Distributionally-Robust and Specification-Robust Fairness for Machine Learning

by

Omid Memarrast
B.Sc., University of Tehran, Tehran, Iran, 2012
M.Sc., University of Illinois Chicago, Chicago, USA, 2021

THESIS

Submitted in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in Computer Science
in the Graduate College of the
University of Illinois Chicago, 2023

Chicago, Illinois

Defense Committee:
Prof. Brian Ziebart, Chair and Advisor
Department of Computer Science, University of Illinois Chicago
Prof. Xinhua Zhang
Department of Computer Science, University of Illinois Chicago
Prof. Ian Kash
Department of Computer Science, University of Illinois Chicago
Prof. Abolfazl Asudeh
Department of Computer Science, University of Illinois Chicago
Prof. Anqi Liu
Department of Computer Science, Johns Hopkins University

Dedicated to my mother, an inspiration for me in her battle against Multiple Sclerosis.

ACKNOWLEDGMENT

First and foremost, I would like to express my gratitude to my advisor, Prof. Brian Ziebart, for his invaluable mentorship and unwavering assistance during my Ph.D. journey. Brian has been more than an advisor to me; he has been a friend who has guided and inspired me with his wisdom, patience, and kindness. He introduced me to the field of Fair Machine Learning and encouraged me to pursue this topic for my dissertation. His remarkable expertise, encouragement, and patience have played a crucial role in my academic and personal development, and have been instrumental to my success. I feel incredibly fortunate to have had the opportunity to work under his supervision and to learn from him.

I would like to thank Prof. Xinhua Zhang for his kind invaluable support and advice. I would like to also thank the rest of my thesis committee members: Prof. Ian Kash, Prof. Abolfazl Asudeh, and Prof. Anqi Liu, for providing their valuable insights on my research.

I want to thank my collaborators: Ashkan Rezaei, Rizal Fathony, Anqi Liu, and Linh Vu for their contributions to get our works published. Furthermore, I thank my fellow lab mates: Mohammad Ali Bashiri, Yeshu Li, Sanket Gaurav, Danyal Saeed, Rushit Shah, Jurat Shayiding, and Nikolaos Agadakos for the encouragement and feedback they provided. And thank you to

ACKNOWLEDGMENT (Continued)

all the friends I have made at UIC, and the beautiful city of Chicago and elsewhere, for all the happiness you've brought me. Thanks to Elnaz, for sharing many happy and sad moments with me, and for helping me become a better version of myself.

Last, but never least, I am thankful to my parents, and my brothers, Reza and Mahdi, for their constant support and love in my life. I wouldn't have been able to go this far without you.

OM

CONTRIBUTIONS OF AUTHORS

In Chapter 3, two published manuscripts (Rezaei et al., 2020; Rezaei et al., 2021) are presented. In (Rezaei et al., 2020), Ashkan Rezaei and Rizal Fathony were the primary authors. Ashkan Rezaei formulated the fair and adversarial robust log loss classification and derived its parametric form. He also designed and implemented the optimization algorithm and conducted most of the experiments under the guidance of our advisor, Prof. Brian Ziebart. Rizal Fathony contributed to the proof of asymptotic convergence properties and the discussion of the algorithm for finding fair thresholds. I contributed to the discussion of the formulation, prepared the datasets for experiments, and ran baselines.

In (Rezaei et al., 2021), Ashkan Rezaei was the primary author. He formulated and derived the fair robust log loss under covariate shift, proved the theorems, designed and implemented the optimization algorithm, and conducted most of the experiments under the supervision of our advisor, Prof. Brian Ziebart. Anqi Liu contributed to the discussion of the formulation of fairness under covariate shift and generating biased samplings for the experiments. I contributed to the discussion of the formulation, prepared the datasets for experiments, and ran baselines.

CONTRIBUTIONS OF AUTHORS (Continued)

In Chapter 4, a published manuscript (Memarrast et al., 2023a) is presented, in which I was the primary author. I formulated and derived fair and adversarial robust learning to rank, designed and implemented the optimization algorithm, and conducted all the experiments under the supervision of my advisor, Prof. Brian Ziebart. Ashkan Rezaei contributed to the formulation of fair and adversarial robust learning to rank. Rizal Fathony contributed to the discussion on the optimization of the algorithm.

Chapter 5 presents another published manuscript (Memarrast et al., 2023b), in which I was the primary author. I formulated and derived a superhuman fairness-aware classifier, designed and implemented the optimization algorithm, and conducted the experiments under the supervision of my advisor, Prof. Brian Ziebart. Linh Vu contributed to the implementation of the algorithm, dataset preparation, and running some baselines.

TABLE OF CONTENTS

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

TABLE OF CONTENTS (Continued)

CHAPTER **PAGE**

LIST OF TABLES

LIST OF FIGURES

LIST OF FIGURES (Continued)

FIGURE **PAGE**

SUMMARY

Developing machine learning methods with high accuracy that also avoid unfair treatment of different groups has become increasingly important for data-driven decision making in social applications. This thesis presents the development of fair machine learning algorithms through two main paradigms. Firstly, we leverage the distributionally robust framework to develop fair algorithms for classification and ranking. Specifically, we propose novel approaches that minimize disparities across different groups while preserving overall classification or ranking performance. Secondly, we introduce a specification-robust approach to fair classification, inspired by the ideas from imitation learning.

Supervised learning often relies on Empirical Risk Minimization (ERM) (Vapnik, 1992) to train models that can generalize well to unseen data. However, ERM methods are susceptible to noise and outliers due to their dependence on sample means. To address this issue, adversarial robust learning (Asif et al., 2015) has emerged as a promising approach, formulating supervised learning as a minimax game between a predictor and an adversary. This thesis investigates how this framework can be applied to fair machine learning tasks and propose two approaches for

fair classification: fair and robust log loss classification (Rezaei et al., 2020) and fair and robust log loss classification under covariate shift (Rezaei et al., 2021).

Furthermore, this thesis leverages adversarial robust learning in developing fair and robust models for structured prediction problems, where the goal is to predict a ranking of items or outputs rather than a single label. It introduces a fair and robust learning-to-rank approach (Memarrast et al., 2023a) that achieves fairness of exposure for protected groups (such as race or gender) while maximizing utility to the users. Overall, this thesis explores the potential of adversarial robust learning for addressing fairness in classification and structured prediction tasks and provides new approaches to building fair and robust machine learning models.

The second part of this thesis discusses the *specification-robust* approach to fairness in machine learning. While most fairness approaches optimize a specified trade-off between performance measures (such as accuracy, log loss, or AUC) and fairness metrics (such as demographic parity or equalized odds), this begs the question of whether the right trade-offs are being specified. To address this issue, this thesis proposes a new approach called *superhuman fairness* (Memarrast et al., 2023b) which recasts fair machine learning as an imitation learning task. Superhuman fairness seeks to simultaneously outperform human decisions on multiple predictive performance and fairness measures, rather than relying on a pre-specified trade-off. The thesis demonstrates the benefits of this approach given suboptimal decisions, showing that it can improve both performance and fairness outcomes.

Finally, we outline further directions of our ongoing and future research.

CHAPTER 1

Introduction

Machine learning algorithms aim at optimizing performance metrics that measure success according to classification accuracy, precision, F-measure, etc. Unfortunately, there are other priorities, concerns, or ethical assumptions that these models ignore in practice, and can be totally violated as a result. Over the last few years, there has been a surge in interest in the fairness and privacy properties of machine learning algorithms, with important implications for practitioners. Social aspects of the contexts in which machine learning algorithms are deployed tend to influence priorities, and fairness properties are a prime example. There are certain demographic groups to which we prefer to extend equal treatment. In a similar vein, we would like some guarantee that individuals with similar characteristics on some dimensions get the same treatment regardless of differences in other characteristics. In certain decision-making applications, this may be preferable or even required by law e.g., admissions decisions

for universities (Lowry and Macpherson, 1988; Chang, 2006; Kabakchieva, 2013), decisions regarding employment and promotions (Lohr, 2013), medical decisions for insurers and hospitals (Shipp et al., 2002; Obermeyer and Emanuel, 2016), sentencing guidelines for civil and criminal cases within the justice system (Moses and Chan, 2014; O'Neil, 2016), the financial industry's loan decision-making process (Shaw and Gentry, 1988; Carter and Catlett, 1987; Bose and Mahapatra, 2001), as well as in a variety of other applications.

The design of machine learning systems now emphasizes equitable behavior, despite existing disparities in the data. Gender imbalances and other disparities may stem from historical discrimination or inherent differences in characteristics. Even non-personal data can be influenced by demographic disparities like income or age when collected by individuals (e.g., smartphones).

Simply omitting protected attributes (race, gender, etc.) is ineffective as they are implicitly represented by proxy variables (e.g., zip code, personality features). Removing protected attributes not only fails to prevent discrimination but also complicates bias mitigation through fair learning methods. These challenges have sparked the emergence of fair machine learning, which addresses these concerns from various angles and approaches.

This thesis develops new methods for constructing predictors with fairness and performance guarantees that are robust to noise in the label distribution and desired fairness-performance trade-offs.

1.1 Fairness Criteria

In general, fairness is defined in two categories: group fairness criteria partition the population into groups, and they use defined statistical measures to ensure that members across groups are

treated equally. Individual fairness (Dwork et al., 2012) on the other hand, is the notion that all individuals should be treated equally regardless of affiliation to a group. There has been extensive research and discussion regarding acceptable definitions of fairness and their accuracy and effect in practice. As the main surveys on algorithms and measures for fairness we refer to (Binns, 2018; Hutchinson and Mitchell, 2019; Verma and Rubin, 2018; Saxena et al., 2019; Mehrabi et al., 2019). In this thesis, we focus on group fairness and develop algorithms that treat members across protected groups e.g. race or gender equally. As group fairness constraints, we mainly focus on *demographic parity* (Calders et al., 2009), *equalized odds*, *equalized opportunity* (Hardt et al., 2016) and *predictive rate parity* (Chouldechova, 2017a). In the next chapter, we will provide formal definitions of fairness measures and illustrate each measure's satisfaction through relevant scenarios.

1.2 Distributionally Robust Learning

1.2.1 Empirical Risk Minimization

Empirical Risk Minimization (ERM) (Vapnik, 1992) is a common principle in machine learning where the goal is to learn models from labeled data that generalize well on unseen data. The true data distribution is typically unknown, so the error is estimated using the sample mean error from training samples. ERM selects a hypothesis function from a function class that minimizes the expected loss on training samples, with regularization used to balance risk minimization and generalization.

However, ERM has limitations when it comes to optimizing loss functions for fitting a hypothesis function. Many desired performance metrics are discrete, non-convex, and non-continuous, making them impractical to optimize under ERM (Höffgen and Simon, 1992; Steinwart and Christmann, 2008). To overcome this, convex loss functions are often used as proxies for the desired metrics, resulting in popular machine learning models like logistic regression (Cox, 1958), support vector machines (SVM) (Cortes and Vapnik, 1995), and conditional random fields (CRF) (Lafferty et al., 2001a).

While optimizing the surrogate loss indirectly optimizes the original loss, Fisher consistency guarantees that the learning method with the surrogate loss produces the Bayes optimal classifier. Probabilistic approaches ensure Fisher consistency but may be less efficient compared to large margin approaches. Structured SVM (Tsochantaridis et al., 2005) is the only model providing flexibility in using the desired loss metric but does not guarantee Fisher consistency.

ERM methods also lack robustness against selection bias, perturbations, noise, and outliers in training data due to their reliance on sample mean error estimation (Christmann and Steinwart, 2004; Minsker and Mathieu, 2019). An alternative approach to ERM is adversarial robust learning (Topsøe, 1979; Grünwald and Dawid, 2004; Asif et al., 2015), which addresses these limitations.

1.2.2 Adversarial Robust Learning

Minimizing the average training error through Empirical Risk Minimization (ERM) can result in absorbing biases and spurious correlations present in the training data (Christmann and Steinwart, 2004; Minsker and Mathieu, 2019). An alternative to ERM is Distributionally

Robust Learning (DRL) (Ben-Tal et al., 2013), where the model selection involves minimizing the worst-case expected loss within an ambiguity set of distributions constrained by the training data distribution. Different approaches define the distribution set using statistical distance metrics such as f-divergence (Ben-Tal et al., 2013; Namkoong and Duchi, 2016), Wasserstein distance (Esfahani and Kuhn, 2018; Shafieezadeh Abadeh et al., 2015), or moment matching (Delage and Ye, 2010; Zymler et al., 2013).

Adversarial robust learning (Asif et al., 2015) is a recent idea in robust learning that presents supervised learning as a minimax game between a predictor minimizing expected loss and an adversary maximizing expected loss under moment matching constraints from empirical data. This framework has been successfully applied for cost-sensitive classification (Asif et al., 2015), multi-class zero-one loss (Fathony et al., 2016), ordinal regression (Fathony et al., 2017) and structured prediction on graphical models (Fathony et al., 2018). This framework not only guarantees robust properties but also can optimize non-convex loss functions by treating the data distribution as uncertain. In practice, avoiding the approximation gap between the surrogate and target loss functions gives further advantages to classification performance. In this thesis, our focus is on employing the adversarial robust learning framework for probabilistic classification and structured prediction tasks. We utilize log loss as the performance metric for classification and NDCG (Normalized Discounted Cumulative Gain) for ranking. Additionally, we aim to ensure fairness in the predictor within the same conditions.

The first part of this thesis discusses *distributionally robust fairness for machine learning*. We first present two fair and robust binary classification techniques that leverage adversarial robust

framework: (1) *fair and robust minimizer of the logarithmic loss under iid assumption*; and building upon this approach (2) *fair and robust minimizer of the logarithmic loss under covariate shift*. Our approach for constructing the learning algorithms is based on the robust adversarial formulation (Topsøe, 1979; Grünwald and Dawid, 2004; Delage and Ye, 2010; Asif et al., 2015), i.e., by focusing on answering the question: *"what predictor best maximizes the performance metric (or minimizes the loss metric) in the worst case given the statistical summaries of the empirical distributions and satisfies statistical group fairness constraints?"* Next, we utilize the adversarial robust framework to develop a structured prediction method: *fair and robust learning-to-rank*. This approach maximizes the ranking utility and guarantees the fairness of exposure properties for different groups of items. We also propose ideas for developing a fair bipartite matching approach for future work by employing the same framework.

1.3 Specification Robust Learning

1.3.1 Performance-Fairness Trade-offs in Fairness Approaches

The social impacts of algorithmic decisions based on machine learning have motivated various group and individual fairness properties that decisions should ideally satisfy (Calders et al., 2009; Hardt et al., 2016). Unfortunately, impossibility results prevent multiple common group fairness properties from being simultaneously satisfied (Kleinberg et al., 2016). Thus, no set of decisions can be universally fair to all groups and individuals for all notions of fairness. Instead, specified weightings, or trade-offs, of different criteria are often optimized (Liu and Vicente, 2022). Identifying an appropriate trade-off to prescribe to these fairness methods is a

daunting task open to application-specific philosophical and ideological debate that could delay or completely derail the adoption of algorithmic methods.

Most fairness approaches optimize a specified trade-off between performance measure(s) (e.g., accuracy, log loss, or AUC) and fairness metric(s) (e.g., demographic parity, equalized odds). This frequently creates a meta-optimization problem: which measures and metrics should be traded off? In the second part of this thesis, rather than seeking a prescriptive answer to this question (using first principles or preference elicitation), we propose a data-driven approach seeking *superhuman fairness*—an amplification of the performance-fairness trade-offs that human decisions reflect.

1.3.2 Superhuman Fairness

In superhuman fairness, we aim to produce decisions with performance and fairness—potentially defined using multiple metrics—that are simultaneously better than reference decisions (e.g., from a human decision-maker or baseline method). We leverage recent advances in imitation learning to develop theory and algorithms for this task, and experimentally demonstrate the benefits of our approach when reference decisions are suboptimal. Our approach extends subdominance minimization inverse optimal control (Ziebart et al., 2022) to the fair binary classification setting (Memarrast et al., 2023c). This algorithm is flexible enough to optimize for multiple performance and fairness metrics simultaneously. Through experiments, we demonstrate its efficiency in reducing multiple fairness criteria while maintaining high performance.

The second part of this thesis discusses *specification-robust fairness for machine learning* where we use subdominance minimization for developing fair machine learning models.

1.4 Outline of the document

This thesis document is organized into six parts: Introduction; Preliminary and Background; Distributionally Robust Fair Classification; Fairness for Robust Learning To Rank; Superhuman Fairness; Conclusion and Future Directions.

Chapter 2 serves as preliminary and background. In this chapter, we present a mathematical formulation for the adversarial learning framework. We illustrate each fairness measure with a practical example and then we provide the mathematical definitions for these measures. Finally, we review the existing fair decision-making approaches in the literature.

Chapter 3 covers two research papers: "Fairness for Robust Log Loss Classification" published at AAAI Conference on Artificial Intelligence 2020 (Rezaei et al., 2020), and "Robust Fairness under Covariate Shift" published at AAAI Conference on Artificial Intelligence 2021 (Rezaei et al., 2021). In this chapter, we address two scenarios where fairness is a critical concern in machine learning. The first scenario involves robust minimization of the logarithmic loss, while considering partial knowledge of the conditional label distribution and an empirical group fairness constraint with known group membership. In the second scenario, we address the challenge of constructing a fair predictor under covariate shift, where the source and target distributions differ, but the conditional label distribution remains unchanged. The results from both scenarios are presented together in this chapter.

Chapter 4 is based on the manuscript titled "Fairness for Robust Learning To Rank" published at The Pacific-Asia Conference on Knowledge Discovery and Data Mining 2023 (Memarrast et al., 2023a). In this work, we investigate the problem of bias and unfairness in ranking as a

structured prediction problem. We develop a new robust learning-to-rank approach that can provide guarantees for fairness of exposure across groups while maximizing utility to the users. The proposed approach is able to trade-off between utility and fairness and is robust to outliers and noisy data.

Chapter 5 is based on a paper titled "Superhuman Fairness" published at the International Conference on Machine Learning 2023 (Memarrast et al., 2023b). In this chapter, we introduce superhuman fairness, an approach to fairness-aware classifier construction based on imitation learning. Our approach avoids explicit performance-fairness trade-off specification or elicitation. Instead, it seeks to unambiguously outperform human decisions across multiple performance and fairness measures with maximal frequency.

Chapter 6 concludes this thesis and discusses our ongoing and future research on fairness in machine learning, particularly in developing a fairness-aware bipartite matching approach. We use the distributionally robust learning framework to derive a fair and robust predictor for the maximum bipartite matching task, ensuring fairness constraints across protected groups.

CHAPTER 2

Preliminary and Background

2.1 Distributionally Robust Supervised Learning

Machine-learning algorithms have the potential to assume too much reliance on "accurate, clean, and well-labeled data" to provide accurate results (Schmelzer, 2019). In recent reports about AI and Machine Learning projects, about 80% of the time spent on data preparation and engineering is devoted to these activities (Minsker and Mathieu, 2019). This is one reason why robust learning is becoming increasingly relevant in practice and has gained much interest in the research community. In robust learning, the goal is to develop techniques that are relatively insensitive to outlier data or stochastic data perturbations. Another desirable property of a robust estimator is that it is efficient when the model assumptions are satisfied, i.e. that the variance is minimal (Christmann and Steinwart, 2004). Several robust learning models have been

proposed for various uncertainty set definitions based on F-divergence measures (Namkoong and Duchi, 2016; Namkoong and Duchi, 2017; Hashimoto et al., 2018), moment matching (Livni et al., 2012; Delage and Ye, 2010), and Wasserstein metric (Shafieezadeh Abadeh et al., 2015; Esfahani and Kuhn, 2018; Chen and Paschalidis, 2018).

Robust Bayes decision theory (Topsøe, 1979; Grünwald and Dawid, 2004), established the relation between the maximum entropy and the worst-case expected loss in minimax game settings, as well as the distributional robustness (Delage and Ye, 2010). Based on this, the adversarial formulation proposed by (Asif et al., 2015) for robust cost-sensitive classification seeks out the predictor that minimizes the worst-case expected loss given the statistical characteristics of the empirical data. This formulation establishes a zero-some game between the predictor distribution (P) minimizing the expected loss and an adversary distribution (Q) maximizing the expected loss while being constrained to satisfy the empirical statistics of the training data.

Definition 1. *The **adversarial formulation** for **robust** supervised learning can be formulated as:*

$$\min_{\mathbb{P}(\hat{y}|\mathbf{x})\in\Delta} \max_{\mathbb{Q}(\check{y}|\mathbf{x})\in\Delta\cap\Xi} \mathbb{E}_{\tilde{P}(\mathbf{x})\mathbb{P}(\hat{y}|\mathbf{x})\mathbb{Q}(\check{y}|\mathbf{x})}[loss(\hat{Y},\check{Y})] \tag{2.1}$$

$$\Xi : \Big\{\mathbb{Q} \mid \mathbb{E}_{\tilde{P}(\mathbf{x});\mathbb{Q}(\hat{y}|\mathbf{x})}[\phi(\mathbf{X},\hat{Y})] = \mathbb{E}_{\tilde{P}(\mathbf{x},y)}\left[\phi(\mathbf{X},Y)\right]\Big\}, \tag{2.2}$$

where $\phi(.)$ is a vector-valued feature function and Δ is the set of conditional probability simplexes (i.e. $\mathbb{P}(y|\mathbf{x}) \geq 0, \sum_y \mathbb{P}(y|\mathbf{x}) = 1, \forall \mathbf{x}, y$).

Note that the loss function in the above formulation can be any general $\mathbb{R}^{|Y|\times|Y|}$ cost matrix (Asif et al., 2015).

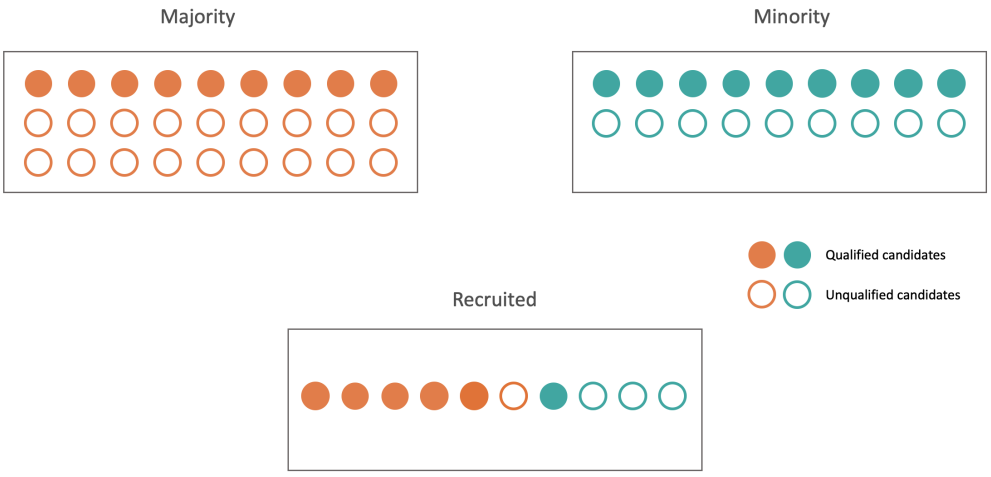
Generally, these formulations are solved by Lagrangizing the moment matching constraints Ξ and solving the dual form by optimizing the dual parameters by convex optimization methods. The loss function can be non-convex using this approach, as long as the inner minimax problem can be solved efficiently. There are two ways to solve the inner minimax game – linear programming (Asif et al., 2015) or double oracles – which relies on finding each player's best response in polynomial time (Wang et al., 2015).

2.1.1 Group Fairness Measures in Classification

We must be able to define *unfairness* before we can detect it in machine learning. We review the primary definitions of group fairness used in this thesis and explain them with a practical example (images are adopted from (Landeau, 2020)). We take into account the process of identifying potential candidates for a job. We will discuss the formal definitions and technical details for group fairness notions in 2.2. For simplicity, we consider gender (male vs female) as protected attribute in a binary classification setting where the goal is to recruit qualified candidates from both genders *fairly*. We describe scenarios where in each one we have a *fair* recruitment process based on a group fairness definition.

2.1.1.1 Demographic Parity

To achieve demographic parity, the distribution of predictions should be the same across subpopulations. In Figure 1 recruitment process satisfies demographic parity. That is, the



		Actual Class	
	Males	Qualified	Unqualified
Predicted	Recruited	5	1
	Not Recruited	4	17

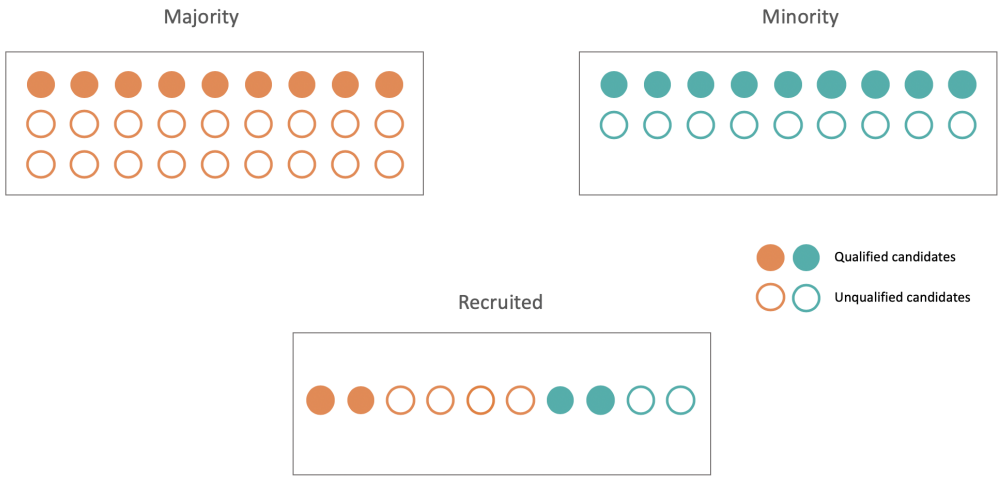
		Actual Class	
	Females	Qualified	Unqualified
	Recruited	1	3
	Not Recruited	8	6

Figure 1: **Demographic Parity**
Percentage of males recruited: $6/27 = 22\%$
Percentage of females recruited: $4/18 = 22\%$

number of candidates recruited from each gender is proportional to their population i.e. 0.22 in this example. Figure 1 also shows the confusion matrix for both groups.

2.1.1.2 Equality of Odds

Regardless of whether an applicant is a male or a female, if the candidate is qualified, he/she is equally likely to get recruited, and if not he/she is equally likely to get rejected. If this holds equal odds is satisfied. As shown in Figure 2 the number of candidates recruited from each gender is proportional to their qualified candidates i.e. 0.22 and the number of candidates not recruited from each gender is proportional to their unqualified candidates i.e. 0.77. Figure 2 shows the confusion matrix for both groups.



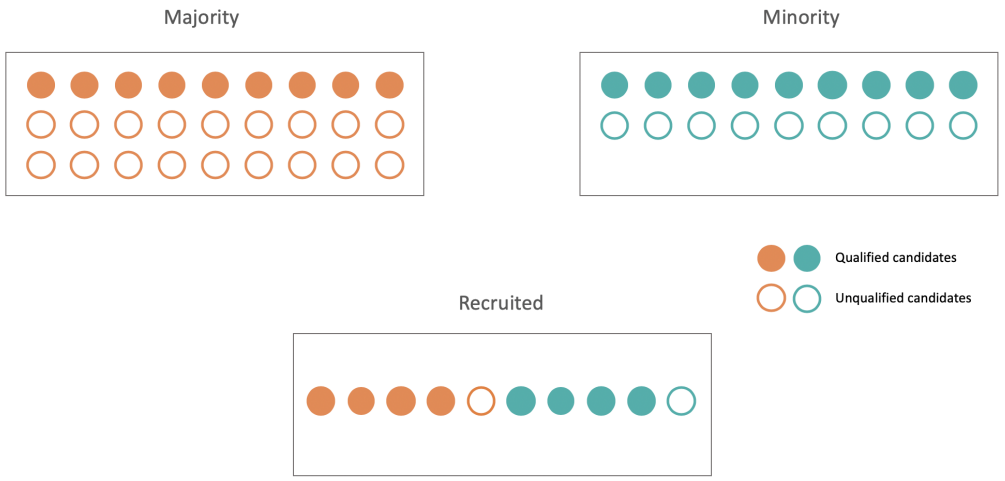
		Actual Class	
	Males	Qualified	Unqualified
Predicted	Recruited	2	4
	Not Recruited	7	14

		Actual Class	
	Females	Qualified	Unqualified
	Recruited	2	2
	Not Recruited	7	7

Figure 2: **Equal Odds**
Percentage of qualified males recruited: $2/9 = 22\%$
Percentage of qualified females recruited: $2/9 = 22\%$
Percentage of unqualified males not recruited: $14/18 = 77\%$
Percentage of unqualified females not recruited: $7/9 = 77\%$

2.1.1.3 Equality of Opportunity

Equality of opportunity is a relaxed version of equality of odds. Equality of opportunity is achieved by only applying the equality of odds constraint to the true positive rate, ensuring that each group has the same opportunity of getting recruited. In Figure 3 recruitment process satisfies equal opportunity. That is, the number of candidates recruited from each gender is proportional to their qualified candidates i.e. 0.44. Figure 3 also shows the confusion matrix for both groups.



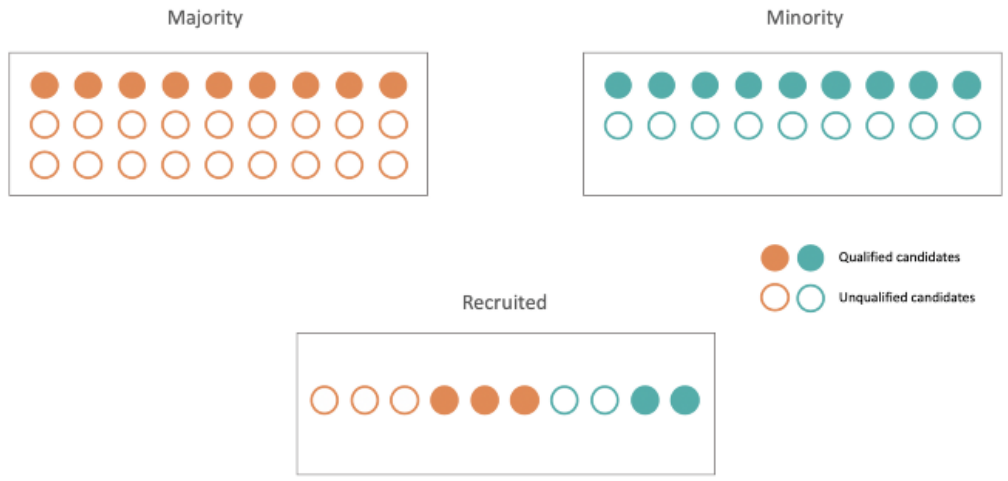
		Actual Class	
	Males	Qualified	Unqualified
Predicted	Recruited	4	1
	Not Recruited	5	17

	Actual Class	
Females	Qualified	Unqualified
Recruited	4	1
Not Recruited	5	8

Figure 3: **Equal Opportunity**
Percentage of qualified males recruited: $4/9 = 44\%$
Percentage of qualified females recruited: $4/9 = 44\%$

2.1.1.4 Predictive Rate Parity

Predictive rate parity is achieved by a model when the chances of a positive outcome in the target variable are the same for all subpopulations, regardless of their predicted positive outcome. As shown in Figure 4 the number of candidates qualified from each gender is proportional to their qualified candidates i.e. 0.50 but the number of candidates unqualified from each gender is not proportional to their not-recruited candidates (0.72 vs 0.50) since it is impossible to satisfy predictive rate parity in this example. Figure 4 shows the confusion matrix for both groups.



		Actual Class	
	Males	Qualified	Unqualified
Predicted	Recruited	3	3
	Not Recruited	6	15

		Actual Class	
	Females	Qualified	Unqualified
	Recruited	2	2
	Not Recruited	7	7

Figure 4: **Predictive Rate Parity**
Percentage of recruited males that are qualified: $3/6 = 50\%$
Percentage of recruited females that are qualified: $2/4 = 50\%$
Percentage of not-recruited males that are unqualified: $15/21 = 72\%$
Percentage of not-recruited females that are unqualified: $7/14 = 50\%$

2.1.2 Group Fairness Measures in Ranking

Similar to classification there are fairness measures introduced in the literature for structured prediction tasks. In this thesis, we focus on exposure-based group fairness measures. As a notable one, demographic parity in ranking is satisfied if average exposure for both groups is equal in top k ranks. Exposure can be measured by a position bias function that gives higher values for top ranks. In Table I two rankings are shown, ranking 1 does not satisfy demographic parity while ranking 2 satisfies this constraint.

Item	m1	m2	m3	m4	m5	m6	m7	f1	f2	f3	m → male	
Relevance	1	1	1	1	1	1	0	1	0	0	f → female	
Position	1	2	3	4	5	6	7	8	9	10	Demographic Parity violation	utility NDCG
Position bias	1	$\frac{1}{2}$	$\frac{1}{3}$	$\frac{1}{4}$	$\frac{1}{5}$	$\frac{1}{6}$	$\frac{1}{7}$	$\frac{1}{8}$	$\frac{1}{9}$	$\frac{1}{10}$	$\|Exp(m) - Exp(f)\|$	-
Ranking 1	m1	m2	m3	m4	m5	m6	f1	m7	f2	f3	0.25	1.0
Ranking 2	m1	f1	m2	f2	m3	m4	m5	f3	m6	m7	0	0.95

TABLE I: Ranking 1 maximizes utility and is oblivious to fairness constraints.
Ranking 2 maximizes utility while ensuring demographic parity.

2.2 Fairness Measures: Formal Definitions

In 1.1 we discussed the difference between group fairness and individual fairness. Additionally, in 2.1.1, we looked at practical examples where group fairness constraints are satisfied. In this section, we review the mathematical definitions for group fairness measures.

For simplicity, we consider a binary decision setting with examples drawn from a distribution: $(\mathbf{X}, A, Y) \sim P$. Here $y = 1$ is viewed as the "advantaged" class for positive decisions to be made. The general decision task is to construct a probabilistic mapping, P, for a distribution over decision variable $\hat{y} \in \{0, 1\}$ given the feature vector $\mathbf{x} \in \mathcal{X}$. Each example also possesses a protected attribute $a \in \{0, 1\}$ that defines membership in one of two groups. We consider three illustrative examples: admissions to medical school, approval of loans, and prescription of medical treatment. The relevant variables are defined for these tasks in Table II. Different forms of fairness may be appropriate in each of these decision tasks.

Setting	$\hat{y}$	y	a
Admissions	Admitted	Would succeed	Sex
Loans	Loan approval	Would re-pay	Age
Treatment	Provided	Would benefit	Race

TABLE II: Variables for three decision-making tasks.

Fairness requires treating the different groups equivalently in various ways. Unfortunately, the naïve approach of excluding the protected attribute from the decision function, e.g., restricting to $\mathbb{P}(\hat{y}|\mathbf{x})$, does not guarantee fairness because the protected attribute a may still be inferred from $\mathbf{x}$ (Dwork et al., 2012). Instead of imposing constraints on the predictor's inputs, definitions of fairness require statistical properties on its decisions to hold.

If student qualifications for medical school admissions are assumed to be the same across sexes, ensuring Demographic Parity (2.2.1) (Calders et al., 2009) may be the most appropriate form of fairness.

Definition 2.2.1. *A classifier satisfies* Demographic Parity (D.P.) *if the output variable* $\widehat{Y}$ *is statistically independent of the protected attribute* A*:*

$$\mathbb{P}(\widehat{Y}=1|A=a) = \mathbb{P}(\widehat{Y}=1), \ \ \forall a \in \{0,1\}. \tag{2.3}$$

If default rates for older (Age ≥ 40) and younger (Age < 40) loan applicants differ, providing the same approval rate to each group (Demographic Parity) may not be desirable. However, providing the same approval rates to individuals who will repay in each group and the same

approval rates to individuals who will not repay (Equalized Odds (Hardt et al., 2016)) may be a desirable fairness guarantee.

Definition 2.2.2. *A classifier satisfies* Equalized Odds (E.Odd.) *if the output variable* $\widehat{Y}$ *is conditionally independent of the protected attribute* A *given the true label* Y*:*

$$\mathbb{P}(\widehat{Y}=1|A=a,Y=y)=\mathbb{P}(\widehat{Y}=1|Y=y),\ \ \forall y,a\in\{0,1\}. \tag{2.4}$$

If there is little benefit from providing a positive decision to a non-advantaged individual, only imposing the above constraint on a particular label (the advantaged class) may be desirable (Equalized Opportunity (Hardt et al., 2016)). For example, guaranteeing the same proportion of people who would benefit from a treatment will receive the treatment in each group may be desirable without also requiring the same rates for people who would not benefit from the treatment.

Definition 2.2.3. *A classifier satisfies* Equalized Opportunity (E.Opp.) *if the output variable* $\widehat{Y}$ *and protected attribute* A *are conditionally independent given* $Y=1$*:*

$$\mathbb{P}(\widehat{Y}=1|A=a,Y=1)=\ \mathbb{P}(\widehat{Y}=1|Y=1),\ \ \forall a\in\{0,1\}. \tag{2.5}$$

By switching $\widehat{Y}$ and Y in 2.2.2, we will achieve Predictive Rate Parity, another prominent fairness metric.

Definition 2.2.4. *A classifier satisfies* Predictive Rate Parity (PRP) *if the true label Y is conditionally independent of the protected attribute A given the output variable $\widehat{Y}$:*

$$\mathbb{P}(Y=1|A=a,\widehat{Y}=y)=\mathbb{P}(Y=1|\widehat{Y}=y),\ \ \forall y,a\in\{0,1\}. \tag{2.6}$$

A relaxation of this metric is *Positive Predictive Value* (PPV) where $\widehat{Y}=1$. In this case, the approval rate needs to have the same *precision* across groups.

2.3 Existing Fair Decision-Making Methods

2.3.1 Fair Classification Techniques

Techniques for constructing predictors with group fairness properties can be broadly categorized into pre-, post-, and in-processing methods. Most methods take an agnostic approach to the actual classification method used.

Pre-processing methods such as reweighting and relabeling (Kamiran and Calders, 2012) transform the input data to remove dependence between the class and protected attribute according to a predefined fairness constraint. Other preprocessing methods (Calmon et al., 2017; Zemel et al., 2013) cast the transformation as an optimization problem to find a randomized mapping that limits the dependence of the transformed outcome on protected attribute while remaining statistically close to the original dataset.

In contrast, post-processing methods adjust the class labels (or label distributions) provided by black box classifiers to satisfy desired fairness criteria (Hardt et al., 2016; Pleiss et al., 2017; Hacker and Wiedemann, 2017). Though guaranteeing equalized odds and calibration at

the same time has been shown by (Kleinberg et al., 2016) and (Chouldechova, 2017b) to be impossible in general, (Pleiss et al., 2017) proposes a calibrated post-processing method that provides a relaxed notion of equalized odds.

In-processing approaches learn to minimize the prediction loss while incorporating the fairness constraints into the training process (Donini et al., 2018; Zafar et al., 2017; Zafar et al., 2017a; Zafar et al., 2017b; Cotter et al., 2018; Goel et al., 2018; Woodworth et al., 2017; Kamishima et al., 2011; Bechavod and Ligett, 2017; Rezaei et al., 2020; Memarrast et al., 2021), generative-adversarial training (Madras et al., 2018; Zhang et al., 2018; Celis and Keswani, 2019; Xu et al., 2018; Adel et al., 2019), reduction-based methods (Agarwal et al., 2018; Cotter et al., 2018), or meta-algorithms (Celis et al., 2019; Menon and Williamson, 2018).

2.3.2 Fair Ranking Methods

We can broadly group existing fair ranking approaches into various categories based on their notions of fairness. Some previous work has focused on compositional-based fairness for items maintaining statistical parity where the objects are positioned (Yang and Stoyanovich, 2017; Zehlike et al., 2017; Celis et al., 2018; Stoyanovich et al., 2018; Asudeh et al., 2019; Geyik et al., 2019; Celis et al., 2020). Metric-based works base their fairness constraints on statistical parity for pairwise ranking across item groups (Beutel et al., 2019; Kallus and Zhou, 2019; Narasimhan et al., 2020; Lahoti et al., 2019). Several works argue that economic opportunities (e.g., exposure, clickthroughs, etc.) should be allocated on the basis of merit, not a winner-take-all strategy (Singh and Joachims, 2018; Biega et al., 2018; Diaz et al., 2020). While fair learning-to-rank approach discussed in chapter 4 falls into this category, none of the existing techniques utilizes

a distributionally robust approach to derive a fair LTR system like ours. As a result, their performance degrades in the presence of training label noise (shown in chapter 4). In addition to item-based approaches, two-sided fair ranking techniques satisfy fairness constraints for both users and items (Do et al., 2021; Patro et al., 2020a; Basu et al., 2020; Patro et al., 2020b).

As opposed to group fairness in ranking, there have been some works focusing on individual fair ranking. Group fair models can be inherently unfair to individuals. Among the models, (Biega et al., 2018) and (Singh and Joachims, 2019) propose individually fair LTR methods that measure the similarity of items through relevance. In a fundamentally different approach, (Bower et al., 2020) builds upon the work of (Dwork et al., 2012) to employ a fair metric on queries.

There have also been recent studies that focus on other aspects of fair ranking. Several works have looked at fair ranking in the presence of noisy protected attributes (Mehrotra and Celis, 2021; Mehrotra and Vishnoi, 2022). Another line of research aims to select individuals distributed across different groups fairly when there is implicit group bias (Kleinberg and Raghavan, 2018; Celis et al., 2020). Recent studies have also investigated how uncertainty about protected attributes, labels, and other features of the machine learning model affect its fairness properties (Ghosh et al., 2021; Prost et al., 2021). Contrary to this line of work, (Singh et al., 2021) takes into account the presence of uncertainty when estimating merits and defining a corresponding merit-based notion of fairness. In a different direction, (Oosterhuis, 2021) proposes a computationally efficient method for fair LTR using Plackett-Luce models, ensuring its viability in real-world scenarios.

2.3.3 Trade-offs in Fair Methods

Following our discussion in 1.3.1 we look at some of the fair techniques that achieve a performance-fairness trade-off. Numerous fair classification algorithms have been developed over the past few years, with most targeting one or two fairness measures (Zafar et al., 2015; Hardt et al., 2016; Goel et al., 2018; Aghaei et al., 2019). With some exceptions (Blum and Stangl, 2019), predictive performance and fairness are typically competing objectives in supervised machine learning approaches (Menon and Williamson, 2018). Thus, though satisfying many fairness properties simultaneously may be naïvely appealing, doing so often significantly degrades predictive performance or even creates infeasibility (Kleinberg et al., 2016).

Given this, many approaches seek to choose parameters θ for (probabilistic) classifier P_θ that balance the competing predictive performance and fairness objectives (Kamishima et al., 2012; Hardt et al., 2016; Menon and Williamson, 2018; Celis et al., 2019; Martinez et al., 2020; Rezaei et al., 2020). Recently, (Hsu et al., 2022) proposed a novel optimization framework to satisfy three conflicting fairness measures (demographic parity, equalized odds, and predictive rate parity) to the best extent possible:

$$\min_{\theta} \mathbb{E}_{\hat{\mathbf{y}} \sim P_\theta} \Big[\text{loss}(\hat{\mathbf{y}}, \mathbf{y}) + \alpha_{\texttt{DP}} \texttt{D.DP}(\hat{\mathbf{y}}, \mathbf{a}) + \alpha_{\texttt{Odds}} \texttt{D.EqOdds}(\hat{\mathbf{y}}, \mathbf{y}, \mathbf{a}) + \alpha_{\texttt{PRP}} \texttt{D.PRP}(\hat{\mathbf{y}}, \mathbf{y}, \mathbf{a}) \Big]. \quad (2.7)$$

CHAPTER 3

Distributionally Robust Fair Classification

(Parts of this chapter were previously published as "Fairness for Robust Log Loss Classification" (Rezaei et al., 2020) in the AAAI Conference on Artificial Intelligence 34 (AAAI 2020) and as "Robust Fairness under Covariate Shift" (Rezaei et al., 2021) in the AAAI Conference on Artificial Intelligence 35 (AAAI 2021).)

3.1 Fairness for Robust Log Loss Classification

The adversarial robust formulation for logarithmic loss has been shown to reduces to constrained entropy maximization (Topsøe, 1979). This, however, will no longer be true under additional constraints on the predictor. We seek to leverage this formulation to re-derive a new classifier from the first principles of distributional robustness that incorporates fairness criteria

into a worst-case logarithmic loss minimization. We again consider a binary decision setting with examples drawn from a population distribution: $(\mathbf{X}, A, Y) \sim P$, with $\tilde{P}(\mathbf{x}, a, y)$ denoting this empirical sample distribution: $\{\mathbf{x}_i, a_i, y_i\}_{i=1:n}$

3.1.1 Robust Log Loss Minimization

The logarithmic loss, $-\sum_{\mathbf{x},y} P(\mathbf{x}, y) \log \mathbb{P}(y|\mathbf{x})$, is an information-theoretic measure of the expected amount of "surprise" (in bits for $\log_2$) that the predictor, $\mathbb{P}(y|\mathbf{x})$, experiences when encountering labels y distributed according to $P(\mathbf{x}, y)$. Robust minimization of the logarithmic loss serves a fundamental role in constructing exponential probability distributions (e.g., Gaussian, Laplacian, Beta, Gamma, Bernoulli (Lisman and Zuylen, 1972)) and predictors (Manning and Klein, 2003). For conditional probabilities, it is equivalent to maximizing the conditional entropy (Jaynes, 1957):

$$\min_{\mathbb{P}(\hat{y}|\mathbf{x})\in\Delta} \max_{\mathbb{Q}(\hat{y}|\mathbf{x})\in\Delta\cap\Xi} -\sum_{\mathbf{x},\hat{y}} \tilde{P}(\mathbf{x})\mathbb{Q}(\hat{y}|\mathbf{x}) \log \mathbb{P}(\hat{y}|\mathbf{x}) \tag{3.1}$$

$$= \max_{\mathbb{P}(\hat{y}|\mathbf{x})\in\Xi} -\sum_{\mathbf{x},\hat{y}} \tilde{P}(\mathbf{x})\mathbb{P}(\hat{y}|\mathbf{x}) \log \mathbb{P}(\hat{y}|\mathbf{x}) = \max_{\hat{P}(\hat{y}|\mathbf{x})\in\Xi} H(\hat{Y}|\mathbf{X}),$$

after simplifications based on the fact that the saddle point solution is $\mathbb{P} = \mathbb{Q}$. When the loss maximizer $\mathbb{Q}$ is constrained to match the statistics of training data (specified using vector-valued feature function ϕ),

$$\Xi : \Big\{ \mathbb{Q} \mid \mathbb{E}_{\tilde{P}(\mathbf{x});\mathbb{Q}(\hat{y}|\mathbf{x})}[\phi(\mathbf{X}, \hat{Y})] = \mathbb{E}_{\tilde{P}(\mathbf{x},y)} \left[\phi(\mathbf{X}, Y)\right] \Big\}, \tag{3.2}$$

the robust log loss minimizer/maximum entropy predictor (Equation 3.1) is the logistic regression model, $P(y|\mathbf{x}) \propto e^{\theta^{\mathrm{T}}\phi(\mathbf{x},y)}$, with θ estimated by maximizing data likelihood (Manning and Klein, 2003). While this distribution technically needs to only be defined at input values in which training data exists (i.e., $\tilde{P}(\mathbf{x}) > 0$), we employ an inductive assumption that generalizes the form of the distribution to other inputs.

3.1.2 Robust and Fair Log Loss Minimization

We recall the fairness definitions 2.2.1, 2.2.2 and 2.2.3. As mentioned in 2.2, The sets of decision functions P satisfying these fairness constraints are convex and can be defined using linear constraints (Agarwal et al., 2018). The general form for these constraints is:

$$\Gamma : \left\{ \mathbb{P} \mid \frac{1}{p_{\gamma_1}} \mathbb{E}_{\substack{\tilde{P}(\mathbf{x},a,y) \\ \mathbb{P}(\hat{y}|\mathbf{x},a,y)}} [\mathbb{I}(\hat{Y}=1 \wedge \gamma_1(A,Y))] = \frac{1}{p_{\gamma_0}} \mathbb{E}_{\substack{\tilde{P}(\mathbf{x},a,y) \\ \mathbb{P}(\hat{y}|\mathbf{x},a,y)}} [\mathbb{I}(\hat{Y}=1 \wedge \gamma_0(A,Y))] \right\}, \tag{3.3}$$

where γ_1 and γ_0 denote some combination of group membership and ground-truth class for each example, while p_{γ_1} and p_{γ_0} denote the empirical frequencies of γ_1 and γ_0: $p_{\gamma_i} = \mathbb{E}_{\tilde{P}(a,y)}[\gamma_i(A,Y)]$. We specify γ_1 and γ_0 in (Equation 3.3) for fairness constraints (Definitions 2.2.1, 2.2.2 and 2.2.3) as:

$$\Gamma_{\text{dp}} \iff \gamma_j(A,Y) = \mathbb{I}(A=j); \tag{3.4}$$

$$\Gamma_{\text{e.opp}} \iff \gamma_j(A,Y) = \mathbb{I}(A=j \wedge Y=1); \tag{3.5}$$

$$\Gamma_{\text{e.odd}} \iff \gamma_j(A,Y) = \begin{bmatrix} \mathbb{I}(A=j \wedge Y=1) \\ \\ \mathbb{I}(A=j \wedge Y=0) \end{bmatrix}. \tag{3.6}$$

Since the fairness constraints and statistic-matching constraints are often not fully compatible (i.e., $\Gamma \not\subseteq \Xi$), the saddle point solution is no longer simple (i.e., $\mathbb{P} \neq \mathbb{Q}$).

3.1.2.1 Formulation and Algorithms

Given fairness requirements for a predictor (Equation 3.3) and partial knowledge of the population distribution provided by a training sample (Equation 3.2), how should a fair predictor be constructed? Like all inductive reasoning approaches, good performance on a known training sample does not ensure good performance on the unknown population distribution. We take a robust estimation perspective by seeking the best solution for the worst-case population distribution under these constraints.

3.1.2.2 Robust and fair log loss minimization

We formulate the robust fair predictor's construction as a minimax game between the predictor and a worst-case approximator of the population distribution. We assume the availability of a set of training samples, $\{(\mathbf{x}_i, a_i, y_i)\}_{i=1:n}$, which we equivalently denote by probability distribution $\tilde{P}(\mathbf{x}, a, y)$.

Definition 3.1.1. *The* **Fair Robust Log-Loss Predictor**, $\mathbb{P}$, *minimizes the worst-case log loss—as chosen by approximator* $\mathbb{Q}$ *constrained to reflect training statistics (denoted by*

set Ξ *of* (Equation 3.2)*)—while providing empirical fairness guarantees*[1] *(denoted by set* Γ *of* (Equation 3.3)*):*

$$\min_{\mathbb{P}\in\Delta\cap\Gamma}\max_{\mathbb{Q}\in\Delta\cap\Xi}\mathbb{E}_{\substack{\tilde{P}(\mathbf{x},a,y)\\ \mathbb{Q}(\hat{y}|\mathbf{x},a,y)}}\left[-\log\mathbb{P}(\hat{Y}|\mathbf{X},A,Y)\right]. \tag{3.7}$$

Though conditioning the decision variable $\hat{Y}$ on the true label Y would appear to introduce a trivial solution ($\hat{Y}=Y$), instead, Y only influences $\hat{Y}$ based on fairness properties due to the robust predictor's construction. Note that if the fairness constraints do not relate Y and $\hat{Y}$, the resulting distribution is conditionally independent (i.e., $\mathbb{P}(\hat{Y}|\mathbf{X},A,Y=0)=\mathbb{P}(\hat{Y}|\mathbf{X},A,Y=1)$), and when all fairness constraints are removed, this formulation reduces to the familiar logistic regression model (Manning and Klein, 2003). Conveniently, this saddle point problem is convex-concave in $\mathbb{P}$ and $\mathbb{Q}$ with additional convex constraints (Γ and Ξ) on each distribution.

3.1.2.3 Parametric Distribution Form

By leveraging strong minimax duality in the "log-loss game" (Topsøe, 1979; Grünwald and Dawid, 2004) and strong Lagrangian duality (Boyd and Vandenberghe, 2004), we derive the parametric form of our predictor.[2]

---

[1] Δ *is the set of conditional probability simplexes (i.e.,* $\mathbb{P}(y|\mathbf{x},a)\geq 0, \sum_{y'}\mathbb{P}(y'|\mathbf{x},a)=1, \forall\mathbf{x},y,a$*).*

[2] The proofs of Theorem 3.1.2 and other theorems in the paper are available in the supplementary material.

Theorem 3.1.2. *The* **Fair Robust Log-Loss Predictor** *(Definition 3.1.1) has equivalent dual formulation:*

$$\begin{aligned}\min_{\theta}\max_{\lambda}\ \frac{1}{n}\sum_{(\mathbf{x},a,y)\in\mathcal{D}}\Bigg\{&\mathbb{E}_{\mathbb{Q}_{\theta,\lambda}(\hat{y}|\mathbf{x},a,y)}\Big[-\log\mathbb{P}_{\theta,\lambda}(\hat{Y}|\mathbf{x},a,y)\Big]\\&+\theta^{\top}\Big(\mathbb{E}_{\mathbb{Q}_{\theta,\lambda}(\hat{y}|\mathbf{x},a,y)}[\phi(\mathbf{x},\hat{Y})]-\phi(\mathbf{x},y)\Big)\\&+\lambda\big(\tfrac{1}{p_{\gamma_1}}\mathbb{E}_{\mathbb{P}_{\theta,\lambda}(\hat{y}|\mathbf{x},a,y)}[\mathbb{I}(\hat{Y}=1\wedge\gamma_1(A,Y))]\\&-\tfrac{1}{p_{\gamma_0}}\mathbb{E}_{\mathbb{P}_{\theta,\lambda}(\hat{y}|\mathbf{x},a,y)}[\mathbb{I}(\hat{Y}=1\wedge\gamma_0(A,Y))]\big)\Bigg\},\end{aligned}\tag{3.8}$$

with Lagrange multipliers θ and λ for moment matching and fairness constraints, respectively, and n samples in the dataset. The parametric distribution of $\mathbb{P}$ is:

$$\mathbb{P}_{\theta,\lambda}(\hat{y}=1|\mathbf{x},a,y)=\tag{3.9}$$

$$\begin{cases}\min\left\{e^{\theta^{\top}\phi(\mathbf{x},1)}/Z_{\theta}(\mathbf{x}),\frac{p_{\gamma_1}}{\lambda}\right\} & \text{if } \gamma_1(a,y)\wedge\lambda>0\\ \max\left\{e^{\theta^{\top}\phi(\mathbf{x},1)}/Z_{\theta}(\mathbf{x}),1-\frac{p_{\gamma_0}}{\lambda}\right\} & \text{if } \gamma_0(a,y)\wedge\lambda>0\\ \max\left\{e^{\theta^{\top}\phi(\mathbf{x},1)}/Z_{\theta}(\mathbf{x}),1+\frac{p_{\gamma_1}}{\lambda}\right\} & \text{if } \gamma_1(a,y)\wedge\lambda<0\\ \min\left\{e^{\theta^{\top}\phi(\mathbf{x},1)}/Z_{\theta}(\mathbf{x}),-\frac{p_{\gamma_0}}{\lambda}\right\} & \text{if } \gamma_0(a,y)\wedge\lambda<0\\ e^{\theta^{\top}\phi(\mathbf{x},1)}/Z_{\theta}(\mathbf{x}) & \text{otherwise,}\end{cases}$$

where $Z_\theta(\mathbf{x}) = e^{\theta^\top \phi(\mathbf{x},1)} + e^{\theta^\top \phi(\mathbf{x},0)}$ *is the normalization constant. The parametric distribution of* $\mathbb{Q}$ *is defined using the following relationship with* $\mathbb{P}$*:*

$$\mathbb{Q}_{\theta,\lambda}(\hat{y}=1|\mathbf{x},a,y) = \mathbb{P}_{\theta,\lambda}(\hat{y}=1|\mathbf{x},a,y)\times \tag{3.10}$$
$$\begin{cases} \left(1+\frac{\lambda}{p_{\gamma_1}}\mathbb{P}_{\theta,\lambda}(\hat{y}=0|\mathbf{x},a,y)\right) & \textit{if } \gamma_1(a,y) \\ \left(1-\frac{\lambda}{p_{\gamma_0}}\mathbb{P}_{\theta,\lambda}(\hat{y}=0|\mathbf{x},a,y)\right) & \textit{if } \gamma_0(a,y) \\ 1 & \textit{otherwise.} \end{cases}$$

Note that the predictor's distribution is a member of the exponential family that is similar to standard binary logistic regression, but with the option to *truncate* the probability based on the value of λ. The truncation of $\mathbb{P}_{\theta,\lambda}(\hat{y}=1|\mathbf{x},a,y)$ is from above when $0 < p_{\gamma_1}/\lambda < 1$ and $\gamma_1(a,y)=1$, and from below when $-1 < p_{\gamma_1}/\lambda < 0$ and $\gamma_1(a,y)=1$. The approximator's distribution is computed from the predictor's distribution using the quadratic function in (Equation 3.10), e.g., in the case where $\gamma_1(a,y)=1$:

$$\mathbb{Q}_{\theta,\lambda}(\hat{y}=1|\mathbf{x},a,y) = \rho(1+\tfrac{\lambda}{p_{\gamma_1}}(1-\rho)) = (1+\tfrac{\lambda}{p_{\gamma_1}})\rho - \tfrac{\lambda}{p_{\gamma_1}}\rho^2,$$

where $\rho \triangleq \mathbb{P}_{\theta,\lambda}(\hat{y}=1|\mathbf{x},a,y)$.

Figure 5 illustrates the relationship between $\mathbb{P}_{\theta,\lambda}(\hat{y}=1|\mathbf{x},a,y)$ and $\mathbb{Q}_{\theta,\lambda}(\hat{y}=1|\mathbf{x},a,y)$ for decisions influencing the fairness of group one (i.e., $\gamma_1(a,y)=1$). When $\lambda/p_{\gamma_1}=0$, the approximator's probability is equal to the predictor's probability as shown in the plot as a

---

[1] `https://github.com/gpleiss/equalized_odds_and_calibration`

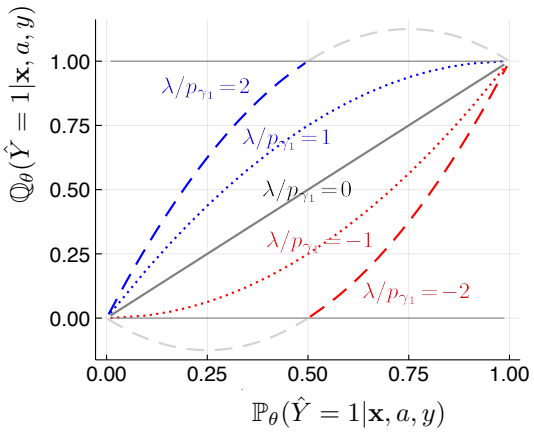


Figure 5: The relationship between predictor and approximator's distributions, $\mathbb{P}$ and $\mathbb{Q}$.

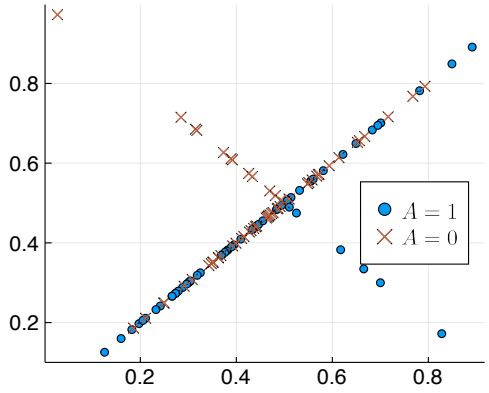


Figure 6: Post-processing correction [1] of logistic regression (Pleiss et al., 2017; Hardt et al., 2016) on the COMPAS dataset.

straight line. Positive values of λ curve the function upward (e.g., $\lambda/p_{\gamma_1}=1$) as shown in the plot. For larger λ (e.g., $\lambda/p_{\gamma_1}=2$), some of the valid predictor probabilities ($0<\mathbb{P}<1$) map to invalid approximator probabilities (i.e., $\mathbb{Q}\geq 1$) according to the quadratic function. In this case (e.g., $\lambda/p_{\gamma_1}=2$ and $\mathbb{P}_{\theta,\lambda}(\hat{y}=1|\mathbf{x},a,y)>0.5$), the predictor's probability is truncated to $p_{\gamma_1}/\lambda=0.5$ according to (Equation 3.9). Similarly, for negative λ, the curve is shifted downward and the predictor's probability is truncated when the quadratic function mapping results in a negative value of $\mathbb{Q}$. When $\gamma_0(a,y)=1$, the reverse shifting is observed, i.e., shifting downward when $\lambda>0$ and shifting upward when $\lambda<0$.

We contrast our reshaping function of the decision distribution (Figure 5) with the *post-processing method* of (Hardt et al., 2016) shown in Figure 6. Here, we use $\mathbb{Q}(\hat{Y}=1|\mathbf{x},a)$ to represent the estimating distributions (the approximator's distribution in our method, and

the standard logistic regression in (Hardt et al., 2016)) and the post-processed predictions as $\mathbb{P}(\hat{Y} = 1|\mathbf{x}, a)$. Both shift the positive prediction rates of each group to provide fairness. However, our approach provides a monotonic and parametric transformation, avoiding the criticisms that (Hardt et al., 2016)'s modification (flipping some decisions) is partially random, creating an unrestricted hypothesis class (Bechavod and Ligett, 2017). Additionally, since our parametric reshaping function is learned within an *in-processing* method, it avoids the noted suboptimalities that have been established for certain population distributions when employing post-processing alone (Woodworth et al., 2017).

3.1.2.4 Enforcing fairness constraints

The inner maximization in (Equation 3.8) finds the optimal λ that enforces the fairness constraint. From the perspective of the parametric distribution of $\mathbb{P}$, this is equivalent to finding threshold points (e.g., p_{γ_1}/λ and $1 - p_{\gamma_0}/\lambda$) in the min and max function of (Equation 3.9) such that the expectation of the truncated exponential probabilities of $\mathbb{P}$ in group γ_1 match the one in group γ_0. Given the value of θ, we find the optimum λ^* directly by finding the threshold points. We first compute the exponential probabilities $P_e(\hat{y} = 1|\mathbf{x}, a, y) = \exp(\theta^\top \phi(\mathbf{x}, 1))/Z_\theta(\mathbf{x})$ for each examples in γ_1 and γ_0. Let E_1 and E_0 be the sets that contain P_e for group γ_1 and γ_0 respectively. Finding λ^* given the sets E_1 and E_0 requires sorting the probabilities for each set, and then iteratively finding the threshold points for both sets simultaneously. We refer to the supplementary material for the detailed algorithm.

3.1.2.5 Learning

Our learning process seeks parameters θ, λ for our distributions ($\mathbb{P}_{\theta,\lambda}$ and $\mathbb{Q}_{\theta,\lambda}$) that match the statistics of the approximator's distribution with training data (θ) and provide fairness (λ), as illustrated in (Equation 3.8). Using our algorithm from the previous subsection to directly compute the best λ given arbitrary values of θ, denoted λ^*_θ, the optimization of (Equation 3.8) reduces to a simpler optimization solely over θ, as described in Theorem 3.1.3.

Theorem 3.1.3. *Given the optimum value of λ^*_θ for θ, the dual formulation in* (Equation 3.8) *reduces to:*

$$\min_\theta \tfrac{1}{n} \textstyle\sum_{(\mathbf{x},a,y)\in\mathcal{D}} \ell_{\theta,\lambda^*_\theta}(\mathbf{x},a,y), \qquad \textit{where:} \tag{3.11}$$

$$\ell_{\theta,\lambda^*}(\mathbf{x},a,y) = -\theta^\top\phi(\mathbf{x},y) +$$

$$\begin{cases} -\log(\frac{p_{\gamma_1}}{\lambda^*_\theta}) + \theta^\top(\phi(\mathbf{x},1)) & \textit{if } \gamma_1(a,y) \wedge T(\mathbf{x},\theta) \wedge \lambda^*_\theta > 0 \\ -\log(\frac{p_{\gamma_0}}{\lambda^*_\theta}) + \theta^\top(\phi(\mathbf{x},0)) & \textit{if } \gamma_0(a,y) \wedge T(\mathbf{x},\theta) \wedge \lambda^*_\theta > 0 \\ -\log(-\frac{p_{\gamma_1}}{\lambda^*_\theta}) + \theta^\top(\phi(\mathbf{x},0)) & \textit{if } \gamma_1(a,y) \wedge T(\mathbf{x},\theta) \wedge \lambda^*_\theta < 0 \\ -\log(-\frac{p_{\gamma_0}}{\lambda^*_\theta}) + \theta^\top(\phi(\mathbf{x},1)) & \textit{if } \gamma_0(a,y) \wedge T(\mathbf{x},\theta) \wedge \lambda^*_\theta < 0 \\ \log Z_\theta(\mathbf{x}) & \textit{otherwise.} \end{cases}$$

Here, $T(\mathbf{x},\theta) \triangleq 1$ if the exponential probability is truncated (for example when $e^{\theta^\top\phi(\mathbf{x},1)}/Z_\theta(\mathbf{x}) > p_{\gamma_1}/\lambda^*_\theta$, $\gamma_1(a,y) = 1$, and $\lambda^*_\theta > 0$), and is 0 otherwise.

We present an important optimization property for our objective function in the following theorem.

Theorem 3.1.4. *The objective function in Theorem 3.1.3* (Equation 3.11) *is convex with respect to* θ.

To improve the generalizability of our parametric model, we employ a standard L2 regularization technique that is common for logistic regression models: $\theta^* = \text{argmin}_\theta \sum_{(\mathbf{x},a,y)\in\mathcal{D}} \ell_{\theta,\lambda^*_\theta}(\mathbf{x}, a, y) + \frac{C}{2}\|\theta\|_2^2$, where C is the regularization constant. We employ a standard batch gradient descent optimization algorithm (e.g., L-BFGS) to obtain a solution for θ^*.[1] We also compute the corresponding solution for the inner optimization, $\lambda^*_{\theta^*}$, and then construct the optimal predictor and approximator's parametric distributions based on the values of θ^* and $\lambda^*_{\theta^*}$.

3.1.2.6 Inference

In the inference step, we apply the optimal parametric predictor distribution $\mathbb{P}_{\theta^*,\lambda^*_{\theta^*}}$ to new example inputs $(\mathbf{x}, a)$ in the testing set. Given the value of θ^* and $\lambda^*_{\theta^*}$, we calculate the predictor's distribution for our new data point using (Equation 3.9). Note that the predictor's parametric distribution also depends on the group membership of the example. For fairness constraints not based on the actual label Y, e.g., D.P., this parametric distribution can be directly applied to make predictions. However, for fairness constraints that depend on the true label, e.g., E.Opp. and E.Odd., we introduce a prediction procedure that estimates the true label using the approximator's parametric distribution.

For fairness constraints that depend on the true label, our algorithm outputs the predictor and approximator's parametric distributions conditioned on the value of true label, i.e., $\mathbb{P}(\hat{y}|\mathbf{x}, a, y)$

---

[1] We refer the reader to the supplementary material for details.

and $\mathbb{Q}(\hat{y}|\mathbf{x}, a, y)$. Our goal is to produce the conditional probability of $\hat{y}$ that does not depend on the true label, i.e., $\mathbb{P}(\hat{y}|\mathbf{x}, a)$. We construct the following procedure to estimate this probability. Based on the marginal probability rule, $\mathbb{P}(\hat{y}|\mathbf{x}, a)$ can be expressed as:

$$\begin{aligned}\mathbb{P}(\hat{y}|\mathbf{x}, a) = \mathbb{P}(\hat{y}|\mathbf{x}, a, y = 1)P(y = 1|\mathbf{x}, a) \\ + \mathbb{P}(\hat{y}|\mathbf{x}, a, y = 0)P(y = 0|\mathbf{x}, a).\end{aligned} \tag{3.12}$$

However, since we do not have access to $P(y|\mathbf{x}, a)$, we cannot directly apply this expression. Instead, we approximate $P(y|\mathbf{x}, a)$ with the approximator's distribution $\mathbb{Q}(\hat{y}|\mathbf{x}, a)$. Using the similar marginal probability rule, we express the estimate as:

$$\begin{aligned}\mathbb{Q}(\hat{y}|\mathbf{x}, a) \approx \mathbb{Q}(\hat{y}|\mathbf{x}, a, y = 1)\mathbb{Q}(\hat{y} = 1|\mathbf{x}, a) \\ + \mathbb{Q}(\hat{y}|\mathbf{x}, a, y = 0)\mathbb{Q}(\hat{y} = 0|\mathbf{x}, a).\end{aligned} \tag{3.13}$$

By rearranging the terms above, we calculate the estimate as:

$$\begin{aligned}\mathbb{Q}(\hat{y}=1|\mathbf{x}, a)=\mathbb{Q}(\hat{y}=1|\mathbf{x}, a, y=0)/(\mathbb{Q}(\hat{y}=0|\mathbf{x}, a, y=1) \\ +\mathbb{Q}(\hat{y}=1|\mathbf{x}, a, y=0)),\end{aligned} \tag{3.14}$$

which is directly computed from the approximator's parametric distribution produced by our model using (Equation 3.10). Finally, to obtain the predictor's conditional probability

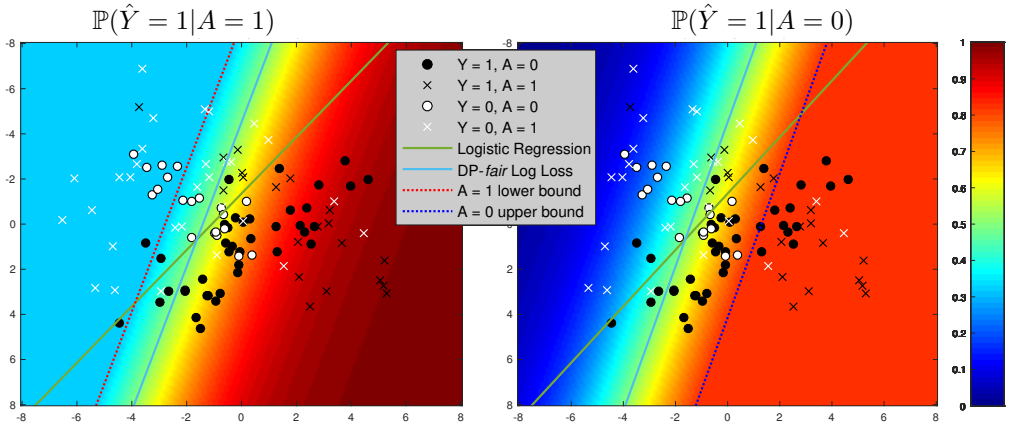


Figure 7: Experimental results on a synthetic dataset with: a heatmap indicating the predictive probabilities of our approach, along with decision and threshold boundaries; and the unfair logistic regression decision boundary.

estimate ($\mathbb{P}(\hat{y}|\mathbf{x}, a)$), we replace $P(y|\mathbf{x}, a)$ in (Equation 3.12) with $\mathbb{Q}(\hat{y}|\mathbf{x}, a)$ calculated from (Equation 3.14).

3.1.3 Experiments

3.1.3.1 Illustrative behavior on synthetic data

We illustrate the key differences between our model and logistic regression with demographic parity requirements on 2D synthetic data in Figure 7. The predictive distribution includes different truncated probabilities for each group: raising the minimum probability for group $A = 1$ and lowering the maximum probability for group $A = 0$. This permits a decision boundary that differs significantly from the logistic regression decision boundary and better realizes the desired fairness guarantees. In contrast, *post-processing methods* using logistic regression as the base classifier (Hardt et al., 2016) are constrained to reshape the given unfair logistic regression

predictions without shifting the decision boundary orientation, often leading to suboptimality (Woodworth et al., 2017).

3.1.3.2 Datasets

We evaluate our proposed algorithm on three benchmark fairness datasets:

1. The **UCI Adult** (Dheeru and Karra Taniskidou, 2017) dataset includes 45,222 samples with an income greater than $50k considered to be a favorable binary outcome. We choose gender as the protected attribute, leaving 11 other features for each example.
2. The ProPublica's **COMPAS** recidivism dataset (Larson et al., 2016) contains 6,167 samples, and the task is to predict the recidivism of an individual based on criminal history, with the binary protected attribute being race (white and non-white) and an additional nine features.
3. The dataset from the Law School Admissions Council's National Longitudinal Bar Passage Study (Wightman, 1998) has 20,649 examples. Here, the favorable outcome for the individual is passing the bar exam, with race (restricted to white and black only) as the protected attribute, and 13 other features.

3.1.3.3 Comparison methods

We compare our method (*Fair Log-loss*) against various baseline/fair learning algorithms that are primarily based on logistic regression as the base classifier:

1. **Unconstrained logistic regression** is a standard logistic regression model that ignores all fairness requirements.
2. The **cost sensitive reduction approach** by (Agarwal et al., 2018) reduces fair classification to learning a randomized hypothesis over a sequence of cost-sensitive classifiers. We use the

sample-weighted implementation of Logistic Regression in scikit-learn as the base classifier, to compare the effect of the reduction approach. We evaluate the performance of the model by varying the constraint bounds across the set $\epsilon \in \{.001, .01, .1\}$.

3. The **constraint-based learning method**[1] of (Zafar et al., 2017; Zafar et al., 2017a) uses a covariance proxy measure to achieve equalized odds (under the name disparate mistreatment) (Zafar et al., 2017a), and improve the disparate impact ratio (Zafar et al., 2017), which we use as a baseline method to evaluate demographic parity violation. They cast the resulting non-convex optimization as a disciplined convex-concave program in training time. We use the logistic regression as the base classifier.

4. For demographic parity, we compare with the **reweighting method** (*reweighting*) of (Kamiran and Calders, 2012), which learns weights for each combination of class label and protected attribute and then uses these weights to resample from the original training data which yields a new dataset with no statistical dependence between class label and protected attribute. The new balanced dataset is then used for training a classifier. We use IBM AIF360 toolkit to run this method.

5. For equalized odds, we also compare with the **post-processing method** of (Hardt et al., 2016) which transforms the classifier's output by solving a linear program that finds a prediction minimizing misclassification errors and satisfying the equalized odds constraint

---

[1] `https://github.com/mbilalzafar/fair-classification`

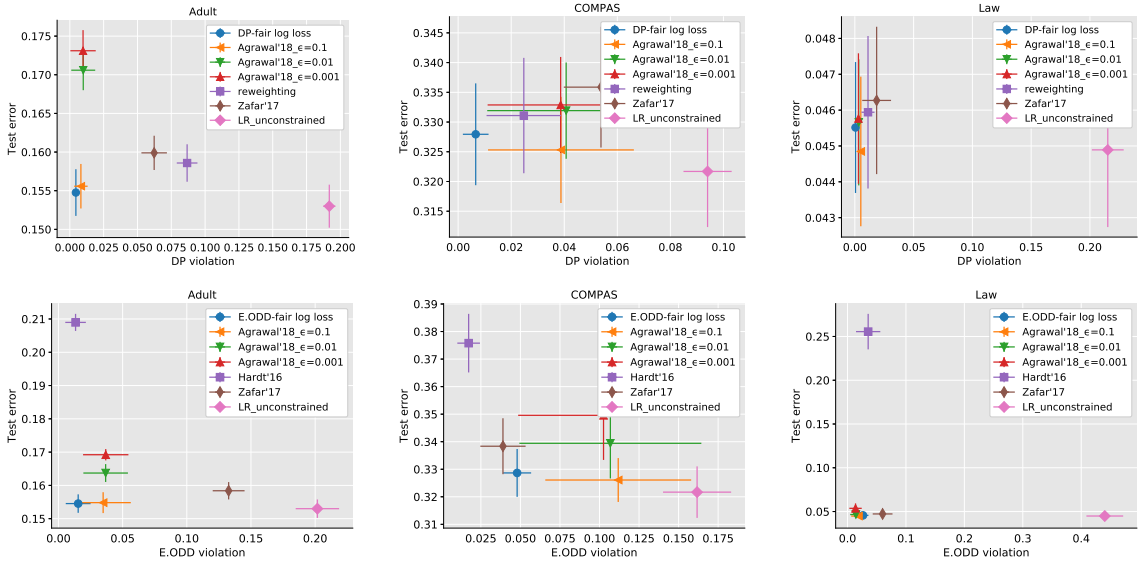


Figure 8: *Test classification error* versus *Demographic Parity* (top row) and *Equalized Odds* (bottom row) constraint violations. The bars indicate standard deviation on 20 random splits of data.

from the set of probability formed by the convex hull of the original classifier's probabilities and the extreme point of probability values (i.e., zero and one).

3.1.3.4 Evaluation measures and setup

Data-driven fair decision methods seek to minimize both prediction error rates and measures of unfairness. We consider the misclassification rate (i.e., the 0-1 loss, $\mathbb{E}[\hat{Y} \neq Y]$) on a withheld test sample to measure prediction error. To quantify the unfairness of each method, we measure the degree of fairness violation for demographic parity (D.P.) as: $\left|\mathbb{E}[\mathbb{I}(\hat{Y}=1)|A=1] - \mathbb{E}[\mathbb{I}(\hat{Y}=1)|A=0]\right|$, and the sum of fairness violations for each class to measure the total violation for equalized odds (E.Odd.) as: $\sum_{y\in\{0,1\}} \left(\left|\mathbb{E}[\mathbb{I}(\hat{Y}=1)|A=1,Y=y] - \mathbb{E}[\mathbb{I}(\hat{Y}=1)|A=0,Y=y]\right|\right)$,

to obtain a level comparison across different methods. We follow the methodology of (Agarwal et al., 2018) to give all methods access to the protected attribute both at training and testing time by including the protected attribute in the feature vector. We perform all of our experiments using 20 random splits of each dataset into a training set (70% of examples) and a testing set (30%). We record the averages over these twenty random splits and the standard deviation. We cross validate our model on a separate validation set using the best logloss to select an L2 penalty from ($\{.001, .005, .01, .05, .1, .2, .3, .4, .5\}$).

3.1.3.5 Experimental Results

Figure 8 provides the evaluation results (test error and fairness violation) of each method for demographic parity and equalized odds on test data from each of the three datasets Fairness can be vacuously achieved by an agnostic predictor that always outputs labels according to independent (biased) coin flips. Thus, the appropriate question to ask when considering these results is: "how much additional test error is incurred compared to the baseline of the unfair logistic regression model for how much of an increase in fairness?"

For demographic parity on the Adult dataset, our *Fair Log-loss* approach outperforms all baseline methods on average for both test error rate and for fairness violation, and on COMPAS dataset it achieves the lowest ratio of increased fairness over increased error. Additionally, the increase in test error over the unfair unconstrained logistic regression model is small. For demographic parity on the Law dataset, the relationship between methods is not as clear, but our *Fair Log-loss* approach still resides in the Pareto optimal set, i.e., there are no other methods that are significantly better than our result on both criteria. For equalized odds, *Fair Log-loss*

provides the lowest ratios of increased fairness over increased error rate for the Adult and COMPAS datasets, and competitive performance on the Law dataset. The post-processing method provides comparable or better fairness at the cost of significantly higher error rates. This shows that the approximation in our prediction procedure does not significantly impact the performance of our method. In terms of the running time, our method is an order of magnitude faster than comparable methods (e.g., the train and test running time on one random split of the Adult dataset takes approximately 5 seconds by our algorithm, 80 seconds for the constraint-based method (Zafar et al., 2017), and 100 seconds for the reduction-based method (Agarwal et al., 2018)).

3.2 Robust Fairness under Covariate Shift

Though many definitions and measures of (un)fairness have been proposed (See (Verma and Rubin, 2018; Mehrabi et al., 2019)), the most widely adopted are group fairness measures of *demographic parity* (Calders et al., 2009), *equalized opportunity*, and *equalized odds* (Hardt et al., 2016). Techniques have been developed as either post-processing steps (Hardt et al., 2016) or in-processing learning methods (Agarwal et al., 2018; Zafar et al., 2017a; Rezaei et al., 2020) seeking to achieve fairness according to these group fairness definitions. These methods attempt to make fair predictions at testing time by strongly assuming that training and testing data are *independently and identically drawn (iid)* from the same distribution, so that providing fairness on the training dataset provides approximate fairness on the testing dataset.

In practice, it is common for data distributions to *shift* between the training data set (*source distribution*) and the testing data set (*target distribution*). For example, the characteristics of loan applicants may differ significantly over time due to macroeconomic trends or changes in the self-selection criteria that potential applicants employ.

Figure 9 illustrates the declining performance of a post-processing method (Hardt et al., 2016) and an in-processing method (Rezaei et al., 2020) that do not consider distribution shift and instead only depend on source fairness measurements. Therefore, relying on the iid assumption, which is often violated in practice, introduces significant limitations for realizing desired fairness in critical applications.

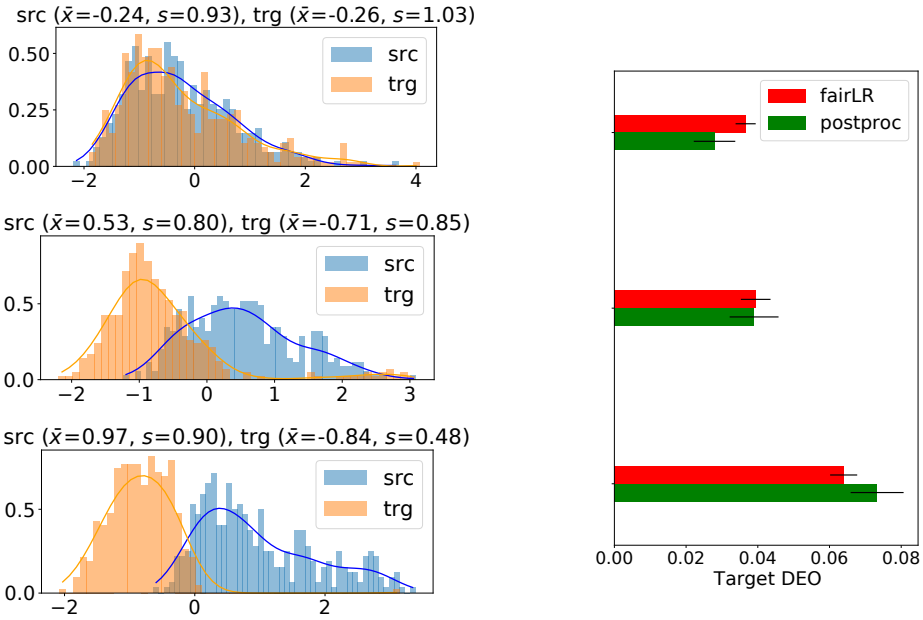


Figure 9: The difference of equalized opportunity between two genders in German UCI dataset evaluated on the target distribution for (a) the post-processing method of (Hardt et al., 2016) and (b) an in-processing fairLR method (Rezaei et al., 2020), that both do not account for distribution shift, and correct true positive rate parity on source data. The histograms on the left show the corresponding distribution shift on first principal component of the covariates between source and target data. The shift intensity has an overall increasing effect on DEO violation of both methods.

We seek to address the task of providing fairness guarantees under the non-iid assumption of covariate shift. Covariate shift is a special case of data distribution shift. It assumes that the relationship between labels and covariates (inputs) is the same for both distributions, while only the source and target covariate distributions differ.

In this work, we propose a robust estimation approach for constructing a fair predictor under covariate shift.

Our model builds on robust classification method of (Liu and Ziebart, 2014) under covariate shift, where the target distribution is estimated by a worst-case adversary that maximizes the log-loss while matching the feature statistics under source distribution. Therefore, if we know the separating set of features, we can incorporate them as constraints for the adversary. However, it is usually difficult to know the exact causal model of the data generating process in practice.

3.2.1 Approach

3.2.1.1 Preliminaries & Notation

We assume a binary classification task $Y, \hat{Y} \in \{0, 1\}$, where Y denotes the true label, and $\hat{Y}$ denotes the prediction for a given instance with features $\mathbf{X} \in \mathcal{X}$ and group attribute $A \in \{0, 1\}$. We consider $y = 1$ as the privileged class (e.g., an applicant who would repay a loan). Further, we assume a given source distribution $(\mathbf{X}, A, Y) \sim P_{\text{src}}$ over features, attribute, and label, and a target distribution $(\mathbf{X}, A) \sim P_{\text{trg}}$ over features and attribute only, throughout our paper.

3.2.1.2 Covariate Shift

In the context of fair prediction, the covariate shift assumption is that the distribution of covariates and group membership can shift between source and target distributions:

$$P_{\text{src}}(\mathbf{x}, a, y) = P_{\text{src}}(\mathbf{x}, a) P(y|\mathbf{x}, a) \tag{3.15}$$

$$P_{\text{trg}}(\mathbf{x}, a, y) = P_{\text{trg}}(\mathbf{x}, a) P(y|\mathbf{x}, a). \tag{3.16}$$

Note that we do not assume how the sensitive group membership a is correlated with other features $\mathbf{x}$.

3.2.1.3 Importance Weighting

A standard approach for addressing covariate shift is to reweight the source data to represent the target distribution (Sugiyama et al., 2007). A desired statistic $f(x, a, y)$ of the target distribution can be obtained using samples from the source distribution $(x_i, a_i, y_i)_{i=1:n}$:

$$\mathbb{E}_{\substack{P_{\text{trg}}(\mathbf{x},a) \\ P(y|\mathbf{x},a)}}\big[f(\mathbf{X}, A, Y)\big] \approx \sum_{i=1}^{n} \frac{P_{\text{trg}}(\mathbf{x}_i, a_i)}{P_{\text{src}}(\mathbf{x}_i, a_i)} f(\mathbf{x}_i, a_i, y_i). \tag{3.17}$$

As long as the source distribution has support for the entire target distribution (i.e., $P_{\text{trg}}(\mathbf{x}, a) > 0 \implies P_{\text{src}}(\mathbf{x}, a) > 0$), this approximation is exact asymptotically as $n \to \infty$. However, the approximation is only guaranteed to have bounded error for finite n if the source distribution's support for target distribution samples is lower bounded (Cortes et al., 2010): $\mathbb{E}_{P_{\text{trg}}(\mathbf{x},a)}\left[P_{\text{trg}}(\mathbf{X}, A)/P_{\text{src}}(\mathbf{X}, A)\right] < \infty$. Unfortunately, this requirement is fairly strict and will not be satisfied even under common and seemingly benign amounts of shift. For example, if source and target samples are drawn from Gaussian distributions with equal (co-)variance, but slightly different means, it is not satisfied.

3.2.1.4 Robust Log Loss Classification under Covaraite Shift

We base our method on the robust approach of (Liu and Ziebart, 2014) for covariate shift, which addresses this fragility of reweighting methods. In this formulation, the probabilistic predictor $\mathbb{P}$ minimizes the log loss on a worst-case approximation of the target distribution

provided by an adversary $\mathbb{Q}$ that maximizes the log loss while matching the feature statistics of the source distribution:

$$\begin{aligned}&\min_{\mathbb{P}(y|\mathbf{x})\in\Delta}\ \max_{\mathbb{Q}(y|\mathbf{x})\in\Delta\cap\Xi}\mathbb{E}_{P_{\text{trg}}(\mathbf{x})\mathbb{Q}(y|\mathbf{x})}[-\log\mathbb{P}(Y|\mathbf{X})]\\ =&\max_{\mathbb{P}(y|\mathbf{x})\in\Delta\cap\Xi}H_{P_{\text{trg}}(\mathbf{x})\mathbb{P}(y|\mathbf{x})}(Y|\mathbf{X}),\end{aligned}\tag{3.18}$$

where a moment-matching constraint set $\Xi = \{\mathbb{Q} \,|\, \mathbb{E}_{P_{\text{src}}(\mathbf{x})\mathbb{Q}(y|\mathbf{x})}[\phi(\mathbf{X}, Y)] = \mathbb{E}_{P_{\text{src}}(\mathbf{x},y)}[\phi(\mathbf{X}, Y)]\}$ on source data is enforced with $\phi(\mathbf{x}, y)$ denoting the feature function, and Δ denoting the conditional probability simplex. A first-order moments feature function, $\phi(\mathbf{x}, y) = [x_1 y, x_2 y, \dots x_m y]^\top$, is typical but higher-order moments, e.g., $yx_1, yx_2^2, yx_3^n, \dots$ or mixed moments, e.g., $yx_1, yx_1x_2, yx_1^2x_2x_3, \dots$, can be included. The saddle point solution under these assumptions is $\mathbb{P} = \mathbb{Q}$ which reduces the formulation to maximizing the target distribution conditional entropy (H) while matching feature statistics of the source distribution. The probabilistic predictor of (Liu and Ziebart, 2014) reduces to the following parametric form:

$$\mathbb{P}_\theta(y|\mathbf{x}) = e^{\frac{P_{\text{src}}(\mathbf{x})}{P_{\text{trg}}(\mathbf{x})}\theta^\top\phi(\mathbf{x},y)} \Big/ \sum_{y'\in\mathcal{Y}} e^{\frac{P_{\text{src}}(\mathbf{x})}{P_{\text{trg}}(\mathbf{x})}\theta^\top\phi(\mathbf{x},y')},\tag{3.19}$$

where the Lagrange multipliers θ maximize the target distribution log likelihood in the dual optimization problem.

3.2.2 Formulation

Our formulation seeks a robust and fair predictor under the covariate shift assumption by playing a minimax game between a minimizing predictor and a worst-case approximator of

the target distribution that matches the feature statistics from the source and marginals of the groups from target. We assume the availability of a set of labeled examples $\{\mathbf{x}_i, a_i, y_i\}_{i=1}^n$ sampled from the source $P_{\text{src}}(\mathbf{x}, a, y)$ and unlabeled examples $\{\mathbf{x}_i, a_i\}_{i=1}^m$ sampled from target distribution $P_{\text{trg}}(\mathbf{x}, a)$ during training.

Definition 2. *The* **Fair Robust Log-Loss Predictor under Covariate Shift**, $\mathbb{P}$ *minimizes the worst-case expected log loss with an μ-weighted expected fairness penalty on target, approximated by adversary $\mathbb{Q}$ constrained to match source distribution statistics (denoted by set Ξ) and group marginals on target (Γ):*

$$\min_{\mathbb{P}\in\Delta} \max_{\mathbb{Q}\in\Delta\cap\Xi\cap\Gamma} \mathbb{E}_{P_{trg}(\mathbf{x},a)\mathbb{Q}(y|\mathbf{x},a)}[-\log \mathbb{P}(Y|\mathbf{X}, A)] \tag{3.20}$$
$$+\mu\, \mathbb{E}_{P_{trg}(\mathbf{x},a)\mathbb{Q}(y'|\mathbf{x},a)\mathbb{P}(y|\mathbf{x},a)}[f(A, Y', Y)]$$

such that:

$$\Xi(\mathbb{Q}) : \mathbb{E}_{\substack{P_{src}(\mathbf{x},a)\\ \mathbb{Q}(y|\mathbf{x},a)}}[\phi(\mathbf{X}, Y)] = \mathbb{E}_{P_{src}(\mathbf{x},a,y)}[\phi(\mathbf{X}, Y)] \text{ and}$$

$$\Gamma(\mathbb{Q}) : \mathbb{E}_{\substack{P_{trg}(\mathbf{x},a)\\ \mathbb{Q}(y|\mathbf{x},a)}}[g_k(A, Y)] = \underbrace{\mathbb{E}_{\substack{P_{trg}(\mathbf{x},a)\\ \tilde{P}_{trg}(y|\mathbf{x},a)}}[g_k(A, Y)]}_{\tilde{g}_k},$$

$\forall k \in \{0, 1\}$, *where ϕ is the feature function, μ is the fairness penalty weight, $g_k(.,.)$ is a selector function for group k according to the fairness definition, i.e., for equalized opportunity:*

$g_k(A,Y) = \mathbb{I}(A{=}k \wedge Y{=}1)$, $\tilde{g}_k$ *the estimated group density on target, and* $f(.,.,.)$ *is a weighting function of the mean score difference between the two groups:*

$$f(A,Y,\hat{Y}) = \begin{cases} \frac{1}{\tilde{g}_1} & \text{if } g_1(A,Y) \wedge \mathbb{I}(\hat{Y}{=}1) \\ -\frac{1}{\tilde{g}_0} & \text{if } g_0(A,Y) \wedge \mathbb{I}(\hat{Y}{=}1) \\ 0 & \text{otherwise.} \end{cases} \tag{3.21}$$

The Γ constraint enforces $\mathbb{Q}$ to be consistent with the marginal probability of the groups on target ($\tilde{g}_k$) for equalized opportunity. This marginal probability is unknown, since the true label Y on target is unavailable. Thus, we estimate these marginal probabilities by employing the robust model (Equation 3.19) as $\tilde{P}_{\text{trg}}(y|\mathbf{x},a)$ in Γ in (Equation 3.20) to first guess the labels under covariate shift ignoring fairness ($\mu = 0$). We penalize the expected difference in true positive rate of groups in target according to our worst-case approximation of each example being positively labeled. This needs to be measured on the entire target example set and requires batch gradient updates to enforce.

Theorem 3.2.1. *Given binary class labels and protected attributes $y, a \in \{0, 1\}$, the fair probabilistic classifier for equalized opportunity robust under covariate shift as defined in* (Equation 3.20) *can be obtained by solving:*

$$\begin{aligned}
&\log \frac{1 - \mathbb{P}(y = 1|\mathbf{x}, a)}{\mathbb{P}(y = 1|\mathbf{x}, a)} + \mu\, \mathbb{E}_{\mathbb{P}(y'|\mathbf{x},a)} \left[f(a, y = 1, Y') \right] \\
&\quad + \frac{P_{src}(\mathbf{x}, a)}{P_{trg}(\mathbf{x}, a)} \theta^{\mathrm{T}} \left(\phi(\mathbf{x}, y = 1) - \phi(\mathbf{x}, y = 0) \right) \\
&\quad + \sum_{k \in \{0,1\}} \lambda_k g_k(a, y = 1) = 0,
\end{aligned} \tag{3.22}$$

where θ and λ are the dual Lagrange multipliers for source feature matching constraints (Ξ) and target group marginal matching (Γ) respectively, and μ is the penalty weight chosen to minimize the expected fairness violation on target.

Given the solution $\mathbb{P}^*$ obtained above, for $\mathbb{Q}$ to be in equilibrium (given θ and λ) it suffices to choose $\mathbb{Q}$ for $y = 1$ such that: $\mathbb{Q}(y|\mathbf{x}, a) =$

$$\frac{\mathbb{P}^*(y|\mathbf{x}, a)}{1 - \mu f(a, y, y)\mathbb{P}^*(y|\mathbf{x}, a) + \mu f(a, y, y)\mathbb{P}^{*^2}(y|\mathbf{x}, a)}, \tag{3.23}$$

where additionally it must hold that $0 \leq \mathbb{Q}(y|x, a) \leq 1$:

$$\implies \begin{cases} 0 \leq \mathbb{P}(y = 1|\mathbf{x}, a) \leq \frac{1}{\mu f(a,1,1)} & \text{if } \mu f(a, 1, 1) > 1 \\ 0 \leq \mathbb{P}(y = 1|\mathbf{x}, a) \leq 1 & \text{otherwise.} \end{cases}$$

Due to monotonicity, $\mathbb{P}$ in (Equation 3.22) is efficiently found using a binary-search in the simplex.

3.2.3 Experiments

We demonstrate the effectiveness of our method on biased samplings from four benchmark datasets:

- The `COMPAS` criminal recidivism risk assessment dataset (Larson et al., 2016).
- UCI `German` dataset (Dheeru and Karra Taniskidou, 2017).
- UCI `Drug` dataset (Fehrman et al., 2017).
- UCI `Arrhythmia` dataset (Dheeru and Karra Taniskidou, 2017).

3.2.3.1 Biased Sampling:

We model a general shift in the distribution of covariates between source and target, i.e., $P_{\text{src}}(\mathbf{x}, a) \neq P_{\text{trg}}(\mathbf{x}, a)$, by creating biased sampling based on the principal components of the covariates. We follow the previous literature on covariate shift (Gretton et al., 2009) and take the following steps to create the covariate shift on each dataset: We normalized all non-categorical features by z-score. We retrieve the first principal component $\mathcal{C}$ of covariates $(\mathbf{x}, a)$ by applying principal component analysis (PCA). We then estimate the mean $\mu(\mathcal{C})$ and standard deviation $\sigma(\mathcal{C})$, and set a Gaussian distribution $D_t(\mu(\mathcal{C}), \sigma(\mathcal{C}))$ for random sampling of target. We choose parameters α, β and set another Gaussian distribution $D_s(\mu(\mathcal{C}) + \alpha, \frac{\sigma(\mathcal{C})}{\beta})$ for random sampling of source data. We fix the sample size for both source and target to 40% of the original dataset; and construct the source data by sampling without replacement in proportion to D_s, and the target data by sampling without replacement from the remaining data in proportion to D_t.

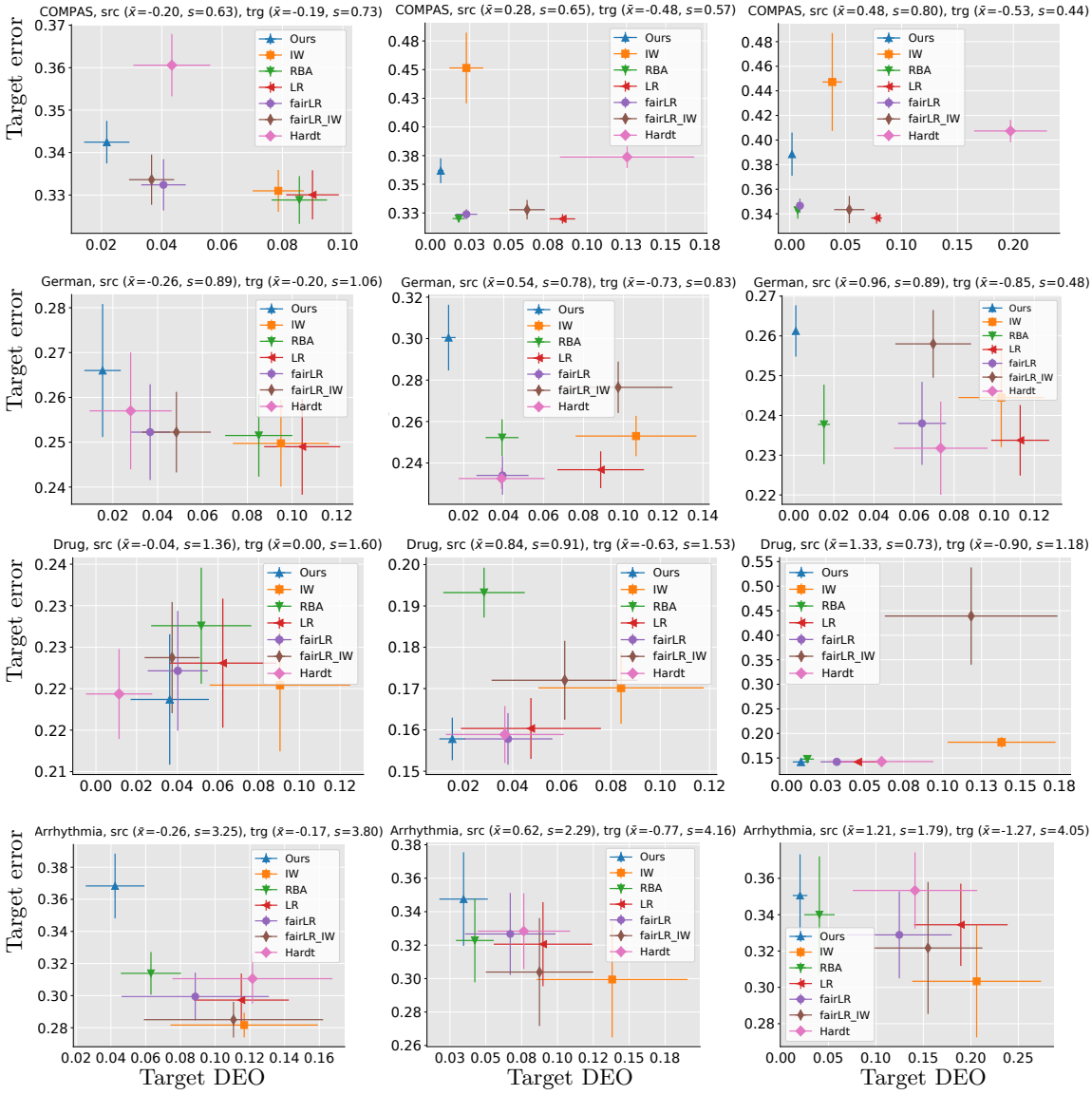


Figure 10: Average *prediction error* versus average *difference of equalized opportunity* (DEO) on target samples. The bar is the 95% confidence interval on ten random biased samplings on the first principal component of the covariates ($P_{\text{src}}(x, a) \neq P_{\text{trg}}(x, a)$).

3.2.3.2 Baseline methods

We evaluate the performance of our model in terms of the trade-off between prediction error and fairness violation on target under various intensities of covariate shift. We focus on equalized opportunity as our fairness definition. We compare against the following baselines:

- **Logistic Regression** (LR) is the standard logistic regression predictor trained on source data, ignoring both covariate shift and desired fairness properties.
- **Robust Bias-Aware Log Loss Classifier** (RBA) (Liu and Ziebart, 2014) in (Equation 3.18) which accounts for the covariate shift but ignores fairness.
- **Sample Re-weighted Logistic Regression** (LR_IW) (Shimodaira, 2000) minimizes the re-weighted log loss on the source data, according to the importance weighting scheme (Equation 3.17): it only accounts for the covariate shift.
- **Post Processing**[1] (HARDT) transforms the logistic regression target output to adjust for true positive rate parity (Hardt et al., 2016); ignores covariate shift.
- **Fair Logistic Regression** (FAIRLR) also optimizes worst-case log loss subject to fairness as linear constraints with observed labels on source data (Rezaei et al., 2020). It accounts for fairness, but ignores the covariate shift.

---

[1]We use the implementation from `https://fairlearn.github.io`.

- **Sample Re-weighted Fair Logistic Regression** (FAIRLR_IW) the fairLR method augmented with importance weighting scheme (Equation 3.17) in training. This baseline account for both fairness and covariate shift.

3.2.3.3 Results

Figure 10 shows our experimental results on three samplings from close to IID (left) to mild (middle) and strong covariate shift (right). Figure 9 provides an example of these samplings on `German`. On the `COMPAS` dataset, our method consistently achieves the lowest DEO while incurring higher loss compared to RBA and FAIRLR. In contrast, the HARDT method's difference of equalized opportunity (DEO) increases with the increasing shift. The optimal μ lies consistently close to zero on larger shifts on this dataset, which explains why RBA and FAIRLR are also very close to our method, indicating that the created shift was positively correlated to fairness. On the `German` dataset, our method provides the lowest and closest to zero average DEO on all shifted samplings, with competitive prediction error compared to other baselines. As the shift intensifies, DEO violation increases for other baselines (except RBA), which shows the negative effect of covariate shift on fairness for this dataset. On the `Drug` dataset, our method incurs higher DEO compared to HARDT's for the IID setting. However, as the shift intensifies, our method achieves the lowest DEO and lowest prediction error. The samples on `Arrhythmia` are much smaller and have relative larger standard deviation of covariates. On this dataset, our method achieves the lowest fairness violation at the cost of incurring slightly higher error compared to RBA and other baselines.

In summary, our method achieves the lowest DEO on 11 out of 12 samplings in our experiments. The prediction error of our method on `Drug` also remains the lowest, while remaining competitive on rest of the datasets.

3.2.4 Conclusions

In this work, we developed a novel adversarial approach for seeking fair decision making under covariate shift. In contrast with importance weighting methods, our approach is designed to operate appropriately even when portions of the shift between source and target distributions are extreme. The key technical challenge we address is the lack of labeled target data points, making target fairness assessment challenging. We instead propose to measure approximated fairness against an worst-case adversary that is constrained by source data properties and group marginals from target. We incorporate fairness as a weighted penalty and tune the weighted penalty to provide fairness against the adversary. More extensive evaluation on naturally-biased datasets and generalization of this approach to decision problems beyond binary classification are both important future directions.

CHAPTER 4

Fairness for Robust Learning To Rank

(This chapter is based on a paper published as "Fairness for Robust Learning To Rank" (Memarrast et al., 2023a) in The Pacific-Asia Conference on Knowledge Discovery and Data Mining 27 (PAKDD 2023).)

4.1 Introduction

Searching for relevant information through large amounts of data is a ubiquitous computing task. Applications include ordering search results (e.g., Google, Bing, or Baidu), personalizing social networks (e.g., Facebook, Instagram or Twitter), product recommendations for e-commerce sites (e.g., Amazon or eBay), and content recommendation for news/media sites (e.g., YouTube

or Netflix). Ranking a subset of items is a crucial component in these applications to help users find relevant information quickly amongst vast amounts of data.

Rankings often have social implications beyond the immediate utility they provide, since higher rankings provide opportunities for individuals and groups associated with the ranked items. As a consequence, biases in ranking systems—whether intentional or not—raise ethical concerns about their long-term economic and societal harming effect (Noble, 2018). Rankings that solely maximize utility or relevance can perpetuate existing societal biases that exist in training data whilst remaining oblivious to the societal detriment they cause by amplifying such biases (O'Neil, 2016).

Conventional ranking algorithms typically produce rankings to best serve the interests of those conducting searches by ordering the items by the probability of relevance so that utility to the users will be maximized (Robertson, 1977). Users are fulfilled, yet being oblivious to certain attributes of items to be ranked can have a harmful effect on minority groups in the items. Consequently, this could lead to further disparities, particularly for socially salient sub-populations, due to historic and current discriminatory practices which have introduced biases into data-driven models (Friedman and Nissenbaum, 1996). Biased outcomes drawn by these models negatively impact items in marginalized protected groups in critical decision making systems such as hiring or housing where items compete for exposure. Being unfair towards one group can lead to winner-takes-all dynamics that reinforce existing disparities (Singh and Joachims, 2018). Protected group definitions vary between different applications, and can include characteristics such as race, gender, religion, etc. In group fairness, algorithms divide

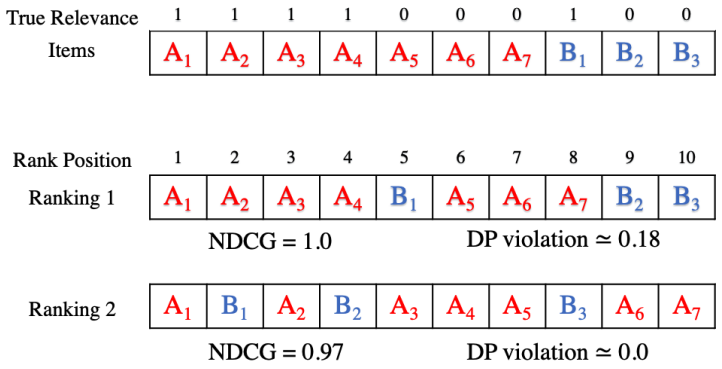


Figure 11: Ranking 1 ignores fairness whereas Ranking 2 satisfies the demographic parity fairness constraint while only slightly decreasing the utility.

the population into groups based on the protected attribute and guarantee the same treatment for members across groups. In ranking, this treatment can be evaluated using statistical metrics defined for measuring fairness. In this paper, we focus primarily on exposure-based group fairness measures. As a notable example, *demographic parity* (DP) in ranking is satisfied if the average exposure for both groups is equal in the top k ranks.

As a motivating example, in Figure 11 we consider two rankings based on items' true relevance and group membership. As a result of ranking 1, the highest utility is achieved, and fairness is ignored. In contrast, ranking 2 satisfies the demographic parity fairness constraint while still preserving high utility.

Fair ranking approaches seeking to provide group fairness properties can be categorized into *post-processing* and *in-processing* methods. *Post-processing* techniques are used to re-rank a given high utility ranking to incorporate fairness constraints while seeking to retain high utility

(Singh and Joachims, 2018; Biega et al., 2018). These methods assume that true relevance labels are available and require other fairness-unaware learning methods (e.g., regression) to predict the true labels as a pre-processing step. Recovering from unfair regression based rankings in the re-ranking step may not be feasible in some circumstances (Yadav et al., 2021).

The fair ranking problem can also be addressed as an *in-processing*, learning-to-rank (LTR) task where the algorithm learns to maximize utility subject to fairness constraints from training data. Our algorithm falls into this category. As a notable fair LTR techinque, DELTR (Zehlike and Castillo, 2020) optimizes a weighted summation of a loss function and a fairness criterion. This algorithm is constrained in how it measures fairness: it only considers the top-1 place in each ranking but not how additional items are ranked. Fair-PG-LTR (Singh and Joachims, 2019), another fair LTR method, prioritizes utility and fairness simultaneously for the full ranking by making use of a policy gradient optimization algorithm. While providing a fairness-utility trade-off, fair LTR approaches need to be robust to outliers and noisy data. For example, the label of recidivism in the COMPAS dataset is regarded to be noisy (Eckhouse, 2017). This makes prediction while incorporating fairness constraints more difficult. With improved robustness properties, a fair LTR can achieve better utility for highly fair rankings, which results in a preferable utility-fairness trade-off.

In this work, we derive a new LTR system based on the first principles of distributional robustness to provide both fairness and robustness to label noise. We formulate a minimax game with the *ranker player* choosing a distribution over rankings constrained to satisfy fairness requirements on the training samples while maximizing utility, and an *adversary player* choosing

a distribution of item relevancies that minimizes utility while being similar to training data properties. Rather than narrowly optimizing the rankings for the specific training data, this approach produces rankings that provide utility and fairness robustly for a family of distributions that includes the training data. We show that this approach is flexible enough to implement a wide range of fairness constraints and that it can be extended to accept generic utility values. To compare our proposed framework to existing fair LTR solutions, we perform empirical evaluations on simulated and real-world datasets that demonstrate the effectiveness and validity of the resulting algorithm. We show that our approach is able to trade-off between utility and fairness much better at high levels of fairness than existing baseline methods. Furthermore, the robustness properties of our approach enable it to outperform existing baselines in the presence of varying degrees of label noise in the training data. To the best of our knowledge, this is the first distributionally robust fair LTR method.

4.2 Learning Fair Robust Ranking

4.2.1 Probabilistic Ranking

To formulate the ranking task, we consider a dataset of ranking problems $\mathcal{D} = \{\mathcal{R}^i\}_{i=1}^N$ for N different queries, where each $\mathcal{R}^i = \{d_j\}_{j=1}^M$ is a candidate item set of size M for a single query. For every item d_j in this set, we denote $rel(d_j)$ as its corresponding relevance judgment. We denote the utility of a ranking (permutation) π for a single query as $\text{Util}(\pi)$. The optimization problem can be written as: $\pi^* = \text{argmax}_{\pi \in \Pi_{\text{fair}}} \text{Util}(\pi)$. Utility measures used for rankings are based on the relevance of the individual items being ranked for a particular

ranking problem, $\mathcal{R}_{|\text{query}} = \{d_j\}_{j=1}^{M}$. For example, the Discounted Cumulative Gain (DCG) (Järvelin and Kekäläinen, 2002), which is a common evaluation measure for ranking systems that discounts the utility for lower-ranked items,

$$DCG(\pi) = \sum_{d_j \in \mathcal{R}} \frac{2^{rel(d_j)} - 1}{\log(1 + \pi_j)} \Rightarrow \text{Util}(\pi) = \sum_{j=1}^{M} u_j v_{\pi_j}, \tag{4.1}$$

is a member of the more general family of linear utility functions where $u_j = 2^{rel(d_j)} - 1$ representing the utility of a single item d_j based on its relevance $rel(.)$ and $v_k = \frac{1}{\log(1+k)}$ providing the degree of attention that item d_j receives by being placed at rank k by permutation π, i.e., $\pi_{d_j} = k$.

The space of all permutations of items is exponential in the number of items, making naïve methods that find a utility-maximizing ranking subject to fairness constraints intractable. To overcome this problem, we consider a probabilistic ranking in which instead of a single ranking, a distribution over rankings is used. We define the probability of positioning item d_j at rank k as $P_{j,k}$. Then $\mathbf{P}$ constructs a doubly stochastic matrix of size $M \times M$ where entries in each row and each column must sum up to 1. By employing the idea of probabilistic ranking, we express the ranking utility in (Equation 4.1) as an expected utility of a probabilistic ranking:

$$\text{U}(\mathbf{P}) = \sum_{j=1}^{M} \sum_{k=1}^{M} \mathbf{P}_{j,k}\, u_j\, v_k = \mathbf{u}^T \mathbf{P} \mathbf{v}, \tag{4.2}$$

which we equivalently express in a vectorized format where $\mathbf{u}$ and $\mathbf{v}$ are both column vectors of size M. Following (Singh and Joachims, 2018), the fair ranking optimization can be expressed as a linear programming problem:

$$\max_{\mathbf{P} \in \Delta \cap \Gamma_{\text{fair}}} \mathbf{u}^T \mathbf{P} \mathbf{v} \tag{4.3}$$

$$\text{where:} \quad \Delta : \mathbf{P}\mathbf{1} = \mathbf{P}^\top \mathbf{1} = \mathbf{1}, \;\; \mathbf{P}_{j,k} \geq 0, \;\; \forall_{1 \leq j,k \leq M}$$

and Γ_{fair} denotes any linear constraint set of the form $\mathbf{f}^\top \mathbf{P} \mathbf{g} = h$. Choosing $\mathbf{f}$ as the utility of items according to groups and $\mathbf{g}$ as the exposure of ranking position enforces equality of exposure across protected groups. In contrast to (Singh and Joachims, 2018), which uses this framework to re-rank the items to satisfy fairness constraints (i.e., a post-processing method), we extend this linear perspective to derive a *learning-to-rank* approach that learns to optimize utility and fairness simultaneously during training (i.e., an in-processing method).

4.2.2 Learning to Rank using an Adversarial Approach

We adopt a distributionally robust approach to the LTR problem by constructing a worst-case adversarial distribution on item utilities. We formulate the robust fair ranking construction as a minimax game between two players: a fair predictor $\mathbf{P}$ that makes a probabilistic prediction over the set of all possible rankings to maximize expected ranking utility; and an adversary $\mathbf{Q} : Q_{j,k'} \triangleq P(u_j = 1, \sum_{d_i \in G_{d_j}} u_i = k')$ that approximates a probability distribution for the utility of items which minimizes the expected ranking utility. The adversary is additionally constrained to match the feature moments of the empirical training distribution. Since we solve

the problem for a given query, the query-dependent terms are omitted from the formulation for simplicity.

In our notation, we represent ranking items d by their feature representation $\mathbf{X} \in \mathbb{R}^{M \times L}$ as a matrix of M items with L features. For a given item set $\mathbf{X}$, the expected ranking utility of a probabilistic ranking $\mathbf{P}$ against a utility distribution $\mathbf{Q}$ can be expressed as:

$$\mathrm{U}(\mathbf{X}, \mathbf{P}, \mathbf{Q}) = \sum_{j=1}^{M} \mathbb{E}_{u_j|\mathbf{X}\sim\mathbf{Q}} \left[u_j \mathbb{E}_{\pi_j|\mathbf{X}\sim\mathbf{P}} \left[v_{\pi_j} \right] \right]. \tag{4.4}$$

Then, the utility-maximizing optimization problem under fairness constraints can be formulated as:

Definition 4.2.1. *Given a training dataset of N ranking problems $\mathcal{D} = \{(\mathbf{X}^i, \mathbf{u}^i)\}_{i=1}^N$, with $\mathbf{u} \in \mathbb{R}^M$ being the true relevance and $\mathbf{X} \in \mathbb{R}^{M \times L}$ the feature representation of ranking problem of size M. The fair probabilistic ranking $\mathbf{P}(\pi) \in \mathbb{R}^{M \times M}$ in adversarial learning-to-rank learns a fair ranking that maximizes the worst-case ranking utility approximated by an adversary $\mathbf{Q}(\check{\mathbf{u}})$, constrained to match the feature statistics of the training data.*

$$\begin{aligned} &\max_{\mathbf{P}(\pi|\mathbf{X})\in\Delta} \min_{\mathbf{Q}(\check{\mathbf{u}}|\mathbf{X})} \mathbb{E}_{\mathbf{X}\sim\widetilde{P}} \left[\mathrm{U}(\mathbf{X}, \mathbf{P}, \mathbf{Q}) \right] + \lambda \mathit{unfair}(\mathbf{P}, \mathbf{Q}) \\ &s.t.\ \mathbb{E}_{\mathbf{X}\sim\tilde{P}} \left[\sum_{j=1}^{M} \mathbb{E}_{\check{u}_j|\mathbf{X}\sim\mathbf{Q}} \left[\check{u}_j \mathbf{X}_{j,:} \right] \right] = \mathbb{E}_{\mathbf{X},\mathbf{u}\sim\widetilde{P}} \left[\sum_{j=1}^{M} u_j \mathbf{X}_{j,:} \right] \end{aligned} \tag{4.5}$$

where $\widetilde{P}$ denotes the empirical distribution over ranking dataset $\mathcal{D} = \{(\mathbf{X}^i, \mathbf{u}^i)\}_{i=1}^N$, $\check{\mathbf{u}}$ denotes the random variable for adversary relevance, and Δ denotes the set of doubly stochastic matrices.

This general adversarial formulation plays a foundational role in constructing probability models and prediction techniques (Grünwald and Dawid, 2004; Farnia and Tse, 2016; Fathony et al., 2016; Fathony et al., 2017). Specifically, when the game is played using the logarithmic loss, distributions from the exponential family and logistic regression are produced (Topsøe, 1979). Extensions to structured prediction, including learning permutations/matchings/rankings provide computational efficiency and Bayes optimal consistency (Fathony et al., 2018). This approach has been utilized to provide fair and robust predictions under covariate shift (Rezaei et al., 2021) as well as for constructing reliable predictors for fair log loss classification (Rezaei et al., 2020). Similar to this line of work, our proposed approach imposes fairness constraints on predictor $\mathbf{P}$. However, we avoid exponential distributions for rankings because exactly normalizing the distribution over $M!$ rankings require a #P-hard matrix permanent calculation (Petterson et al., 2009b).

Our formulation in Definition 4.2.1 accepts generic utility values. In our work, we focus on binary utility, which is one of the common applications of the ranking problem, where the utility label indicates if a particular item is relevant or not. However, as described in the appendix, the extension of the method to other settings is easy.

For the binary utility problem, the expected utility can be further simplified as:

$$\begin{aligned}
\mathrm{U}(\mathbf{X},\mathbf{P},\mathbf{Q}) &= \sum_{j=1}^{M} \mathbb{E}_{u_j|\mathbf{X}\sim\mathbf{Q}}\left[u_j \mathbb{E}_{\pi_j|\mathbf{X}\sim\mathbf{P}}\left[v_{\pi_j}\right]\right] \\
&= \sum_{j=1}^{M}\sum_{k=1}^{M}\left(\sum_{k'=1}^{M}\mathbf{Q}(u_j=1,\sum_{d_i\in G_{d_j}} u_i = k'|\mathbf{X})\right)\mathbf{P}(\pi_j=k|\mathbf{X})v_k \\
&= (\mathbf{Q1})^\top \mathbf{P}\mathbf{v} = \mathbf{q}^\top\mathbf{P}\mathbf{v}, \qquad (4.6)
\end{aligned}$$

where the entries in the vector $\mathbf{q}$ contains the relevance probability of item d_j. In the following sections, we use this vector notation to simplify the optimization formulation.

4.2.3 Fairness of Exposure in Ranking

Our approach is flexible enough to implement a wide range of fairness constraints. In general, fairness criteria defined bilinearly in $\mathbf{P}$ and $\mathbf{Q}$ can be applied (separately) to the optimization framework. In this work, we focus on exposure-based fairness constraints such as demographic parity, disparate treatment and disparate impact (Singh and Joachims, 2018).

Demographic parity only need to be expressed linearly in $\mathbf{P}$. For a set of disjoint group members $G_1,\ldots,G_{|S|}$, the constraint requires $\frac{1}{|G_s|}\sum_{d_j\in G_s}\sum_{k=1}^{M}\mathbf{P}_{j,k}v_k = \frac{1}{|G_{s'}|}\sum_{d_j\in G_{s'}}\sum_{k=1}^{M}\mathbf{P}_{j,k}v_k = \tau, \quad \forall s,s'\in S$. This can be compactly written in the vector form as $\mathbf{f}_s^\top\mathbf{P}\mathbf{v}=\tau, \quad s\in S$.

For disparate treatment and disparate impact fairness constraints, the sufficient statistics depend on the number of relevant items within each group. As shown in the appendix, these fairness constraints can be expressed bilinearly in $\mathbf{P}$ and $\mathbf{Q}$ variables, so they can be incorporated

in the optimization framework. For the rest of the chapter, we focus on demographic parity in our exposition and experiments.

4.3 Augmented marginal approach to fairness optimization

For some fairness definitions, the sufficient statistics depend on the number of relevant items within each group. We let $\mathbf{q}_k(x)$ denote the marginal probability of relevance level x for each item when the group to which that item belongs has k total relevant items:

$$\mathbf{q}_k(x) \triangleq \begin{bmatrix} P(u_1 = x, \sum_{d_j \in G(d_1)} u_j = k) \\ P(u_2 = x, \sum_{d_j \in G(d_2)} u_j = k) \\ \vdots \\ P(u_n = x, \sum_{d_j \in G(d_n)} u_j = k). \end{bmatrix} \tag{4.7}$$

For these sets of marginals to be valid, they must satisfy:

$$\Gamma \triangleq \forall \text{ groups } s, \text{ sizes } k, \sum_{d_i \in G_s} P\left(u_i = 1 \middle| \sum_{d_j \in G_s} u_j = k\right) = k \tag{4.8}$$

$$\iff \forall \text{ groups } s, \text{ sizes } k, \mathbf{q}_k(1)^{\mathrm{T}} \mathbf{1}_{d_j \in G_s} = k((\mathbf{q}_k(0) + \mathbf{q}_k(1))^{\mathrm{T}} \mathbf{1}_{d_j \in G_s}), \tag{4.9}$$

which are linear constraints on $\mathbf{q}_k$.

For notational convenience, we defined matrix $\mathbf{Q} \triangleq \left[\mathbf{q}_0(1)\ \mathbf{q}_1(1)\ \mathbf{q}_2(1) \ldots\right]$ by concatenating all relevant marginal probabilities. Each row in matrix $\mathbf{Q}$ consist of marginal probabilities of an item being relevant when its corresponding group has different total relevant items. Therefore, if we sum each row in $\mathbf{Q}$ we get the probability of the corresponding item being relevant which

is the same probability in vector $\mathbf{q}$ in (Equation 4.11). Hence, $\mathbf{q}$ can be inferred from $\mathbf{Q}$ using $\mathbf{q} = \mathbf{Q1}$.

The Disparate Treatment constraint enforces that exposure of two groups to be proportional to their average utility. The constraint for group s and s' can then be expressed as:

$$
\begin{aligned}
&\sum_k \left(\frac{1}{k} \left(\sum_{d_i \in G_s} \mathbf{Q}_{i,k} \right) \sum_j \mathbf{P}_{i,j} v_j \right) = \sum_k \left(\frac{1}{k} \left(\sum_{d_i \in G_{s'}} \mathbf{Q}_{i,k} \right) \sum_j \mathbf{P}_{i,j} v_j \right) \\
\iff &\sum_k \left(\frac{1}{k} \left(\sum_{d_i \in G_s} \mathbf{Q}_{i,k} - \sum_{d_i \in G_{s'}} \mathbf{Q}_{i,k} \right) \sum_j \mathbf{P}_{i,j} v_j \right) = 0 \\
\iff &\sum_k \left(\sum_i \frac{1}{k} \mathbf{Q}_{i,k} \underbrace{\left(\mathbf{1}_{d_i \in G_s} - \mathbf{1}_{d_i \in G_{s'}} \right)}_{\mathbf{f}} \sum_j \mathbf{P}_{i,j} v_j \right) = 0 \\
\iff &\underbrace{\left((\mathbf{Qg}) \circ \mathbf{f} \right)^T \mathbf{Pv}}_{\mathrm{L}^{\mathrm{DT}}_{\mathrm{s,s'}}(\mathbf{P},\mathbf{Q})} = 0
\end{aligned}
$$

which is bilinear in $\mathbf{Q}$ and $\mathbf{P}$ variables. Note that $\circ$ is the Hadamard-product (element-wise) operator. The vector $\mathbf{f}$ encodes the group membership: $\mathbf{f}_i = \begin{cases} 1 & \text{if } d_i \in G_s \\ -1 & \text{if } d_i \in G_{s'} \\ 0 & \text{otherwise.} \end{cases}$, and the vector $\mathbf{g}$ is the weight ($\frac{1}{k}$) used for each column in matrix $\mathbf{Q}$: $\mathbf{g}_k = \frac{1}{k}$.

The Disparate Impact constraint requires that the average "click-through rate" for two groups be proportional to their average utility. This constraint similarly for groups s and s' can be expressed as:

$$\sum_k \left(\frac{1}{k} \sum_{d_i \in G_s} \left(\mathbf{Q}_{i,k} \sum_j \mathbf{P}_{i,j} v_j \right) \right) = \sum_k \left(\frac{1}{k} \sum_{d_i \in G_{s'}} \left(\mathbf{Q}_{i,k} \sum_j \mathbf{P}_{i,j} v_j \right) \right)$$

$$\iff \sum_k \left(\frac{1}{k} \sum_{d_i \in G_s} \left(\mathbf{Q}_{i,k} \sum_j \mathbf{P}_{i,j} v_j \right) - \frac{1}{k} \sum_{d_i \in G_{s'}} \left(\mathbf{Q}_{i,k} \sum_j \mathbf{P}_{i,j} v_j \right) \right) = 0$$

$$\iff \left(\sum_i \left(\mathbf{1}_{d_i \in G_s} - \mathbf{1}_{d_i \in G_{s'}} \right) \left(\sum_k \frac{1}{k} \mathbf{Q}_{i,k} \sum_j \mathbf{P}_{i,j} v_j \right) \right) = 0$$

$$\iff \underbrace{\mathbf{f}^T \left((\mathbf{Q}\mathbf{g}) \circ \mathbf{P}\mathbf{v} \right)}_{\mathrm{L}^{\mathrm{DI}}_{\mathrm{s,s'}}(\mathbf{P},\mathbf{Q})} = 0$$

which is bilinear in $\mathbf{Q}$ and $\mathbf{P}$ variables.

The optimization in (Equation 4.11) can be modified to accept these fairness constraints. In the new optimization we substitute $\mathbf{q}$ with $\mathbf{Q1}$ and optimize over matrix $\mathbf{Q}$:

$$\max_{\theta} \ \mathbb{E}_{\mathbf{x},\mathbf{u} \sim \widetilde{P}} \left[\max_{\mathbf{P} \in \Delta} \min_{\mathbf{Q} \in \Gamma} (\mathbf{Q1})^\top \mathbf{P}\mathbf{v} + \left\langle \mathbf{Q1} - \mathbf{u}, \sum\nolimits_l \theta_l \mathbf{X}_{:,l} \right\rangle \right] \tag{4.10}$$

$$\text{s.t.} \quad \mathrm{L}_{\mathrm{s,s'}}(\mathbf{P}, \mathbf{Q}) = 0, \quad \forall s, s' \in S,$$

4.4 Optimization

We solve the constrained minimax formulation in Definition 4.2.1 in Lagrangian dual form, where we optimize the dual parameters $\boldsymbol{\theta} \in \mathbb{R}^{L\times 1}$ for the feature matching constraint of L features by gradient descent. Rewriting the optimization in matrix notation yields:

$$\max_{\theta}\ \mathbb{E}_{\mathbf{x},\mathbf{u}\sim\widetilde{P}}\left[\max_{\mathbf{P}\in\Delta}\min_{0\leq\mathbf{q}\leq 1}\ \mathbf{q}^{\top}\mathbf{P}\mathbf{v}+\left\langle \mathbf{q}-\mathbf{u},\sum\nolimits_{l}\theta_{l}\mathbf{X}_{:,l}\right\rangle\right] \tag{4.11}$$
$$\text{s.t.}\quad \mathbf{f}_{s}^{\top}\mathbf{P}\mathbf{v}=\tau,\quad s\in S,$$

where: $\mathbf{P}(\pi)\in\mathbb{R}^{M\times M}$ is a doubly stochastic matrix, and the value of cell $\mathbf{P}_{j,k}$ represents the probability that $\pi_j = k$; $\mathbf{u}\in\mathbb{R}^{M\times 1}$ is a vector of true labels whose j^{th} values is 1 when the item j is relevant to the query, i.e., $u_j = 1$ and 0 otherwise; $\mathbf{q}\in\mathbb{R}^{M\times 1}$ is a probability vector of the adversary's estimation of each item being relevant; $\mathbf{X}_{:,l}\in\mathbb{R}^{M\times 1}$ denotes the l^{th} feature of M samples; S is the set of protected attributes; and $\mathbf{v}\in\mathbb{R}^{M\times 1}$ is a vector containing the values of position bias function for each position. To denote the Frobenius inner product between two matrices $\langle .,.\rangle$ is used, i.e., $\langle A,B\rangle = \sum_{i,j} A_{i,j}B_{i,j}$.

For optimization purposes, using strong duality, we push the minimization over $\mathbf{q}$ to the outermost level in (Equation 4.11). Since the objective is non-smooth, for both $\mathbf{P}$ and $\mathbf{q}$, we add strongly convex prox-functions to make the objective smooth. Furthermore, to make our approach handle feature sampling error, we add a regularization penalty to the parameter $\boldsymbol{\theta}$. To

apply (Equation 4.11) on training data, we replace empirical expectation with an average over all training samples. The new formulation is as follows:

$$\min_{\{0\leq \mathbf{q}^i\leq 1\}_{i=1}^N} \max_{\theta} \frac{1}{N}\sum_{i=1}^{N} \max_{\mathbf{P}^i\in\Delta} \left[{\mathbf{q}^i}^\top \mathbf{P}^i\mathbf{v}^i - \left\langle \mathbf{q}^i - \mathbf{u}^i, \sum\nolimits_l \theta_l \mathbf{X}^i_{:,l} \right\rangle \right.$$
$$\left. +\lambda {\mathbf{f}^i}^\top \mathbf{P}^i\mathbf{v}^i - \frac{\mu}{2}\left\|\mathbf{P}^i\right\|_F^2 + \frac{\mu}{2}\left\|\mathbf{q}^i\right\|_2^2 \right] - \frac{\gamma}{2}\left\|\boldsymbol{\theta}\right\|_2^2, \tag{4.12}$$

where superscript i is the i^{th} sample from N ranking problems in the training set. We denote λ, γ and μ as the fairness penalty parameter (which can be adjusted to obtain different trade-offs between fairness and utility, rather than strictly optimized), a regularization penalty parameter and a smoothing penalty parameter, respectively. The inner maximization over $\mathbf{P}$ and $\boldsymbol{\theta}$ can be solved separately, given a fixed $\mathbf{q}$. The maximization over $\boldsymbol{\theta}$ has a closed-form solution where the l^{th} element of $\boldsymbol{\theta}^*$ is:

$$\theta_l^* = -\frac{1}{\gamma N}\sum_{i=1}^{N} \left\langle \mathbf{q}^i - \mathbf{u}^i, \mathbf{X}^i_{:,l} \right\rangle. \tag{4.13}$$

Independently from $\boldsymbol{\theta}$, we can solve the inner maximization over $\mathbf{P}$ for every training sample using a projection technique. The optimal $\mathbf{P}$ for i^{th} training sample (i.e., $\mathbf{P}^{i^*}$) is:

$$\mathbf{P}^{i^*} = \underset{\mathbf{P}^i\in\Delta}{\text{argmax}}\ {\mathbf{q}^i}^\top \mathbf{P}^i\mathbf{v}^i + \lambda {\mathbf{f}^i}^\top \mathbf{P}^i\mathbf{v}^i - \frac{\mu}{2}\left\|\mathbf{P}^i\right\|_F^2$$
$$\mathbf{P}^{i^*} = \underset{\mathbf{P}^i\in\Delta}{\text{argmin}}\ \frac{\mu}{2}\left\|\mathbf{P}^i - \frac{1}{\mu}(\mathbf{q}^i + \lambda\mathbf{f}^i){\mathbf{v}^i}^\top\right\|_F^2 - \frac{1}{2\mu}\left\|\mathbf{q}^i{\mathbf{v}^i}^\top\right\|_F^2. \tag{4.14}$$

As derived in (Equation 4.14), the minimization takes the form of $\min_{\mathbf{P}\geq 0} \|\mathbf{P}-\mathbf{R}\|_F^2$, s.t. : $\mathbf{P1} = \mathbf{P}^\top \mathbf{1} = \mathbf{1}$, and can be interpreted as projecting matrix $\frac{1}{\mu}(\mathbf{q}^i + \lambda \mathbf{f})\mathbf{v}^{i^\top}$ into the set of doubly-stochastic matrices. The projection from an arbitrary matrix $\mathbf{R}$ to the set of doubly-stochastic matrices can be solved using the ADMM projection algorithm (Boyd et al., 2011).

Since each entry in $\mathbf{q}$ represents a probability, the outer optimization over $\mathbf{q}$ is solved using the L-BFGS-B algorithm with a bounded constraint of the probability simplex (Byrd et al., 1995). The algorithm optimizes the quadratic approximation of the objective function (using limited memory Quasi-Newton) on the convex set with each iteration. In each update step, a projection to the probability simplex is needed. Based on the above optimization, the adversary's optimal relevance probability $\mathbf{q}^*$ can be obtained. Following (Equation 4.13) we compute the θ^* over the optimal $\mathbf{q}^*$. As weights for the features, we use the θ^* that our model learns from this optimization for making predictions for test examples. Algorithm 1 shows the steps for training.

Algorithm 1: The Fair-Robust LTR

Input: Training dataset $\mathcal{D} = \{(\mathbf{X}^i, \mathbf{u}^i)\}_{i=1}^N$, fairness penalty parameter λ.
Output: $\theta^*, \mathbf{P}^*, \mathbf{q}^*$
$\mathbf{q} \leftarrow$ random initialization;
repeat
  update θ by (Equation 4.13) with $\mathbf{q}$.
  update $\mathbf{P}$ by (Equation 4.14) with $\mathbf{q}$.
  update $\mathbf{q}$ by (Equation 4.12) with $\{\mathbf{P}, \theta\}$.
until *convergence*;

4.4.1 Inference and Runtime Analysis

For prediction, we use $\boldsymbol{\theta}$ and μ learned from training data while performing the optimization in (Equation 4.12). After removing the constant terms, we solve a similar optimization problem for test data. That is:

$$\min_{\{0\leq \mathbf{q}^i \leq 1\}_{i=1}^{N^{\text{test}}}} \frac{1}{N^{\text{test}}} \sum_{i=1}^{N^{\text{test}}} \max_{\mathbf{P}^i \in \Delta} \left[{\mathbf{q}^i}^\top \mathbf{P}^i \mathbf{v}^i - \left\langle \mathbf{q}^i, \sum\nolimits_l \theta_l^* \mathbf{X}_{:,l}^i \right\rangle \right.$$
$$\left. + \lambda {\mathbf{f}^i}^\top \mathbf{P}^i \mathbf{v}^i - \frac{\mu}{2} \left\| \mathbf{P}^i \right\|_F^2 + \frac{\mu}{2} \left\| \mathbf{q}^i \right\|_2^2 \right], \tag{4.15}$$

where superscript i pertains to the i^{th} ranking problem in the the test set of size N^{test}. We follow the steps for solving the optimization in training. Although we play a minimax game between predictor and adversary in inference, we emphasize that there is no gradient learning of θ as in training, and true relevance labels ($\mathbf{u}$) are not used in inference. After convergence, we use the resulting $\mathbf{P}^*$ from the optimization to predict the ranking of items in the test set. We employ the Hungarian algorithm (Kuhn, 1955) to solve the problem of matching items to positions. Algorithm 2 shows the steps for inference.

4.4.1.1 Runtime Analysis.

In the training step, solving the optimization in (Equation 4.12) involves running a projected gradient descent algorithm. In each iteration, it requires the computation of the gradient and the projection to box constraints. The box constraint projection's runtime is linear in terms of the number of variables, hence costing $\mathcal{O}(NM)$. The gradient computation requires solving for θ^*, which costs $\mathcal{O}(NML)$ from the dot product computations; and solving for $\mathbf{P}^*$, which

Algorithm 2: Inference Algorithm for Fair-Robust LTR

Input: Test dataset $\mathcal{D} = \{(\mathbf{X}^i, \mathbf{u}^i)\}_{i=1}^{N^{\text{test}}}$, fairness penalty parameter λ, dual parameter θ^*

Output: $\mathbf{P}^*, \mathbf{q}^*$, final_rankings = HungarianAlg($\mathbf{P}^*$)

$\mathbf{q} \leftarrow$ random initialization ;

repeat

  update $\mathbf{P}$ by (Equation 4.14) with $\mathbf{q}$.

  update $\mathbf{q}$ by (Equation 4.12) with $\{\mathbf{P}, \theta\}$.

until *convergence*;

can be posed as a doubly-stochastic matrix projection. We employ an ADMM algorithm to perform the projection to doubly stochastic matrix, which has linear convergence due to the strong convexity of the objective (Deng and Yin, 2016). Each step inside the ADMM consists of M projections to M-element simplex, hence costing $\mathcal{O}(M^2)$ computations in total for N ranking problems (Duchi et al., 2008). For the inference step, the Hungarian algorithm requires a cubic runtime, $\mathcal{O}(M^3)$. However, in the case of large M, given that exposure reduces with a logarithmic function and lower ranks have very similar exposure, we find the ranking of items using feature function ($\theta.\mathbf{X}$) and then run the Hungarian algorithm only for the top k ranked items to satisfy fairness constraint for those items ($k = 10$). This reduces the complexity of the algorithm from $\mathcal{O}(M^3)$ to $\mathcal{O}(k^3)$.

4.5 Experiments

In this section, we apply our fair adversarial ranking framework to the task of *learning-to-rank* under *group fairness* constraints. In order to compare our proposed framework with existing fair LTR solutions, we use simulated and real-world datasets to carry out in-depth empirical

evaluations. The learning task is to determine the feature function in the training based on the items' ground truth utilities and fairness constraints. At testing time, this feature function coupled with a penalty for fairness violation is used to determine the ranking for the items in the test set with maximum utility while satisfying fairness constraints.

4.5.1 Fairness Benchmark Datasets

4.5.1.1 Setup

We perform experiments on three benchmark datasets where we follow steps discussed in (Singh and Joachims, 2019) to adapt German, Adult, and COMPAS datasets to an LTR task. These datasets are inherently biased, making them viable alternatives for evaluation when no real world datasets exist for a fair LTR task. First, we split each dataset randomly into a disjoint train and test set. Then from each train/test set, we construct a corresponding LTR train/test set. For each query, we sample randomly with replacement a set of 10 candidates each, representative of both relevant and irrelevant items, where, on average four individuals are relevant. Each individual in the candidate set is a member of a group G_s based on its protected attribute. The training data consists of 500 ranking problems. We evaluate our learned model on 100 separate ranking problems serving as the test set. We repeat this process 10 times and report the 95% confidence interval in the results. The regularization constant γ and smoothing penalty parameter μ in (Equation 4.12) are chosen by 3-fold cross validation. We describe datasets used in our experiments:

- UCI `Adult`, census income dataset (Dheeru and Karra Taniskidou, 2017). The goal is to predict whether income is above \$50K/yr on the basis of census results.

TABLE III: Dataset characteristics.

Dataset	n	Features	Attribute
`Adult`	45,222	12	Gender
`COMPAS`	6,167	10	Race
`German`	1,000	20	Gender

- The `COMPAS` criminal recidivism risk assessment dataset (Larson et al., 2016) is designed to predict whether a defendant is likely to re-offend based on criminal history.
- UCI `German` dataset (Dheeru and Karra Taniskidou, 2017). Based on personal information and credit history, the goal is to classify good and bad credit.

Table III shows the statistics of each dataset with their protected attributes.

4.5.1.2 Baseline methods

To evaluate the performance of our model, we compare it against three different baselines that have similarities to and differences from our model: FAIR-PGRank (Singh and Joachims, 2019) and DELTR (Zehlike and Castillo, 2020) are in-processing, LTR methods, like ours; the Post-Processing method of (Singh and Joachims, 2018) employs the fairness constraint formulation that we build our optimization framework based on. We also add a Random baseline that ranks items in each query randomly to give context to NDCG. We discuss baseline methods in more details[1]:

[1]We use the implementation from `https://github.com/ashudeep/Fair-PGRank` for all baselines.

- **Post Processing** (Post-Proc) (Singh and Joachims, 2018) In order to make a fair comparison with in-processing LTR approaches, we first learn a linear regression model that is trained on all query-item sets in the training data and predicts the relevance of an item to a query in test set. Then, these estimated relevances are used as input to the linear program optimization described in (Singh and Joachims, 2018) with a demographic parity constraint for group fairness.
- **Fair Policy Ranking** (Fair-PGRank) (Singh and Joachims, 2019) An end-to-end, in-processing LTR approach that uses a policy gradient method, directly optimizing for both utility and fairness measures.
- **Reducing Disparate Exposure** (DELTR) (Zehlike and Castillo, 2020) An in-processing LTR method optimizing a weighted sum of a loss function and a fairness criterion. The loss function is a cross entropy designed for ranking (Cao et al., 2007) and fairness objective is a squared hinge loss based on disparate exposure.

4.5.1.3 Evaluation Metrics

We use the *normalized discounted cumulative gain* (NDCG) (Järvelin and Kekäläinen, 2002), as the utility measure. This is defined as: $NDCG(\pi) = DCG(\pi)/Z$, where Z is the DCG for ideal ranking and is used to normalize the ranking so that a perfect ranking would give a NDCG score of 1.

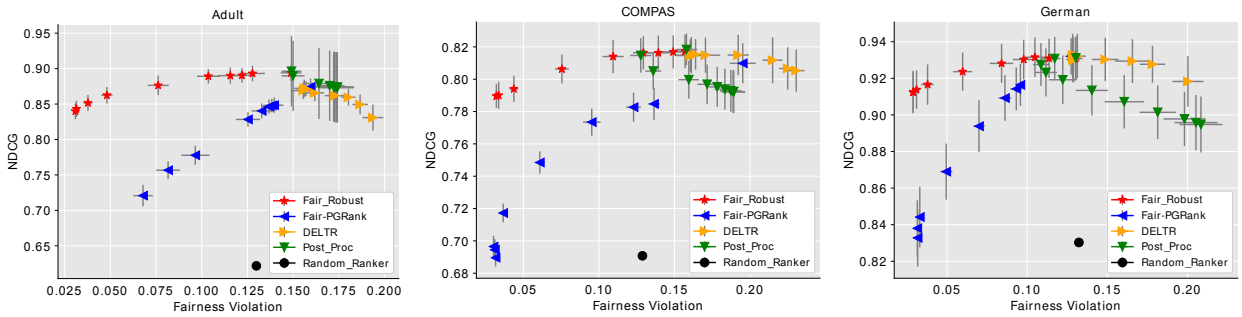


Figure 12: Average *NDCG* versus average *difference of demographic parity* (DP) on test samples, for increasing degrees of fairness penalty λ in each method. FAIR-ROBUST: $\lambda \in [0, 20]$, FAIR-PGRANK: $\lambda \in [0, 20]$, DELTR: $\lambda \in [0, 10^6]$, POST-PROC: $\lambda \in [0, 0.2]$.

For the fairness evaluation in our approach we use *demographic parity* as our fairness violation metric which is based on disparity of average exposure across two groups:

$$\hat{D}_{group}(\mathbf{P}) = |\text{Ex}(G_0|\mathbf{P}) - \text{Ex}(G_1|\mathbf{P})|. \tag{4.16}$$

4.5.1.4 Results

Figure 12 shows the performance of our model (FAIR-ROBUST) against baselines on the three benchmark datasets. We observe a trade-off between fairness and utility in both FAIR-PGRANK and FAIR-ROBUST, i.e., as we increase the fairness penalty parameter (λ), demographic parity difference (as a measure of fairness violation) and NDCG both drop. While DELTR and POST-PROC achieve comparable NDCG when $\lambda = 0$, they fail to satisfy demographic parity as we increase λ and are unable to provide a sufficient utility-fairness trade-off when high levels of fairness are desired. In all three datasets, FAIR-ROBUST outperforms FAIR-PGRANK in terms of

ranking utility when fairness is a priority. When comparing the utility-fairness trade-off between the two approaches, we observe that FAIR-ROBUST can retain higher NDCG in high levels of fairness and provides a preferable trade-off. One notable point is that, even in a noisy dataset like the `COMPAS` dataset, our approach performs better than other methods due to its robustness.

4.5.1.5 Robustness Test

One key benefit of our approach is its robustness to label noise in the learning process. This allows our method can be trained on data with noisy labels and outliers, and still perform well on the test data. To test this property, we repeat the previous experiment with noise added to the training data. After sampling rankings for the training and test sets, we randomly flip $x\%$ of the labels in each ranking problem in the training set. In our experiments, we test various amounts of noise in the training data where x can be 20%, 30%, or 40%. Figure 13 shows the results for the robustness test. Similar to the previous experiment, we observe a trade-off between fairness and utility for FAIR-ROBUST. As the amount of the noise increases FAIR-PGRANK performs poorly and can't maintain its trade-off. Note that when $\lambda = 0$, FAIR-PGRANK still performs well but for other values of λ its NDCG gets close to random ranking. We refer the reader to the appendix for more results on the robustness property.

4.5.2 Microsoft Learning to Rank Dataset

4.5.2.1 Setup

In the previous experiments, we used datasets with inherent demographic biases but the LTR tasks were simulated and constructed from a classification task. To better understand the effectiveness of our approach, we evaluate its performance on a real world LTR dataset.

Due to the lack of standard datasets with demographic biases for LTR task, we use Microsoft's Learning to Rank dataset (Qin and Liu, 2013). We follow the steps discribed in (Yadav et al., 2021) to pre-process the dataset.In this experiment, we compare our method to Fair-PGRank, as both methods are able to trade-off between fairness and utility. Additionally, we include a random baseline, which sorts each item in a query randomly, to give context to NDCG. Similar to the previous experiments, we use NDCG as the utility measure and *demographic parity* as our fairness violation metric, which is based on the disparity of average exposure across two groups.

4.5.2.2 Results

Figure 14 shows the fairness and accuracy trade-off on the test set. With large fairness regularization, Fair-PGRank drops below a random ranking in terms of NDCG, making it inconsistent. This plot shows that Fair-Robust smoothly trades-off group fairness for NDCG. Fair-PGRank's NDCG and group exposure, on the other hand, deteriorate for increasing regularization strength, as (Yadav et al., 2021) also observed.

4.6 Conclusions

In this work, we developed a new LTR system that achieves fairness of exposure for protected groups while maximizing utility to the users. Our adversarial approach constructs a minimax game with the ranker player choosing a distribution over rankings constrained to provide fairness while maximizing utility and an adversary player choosing a distribution of item relevancies that minimizes utility while being similar to training data properties. We show that our method is able to trade-off between utility and fairness much better at high levels of fairness than existing baseline methods. Our work addresses the problem of providing more robust fairness given a

chosen fairness criterion, but does not answer the broader question of which fairness criterion is appropriate for a particular ranking application. Since optimizing one fairness criterion can be detrimental to other fairness criteria, this is an important practical consideration with societal implications. More extensive evaluations based on incorporating other fairness metrics, such as disparate treatment, and generalization of this approach beyond binary utility are two important future directions.

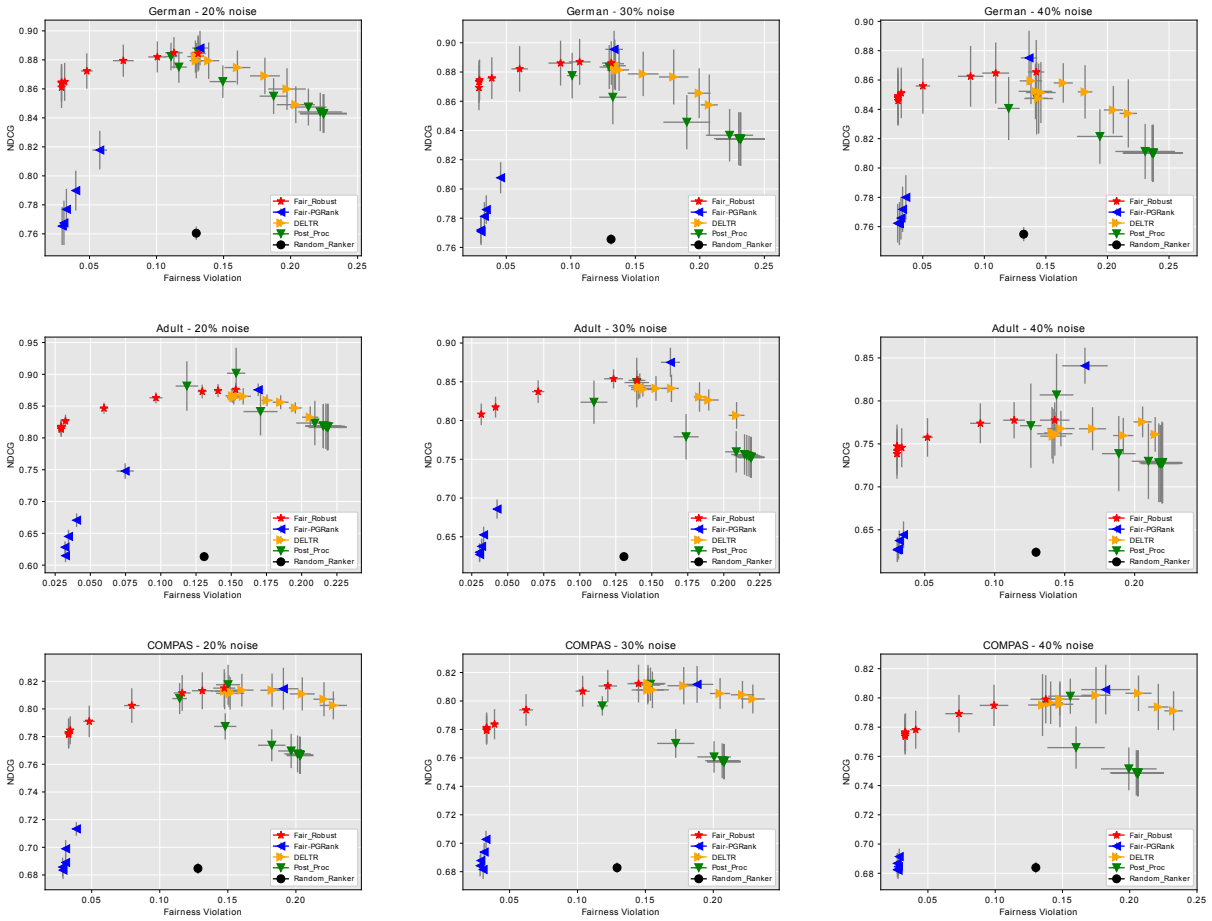


Figure 13: Robustness test on `German`, `Adult` and `COMPAS` datasets with varying degrees of noise in the training data.

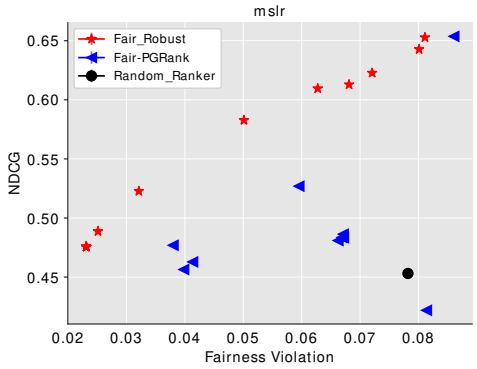


Figure 14: Average *NDCG* versus average *difference of demographic parity* (DP) on test samples, for increasing degrees of fairness penalty λ in each method. FAIR-ROBUST: $\lambda \in [0, 10]$, FAIR-PGRANK: $\lambda \in [0, 10]$.

CHAPTER 5

Superhuman Fairness

(This chapter is based on a paper published as "Superhuman Fairness" (Memarrast et al., 2023b) in the International Conference on Machine Learning 40 (ICML 2023).)

5.1 Introduction

impossibility results prevent multiple common group fairness properties from being simultaneously satisfied (Kleinberg et al., 2016). Thus, no set of decisions can be universally fair to all groups and individuals for all notions of fairness. Instead, specified weightings, or trade-offs, of different criteria are often optimized (Liu and Vicente, 2022). Identifying an appropriate trade-off to prescribe to these fairness methods is a daunting task open to application-specific philosophical and ideological debate that could delay or completely derail the adoption of algorithmic methods.

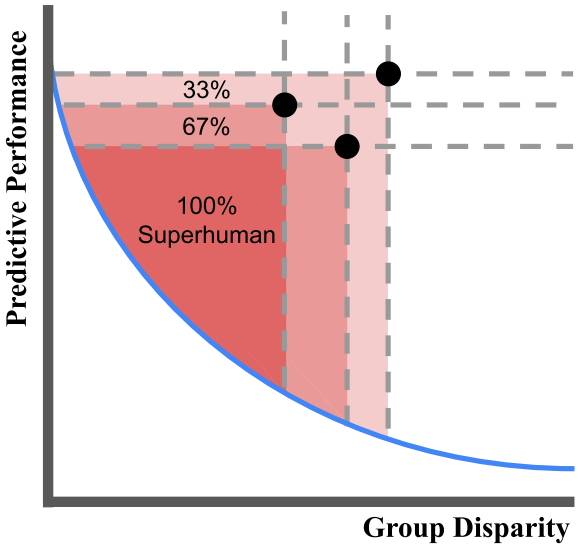


Figure 15: Three sets of decisions (black dots) with different predictive performance and group disparity values defining the sets of 100%-, 67%-, and 33%-superhuman fairness-performance values (red shades) based on Pareto dominance.

We consider the motivating scenario of multiple (error-prone) stakeholders with different notions of fairness and desired performance-fairness trade-offs collaboratively producing decisions. Preference elicitations (Hiranandani et al., 2020) is of limited use since knowing the stakeholder tradeoffs still leaves the question of how different stakeholders preferences should be prioritized. Rather than seeking optimal decisions for specific performance-fairness (meta-)trade-offs, we propose a more modest, yet more practical objective: **produce decisions preferred by all stakeholders over human-produced decisions with maximal frequency.** This provides an opportunity for **superhuman decisions** that Pareto dominate human decisions across predictive performance and fairness measures (Figure Figure 15) *without identifying an explicit desired trade-off*. We argue that for many algorithmic fairness tasks, frequently outperforming

human decisions across all relevant predictive performance and fairness measures may be sufficient for replacing human decision-makers with algorithmic decision-makers.

To the best of our knowledge, this paper is the first to define fairness objectives for supervised machine learning with respect to noisy human decisions rather than using prescriptive trade-offs or hard constraints. We leverage and extend a recently-developed imitation learning method for **subdominance minimization** (Ziebart et al., 2022). Instead of using the subdominance to identify a target trade-off, as previous work does in the inverse optimal control setting of sequential decision-making to estimate a cost function, we use it to directly optimize our fairness-aware classifier. We develop a method based on policy gradient optimization (Sutton and Barto, 2018) that allows flexible classes of probabilistic decision policies (e.g., aware or unaware of protected group membership status) to be optimized for given sets of performance/fairness measures and demonstrations.

We conduct extensive experiments on standard fairness datasets (`Adult` and `COMPAS`) using accuracy as a performance measure and three conflicting fairness definitions: Demographic Parity (Calders et al., 2009), Equalized Odds (Hardt et al., 2016), and Predictive Rate Parity (Chouldechova, 2017a). Though our motivation is to outperform human decisions, we employ a synthetic decision-maker with differing amounts of label and group membership noise to identify sufficient conditions for superhuman fairness of varying degrees. We find that our approach achieves high levels of superhuman performance that increase rapidly with reference decision noise and significantly outperform the superhumanness of other methods that are based on more narrow fairness-performance objectives.

5.2 Fairness, Elicitation, and Imitation

5.2.1 Group Fairness Measures

Group fairness measures are primarily defined by confusion matrix statistics (based on labels $y_i \in \{0, 1\}$ and decisions/predictions $\hat{y}_i \in \{0, 1\}$ produced from inputs $\mathbf{x}_i \in \mathbb{R}^M$) for examples belonging to different protected groups (e.g., $a_i \in \{0, 1\}$).

We focus on three prevalent fairness properties in this paper:

- **Demographic Parity** (DP) (Calders et al., 2009) requires equal positive rates across protected groups:

$$\mathrm{P}(\hat{Y} = 1|A = 1) = \mathrm{P}(\hat{Y} = 1|A = 0);$$

- **Equalized Odds** (EqOdds) (Hardt et al., 2016) requires equal true positive rates and false positive rates across groups, i.e.,

$$\mathrm{P}(\hat{Y}=1|Y=y, A=1) = \mathrm{P}(\hat{Y}=1|Y=y, A=0), \quad y \in \{0, 1\};$$

- **Predictive Rate Parity** (PRP) (Chouldechova, 2017a) requires equal positive predictive value ($\hat{y} = 1$) and negative predictive value ($\hat{y} = 0$) across groups:

$$\mathrm{P}(Y=1|A=1, \hat{Y}=\hat{y}) = \mathrm{P}(Y=1|A=0, \hat{Y}=\hat{y}), \quad \hat{y} \in \{0, 1\}.$$

Violations of these fairness properties can be measured as differences:

$$\texttt{D.DP}(\hat{\mathbf{y}}, \mathbf{a}) = \left| \frac{\sum_{i=1}^{N} \mathbb{I}\left[\hat{y}_i = 1, a_i = 1\right]}{\sum_{i=1}^{N} \mathbb{I}\left[a_i = 1\right]} - \frac{\sum_{i=1}^{N} \mathbb{I}\left[\hat{y}_i = 1, a_i = 0\right]}{\sum_{i=1}^{N} \mathbb{I}\left[a_i = 0\right]} \right|; \tag{5.1}$$

$$\texttt{D.EqOdds}(\hat{\mathbf{y}}, \mathbf{y}, \mathbf{a}) = \max_{y \in \{0,1\}} \left| \frac{\sum_{i=1}^{N} \mathbb{I}\left[\hat{y}_i = 1, y_i = y, a_i = 1\right]}{\sum_{i=1}^{N} \mathbb{I}\left[a_i = 1, y_i = y\right]} - \frac{\sum_{i=1}^{N} \mathbb{I}\left[\hat{y}_i = 1, y_i = y, a_i = 0\right]}{\sum_{i=1}^{N} \mathbb{I}\left[a_i = 0, y_i = y\right]} \right|; \tag{5.2}$$

$$\texttt{D.PRP}(\hat{\mathbf{y}}, \mathbf{y}, \mathbf{a}) = \max_{y \in \{0,1\}} \left| \frac{\sum_{i=1}^{N} \mathbb{I}\left[y_i = 1, \hat{y}_i = y, a_i = 1\right]}{\sum_{i=1}^{N} \mathbb{I}\left[a_i = 1, \hat{y}_i = y\right]} - \frac{\sum_{i=1}^{N} \mathbb{I}\left[y_i = 1, \hat{y}_i = y, a_i = 0\right]}{\sum_{i=1}^{N} \mathbb{I}\left[a_i = 0, \hat{y}_i = y\right]} \right|. \tag{5.3}$$

5.2.2 Preference Elicitation & Imitation Learning

Preference elicitation (Chen and Pu, 2004) is one natural approach to identifying desirable performance-fairness trade-offs. Preference elicitation methods typically query users for their pairwise preference on a sequence of pairs of options to identify their utilities for different characteristics of the options. This approach has been extended and applied to fairness measure elicitation (Hiranandani et al., 2020), allowing efficient learning of linear (e.g., (Equation 5.4)) and non-linear measures from finite and noisy preference feedback.

$$\min_{\theta} \mathbb{E}_{\hat{\mathbf{y}} \sim P_\theta} \Big[\text{loss}(\hat{\mathbf{y}}, \mathbf{y}) + \alpha_{\texttt{DP}} \texttt{D.DP}(\hat{\mathbf{y}}, \mathbf{a}) + \alpha_{\texttt{Odds}} \texttt{D.EqOdds}(\hat{\mathbf{y}}, \mathbf{y}, \mathbf{a}) + \alpha_{\texttt{PRP}} \texttt{D.PRP}(\hat{\mathbf{y}}, \mathbf{y}, \mathbf{a}) \Big]. \tag{5.4}$$

When decisions are made jointly by multiple stakeholders (Donaldson and Preston, 1995) rather than a single individual, preference elicitation may not be very informative. Each stakeholder's preferences could be elicited, for example, but how those sets of preferences should be prioritized

to determine joint outcomes can remain unclear without strong additional assumptions about the decision-making process (e.g., outcomes determined by a majority vote) (Dowling et al., 2016).

Imitation learning (Osa et al., 2018) is a type of supervised machine learning that seeks to produce a general-use policy $\hat{\pi}$ based on demonstrated trajectories of states and actions, $\tilde{\xi} = (\tilde{s}_1, \tilde{a}_1, \tilde{s_2}, \ldots, \tilde{s}_T)$. Inverse reinforcement learning methods (Abbeel and Ng, 2004; Ziebart et al., 2008) seek to rationalize the demonstrated trajectories as the result of (near-) optimal policies on an estimated cost or reward function. Feature matching (Abbeel and Ng, 2004) plays a key role in these methods, guaranteeing if the expected feature counts match, the estimated policy $\hat{\pi}$ will have an expected cost under the demonstrator's unknown fixed cost function weights $\tilde{w} \in \mathbb{R}^K$ equal to the average of the demonstrated trajectories:

$$\begin{aligned} \mathbb{E}_{\xi\sim\hat{\pi}}\left[f_k(\xi)\right] &= \frac{1}{N}\sum_{i=1}^{N} f_k\left(\tilde{\xi}_i\right), \forall k \\ \implies \mathbb{E}_{\xi\sim\hat{\pi}}\left[\text{cost}_{\tilde{w}}(\xi)\right] &= \frac{1}{N}\sum_{i=1}^{N} \text{cost}_{\tilde{w}}\left(\tilde{\xi}_i\right), \end{aligned} \tag{5.5}$$

where $f_k(\xi) = \sum_{s_t\in\xi} f_k\left(s_t\right)$.

(Syed and Schapire, 2007) seeks to outperform the set of demonstrations when the signs of the unknown cost function are known, $\tilde{w}_k \geq 0$, by making the inequality,

$$\mathbb{E}_{\xi\sim\pi}\left[f_k(\xi)\right] \leq \frac{1}{N}\sum_{i=1}^{N} f_k\left(\tilde{\xi}_i\right), \forall k, \tag{5.6}$$

strict for at least one feature. Subdominance minimization (Ziebart et al., 2022) seeks to produce trajectories that outperform each demonstration by a margin:

$$f_k(\xi) + m_k \leq f_k(\tilde{\xi}_i), \forall i, k, \tag{5.7}$$

under the same assumption of known cost weight signs. However, since this is often infeasible, the approach instead minimizes the subdominance, which measures the α-weighted violation of this inequality:

$$\text{subdom}_\alpha(\xi, \tilde{\xi}) \triangleq \sum_k \left[\alpha_k \left(f_k(\xi) - f_k(\tilde{\xi})\right) + 1\right]_+, \tag{5.8}$$

where $[f(x)]_+ \triangleq \max(f(x), 0)$ is the hinge function and the per-feature margin has been reparameterized as α_k^{-1}. Previous work (Ziebart et al., 2022) has employed subdominance minimization in conjunction with inverse optimal control:

$$\min_{\mathbf{w}} \min_{\alpha} \sum_{i=1}^{N} \sum_{k=1}^{K} \text{subdom}_\alpha(\xi^*(\mathbf{w}), \tilde{\xi}_i), \text{where:}$$
$$\xi^*(\mathbf{w}) = \underset{\xi}{\text{argmin}} \sum_k w_k f_k(\xi),$$

learning the cost function parameters $\mathbf{w}$ for the optimal trajectory $\xi^*(\mathbf{w})$ that minimizes subdominance. One contribution of this paper is extending subdominance minimization to the more flexible prediction models needed for fairness-aware classification that are not directly conditioned on cost features or performance/fairness metrics.

5.3 Subdominance Minimization for Improved Fairness-Aware Classification

We approach fair classification from an imitation learning perspective (Ziebart et al., 2022). We assume vectors of (human-provided) reference decisions are available that may have been produced collaboratively by multiple stakeholders with competing predictive performance-fairness tradeoffs. Our goal is to construct a fairness-aware classifier that outperforms reference decisions on all performance and fairness measures on withheld data as frequently as possible, which also provides guarantees to all stakeholders.

5.3.1 Superhumanness and Subdominance

We consider reference decisions $\tilde{\mathbf{y}} = \{\tilde{y}_j\}_{j=1}^{\mathrm{M}}$ that are drawn from an (unknown) human decision-making process or baseline method $\tilde{\mathbb{P}}$, on a set of M items, $\mathbf{X}_{\mathrm{M}\times\mathrm{L}} = \{\mathbf{x}_j\}_{j=1}^{\mathrm{M}}$, where L is the number of attributes in each of M items $\mathbf{x}_j$. Group membership attributes a_m from vector $\mathbf{a}$ indicate to which group item m belongs.

The predictive performance and fairness of decisions $\hat{\mathbf{y}}$ for each item are assessed based on ground truth $\mathbf{y}$ and group membership $\mathbf{a}$ using a set of predictive loss and unfairness measures[1] $\{f_k(\hat{\mathbf{y}}, \mathbf{y}, \mathbf{a})\}$ (e.g., Equation 5.1, Equation 5.2, Equation 5.3). Without loss of generality, we assume that larger values for these measures are less desirable. Ideally, the set of these measures should cover the bases of all stakeholder preference functions (i.e., stakeholder cost functions for

[1]These measures take the place of features used to define cost/reward function in imitation learning methods. We instead use features to describe the inputs to our fairness-aware decision model, $\hat{\mathbb{P}}_{\boldsymbol{\theta}}$.

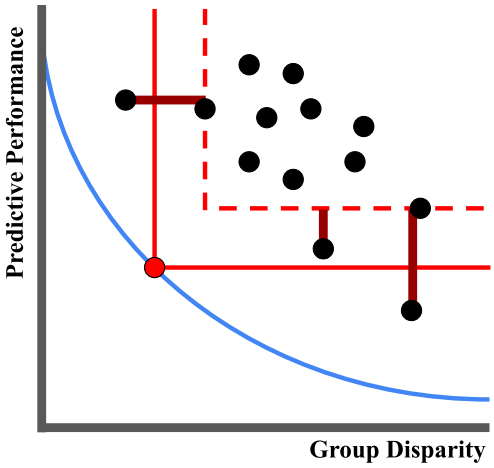


Figure 16: A Pareto frontier for possible $\hat{P}_\theta$ (blue) optimally trading off predictive performance (e.g., inaccuracy) and group unfairness. The model-produced decision (red point) defines dominance boundaries (solid red) and margin boundaries (dashed red), which incur subdominance (maroon lines) on three examples.

evaluating different vectors of decisions can be expressed as summed monotonic transformations of $\{f_k(\hat{\mathbf{y}}, \mathbf{y}, \mathbf{a})\}$ measures).

Definition 5.3.1. *A fairness-aware classifier is considered γ-superhuman for a given set of predictive loss and unfairness measures $\{f_k\}$ if its decisions $\hat{\mathbf{y}}$ satisfy:*

$$P\left(\boldsymbol{f}\left(\hat{\mathbf{y}}, \mathbf{y}, \mathbf{a}\right) \preceq \boldsymbol{f}\left(\tilde{\mathbf{y}}, \mathbf{y}, \mathbf{a}\right)\right) \geq \gamma.$$

If strict Pareto dominance is required to be γ-superhuman, which is often effectively true for continuous domains, then by definition, at most $(1-\gamma)\%$ of human decision makers could be γ-superhuman. However, far fewer than $(1-\gamma)$ may be γ-superhuman if pairs of human decisions do not Pareto dominate one another in either direction (i.e., neither $\boldsymbol{f}\left(\tilde{\mathbf{y}}, \mathbf{y}, \mathbf{a}\right) \preceq \boldsymbol{f}\left(\tilde{\mathbf{y}}', \mathbf{y}, \mathbf{a}\right)$ nor $\boldsymbol{f}\left(\tilde{\mathbf{y}}', \mathbf{y}, \mathbf{a}\right) \preceq \boldsymbol{f}\left(\tilde{\mathbf{y}}, \mathbf{y}, \mathbf{a}\right)$ for pairs of human decisions $\tilde{\mathbf{y}}$ and $\tilde{\mathbf{y}}'$). From this perspective,

any decisions with γ-superhuman performance more than $(1-\gamma)\%$ of the time exceed the performance limit for the distribution of human demonstrators.

Unfortunately, directly maximizing γ is difficult in part due to the discontinuity of Pareto dominance ($\preceq$). The subdominance (Ziebart et al., 2022) serves as a convex upper bound for non-dominance in each metric $\{f_k\}$ and on $1-\gamma$ in aggregate:

$$\begin{aligned}
\text{subdom}^k_{\alpha_k}(\hat{\mathbf{y}}, \tilde{\mathbf{y}}, \mathbf{y}, \mathbf{a}) &\triangleq \Big[\alpha_k \left(f_k(\hat{\mathbf{y}}, \mathbf{y}, \mathbf{a}) - f_k(\tilde{\mathbf{y}}, \mathbf{y}, \mathbf{a})\right) + 1\Big]_+ . \\
\text{subdom}_{\boldsymbol{\alpha}}(\hat{\mathbf{y}}, \tilde{\mathbf{y}}, \mathbf{y}, \mathbf{a}) &\triangleq \sum_k \text{subdom}^k_{\alpha_k}(\hat{\mathbf{y}}, \tilde{\mathbf{y}}, \mathbf{y}, \mathbf{a}).
\end{aligned} \tag{5.9}$$

Given N vectors of reference decisions as demonstrations, $\tilde{\boldsymbol{\mathcal{Y}}} = \{\tilde{\mathbf{y}}_i\}_{i=1}^{\mathrm{N}}$, the subdominance for decision vector $\hat{\mathbf{y}}$ with respect to the set of demonstrations is[1]

$$\text{subdom}_{\boldsymbol{\alpha}}(\hat{\mathbf{y}}, \tilde{\boldsymbol{\mathcal{Y}}}, \mathbf{y}, \mathbf{a}) = \frac{1}{N} \sum_{\tilde{\mathbf{y}} \in \tilde{\boldsymbol{\mathcal{Y}}}} \text{subdom}_{\boldsymbol{\alpha}}(\hat{\mathbf{y}}, \tilde{\mathbf{y}}, \mathbf{y}, \mathbf{a}),$$

where $\hat{\mathbf{y}}_i$ is the predictions produced by our model for the set of items $\mathbf{X}_i$, and $\hat{\boldsymbol{\mathcal{Y}}}$ is the set of these prediction sets, $\hat{\boldsymbol{\mathcal{Y}}} = \{\hat{\mathbf{y}}_i\}_{i=1}^{\mathrm{N}}$. The subdominance is illustrated by Figure 16. Following concepts from support vector machines (Cortes and Vapnik, 1995), reference decisions $\tilde{\mathbf{y}}$ that

---

[1]For notational simplicity, we assume all demonstrated decisions $\tilde{\mathbf{y}} \in \tilde{\boldsymbol{\mathcal{Y}}}$ correspond to the same M items represented in $\mathbf{X}$. Generalization to unique $\mathbf{X}$ for each demonstration is straightforward.

actively constrain the predictions $\hat{\mathbf{y}}$ in a particular feature dimension, k, are referred to as *support vectors* and denoted as:

$$\tilde{\boldsymbol{\mathcal{Y}}}_{\mathrm{SV}_k}(\hat{\mathbf{y}}, \alpha_k) = \Big\{\tilde{\mathbf{y}} | \alpha_k(f_k(\hat{\mathbf{y}}) - f_k(\tilde{\mathbf{y}})) + 1 \geq 0\Big\}.$$

5.3.2 Performance-Fairness Subdominance Minimization

We consider probabilistic predictors $\mathbb{P}_\theta : \mathcal{X}^M \to \Delta_{\mathcal{Y}^M}$ that make structured predictions over the set of items in the most general case, but can also be simplified to make conditionally independent decisions for each item.

Definition 5.3.2. *The minimally subdominant fairness-aware classifier $\hat{\mathbb{P}}_{\boldsymbol{\theta}}$ has model parameters $\boldsymbol{\theta}$ chosen by:*

$$\underset{\boldsymbol{\theta}}{\operatorname{argmin}} \min_{\boldsymbol{\alpha} \succeq 0} \mathbb{E}_{\hat{\mathbf{y}}|\boldsymbol{X} \sim P_{\boldsymbol{\theta}}} \left[\mathrm{subdom}_{\boldsymbol{\alpha},1}\left(\hat{\mathbf{y}}, \tilde{\boldsymbol{\mathcal{Y}}}, \mathbf{y}, \mathbf{a}\right)\right] + \lambda \|\boldsymbol{\alpha}\|_1.$$

Hinge loss slopes $\boldsymbol{\alpha} \triangleq \{\alpha_k\}_{k=1}^{\mathrm{K}}$ are also learned from the data during training. For the subdominance of the k_{th} measure, α_k indicates the degree of sensitivity to how much the algorithm fails to sufficiently outperform demonstrations in that measure. When α_k value is higher, reducing underperformance on that measure minimizes the overall subdominance more than reducing underperformance on other measures.

The bi-level optimization of $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ differs from single-level support vector machine optimization (of $\boldsymbol{\theta}$ alone), which is a convex optimization problem (Cortes and Vapnik, 1995).

Instead, subdominance is a quasi-convex function, which similarly implies that there are no local optima as a function of the realized predictive performance/fairness measures.

Theorem 5.3.3. *The α_k-minimized subbdominance,*

$$\sum_k \overbrace{\min_{\alpha_k \geq 0} \left(\text{subdom}^k_{\alpha_k} \left(\hat{\mathbf{y}}, \tilde{\boldsymbol{\mathcal{Y}}}, \mathbf{y}, \mathbf{a} \right) + \lambda_k \alpha_k \right)}^{\Gamma_k(\hat{\mathbf{y}}, \tilde{\boldsymbol{\mathcal{Y}}}, \mathbf{y}, \mathbf{a})}, \tag{5.10}$$

is a quasiconvex function in terms of the set of measures, $\{f_k(\hat{\mathbf{y}}, \mathbf{y}, \mathbf{a})\}$.

We use the gradient of the expected subdominance with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\alpha}$ to update these variables iteratively, and after convergence, the best learned weights $\boldsymbol{\theta}^*$ are used in the final model $\hat{\mathbb{P}}_{\boldsymbol{\theta}^*}$. Though subdominance minimization is not necessarily quasiconvex in terms of model parameters $\boldsymbol{\theta}$, particularly for complex, nonlinear models, stochastic gradient methods are often effective in avoiding local optima. A commonly used linear model like logistic regression can be used for $\hat{\mathbb{P}}_{\boldsymbol{\theta}}$ to simplify the overall optimization.

Theorem 5.3.4. *The gradient of expected subdominance under $\hat{\mathbb{P}}_\theta$ with respect to the set of reference decisions $\{\tilde{\mathbf{y}}_i\}_{i=1}^N$ is:*

$$\begin{aligned} &\nabla_\theta \mathbb{E}_{\hat{\mathbf{y}}|\mathbf{X} \sim \hat{P}_\theta} \left[\sum_k \Gamma_k \left(\hat{\mathbf{y}}, \tilde{\boldsymbol{\mathcal{Y}}}, \mathbf{y}, \mathbf{a} \right) \right] \\ &= \mathbb{E}_{\hat{\mathbf{y}}|\mathbf{X} \sim \hat{P}_\theta} \left[\left(\sum_k \Gamma_k(\hat{\mathbf{y}}, \tilde{\boldsymbol{\mathcal{Y}}}, \mathbf{y}, \mathbf{a}) \right) \nabla_\theta \log \hat{\mathbb{P}}_\theta(\hat{\mathbf{y}}|\mathbf{X}) \right], \end{aligned}$$

where the optimal α_k *for each* Γ_k (Equation 5.10) *is obtained from:*

$$\alpha_k = \underset{\alpha_k^{(m)}}{\operatorname{argmin}}\, m \text{ such that } f_k\left(\hat{\mathbf{y}}\right) + \lambda \leq \frac{1}{m}\sum_{j=1}^{m} f_k\left(\tilde{\mathbf{y}}^{(j)}\right),$$

using $\alpha_k^{(j)} = \frac{1}{f_k\left(\hat{\mathbf{y}}^{(j)}\right) - f_k\left(\tilde{\mathbf{y}}^{(j)}\right)}$ *to represent the* α_k *value that would make the demonstration with the* j_{th} *smallest* f_k *measure,* $\tilde{\mathbf{y}}^{(j)}$*, a support vector with zero subdominance.*

Using gradient descent, we update the model weights $\boldsymbol{\theta}$ using an approximation of the gradient based on a set of sampled predictions $\hat{\mathbf{y}} \in \hat{\boldsymbol{\mathcal{Y}}}$ from the model $\hat{\mathbb{P}}_{\boldsymbol{\theta}}$:

$$\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \eta \left(\sum_{\hat{\mathbf{y}} \in \hat{\boldsymbol{\mathcal{Y}}}} \left(\sum_k \Gamma_k(\hat{\mathbf{y}}, \tilde{\boldsymbol{\mathcal{Y}}}, \mathbf{y}, \mathbf{a}) \right) \nabla_\theta \log \hat{\mathbb{P}}_\theta(\hat{\mathbf{y}}|\mathbf{X}) \right),$$

We show the steps for training our model in Algorithm 3. *Reference decisions* $\{\tilde{\mathbf{y}}_i\}_{i=1}^{\mathrm{N}}$ from a human decision-making process or baseline method $\tilde{\mathbb{P}}$ are provided as input to the algorithm. $\boldsymbol{\theta}$ is set to an initial value. In each iteration of the algorithm, we first sample a set of *model predictions* $\{\hat{\mathbf{y}}_i\}_{i=1}^{\mathrm{N}}$ from $\hat{\mathbb{P}}_{\boldsymbol{\theta}}(.|\mathbf{X}_i)$ for the matching items used for *reference decisions* $\{\tilde{\mathbf{y}}_i\}_{i=1}^{\mathrm{N}}$. We then find the new $\boldsymbol{\theta}$ (and $\boldsymbol{\alpha}$) based on the algorithms discussed in Theorem 5.3.4.

5.3.3 Generalization Bounds

A fairness-aware classifier with a relatively small number of support vectors has important generalization guarantees under iid assumptions.

Theorem 5.3.5. *A classifier* $\hat{\mathbb{P}}_{\boldsymbol{\theta}}$ *from a family with a convex realizable space of measures* $\{f_k(\hat{\mathbf{y}}, \tilde{\mathbf{y}}, \mathbf{y}, \mathbf{a})\}$ *minimizing* $\sum_i \mathrm{subdom}_{\boldsymbol{\alpha}}\left(\hat{\mathbf{y}}, \tilde{\mathbf{y}}_i, \mathbf{y}_i, \mathbf{a}\right)$ *on a set of* N *iid reference decisions with*

Algorithm 3: Subdominance policy gradient optimization

Draw N set of reference decisions $\{\tilde{\mathbf{y}}_i\}_{i=1}^{\mathrm{N}}$ from a human decision-maker or baseline method $\tilde{\mathbb{P}}$. Initialize: $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta}_0$;

while $\boldsymbol{\theta}$ *not converged* **do**

  Sample model predictions $\{\hat{\mathbf{y}}_i\}_{i=1}^{\mathrm{N}}$ from $\hat{\mathbb{P}}_{\boldsymbol{\theta}}(.|\mathbf{X}_i)$ for the matching items used in reference decisions $\{\tilde{\mathbf{y}}_i\}_{i=1}^{\mathrm{N}}$;

  for $k \in \{1, ..., K\}$ **do**

    Sort reference decisions $\{\tilde{\mathbf{y}}_i\}_{i=1}^{\mathrm{N}}$ in ascending order by k^{th} measure value $f_k(\tilde{\mathbf{y}}_i)$: $\{\tilde{\mathbf{y}}^{(j)}\}_{j=1}^{\mathrm{N}}$;

    Compute $\alpha_k^{(j)} = \frac{1}{f_k\left(\tilde{\mathbf{y}}^{(j)}\right) - f_k\left(\hat{\mathbf{y}}^{(j)}\right)}$;

    $\alpha_k = \underset{\alpha_k^{(m)}}{\operatorname{argmin}}\, m$ such that

    $f_k(\hat{\mathbf{y}}) + \lambda \leq \frac{1}{m} \sum_{j=1}^{m} f_k\left(\tilde{\mathbf{y}}^{(j)}\right)$;

    Compute $\Gamma_k(\hat{\mathbf{y}}_i, \tilde{\boldsymbol{\mathcal{Y}}}, \mathbf{y}, \mathbf{a})$;

  $\boldsymbol{\theta} \leftarrow \boldsymbol{\theta} + \frac{\eta}{N} \sum_i \left(\sum_k \Gamma_k(\hat{\mathbf{y}}_i, \tilde{\boldsymbol{\mathcal{Y}}}, \mathbf{y}, \mathbf{a})\right) \nabla_\theta \log \hat{\mathbb{P}}_\theta(\hat{\mathbf{y}}_i|\mathbf{X}_i)$;

support vector sets $\left\{\tilde{\boldsymbol{\mathcal{Y}}}_{SV_k}(\hat{\mathbf{y}}, \alpha_k)\right\}$ *is on average* γ-**superhuman** *on the population distribution with:* $\gamma = 1 - \frac{1}{N}\left\|\bigcup_{k=1}^{K} \tilde{\boldsymbol{\mathcal{Y}}}_{S\ V_k}(\hat{\mathbf{y}}, \alpha_k)\right\|$.

The proof for this generalization bound (see 5.4) is an extension to our setting of the generalization bound based on support vectors developed for inverse optimal control subdominance minimization (Ziebart et al., 2022). It requires that the realizable set of measures $\{f_k(\hat{\mathbf{y}}, \mathbf{y}, \mathbf{a})\}$ is convex and that the (deterministic) P_θ with measures globally minimizing subdominance can be found. This may be unrealistic for complex P_θ models (e.g., multilayer neural networks).

Importantly, superhuman performance provides comparative satisfaction guarantees for stakeholders. Specifically, stakeholders will prefer the algorithmic decisions with at least γ frequency for a fairly wide range of cost functions defined in terms of the measures $\{f_k(\hat{\mathbf{y}}, \mathbf{y}, \mathbf{a})\}$.

Corollary 5.3.6. *For any stakeholder with a cost function, cost*$(\mathbf{f}, \mathbf{X})$ *such that:*

$$\mathbf{f}_1 \preceq \mathbf{f}_2 \implies \textit{cost}(\mathbf{f}_1, \mathbf{X}) \leq \textit{cost}(\mathbf{f}_2, \mathbf{X}),$$

a γ*-superhuman classifier will be preferable in expectation with probability at least:*

$$P\left(\textit{cost}(\mathbf{f}(\hat{\mathbf{y}}, \mathbf{y}, \mathbf{a}), X) \leq \textit{cost}(\mathbf{f}(\tilde{\mathbf{y}}, \mathbf{y}, \mathbf{a})), \mathbf{X})\right) \geq \gamma.$$

5.4 Proofs of Theorems

Proof of Theorem 5.3.3. We first establish that the average α_k-minimized subdominance of a single measure k,

$$\frac{1}{N}\sum_{\tilde{\mathbf{y}}} \min_{\alpha_k} \text{subdom}_{\alpha_k}^{k}(\hat{\mathbf{y}}, \tilde{\mathbf{y}}, \mathbf{y}, \mathbf{a}) = \frac{1}{N}\sum_{\tilde{\mathbf{y}}} \left[\alpha_k^* \left(\hat{f}_k - f_k(\tilde{\mathbf{y}}, \mathbf{y}, \mathbf{a})\right) + 1\right]_+ ., \tag{5.11}$$

is a monotonic (increasing) function of $\hat{f}_k \triangleq f_k(\hat{\mathbf{y}}, \mathbf{y}, \mathbf{a})$.

When $\alpha_k^* \geq 0$ is nonzero, it is minimized by defining a margin boundary at the largest support vector, $\tilde{\mathbf{y}}_{(j)}$:

$$\alpha_k^* = \frac{1}{f_k(\tilde{\mathbf{y}}_{(j)}, \mathbf{y}, \mathbf{a}) - \hat{f}_k}.$$

When summed over all examples, (Equation 5.11) can be expressed as:

$$\frac{j}{N}\left(\frac{\left(\hat{f}_k - \overline{f_k(\tilde{\mathbf{y}}_{(1:j)}, \mathbf{y}, \mathbf{a})}\right)}{f_k(\tilde{\mathbf{y}}_{(j)}, \mathbf{y}, \mathbf{a}) - \hat{f}_k} + 1\right) = \frac{j}{N}\left(\frac{\left(f_k(\tilde{\mathbf{y}}_{(j)}, \mathbf{y}, \mathbf{a}) - \overline{f_k(\tilde{\mathbf{y}}_{(1:j)}, \mathbf{y}, \mathbf{a})}\right)}{f_k(\tilde{\mathbf{y}}_{(j)}, \mathbf{y}, \mathbf{a}) - \hat{f}_k}\right). \tag{5.12}$$

From the left-hand side of (Equation 5.12), we can see that when $\hat{f}_k$ is equal to the average features of the j (smallest) support vectors, $\overline{f_k(\tilde{\mathbf{y}}_{(1:j)}, \mathbf{y}, \mathbf{a})}$, the subdominance is equal to the support vector frequency (j/N). This is also precisely the value of $\hat{f}_k$ at which a new support vector with measure value $f_k(\tilde{\mathbf{y}}_{(j+1)}, \mathbf{y}, \mathbf{a})$, is added. Starting from the left-hand side of (Equation 5.12), we show that this has the same value of j/N for the subdominance when $\hat{f}_k = \overline{f_k(\tilde{\mathbf{y}}_{(1:j)}, \mathbf{y}, \mathbf{a})}$:

$$
\begin{aligned}
&\frac{j+1}{N}\left(\frac{\overline{f_k(\tilde{\mathbf{y}}_{(1:j)}, \mathbf{y}, \mathbf{a})} - \overline{f_k(\tilde{\mathbf{y}}_{(1:j+1)}, \mathbf{y}, \mathbf{a})}}{f_k(\tilde{\mathbf{y}}_{(j+1)}, \mathbf{y}, \mathbf{a}) - \overline{f_k(\tilde{\mathbf{y}}_{(1:j)}, \mathbf{y}, \mathbf{a})}} + 1\right) \\
&= \frac{j+1}{N}\left(\frac{\overline{f_k(\tilde{\mathbf{y}}_{(1:j)}, \mathbf{y}, \mathbf{a})} - \overline{f_k(\tilde{\mathbf{y}}_{(1:j+1)}, \mathbf{y}, \mathbf{a})} + f_k(\tilde{\mathbf{y}}_{(j+1)}, \mathbf{y}, \mathbf{a}) - \overline{f_k(\tilde{\mathbf{y}}_{(1:j)}, \mathbf{y}, \mathbf{a})}}{f_k(\tilde{\mathbf{y}}_{(j+1)}, \mathbf{y}, \mathbf{a}) - \overline{f_k(\tilde{\mathbf{y}}_{(1:j)}, \mathbf{y}, \mathbf{a})}}\right) \\
&= \frac{j+1}{N}\left(\frac{-\overline{f_k(\tilde{\mathbf{y}}_{(1:j+1)}, \mathbf{y}, \mathbf{a})} + f_k(\tilde{\mathbf{y}}_{(j+1)}, \mathbf{y}, \mathbf{a})}{f_k(\tilde{\mathbf{y}}_{(j+1)}, \mathbf{y}, \mathbf{a}) - \overline{f_k(\tilde{\mathbf{y}}_{(1:j)}, \mathbf{y}, \mathbf{a})}}\right) \\
&= \frac{1}{N}\left(\frac{-(j+1)\overline{f_k(\tilde{\mathbf{y}}_{(1:j+1)}, \mathbf{y}, \mathbf{a})} + f_k(\tilde{\mathbf{y}}_{(j+1)}, \mathbf{y}, \mathbf{a}) + j f_k(\tilde{\mathbf{y}}_{(j+1)}, \mathbf{y}, \mathbf{a})}{f_k(\tilde{\mathbf{y}}_{(j)}, \mathbf{y}, \mathbf{a}) - \overline{f_k(\tilde{\mathbf{y}}_{(1:j)}, \mathbf{y}, \mathbf{a})}}\right) \\
&\overset{(a)}{=} \frac{1}{N}\left(\frac{-j\overline{f_k(\tilde{\mathbf{y}}_{(1:j)}, \mathbf{y}, \mathbf{a})} + j f_k(\tilde{\mathbf{y}}_{(j+1)}, \mathbf{y}, \mathbf{a})}{f_k(\tilde{\mathbf{y}}_{(j+1)}, \mathbf{y}, \mathbf{a}) - \overline{f_k(\tilde{\mathbf{y}}_{(1:j)}, \mathbf{y}, \mathbf{a})}}\right) = \frac{j}{N},
\end{aligned}
\tag{5.13}
$$

where step (a) follows from $(j+1)\overline{f_k(\tilde{\mathbf{y}}_{(1:j+1)}, \mathbf{y}, \mathbf{a})} - f_k(\tilde{\mathbf{y}}_{(j+1)}, \mathbf{y}, \mathbf{a}) = j\overline{f_k(\tilde{\mathbf{y}}_{(1:j)}}$. This shows that at its non-smooth points, the subdominance is not decreasing.

Differentiating the right-hand side of (Equation 5.12) yields:

$$
j\left(\frac{\left(f_k(\tilde{\mathbf{y}}_{(j)}, \mathbf{y}, \mathbf{a}) - \overline{f_k(\tilde{\mathbf{y}}_{(1:j)}, \mathbf{y}, \mathbf{a})}\right)}{(f_k(\tilde{\mathbf{y}}_{(j)}, \mathbf{y}, \mathbf{a}) - \hat{f}_k)^2}\right), \tag{5.14}
$$

which is nonnegative as long as $f_k(\tilde{\mathbf{y}}_{(j)}) \geq \overline{f_k(\tilde{\mathbf{y}}_{(1:j)}, \mathbf{y}, \mathbf{a})}$, a condition that is always true by definition of the ordered support vectors. Thus, since subdominance is non-decreasing at both its smooth and nonsmooth portions, it is a monotonic (increasing) function of $\hat{f}_k$ in each dimension k.

Since the per-measure subdominances are independent and combined via summation over all the dimensions k to form the entire subdominance, the sublevel sets must be convex, and the subdominance overall is therefore a quasiconvex function of $\hat{\mathbf{f}}$. □

Proof of Theorem 5.3.4. The gradient of the training objective with respect to model parameters θ is:

$$\nabla_\theta \mathbb{E}_{\hat{\mathbf{y}}|\mathbf{X}\sim\hat{P}_\theta}\left[\sum_k \overbrace{\min_{\alpha_k}\left(\text{subdom}^k_{\alpha_k}\left(\hat{\mathbf{y}}, \tilde{\boldsymbol{\mathcal{Y}}}, \mathbf{y}, \mathbf{a}\right) + \lambda_k\alpha_k\right)}^{\Gamma_k(\hat{\mathbf{y}},\tilde{\boldsymbol{\mathcal{Y}}},\mathbf{y},\mathbf{a})}\right] = \mathbb{E}_{\hat{\mathbf{y}}|\mathbf{X}\sim\hat{P}_\theta}\left[\left(\sum_k \Gamma_k(\hat{\mathbf{y}}, \tilde{\boldsymbol{\mathcal{Y}}}, \mathbf{y}, \mathbf{a})\right)\nabla_\theta \log \hat{\mathbb{P}}_\theta(\hat{\mathbf{y}}|\mathbf{X})\right],$$

which follows directly from a property of gradients of logs of function:

$$\nabla_\theta \log \hat{\mathbb{P}}(\hat{\mathbf{y}}|\mathbf{X}) = \frac{1}{\hat{\mathbb{P}}(\hat{\mathbf{y}}|\mathbf{X})}\nabla_\theta \hat{\mathbb{P}}(\hat{\mathbf{y}}|\mathbf{X}) \implies \nabla_\theta \hat{\mathbb{P}}_\theta(\hat{\mathbf{y}}|\mathbf{X}) = \hat{\mathbb{P}}(\hat{\mathbf{y}}|\mathbf{X})\nabla_\theta \log \hat{\mathbb{P}}(\hat{\mathbf{y}}|\mathbf{X}). \tag{5.15}$$

We note that this is a well-known approach employed by policy-gradient methods in reinforcement learning (Sutton and Barto, 2018).

Next, we consider how to obtain the $\alpha-$minimized subdominance for a particular tuple $(\hat{\mathbf{y}},\tilde{\boldsymbol{\mathcal{Y}}},\mathbf{y},\mathbf{a})$, $\Gamma_k\left(\hat{\mathbf{y}}, \tilde{\boldsymbol{\mathcal{Y}}}, \mathbf{y}, \mathbf{a}\right) = \min_{\alpha_k}\left(\text{subdom}^k_{\alpha_k}\left(\hat{\mathbf{y}}, \tilde{\boldsymbol{\mathcal{Y}}}, \mathbf{y}, \mathbf{a}\right) + \lambda_k\alpha_k\right)$, analytically.

First, we note that $\text{subdom}^k_{\alpha_k}\left(\hat{\mathbf{y}}, \tilde{\boldsymbol{\mathcal{Y}}}, \mathbf{y}, \mathbf{a}\right) + \lambda_k \alpha_k$ is comprised of hinged linear functions of α_k, making it a convex and piece-wise linear function of α_k. This has two important implications: (1) any point of the function for which the subgradient includes 0 is a global minimum of the function (Boyd and Vandenberghe, 2004); (2) an optimum must exist at a corner of the function: $\alpha_k = 0$ or where one of the hinge functions becomes active:

$$\alpha_k(f_k(\hat{\mathbf{y}}_i) - f_k(\tilde{\mathbf{y}}_i)) + 1 = 0 \implies \alpha_k = \frac{1}{f_k(\tilde{\mathbf{y}}_i) - f_k(\hat{\mathbf{y}}_i)}. \tag{5.16}$$

The subgradient for the j^{th} of these points (ordered by f_k value from smallest to largest and denoted $f_k(\tilde{\mathbf{y}}^{(j)})$ for the demonstration) is:

$$\begin{aligned}
&\partial_{\alpha_k} \text{subdom}^k_{\alpha_k}\left(\hat{\mathbf{y}}, \tilde{\boldsymbol{\mathcal{Y}}}, \mathbf{y}, \mathbf{a}\right)\Big|_{\alpha_k = (f_k(\hat{\mathbf{y}}) - f_k(\tilde{\mathbf{y}}^{(j)}))^{-1}} = \partial_{\alpha_k}\left(\frac{1}{N}\sum_{i=1}^{j}\left[\alpha_k\left(f_k(\hat{\mathbf{y}}) - f_k(\tilde{\mathbf{y}}^{(i)})\right) + 1\right]_+ + \lambda\alpha_k\right) \\
&= \lambda + \frac{1}{N}\sum_{i=1}^{j-1}\left(f_k(\hat{\mathbf{y}}) - f_k(\tilde{\mathbf{y}}^{(i)})\right) + \left[0, f_k(\hat{\mathbf{y}}) - f_k(\tilde{\mathbf{y}}^{(j)})\right],
\end{aligned}$$

where the final bracketed expression indicates the range of values added to the constant value preceding it.

The smallest j for which the largest value in this range is positive must contain the 0 in its corresponding range, and is thus the provides the j value for the optimal α_k value. □

Proof of Theorem 5.3.5. We first recall generalization guarantees for support vector machines (SVMs) (Cortes and Vapnik, 1995) based on leave-one-out cross validation (LOOCV) that our approach leverages. For support vector machines, examples that are not support vectors incur

zero loss and do not actively constrain the SVM parameters. Thus, when these examples are removed, the decision boundary does not change and therefore no cross validation loss is incurred on any left-out example during LOOCV. Due to this, the support vector frequency is an upper bound on the leave-one-out cross validation error, which is an (almost) unbiased estimate of the generalization inaccuracy (Vapnik and Chapelle, 2000).

Since subdominance is quasiconvex instead of convex, this analysis is slightly more complicated. Specifically, it requires the set of realizable $\mathbf{f}$ measures to be convex. The intersection of the sublevel sets of the quasiconvex subdominance (Theorem 5.3.3 with a convex set of feasible measures is also convex, so the constrained subdominance minimization problem (minimizing subdominance over the set of realizable features for the family of possible P_θ) is also quasiconvex. As a result, no local optima exist that are not global optima. Since the non-support vectors do not actively constrain the global optima, removing them does not change the global optima and therefore they do not contribute any loss to the leave-one-out cross validation error. The remaining argument then follows directly from the SVM LOOCV analysis.

□

5.5 Experiments

The goal of our approach is to produce a fairness-aware prediction method that outperforms reference (human) decisions on multiple fairness/performance measures. In this section, we

discuss our experimental design to synthesize reference decisions with varying levels of noise, evaluate our method, and provide comparison baselines.[1]

5.5.1 Training and Testing Dataset Construction

To emulate human decision-making with various levels of noise, we add noise to benchmark fairness datasets and apply fair learning methods over repeated randomized dataset splits. We describe this process in detail in the following section.

5.5.1.0.1 Datasets

We perform experiments on two benchmark fairness datasets:

- UCI `Adult` dataset (Dheeru and Karra Taniskidou, 2017) considers predicting whether a household's income exceeds \$50K/yr based on census data. Group membership is based on gender. The dataset consists of 45,222 items.
- `COMPAS` dataset (Larson et al., 2016) considers predicting recidivism with group membership based on race. It consists of 6,172 examples.

5.5.1.1 Partitioning the data

We first split the entire dataset randomly into a disjoint train (`train-all`) and test (`test-all`) set of equal size. The test set (`test-all`) is entirely withheld from the training procedure and ultimately used solely for evaluation. To produce each demonstration (a vector of reference decisions), we split the (`train-all`) set randomly into a disjoint train (`train-demo`) and test (`test-demo`) set of equal size.

---

[1]Our code is publicly available at `https://github.com/omidMemari/superhumn-fairness`.

5.5.1.2 Noise insertion

We randomly flip $\epsilon\%$ of the ground truth labels $\mathbf{y}$ and group membership attributes $\mathbf{a}$ to add noise to our demonstration-producing process.

5.5.1.3 Fair classifier $\tilde{\mathbb{P}}$

Using the noisy data, we provide existing fairness-aware methods with labeled `train-demo` data and unlabeled `test-demo` to produce decisions on the `test-demo` data as demonstrations $\tilde{\mathbf{y}}$. Specifically, we employ:

- The **Post-processing** method of (Hardt et al., 2016), which aims to reduce both *prediction error* and {*demographic parity* or *equalized odds*} at the same time. We use *demographic parity* as the fairness constraint. We produce demonstrations using this method for `Adult` dataset.
- **Robust fairness for logloss-based classification** (Rezaei et al., 2020) employs distributional robustness to match target fairness constraint(s) while robustly minimizing the log loss. We use *equalized odds* as the fairness constraint. We employ this method to produce demonstrations for `COMPAS` dataset.

We repeat the process of partitioning `train-all` $N = 50$ times to create randomized partitions of `train-demo` and `test-demo` and to then produce a set of demonstrations $\{\tilde{\mathbf{y}}\}_{i=1}^{50}$.

5.5.2 Evaluation Metrics and Baselines

5.5.2.1 Predictive Performance and Fairness Measures

Our focus for evaluation is on outperforming demonstrations in multiple fairness and performance measures. We use $K = 4$ measures: *inaccuracy* (`Prediction error`), *difference from demographic parity* (`D.DP`), *difference from equalized odds* (`D.EqOdds`), *difference from predictive rate parity* (`D.PRP`).

5.5.2.2 Baseline methods

As baseline comparisons, we train five different models on the entire train set (`train-all`) and then evaluate them on the withheld test data (`test-all`):

- The **Post-processing** model of (Hardt et al., 2016) with {*demographic parity* or *equalized odds*} as the fairness constraint (`post_proc_dp` and `post_proc_eqodds`).
- The **Robust Fair-logloss** model of (Rezaei et al., 2020) with {*demographic parity* or *equalized odds*} as the fairness constraint (`fair_logloss_dp` and `fair_logloss_eqodds`).
- The **Multiple Fairness Optimization** framework of (Hsu et al., 2022) which is designed to satisfy three conflicting fairness measures {*demographic parity, equalized odds*, and *predictive rate parity*} to the best extent possible (`MFOpt`).

5.5.2.3 Hinge Loss Slopes

As discussed previously, each α_k value corresponds to the hinge loss slope, which defines the sensitivity of produced decision not sufficiently outperforming the demonstrations on the k_{th}

measure. When the α_k is large, the model heavily weights support vector reference decisions for that particular k when minimizing subdominance. We report these values in our experiments.

5.5.3 Superhuman Model Specification and Updates

We use a *logistic regression* model $\mathbb{P}_{\boldsymbol{\theta}_0}$ with first-order moment feature functions, $\phi(y, \mathbf{x}) = [x_1 y, x_2 y, \ldots x_m y]^\top$, and weights $\boldsymbol{\theta}$ applied independently on each item as our decision model. During the training process, we update the model parameter $\boldsymbol{\theta}$ to reduce subdominance.

5.5.3.1 Sample from model

In each iteration of the algorithm, we first sample *prediction vectors* $\{\hat{\mathbf{y}}_i\}_{i=1}^{\mathrm{N}}$ from $\hat{\mathbb{P}}_{\boldsymbol{\theta}}(.|\mathbf{X}_i)$ for the matching items used in demonstrations $\{\tilde{\mathbf{y}}_i\}_{i=1}^{\mathrm{N}}$. In the implementation, to produce the i_{th} sample, we look up the indices of the items used in $\tilde{\mathbf{y}}_i$, which constructs item set $\mathbf{X}_i$. Now we make predictions using our model on this item set $\hat{\mathbb{P}}_{\boldsymbol{\theta}}(.|\mathbf{X}_i)$. The model produces a probability distribution for each item which can be sampled and used as a prediction $\{\hat{\mathbf{y}}_i\}_{i=1}^{\mathrm{N}}$.

5.5.3.2 Update model parameters

We update $\boldsymbol{\theta}$ until convergence using Algorithm 3. For our logistic regression model, the gradient is:

$$\nabla_\theta \log \hat{\mathbb{P}}_\theta(\hat{\mathbf{y}}|\mathbf{X}) = \phi(\hat{\mathbf{y}}, \mathbf{X}) - \mathbb{E}_{\hat{\mathbf{y}}|\mathbf{X}\sim\hat{P}_\theta}\left[\phi(\hat{\mathbf{y}}, \mathbf{X})\right],$$

where ϕ denotes the feature function and $\phi(\hat{\mathbf{y}}, \mathbf{X}) = \sum_{m=1}^{\mathrm{M}} \phi(\hat{\mathbf{y}}_m, \mathbf{x}_m)$ is the corresponding feature function for the i^{th} set of reference decisions. We employ a learning rate of $\eta = 0.01$.

TABLE IV: Experimental results on noise-free datasets, along with the α_k values learned for each feature in subdominance minimization.

Method \ Dataset	Adult				COMPAS			
	Prediction error	DP diff	EqOdds diff	PRP diff	Prediction error	DP diff	EqOdds diff	PRP diff
α_k	29.63	10.77	5.83	13.42	29.33	4.51	3.34	57.74
γ-superhuman	100%	100%	100%	100%	100%	100%	100%	98%
MinSub-Fair (ours)	**0.1937**	0.0310	**0.0093**	0.1760	0.3600	0.0320	0.0367	0.1723
MFOpt	0.3157	**0.0132**	0.0225	0.2092	0.4597	0.0919	0.0397	0.1533
post_proc_dp	0.2265	0.1442	0.0879	0.2304	0.3532	0.0879	0.0884	0.1605
post_proc_eqodds	0.2176	0.1572	0.1396	0.1451	**0.3513**	0.1442	0.1584	**0.1485**
fair_logloss_dp	0.3835	0.0246	0.0577	**0.1158**	0.4846	**0.0053**	0.1455	0.1832
fair_logloss_eqodds	0.3776	0.1179	0.0238	0.1380	0.4870	0.1272	**0.0119**	0.1539

TABLE V: Experimental results on datasets with noisy demonstrations, along with the α_k values learned for each feature.

Method \ Dataset	Adult				COMPAS			
	Prediction error	DP diff	EqOdds diff	PRP diff	Prediction error	DP diff	EqOdds diff	PRP diff
α_k	29.63	10.77	5.83	13.42	29.33	4.51	3.34	57.74
γ-superhuman	100%	100%	100%	100%	100%	100%	100%	98%
MinSub-Fair (ours)	**0.1937**	0.0310	**0.0093**	0.1760	0.3600	0.0320	0.0367	0.1723
MFOpt	0.3157	**0.0132**	0.0225	0.2092	0.4597	0.0919	0.0397	0.1533
post_proc_dp	0.2265	0.1442	0.0879	0.2304	0.3532	0.0879	0.0884	0.1605
post_proc_eqodds	0.2176	0.1572	0.1396	0.1451	**0.3513**	0.1442	0.1584	**0.1485**
fair_logloss_dp	0.3835	0.0246	0.0577	**0.1158**	0.4846	**0.0053**	0.1455	0.1832
fair_logloss_eqodds	0.3776	0.1179	0.0238	0.1380	0.4870	0.1272	**0.0119**	0.1539

5.5.4 Experimental Results

After training each model, e.g., obtaining the best model weight $\boldsymbol{\theta}^*$ from the training data (`train-all`) for `superhuman`, we evaluate each on unseen test data (`test-all`). We employ hard predictions (i.e., the most probable label) using our approach at test time rather than randomly sampling.

TABLE VI: Percentage of reference demonstrations that each method outperforms in all prediction/fairness measures.

	`Adult` $\epsilon = 0.0$	`Adult` $\epsilon = 0.2$	`COMPAS` $\epsilon = 0.0$	`COMPAS` $\epsilon = 0.2$
`MinSub-Fair` (ours)	**96%**	**100%**	**100%**	**98%**
`MFOpt`	42%	0%	18%	18%
`post_proc_dp`	16%	86%	**100%**	80%
`post_proc_eqodds`	0%	66%	**100%**	88%
`fair_logloss_dp`	0%	0%	0%	0%
`fair_logloss_eqodds`	0%	0%	0%	0%

5.5.4.1 Noise-free reference decisions

Our first set of experiments considers learning from reference decisions with no added noise.[1] The results are shown in Figure 17 and Figure 18. We observe that our approach outperforms demonstrations in all fairness measures and shows comparable performance in *accuracy*. The (`post_proc_dp`) performs comparably to the average of demonstrations in all dimensions, hence our approach can outperform it in all fairness measures. In comparison to (`post_proc_dp`), our approach can outperform in all fairness measures but is slightly worse in *prediction error*.

We show the experiment results along with α_k values in Table IV. Note that the margin boundaries (dotted red lines) in Figure 17 and Figure 18 are equal to $\frac{1}{\alpha_k}$ for measure k, hence there is reverse relation between α_k and margin boundary for measure k. We observe larger values of α_k for *prediction error* and *demographic parity difference*. The reason is that these

[1]Added noise does not imply the original dataset is noise-free.

measures are already optimized in demonstrations and our model has to increase α_k values for those measures to sufficiently outperform them.

5.5.4.2 Noisy reference decisions

In our second set of experiments, we introduce significant amounts of noise ($\epsilon = 0.2$) into our reference decisions. We similarly add this noise to the training datasets (`train-all`) of the baseline methods. The results for these experiments are shown in Figure 19 and Figure 20. We observe that in the case of learning from noisy demonstrations, our approach still outperforms the reference decisions.

The main difference here is that due to the noisy setting, demonstrations have worse *prediction error* but regardless of this issue, our approach still can achieve a competitive *prediction error*. We show the experimental results along with α_k values in Table V.

5.5.4.3 Relationship of noise to superhuman performance

We also evaluate the relationship between the amount of augmented noise in the label and protected attribute of demonstrations, with achieving γ-superhuman performance in our approach. As shown in Figure 21, with slightly increasing the amount of noise in demonstrations, our approach can outperform 100% of demonstrations and reach 1-superhuman performance. In Table VI we show the percentage of demonstrations that each method can outperform across all prediction/fairness measures (i.e., the $\gamma-$superhuman value).

5.5.4.4 Experiment with more measures

Since our approach is flexible enough to accept wide range of fairness/performance measures, we extend the experiment on `Adult` to $K = 5$ features. In this experiment we use *Demographic*

Parity (`D.DP`), *Equalized Odds* (`D.EqOdds`), *False Negative Rate* (`D.FNR`), *False Positive Rate* (`D.FPR`) and *Prediction Error* as the features to outperform reference decisions on. The results are shown in Figure 22.

5.6 Conclusions

In this paper, we introduce superhuman fairness, an approach to fairness-aware classifier construction based on imitation learning. Our approach avoids explicit performance-fairness trade-off specification or elicitation. Instead, it seeks to unambiguously outperform human decisions across multiple performance and fairness measures with maximal frequency. When successful, this provides important guarantees for stakeholders with a broad set of possible preferences for performance and fairness measures. We develop a general framework for pursuing this based on subdominance minimization (Ziebart et al., 2022) and policy gradient optimization methods (Sutton and Barto, 2018) that enable a broad class of probabilistic fairness-aware classifiers to be learned. Our experimental results show the effectiveness of our approach in outperforming synthetic decisions corrupted by small amounts of label and group-membership noise when evaluated using multiple fairness criteria combined with predictive accuracy.

5.6.1 Societal impacts

By design, our approach has the potential to identify fairness-aware decision-making tasks in which human decisions can frequently be outperformed by a learned classifier on a set of provided performance and fairness measures. This has the potential to facilitate a transition from manual to automated decisions that are preferred by all interested stakeholders, so long as their interests are reflected in some of those measures. Since the formulation only provides

preference guarantees for stakeholders with nonnegatively-weighted combinations of performance and fairness measures, it may reduce the negative impact of stakeholders in human-produced decision-making from successfully seeking negative outcomes for specific groups.

Despite these benefits, our approach also has limitations. First, when performance-fairness tradeoffs can either be fully specified (e.g., based on first principles) or effectively elicited, fairness-aware classifiers optimized for those trade-offs should produce better results than our approach, which operates under greater uncertainty cast by the noisiness of human decisions. Second, if target fairness concepts lie outside the set of measures we consider, our resulting fairness-aware classifier will be oblivious to them. Third, our approach assumes human-demonstrated decision are well-intentioned, noisy reflections of desired performance-fairness trade-offs. If this is not the case, then our methods could succeed in outperforming them across all fairness measures, but still not provide an adequate degree of fairness.

5.6.2 Future directions

We have conducted experiments with a relatively small number of performance/fairness measures using a simplistic logistic regression model. Scaling our approach to much larger numbers of measures and classifiers with more expressive representations are both of great interest. Additionally, we plan to pursue experimental validation using human-provided fairness-aware decisions in addition to the synthetically-produced decisions we consider in this paper. More broadly, other techniques that can minimize subdominance or provide generalization guarantees for stakeholders adoption preferences of algorithmic decision-making are of significant interest.

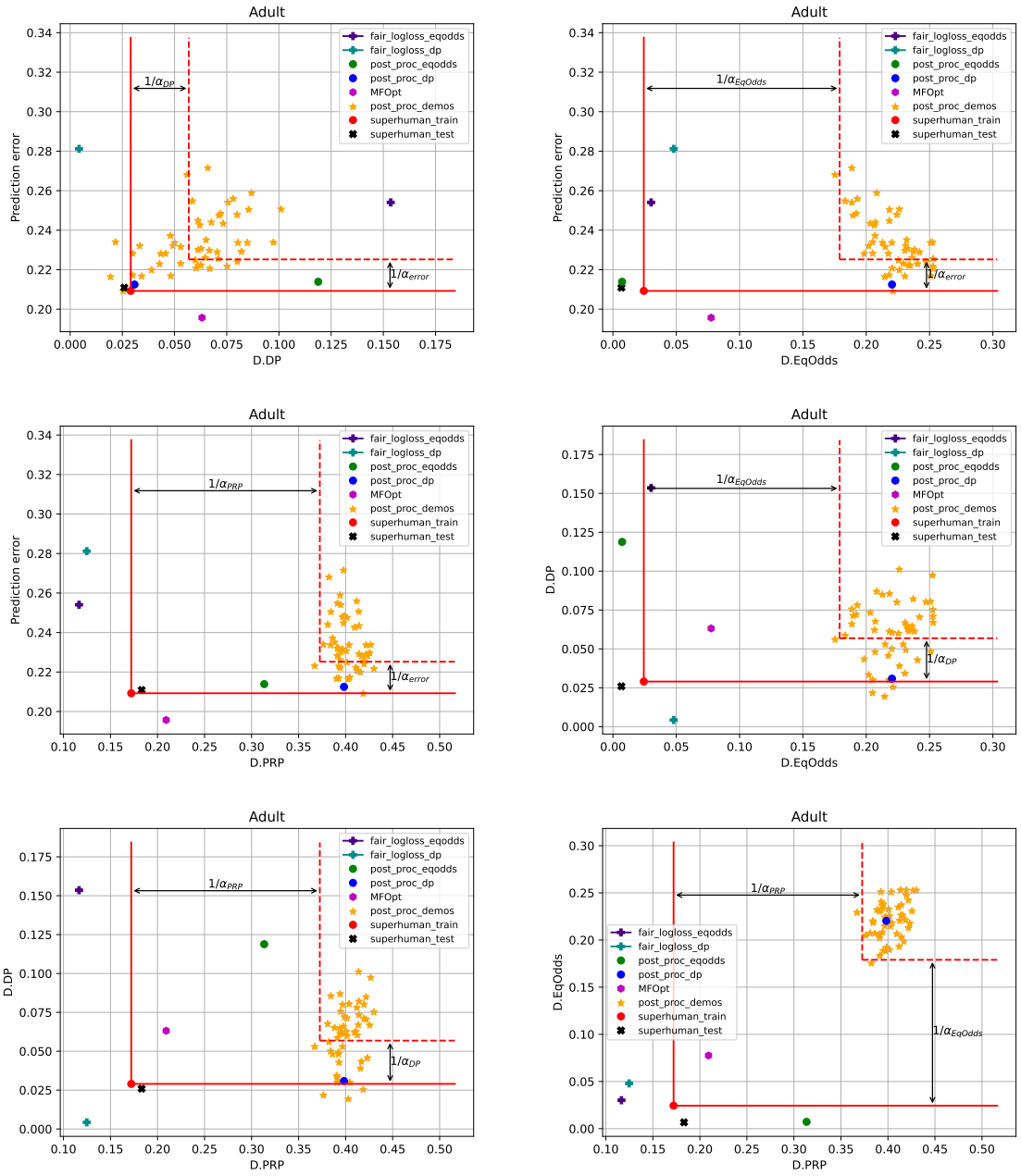


Figure 17: *Prediction error* versus *difference of*: *Demographic Parity* (`D.DP`), *Equalized Odds* (`D.EqOdds`) and *Predictive Rate Parity* (`D.PR`) on test data using noiseless training data ($\epsilon = 0$) for `Adult` dataset.

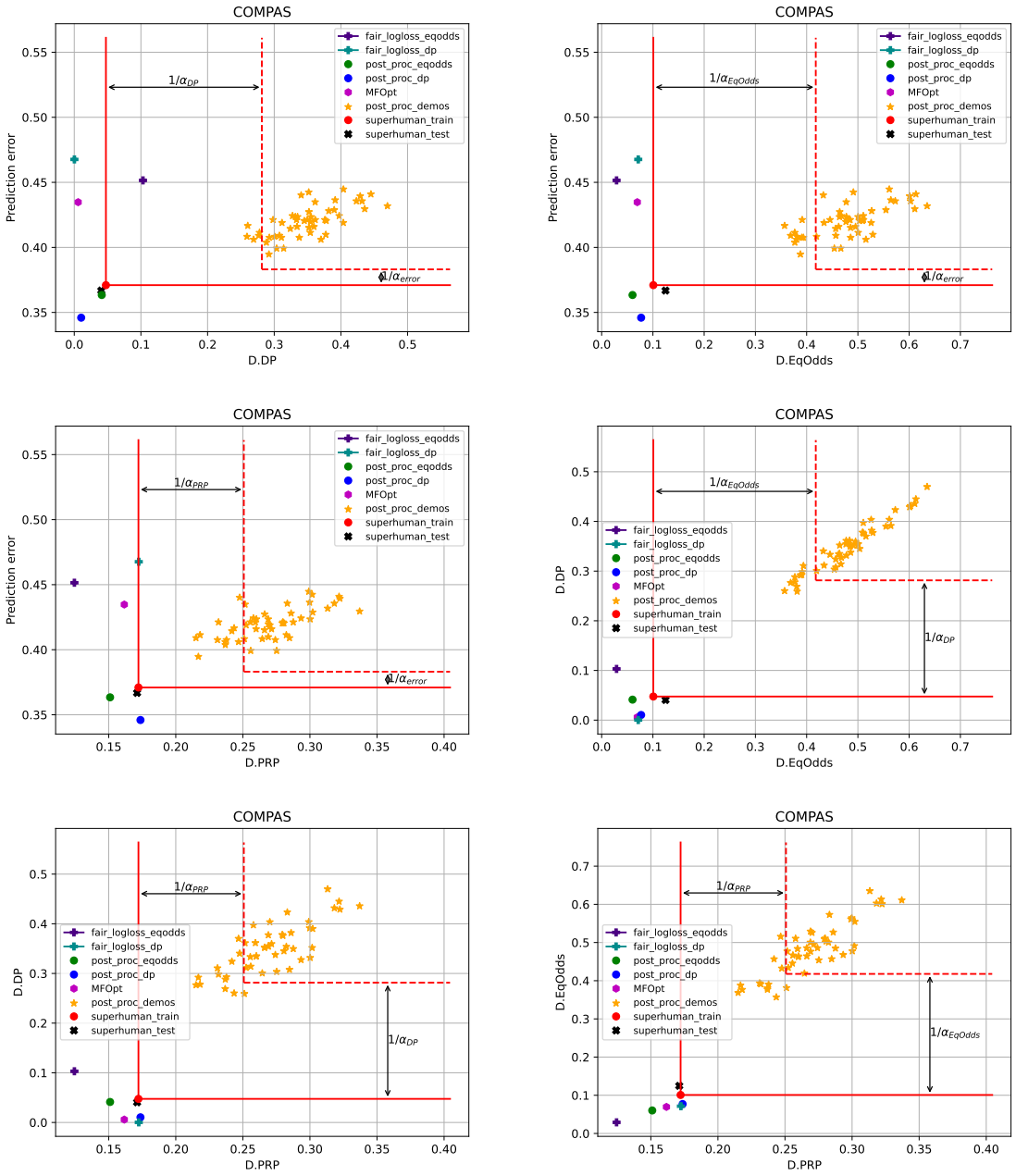


Figure 18: *Prediction error* versus *difference of*: *Demographic Parity* (`D.DP`), *Equalized Odds* (`D.EqOdds`) and *Predictive Rate Parity* (`D.PR`) on test data using noiseless training data ($\epsilon = 0$) for `COMPAS` dataset.

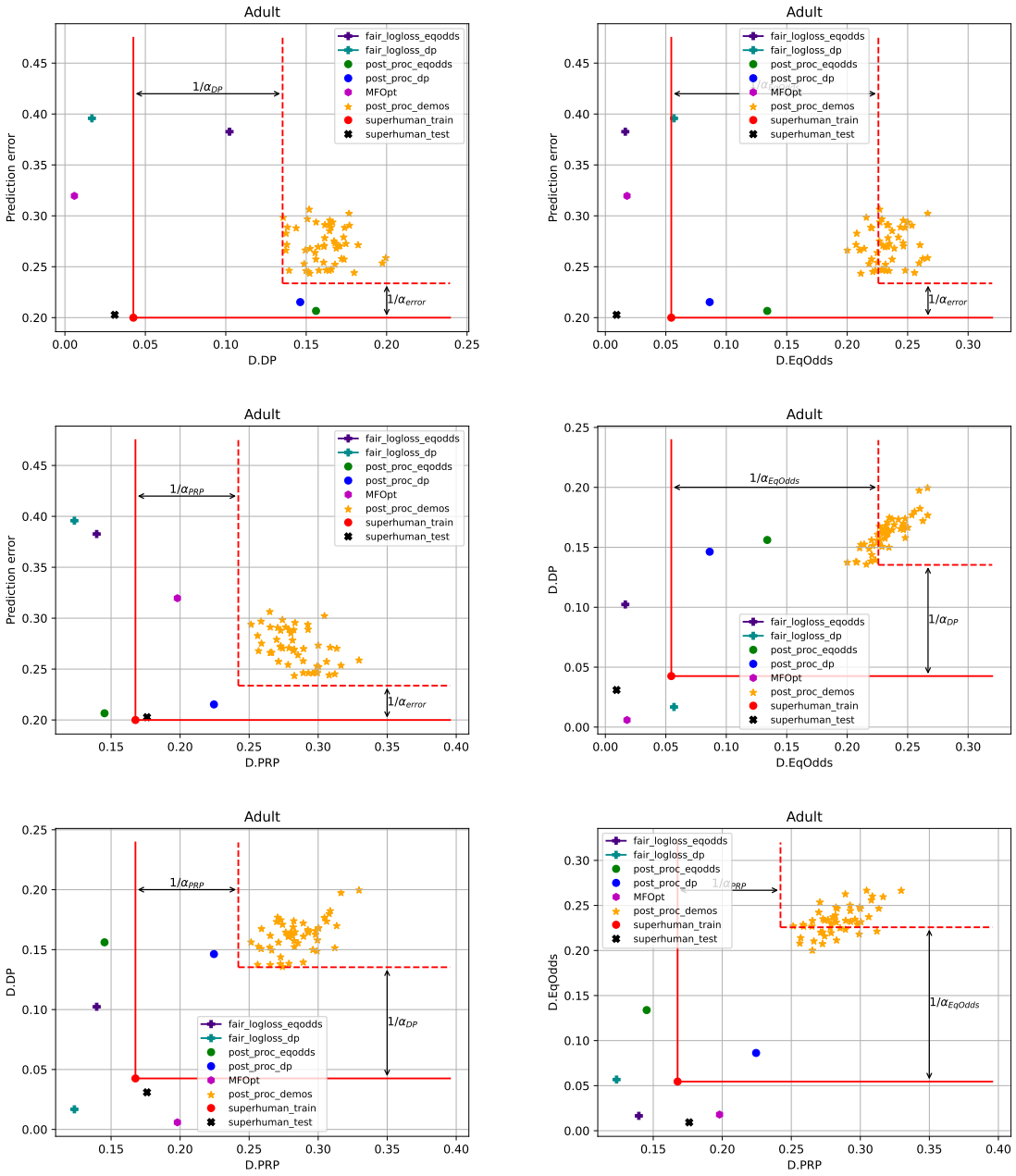


Figure 19: Experimental results on the `Adult` dataset with noisy demonstrations ($\epsilon = 0.2$). Margin boundaries are shown with dotted red lines. Each plot shows the relationships between two features.

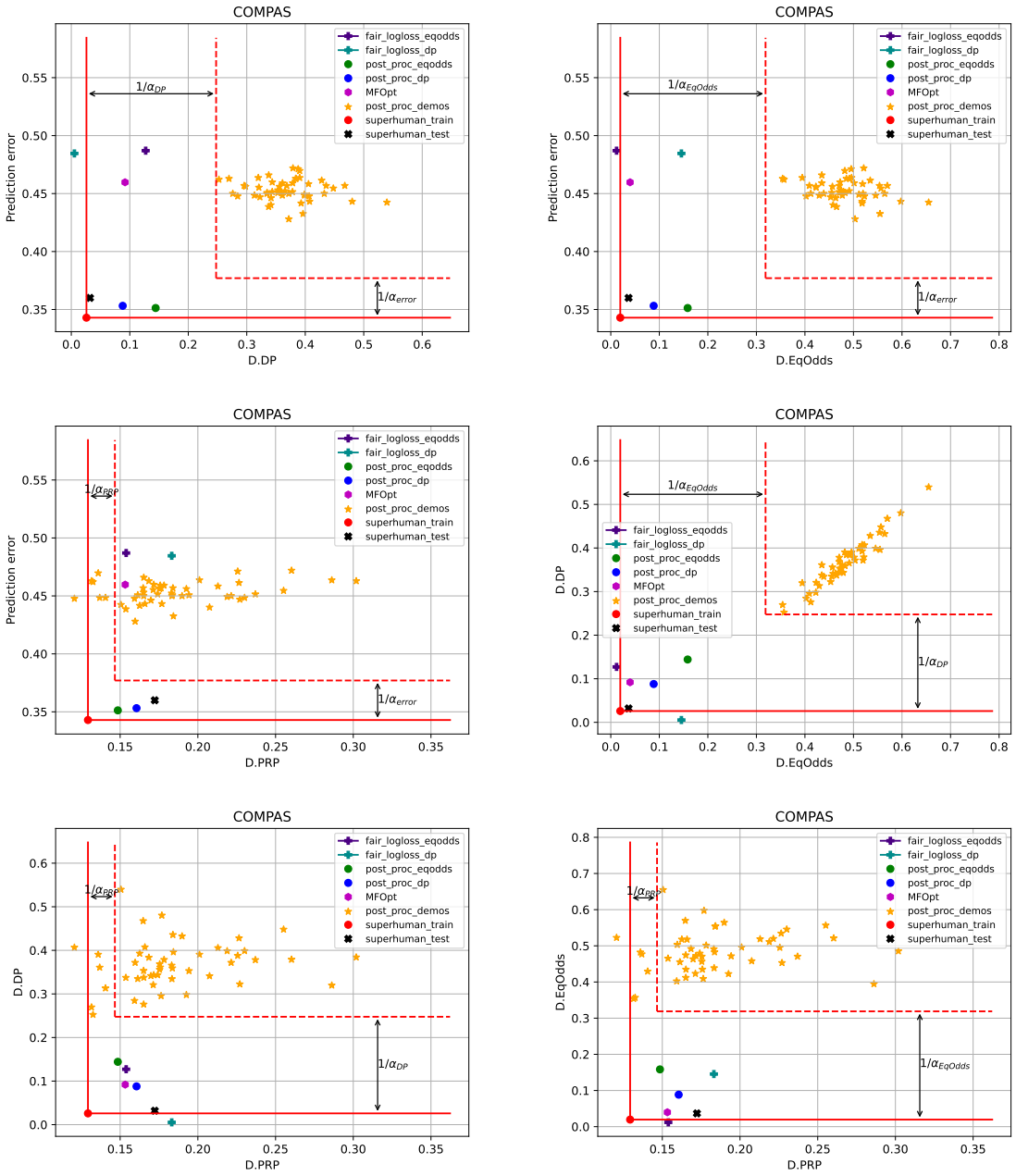


Figure 20: Experimental results on the `COMPAS` dataset with noisy demonstrations ($\epsilon = 0.2$). Margin boundaries are shown with dotted red lines. Each plot shows the relationships between two features.

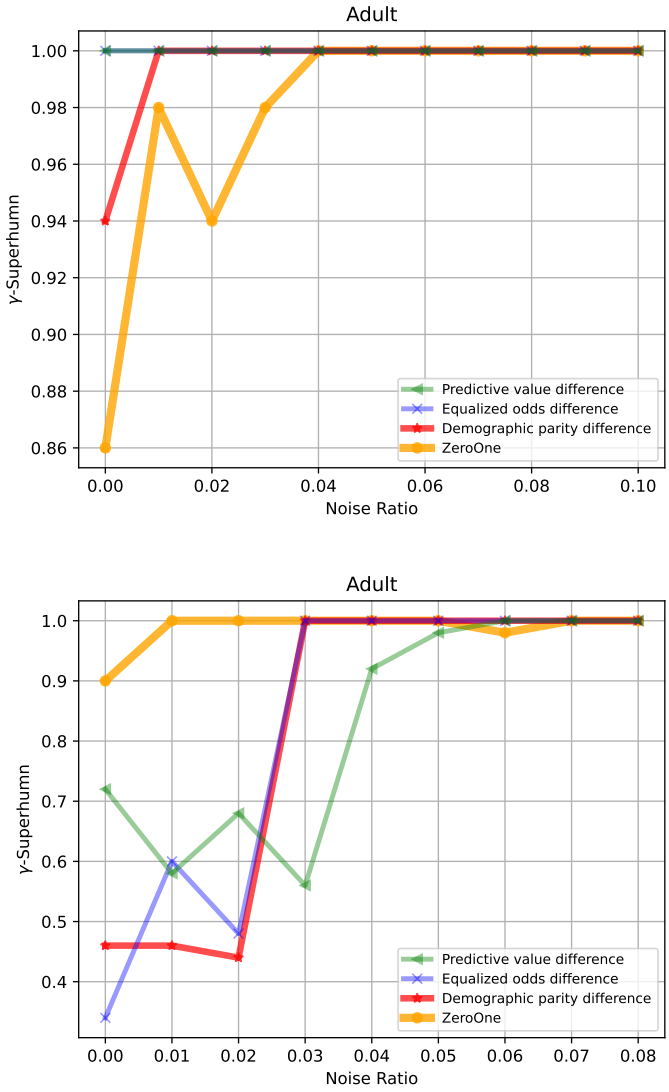


Figure 21: The relationship between the ratio of augmented noise in the label and the protected attribute of reference decisions produced by post-processing (upper) and fairlogloss (lower) and achieving γ-superhuman performance in our approach.

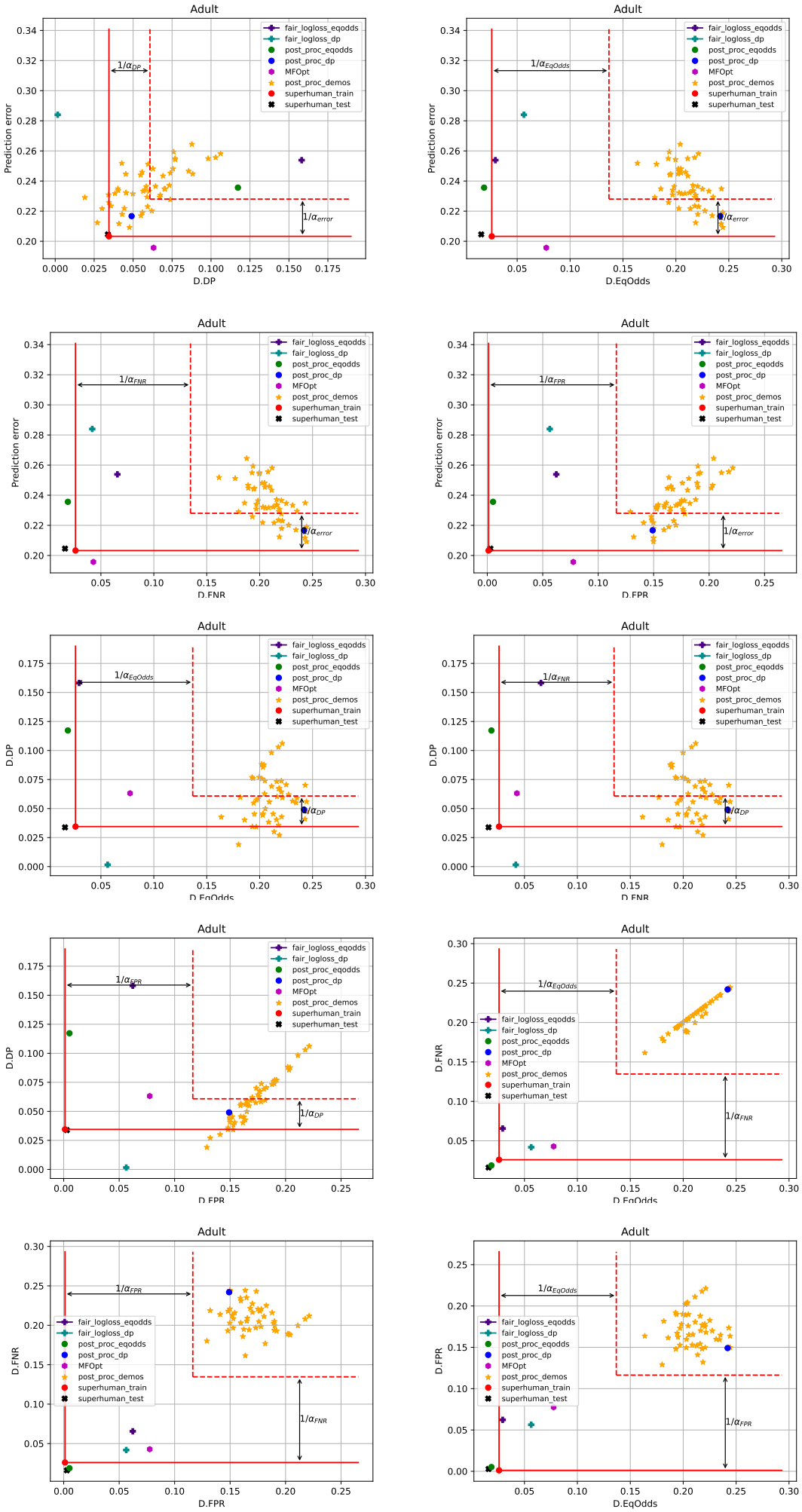


Figure 22: The trade-off between each pair of: *difference of Demographic Parity* (`D.DP`), *Equalized Odds* (`D.EqOdds`), *False Negative Rate* (`D.FNR`), *False Positive Rate* (`D.FPR`) and *Prediction Error* on test data using noiseless training data ($\epsilon = 0$) for `Adult` dataset.

CHAPTER 6

Conclusion and Future Work

6.1 Conclusion

This thesis has examined the problem of fairness in machine learning models through multiple perspectives, incorporating four distinct approaches to tackle the problem. The first three works leverage an adversarial robust framework, with two focusing on fair classification and the third addressing a structured prediction task. The fourth work takes a different approach and leverages ideas from imitation learning to build a fair classification algorithm.

In the first work, we derived a new classifier from the first principles of distributionally robust estimation, which formulated a learning objective imposing fairness requirements on the predictor and maintained training data characteristics through feature-matching constraints. The resulting parametric exponential family conditional distribution resembled a truncated logistic regression model and performed well in benchmark fairness datasets, making it quite

general due to the flexibility of its construction as a robust estimation. In the second work, we developed a novel adversarial approach for seeking fair decision-making under covariate shift, which operated appropriately even when portions of the shift between source and target distributions were extreme, in contrast with importance weighting methods. To address the lack of labeled target data points, we proposed measuring approximated fairness against a worst-case adversary constrained by source data properties and group marginals from the target. We incorporated fairness as a weighted penalty and tuned the weighted penalty to provide fairness against the adversary.

Another work leverages the adversarial robust learning framework in solving structured prediction problems. In this work, we developed a learning-to-rank system that achieves fairness of exposure for protected groups while maximizing utility to the users. Our approach constructs a minimax game with the ranker player choosing a distribution over rankings constrained to provide fairness while maximizing utility, and an adversary player choosing a distribution of item relevancies that minimizes utility while being similar to training data properties. Our method was able to trade-off between utility and fairness much better at high levels of fairness than existing baseline methods. Although our work addressed the problem of providing more robust fairness given a chosen fairness criterion, it did not provide an answer to the broader question of which fairness criterion is appropriate for a particular ranking application.

The final work takes a distinct approach, incorporating ideas from imitation learning to create a fair classification algorithm. The work introduces the concept of "superhuman fairness," which seeks to surpass human decision-making in terms of both performance and

fairness across a range of measures. This approach does not require explicit specification or elicitation of performance-fairness trade-offs. Instead, it employs subdominance minimization and policy gradient optimization methods to enable the learning of a broad class of probabilistic fairness-aware classifiers. Experimental results demonstrate the effectiveness of the approach in outperforming synthetic decisions that are affected by small amounts of label and group-membership noise when evaluated using multiple fairness criteria in conjunction with predictive accuracy.

Together, these works provide a comprehensive view of the problem of fairness in machine learning, addressing various challenges from different angles. Our contributions include novel approaches to fairness-aware log loss classification, addressing fairness under covariate shift, fair learning-to-rank, and superhuman fair classification. While our work provides important insights into ensuring fairness in machine learning models, there is still much work to be done to address the broader societal implications of these models and the trade-offs between different fairness criteria in practical applications.

In the next section, we discuss our ongoing work on Fairness-aware Bipartite Matching problem where we leverage distributionally robust learning framework.

6.2 Fairness-aware Bipartite Matching

Formulation and polynomial time complexity of solving weighted bipartite matching problems has made this task a suitable framework for a wide variety of applications: recognizing correspondences in similar images (Belongie et al., 2002; Liu et al., 2008; Rui et al., 2007), word alignment (Chan and Ng, 2008), providing ranked lists of items for information retrieval tasks (Amini et al., 2008), to name but a few. In this definition, given the two sets of elements, the goal is to find the one-to-one matching, which has the largest sum of pairwise utilities.

The machine learning modeling slightly differs from the classical combinatorial definition; instead of having weights, feature vectors corresponding to each edge are given. Therefore, machine learning methods seek to estimate the pairwise utilities of bipartite graphs so that the maximum weighted complete matching is most compatible with the (distribution of) ground truth matchings of training data.

The existing algorithms that minimize the loss are prone to amplifying existing stereotypes towards protected groups which ultimately leads to a more favorable outcome for one of the groups. For example, in the problem of applicants to jobs matching where applicants divide into male and female groups, as shown in Figure 23 (left), solving the weighted bipartite matching learned from the weights results in a matching where the difference between the sum of weights belong one protected group (males) to the other (females) is relatively high. This indicates that applicants in one of the groups are getting matched with jobs that are more aligned with their

desired attributes (e.g., jobs with higher wages). On the other hand, in Figure 23 (right), the sum of the weights for both protected groups are equal.

Aside from fairness concerns, the exponentiated potential fields models (Lafferty et al., 2001b; Petterson et al., 2009a) and maximum margin methods based on hinge loss surrogate (Taskar et al., 2005; Tsochantaridis et al., 2005) suffer from intractability when sets are large and lack fisher consistency, respectively. To address consistency and efficiency deficiencies, Fathony et al. (Fathony et al., 2018) have proposed a new approach by using an adversarial min-max game. We use this framework since doing so will enable us to introduce fairness constraints to the optimization objective. To the best of our knowledge, this is the first work that investigates fairness in maximum bipartite matching problems.

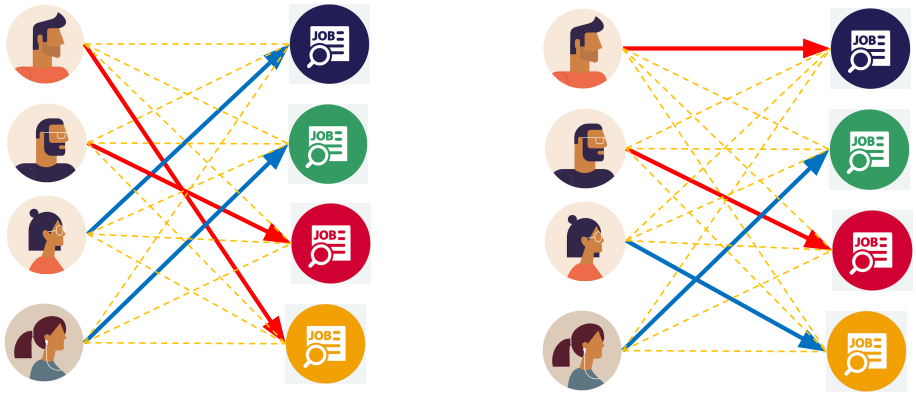
6.2.1 Approach

We propose an adversarial approach to address unfairness towards demographic groups in maximum bipartite matching problems. Our approach efficiently solves the maximum bipartite matching task while ensuring group fairness.

6.2.1.1 Adversarial Approach

In bipartite matching task, training data consists of two sets of nodes (sets M and N) with equal size and π as matching assignment. To demonstrate joint feature representations (edges) in bipartite graph x we use ψ function where $\psi_i(x, \pi_i = j)$ denotes edge weight connecting i_{th} node from set M to j_{th} node from set N. In this setting, $\psi(x, \pi)$ is defined additively over each node assignment in graph x, i.e. $\psi(x, \pi) = \sum_{i=1}^{n} \psi_i(x, \pi_i)$,

In our proposed adversarial framework, the goal is to find a predictor that 1) robustly minimizes



Matching: $\pi_1 = 4$, $\pi_2 = 3$, $\pi_3 = 1$, $\pi_4 = 2$
Given $\psi_{14} = 8$, $\psi_{23} = 6$, $\psi_{31} = 4$, $\psi_{42} = 3$
Sum: $\Sigma_i \psi_i(\pi_i) = 8 + 6 + 4 + 3 = 21$
For males: $\Sigma_i \psi_i(\pi_i) = 8 + 6 = 14$
For females: $\Sigma_i \psi_i(\pi_i) = 4 + 3 = 7$

Matching: $\pi_1 = 1$, $\pi_2 = 3$, $\pi_3 = 4$, $\pi_4 = 2$
Given $\psi_{11} = 4$, $\psi_{23} = 6$, $\psi_{34} = 4$, $\psi_{42} = 3$
Sum: $\Sigma_i \psi_i(\pi_i) = 4 + 6 + 7 + 3 = 20$
For males: $\Sigma_i \psi_i(\pi_i) = 4 + 6 = 10$
For females: $\Sigma_i \psi_i(\pi_i) = 7 + 3 = 10$

Figure 23: Two matchings: the left one maximizes sum of weights but ignores demographic parity; Matching on the right maximizes sum of weights while satisfying demographic parity.

the Hamming loss against the worst-case permutation mixture probability that is consistent with the statistics of the training data, and 2) satisfies group fairness constraints on different demographic groups (e.g. males vs females) existing on one or both set of nodes. In adversarial approach, a predictor makes a probabilistic prediction over the set of all possible assignments (denoted as $\hat{P}$). Instead of evaluating the predictor with empirical distribution, the predictor is pitted against and adversary that also makes a probabilistic prediction (denoted as $\check{P}$). The predictor's objective is to minimize the expected loss function calculated from the predictor's and adversary's probabilistic predictions, while the adversary seeks to maximize the loss. The adversary is constrained to select a probabilistic prediction that matches the statistical summaries

of the empirical training distribution (denoted as $\tilde{P}$) via moment matching constraints on joint features $\psi(x, \pi)$. To mitigate fairness issue, predictor is constrained to ensure joint feature representations are equal across two demographic groups (G_0 and G_1). For now, we define fairness constraints for one set of nodes. The formulation can be written as follows:

$$\min_{\hat{P}(\hat{\pi}|x)} \max_{\check{P}(\check{\pi}|x)} \mathbb{E}_{x\sim\tilde{P};\hat{\pi}|x\sim\hat{P};\check{\pi}|x\sim\check{P}}[\mathbf{loss}(\hat{\pi}, \check{\pi})] \tag{6.1}$$

$$\text{s.t.} \quad \|\mathbb{E}_{x\sim\tilde{P};\check{\pi}|x\sim\check{P}}[\sum_{i=1}^{n} \psi_i(x, \check{\pi}_i)] - \mathbb{E}_{(x,\pi)\sim\tilde{P}}[\sum_{i=1}^{n} \psi_i(x, \pi_i)]\| \leq \varepsilon \tag{6.2}$$

$$\|\mathbb{E}_{x\sim\tilde{P};\hat{\pi}|x\sim\hat{P}}[\sum_{i\in G_0} \phi_i(x, \hat{\pi}_i)] - \mathbb{E}_{x\sim\tilde{P};\hat{\pi}|x\sim\hat{P}}[\sum_{i\in G_1} \phi_i(x, \hat{\pi}_i)]\| \leq \delta \tag{6.3}$$

where the inequality Equation 6.3 denotes fairness constraint and δ is a threshold that indicates maximum allowed unfairness. To solve the optimization in Equation 6.1 we can use method of Lagrangian multipliers and strong duality for convex-concave saddle point problems (Von Neumann and Morgenstern, 1945; Sion, 1958). The equivalent dual formulation can be written as:

$$\min_{\theta} \max_{\lambda} \min_{\hat{P}(\hat{\pi}|x)} \max_{\check{P}(\check{\pi}|x)} \mathbb{E}_{x\sim\tilde{P};\hat{\pi}|x\sim\hat{P};\check{\pi}|x\sim\check{P}}[\mathbf{loss}(\hat{\pi}, \check{\pi})] \tag{6.4}$$

$$+ \theta^{\intercal}(\mathbb{E}_{x\sim\tilde{P};\check{\pi}|x\sim\check{P}}[\sum_{i=1}^{n} \psi_i(x, \check{\pi}_i)] - \mathbb{E}_{(x,\pi)\sim\tilde{P}}[\sum_{i=1}^{n} \psi_i(x, \pi_i)]) \tag{6.5}$$

$$+ \lambda^{\intercal}(\mathbb{E}_{x\sim\tilde{P};\hat{\pi}|x\sim\hat{P}}[\sum_{i\in G_0} \phi_i(x, \hat{\pi}_i)] - \mathbb{E}_{x\sim\tilde{P};\hat{\pi}|x\sim\hat{P}}[\sum_{i\in G_1} \phi_i(x, \hat{\pi}_i)]) \tag{6.6}$$

$$+ \varepsilon\|\theta\|_* + \delta\|\lambda\|_* \tag{6.7}$$

where θ is Lagrange dual variable for moment matching constraints and λ is penalty parameter for fairness constraints. For the loss function we use Hamming distance, $\text{loss}(\hat{\pi}, \check{\pi}) = \sum_{i=1}^{n} 1(\hat{\pi}_i \neq \check{\pi}_i)$. In this setting, if the size of game is n (number of nodes in each set), there exist $n!$ actions (permutations) for both predictor player $\hat{\pi}$ and adversarial approximator player $\check{\pi}$. This results in $\mathcal{O}(n!)$ sized game which is intractable for modestly-sized problems.

6.2.1.2 Marginal Distribution Formulation

To improve the computational efficiency of the adversarial approach, we use a marginal distribution formulation that depends only on the marginal probabilities of the assignment. This formulation leverages the Birkhoff-von Neumann theorem (Birkhoff, 1946; Von Neumann, 1953), which states that the convex hull of a set of permutation matrices constructs a convex polytope in $\mathcal{R}^{n^2}$. In this setting, the number of quantities we optimize grows quadratically, which is a significant improvement over the non-marginal approach where the space of distributions over permutations of n objects grows factorially. We use $\mathbf{P}$ and $\mathbf{Q}$ as the marginal probability matrices for the predictor and adversary, respectively, and $\mathbf{Y}$ as the ground truth permutation in the training data. We also use $\mathbf{X}_k$ as a $n \times n$ matrix to represent the k_{th} feature of $\psi_i(x, \pi_i = j)$.

Appendix A

Appendix

A.1 Copyright Policy of Association for the Advancement of Artificial Intelligence (AAAI)

1. Author(s) agree to transfer their copyrights in their article/paper to the Association for the Advancement of Artificial Intelligence (AAAI), in order to deal with future requests for reprints, translations, anthologies, reproductions, excerpts, and other publications. This grant will include, without limitation, the entire copyright in the article/paper in all countries of the world, including all renewals, extensions, and reversions thereof, whether such rights current exist or hereafter come into effect, and also the exclusive right to create electronic versions of the article/paper, to the extent that such right is not subsumed under copyright.

APPENDIX (Continued)

2. The author(s) warrants that they are the sole author and owner of the copyright in the above article/paper, except for those portions shown to be in quotations; that the article/paper is original throughout; and that the undersigned right to make the grants set forth above is complete and unencumbered.

3. The author(s) agree that if anyone brings any claim or action alleging facts that, if true, constitute a breach of any of the foregoing warranties, the author(s) will hold harmless and indemnify AAAI, their grantees, their licensees, and their distributors against any liability, whether under judgment, decree, or compromise, and any legal fees and expenses arising out of that claim or actions, and the undersigned will cooperate fully in any defense AAAI may make to such claim or action. Moreover, the undersigned agrees to cooperate in any claim or other action seeking to protect or enforce any right the undersigned has granted to AAAI in the article/paper. If any such claim or action fails because of facts that constitute a breach of any of the foregoing warranties, the undersigned agrees to reimburse whomever brings such claim or action for expenses and attorneys' fees incurred therein.

4. Author(s) retain all proprietary rights other than copyright (such as patent rights).

5. Author(s) may make personal reuse of all or portions of the above article/paper in other works of their own authorship.

6. Author(s) may reproduce, or have reproduced, their article/paper for the author's personal use, or for company use provided that AAAI copyright and the source are indicated, and that the copies are not used in a way that implies AAAI endorsement of a product or service of an employer, and that the copies per se are not offered for sale. The foregoing right shall not permit

the posting of the article/paper in electronic or digital form on any computer network, except by the author or the author's employer, and then only on the author's or the employer's own web page or ftp site. Such web page or ftp site, in addition to the aforementioned requirements of this Paragraph, must provide an electronic reference or link back to the AAAI electronic server, and shall not post other AAAI copyrighted materials not of the author's or the employer's creation (including tables of contents with links to other papers) without AAAI's written permission.

7. Author(s) may make limited distribution of all or portions of their article/paper prior to publication.

8. In the case of work performed under U.S. Government contract, AAAI grants the U.S. Government royalty-free permission to reproduce all or portions of the above article/paper, and to authorize others to do so, for U.S. Government purposes.

9. In the event the above article/paper is not accepted and published by AAAI, or is withdrawn by the author(s) before acceptance by AAAI, this agreement becomes null and void.

A.2 Copyright Policy of the International Conference on Machine Learning (ICML)

The International Conference on Machine Learning (ICML) conference's proceeding is published by the Proceedings of Machine Learning Research (PMLR).

The Proceedings of Machine Learning Research (formerly JMLR Workshop and Conference Proceedings) is a series aimed specifically at publishing machine learning research presented at workshops and conferences. Each volume is separately titled and associated with a particular

APPENDIX (Continued)

workshop or conference and will be published online on the PMLR web site. Authors will retain copyright and individual volume editors are free to make additional hardcopy publishing arrangements (see for example the Challenges in Machine Learning series which includes free PDFs and low cost hard copies), but PMLR will not produce hardcopies of these volumes.

Licence to Publish
Proceedings Papers

SPRINGER NATURE

Licensee	Springer Nature Switzerland AG	(the 'Licensee')
Title of the Proceedings Volume/Edited Book or Conference Name:	Proceedings of The 27th Pacific-Asia Conference on Knowledge Discovery and Data Mining (PAKDD 2023)	(the 'Volume')
Volume Editor(s) Name(s):	Hisashi Kashima, Wen-Chih Peng, Tsuyoshi Ide	
Proposed Title of the Contribution:	Fairness for Robust Learning to Rank	(the 'Contribution')
Series: The Contribution may be published in the following series	A Springer Nature Computer Science book series (CCIS, LNAI, LNBI, LNBIP or LNCS)	
Author(s) Full Name(s):	Omid Memarrast, Ashkan Rezaei, Rizal Fathony, Brian Ziebart	(the 'Author')

When Author is more than one person the expression "Author" as used in this Agreement will apply collectively unless otherwise indicated.

Corresponding Author Name:	Omid Memarrast	
Instructions for Authors	https://resource-cms.springernature.com/springer-cms/rest/v1/content/19242230/data/	(the 'Instructions for Authors')

1 Grant of Rights

a) For good and valuable consideration, the Author hereby grants to the Licensee the perpetual, exclusive, world-wide, assignable, sublicensable and unlimited right to: publish, reproduce, copy, distribute, communicate, display publicly, sell, rent and/or otherwise make available the contribution identified above, including any supplementary information and graphic elements therein (e.g. illustrations, charts, moving images) (the 'Contribution') in any language, in any versions or editions in any and all forms and/or media of expression (including without limitation in connection with any and all end-user devices), whether now known or developed in the future. Without limitation, the above grant includes: (i) the right to edit, alter, adapt, adjust and prepare derivative works; (ii) all advertising and marketing rights including without limitation in relation to social media; (iii) rights for any training, educational and/or instructional purposes; (iv) the right to add and/or remove links or combinations with other media/works; and (v) the right to create, use and/or license and/or sublicense content data or metadata of any kind in relation to the Contribution (including abstracts and summaries) without restriction. The above rights are granted in relation to the Contribution as a whole or any part and with or in relation to any other works.

b) Without limiting the rights granted above, Licensee is granted the rights to use the Contribution for the purposes of analysis, testing, and development of publishing- and research-related workflows, systems, products, projects, and services; to confidentially share the Contribution with select third parties to do the same; and to retain and store the Contribution and any associated correspondence/files/forms to maintain the historical record, and to facilitate research integrity investigations. The grant of rights set forth in

this clause (b) is irrevocable.

c) If the Licensee elects not to publish the Contribution for any reason, all publishing rights under this Agreement as set forth in clause 1a above will revert to the Author.

2 Copyright

Ownership of copyright in the Contribution will be vested in the name of the Author. When reproducing the Contribution or extracts from it, the Author will acknowledge and reference first publication in the Volume.

3 Use of Contribution Versions

a) For purposes of this Agreement: (i) references to the "Contribution" include all versions of the Contribution; (ii) "Submitted Manuscript" means the version of the Contribution as first submitted by the Author prior to peer review; (iii) "Accepted Manuscript" means the version of the Contribution accepted for publication, but prior to copy-editing and typesetting; and (iv) "Version of Record" means the version of the Contribution published by the Licensee, after copy-editing and typesetting. Rights to all versions of the Manuscript are granted on an exclusive basis, except for the Submitted Manuscript, to which rights are granted on a non-exclusive basis.

b) The Author may make the Submitted Manuscript available at any time and under any terms (including, but not limited to, under a CC BY licence), at the Author's discretion. Once the Contribution has been published, the Author will include an acknowledgement and provide a link to the Version of Record on the publisher's website: "This preprint has not undergone peer review (when applicable) or any post-submission improvements or corrections. The Version of Record of this contribution is published in [insert volume title], and is available online at https://doi.org/[insert DOI]".

c) The Licensee grants to the Author (i) the right to make the Accepted Manuscript available on their own personal, self-maintained website immediately on acceptance, (ii) the right to make the Accepted Manuscript available for public release on any of the following twelve (12) months after first publication (the "Embargo Period"): their employer's internal website; their institutional and/or funder repositories. Accepted Manuscripts may be deposited in such repositories immediately upon acceptance, provided they are not made publicly available until after the Embargo Period.
The rights granted to the Author with respect to the Accepted Manuscript are subject to the conditions that (i) the Accepted Manuscript is not enhanced or substantially reformatted by the Author or any third party, and (ii) the Author includes on the Accepted Manuscript an acknowledgement in the following form, together with a link to the published version on the publisher's website: "This version of the contribution has been accepted for publication, after peer review (when applicable) but is not the Version of Record and does not reflect post-acceptance improvements, or any corrections. The Version of Record is available online at: http://dx.doi.org/[insert DOI]. Use of this Accepted Version is subject to the publisher's Accepted Manuscript terms of use https://www.springernature.com/gp/open-research/policies/accepted-manuscript-terms".
Under no circumstances may an Accepted Manuscript be shared or distributed under a Creative Commons or other form of open access licence.
Any use of the Accepted Manuscript not expressly permitted under this subclause (c) is

subject to the Licensee's prior consent.

d) The Licensee grants to Author the following non-exclusive rights to the Version of Record, provided that, when reproducing the Version of Record or extracts from it, the Author acknowledges and references first publication in the Volume according to current citation standards. As a minimum, the acknowledgement must state: "First published in [Volume, page number, year] by Springer Nature".

i. to reuse graphic elements created by the Author and contained in the Contribution, in presentations and other works created by them;

ii. the Author and any academic institution where they work at the time may reproduce the Contribution for the purpose of course teaching (but not for inclusion in course pack material for onward sale by libraries and institutions);

iii. to reuse the Version of Record or any part in a thesis written by the same Author, and to make a copy of that thesis available in a repository of the Author(s)' awarding academic institution, or other repository required by the awarding academic institution. An acknowledgement should be included in the citation: "Reproduced with permission from Springer Nature";

iv. to reproduce, or to allow a third party to reproduce the Contribution, in whole or in part, in any other type of work (other than thesis) written by the Author for distribution by a publisher after an embargo period of 12 months; and

v. to publish an expanded version of their Contribution provided the expanded version (i) includes at least 30% new material (ii) includes an express statement specifying the incremental change in the expanded version (e.g., new results, better description of materials, etc.).

4 Warranties & Representations

Author warrants and represents that:

a)

i. the Author is the sole copyright owner or has been authorised by any additional copyright owner(s) to grant the rights defined in clause 1,

ii. the Contribution does not infringe any intellectual property rights (including without limitation copyright, database rights or trade mark rights) or other third party rights and no licence from or payments to a third party are required to publish the Contribution,

iii. the Contribution has not been previously published or licensed, nor has the Author committed to licensing any version of the Contribution under a licence inconsistent with the terms of this Agreement,

iv. if the Contribution contains materials from other sources (e.g. illustrations, tables, text quotations), Author has obtained written permissions to the extent necessary from the copyright holder(s), to license to the Licensee the same rights as set out in clause 1 but on a non-exclusive basis and without the right to use any graphic

elements on a stand-alone basis and has cited any such materials correctly;

b) all of the facts contained in the Contribution are according to the current body of research true and accurate;

c) nothing in the Contribution is obscene, defamatory, violates any right of privacy or publicity, infringes any other human, personal or other rights of any person or entity or is otherwise unlawful and that informed consent to publish has been obtained for any research participants;

d) nothing in the Contribution infringes any duty of confidentiality owed to any third party or violates any contract, express or implied, of the Author;

e) all institutional, governmental, and/or other approvals which may be required in connection with the research reflected in the Contribution have been obtained and continue in effect;

f) all statements and declarations made by the Author in connection with the Contribution are true and correct;

g) the signatory who has signed this Agreement has full right, power and authority to enter into this Agreement on behalf of all of the Authors; and

h) the Author complies in full with: i. all instructions and policies in the Instructions for Authors, ii. the Licensee's ethics rules (available at https://www.springernature.com/gp/authors/book-authors-code-of-conduct), as may be updated by the Licensee at any time in its sole discretion.

5 Cooperation

a) The Author will cooperate fully with the Licensee in relation to any legal action that might arise from the publication of the Contribution, and the Author will give the Licensee access at reasonable times to any relevant accounts, documents and records within the power or control of the Author. The Author agrees that any Licensee affiliate through which the Licensee exercises any rights or performs any obligations under this Agreement is intended to have the benefit of and will have the right to enforce the terms of this Agreement.

b) Author authorises the Licensee to take such steps as it considers necessary at its own expense in the Author's name(s) and on their behalf if the Licensee believes that a third party is infringing or is likely to infringe copyright in the Contribution including but not limited to initiating legal proceedings.

6 Author List

Changes of authorship, including, but not limited to, changes in the corresponding author or the sequence of authors, are not permitted after acceptance of a manuscript.

7 Post Publication Actions

The Author agrees that the Licensee may remove or retract the Contribution or publish a correction or other notice in relation to the Contribution if the Licensee determines that such

actions are appropriate from an editorial, research integrity, or legal perspective.

8 Controlling Terms

The terms of this Agreement will supersede any other terms that the Author or any third party may assert apply to any version of the Contribution.

9 Governing Law

This Agreement shall be governed by, and shall be construed in accordance with, the laws of Switzerland. The courts of Zug, Switzerland shall have the exclusive jurisdiction.

Signed for and on behalf of the Author	Print Name:	Date:
	Omid Memarrast	03/22/2023

Address:	500 W Fullerton Pkwy Apt 301, Chicago, IL 60614
Email:	omemar2@uic.edu

CITED LITERATURE

Abbeel, P. and Ng, A. Y.: Apprenticeship learning via inverse reinforcement learning. In Proceedings of the International Conference on Machine Learning, pages 1–8, 2004.

Adel, T., Valera, I., Ghahramani, Z., and Weller, A.: One-network adversarial fairness. In AAAI, 2019.

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. M.: A reductions approach to fair classification. In ICML, 2018.

Aghaei, S., Azizi, M. J., and Vayanos, P.: Learning optimal and fair decision trees for non-discriminative decision-making. In AAAI Conference on Artificial Intelligence, volume 33, pages 1418–1426, 2019.

Amini, M., Truong, T., and Goutte, C.: A boosting algorithm for learning bipartite ranking functions with partially labeled data. In Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2008, Singapore, July 20-24, 2008, eds. S. Myaeng, D. W. Oard, F. Sebastiani, T. Chua, and M. Leong, pages 99–106. ACM, 2008.

Asif, K., Xing, W., Behpour, S., and Ziebart, B. D.: Adversarial cost-sensitive classification. In UAI, 2015.

Asudeh, A., Jagadish, H., Stoyanovich, J., and Das, G.: Designing fair ranking schemes. In Proceedings of the 2019 International Conference on Management of Data, pages 1259–1276, 2019.

Basu, K., DiCiccio, C., Logan, H., and Karoui, N. E.: A framework for fairness in two-sided marketplaces. arXiv preprint arXiv:2006.12756, 2020.

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

Belongie, S. J., Malik, J., and Puzicha, J.: Shape matching and object recognition using shape contexts. IEEE Trans. Pattern Anal. Mach. Intell., 24(4):509–522, 2002.

Ben-Tal, A., Den Hertog, D., De Waegenaere, A., Melenberg, B., and Rennen, G.: Robust solutions of optimization problems affected by uncertain probabilities. Management Science, 59(2):341–357, 2013.

Beutel, A., Chen, J., Doshi, T., Qian, H., Wei, L., Wu, Y., Heldt, L., Zhao, Z., Hong, L., Chi, E. H., et al.: Fairness in recommendation ranking through pairwise comparisons. In ACM SIGKDD International Conference on Knowledge Discovery & Data Mining(KDD), pages 2212–2220, 2019.

Biega, A. J., Gummadi, K. P., and Weikum, G.: Equity of attention: Amortizing individual fairness in rankings. In The 41st international acm sigir conference on research & development in information retrieval, pages 405–414, 2018.

Binns, R.: Fairness in machine learning: Lessons from political philosophy. In Conference on Fairness, Accountability and Transparency, pages 149–159. PMLR, 2018.

Birkhoff, G.: Three observations on linear algebra. Univ. Nac. Tacuman, Rev. Ser. A, 5:147–151, 1946.

Blum, A. and Stangl, K.: Recovering from biased data: Can fairness constraints improve accuracy? arXiv preprint arXiv:1912.01094, 2019.

Bose, I. and Mahapatra, R. K.: Business data mining—a machine learning perspective. Information & management, 2001.

Bower, A., Eftekhari, H., Yurochkin, M., and Sun, Y.: Individually fair rankings. In International Conference on Learning Representations, 2020.

Boyd, S., Parikh, N., and Chu, E.: Distributed optimization and statistical learning via the alternating direction method of multipliers. Now Publishers Inc, 2011.

Boyd, S. and Vandenberghe, L.: Convex optimization. Cambridge University Press, 2004.

Byrd, R. H., Lu, P., Nocedal, J., and Zhu, C.: A limited memory algorithm for bound constrained optimization. SIAM Journal on scientific computing, 16(5):1190–1208, 1995.

Calders, T., Kamiran, F., and Pechenizkiy, M.: Building classifiers with independency constraints. In ICDMW '09, 2009.

Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., and Varshney, K. R.: Optimized pre-processing for discrimination prevention. In NeurIPS, 2017.

Cao, Z., Qin, T., Liu, T.-Y., Tsai, M.-F., and Li, H.: Learning to rank: from pairwise approach to listwise approach. In Proceedings of the 24th international conference on Machine learning, pages 129–136, 2007.

Carter, C. and Catlett, J.: Assessing credit card applications using machine learning. IEEE Expert, 1987.

Celis, L. E., Huang, L., Keswani, V., and Vishnoi, N. K.: Classification with fairness constraints: A meta-algorithm with provable guarantees. In ACM FAT*, 2019.

Celis, L. E. and Keswani, V.: Improved adversarial learning for fair classification. arXiv preprint arXiv:1901.10443, 2019.

Celis, L. E., Mehrotra, A., and Vishnoi, N. K.: Interventions for ranking in the presence of implicit bias. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pages 369–380, 2020.

Celis, L. E., Straszak, D., and Vishnoi, N. K.: Ranking with fairness constraints. In 45th International Colloquium on Automata, Languages, and Programming (ICALP 2018). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

Chan, Y. S. and Ng, H. T.: MAXSIM: A maximum similarity metric for machine translation evaluation. In ACL 2008, Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics, June 15-20, 2008, Columbus, Ohio, USA, eds. K. R. McKeown, J. D. Moore, S. Teufel, J. Allan, and S. Furui, pages 55–62. The Association for Computer Linguistics, 2008.

Chang, L.: Applying data mining to predict college admissions yield: A case study. NDIR, 2006.

Chen, L. and Pu, P.: Survey of preference elicitation methods. Technical report, EPFL, 2004.

Chen, R. and Paschalidis, I. C.: A robust learning approach for regression models based on distributionally robust optimization. Journal of Machine Learning Research, 19(13), 2018.

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

Chouldechova, A.: Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. CoRR, abs/1703.00056, 2017.

Christmann, A. and Steinwart, I.: On robustness properties of convex risk minimization methods for pattern recognition. The Journal of Machine Learning Research, 5:1007–1034, 2004.

Cortes, C., Mansour, Y., and Mohri, M.: Learning bounds for importance weighting. In Advances in neural information processing systems, pages 442–450, 2010.

Cortes, C. and Vapnik, V.: Support-vector networks. Machine learning, 20:273–297, 1995.

Cotter, A., Jiang, H., Wang, S., Narayan, T., Gupta, M., You, S., and Sridharan, K.: Optimization with non-differentiable constraints with applications to fairness, recall, churn, and other goals. arXiv preprint, 2018.

Cox, D. R.: The regression analysis of binary sequences. Journal of the Royal Statistical Society: Series B (Methodological), 20(2):215–232, 1958.

Delage, E. and Ye, Y.: Distributionally robust optimization under moment uncertainty with application to data-driven problems. Operations research, 58(3):595–612, 2010.

Deng, W. and Yin, W.: On the global and linear convergence of the generalized alternating direction method of multipliers. Journal of Scientific Computing, 66(3), 2016.

Dheeru, D. and Karra Taniskidou, E.: UCI machine learning repository, 2017. Accessed on June 2020.

Diaz, F., Mitra, B., Ekstrand, M. D., Biega, A. J., and Carterette, B.: Evaluating stochastic rankings with expected exposure. In Proceedings of the 29th ACM International Conference on Information & Knowledge Management, pages 275–284, 2020.

Do, V., Corbett-Davies, S., Atif, J., and Usunier, N.: Two-sided fairness in rankings via lorenz dominance. Advances in Neural Information Processing Systems, 34:8596–8608, 2021.

Donaldson, T. and Preston, L. E.: The stakeholder theory of the corporation: Concepts, evidence, and implications. Academy of management Review, 20(1):65–91, 1995.

Donini, M., Oneto, L., Ben-David, S., Shawe-Taylor, J. S., and Pontil, M.: Empirical risk minimization under fairness constraints. In NeurIPS, 2018.

Dowling, A. W., Ruiz-Mercado, G., and Zavala, V. M.: A framework for multi-stakeholder decision-making and conflict resolution. Computers & Chemical Engineering, 90:136–150, 2016.

Duchi, J., Shalev-Shwartz, S., Singer, Y., and Chandra, T.: Efficient projections onto the l 1-ball for learning in high dimensions. In Proceedings of the 25th international conference on Machine learning, pages 272–279. ACM, 2008.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R.: Fairness through awareness. In ITCS, 2012.

Eckhouse, L.: Big data may be reinforcing racial bias in the criminal justice system. The Washington Post, 2017.

Esfahani, P. M. and Kuhn, D.: Data-driven distributionally robust optimization using the wasserstein metric: Performance guarantees and tractable reformulations. Mathematical Programming, 171(1):115–166, 2018.

Farnia, F. and Tse, D.: A minimax approach to supervised learning. Advances in Neural Information Processing Systems, 29:4240–4248, 2016.

Fathony, R., Asif, K., Liu, A., Bashiri, M. A., Xing, W., Behpour, S., Zhang, X., and Ziebart, B. D.: Consistent robust adversarial prediction for general multiclass classification. arXiv preprint, 2018.

Fathony, R., Bashiri, M. A., and Ziebart, B.: Adversarial surrogate losses for ordinal regression. In NeurIPS, 2017.

Fathony, R., Behpour, S., Zhang, X., and Ziebart, B. D.: Efficient and consistent adversarial bipartite matching. In Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018, eds. J. G.

Dy and A. Krause, volume 80 of Proceedings of Machine Learning Research, pages 1456–1465. PMLR, 2018.

Fathony, R., Liu, A., Asif, K., and Ziebart, B.: Adversarial multiclass classification: A risk minimization perspective. In NeurIPS, 2016.

Fathony, R., Rezaei, A., Bashiri, M., Zhang, X., and Ziebart, B. D.: Distributionally robust graphical models. In NeurIPS, 2018.

Fehrman, E., Muhammad, A. K., Mirkes, E. M., Egan, V., and Gorban, A. N.: The five factor model of personality and evaluation of drug consumption risk. In Data science, pages 231–242. Springer, 2017.

Friedman, B. and Nissenbaum, H.: Bias in computer systems. ACM Transactions on Information Systems (TOIS), 14(3):330–347, 1996.

Geyik, S. C., Ambler, S., and Kenthapadi, K.: Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In International Conference on Knowledge Discovery & Data Mining, 2019.

Ghosh, A., Dutt, R., and Wilson, C.: When fair ranking meets uncertain inference. arXiv preprint arXiv:2105.02091, 2021.

Goel, N., Yaghini, M., and Faltings, B.: Non-discriminatory machine learning through convex fairness criteria. In AAAI, 2018.

Gretton, A., Smola, A., Huang, J., Schmittfull, M., Borgwardt, K., and Schölkopf, B.: Covariate shift by kernel mean matching. Dataset shift in machine learning, 3(4):5, 2009.

Grünwald, P. D. and Dawid, A. P.: Game theory, maximum entropy, minimum discrepancy, and robust Bayesian decision theory. Annals of Statistics, 32, 2004.

Hacker, P. and Wiedemann, E.: A continuous framework for fairness. arXiv preprint, 2017.

Hardt, M., Price, E., and Srebro, N.: Equality of opportunity in supervised learning. In NeurIPS, 2016.

Hashimoto, T., Srivastava, M., Namkoong, H., and Liang, P.: Fairness without demographics in repeated loss minimization. In ICML, 2018.

Hiranandani, G., Narasimhan, H., and Koyejo, S.: Fair performance metric elicitation. Advances in Neural Information Processing Systems, 33:11083–11095, 2020.

Höffgen, K.-U. and Simon, H. U.: Robust trainability of single neurons. In Proceedings of the fifth annual workshop on Computational learning theory, pages 428–439, 1992.

Hsu, B., Mazumder, R., Nandy, P., and Basu, K.: Pushing the limits of fairness impossibility: Who's the fairest of them all? In Advances in Neural Information Processing Systems, 2022.

Hutchinson, B. and Mitchell, M.: 50 years of test (un) fairness: Lessons for machine learning. In Proceedings of the Conference on Fairness, Accountability, and Transparency, pages 49–58, 2019.

Järvelin, K. and Kekäläinen, J.: Cumulated gain-based evaluation of ir techniques. ACM Transactions on Information Systems (TOIS), 20(4):422–446, 2002.

Jaynes, E. T.: Information theory and statistical mechanics. Physical review, 106(4), 1957.

Kabakchieva, D.: Predicting student performance by using data mining methods for classification. Cybernetics and Information Technologies, 13(1), 2013.

Kallus, N. and Zhou, A.: The fairness of risk scores beyond classification: Bipartite ranking and the xauc metric. In Advances in Neural Information Processing Systems, pages 3438–3448, 2019.

Kamiran, F. and Calders, T.: Data preprocessing techniques for classification without discrimination. Knowledge and Information Systems, 33(1), Oct 2012.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J.: Fairness-aware classifier with prejudice remover regularizer. In Joint European Conference on Machine Learning and Knowledge Discovery in Databases, pages 35–50. Springer, 2012.

Kamishima, T., Akaho, S., and Sakuma, J.: Fairness-aware learning through regularization approach. In ICDMW, 2011.

Kleinberg, J., Mullainathan, S., and Raghavan, M.: Inherent trade-offs in the fair determination of risk scores. arXiv preprint arXiv:1609.05807, 2016.

Kleinberg, J. and Raghavan, M.: Selection problems in the presence of implicit bias. In 9th Innovations in Theoretical Computer Science Conference (ITCS 2018), 2018.

Kleinberg, J. M., Mullainathan, S., and Raghavan, M.: Inherent trade-offs in the fair determination of risk scores. CoRR, abs/1609.05807, 2016.

Kuhn, H. W.: The hungarian method for the assignment problem. Naval research logistics quarterly, 2(1-2):83–97, 1955.

Lafferty, J., McCallum, A., and Pereira, F. C.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. 2001.

Lafferty, J. D., McCallum, A., and Pereira, F. C. N.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the Eighteenth International Conference on Machine Learning (ICML 2001), Williams College, Williamstown, MA, USA, June 28 - July 1, 2001, eds. C. E. Brodley and A. P. Danyluk, pages 282–289. Morgan Kaufmann, 2001.

Lahoti, P., Gummadi, K. P., and Weikum, G.: Operationalizing individual fairness with pairwise fair representations. Proceedings of the VLDB Endowment, 13(4):506–518, 2019.

Landeau, A.: Measuring fairness in machine learning models. `https://blog.dataiku.com/measuring-fairness-in-machine-learning-models`, 2020. (Accessed on 07/17/2023).

Larson, J., Mattu, S., Kirchner, L., and Angwin, J.: How we analyzed the compas recidivism algorithm. ProPublica, 9, 2016.

Lisman, J. and Zuylen, M. v.: Note on the generation of most probable frequency distributions. Statistica Neerlandica, 26(1), 1972.

Liu, A. and Ziebart, B.: Robust classification under sample selection bias. In NeurIPS, 2014.

Liu, L., Sun, L., Rui, Y., Shi, Y., and Yang, S.: Web video topic discovery and tracking via bipartite graph reinforcement model. In Proceedings of the 17th International Conference on World Wide Web, WWW 2008, Beijing, China, April 21-25, 2008, eds. J. Huai, R. Chen, H. Hon, Y. Liu, W. Ma, A. Tomkins, and X. Zhang, pages 1009–1018. ACM, 2008.

Liu, S. and Vicente, L. N.: Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach. Computational Management Science, pages 1–25, 2022.

Livni, R., Crammer, K., and Globerson, A.: A simple geometric interpretation of svm using stochastic adversaries. In Artificial Intelligence and Statistics, pages 722–730. PMLR, 2012.

Lohr, S.: Big data, trying to build better workers. The New York Times, 21, 2013.

Lowry, S. and Macpherson, G.: A blot on the profession. British Medical Journal (Clinical research ed.), 1988.

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

Manning, C. and Klein, D.: Optimization, maxent models, and conditional estimation without magic. In NAACL, 2003.

Martinez, N., Bertran, M., and Sapiro, G.: Minimax Pareto fairness: A multi objective perspective. In Proceedings of the International Conference on Machine Learning, pages 6755–6764. PMLR, 13–18 Jul 2020.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A.: A survey on bias and fairness in machine learning. arXiv preprint arXiv:1908.09635, 2019.

Mehrotra, A. and Celis, L. E.: Mitigating bias in set selection with noisy protected attributes. In Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, pages 237–248, 2021.

Mehrotra, A. and Vishnoi, N.: Fair ranking with noisy protected attributes. Advances in Neural Information Processing Systems, 35:31711–31725, 2022.

Memarrast, O., Rezaei, A., Fathony, R., and Ziebart, B.: Fairness for robust learning to rank. arXiv preprint arXiv:2112.06288, 2021.

Memarrast, O., Rezaei, A., Fathony, R., and Ziebart, B.: Fairness for robust learning to rank. In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pages 544–556. Springer, 2023.

Memarrast, O., Vu, L., and Ziebart, B. D.: Superhuman fairness. In International Conference on Machine Learning, pages 24420–24435. PMLR, 2023.

Memarrast, O., Vu, L., and Ziebart, B. D.: Superhuman fairness. In ICLR 2023 Workshop on Pitfalls of limited data and computation for Trustworthy ML, 2023.

Menon, A. K. and Williamson, R. C.: The cost of fairness in binary classification. In ACM FAT*, 2018.

Minsker, S. and Mathieu, T.: Excess risk bounds in robust empirical risk minimization. arXiv preprint arXiv:1910.07485, 2019.

Moses, L. B. and Chan, J.: Using big data for legal and law enforcement decisions: Testing the new tools. UNSWLJ, 2014.

Namkoong, H. and Duchi, J. C.: Stochastic gradient methods for distributionally robust optimization with f-divergences. In NIPS, volume 29, pages 2208–2216, 2016.

Namkoong, H. and Duchi, J. C.: Variance-based regularization with convex objectives. In NIPS, 2017.

Narasimhan, H., Cotter, A., Gupta, M. R., and Wang, S.: Pairwise fairness for ranking and regression. In AAAI, pages 5248–5255, 2020.

Noble, S. U.: Algorithms of oppression: How search engines reinforce racism. nyu Press, 2018.

Obermeyer, Z. and Emanuel, E. J.: Predicting the future—big data, machine learning, and clinical medicine. The New England Journal of Medicine, 375(13), 2016.

O'Neil, C.: Weapons of math destruction: How big data increases inequality and threatens democracy. Broadway Books, 2016.

Oosterhuis, H.: Computationally efficient optimization of plackett-luce ranking models for relevance and fairness. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 1023–1032, 2021.

Osa, T., Pajarinen, J., Neumann, G., Bagnell, J. A., Abbeel, P., Peters, J., et al.: An algorithmic perspective on imitation learning. Foundations and Trends® in Robotics, 7(1-2):1–179, 2018.

Patro, G. K., Biswas, A., Ganguly, N., Gummadi, K. P., and Chakraborty, A.: Fairrec: Two-sided fairness for personalized recommendations in two-sided platforms. In Proceedings of The Web Conference 2020, pages 1194–1204, 2020.

Patro, G. K., Chakraborty, A., Ganguly, N., and Gummadi, K.: Incremental fairness in two-sided market platforms: on smoothly updating recommendations. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 181–188, 2020.

Petterson, J., Caetano, T. S., McAuley, J. J., and Yu, J.: Exponential family graph matching and ranking. In Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada, eds. Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, pages 1455–1463. Curran Associates, Inc., 2009.

Petterson, J., Yu, J., McAuley, J., and Caetano, T.: Exponential family graph matching and ranking. Advances in Neural Information Processing Systems, 22, 2009.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q.: On fairness and calibration. In NeurIPS, 2017.

Prost, F., Awasthi, P., Blumm, N., Kumthekar, A., Potter, T., Wei, L., Wang, X., Chi, E. H., Chen, J., and Beutel, A.: Measuring model fairness under noisy covariates: A theoretical perspective. arXiv preprint arXiv:2105.09985, 2021.

Qin, T. and Liu, T.-Y.: Introducing letor 4.0 datasets. arXiv preprint arXiv:1306.2597, 2013.

Rezaei, A., Fathony, R., Memarrast, O., and Ziebart, B.: Fairness for robust log loss classification. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, pages 5511–5518, 2020.

Rezaei, A., Liu, A., Memarrast, O., and Ziebart, B. D.: Robust fairness under covariate shift. In Proceedings of the AAAI Conference on Artificial Intelligence, volume 35, pages 9419–9427, 2021.

Robertson, S. E.: The probability ranking principle in ir. Journal of documentation, 1977.

Rui, X., Li, M., Li, Z., Ma, W., and Yu, N.: Bipartite graph reinforcement model for web image annotation. In Proceedings of the 15th International Conference on Multimedia 2007, Augsburg, Germany, September 24-29, 2007, eds. R. Lienhart, A. R. Prasad, A. Hanjalic, S. Choi, B. P. Bailey, and N. Sebe, pages 585–594. ACM, 2007.

Saxena, N. A., Huang, K., DeFilippis, E., Radanovic, G., Parkes, D. C., and Liu, Y.: How do fairness definitions fare? examining public attitudes towards algorithmic definitions of

fairness. In Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, pages 99–106, 2019.

Schmelzer, R.: The achilles' heel of ai, Mar 2019.

Shafieezadeh Abadeh, S., Mohajerin Esfahani, P. M., and Kuhn, D.: Distributionally robust logistic regression. Advances in Neural Information Processing Systems, 28:1576–1584, 2015.

Shaw, M. J. and Gentry, J. A.: Using an expert system with inductive learning to evaluate business loans. Financial Management, 1988.

Shimodaira, H.: Improving predictive inference under covariate shift by weighting the log-likelihood function. Journal of statistical planning and inference, 90(2):227–244, 2000.

Shipp, M. A., Ross, K. N., Tamayo, P., Weng, A. P., Kutok, J. L., Aguiar, R. C., Gaasenbeek, M., Angelo, M., Reich, M., Pinkus, G. S., et al.: Diffuse large b-cell lymphoma outcome prediction by gene-expression profiling and supervised machine learning. Nature medicine, 8(1), 2002.

Singh, A. and Joachims, T.: Fairness of exposure in rankings. In SIGKDD. ACM, 2018.

Singh, A. and Joachims, T.: Policy learning for fairness in ranking. In Advances in Neural Information Processing Systems, pages 5426–5436, 2019.

Singh, A., Kempe, D., and Joachims, T.: Fairness in ranking under uncertainty. Advances in Neural Information Processing Systems, 34, 2021.

Sion, M.: On general minimax theorems. Pacific Journal of Mathematics, 8(1), 1958.

Steinwart, I. and Christmann, A.: Support vector machines. Springer Science & Business Media, 2008.

Stoyanovich, J., Yang, K., and Jagadish, H.: Online set selection with fairness and diversity constraints. In Proceedings of the EDBT Conference, 2018.

Sugiyama, M., Krauledat, M., and Müller, K.-R.: Covariate shift adaptation by importance weighted cross validation. Journal of Machine Learning Research, 8(May):985–1005, 2007.

Sutton, R. S. and Barto, A. G.: Reinforcement learning: An introduction. MIT press, 2018.

Syed, U. and Schapire, R. E.: A game-theoretic approach to apprenticeship learning. Advances in neural information processing systems, 20, 2007.

Taskar, B., Chatalbashev, V., Koller, D., and Guestrin, C.: Learning structured prediction models: a large margin approach. In Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005, eds. L. D. Raedt and S. Wrobel, volume 119 of ACM International Conference Proceeding Series, pages 896–903. ACM, 2005.

Topsøe, F.: Information-theoretical optimization techniques. Kybernetika, 15(1):8–27, 1979.

Tsochantaridis, I., Joachims, T., Hofmann, T., and Altun, Y.: Large margin methods for structured and interdependent output variables. J. Mach. Learn. Res., 6:1453–1484, 2005.

Vapnik, V.: Principles of risk minimization for learning theory. In Advances in neural information processing systems, pages 831–838, 1992.

Vapnik, V. and Chapelle, O.: Bounds on error expectation for support vector machines. Neural computation, 12(9):2013–2036, 2000.

Verma, S. and Rubin, J.: Fairness definitions explained. In 2018 IEEE/ACM International Workshop on Software Fairness (FairWare), pages 1–7. IEEE, 2018.

Von Neumann, J.: A certain zero-sum two-person game equivalent to the optimal assignment problem. Contributions to the Theory of Games, 2(0):5–12, 1953.

Von Neumann, J. and Morgenstern, O.: Theory of games and economic behavior. Bull. Amer. Math. Soc, 51(7), 1945.

Wang, H., Xing, W., Asif, K., and Ziebart, B.: Adversarial prediction games for multivariate losses. In Advances in Neural Information Processing Systems, pages 2710–2718, 2015.

Wightman, L. F.: LSAC national longitudinal bar passage study, 1998.

Woodworth, B., Gunasekar, S., Ohannessian, M. I., and Srebro, N.: Learning non-discriminatory predictors. In COLT, 2017.

Xu, D., Yuan, S., Zhang, L., and Wu, X.: FairGAN: Fairness-aware generative adversarial networks. In IEEE Big Data, 2018.

Yadav, H., Du, Z., and Joachims, T.: Policy-gradient training of fair and unbiased ranking functions. In Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, pages 1044–1053, 2021.

Yang, K. and Stoyanovich, J.: Measuring fairness in ranked outputs. In Proceedings of the 29th International Conference on Scientific and Statistical Database Management, pages 1–6, 2017.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P.: Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In WWW, 2017.

Zafar, M. B., Valera, I., Rodriguez, M., Gummadi, K., and Weller, A.: From parity to preference-based notions of fairness in classification. In NeurIPS, 2017.

Zafar, M. B., Valera, I., Rodriguez, M. G., and Gummadi, K. P.: Fairness constraints: Mechanisms for fair classification. arXiv preprint arXiv:1507.05259, 2015.

Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P.: Fairness constraints: Mechanisms for fair classification. In AISTATS, 2017.

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

Zehlike, M. and Castillo, C.: Reducing disparate exposure in ranking: A learning to rank approach. In Proceedings of The Web Conference 2020, pages 2849–2855, 2020.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C.: Learning fair representations. In ICML, 2013.

Zhang, B. H., Lemoine, B., and Mitchell, M.: Mitigating unwanted biases with adversarial learning. In AIES, 2018.

Ziebart, B., Choudhury, S., Yan, X., and Vernaza, P.: Towards uniformly superhuman autonomy via subdominance minimization. In International Conference on Machine Learning, pages 27654–27670. PMLR, 2022.

Ziebart, B. D., Maas, A. L., Bagnell, J. A., Dey, A. K., et al.: Maximum entropy inverse reinforcement learning. In AAAI, volume 8, pages 1433–1438, 2008.

Zymler, S., Kuhn, D., and Rustem, B.: Distributionally robust joint chance constraints with second-order moment information. Mathematical Programming, 137:167–198, 2013.

VITA

NAME	Omid Memarrast
EDUCATION	**Ph.D., Computer Science**, University of Illinois Chicago, 2023
	M.Sc., Computer Science, University of Illinois Chicago, 2021
	B.Sc., Software Engineering, University of Tehran, Iran, 2012
EXPERIENCE	**Research Assistant**, University of Illinois Chicago, IL, 2019 - 2023
	Teaching Assistant, University of Illinois Chicago, IL, 2017 - 2021
	Machine Learning Research Intern, LinkedIn, Sunnyvale, CA, 2020
	Data Science Intern, Morningstar, Inc., Chicago, IL, 2018
PUBLICATIONS	**Omid Memarrast**, Linh Vu, Brian Ziebart. "Superhuman Fairness" In Proceedings of the International Conference on Machine Learning, ICML 2023.
	Omid Memarrast, Ashkan Rezaei, Rizal Fathony, Brian Ziebart. "Fairness for Robust Learning to Rank" In Proceedings of Advances in Knowledge Discovery and Data Mining: Pacific-Asia Conference, PAKDD 2023.

Ashkan Rezaei, Anqi Liu, **Omid Memarrast**, Brian Ziebart."Robust Fairness Under Covariate Shift" In Proceedings of the AAAI Conference on Artificial Intelligence 2021.

Ashkan Rezaei, Rizal Fathony, **Omid Memarrast**, Brian Ziebart. "Fairness for Robust Log Loss Classification" In Proceedings of the AAAI Conference on Artificial Intelligence 2020.

Daniel Khashabi, Arman Cohan, Siamak Shakeri, Pedram Hosseini et al. "Parsinlu: a suite of language understanding challenges for persian." Transactions of the Association for Computational Linguistics 2021.

WORKSHOPS

Omid Memarrast, Linh Vu, and Brian Ziebart. "Superhuman Fairness" In ICLR Workshop on Pitfalls of limited data and computation for Trustworthy ML, 2023.

Omid Memarrast, Ashkan Rezaei, Rizal Fathony, Brian Ziebart, "Fairness for Robust Learning to Rank" in NeurIPS Workshop: Algorithmic Fairness through the Lens of Causality and Robustness, 2021.

SERVICE

Reviewer, **NeurIPS 2021, NeurIPS 2022, NeurIPS 2023**

Reviewer, **ICML 2022**

Program Committee, **IJCAI 2021, IJCAI 2022**