the query output at threshold t_i for sensitive group $A = a$ on the $\texttt{ROC}_s^a$. $\mathcal{Q}^a(t_i) \triangleq ROC_{H_s^a, G_s^a}(t_i)$.

Abusing notations, we use $\mathcal{Q}^a(t_i)$ and $\mathcal{Q}_i^a$ interchangeably. Let $\mathcal{Q}^a = (\mathcal{Q}_1^a, \ldots, \mathcal{Q}_k^a)$ be the sequence of all query outputs for group a. In the next section, we construct the piecewise linear approximation of the group-wise ROC curves using the group-wise query outputs $\mathcal{Q}^a$.

3.4 Piecewise Linear Approximation (PLA) of ROC-curves

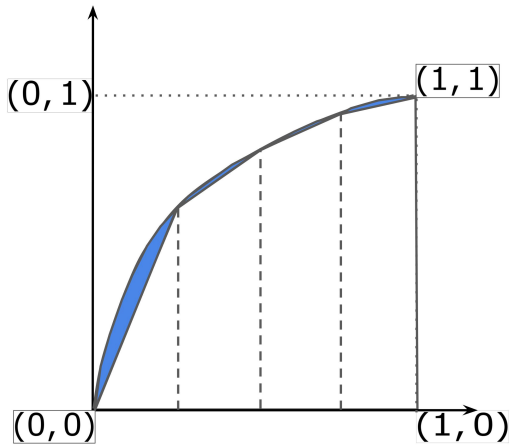


Figure 3.1 Piecewise Linear Approximation (PLA) of ROC-curve

To obtain the piecewise linear approximation (PLA), we sample k points from ROC and construct a straight line from $\mathcal{Q}_i^a$ to $\mathcal{Q}_{i+1}^a$ for all $i = 1 \dots k-1$. Lastly, we join $(0,0)$ to $\mathcal{Q}_1^a$ (see Figure 3.1). Following these steps on the query sets $\mathcal{Q}^a$ will generate the PLAs for protected groups $a \in \{0,1\}$. We denote by $\widehat{G_s^a}, \widehat{H_s^a}$ the cumulative distributions induced by the linear approximation of the ROC-curve on s.

Due to PLA, we incur a loss $\mathcal{L}_{LPA}$ in AUC_{H_s,G_s}(shaded region in Figure (3.1)). $\mathcal{L}_{LPA}$ is inversely proportional to the number of queries k, see Section 4.1 for bounds on this loss. Hence, we shall ignore this loss in our fairness analysis as it can be brought arbitrarily close to 0 by increasing k.

3.5 Algorithm Description

Since we are using a post-processing technique to ensure fairness, it is impossible to shift any ROC above its current position, i.e., build a classifier corresponding to any point in the epigraph (the points above the ROC curve) of $\texttt{ROC}_s$ just with the help of s. Interestingly, a classifier representing a point in the hypograph (points below the curve) of $s \cap \mathcal{S}$ can be obtained through randomization on the predicted scores (see Chapter 3 in Barocas et al. [2023]). The key idea involves abstracting out the convex hull (Fig 2.2) formed by the three points $(0,0)$, $(1,1)$ and $\mathcal{Q}_i^{up}$, and sampling outcomes from classifiers representing $(0,0)$, $(1,1)$[1] and $\mathcal{Q}_i^{up}$ with specific probabilities. By taking convex combinations of the three aforementioned points in the ROC space, we can represent any point lying in their convex hull. The exact convex combinations are described in **C2**. We leverage this property to achieve ε_1-Equalized

[1]Note that $(0,0)$ and $(1,1)$ represent 'always reject' and 'always accept' classifiers.

ROC. We denote this space as *ROC-space of* $s - \mathcal{S}|_s$. Each point in $\mathcal{S}|_s$ represents a binary classifier in terms of its performance at a certain threshold t. Each point is of the form $(FPR(t), TPR(t))$.

In the realm of binary classification, it is a common occurrence for one group to be subject to discrimination. Specifically, if we plot $\texttt{ROC}_s^0, \texttt{ROC}_s^1$, we will find that one of the ROCs is notably situated below the other. For this study, the ROC predominantly above the other will be designated as ROC_{up}, while the other ROC will be referred to as ROC_{down}. We believe this is a reasonable assumption because we observed that in most classifiers (for which we present the results and others we explored on the datasets mentioned in Section E3), the ROCs don't intersect or intersect at regions where $FPR \leq 0.2$ or $TPR \geq 0.5$. Typically, no practitioner will work in those areas of ROCs. We leave for future work to address intersecting ROCs.

Let $\mathcal{Q}^{up}, \mathcal{Q}^{down}$ be the corresponding set of query points for $\texttt{ROC}_{up}, \texttt{ROC}_{down}$ respectively. We also denote their fair counterparts by $\widetilde{\mathcal{Q}}^{up}, \widetilde{\mathcal{Q}}^{down}$.

3.5.1 Algorithm Defintions

We need to transport $\texttt{ROC}_{up}$ towards $\texttt{ROC}_{down}$ such that the new ROCs are within ε distance of each other. Our approach is geometric. We need to identify certain points/curves in the epigraph of $\texttt{ROC}_{down}$ as follows.

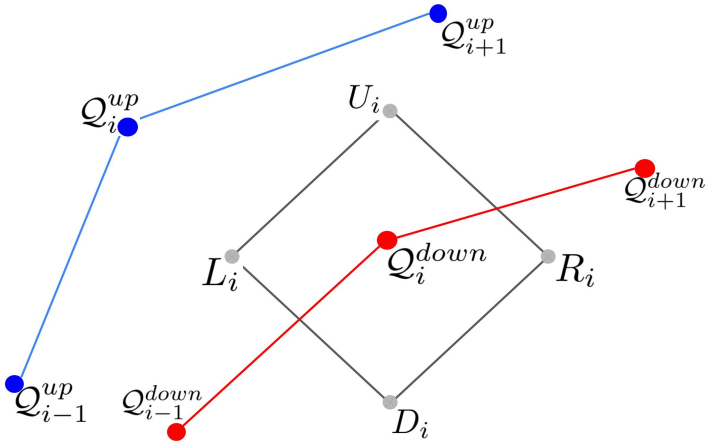


Figure 3.2 Norm Boundary and Boundary Cut

Definition 3.5.1 (Norm Boundary)**.** *The set of all points within* ε *distance* (ℓ_1 *norm) from* $\mathcal{Q}_i^{down}$ *is known as the* norm set $\mathfrak{C}_i$. *Formally, we have*

$$\mathfrak{C}_i \triangleq \{x : x \in [0,1]^2, ||x - \mathcal{Q}_i^{down}||_1 \leq \varepsilon\}$$

The set of all points exactly ε *distance (in* $\mathcal{L}_1$ *norm) from* $\mathcal{Q}_i^a$ *is known as* Norm Boundary $\mathfrak{B}_i$. *Formally,*

$$\mathfrak{B}_i \triangleq \{x : x \in [0,1]^2, ||x - \mathcal{Q}_i^{down}||_1 = \varepsilon\}$$

Additionally, we denote the vertices of the Norm Boundary Rhombus (starting from the topmost point and moving clockwise) as U_i, R_i, D_i, and L_i.

We say that an index $i \in [1, 2, \ldots, k]$ is a Boundary Cut index when ROC_{up} intersects the Norm Boundary $\mathfrak{B}_i$. Formally,

Definition 3.5.2 (Boundary Cut)**.** *Index $i \in [1, 2, \ldots, k]$ is a* Boundary Cut index *when $\mathfrak{B}_i \cap ROC_{up} \neq \phi$.*

3.6 Shifts

We now introduce and formally define three distinct types of shifts that play a central role in our algorithm. These shifts correspond to specific operations on the threshold points of the upper ROC curve, denoted as ROC_{up}. Each shift captures a particular mode of movement or adjustment applied to these threshold points. Collectively, these operations characterize how ROC_{up} is transformed or transported throughout the algorithm. In essence, the progression of ROC_{up} under these shifts models the underlying mechanism by which the algorithm enforces or adapts fairness constraints.

3.6.1 UpShift

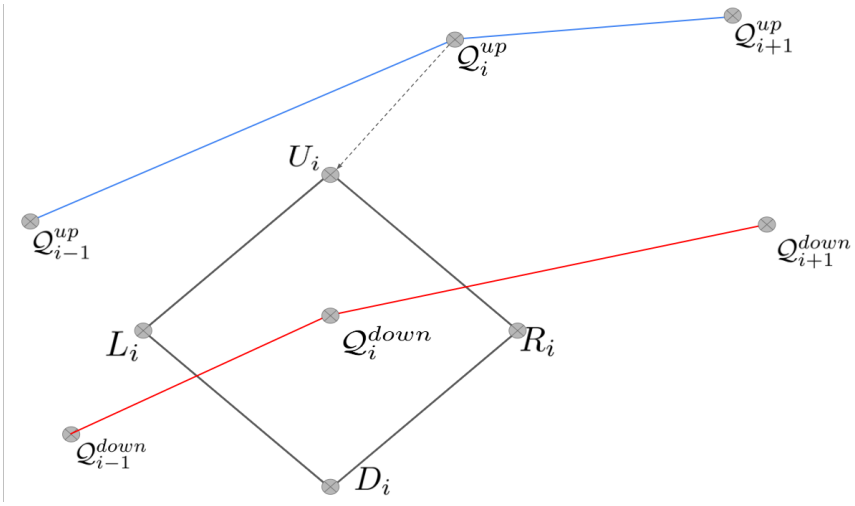


Figure 3.3 Upshift

For a given $i \in [1, 2, \ldots, k]$, Upshift is the transportation of $\mathcal{Q}_i^{up}$ to the point U_i.

Definition 3.6.1 (UpShift)**.** *For a given $i \in [1, 2, \ldots, k]$, Upshift is the transportation of $\mathcal{Q}_i^{up}$ to the point U_i. Formally,* UpShift *can be defined as the function that returns a fair threshold $\widetilde{\mathcal{Q}}_i^{up}$ (i.e., U_i) by taking the $\mathcal{Q}_i^{down}$ and ε as the arguments.*

3.6.2 LeftShift

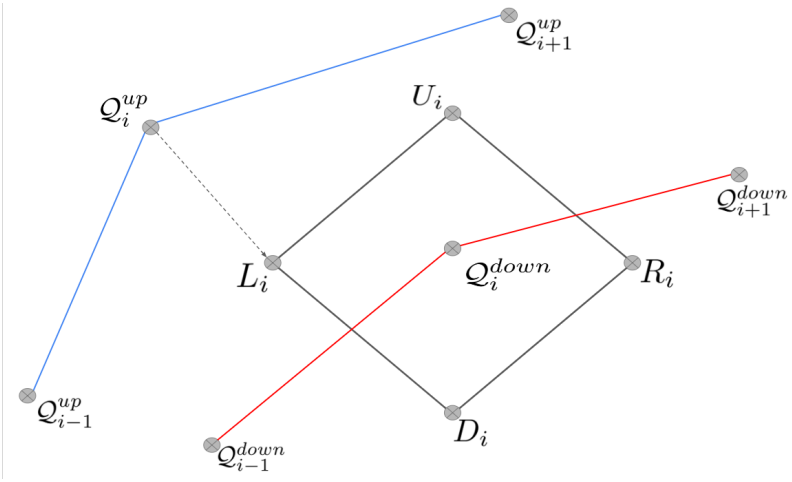


Figure 3.4 LeftShift

For a given $i \in [1, 2, \ldots, k]$, Leftshift is the transportation of $\mathcal{Q}_i^{up}$ to the point L_i. Formally,

Definition 3.6.2 (LeftShift)**.** *LeftShift is a function that returns a fair threshold $\widetilde{\mathcal{Q}}_i^{up}$ (i.e. L_i) by taking the $\mathcal{Q}_i^{down}$ and ε as the arguments.*

3.6.3 CutShift

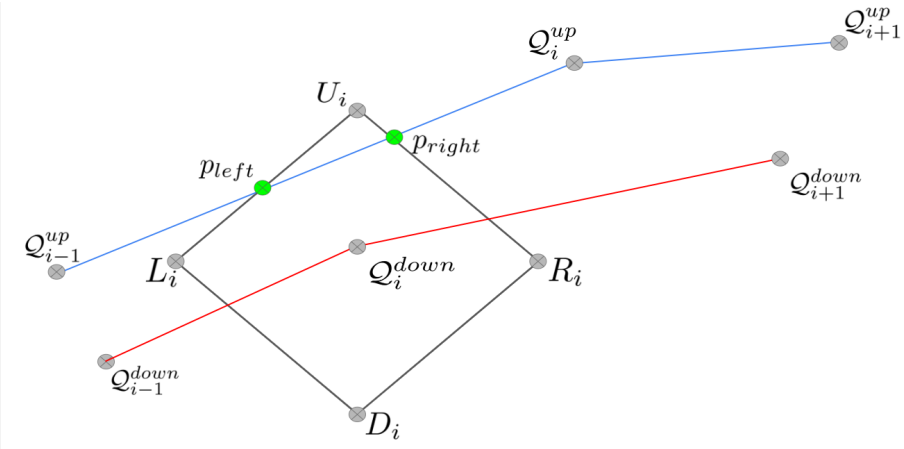


Figure 3.5 CutShift

Definition 3.6.3 (CutShift)**.** *For a given $i \in [1, 2, \ldots, k]$ (representing the index of the ROC_{down}), we run through all the points of the ROC_{up} and return the set of all points that intersect the Norm Boundary $\mathfrak{B}_i$. Formally, we define* Cutshift *as a function that takes $\mathcal{Q}_i^{down}$ and ε as the arguments and*

*returns $ROC_{up} \cap \mathfrak{B}_i$. The set $ROC_{up} \cap \mathfrak{B}_i$ can be represented as $\{p_{left}, p_{right}\}$ denoting the points at the intersection of ROC_{up} at the **left-side** of the Norm Boundary and the **right-side** of the Norm Boundary respectively.*

Now, we elaborate on the above procedure to transport points from ROC_{up} towards ROC_{down}.

3.6.4 Algorithm

We provide a geometric algorithm that returns a classifier equivalent to the scoring function h^* in $\mathcal{S}|_s$.

Algorithm 10 FROC ALGORITHM

Require: ROC_{up}, ROC_{down}, ε

Ensure: $FairROC_{up}$, $FairROC_{down}$

Initialize $i \leftarrow 1$, $k \leftarrow$ length(ROC_{up})
$FairROC_{up} \leftarrow \emptyset$, $FairROC_{down} \leftarrow ROC_{down}$
while $i < k - 1$ **do**
    $i \leftarrow i + 1$
    if BOUNDARYCUT(i, ε) == TRUE **then**
        $p_{left}, p_{right} \leftarrow$ CUTSHIFT$(i, ROC_{up}, ROC_{down})$
        if $FPR(\mathcal{Q}_i^{up}) \geq FPR(\mathcal{Q}_i^{down})$ **then**
            $\widetilde{\mathcal{Q}}_i^{up} \leftarrow p_{right}$
        else
            $\widetilde{\mathcal{Q}}_i^{up} \leftarrow p_{left}$
        end if
    else if $\mathcal{Q}_i^{up} \in$ HYPOGRAPH(ROC_{down}) **then**
        $\widetilde{\mathcal{Q}}_i^{up} \leftarrow \mathcal{Q}_i^{up}$
        continue
    else
        if Area$(\square \mathcal{Q}_{i+1}^{up} \mathcal{Q}_i^{up} \mathcal{Q}_{i-1}^{up} L_i) \geq$ Area$(\square \mathcal{Q}_{i+1}^{up} \mathcal{Q}_i^{up} \mathcal{Q}_{i-1}^{up} U_i)$ **then**
            $\widetilde{\mathcal{Q}}_i^{up} \leftarrow U_i$
        else
            $\widetilde{\mathcal{Q}}_i^{up} \leftarrow L_i$
        end if
    end if
    $FairROC_{up} \leftarrow$ APPEND$(\widetilde{\mathcal{Q}}_i^{up})$
end while

Note that Algorithm 10 treats ROC_{down} as *implicitly* fair. Also, by $Area(\square ABCD)$, we denote the area of the quadrilateral whose vertices are A, B, C, and D. This area is easily found in this context by

splitting $\square ABCD$ into two disjoint triangles- ΔABC and ΔACD and using Heron's formula Kendig [2000] on each triangle.

For example, consider $Area(\Delta \mathcal{Q}_i^{up} \mathcal{Q}_{i-1}^{up} L_i)$. Let $a = ||\mathcal{Q}_i^{up} \mathcal{Q}_{i-1}^{up}||_2$, $b = ||\mathcal{Q}_i^{up} L_i||_2$ and $c = ||\mathcal{Q}_{i-1}^{up} L_i||_2$. Additionally, we define $s = \frac{a+b+c}{2}$. Then, it is true that:

$$Area(\Delta \mathcal{Q}_i^{up} \mathcal{Q}_{i-1}^{up} L_i) = \sqrt{s(s-a)(s-b)(s-c)}$$

3.6.5 Obtaining fair classifier from the updated ROCs

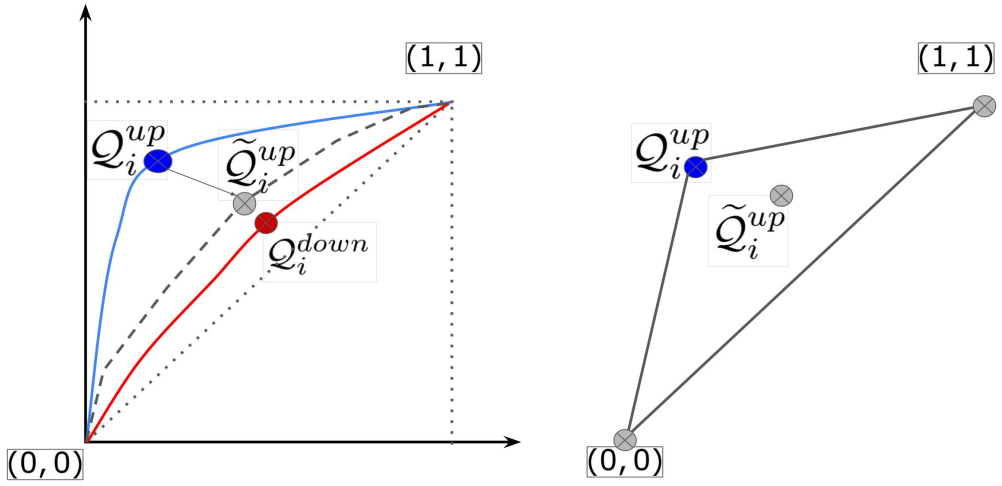


Figure 3.6 Convex Combinations

The algorithm described in the previous subsection returns the fair ROC curves according to ε_1-Equalized ROC. As a final step, we need to find the transformed classifier. We call it

$$\texttt{ConstructClassifier}(FairROC_{up}, FairROC_{down}, \texttt{ROC}_s^0, \texttt{ROC}_s^1)$$

which returns a probabilistic binary classifier representing $h = \mathcal{H}(s)$ such that it represents the FairROCs. We construct one using the procedure explained in Section 3.6.5.

3.7 Summary

In this chapter, we presented a comprehensive formulation of our fairness-aware post-processing framework. We began by introducing the concept of the ε-equalized ROC, a principled extension of Equalized Odds that enforces fairness not at a single decision threshold but uniformly across the entire

ROC curve. This generalization allows us to account for the full spectrum of classifier behavior, offering a more robust and threshold-independent fairness criterion.

We then formalized the problem setting and introduced a query-based model that enables access to classifier performance at various thresholds. To facilitate computational tractability, we approximated the ROC curves using a piecewise linear representation, which serves as a practical surrogate for continuous ROCs.

The chapter culminated in a detailed description of our algorithmic framework. We defined a set of shift operations that transform the ROC curves while preserving as much predictive performance as possible. By applying these shifts iteratively and in a structured manner, our method constructs updated, fair ROC curves from which a final post-processed classifier is derived.

In the next chapter, we will analyze and prove the theoretical properties of the proposed algorithm.

Chapter 4

Theoretical Analysis

> Algebra is like sheet music; the important thing isn't whether you can read it, but whether you can hear it.
>
> – Neils Bohr in *Oppenheimer*

Abstract

This chapter provides a rigorous theoretical analysis of the post-processing framework introduced in the previous chapter. We begin by analyzing the approximation guarantees offered by the piece-wise linear representation (PLA) of ROC curves, which forms the computational foundation of our algorithm. We then examine the trade-offs in predictive performance introduced by our fairness constraints, through a detailed AUC loss analysis.

Subsequently, we present optimality results for specific transformation operations, including *CutShift*, *UpShift*, and *LeftShift*, which collectively define the algorithm's behavior. These results highlight boundary conditions and establish when certain operations achieve provable optimality. We also explore the sample complexity of the method, offering guarantees on the generalization of fairness and utility under finite data regimes. Finally, we discuss extensions and variants that build upon our core framework, pointing toward broader applicability. This chapter thus solidifies the theoretical underpinnings of our algorithm and sets the stage for empirical validation.

Having introduced the algorithmic framework in the previous chapter, we now turn our attention to its theoretical foundations. This chapter provides the mathematical rigor necessary to understand the guarantees, trade-offs, and limitations of our approach. We begin by analyzing the approximation properties of the piece-wise linear representation used to model ROC curves, which underpins the algorithm's structure. This sets the stage for a deeper exploration of fairness-performance trade-offs and the optimality of the transformation operations introduced earlier.

As described in Section (3.4), we work with PLA of the ROC curves $ROC_{H_s^a, G_s^a}$, $a \in \{0, 1\}$. This causes a *loss* in area under ROC. We denote this loss by $\mathcal{L}_{PLA}$ and is quantified as the difference in AUCs of $ROC_{H_s^a, G_s^a}$ and $ROC_{\widehat{H_s^a}, \widehat{G_s^a}}$.

In Section 3.6.4, transporting the ROC query points, $\mathcal{Q}^{up}$, introduces a decrease of the area under the ROC curve due to the transformation of the scoring function s to h. We denote this loss by $\mathcal{L}_{AUC}$. This loss can be quantified as the difference in AUCs of $ROC_{\widehat{H_s^a}, \widehat{G_s^a}}$ and $ROC_{H_h^a, G_h^a}$ The total loss in AUC, $\mathcal{L}$, induced by `FROC` is given by: $\mathcal{L} = \mathcal{L}_{PLA} + \mathcal{L}_{AUC}$

4.1 PLA Analysis

We start our analysis by making a few standard assumptions regarding the continuity and differentiability of the cumulative distributions on the family of scoring functions $\mathcal{S}$. We adopt a less stringent assumption than that presented in Vogel et al. [2021], as we impose only an upper bound on the slopes. This contrasts with the approach in Vogel et al. [2021], which necessitates both an upper and lower bound on the slopes.

Assumption 4.1.1. *We assume that the rate of change (with respect to the thresholds t) of the $TPRs$ and $FPRs$ is upper bounded. I.e. we assume that $\exists\, u_T, u_F \in \mathbb{R}$ such that $\frac{d\,TPR}{dt} \leq u_T$ and $\frac{d\,FPR}{dt} \leq u_F$.*

Theorem 4.1.1. *Let $ROC_{\widehat{H_s^a}, \widehat{G_s^a}}$ be the* Piecewise Linear Approximation (PLA) *of the ROC curve $ROC_{H_s^a, G_s^a}$, constructed over a query set of k equidistant thresholds,*

$$\mathcal{T} = \{t_i \mid t_i = i/k, \quad \forall i \in [k]\}.$$

Then, the corresponding approximation loss, denoted as $\mathcal{L}_{PLA}$, is bounded by:

$$\mathcal{L}_{PLA} \leq \frac{1}{2} \frac{u_T u_F}{k^2} \times k = \frac{1}{2} \frac{u_T u_F}{k}. \tag{4.1}$$

Proof. The approximation loss $\mathcal{L}_{PLA}$ is represented by the shaded region in Figure 4.1.

To derive an upper bound on $\mathcal{L}_{PLA}$, we consider the worst-case scenario where the true ROC curve $ROC_{H_s^a, G_s^a}$ deviates maximally from its PLA $ROC_{\widehat{H_s^a}, \widehat{G_s^a}}$. The maximum possible deviation occurs when the ROC curve is stretched to the dotted line shown in Figure 4.2.

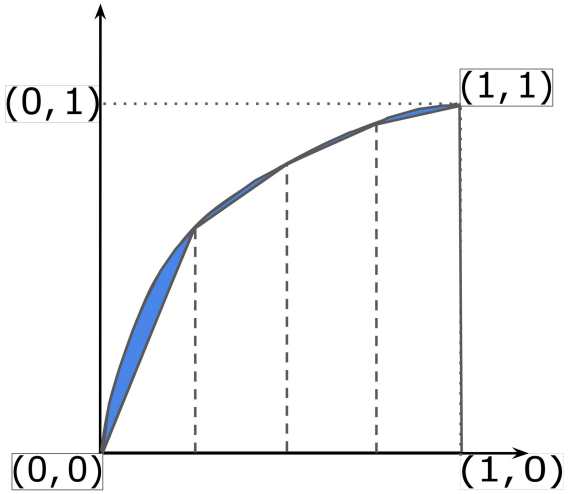


Figure 4.1 Illustration of the approximation loss $\mathcal{L}_{PLA}$.

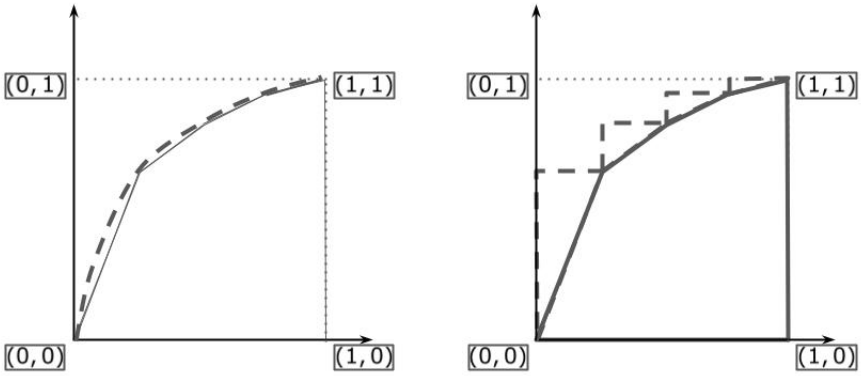


Figure 4.2 Maximally stretched ROC curve (dotted line).

However, the area cannot exceed the upper bound imposed by the dotted line, as ROCs are one-to-one and monotonically increasing functions. This constraint ensures that the worst-case scenario for approximation loss is captured by the region shaded in Figure 4.3.

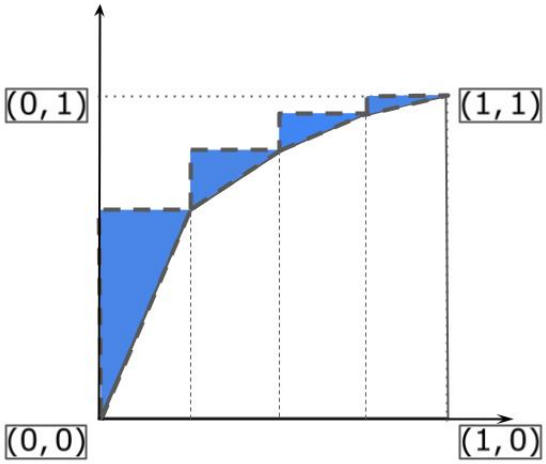


Figure 4.3 The maximum possible loss of AUC due to linear interpolation, represented by the dark blue shaded area.

The goal now is to bound the total area of the blue-shaded triangular regions formed by the linear interpolation error. Each triangle has:

- A **base length** of $\frac{1}{k} \times u_F$ (since there are k thresholds and the maximum slope of the false positive rate (FPR) with respect to thresholds is u_F).
- A **maximum possible height** of $\frac{1}{k} \times u_T$ (since there are k thresholds and the maximum slope of the true positive rate (TPR) with respect to thresholds is u_T).

Thus, the maximum possible area of each triangular segment is given by:

$$\frac{1}{2}\frac{u_T u_F}{k^2}. \tag{4.2}$$

Since there are k such intervals along the ROC curve, the total approximation loss accumulates as follows:

$$\mathcal{L}_{PLA} \leq \sum_{i=1}^{k} \frac{1}{2}\frac{u_T u_F}{k^2} = \frac{1}{2}\frac{u_T u_F}{k^2} \times k = \frac{1}{2}\frac{u_T u_F}{k}. \tag{4.3}$$

□

Thus, in the limit as $k \to \infty$, the approximation loss vanishes:

$$\lim_{k\to\infty} \mathcal{L}_{PLA} = \lim_{k\to\infty} \frac{1}{2}\frac{u_T u_F}{k} = 0. \tag{4.4}$$

4.2 AUC Loss Analysis

We start our analysis by making a few assumptions regarding the spacing of the ROC thresholds and the ROC curve.

Assumption 4.2.1. *We have two assumptions:*

- $\forall i \in \{1, 2, \ldots, k\}$*, we assume that* $FPR(\mathcal{Q}_{i-1}^{down}) \leq FPR(\mathcal{Q}_i^{up}) \leq FPR(\mathcal{Q}_{i+1}^{down})$.
- *We assume that the* ROC_{up} *can intersect any Norm boundary (i.e.* $(\mathfrak{B}_i)_{i\in\{1,2,\ldots,k\}}$*) at most 2 times.*

We note that even if **Assumption 4.2.1** does not hold, `FROC` remains operational and continues to produce outputs that are ε_1-Equalized ROC fair. However, under these conditions, the optimality with respect to AUC is not guaranteed, as **Theorem 4.4** no longer applies. The necessity of these assumptions is discussed in greater detail in the extended version of this paper.

While we currently rely on Assumption 4.2.1 due to the use of uniformly chosen thresholds, it is not strictly necessary. If thresholds are selected appropriately, specifically, in a non-uniform manner, Assumption 4.2.1 can always be satisfied by construction. For instance, one could employ a binary search procedure to identify suitable threshold values that ensure the assumption holds. However, for the sake of simplicity and to maintain focus on the main exposition, we do not delve into those details here. The key point we wish to emphasize is that, if needed, Assumption 4.2.1 can be eliminated without compromising the framework.

4.2.1 Boundary Optimality

This section proves that all optimally fair points must lie on some Norm Boundary. We do this by establishing that the performance of any point in the Norm Set can be improved by appropriate transportation to a point on the Norm Boundary.

All Optimal Points Lie on the Norm Boundary

Theorem 4.2.1 (Norm Boundary)**.** *Let* $(\widetilde{\mathcal{Q}}_i^{up})_{i\in\{1,2,\ldots,k\}}$ *denote the set of optimal fair points, i.e., the set of threshold points that maximize the AUC while satisfying the* ε*-fairness criterion. Then, these optimal fair points must necessarily be a subset of the norm boundary, formally:*

$$(\widetilde{\mathcal{Q}}_i^{up})_{i\in\{1,2,\ldots,k\}} \subseteq (\mathfrak{B}_i)_{i\in\{1,2,\ldots,k\}}.$$

Proof. We proceed by contradiction. Suppose that there exists an optimal fair point C that lies in the *interior* of the norm set rather than on its boundary. That is, C leads to an ROC curve with the maximum possible AUC while satisfying the ε_1-Equalized ROC fairness criterion.

In Figure 4.4, the point $\mathcal{Q}_i^{up}$ has been transported to an interior point C within the norm set. The shaded region in blue represents the AUC loss incurred due to this transformation.

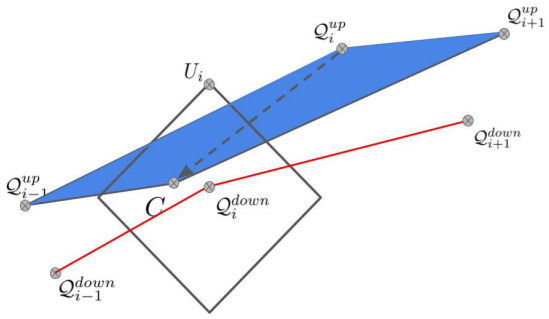


Figure 4.4 The blue-colored region represents the AUC loss incurred when the point is in the interior of the norm set.

However, as illustrated in Figure 4.5, the AUC loss can be reduced by selecting a point on the norm boundary instead. Specifically, choosing a CutShift point (denoted as A) on the boundary results in a strictly lower AUC loss compared to selecting the interior point C.

Formally, the AUC loss associated with the interior point C is given by the area:

$$\mathcal{L}_C = \text{Area}(\square \mathcal{Q}_{i-1}^{up} C \mathcal{Q}_{i+1}^{up} \mathcal{Q}_i^{up}).$$

On the other hand, the AUC loss when selecting the boundary point A is:

$$\mathcal{L}_A = \text{Area}(\square \mathcal{Q}_i^{up} A \mathcal{Q}_{i+1}^{up}).$$

By construction, these areas satisfy the following relation:

$$\text{Area}(\square \mathcal{Q}_{i-1}^{up} C \mathcal{Q}_{i+1}^{up} \mathcal{Q}_i^{up}) = \text{Area}(\square \mathcal{Q}_i^{up} A \mathcal{Q}_{i+1}^{up}) + \text{Area}(\square \mathcal{Q}_{i-1}^{up} C \mathcal{Q}_{i+1}^{up} A). \tag{4.5}$$

Since areas are always non-negative, it follows that:

$$\text{Area}(\square \mathcal{Q}_{i-1}^{up} C \mathcal{Q}_{i+1}^{up} \mathcal{Q}_i^{up}) \geq \text{Area}(\square \mathcal{Q}_i^{up} A \mathcal{Q}_{i+1}^{up}). \tag{4.6}$$

This contradicts the assumption that C was an optimal fair point because the selection of A strictly reduces the AUC loss. Therefore, C cannot be an optimal fair point, implying that all optimal fair points must lie on the norm boundary. □

4.2.2 CutShift Optimality

Theorem 4.2.2. *If i is a boundary cut point, then the CutShift operation must be performed. Given the two points returned by the CutShift operation, denoted as p_{left} and p_{right}, the point that is closer to $\mathcal{Q}_i^{up}$ must be chosen, formally:*

$$\widetilde{\mathcal{Q}}_i^{up} = \arg \min_{p \in \{p_{left}, p_{right}\}} |FPR(\mathcal{Q}_i^{up}) - FPR(p)|.$$

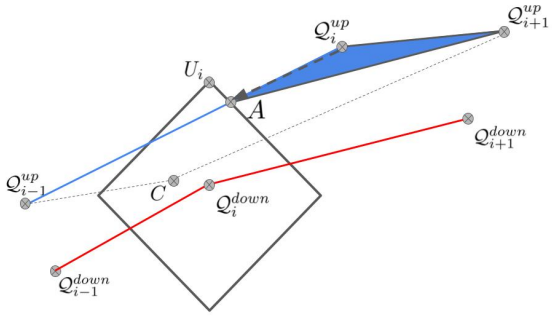


Figure 4.5 The dark blue region represents the new AUC loss after selecting a point on the norm boundary. The light blue region represents the previous AUC loss.

Proof. We proceed by contradiction. Suppose that an optimal fair point C lies on the norm boundary rather than being selected via the CutShift operation. That is, C leads to an ROC curve with the maximum possible AUC while satisfying the ε_1-Equalized ROC fairness criterion.

As illustrated in Figure 4.6, the point $\mathcal{Q}_i^{up}$ has been transported to C within the norm set. The shaded light blue region represents the AUC loss incurred due to this transformation.

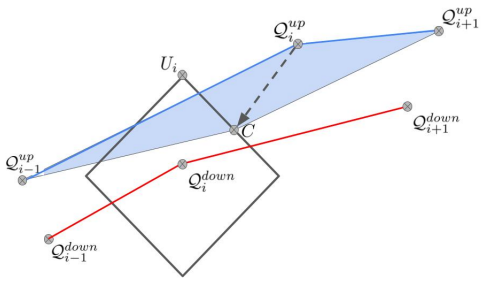


Figure 4.6 The CutShift operation is not followed. The light blue region represents the AUC loss due to this operation.

However, as shown in Figure 4.7, selecting a point on the norm boundary via the CutShift operation leads to a lower AUC loss. Specifically, selecting the CutShift point A reduces the AUC loss relative to C.

Formally, the AUC loss associated with the selection of point C is given by:

$$\mathcal{L}_C = \text{Area}(\square \mathcal{Q}_{i-1}^{up} C \mathcal{Q}_{i+1}^{up} \mathcal{Q}_i^{up}).$$

In contrast, the AUC loss when selecting the CutShift point A is:

$$\mathcal{L}_A = \text{Area}(\square \mathcal{Q}_i^{up} A \mathcal{Q}_{i+1}^{up}).$$

Since:

$$\text{Area}(\square \mathcal{Q}_{i-1}^{up} C \mathcal{Q}_{i+1}^{up} \mathcal{Q}_i^{up}) = \text{Area}(\square \mathcal{Q}_i^{up} A \mathcal{Q}_{i+1}^{up}) + \text{Area}(\square \mathcal{Q}_{i-1}^{up} C \mathcal{Q}_{i+1}^{up} A), \tag{4.7}$$

and the last term is non-negative:

$$\text{Area}(\square \mathcal{Q}_{i-1}^{up} C \mathcal{Q}_{i+1}^{up} A) \geq 0, \tag{4.8}$$

It follows that:

$$\mathcal{L}_C \geq \mathcal{L}_A. \tag{4.9}$$

This contradicts the assumption that C is an optimal fair point, as selecting A via CutShift strictly decreases AUC loss. Therefore, all optimal fair points must be selected using the CutShift operation. □

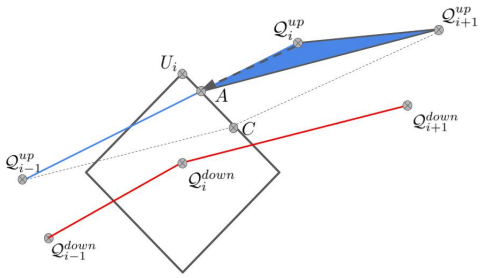


Figure 4.7 The CutShift operation is followed. The dark blue region represents the AUC loss, which is lower than the AUC loss in Figure 4.6.

4.2.3 UpShift and LeftShift

Theorem 4.2.3 (UpShift)**.** *If i is not a boundary cut point and if:*

$$\text{Area}(\square \mathcal{Q}_{i+1} \mathcal{Q}_i \mathcal{Q}_{i-1} L_i) \geq \text{Area}(\square \mathcal{Q}_{i+1} \mathcal{Q}_i \mathcal{Q}_{i-1} U_i),$$

Then the UpShift operation must be performed, and the resulting point U_i is the new fair point $\widetilde{\mathcal{Q}}_i^{up}$. Otherwise, the LeftShift operation must be performed, and the resulting point L_i is the new fair point $\widetilde{\mathcal{Q}}_i^{up}$.

Proof. By a similar argument as in the previous proofs, we argue (through Figures 4.8, 4.9, and 4.11) that either the point recommended by UpShift (U_i) or LeftShift (L_i) is the optimal fair point.

To determine which shift operation should be performed, we compute and compare the AUC losses using Heron's formula. Given a quadrilateral $\square ABCD$, we compute the area by splitting it into two disjoint triangles, $\triangle ABC$ and $\triangle ACD$.

For instance, consider the area of $\triangle \mathcal{Q}_i^{up}\mathcal{Q}_{i-1}^{up}L_i$. Let:

$$a = ||\mathcal{Q}_i^{up}\mathcal{Q}_{i-1}^{up}||_2,$$
$$b = ||\mathcal{Q}_i^{up}L_i||_2,$$
$$c = ||\mathcal{Q}_{i-1}^{up}L_i||_2.$$

Defining the semi-perimeter as:

$$s = \frac{a+b+c}{2},$$

The area is given by Heron's formula:

$$\text{Area}(\triangle \mathcal{Q}_i^{up}\mathcal{Q}_{i-1}^{up}L_i) = \sqrt{s(s-a)(s-b)(s-c)}.$$

By computing and comparing the areas of both candidate quadrilaterals, we select the shift operation that minimizes AUC loss, ensuring an optimal fairness transformation. □

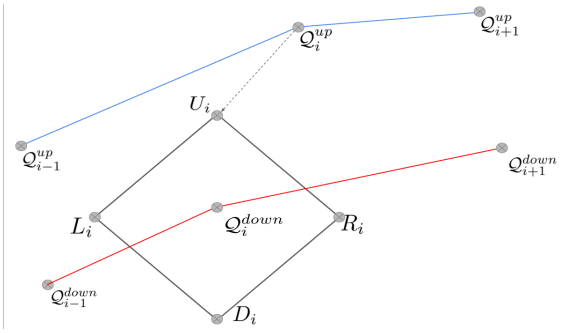


Figure 4.8 UpShift operation: The dotted arrow represents the movement from $\mathcal{Q}_i^{up}$ to U_i.

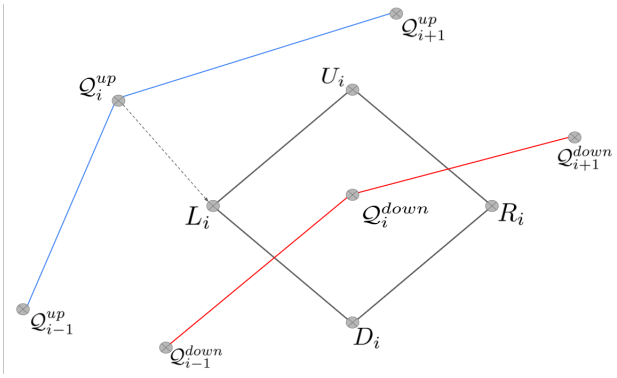


Figure 4.9 LeftShift operation: The dotted arrow represents the movement from $\mathcal{Q}_i^{up}$ to L_i.

We now present the main theorem's proof:

Step 1: We prove that all optimally fair points $(\widetilde{\mathcal{Q}}_i^{up})_{i\in\{1,2,\ldots,k\}}$ must lie on the Norm Boundaries of the

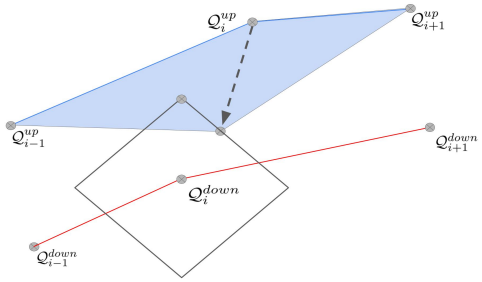


Figure 4.10 UpShift operation is not followed. The light blue region represents the AUC loss.

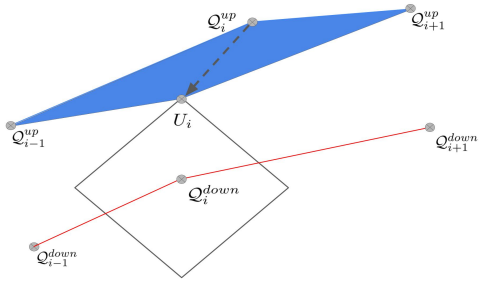


Figure 4.11 UpShift operation is followed. The dark blue region represents the AUC loss, which is lower than the previous AUC loss (Figure 4.10).

corresponding $\mathcal{Q}_i^{down}$. (i.e. $(\mathfrak{B}_i)_{i\in\{1,2,\ldots,k\}}$)
Step 2: We then prove that if $\mathfrak{B}_i \cap ROC_{up} \neq \phi$, then the CutShift transportation is the optimal transportation.
Step 3: We then prove that if $\mathfrak{B}_i \cap ROC_{up} = \phi$, then, based on the Cover and aforementioned area condition, the UpShift or the LeftShift transportation is the optimal transportation.

4.2.4 Sample Complexity

If the **Assumption 4.2** holds true, then we have the following analysis:

- All UpShift Operations will be constant time ($O(1)$).
- All CutShift Operations will also be constant time ($O(1)$). This is because **Assumption 4.2** ensures that we do not have to run through the entire length of ROC_{up} to find the intersection points, i.e., p_{left} and p_{right}.

Therefore, the running time of `FROC` is $O(k)$. However, when no assumptions are made, then the CutShift operation is no longer $O(1)$. We may have to run through the entire length of ROC_{up} to find the intersection points, i.e., p_{left} and p_{right}. This makes the CutShift operation $O(k)$. Therefore, the time complexity of `FROC` is $O(k^2)$.

4.3 Further Variants

4.3.1 Multiple Protected Groups

Our approach is extendable to scenarios involving multiple protected groups. The procedure begins by applying the FROC algorithm to the ROC curve that is immediately above the bottom-most ROC curve. Subsequently, FROC is applied to the ROC curve directly above the one previously processed. This iterative application continues until the top ROC curve is reached. While this method ensures -Equalized ROCfairness across all protected groups, the proof of optimality remains an open question.

4.3.2 Intersection of ROC Curves

In cases where the ROC curves intersect more than twice, our algorithm will still produce a fair output. However, the existing optimality theorems do not apply in such scenarios. When intersections occur, the FROC algorithm can be applied to the dominant segments of the ROC curves—those portions where no intersections are present.

4.4 Summary

In this chapter, we developed the theoretical foundations of our fairness-aware post-processing framework. Through detailed analysis of the piece-wise linear approximation (PLA), we demonstrated that our discretized model retains meaningful guarantees with respect to the continuous ROC curve. This approximation is essential to the computational feasibility of the algorithm.

We then investigated the impact of fairness interventions on predictive utility, quantifying the potential AUC loss introduced by our method. This led to a deeper understanding of the inherent trade-offs in enforcing fairness across decision thresholds.

Building on this, we established optimality results for a suite of transformation operations—namely CutShift, UpShift, and LeftShift—by characterizing their behavior at boundary points and under specific constraints. These results provide formal guarantees for the building blocks of our algorithm.

Furthermore, we analyzed the sample complexity of our approach, deriving bounds on the number of samples required to ensure reliable fairness enforcement and generalization. Lastly, we introduced several theoretical variants that offer extensions or refinements to the core algorithm.

Together, these contributions validate the theoretical soundness of our proposed method and lay the groundwork for the empirical evaluations presented in the following chapter.

Chapter 5

Empirical Analysis

> Theory can only take you so far....
>
> – J. Robert Oppenheimer

> In theory, there is no difference between theory and practice. In practice, there is.
>
> – Attributed to Yogi Berra

> Experiment is the sole source of truth. It alone can teach us something new; it alone can give us certainty.
>
> –Henri Poincaré

Abstract

In this chapter, we present an empirical evaluation of FROC, comparing its performance with existing post-processing techniques across multiple benchmark datasets. We describe our *experimental setup*, outline the *datasets and classifiers* used, and analyze the *results* in terms of fairness and accuracy.

Having developed and analyzed our framework in the previous chapters, we now turn to its empirical evaluation. This chapter aims to validate the practical utility of FROC by benchmarking it against existing post-processing methods across standard datasets. Through carefully designed experiments, we assess its effectiveness in improving fairness while maintaining competitive accuracy. We begin by detailing the experimental setup, including the datasets and classifiers used in our evaluation.

5.1 Experimental Setup

Datasets: We evaluate the performance of our proposed method using two widely recognized benchmark datasets:

- **ADULT** [Becker and Kohavi, 1996]: This dataset comprises U.S. census data and is commonly employed for income prediction tasks. In our experiments, we designate *MALE* and *FEMALE* as the protected demographic groups.
- **COMPAS** [Angwin et al., 2016]: This dataset is utilized for predicting recidivism risk. We define the protected groups as *BLACK* and *OTHERS*.

In addition to the above, we perform supplementary evaluations on the CelebA dataset, with detailed results presented in Appendices E and F.

Classifiers: We apply `FROC` to Receiver Operating Characteristic (ROC) curves derived from the following classification models:

- **C1: Fair Neural Network Classifier (FNNC)** [Padala and Gujar, 2020] – A neural network-based model that incorporates fairness constraints during training.
- **C2: Logistic Regression** – A linear classification model widely used due to its interpretability and efficiency.
- **C3: Random Forest** – An ensemble learning method based on decision trees, known for its robustness and predictive performance.

The implementation of C1 follows the original source provided by the authors. For C2 and C3, we use the standard implementations available in the `scikit-learn` library.

Post-Processing Baselines: We compare `FROC` against existing fairness-aware post-processing techniques, including:

- **B1: FairProjection-CE and FairProjection-KL** [Alghamdi et al., 2022] – These methods adjust the score distribution through information projection techniques to achieve mean equalized odds fairness.

5.2 Experiments

We conduct empirical evaluations by training classifier C1 on both the ADULT and COMPAS datasets, while classifiers C2 and C3 are trained solely on the ADULT dataset. For each protected group, we generate Receiver Operating Characteristic (ROC) curves. Notably, the Fair Neural Network

Classifier (FNNC) is trained *without* enforcing its inherent fairness constraints, and the corresponding ROC curves are obtained.

Subsequently, we apply FROC across a range of fairness thresholds (specified by different values of ε), selecting the operating point that yields optimal classification accuracy. The classifiers adjusted via this procedure are denoted as C1-FROC, C2-FROC, and C3-FROC.

Given that FROC guarantees compliance with ε-Equalized Odds, we evaluate FNNC using the same set of ε thresholds to ensure consistency in fairness constraints across methods.

Baseline Comparison: We assess the performance of FROC and the baseline method B1 on the ADULT dataset using the *mean equalized odds* fairness metric, as defined in Alghamdi et al. [2022] (see Figure 5.3). To ensure a fair and meaningful comparison, we adopt identical training configurations for all base classifiers, following the setup established in the prior work of Alghamdi et al. [2022].

5.2.1 Results

We present experimental results on the COMPAS and ADULT datasets, comparing the performance of FROC against the Fair Neural Network Classifier (FNNC) and established post-processing baselines.

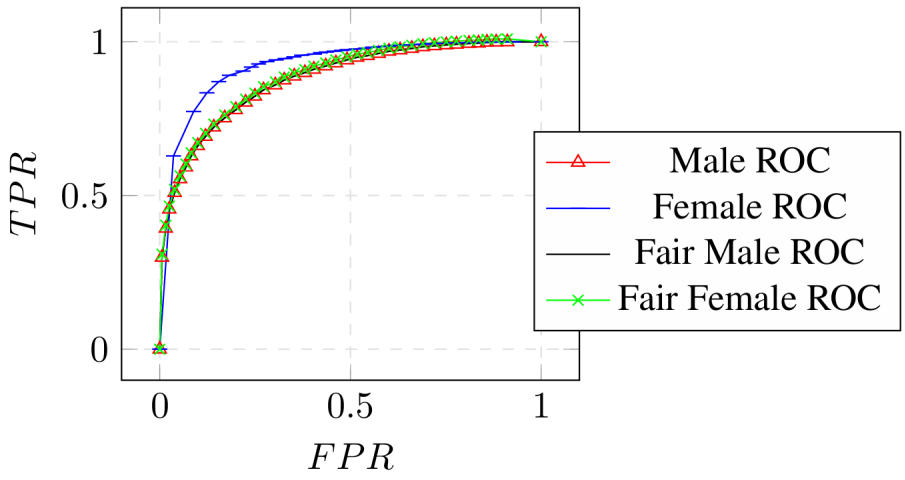


Figure 5.1 ROC curves of classifier C2 on the ADULT dataset, before and after applying FROC.

Figure 5.1 illustrates the ROC curves for classifier C2 on the ADULT dataset, both prior to and following the application of FROC. The ROC curve corresponding to the female subgroup consistently lies above that of the male subgroup, indicating a potential *systematic advantage for male individuals*. Based on this observation, we designate the male ROC as ROC_0 and apply FROC to adjust the female ROC, denoted ROC_1.

Prior to post-processing, the maximum disparity between the two ROC curves is observed to be *0.08*. After applying FROC with a fairness threshold of $\varepsilon = 0.05$, the overall classification accuracy exhibits

a negligible reduction—less than *0.1%*. These results highlight the ability of FROC to substantially improve fairness with minimal compromise in predictive performance.

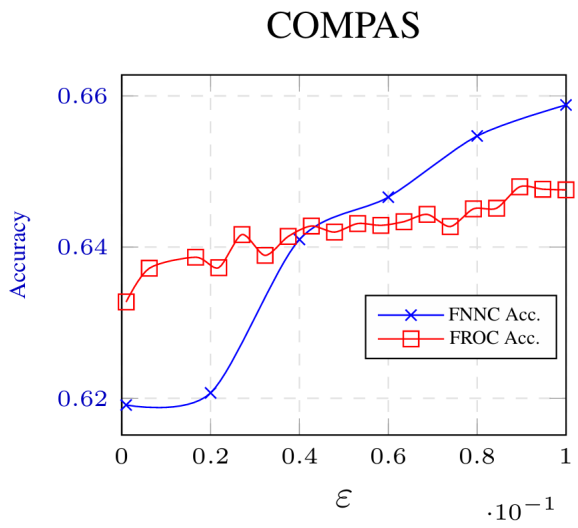


Figure 5.2 ROC comparison: FNNC vs. FNNC-FROC.

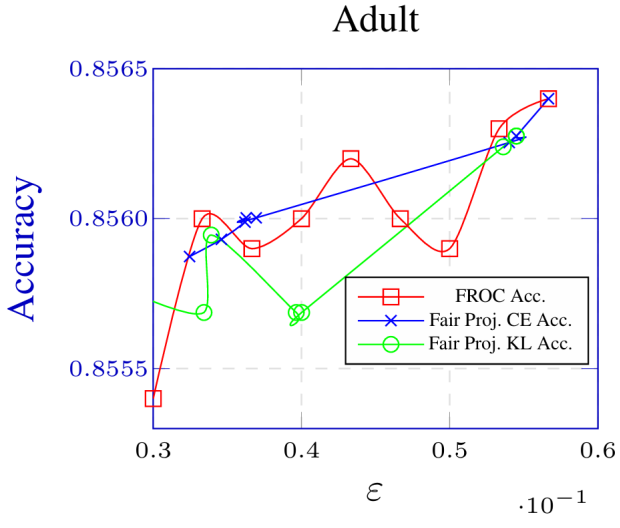


Figure 5.3 C3-FairProjection vs. C3-FROC.

Summary of Observations: Across all experimental settings, we consistently observe the following:

- *Fairness improvements of approximately 7–8%*, accompanied by a maximum reduction in classification accuracy of no more than *2%*.
- As illustrated in Figures 5.2 and 5.3, for small values of the fairness parameter ε, FROC consistently matches or surpasses the performance of FNNC and existing post-processing baselines in terms of fairness-accuracy trade-offs.
- The FNNC model and other prior approaches often *overcompensate for fairness*, resulting in unnecessarily large sacrifices in predictive performance (see Table 2 in Padala and Gujar [2020]).

These findings underscore the strength of FROC as a post-processing technique. By explicitly minimizing AUC loss while satisfying ε-Equalized Odds constraints, FROC provides a principled and efficient approach for achieving fair classification with minimal impact on accuracy.

5.3 Additional Experiments

In this section, we present supplementary experimental results that reinforce and extend the key findings reported in the main text. These additional analyses provide further evidence of the robustness and generalizability of our proposed method across various settings.

5.3.1 Adult Dataset - Weighted ensemble L2

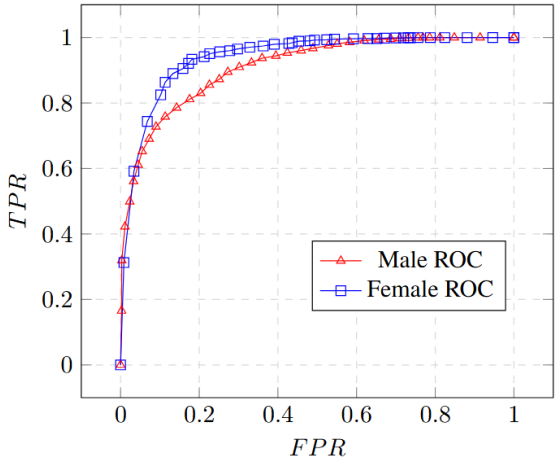


Figure 5.4 Weighted Ensemble L2 baseline ROCs for the Adult dataset.

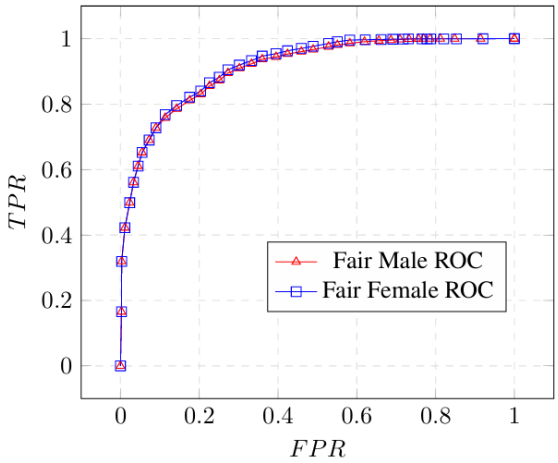


Figure 5.5 (Fair $\varepsilon_1 = 0.01$) Weighted Ensemble L2-FROCROCs for the Adult dataset.

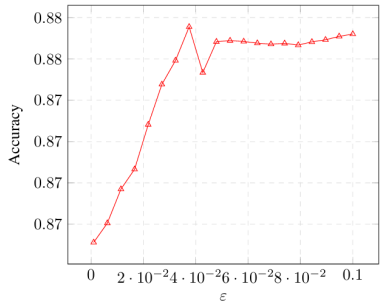


Figure 5.6 Accuracy vs. ε_1 for Weighted Ensemble L2-FROC (Adult)

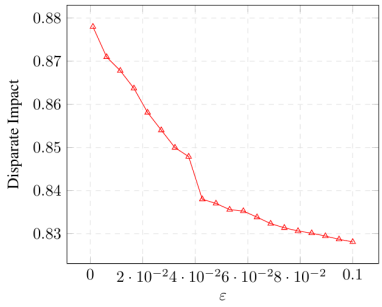


Figure 5.7 Disparate Impact vs. ε_1 for Weighted Ensemble L2-FROC (Adult)

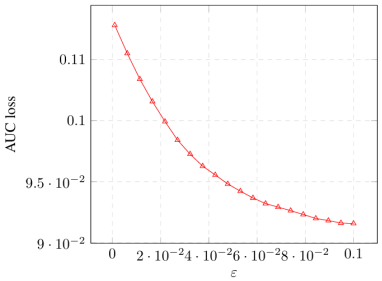


Figure 5.8 AUC loss vs. ε_1 for Weighted Ensemble L2-FROC (Adult)

- **Figure** 5.5 illustrates the application of FROC with a fairness parameter of $\varepsilon = 0.01$. As expected, the resulting ROC curves for different groups are significantly closer, indicating improved alignment across subpopulations.
- **Figures** 5.6 and 5.7 present plots of classification Accuracy and Disparate Impact, respectively, as functions of the fairness parameter ε_1.
- From this analysis, we observe a maximum variance of 1.88×10^{-6} and a maximum Coefficient of Variation (CoV) of 0.15% in Accuracy, indicating high stability in performance across fairness settings.

- Regarding Disparate Impact, the results yield a maximum variance of 2.25×10^{-5} and a CoV of 0.55%, suggesting moderate variability under changing fairness thresholds.
- Notably, the plots reveal that a marginal reduction of 1% in Accuracy corresponds to an improvement of approximately 5% in Disparate Impact, highlighting a favorable trade-off between fairness and performance.
- Finally, **Figure 5.8** displays the Area Under the Curve (AUC) loss as a function of ε. As shown, the AUC loss progressively approaches zero as the fairness constraint is relaxed, confirming that performance is preserved in the limit of unconstrained optimization.

5.3.2 Adult Dataset - Random Forest Gini

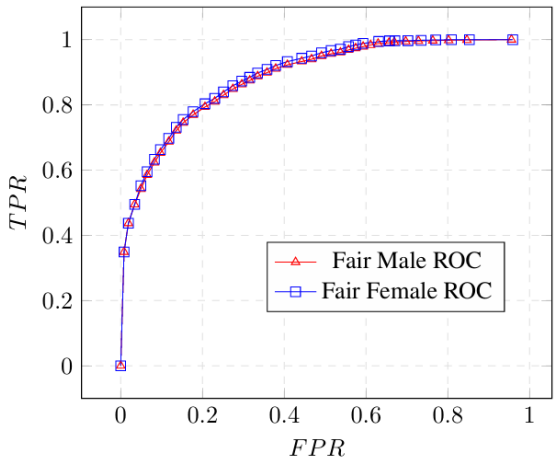


Figure 5.9 Random Forest (Gini) Baseline ROCs for Adult Dataset

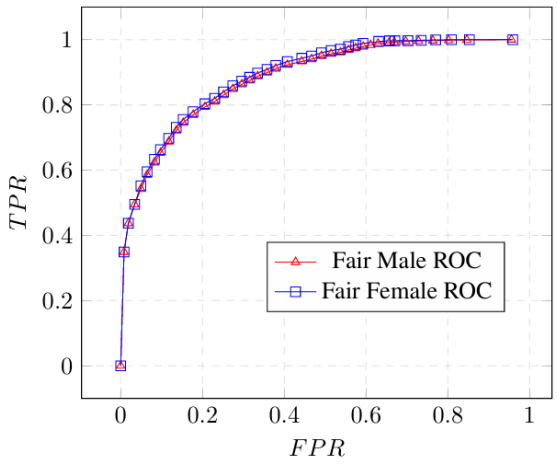


Figure 5.10 (Fair $\varepsilon_1 = 0.01$) Random Forest (Gini)-FROC ROCs for Adult Dataset

- **Figure 5.9** shows the ROC curves obtained by applying FROC with a fairness parameter of $\varepsilon = 0.01$. As anticipated, the post-processed ROC curves exhibit greater alignment, indicating enhanced fairness across protected groups.
- **Figures 5.11** and **5.12** present the relationships between ε_1 and classification Accuracy, as well as ε_1 and Disparate Impact, respectively.
- The analysis yields a maximum variance of 8.3×10^{-7} and a maximum Coefficient of Variation (CoV) of 0.1% for Accuracy, demonstrating exceptional stability in predictive performance across different fairness settings.
- For Disparate Impact, the observed maximum variance is 7.59×10^{-6}, with a corresponding CoV of 0.75%, reflecting a modest level of variability.

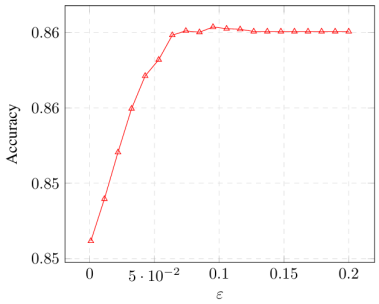


Figure 5.11 Accuracy vs. ε_1 for Random Forest (Gini)-FROC (Adult)

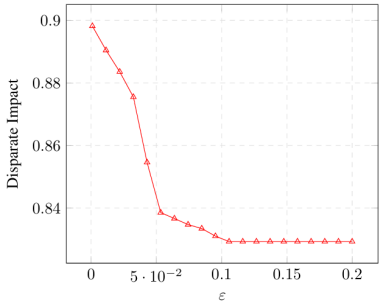


Figure 5.12 Disparate Impact vs. ε_1 for Random Forest (Gini)-FROC (Adult)

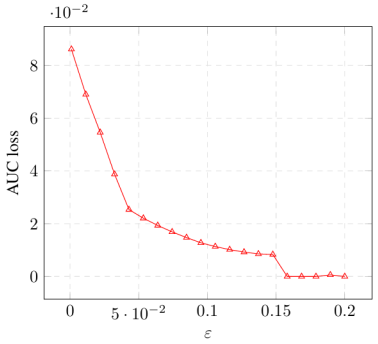


Figure 5.13 AUC loss vs. ε_1 for Random Forest (Gini)-FROC (Adult)

- Importantly, the plots reveal that a 1% decrease in Accuracy results in a 7% improvement in Disparate Impact, indicating a highly favorable trade-off between fairness and performance.

- **Figure 5.13** illustrates the Area Under the Curve (AUC) loss as a function of ε_1. The figure confirms that AUC loss converges toward zero as the fairness constraint is relaxed, suggesting minimal long-term cost in predictive quality.

5.3.3 Adult Dataset - FNNC

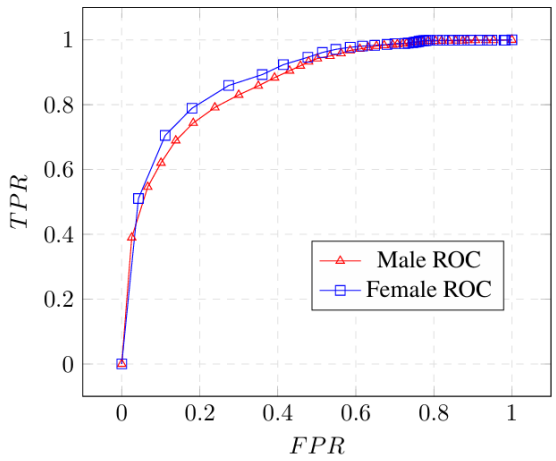


Figure 5.14 FNNC baseline ROCs for the Adult dataset.

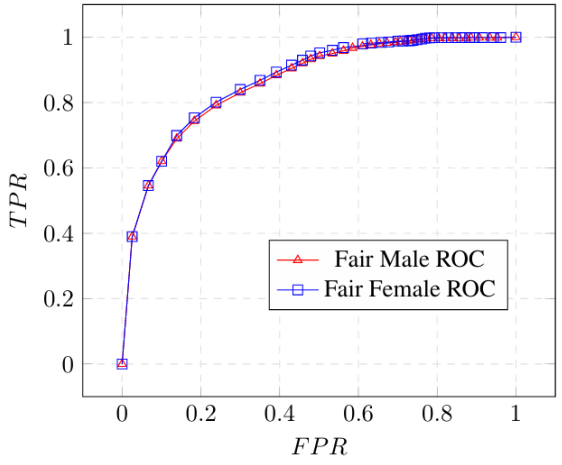


Figure 5.15 (Fair $\varepsilon_1 = 0.01$) FNNC-FROC ROCs for the Adult dataset.

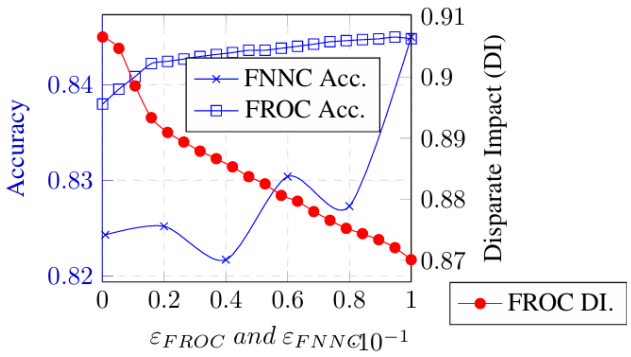


Figure 5.16 FNNC-FROC accuracy vs. ε_1 (Adult).

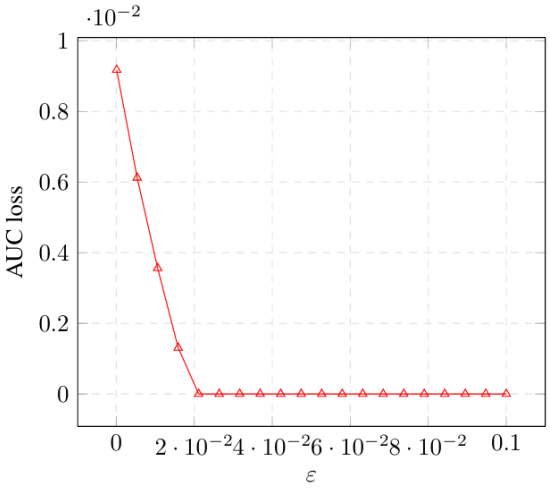


Figure 5.17 FNNC-FROC AUC loss vs. ε_1 (Adult).

- **Figure 5.24** presents the ROC curves obtained by applying FROC with a fairness parameter of $\varepsilon_1 = 0.01$. As expected, the post-processed curves exhibit greater alignment, reflecting improved parity between protected groups.

- **Figure 5.16** includes three plots: Accuracy vs. ε_1, Disparate Impact vs. ε_1, and a comparative plot of $\varepsilon_{\text{FNNC}}$ vs. $\varepsilon_{\text{FROC}}$.

- The results indicate that FNNC performs slightly worse than FROC in terms of accuracy. This performance gap is attributed to FNNC potentially overcorrecting for fairness at lower values of ε_1, a phenomenon also reported in Table 2 of Padala and Gujar [2020]. In contrast, FROC achieves the target fairness with only minimal AUC degradation.

- The analysis shows a maximum variance of 6.6×10^{-7} and a maximum Coefficient of Variation (CoV) of 0.09% for Accuracy, indicating high robustness in predictive performance.

- For Disparate Impact, the maximum variance observed is 1×10^{-4}, with a corresponding CoV of 1.26%, suggesting slightly greater sensitivity to fairness parameter changes.

- As illustrated in the plots, a modest reduction of 1% in Accuracy corresponds to an improvement of approximately 5% in Disparate Impact, demonstrating a beneficial trade-off.

- **Figure 5.17** shows the AUC loss as a function of ε_1. The loss asymptotically approaches zero as the fairness constraint is relaxed, confirming the efficiency of FROC in preserving classification performance.

5.3.4 COMPAS Dataset - Weighted ensemble L2

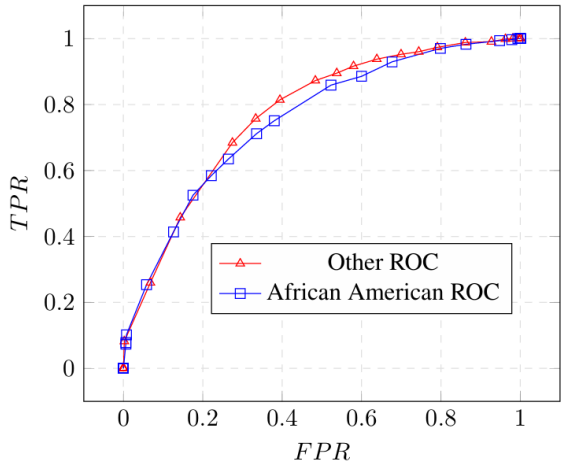


Figure 5.18 Weighted Ensemble L2 baseline ROCs for COMPAS dataset.

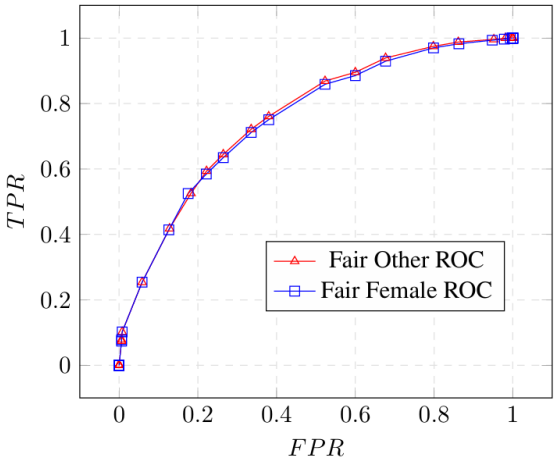


Figure 5.19 (Fair $\varepsilon_1 = 0.01$) Weighted Ensemble L2-FROC ROCs for COMPAS dataset.

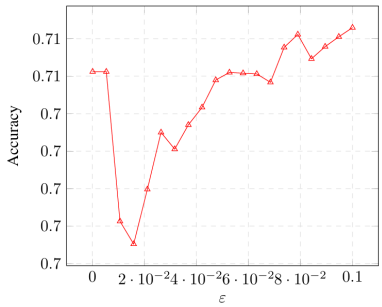


Figure 5.20 Accuracy vs. ε_1 for Weighted Ensemble L2-FROC (COMPAS).

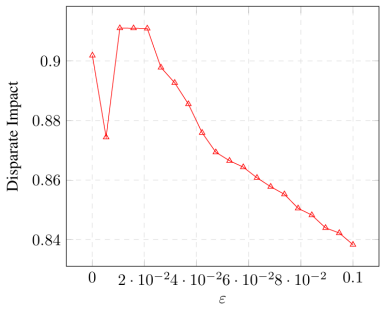


Figure 5.21 Disparate Impact vs. ε_1 for Weighted Ensemble L2-FROC (COMPAS).

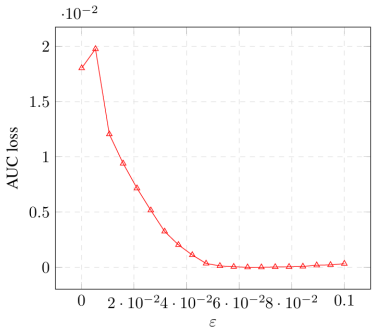


Figure 5.22 AUC loss vs. ε_1 for Weighted Ensemble L2-FROC (COMPAS).

- **Figure 5.19** displays the ROC curves after applying FROC with a fairness parameter of $\varepsilon_1 = 0.01$. As expected, the post-processed curves show improved alignment, demonstrating reduced disparity between groups.
- **Figures 5.20** and **5.21** present the relationship between ε_1 and two key metrics: classification Accuracy and Disparate Impact.
- The analysis reveals a maximum variance of 1.44×10^{-5} and a maximum Coefficient of Variation (CoV) of 0.54% for Accuracy, indicating stable performance across varying levels of the fairness constraint.

- For Disparate Impact, the analysis yields a maximum variance of 1.6×10^{-4} and a maximum CoV of 1.69%, reflecting moderate variability.
- The plots further demonstrate that a 1% reduction in Accuracy corresponds to a 7% improvement in Disparate Impact, indicating a favorable trade-off between fairness and predictive performance.
- **Figure 5.22** illustrates the AUC loss as a function of ε_1. As shown, the AUC loss approaches zero as the fairness constraint is relaxed, highlighting the efficiency of FROC in preserving model performance.

5.3.5 COMPAS Dataset - Random Forest Gini

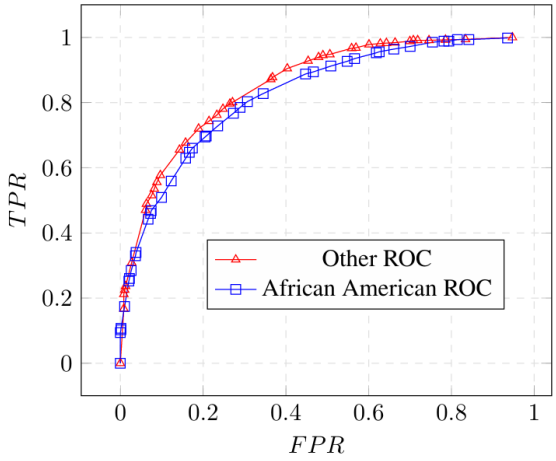


Figure 5.23 Random Forest (Gini) baseline ROCs for the COMPAS dataset.

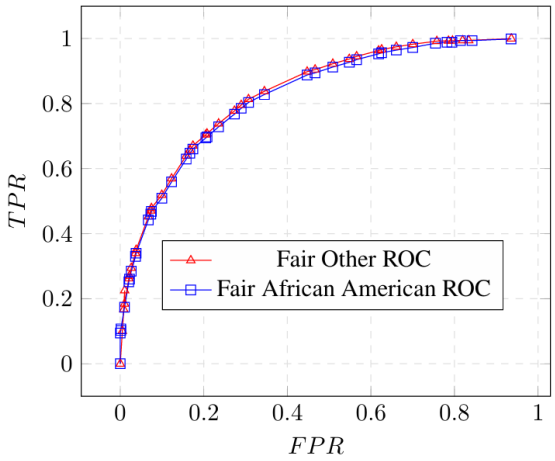


Figure 5.24 (Fair $\varepsilon_1 = 0.01$) Random Forest (Gini)-FROC ROCs for the COMPAS dataset.

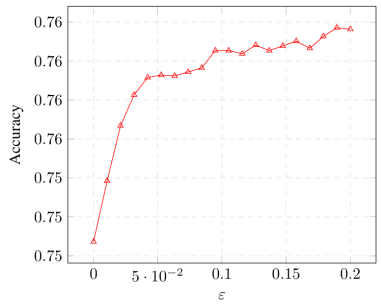


Figure 5.25 Accuracy vs. ε_1 for Random Forest (Gini)-FROC (COMPAS).

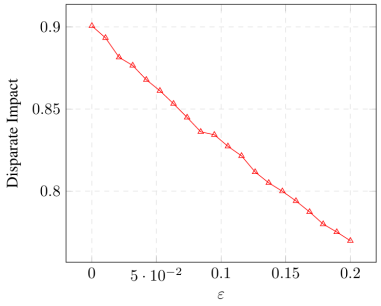


Figure 5.26 Disparate Impact vs. ε_1 for Random Forest (Gini)-FROC (COMPAS).

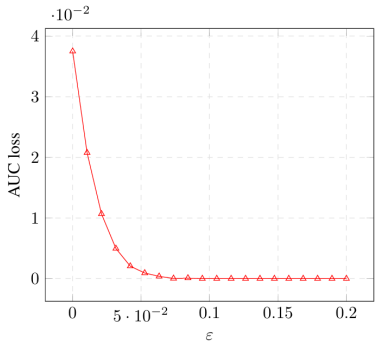


Figure 5.27 AUC loss vs. ε_1 for Random Forest (Gini)-FROC (COMPAS).

- **Figure 5.24** shows the ROC curves resulting from the application of FROC with a fairness parameter of $\varepsilon = 0.01$. As expected, the curves are more closely aligned, indicating improved fairness between the protected groups.
- **Figures 5.25** and **5.26** depict the relationships between ε_1 and two evaluation metrics: classification Accuracy and Disparate Impact, respectively.
- The results demonstrate a maximum variance of 9.63×10^{-6} and a maximum Coefficient of Variation (CoV) of 0.44% for Accuracy, indicating high consistency in predictive performance across varying fairness thresholds.
- For Disparate Impact, the analysis yields a maximum variance of 2×10^{-4} and a maximum CoV of 1.56%, reflecting moderate variability in fairness outcomes.
- As illustrated in the plots, a modest 1% decrease in Accuracy leads to an approximate 7% improvement in Disparate Impact, demonstrating an effective trade-off between predictive performance and fairness.
- **Figure 5.27** presents the AUC loss as a function of ε_1. The results indicate that the AUC loss approaches zero as the fairness constraint is relaxed, confirming that FROC achieves fairness improvements with minimal degradation in model performance.

5.3.6 COMPAS Dataset - FNNC

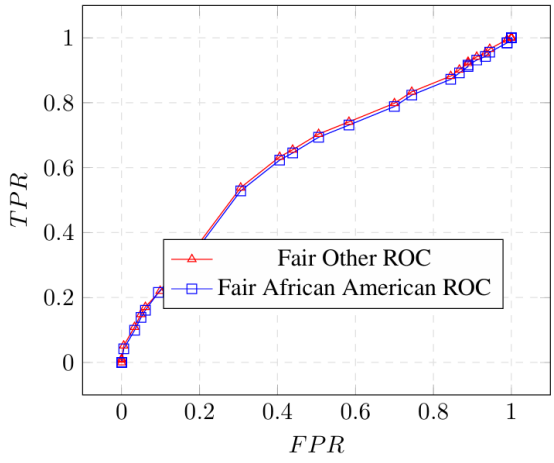


Figure 5.28 FNNC baseline ROCs for the COMPAS dataset.

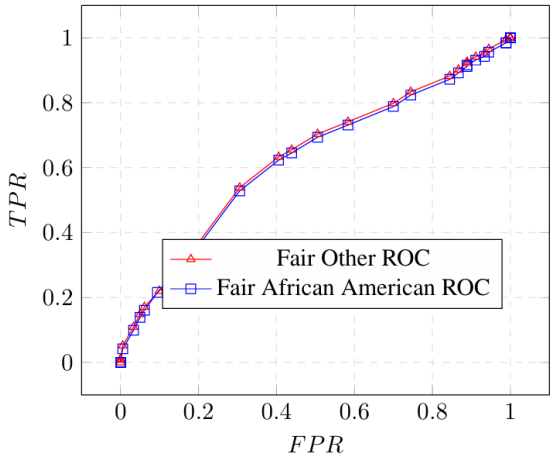


Figure 5.29 (Fair $\varepsilon_1 = 0.01$) FNNC-FROC ROCs for the COMPAS dataset.

- **Figure 5.29** illustrates the ROC curves obtained after applying FROC with a fairness parameter of $\varepsilon_1 = 0.01$. As expected, the post-processed ROC curves are significantly closer, indicating reduced disparity between protected groups.

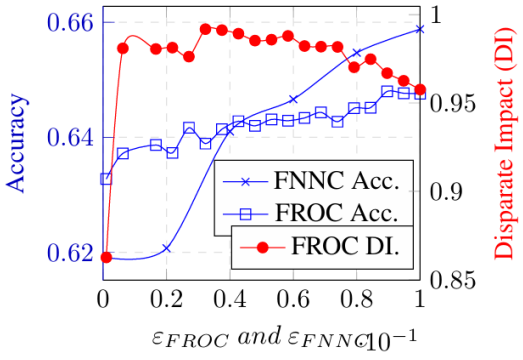


Figure 5.30 FNNC-FROC accuracy vs. ε_1 (COMPAS).

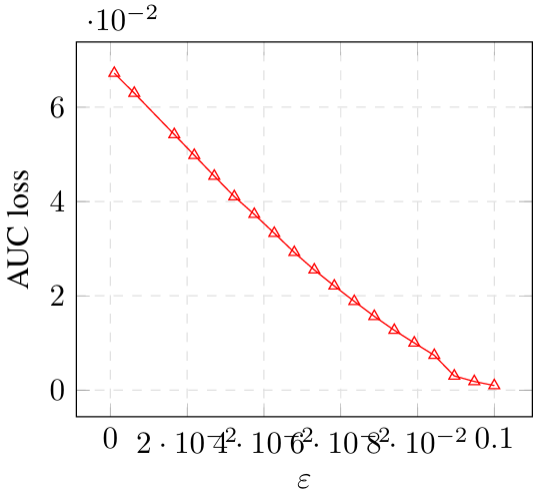


Figure 5.31 FNNC-FROC AUC loss vs. ε_1 (COMPAS).

- **Figure 5.30** presents two plots: Accuracy vs. ε_1 and Disparate Impact vs. ε_1.
- The results indicate that FNNC yields slightly lower accuracy compared to FROC. This performance gap is likely due to FNNC overcorrecting for fairness at small values of $\varepsilon_{\text{FNNC}}$, a behavior previously noted in Table 2 of Padala and Gujar [2020]. In contrast, FROC achieves target fairness with minimal reduction in AUC.
- The analysis reports a maximum variance of 4.83×10^{-6} and a maximum Coefficient of Variation (CoV) of 0.43% for Accuracy, demonstrating stable model performance across fairness parameter values.
- For Disparate Impact, the maximum variance is 2.48×10^{-5} with a CoV of 0.5%, indicating low variability in fairness outcomes.
- The plots reveal that a 1% decrease in Accuracy leads to a 3% improvement in Disparate Impact, reflecting a modest yet positive fairness-performance trade-off.
- **Figure 5.31** shows the AUC loss as a function of ε_1. As expected, the loss diminishes and approaches zero as the fairness constraint is relaxed, affirming the efficiency of FROC in maintaining model quality.

5.3.7 CelebA Dataset

- **Figure 5.33** presents the ROC curves following the application of FROC with a fairness parameter of $\varepsilon_1 = 0.01$. As expected, the resulting ROC curves are more closely aligned, indicating enhanced fairness between groups.

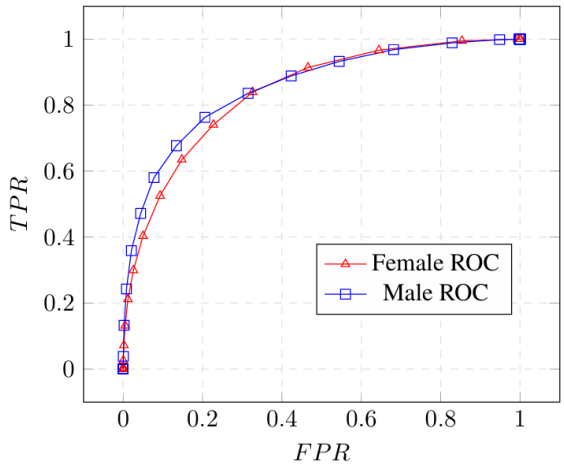


Figure 5.32 ResNet baseline ROCs for the CelebA dataset.

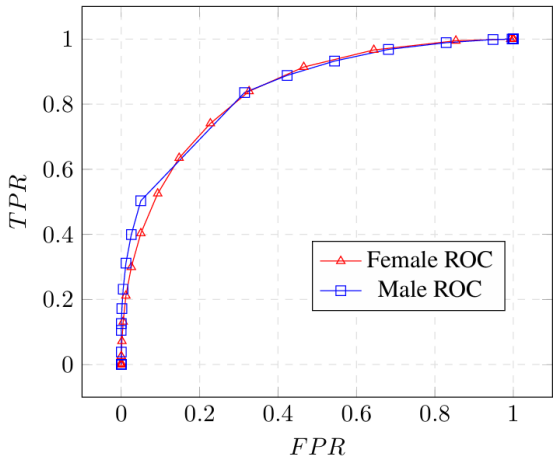


Figure 5.33 (Fair $\varepsilon_1 = 0.01$) ResNet-FROC ROCs for the CelebA dataset.

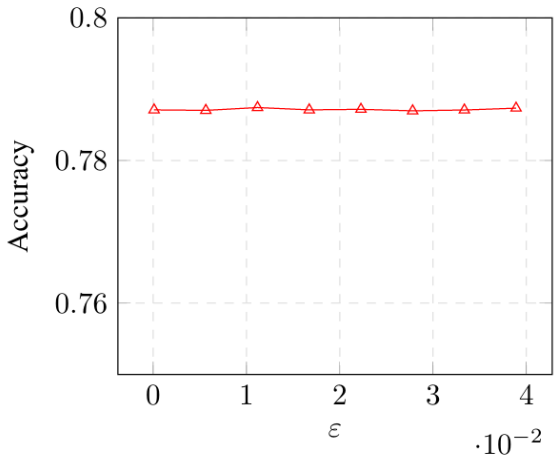


Figure 5.34 ResNEt-FROC Accuracy vs. ε_1 (CelebA)

- The analysis of classification Accuracy, shown in **Figure 5.34**, reveals a maximum variance of 1.9×10^{-7} and a maximum Coefficient of Variation (CoV) of 0.07%, reflecting exceptionally stable performance across varying fairness levels.

- Regarding Disparate Impact, we observe minimal change, as the initial ROC curves were already closely aligned. Consequently, the fairness improvement achieved by FROC is limited in this setting.

- A similar trend is observed for AUC, which remains relatively unchanged with no consistent pattern emerging across different values of ε_1.

Summary

Our empirical analysis demonstrates that `FROC` achieves fairness with minimal loss in AUC, outperforming or on par with traditional fair ML methods in preserving accuracy while meeting fairness constraints.

Chapter 6

Conclusion

> The supreme gift of the artist is the knowledge of when to stop.
>
> - Sherlock Holmes

> Every new beginning comes from some other beginning's end.
>
> - Seneca

This thesis investigated the challenge of ensuring fairness in probabilistic binary classification under the presence of binary protected groups. Specifically, we focused on fairness in the ROC space—a domain particularly relevant when classification decisions are threshold-dependent and made post hoc by practitioners aiming to balance false positives and false negatives.

We introduced a new fairness criterion, ε_p-Equalized ROC, which requires that the $\mathcal{L}_p$ norm between the false positive rates (FPRs) and true positive rates (TPRs) of protected groups remain within a specified bound ε across all possible decision thresholds. This threshold-independent view of fairness strengthens existing definitions by ensuring robustness to downstream decision-making procedures.

To address the case where a classifier does not satisfy ε_1-Equalized ROC, we proposed a novel post-processing approach that transforms the classifier's outputs into a randomized yet fair predictor. This transformation is governed by a query model operating over ROC curves, which introduces necessary modifications to the classifier's group-conditional behavior.

We analyzed the theoretical limitations of this approach, demonstrating that any post-processing method satisfying ε_1-Equalized ROCwill incur some degree of performance degradation, typically in the form of AUC loss. To quantify this, we developed a formal characterization of the minimal AUC loss required to achieve fairness. Building upon this foundation, we designed a linear-time algorithm, `FROC`, and proved that it satisfies ε_1-Equalized ROCwhile achieving theoretical optimality under certain assumptions.

Our empirical analysis, conducted across multiple benchmark datasets (Adult, COMPAS, CelebA) and classifiers (Weighted Ensemble L2, Random Forest with Gini index, FNNC), demonstrated that `FROC` consistently achieves fairness with minimal loss in predictive utility. These results support the practicality of our framework and its applicability in real-world fairness-sensitive decision systems.

Future Work

While this thesis provides a principled foundation for fair post-processing via ROC-based transformations, several avenues remain open for future exploration:

- **Extension to Multiclass and Multigroup Settings:** Our current formulation assumes binary classification with binary protected groups. Extending the framework to multiclass settings or multiple protected attributes (and their intersections) is a promising direction.
- **Adaptive Thresholding with User Constraints:** Incorporating user-specified constraints—such as operating points or cost-sensitive thresholds—into the fairness formulation may improve alignment with practical decision-making contexts.
- **Integration with In-processing Techniques:** While we focus on post-processing, combining this framework with in-processing algorithms may yield stronger overall guarantees and less utility degradation.
- **Robustness and Uncertainty Quantification:** Understanding how uncertainty in group membership or score estimates affects fairness and performance guarantees remains an important and underexplored area.

In conclusion, this thesis contributes both theoretical and practical insights into the design of fair classifiers that remain robust under threshold variation. The proposed methods enable the development of equitable machine learning systems without requiring retraining or full access to model internals—thus offering a versatile tool for practitioners and researchers alike.

Bibliography

Eleni Adamopoulou and Lefteris Moussiades. Chatbots: History, technology, and applications. *Machine Learning with Applications*, 2:100006, 2020.

Meta AI. Llama 2: Open foundation and fine-tuned chat models, 2023. https://ai.meta.com/llama/.

Wael Alghamdi, Hsiang Hsu, Haewon Jeong, Hao Wang, Peter Michalak, Shahab Asoodeh, and Flavio Calmon. Beyond adult and compas: Fair multi-class prediction via information projection. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 38747–38760. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/fd5013ea0c3f96931dec77174eaf9d80-Paper-Conference.pdf.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*, 2016. URL https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. In *Ethics of data and analytics*, pages 254–264. Auerbach Publications, 2022.

Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Calif. L. Rev.*, 104:671, 2016.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT press, 2023.

Barry Becker and Ronny Kohavi. Adult. UCI Machine Learning Repository, 1996. DOI: https://doi.org/10.24432/C5XW20.

Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H Chi, et al. Fairness in recommendation ranking through pairwise comparisons. In *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 2212–2220, 2019.

Daniel Borkan, Lucas Dixon, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Nuanced metrics for measuring unintended bias with real data for text classification. In *Companion proceedings of the 2019 world wide web conference*, pages 491–500, 2019.

Tom B Brown, Benjamin Mann, Nick Ryder, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Niklas Bussmann, Paolo Giudici, Daniele Marinelli, Jochen Papenbrock, Christoph Ruse, and Alex Sasson. Explainable ai in fintech risk management: A case study of lloyds banking group. *Frontiers in Artificial Intelligence*, 4:26, 2021.

Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

Mingliang Chen and Min Wu. Towards threshold invariant fair classification. In *Conference on Uncertainty in Artificial Intelligence*, pages 560–569. PMLR, 2020.

Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Stéphan Clémençon, Gábor Lugosi, and Nicolas Vayatis. Ranking and empirical minimization of u-statistics, 2006. URL https://arxiv.org/abs/math/0603123.

Corinna Cortes and Mehryar Mohri. Auc optimization vs. error rate minimization. *Advances in neural information processing systems*, 16, 2003.

André F Cruz and Moritz Hardt. Unprocessing seven years of algorithmic fairness. *arXiv preprint arXiv:2306.07261*, 2023.

Alexandru Ţifrea, Preethi Lahoti, Ben Packer, Yoni Halpern, Ahmad Beirami, and Flavien Prost. FrappÉ: a group fairness framework for post-processing everything. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Sen Cui, Weishen Pan, Changshui Zhang, and Fei Wang. Towards model-agnostic post-hoc adjustment for balancing ranking fairness and algorithm utility. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 207–217, 2021.

DeepSeek. Deepseek v2: Scaling sparse mixture of experts to 236b parameters, 2024. URL https://www.deepseek.com/blog/deepseek-v2-release-note. Accessed April 2025.

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*, pages 214–226, 2012.

Andre Esteva, Alexandre Robicquet, Bharath Ramsundar, Volodymyr Kuleshov, Mark DePristo, Katherine Chou, Claire Cui, Greg Corrado, Sebastian Thrun, and Jeff Dean. A guide to deep learning in healthcare. *Nature Medicine*, 25(1):24–29, 2019.

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*, pages 259–268, 2015.

Sruthi Gorantla, Amit Deshpande, and Anand Louis. On the problem of underranking in group-fair ranking. In *International Conference on Machine Learning*, pages 3777–3787. PMLR, 2021.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, volume 29, 2016.

Jin Huang and Charles X Ling. Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, 17(3):299–310, 2005.

Taeuk Jang, Pengyi Shi, and Xiaoqian Wang. Group-aware threshold adaptation for fair classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6988–6995, 2022.

Michael I Jordan and Tom M Mitchell. Machine learning: Trends, perspectives, and prospects. *Science*, 349(6245):255–260, 2015.

Nathan Kallus and Angela Zhou. The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric. *Advances in neural information processing systems*, 32, 2019.

Andreas Kamilaris and Francesc Xavier Prenafeta-Boldú. Deep learning in agriculture: A survey. *Computers and Electronics in Agriculture*, 147:70–90, 2018.

Faisal Kamiran and Toon Calders. Classifying without discriminating. In *2nd International Conference on Computer, Control and Communication*, 2009.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*, pages 924–929. IEEE, 2012.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Takuya Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases*, pages 35–50, 2012.

Keith Kendig. Is a 2000-year-old formula still keeping some secrets? *The American Mathematical Monthly*, 107(5):402–415, 2000.

Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In Christos H. Papadimitriou, editor, *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*, volume 67 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 43:1–43:23. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, 2017. ISBN 978-3-95977-029-3. doi: 10.4230/LIPIcs.ITCS.2017.43. URL http://drops.dagstuhl.de/opus/volltexte/2017/8156.

Matt J. Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

Sharon Li. The true cost of training large ai models, 2023. https://spectrum.ieee.org/ai-training-costs.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pages 3384–3393. PMLR, 2018.

Mastercard. Ai and machine learning help stop fraud before it happens, 2021. URL https://www.mastercard.com/news/press/2021/ai-and-machine-learning-help-stop-fraud-before-it-happens/.

Scott M. McKinney, Marcin Sieniek, Vishal Godbole, et al. International evaluation of an ai system for breast cancer screening. *Nature*, 577(7788):89–94, 2020. URL https://www.nature.com/articles/s41586-019-1799-6.

Ninareh Mehrabi, Fred Morstatter, Nitesh Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

Alan Mishler, Edward H Kennedy, and Alexandra Chouldechova. Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 386–400, 2021.

Brent Daniel Mittelstadt, Patrick Allo, Mariarosaria Taddeo, Sandra Wachter, and Luciano Floridi. The ethics of algorithms: Mapping the debate. *Big Data & Society*, 3(2):1–21, 2016.

Preetam Nandy, Cyrus DiCiccio, Divya Venugopalan, Heloise Logan, Kinjal Basu, and Noureddine El Karoui. Achieving fairness via post-processing in web-scale recommender systems. *2022 ACM Conference on Fairness, Accountability, and Transparency*, Jun 2022. doi: 10.1145/3531146.3533136. URL http://dx.doi.org/10.1145/3531146.3533136.

Manisha Padala and Sujit Gujar. Fnnc: Achieving fairness through neural networks. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence,{IJCAI-20}, International Joint Conferences on Artificial Intelligence Organization*, 2020.

David Patterson, Joseph Gonzalez, Quoc Le, et al. Carbon emissions and large neural networks, 2021. https://arxiv.org/abs/2104.10350.

Foster Provost. Machine learning from imbalanced data sets 101. In *Proceedings of the AAAI'2000 workshop on imbalanced data sets*, volume 68, pages 1–3. AAAI Press, 2000.

Inioluwa Deborah Raji, Andrew Smart, Rebecca White, Margaret Mitchell, Timnit Gebru, and Ben Hutchinson. Closing the ai accountability gap: Defining an end-to-end framework for internal algorithmic auditing. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 33–44, 2020.

D Sleeman, Michalis Rissakis, Susan Craw, Nicolas Graner, and Sunil Sharma. Consultant-2: Pre-and post-processing of machine learning applications. *International journal of human-computer studies*, 43(1):43–63, 1995.

Blue River Technology. See & spray technology, 2023. URL https://www.bluerivertechnology.com/.

Tesla. Tesla vehicle safety report q4 2023, 2023. URL https://www.tesla.com/VehicleSafetyReport.

Sahil Verma and Julia Rubin. Fairness definitions explained. In *Proceedings of the international workshop on software fairness*, pages 1–7, 2018.

Robin Vogel, Aurélien Bellet, and Stephan Clémençon. Learning fair scoring functions: Bipartite ranking under roc-based fairness constraints. In *International conference on artificial intelligence and statistics*, pages 784–792. PMLR, 2021.

Dennis Wei, Karthikeyan Natesan Ramamurthy, and Flavio P Calmon. Optimized score transformation for fair classification. *Proceedings of Machine Learning Research*, 108, 2020.

Liyuan Xu, Tianlu Yuan, Xiaokang Wu, and Yuexian Wu. Fairgan: Fairness-aware generative adversarial networks. In *IEEE Big Data*, pages 570–575, 2018.

Zhenhuan Yang, Yan Lok Ko, Kush R. Varshney, and Yiming Ying. Minimax auc fairness: efficient algorithm with provable convergence. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press, 2023. ISBN 978-1-57735-880-0. doi: 10.1609/aaai.v37i10.26405. URL https://doi.org/10.1609/aaai.v37i10.26405.

Mehdi Yazdani-Jahromi, Ali Khodabandeh Yalabadi, AmirArsalan Rajabi, Aida Tayebi, Ivan Garibay, and Ozlem Garibay. Fair bilevel neural network (fairbinn): On balancing fairness and accuracy via

stackelberg equilibrium. *Advances in Neural Information Processing Systems*, 37:105780–105818, 2024.

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*, pages 1171–1180, 2017.

Meike Zehlike, Ke Yang, and Julia Stoyanovich. Fairness in ranking: A survey. *arXiv preprint arXiv:2103.14000*, 2021.

Richard Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International Conference on Machine Learning*, pages 325–333, 2013.

Brian Hu Zhang, Benjamin Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340, 2018.

Han Zhao. Fair and optimal prediction via post-processing. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(20):22686–22686, Mar. 2024. doi: 10.1609/aaai.v38i20.30302. URL https://ojs.aaai.org/index.php/AAAI/article/view/30302.

Zhi-Hua Zhou and Xu-Ying Liu. Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Transactions on knowledge and data engineering*, 18(1):63–77, 2005.