5-2020

Achieving Causal Fairness in Machine Learning

Yongkai Wu
University of Arkansas, Fayetteville

Achieving Causal Fairness in Machine Learning

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy in Computer Science

by

Yongkai Wu
Tsinghua University
Bachelor of Engineering in Electronic Information Science and Technology, 2014
University of Arkansas
Master of Science in Computer Science, 2018

May 2020
University of Arkansas

This dissertation is approved for recommendation to the Graduate Council.

Xintao Wu, Ph.D.
Dissertation Director

Qinghua Li, Ph.D.
Committee member

Brajendra Panda, Ph.D.
Committee member

Song Yang, Ph.D.
Committee member

Lu Zhang, Ph.D.
Committee member

ABSTRACT

Fairness is a social norm and a legal requirement in today's society. Many laws and regulations (e.g., the Equal Credit Opportunity Act of 1974) have been established to prohibit discrimination and enforce fairness on several grounds, such as gender, age, sexual orientation, race, and religion, referred to as sensitive attributes. Nowadays machine learning algorithms are extensively applied to make important decisions in many real-world applications, e.g., employment, admission, and loans. Traditional machine learning algorithms aim to maximize predictive performance, e.g., accuracy. Consequently, certain groups may get unfairly treated when those algorithms are applied for decision-making. Therefore, it is an imperative task to develop fairness-aware machine learning algorithms such that the decisions made by them are not only accurate but also subject to fairness requirements. In the literature, machine learning researchers have proposed association-based fairness notions, e.g., statistical parity, disparate impact, equality of opportunity, etc., and developed respective discrimination mitigation approaches. However, these works did not consider that fairness should be treated as a causal relationship. Although it is well known that association does not imply causation, the gap between association and causation is not paid sufficient attention by the fairness researchers and stakeholders.

The goal of this dissertation is to study fairness in machine learning, define appropriate fairness notions, and develop novel discrimination mitigation approaches from a causal perspective. Based on Pearl's structural causal model, we propose to formulate discrimination as causal effects of the sensitive attribute on the decision. We consider different types of causal effects to cope with different situations, including the path-specific effect for direct/indirect discrimination, the counterfactual effect for group/individual discrimination, and the path-specific counterfactual effect for general cases. In the attempt to measure discrimination, the unidentifiable situations pose an inevitable barrier to the accurate causal inference. To address this challenge, we propose novel bounding methods to accurately estimate the strength of unidentifiable fairness notions, including path-specific fairness, counterfactual fairness, and

path-specific counterfactual fairness. Based on the estimation of fairness, we develop novel and efficient algorithms for learning fair classification models. Besides classification, we also investigate the discrimination issues in other machine learning scenarios, such as ranked data analysis.

ACKNOWLEDGEMENTS

During my Ph.D. study, I have been fortunate to receive plenty of help and support from many ends. I hereby express my sincere gratitude to all of them.

First and foremost, I am greatly indebted to my advisor, Dr. Xintao Wu. I appreciate all his time, guidance, support, and patience throughout the period of my Ph.D. study. He is my role model as an excellent researcher and successful professor who has a strong passion for research and teaching. His enthusiasm has motivated me in the past five years and will continue motivating me in my future career.

My gratitude also goes to Dr. Lu Zhang with whom I have been truly fortunate to collaborate throughout my Ph.D. journey. His advice, experience, and ideas are vital in our research projects. I have benefited tremendously from the collaboration and discussion with Dr. Lu Zhang.

I would like to thank my dissertation committee members, Dr. Qinghua Li, Dr. Brajendra Panda, and Dr. Song Yang. Their constructive comments significantly improve the quality of this dissertation.

I also thank my friends and colleagues in the Social Awareness and Intelligent Learning Lab at the University of Arkansas. I appreciate the collaboration with Srinidhi Katla, Depeng Xu, Qiuping Pan, Dr. Shuhan Yuan, Wen Huang. Thanks also goes to the other friends and colleagues who give me consistent encouragement: Dr. Panpan Zheng, Wei Du, and Dr. Kevin C. Labille. I have been truly honored to work with these excellent researchers.

Finally, I would express my deepest gratitude to my family for their unconditional love and support. This dissertation is dedicated to them.

TABLE OF CONTENTS

1 Introduction . . . 1
1.1 Motivation . . . 1
1.2 Background . . . 2
1.3 Research Challenges . . . 3
1.4 Contributions . . . 5
1.5 Organization . . . 9

2 Related Work . . . 11
2.1 Discrimination Discovery . . . 11
2.2 Discrimination Removal . . . 14
2.3 Summary . . . 17

3 Preliminaries . . . 18
3.1 Data and Attributes . . . 18
3.2 Association-based Fairness Notions . . . 18
3.2.1 Demographic Parity . . . 19
3.2.2 Equality of Opportunity . . . 19
3.3 Structural Causal Model . . . 20
3.3.1 Intervention and Causal Inference . . . 22
3.3.2 Identification of Causal Quantities . . . 23
3.3.3 Total Causal Effect . . . 24

4 Discrimination Discovery and Removal from Classification Data . . . 25
4.1 Introduction . . . 25
4.2 Related Work . . . 27
4.3 Preliminaries . . . 29
4.4 Discrimination Discovery and Removal . . . 31
4.4.1 Modeling Direct/Indirect Discrimination as Path-Specific Effects . . . 31
4.4.2 Discovery Algorithm . . . 36
4.4.3 Removal Algorithm . . . 38
4.5 Dealing with Unidentifiable Situation . . . 40
4.5.1 Preliminaries . . . 41
4.5.2 Bounding Indirect Discrimination . . . 42
4.5.3 Algorithms for Unidentifiable Situation . . . 49
4.6 Experiments . . . 52
4.6.1 Discrimination Discovery . . . 53
4.6.2 Discrimination Removal . . . 55
4.6.3 Unidentifiable Situation . . . 56
4.7 Summary . . . 59

5 Discrimination Discovery and Removal from Ranked Data . . . 60
5.1 Introduction . . . 60
5.2 Related Work . . . 62
5.3 Preliminaries . . . 63
5.4 Modeling Direct and Indirect Discrimination in Ranked Data . . . 64
5.4.1 Building Causal Graph for Ranked Data . . . 65
5.4.2 Quantitative Measurement . . . 66
5.4.3 Relationship between Ranking and Binary Decision . . . 70
5.5 Discovery and Removal Algorithms . . . 74
5.6 Experiments . . . 77
5.6.1 Experimental Setup . . . 77
5.6.2 Discrimination Discovery . . . 79
5.6.3 Discrimination Removal . . . 80
5.7 Summary . . . 82

6 Counterfactual Fairness . . . 83
6.1 Introduction . . . 83
6.2 Preliminaries . . . 84
6.2.1 Counterfactual Inference and Unidentification . . . 84
6.3 Quantifying and Bounding Counterfactual Fairness . . . 85
6.3.1 Identification of Counterfactual Quantity . . . 87
6.3.2 Bounding Unidentifiable Counterfactual Quantity . . . 90
6.3.3 Extending to General Case . . . 91
6.4 Achieving Counterfactual Fairness in Classification . . . 91
6.5 Experiments . . . 94
6.5.1 Datasets . . . 94
6.5.2 Experiment on the Synthetic Dataset . . . 95
6.5.3 Experiment on the *Adult* Dataset . . . 97
6.6 Summary . . . 97

7 Path-specific Counterfactual Fairness . . . 99
7.1 Introduction . . . 99
7.2 Preliminaries . . . 101
7.3 Path-specific Counterfactual Fairness . . . 101
7.4 Measuring Path-specific Counterfactual Fairness . . . 103
7.4.1 Response-function Variable . . . 104
7.4.2 Expressing Path-specific Counterfactual Fairness . . . 106
7.4.3 Bounding Path-specific Counterfactual Fairness . . . 109
7.5 Experiments . . . 110
7.6 Summary . . . 113

8 Convexity and Bounds of Fairness-aware Classification . . . 114
8.1 Introduction . . . 114
8.2 Fairness-aware Classification . . . 115
8.2.1 Classification Problem . . . 116

8.2.2 Fairness-aware Classification Problem . . . 117
8.3 Convex Fairness Classification Framework . . . 119
8.3.1 Constraint-free Criterion . . . 120
8.3.2 Convex Fairness-aware Classification . . . 122
8.3.3 Refined Fairness-aware Classification . . . 124
8.4 Extension to Other Fairness Notions . . . 130
8.5 Experiments . . . 131
8.5.1 Experimental Setup . . . 131
8.5.2 Constraint-free Criterion of Ensuring Fairness . . . 132
8.5.3 Learning Fair Classifiers . . . 133
8.6 Summary . . . 134

9 Conclusions and Future Work . . . 135
9.1 Conclusions . . . 135
9.2 Future Work . . . 138

Bibliography . . . 141

A Appendix . . . 151
A.1 Datasets . . . 151
A.1.1 The *Adult* Dataset . . . 151
A.1.2 The *Dutch Census of 2001* Dataset . . . 151
A.1.3 The *German Credit* Dataset . . . 152
A.2 Software . . . 152

LIST OF FIGURES

Figure 3.1: Causal graphs of a Markovian model. 21
Figure 3.2: Causal graphs of a semi-Markovian model. 21

Figure 4.1: The toy model. 26
Figure 4.2: The "kite" pattern. 30
Figure 4.3: The recanting witness criterion satisfied. 30
Figure 4.4: π_i-specific effect satisfying recanting witness criterion. 46
Figure 4.5: The constructed causal graphs. 54
Figure 4.6: The "kite" pattern when treating `edu_leve` as redlining. 58

Figure 5.1: A toy example produced by two rankers. 60
Figure 5.2: The causal graph of the toy example involving. 66
Figure 5.3: The causal graph of $\mathbf{D}$. 79

Figure 6.1: (a) Causal Graph $\mathcal{G}$. (b) Counterfactual Graph $\mathcal{G}'$ for $P(\hat{y}_s|s', \mathbf{z})$. 87
Figure 6.2: Causal graphs for the synthetic dataset and the *Adult* dataset. 96

Figure 7.1: The "bow" graph. 103
Figure 7.2: The "kite" graph. 103
Figure 7.3: The "w" graph. 103
Figure 7.4: The causal graph for a semi-Markovian model. 103
Figure 7.5: A causal graph with unidentifiable path-specific counterfactual fairness. . 109
Figure 7.6: The causal graph for the synthetic dataset $\mathcal{D}_1$. 111
Figure 7.7: The causal graph for the synthetic dataset $\mathcal{D}_2$. 111

Figure 8.1: Two classifiers and their predictions. 119
Figure 8.2: Curves of examples for $\kappa(\cdot)$ and $\delta(\cdot)$. 123
Figure 8.3: Comparison of fair classifiers on two datasets. 134

LIST OF TABLES

Table 4.1: Comparison of removal algorithms the *Adult* dataset. 55
Table 4.2: Comparison of removal algorithms the *Dutch Census of 2001* dataset. . . 56
Table 4.3: Discrimination in prediction for the *Adult* dataset. 57
Table 4.4: Discrimination in prediction for the *Dutch Census of 2001* dataset. 57
Table 4.5: Discrimination of the *Adult* dataset with unidentification. 57
Table 4.6: Discrimination removal for the *Adult* dataset with unidentification. . . . 59

Table 5.1: Comparison of methods for the discrimination discovery. 80
Table 5.2: Comparison of methods for the discrimination removal. 81
Table 5.3: Comparison of *FRank* with varied τ. 82

Table 6.1: Bounds and ground truth of counterfactual fairness. 96
Table 6.2: Counterfactual fairness for prediction of the synthetic dataset. 97
Table 6.3: Prediction accuracy for the synthetic dataset. 97
Table 6.4: Counterfactual fairness for prediction of the *Adult* dataset. 98
Table 6.5: Prediction accuracy for the *Adult* dataset. 98

Table 7.1: Connection between previous fairness notions and PC fairness 103
Table 7.2: Bounds and ground truth of PC fairness on $\mathcal{D}_1$. 111
Table 7.3: Comparison with existing methods on $\mathcal{D}_2$. 112
Table 7.4: Comparison with the existing method on the *Adult* dataset. 112

Table 8.1: Some common surrogate functions. 130
Table 8.2: $\mathbb{RD}^+, \mathbb{RD}^-$ and risk differences of classic classifiers. 133

Table A.1: Download links of datasets. 151
Table A.2: Links to the implementations of proposed methods. 152

1 Introduction

This chapter introduces the motivation as well as necessary background and summarizes the contributions of this research. The structure of this dissertation is provided at the end of this chapter.

1.1 Motivation

Discrimination refers to an act of making unjustified distinctions among individuals based on their membership or perceived membership, in a certain group, and often occurs when the group is treated less favorably than others. A large number of laws and regulations have been established to prohibit discrimination in many countries and regions. For example, in the USA, the Civil Rights Act of 1964 prohibits employment discrimination based on race, color, religion, sex, or national origin. In the European Union, Council Directive 76/207/EEC implements the principle of equal treatment for men and women as regards access to employment, vocational training and promotion, and working conditions. Although anti-discrimination laws and regulations have been established, anti-discrimination is still an active research topic across multiple disciplines, e.g., social sciences, psychology, and economics, due to the challenges in detecting and eliminating discrimination.

Machine learning has drawn tremendous attention in recent years due to its powerful and effective abilities to tackle some of the world's most challenging problems. Various machine learning algorithms have been designed and deployed to make important decisions in many real-world applications, e.g., employment, admission to universities, and loans from banks. However, traditional machine learning algorithms aim to maximize predictive performance, e.g., accuracy, with regard to the historical training data. Consequently, the resultant models may make unfair and undesired predictions, e.g., some individuals are unfavorably treated due to some sensitive information. An established example of machine learning discrimination is Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) [1], a

commercial software that assesses the risk of recidivism. The risk scores produced by this software have been used in many U.S. courts to make important decisions, e.g., whether an offender is set free or not. However, discrimination against the African-Americans has been identified in COMPAS as the blacks are more likely to be assessed as higher risk than the whites. Discrimination and bias caused by machine learning algorithms may cause severe damage to the unfavorable groups, even perpetuate and aggravate existing prejudices and social inequalities. Therefore, it is imperative to develop fairness-aware machine learning algorithms so that the decisions made by those algorithms achieve fairness and high predictive performance simultaneously.

1.2 Background

In the following, we give some necessary background on fairness-aware machine learning, including several key definitions and categorizations of discrimination.

A *protected* or *sensitive* attribute is the one that is potentially used to unfairly treat individuals. The protected attribute is defined in laws, e.g., race, sex, sexual orientation, age, physical or mental disability, etc. Among groups specified by the values of the protected attribute, a group of people is considered as a *protected* group if they are qualified for special protection by laws or policies. For instance, females are considered as a protected group in employment. Usually, a protected group is also known as an *unfavorable* group. Apart from the protected group, the rest is considered as an *unprotected* group or a *favorable* group. A *decision* attribute is the one that machine learning models attempt to predict. In the field of fairness-aware machine learning, researchers focus on the classification task where the decision attribute is usually assumed to be binary: the positive value corresponding to a beneficial decision and a negative value corresponding to a detrimental decision. For example, being admitted is considered positive and being rejected is negative in the example of admission to universities. With the exception of the protected attribute and the decision attribute, the rest attributes are known as *non-protected*, *non-sensitive*, or *other* attributes.

Discrimination has been studied from different perspectives in many fields. Several categorizations of fairness notions have been proposed. Discrimination is legally divided into direct and indirect discrimination from the perspective of the way discrimination occurs. Direct discrimination occurs when individuals receive less favorable treatment explicitly based on the sensitive attribute. An example of direct discrimination is that a qualified female applicant is rejected solely due to her gender. Indirect discrimination refers to the situation where the treatment is based on apparently neutral non-sensitive attributes but still results in unjustified distinctions against individuals from the protected group. A well-known example of indirect discrimination is redlining, where the residential ZIP code of an individual is used for making decisions such as granting a loan. Although the ZIP code is apparently a neutral attribute, it correlates with race due to the racial composition of residential areas. Thus, the use of the ZIP code may indirectly lead to racial discrimination. From the perspective of levels of granularity in studying, discrimination can be divided into system-level, group-level, and individual-level. System-level discrimination refers to the average unjustified distinction across the whole system. An example of system-level discrimination is that all female applications to a university are unequally treated. Group-level discrimination deals with discrimination that occurs within one particular group specified by the non-protected attributes, e.g., the applicants to a particular major. Individual-level discrimination deals with discrimination against one particular individual. Of course, discrimination can be described as a combination from two perspectives, e.g., direct discrimination at the system level.

1.3 Research Challenges

Fairness-aware machine learning has been actively investigated in the past few years. The research community has developed various approaches for unveiling discrimination by analyzing historical data as well as predictions, and ensuring fairness by modifying the biased data, vulnerable models, or the unfair predictions. Many fairness notions have been proposed to quantify the strength of unfairness by inspecting the inequality of the decisions

among the different demographic groups. Most of these notions are based on association or correlation. However, discrimination is causal, which means that to prove discrimination one needs to derive a causal relationship rather than an association relationship. Although it is well-known that association does not mean causation, the gap between association and causation is not paid enough attention by the fairness-aware machine learning researchers. Due to this gap, most existing fairness notions fail to accurately detect the existence of discrimination. Further, the developed mitigation algorithms based on inappropriate notions cannot successfully eliminate discrimination from the machine learning pipelines. What is worse, these mitigation algorithms result in plenty of data utility loss and predictive errors.

We present several crucial challenges which motive us to conduct research in the fairness-aware machine learning field.

First, most existing works in the field of fairness-aware machine learning rely on correlation/association-based definitions of discrimination. Fairness cannot be well assessed based on the simple concepts of correlation or association. An empirical counterexample is Simpson's paradox, where the statistical conclusions drawn from the sub-groups disagree with that from the whole population. Usually, discrimination claims require plaintiffs to demonstrate a causal connection between the challenged decision and a sensitive characteristic. In order to prove the existence of discrimination or not, it is necessary to examine the causal relationship between the sensitive attribute and the decision rather than the associated relationship. Second, discrimination can be classified into many types, such as system-/group-/individual-level discrimination and direct/indirect discrimination. Existing works usually tackle one or two types of discrimination. In many practical situations, several types of discrimination may simultaneously exist. Thus, it is necessary to develop a unified framework that is able to deal with all types of discrimination. Third, unidentification is a key challenge in causal inference, which means a causal quantify is impossible to be uniquely quantified in theory. It also poses a crucial barrier when researchers apply causal fairness notions into practice. Researchers make do with some intuitive and simplified

assumptions to avoid unidentifiable situations. Consequently, these assumptions significantly reduce the performance of machine learning algorithms. It deserves further investigation on the unidentification of causal fairness notions. Fourth, most existing works focus on the classification task and utilize statistical parity as the fairness metric. Ranked data analysis is another important task where the decisions are a series of unique, concatenating integers that cannot be treated as normal categorical random variables. The causality-based fairness notions and approaches developed for classification cannot be readily adapted to ranked data. Measuring causality-based fairness in the ranked data is an open problem. Last but not least, it has been studied that the fairness constraints are incorporated with machine learning procedures, e.g., training of classification models. Usually, the fairness constraints are non-convex, making the training computationally difficult. More important, there is a lack of theoretical fairness guarantee in the obtained models.

1.4 Contributions

Inspired by the above challenges, the goal of this dissertation is to formulate appropriate fairness notions and develop novel discrimination mitigation approaches from a causality perspective. We leverage the structural causal model, a graphical equation-based mathematical language that describes the causal mechanisms of a system. Within the structural causal model, we formulate various kinds of discrimination as causal effects of the sensitive attribute on the decision, e.g., the path-specific effect and the path-specific counterfactual effect. Then, we target the unidentification problem and develop bounding methods to make accurate fairness judgment. We further develop novel and efficient mitigation algorithms based on the causal fairness notions and bounding methods. Besides that, we investigate the discrimination issues in many other machine learning scenarios, e.g., ranked data analysis and empirical risk minimization classification.

The overall contributions of this dissertation are summarized as follows.

In the fields of law and social science, discrimination is divided into direct and indirect

discrimination. Direct discrimination occurs when individuals receive less favorable treatment explicitly based on the sensitive attribute. Indirect discrimination refers to the situation where the treatment is based on apparently neutral non-sensitive attributes but still results in unjustified distinctions against individuals from the sensitive group. Besides two types of discrimination, some effects between two variables can be reasonably justified and should not be considered as discrimination (e.g., the difference in the average income of females and males caused by their different working hours per week). Previous methods cannot explicitly and correctly identify the three different effects when measuring and removing discrimination. They either incorrectly measure them or crudely consider all connections between the sensitive attribute and the decision as discrimination. Facilitated by the causal modeling, we leverage the path-specific effect [2, 3] to capture the three effects as the causal effects of the sensitive attribute on the decision that are transmitted along different causal paths in the causal graph. For certain situations where indirect discrimination cannot be exactly measured due to the unidentifiability of some path-specific effects, we develop an upper bound and a lower bound of indirect discrimination. Based on that, we develop effective algorithms for discovering direct and indirect discrimination as well as algorithms for precisely removing both types of discrimination while retaining good data utility. The early versions of this work have been published in IJCAI 2017 [4] and TKDE 2019 [5].

Existing causal-based fair machine learning methods focus on the classification problems. Ranked data analysis is another important machine learning task where the decisions are a series of unique, concatenating integers that cannot be treated as normal categorical random variables. Thus, existing methods developed for classification cannot be directly applied to handle the ranked data. Besides, existing methods in fairness-aware ranking are mainly based on statistical parity that cannot accurately measure the discriminatory effects since discrimination is causal. To address these limitations, we propose to map the rank decision to a continuous score variable that represents the qualification of the candidates. Then, we build a causal graph that consists of both the discrete profile attributes and the

continuous score. The path-specific effect technique [4] is extended to the mixed-variable causal graph to identify both direct and indirect discrimination. The relationship between the path-specific effects for the ranked data and those for the binary classification is theoretically analyzed. Finally, algorithms for discovering and removing discrimination from ranked data are developed. This work has been published in KDD 2018 [6].

As discussed in Chapter 1.2, discrimination can be divided into system-level, group-level, and individual-level. In the study of path-specific fairness, we investigate the causal fairness at the system-level where the discrimination is formulated as the aggregated causal effect across all individuals. At the group/individual level, fairness means an individual is equally treated even if it had been from another sensitive group, e.g., a specific male is equally treated if he had been a female. Thus, counterfactual fairness [7] has been defined to capture the group/individual-level discrimination, through evaluating counterfactual effect within a particular group or individual specified by the observation of a set of profile attributes. However, an inherent challenge in the counterfactual fairness is unidentifiability, i.e., the counterfactual quantity cannot be uniquely computed from observed distributions. Existing methods in [7] simply omit some attributes when building the predictive model or postulate the distributions of variables or the causal mechanisms, which may violate the underlying causal model and degrade the prediction performance. To alleviate the challenge of unidentification in counterfactual fairness, we first develop a graphical criterion for determining whether the counterfactual effect is identifiable. For the unidentifiable situations, we derive the lower and upper bounds for the counterfactual quantities and design a fairness criterion based on the bounds. Finally, we develop a post-processing method to reconstruct arbitrary classifiers in order to achieve counterfactual fairness. We formulate the reconstruction problem as a linear constrained optimization problem with the bounded counterfactual fairness criterion as the constraints. This work has appeared in IJCAI 2019 [8].

Various fairness notions have been proposed recently, e.g., path-specific fairness [4, 5], counterfactual fairness [7], on the basis of causality. However, one common challenge of all

causality-based fairness notions is unidentifiability, which is a critical barrier to applying these notions to real-world situations. In the previous study, we develop methods to bound the path-specific and counterfactual fairness. Nevertheless, the tightness of these methods is not analyzed. In addition, it is not clear whether these methods can be applied to other unidentifiable situations, and more importantly, a combination of multiple unidentifiable situations. Motivated by these challenges, we develop a unified framework for handling various causality-based fairness notions. We first propose a general representation of all types of causal effects, i.e., the path-specific counterfactual effect, based on which we define a unified fairness notion that covers most previous causality-based fairness notions, namely the path-specific counterfactual fairness (PC fairness). Then, we develop a constrained optimization problem for bounding the PC fairness, which is motivated by the method in [9] for bounding confounded causal effects. The key idea is to parameterize the causal model using so-called response-function variables, whose distribution captures all randomness encoded in the causal model, so that we can explicitly traverse all possible causal models to find the tightest possible bounds. In the experiments, we evaluate the proposed method and compare it with previous bounding methods using both synthetic and real-world datasets. The results show that our method is capable of bounding causal effects under any unidentifiable situation or combinations. This work has been published in NeurIPS 2019 [10].

Fair classification is receiving increasing attention in the machine learning field. Several recent works have proposed to formulate the fairness-aware classification as constrained optimization problems. Generally, they aim to minimize the empirical risk function subject to certain fairness constraints, e.g., demographic parity (i.e., the proportion difference of the positive predictions between the favorable group and non-favorable group is less than a threshold). However, most quantitative fairness metrics, such as demographic parity and mistreatment parity, are non-convex due to the use of the indicator function, thus making the optimization problem intractable. A widely-used strategy to achieve convexity in optimization is to adopt surrogate functions for both loss function and constraints. One

challenge is that, when surrogate functions are used to convert non-convex functions to convex functions, estimation errors must exist due to the difference between the surrogate function and the original non-convex function. Thus, satisfying the fairness constraints represented by surrogate functions does not sufficiently guarantee achieving the real fairness. Hence, how to perform fairness-aware classification via constrained optimization remains an open problem. We design a general framework for fairness-aware classification which addresses the gap incurred by the estimation errors due to the surrogate functions. Within the framework, we first present a constraint-free criterion (derived from the training data) which ensures that any classifier learned from the data will guarantee to be fair in terms of the specified fairness metric. Thus, when the criterion is satisfied, there is no need to add any fairness constraint into the empirical risk minimization for learning fair classifiers. When the criterion is not satisfied, we formulate various commonly-used fairness metrics (risk difference, risk ratio, and equal odds) as convex constraints that are then directly incorporated into classic classification models. Thanks to the convexity of constraints and the objective function, the constrained optimization problem can be efficiently solved. To connect the surrogated fairness constraints to the original non-convex fairness metric, we further derive the lower and upper bounds of the real fairness measure based on the surrogate function, and develop the refined fairness constraints. This means that, if the refined constraints are satisfied, then it is guaranteed that the real fairness measure is also bounded within the given interval. The bounds work for any surrogate function that is convex and differentiable at zero with the derivative larger than zero. This work has been published in The Web Conference 2019 (formerly known as WWW) [11].

1.5 Organization

The remainder of this dissertation is organized as follows. In Chapter 2, we discuss related work in a wide scope of fairness-aware machine learning to provide a general review of research achievements in this research field. We introduce the preliminary background

for causal inference in Chapter 3 which is fundamental and necessary for all the proposed research. The extra highly related work and preliminaries are given at the beginning of each research chapter, as necessary.

We present the main body of this dissertation in Chapter 4 - Chapter 8. In Chapter 4, we introduce methods for identifying the direct/indirect discrimination by leveraging the path-specific effect and present our efficient mitigation algorithm to achieve both direct and indirect fairness. Chapter 5 elaborates on the extension of the path-specific effect to ranked data and the connection between fairness-aware classification and fairness-aware ranking. Chapter 6 introduces the bounding method to unidentifiable counterfactual fairness and a theoretical sound algorithm to build a counterfactually fair classifier. Chapter 7 shows a unified framework for the path-specific counterfactual fairness and develops a general bounding method for unidentifiable situations. In Chapter 8, we study the convexity and bounding problem of training a classifier with the fairness constraints.

We conclude this dissertation with a discussion of future work in Chapter 9.

2 Related Work

There are two tasks in fairness-aware machine learning: (1) discrimination discovery is the task of unveiling discriminatory practices by analyzing historical datasets or predictions made by predictive models, (2) discrimination prevention aims to remove discrimination by modifying biased data, tweaking predictive model, or manipulating predictions. This chapter reviews the literature, including research works presented in this dissertation, from these two aspects. Some literature particularly related to a specific chapter is discussed in the corresponding chapter. A summary of this review is given at the end of this chapter.

2.1 Discrimination Discovery

How to discover discrimination from data has been studied and many techniques have been proposed in the literature. Among them a widely adopted concept is called *statistical parity*, which means that the demographics of a set of individuals receiving positive (or negative) decisions are identical to the demographics of the population as a whole. Based on statistical parity, the classic statistical metrics of discrimination consider the difference among the proportion of having positive decision for the non-protected group (p_1), that for the protected group p_2, and that for the whole population (p). According to how the difference is measured, these metrics can be distinguished into $p_1 - p_2$ (a.k.a. risk difference), $\frac{p_1}{p_2}$ (a.k.a. risk ratio), $\frac{1-p_1}{1-p_2}$ (a.k.a. relative chance), $\frac{p_1(1-p_2)}{p_2(1-p_1)}$ (a.k.a. odds ratio), $p_1 - p$ (a.k.a. extended risk difference), $\frac{p_1}{p}$ (a.k.a. extended risk ratio), $\frac{1-p_1}{1-p}$ (a.k.a. extended change), etc. Fairness notions based on statistical parity are commonly used and compatible with many laws and regulations for anti-discrimination. For instance, the U.S. legislation for employment discrimination sets the risk ratio threshold as 1.25 (known as the four-fifths rule). The idea of statistical parity can be naturally extended to group levels, e.g., the risk difference of admission between male and female for individuals applying for the computer science major.

Inspired by the statistical parity, Hardt et al. [12], Zafar et al. [13], and Corbett-Davis

et al. [14] proposed *equality of opportunity*, *mistreatment parity*, and *predictive equality*. These notions are designed for predictive models, e.g., a classifier. The general idea is the accuracy of predictions among different demographic groups is equal or similar. Technically, equal opportunity is satisfied if the true positive rates are the same for favorable and unfavorable groups. More strictly, equal odds require that both the true positive rates and false positive rates are respectively the same for favorable and unfavorable groups. Further, Kleinberg et al. [15] showed that *equalized odds* and *calibration*, where a algorithm makes the positive prediction associated with a probability p for a group, then the fraction of positive decisions among this group should be p, cannot be satisfied at the same time except in special cases where the classifier is perfect or the sensitive attribute is independent of the decision attribute.

Individual fairness means the similar individual should be treated similarly [16]. Luong et al. [17] exploited the idea of situation testing to discover individual-level discrimination. For each member of the protected group with a negative decision outcome, testers with similar characteristics are found from a historical dataset. When there are significantly different decision outcomes between the testers of the protected group and the testers of the non-protected group, the negative decision can be considered as discrimination. Conditional discrimination, i.e., part of discrimination may be explained by other legally grounded attributes, was studied by Zliobaite et al. [18]. The task was to evaluate to which extent the discrimination apparent for a group is explainable on a legal ground. The metric is still based on the difference of the positive decision proportions for the protected and non-protected groups.

Data mining techniques have been also studied for measuring discrimination. Pedreschi et al. and Ruggieri et al. extracted from the dataset classification rules which represent certain discrimination patterns [19–21]. If the presence of the protected attribute increases the confidence of a classification rule, it indicates possible discrimination in the dataset. Based on that, Mancuhan and Clifton [22] further proposed to use Bayesian networks to compute the confidence of the classification rules for detecting discrimination. Hajian and Domingo [23]

quantified the direct and indirect discrimination using *extend lift* (*elift*) over association rules. Direct discrimination is identified if the *elift* of the sensitive attribute and the context attribute to the decision attribute is larger than a threshold. Indirect discrimination exists if the *elift* of two context attributes that are strongly correlated with the sensitive attribute to the decision attribute is significant.

Most existing research is often limited to examining the single relationship between one decision attribute and one protected attribute and does not sufficiently incorporate the effects caused by other non-protected attributes. Wu and Wu [24] developed a unified framework that aims to capture and measure discrimination between multiple decision attributes and protected attributes in addition to a set of non-protected attributes. The proposed approach is based on loglinear modeling. The coefficient values of the fitted loglinear model provide quantitative evidence of discrimination in decision making. The conditional independence graph derived from the fitted graphical loglinear model can be used to effectively capture the existence of discrimination patterns based on Markov properties.

All of the above works are mainly based on correlation or association. Recently, several studies have been devoted to analyzing discrimination from the causal perspective. Bonchi et al. [25] developed a framework based on the Suppes-Bayes causal network and several random-walk-based methods to detect different types of discrimination. However, it is unclear how the number of random walks is related to practical discrimination metrics. In addition, the construction of the Suppes-Bayes causal network is impractical with the large number of attribute-value pairs. Studies in [26–28] are built on causal modeling and the associated causal graph, but cannot deal with indirect discrimination. Leveraging the path-specific effect [2], Zhang et al. [4, 5], Nabi et al. [29], and Zhang and Bareinboim [30] developed causal fairness notions to quantifying direct and indirect discrimination. Kilbertus et al. [31] proposed similar discrimination criteria that also consider indirect discrimination. However, it is simplified in order to avoid the complexity in measuring path-specific effects and the proposed discrimination criteria can only qualitatively determine the existence of

the discrimination, but cannot quantitatively measure the amount of discriminatory effects. Huang et al. [32] utilized causal modeling and developed *equality of effort* to capture the difference of effort to achieve the same outcome. Huang et al. [33] studied the multi-cause discrimination where several protected attributes and redlining attributes are presented in a causal model. Khademi et al. [34] introduced two fairness definitions, *fair on average effect (FACE)* and *fair on average causal effect on the treated (FACT)*, based on the potential outcome framework. Kusner et al. [7] initiated the idea of counterfactual fairness which is designed to evaluate the fairness at the group level and the individual level. Counterfactual fairness means the decision toward a individual in the actual world is identical to that in a counterfactual world where the individual had belonged to a different demographic group. Nevertheless, there is a crucial challenge in the quantification of counterfactual fairness posed by unidentification. To address the challenge of unidentification, Wu et al. [8] developed bounding method to estimate the strength of counterfactual fairness. Similarly, Kilbertus et al. [35] studied the unidentification challenge in the unmeasured confounding situations and designed tools to assess the sensitivity of counterfactual fairness. To unify the path-specific fairness and counterfactual fairness, Wu et al. [10] proposed Path-specific Counterfactual fairness (PC fairness) and developed a general method to deal with the unidentification of PC fairness under complicated circumstances, e.g., hidden confounders, "kite" structures, "w" structures, etc.

2.2 Discrimination Removal

Discrimination removal is also an important task. Existing methods for discrimination removal are categorized into three types: pre-processing, in-processing, and post-processing.

Pre-processing methods [18, 23, 24, 36, 37] modify the historical data to remove discriminatory patterns. For example, [18, 36] developed several methods for modifying data, including *massaging*, which changes the labels of some individuals in the dataset to remove discrimination, *reweighting*, which assigns weights to individuals to balance the dataset,

and *sampling*, which changes the sample sizes of different subgroups to make the dataset discrimination-free. Feldman et al. [37] studied how to remove indirect discrimination from data. The authors proposed to modify all the non-sensitive attributes to ensure that the sensitive attribute cannot be predicted from the non-sensitive attributes. As a result, indirect discrimination is removed since the decision, which is determined by the non-sensitive attributes, cannot be used to predict the sensitive attribute. Wu and Wu [24] leveraged loglinear modeling to capture and measure discrimination, and developed a method for discrimination prevention by modifying significant coefficients of the fitted loglinear model and generate unbiased datasets.

Learning fair representation [16, 38–40] is also one of pre-processing methods. Zemel et al. [16] formulated learning fair representation as an optimization problem where the objective function is a combination of the construction error, statistical disparity, and the prediction error. The obtained representation encodes the original data as well as possible and obfuscates the sensitive information. Edwards and Storkey [38] designed an adversarial approach where an adversary tries to recover the sensitive attribute from the representation and the encoder tries to make the sensitive attribute impossible to recover. Through the adversarial training, the encoder can provide discrimination-free representation. Zhang et al. [40] developed an adversarial framework where an adversary attempts to model the sensitive attribute solely from the predictions rather than the representation.

In-processing methods [11, 13, 14, 41–45] tweak the predictive models. Researchers have developed tweaking methods for the widely used data mining models, e.g., the decision tree classifier [44], the naive Bayes classifier [41], and the logistic regression classifier [42]. For example, Kamiran et al. [44] developed a strategy for relabeling the leaf nodes of a decision tree to make it discrimination-free. Calders and Verwer [41] presented three approaches for the naive Bayes classifier: (1) modifying the conditional probability distribution, (2) training different models for different groups, (3) adding a latent variable to the Bayesian model. Kamishima et al. [45] added a regularization term to probabilistic discriminative models, e.g.,

the logistic regression classifier. Recently, researchers [11, 13, 43, 46] have formulated the in-processing fairness-aware classification as constrained optimization problem. Dwork et al. [46] addressed the problem through constructing a predictive model that achieves both statistical parity and individual fairness, i.e., similar individuals should be treated similarly. Zafar et al. [13] added fairness constraints into the classification so that the classifier learned reduces discrimination. Wu et al. [11] formulated the fairness requirements as convex constraints and provided theoretical guarantees for the obtained classifiers.

Post-processing methods manipulate the predictions produced by predictive models. Technically, the pre-processing methods are applicable in the post-processing phase, e.g., *massaging*, *reweighting*, and *sampling*. Moreover, researchers have developed specific methods for post-processing. Kamiran et al. [47] manipulated the predictions for the individuals that are close to the decision boundary. Thus, they developed two methods, Reject Option based Classifier (ROC) for probabilistic classifiers and Discrimination-Aware Ensemble (DAE) for ensemble classifiers. Hardt et al. [12] derived a general post-processing method for any arbitrary classifier. They formulated a mapping function from the predictions made by a classifier to a new prediction. The fairness requirements are formulated as constraints attached into the mapping function. The resultant optimization is linear and can be solved efficiently.

Recently, learning fair generative models [48–50] becomes a topical research trend. Instead of modifying the training data to remove discriminatory effect, Xu et al. [48] designed a generative model, FairGAN, which can directly generate fair data. The generative model able to generate high quality synthetic data that are similar to real data and prevent the discrimination in the generated data. In the latest version, FairGAN$^+$ [50] contains an extra classifier to simultaneously achieve fair data generation and accurate classification. Xu et al. [51] designed a causal fairness-aware generative adversarial networks (CFGAN) to generate a distribution similar to the given real data as well as subject to various causal fairness criteria.

2.3 Summary

A large amount of discrimination discovery notions and approaches have been developed in the past several years, mostly based on association or correlation. These notions are designed to capture numerous kinds of discrimination, e.g., direct discrimination and indirect discrimination, from various aspects, e.g., system-level, group-level, and individual-level. Based on the notions, enormous discrimination mitigation algorithms have been designed. These algorithms attempt to preserve the data utility or predictive performance as well as achieve fairness with regard to the specified fairness notions.

However, there is a huge gap between association and causation which is not paid enough attention by the research community. Without the support of causation, it is difficult to derive accurate discrimination quantification. Based on the inappropriate quantification, the developed removal approaches may aggravate existing discrimination or suffer from crucial utility loss and predictive errors.

3 Preliminaries

In this chapter, we present the essential notations and fundamental background for the whole dissertation. We start with the notations of describing data and distributions. Then we continue with the necessary fairness notions based on association. Last, we present the structural causal model and intervention which are necessary for causal effect estimation and our proposed frameworks.

3.1 Data and Attributes

We consider a dataset $\mathcal{D}$ with finite samples that are randomly extracted from a joint distribution $\mathcal{P}$. In this dataset, each column represents an attribute, corresponding to a variable in this joint distribution. Throughout this manuscript, we use "attribute" and "variable" interchangeably. We denote an attribute by an uppercase alphabet, e.g., X; denote a subset of attributes by a bold uppercase alphabet, e.g., $\mathbf{X}$. Commonly, an upper letter with i in the subscript is referred to as the i-th variable in a variable set. We denote a domain value of attribute X by a lowercase alphabet, e.g., x; denote a value assignment of attributes $\mathbf{X}$ by a bold lowercase alphabet, e.g., $\mathbf{x}$. When there are multiple values used for one variable, the numeric subscript is adopted, e.g., x_i and x_j represent two arbitrary values of X.

3.2 Association-based Fairness Notions

Association-based notions have been widely adopted into measuring the strength of discrimination and making the judgment of fairness. Technically, these notions measure the association between the sensitive attribute and the decision attribute. In this chapter, the sensitive attribute is denoted by S and the decision attribute is denoted by Y. For the sake of simplicity, S and Y are binary, i.e., s^+ and s^- representing the unprotected/favorable group (e.g., male) and protected/unfavorable group (e.g., female), y^+ and y^- representing the positive decision (e.g., being admitted) and the negative decision (e.g., being rejected).

In this dissertation, we involve two common association-based notions: demographic parity and equality of opportunity. These two fairness notions are described as follows.

3.2.1 Demographic Parity

The main idea of *demographic parity* [18, 19, 43, 52] is the proportions of receiving a positive decision are similar among the demographic groups. To measure the strength of disparity, *risk difference* and *risk ratio* are the most common metrics. The proportion p_1 of receiving a positive decision for the favorable group is denoted by a conditional probability:

$$p_1 = P(Y = y + | S = s^+).$$

Similarly, the proportion p_2 for the unfavorable group is defined as:

$$p_2 = P(Y = y + | S = s^-).$$

Thus, *risk difference* RD is defined as the difference of two proportions:

$$RD = p_1 - p_2.$$

If *risk difference* is small, e.g., close to zero, it implies fairness.

Similarly, *risk ratio* RR is the ratio of two proportions:

$$RR = \frac{p_1}{p_2}.$$

If *risk ratio* is approximate to 1, it implies fairness.

3.2.2 Equality of Opportunity

Equality of opportunity [13] a criterion for measuring discrimination in supervised learning. In supervised machine learning, predictions $\hat{Y}$ are made by predictive functions.

Equality of opportunity means the parity of true positive rate for all demographic groups. In a binary classification model, *equality of opportunity* is satisfied if the following equation holds:

$$P(\hat{Y} = y^+|S = s^+, Y = y^+) = P(\hat{Y} = y^+|S = s^-, Y = y^+).$$

A more rigorous criterion, *equality of odds*, requires the parity of both true positive rate and false positive rate for all demographic groups:

$$P(\hat{Y} = y^+|S = s^+, Y = y) = P(\hat{Y} = y^+|S = s^-, Y = y), y \in \{y^+, y^-\}.$$

The association-based notions have been well studied and plenty of notions have been proposed. Due to the space limitation, we only introduce the necessary notions in this chapter. A detailed discussion and comparison can be found in the tutorial [53].

3.3 Structural Causal Model

In order to investigate the causality problems, Judea Pearl [54] has mathematically developed the concept of the *Structural Causal Model (SCM)*, which describes the mechanism by which the variables are determined.

Definition 1 (Structural Causal Model (*SCM*) [54])**.** *A structural causal model $\mathcal{M}$ is represented by a tuple $\langle \mathbf{U}, \mathbf{V}, \mathbf{F}, P(\mathbf{U}) \rangle$ where*

- $\mathbf{U}$ *is a set of exogenous variables that are determined by factors outside the model. A joint probability distribution $P(\mathbf{U})$ is defined over the variables in $\mathbf{U}$.*
- $\mathbf{V}$ *is a set of endogenous variables that are determined by variables in $\mathbf{U} \cup$ $mathbf V$.*
- $\mathbf{F}$ *is a set of structural equations from $\mathbf{U} \cup \mathbf{V}$ to $\mathbf{V}$. Specifically, for each $V \in \mathbf{V}$, there is a function $f_V \in \mathbf{F}$ mapping from $\mathbf{U} \cup (\mathbf{V} \backslash V)$ to V, i.e., $v = f_V(\mathbf{pa}_V, u_V)$, where $\mathbf{pa}_V$ is a realization of a set of endogenous variables $\mathbf{Pa}_V \in \mathbf{V} \setminus V$ that directly determines V, and u_V is a realization of a set of exogenous variables that directly determines V.*

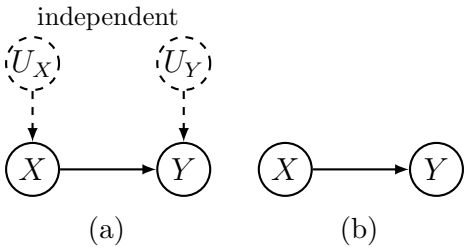


Figure 3.1: Causal graphs of a Markovian model.

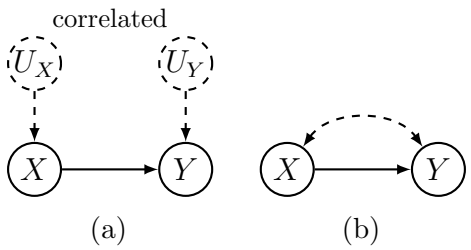


Figure 3.2: Causal graphs of a semi-Markovian model.

If all exogenous variables in $\mathbf{U}$ are mutually independent, then the causal model is called a *Markovian model*. If any pair of exogenous variables in $\mathbf{U}$ is not independent, the causal model is called a *semi-Markovian model*.

Each causal model $\mathcal{M}$ is associated with a graphical causal model, referred to as a *causal graph* $\mathcal{G} = \langle \mathcal{V}, \mathcal{E} \rangle$, where $\mathcal{V}$ is a set of nodes and $\mathcal{E}$ is a set of edges. Each node in $\mathcal{V}$ corresponds to a variable in $\mathbf{V} \cup \mathbf{U}$. Each edges in $\mathcal{E}$ is directed, denoted by a single arrowhead arc $\rightarrow$, and points from each member of $\mathbf{Pa}_X$ toward X to represent the direct causal relationship from this member of $\mathbf{Pa}_X$ toward X. In many applications, the exogenous variables $\mathbf{U}$ (including the outgoing arrows) are omitted from the causal graph, resulting a simple causal graph. In the graphs of Markovian models, e.g., Figure 3.1a, the omission is made without any other changes. A simplified causal graph is shown in Figure 3.1b. However, if any pair of exogenous variables is correlated, a bi-directed edges, denoted by a dotted arc with two arrowheads $\leftarrow\!\!\cdots\!\!\rightarrow$, is drawn between them to represent the existence of a common cause (a.k.a a con-founder). Thus, in the graphs of semi-Markovian models, e.g., Figure 3.2a, the omission is made and a bi-directed edge is added as shown in Figure 3.2b. In the rest of this dissertation, a causal graph refers to the one where the exogenous variables are omitted. In this setting, the nodes $\mathcal{V}$ in a causal graph and variables $\mathbf{V}$ in data $\mathcal{D}$ are used interchangeably. The causal graph associated with a Markovian model is a *Directed Acyclic Graph (DAG)*. The causal graph with a semi-Markovian model is an acyclic graph with dotted bi-directed edges.

Standard terminologies in the graph theory are applicable in the causal graph, e.g.,

parent, child, etc. For a node X, we also use the symbol $\mathbf{Pa}_X$ to denote its parental nodes and use $\mathbf{Ch}_X$ to denote its children. A *path* in a graph is a sequence of edges which concatenates a sequence of nodes. A *directed path* is one where all edges are directed in the same direction. A *causal path* from X to Y is a directed path which starts from X and ends with Y.

In the Markovian model, the directed acyclic causal graph allows one efficiently decomposes the joint distributions $P(\mathbf{x})$ into conditional probabilities using the factorization formula [55]

$$P(\mathbf{x}) = \prod_{x_i \in \mathbf{x}} P(x_i \mid \mathbf{pa}_{X_i}), \tag{3.1}$$

where $P(x_i \mid \mathbf{pa}_{X_i})$ is the conditional probability associated with X_i.

3.3.1 Intervention and Causal Inference

In the causal model, the *do*-operator [54] simulates the physical interventions that force some variables $\mathbf{X}$ to take certain constants $\mathbf{x}$. Formally, the intervention that sets the values of $\mathbf{X}$ to $\mathbf{x}$ is denoted by $do(\mathbf{X} = \mathbf{x})$. The intervention $do(\mathbf{X} = \mathbf{x})$ manipulates the structural causal model and the graphical causal model (a.k.a the causal graph). In the structural causal model $\mathcal{M}$, this intervention substitutes the original equation $X = f(\mathbf{Pa}_X, U_X)$ with $X = x$ for every $X \in \mathbf{X}$. The causal model after performing $do(\mathbf{x})$ is referred to as a sub-model, denoted by $\mathcal{M}_{\mathbf{x}}$. The causal graph $\mathcal{G}_{\mathbf{x}}$ associated with $\mathcal{M}_{\mathbf{x}}$ is a variant of $\mathcal{G}$ where this intervention deletes all the incoming edges to the nodes $\mathbf{X}$ and sets $\mathbf{X}$ to $\mathbf{x}$. For any endogenous variables $\mathbf{Y} \in \mathbf{V} \backslash \mathbf{X}$ which are affected by this intervention, their post-interventional variants in sub-model $\mathcal{M}_{\mathbf{x}}$ are denoted by $\mathbf{Y}_{\mathbf{x}}$. The distribution of $\mathbf{Y}_{\mathbf{x}}$ is called the post-intervention distribution of $\mathbf{Y}$ under $do(\mathbf{x})$, denoted by $P(\mathbf{Y} = \mathbf{y} \mid do(\mathbf{X} = \mathbf{x}))$, $P(\mathbf{y} \mid do(\mathbf{x}))$ or simply $P(\mathbf{y}_{\mathbf{x}})$.

Causal inference is a process of estimating the causal quantities, e.g., the post-interventional distribution $P(\mathbf{y} \mid do(\mathbf{x}))$, from purely observational data and the causal graph. For instance, the post-interventional distribution $P(\mathbf{y} \mid do(\mathbf{x}))$ for any Markovian model can

be expressed as a truncated factorization formula [54]

$$P(\mathbf{y} \mid do(\mathbf{x})) = \prod_{Y \in \mathbf{Y}} P(y \mid \mathbf{pa}_Y)\delta_{\mathbf{X}=\mathbf{x}}, \tag{3.2}$$

where $\delta_{\mathbf{X}=\mathbf{x}}$ means assigning variables in $\mathbf{X}$ involved in the term ahead with the corresponding values in $\mathbf{x}$. Specifically, the post-intervention distribution of a single variable Y given an intervention on a single variable X is given by

$$P(y \mid do(x)) = \sum_{\mathbf{v}'} \prod_{V \in \mathbf{V}\backslash\{X\}} P(v \mid \mathbf{pa}_V)\delta_{X=x}, \tag{3.3}$$

where the summation is a marginalization that traverses all value combinations of $\mathbf{V}' = \mathbf{V}\backslash\{X, Y\}$.

The truncated factorization formula enables the estimation of post-interventional distributions from the observational data in Markovian models. Yet a more challenging problem lies in the semi-Markovian model where the bi-directed edges imply the existence of hidden con-founders and the post-interventional quantities are not unique. It is referred to as *identification* whether a causal quantity can be uniquely estimated from the observational data.

3.3.2 Identification of Causal Quantities

Identification is essential for causal inference as it determines whether a causal quantity, e.g., $P(\mathbf{y} \mid do(\mathbf{x}))$, is consistently derived from the observed data without specifying the whole causal model $\mathcal{M}$.

The definition of *identifiability* is given as follows.

Definition 2 (Identifiability [54])**.** *Let $\mathcal{Q}(\cdot)$ be any computable quantity of a class of models. $\mathcal{Q}$ is identifiable if, for any pair of models $\mathcal{M}_1$ and $\mathcal{M}_2$ from this class, $\mathcal{Q}(\mathcal{M}_1) = \mathcal{Q}(\mathcal{M}_2)$ whenever $P_{\mathcal{M}_1}(\mathbf{v}) = P_{\mathcal{M}_1}(\mathbf{v})$.*

In the context of causal inference, $\mathcal{Q}$ is an arbitrary causal quantity, e.g., the post-interventional distribution $P(\mathbf{y} \mid do(\mathbf{x}))$. According to Definition 2, a causal quantity is identifiable if the estimation is unique given the observational data which are compatible with many potential contradictory causal models. In other words, an unidentifiable quantity would obtain two or more contradictory values given the observational data and the causal graph and in theory, it is impossible to distinguish which one is true. This definition of identifiability is applicable to other types of quantities, e.g., path-specific quantities and counterfactual quantities.

3.3.3 Total Causal Effect

The ultimate task of causal inference is to uncover the cause-effect relationships between variables. Thanks to the *do*-operator, the total causal effect of X on Y is defined in Definition 3 [54]. Note that in this definition, the effect of the intervention is transmitted along all causal paths from the cause X to the effect Y.

Definition 3 (Total causal effect)**.** *The total causal effect $TE(x_2, x_1)$ measures the effect of the change of X from x_1 to x_2 on $Y = y$ transmitted along all causal paths from X to Y. It is given by*

$$TE(x_2, x_1) = P\left(y \mid do(x_2)\right) - P\left(y \mid do(x_1)\right).$$

In the total causal effect, the interventions are performed for all individuals and all variables, thus the effect is aggregated over the whole populations and transmitted via all causal paths. By specifying individuals and causal paths on which the interventions are performed, the customized effects can be defined and evaluated, e.g., the path-specific effect and the counterfactual effect. Since the path-specific effect and the counterfactual effect are only related to partial chapters, we introduce them in the corresponding chapters.

4 Discrimination Discovery and Removal from Classification Data

4.1 Introduction

In the legal and social science fields, discrimination is divided into direct and indirect discrimination. Many approaches have been proposed to deal with both direct and indirect discrimination but significant issues exist. Technically, the difference in decisions across the protected and non-protected groups is a combined (not necessarily linear) effect of direct discrimination, indirect discrimination, and explainable effect that should not be considered as discrimination (e.g., the difference in average income of females and males caused by their different working hours per week). However, existing methods cannot explicitly and correctly identify the three different effects when measuring discrimination. For example, the classic metrics *risk difference*, *risk ratio*, *relative chance*, *odds ratio*, etc. [56] treat all the difference in decisions as discrimination. *Conditional discrimination* [18] realized the explainable effect but failed to correctly measure it. They also failed to distinguish the effects of direct and indirect discrimination. For discrimination removal, a general requirement is to preserve the data utility, i.e., how the distorted data is close to the original one, while achieving non-discrimination. As we shall show in the experiments, a crude method that totally removes all connections between the sensitive attribute and decision (e.g., in [37]) can eliminate discrimination but may suffer significant utility loss. To maximize the data utility, it is necessary to first accurately measure the discriminatory effects.

The causal modeling-based discrimination detection has been proposed most recently [26–28] for improving the correlation based approaches. However, these work also do not tackle indirect discrimination. In this chapter, we develop a framework for discovering and removing both direct and indirect discrimination based on the causal model. A causal model [54] is a structural equation-based mathematical object that describes the causal mechanisms of a system. Each causal model is associated with a causal graph for friendly

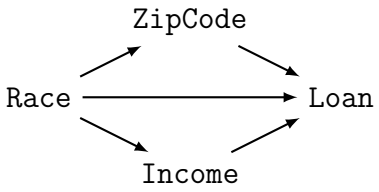


Figure 4.1: The toy model.

causal inference, where causal effects are carried by the *causal paths* that trace arrows pointing from the cause to the effect. Using the causal model, direct and indirect discrimination can be respectively captured by the causal effects of the sensitive attribute on the decision transmitted along different causal paths. To be specific, direct discrimination is modeled as the causal effect transmitted along the direct path from the sensitive attribute to the decision. Indirect discrimination, on the other hand, is modeled as the causal effect transmitted along other causal paths that contain any unjustified attribute.

For example, consider a toy model of a loan application system shown in Figure 4.1. Assume that we treat `Race` as the sensitive attribute, `Loan` as the decision, and `ZipCode` as the unjustified attribute that triggers redlining. Direct discrimination is then transmitted along path `Race` $\rightarrow$ `Loan`, and indirect discrimination is transmitted along path `Race` $\rightarrow$ `ZipCode` $\rightarrow$ `Loan`. Assume that the use of `Income` can be objectively justified as it is reasonable to deny a loan if the applicant has low income. In this case, path `Race` $\rightarrow$ `Income` $\rightarrow$ `Loan` is explainable, which means that part of the difference in loan issuance across different race groups can be explained by the fact that some race groups in the dataset tend to be under-paid.

As shown above, measuring discrimination based on the causal graph requires to measure the causal effect transmitted along certain causal paths. To this end, we employ the technique of the *path-specific effect* [2,3]. We define direct/indirect discrimination as different path-specific effects, and attempt to compute them using the observational data. In theory, the path-specific effect is not always able to be computed from the observational data. This situation is referred to as the unidentifiability of the path-specific effect. We show that direct

discrimination is always identifiable, but indirect discrimination is not identifiable in some cases. For the unidentifiable situation, we provide an upper bound and a lower bound to the effect of indirect discrimination, which is achieved by representing the unidentifiable effect as the expression of counterfactual statements and then scaling up and down specific components of the expression. Based on the theoretical results, we propose effective algorithms that can deal with both identifiable and unidentifiable situations, including algorithms for discovering direct/indirect discrimination, as well as algorithms for precisely removing both types of discrimination while retaining good data utility. The experiments using real datasets show that our approaches are effective in discovering and removing discrimination, ensuring that all types of discrimination are removed while only small utility loss is incurred.

The rest of the chapter is organized as follows. Section 4.2 summarizes the related work. Section 4.4 proposes the criteria and algorithms for discovering and removing both direct and indirect discrimination based on the path-specific effect. Section 4.5 deals with the situation where the indirect discrimination cannot be exactly measured from the observational data according to the unidentifiability of the path-specific effect. The experimental setup and results are discussed in Section 4.6. Finally, Section 4.7 concludes the chapter.

4.2 Related Work

Discrimination discovery has been widely studied and many techniques have been proposed in the literature. A general discussion about the literature is given in Chapter 2. In this section, we focus on direct discrimination, indirect discrimination, and explainable effect.

In 2011, Zliobaite et al. [18] proposed "conditional discrimination", i.e., part of discrimination may be explained by other legally grounded attributes. The task was to evaluate to which extent the discrimination apparent for a group is explainable on a legal ground. The metric is based on the difference of the positive decision proportions for the protected and non-protected groups. Hajian and Domingo [23] quantified the direct and indirect discrimination using *extend lift* (*elift*) over association rules. Direct discrimination

is identified if the *elift* of the sensitive attribute and the context attribute to the decision attribute is larger than a threshold. Indirect discrimination exists if the *elift* of two context attributes that are strongly correlated with the sensitive attributes to the decision attribute is significant. Feldman et al. [37] studied how to remove indirect discrimination from data. The authors modify all the non-sensitive attributes to ensure that the sensitive attribute cannot be predicted from the non-sensitive attributes. As a result, indirect discrimination is removed since the decision, which is determined by the non-sensitive attributes, cannot be used to predict the sensitive attribute.

All of the above works are mainly based on correlation or association. Recently, several studies have been devoted to analyzing discrimination from the causal perspective. Bonchi et al. [25] proposed a framework based on the Suppes-Bayes causal network and developed several random-walk-based methods to detect different types of discrimination. However, it is unclear how the number of random walks is related to practical discrimination metrics. In addition, the construction of the Suppes-Bayes causal network is impractical with the large number of attribute-value pairs. Studies in [26–28] are built on causal modeling and the associated causal graph, but cannot deal with indirect discrimination. The causal model [54] is a mathematical object that describes the causal mechanisms of a system as a set of structural equations. With well-established conceptual and algorithmic tools, the causal model provides a general, formal, yet friendly calculus of causal effects. In this chapter, we adopt the causal model for the quantitative measuring of both direct/indirect discrimination. Specifically, we focus on the technique of path-specific effect [2] that measures the causal effect that is transmitted along certain paths in the causal graph. A recent work [31] proposes similar discrimination criteria that also consider indirect discrimination. However, they are more simplified in order to avoid the complexity in measuring path-specific effects. In addition, [31] suffers inherent limitations: (1) its proposed discrimination criteria can only qualitatively determine the existence of the discrimination, but cannot quantitatively measure the amount of discriminatory effects as we do; (2) its proposed algorithms for avoiding discrimination

proposed only work under the linearity assumptions about the underlying causal model while our methods make no assumption.

For the unidentifiability of the path-specific effect, a recent work [29] proposes three principled approaches: (1) obtaining the data on exogenous variables $\mathbf{U}$; (2) considering an identifiable path-specific effect that includes the paths of interest and some other paths; and (3) deriving bounds for unidentifiable path-specific effects, which is claimed to be an open problem in general. In this chapter, we deal with this issue by adopting the third approach.

4.3 Preliminaries

In Chapter 3, the causal model, the causal graph, intervention, and the total causal effect have been introduced. In this chapter, it is further assumed that the causal model is a *Markovian* model, which means that all exogenous variables in $\mathbf{U}$ are mutually independent.

The total causal effect measures the aggregated causal effect between two attributes over all individuals through all causal paths. The path-specific effect is an extension to the total causal effect in the sense that the effect of the intervention is transmitted only along a subset of causal paths from X to Y [2]. Denote a subset of causal paths by π. The π-specific effect considers a counterfactual situation where the effect of X on Y with the intervention is transmitted along π, while the effect of X on Y without the intervention is transmitted along paths not in π, i.e., $\bar{\pi}$. We denote by $P(y \mid do(x_2|_\pi, x_1|_{\bar{\pi}}))$ the distribution of Y after an intervention of changing X from x_1 to x_2 with the effect transmitted along π. Then, the π-specific effect of X on Y is described as follows.

Definition 4 (Path-specific effect)**.** *Given a path set π, the π-specific effect $PSE_\pi(x_2, x_1)$ measures the effect of the change of X from x_1 to x_2 on $Y = y$ transmitted along π. It is given by*

$$PSE_\pi(x_2, x_1) = P\left(y \mid do(x_2|_\pi, x_1|_{\bar{\pi}})\right) - P\left(y \mid do(x_1)\right).$$

The identifiability of path-specific effect $PSE_\pi(x_2, x_1)$, i.e., whether it can be computed

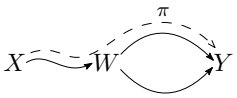


Figure 4.2: The "kite" pattern.

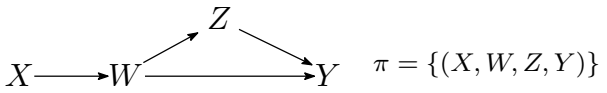


Figure 4.3: The recanting witness criterion satisfied.

from the observational data, depends on the identifiability of $P(y \mid do(x_2|_\pi, x_1|_{\bar{\pi}}))$. The authors in [2] have given the necessary and sufficient condition for $P(y \mid do(x_2|_\pi, x_1|_{\bar{\pi}}))$ to be identifiable, known as the recanting witness criterion.

Definition 5 (Recanting witness criterion)**.** *Given a path set π pointing from X to Y, let W be a node in $\mathcal{G}$ such that: 1) there exists a path from X to W which is a segment of a path in π; 2) there exists a path from W to Y which is a segment of a path in π; 3) there exists another path from W to Y which is not a segment of any path in π. Then, the recanting witness criterion for the π-specific effect is satisfied with W as a witness.*

The graphical pattern of the recanting witness criterion is known as the "kite" pattern, as shown in Figure 4.2. Figure 4.3 shows an example where $\pi = \{(X, W, Z, Y)\}$. It is easy to see that the recanting witness criterion is satisfied with W as the witness.

Theorem 1 (Identifiability of Path-specific Effect)**.** *For path-specific effect $PSE_\pi(x_2, x_1)$, $P(y \mid do(x_2|_\pi, x_1|_{\bar{\pi}}))$ can be computed from the observational data if and only if the recanting witness criterion for the π-specific effect is not satisfied.*

If the recanting witness criterion is not satisfied, $P(y \mid do(x_2|_\pi, x_1|_{\bar{\pi}}))$ can be computed as shown in Theorem 2 [3].

Theorem 2. *For the path-specific effect $PSE_\pi(x_2, x_1)$, if the recanting witness criterion is not satisfied, then $P(y \mid do(x_2|_\pi, x_1|_{\bar{\pi}}))$ can be computed in following steps. Firstly, express $P(y \mid do(x_1))$ as the truncated factorization formula according to Eq.* (3.3)*. Secondly, divide the children of X other than Y into two sets $\mathbf{S}_\pi$ and $\bar{\mathbf{S}}_\pi$, i.e., $\mathbf{Ch}_X \backslash \{Y\} = \mathbf{S}_\pi \cup \bar{\mathbf{S}}_\pi$. Let $\mathbf{S}_\pi$ contain X's each child S where edge $X \to S$ is a segment of a path in π; let $\bar{\mathbf{S}}_\pi$ contain X's each child S where either S is not included in any path from C to E, or edge $X \to S$ is a*

segment of a path not in π. Finally, replace values x_1 with x_2 for the terms corresponding to nodes in $\mathbf{S}_\pi$, and keep values x_1 unchanged for the terms corresponding to nodes in $\bar{\mathbf{S}}_\pi$.

Note that the above computation requires $\mathbf{S}_\pi \cap \bar{\mathbf{S}}_\pi = \emptyset$. Theorem 1 is reflected here in the sense that: $\mathbf{S}_\pi \cap \bar{\mathbf{S}}_\pi = \emptyset$ if and only if the recanting witness criterion for the π-specific effect is not satisfied.

4.4 Discrimination Discovery and Removal

4.4.1 Modeling Direct/Indirect Discrimination as Path-Specific Effects

Consider a historical dataset $\mathcal{D}$ that contains a group of tuples, each of which describes the profile of an individual. Each tuple is specified by a set of attributes $\mathbf{V}$, including the sensitive attributes, the decision, and the non-sensitive attributes. Among the non-sensitive attributes, assume there is a set of attributes that cannot be objectively justified if used in the decision making process, which we refer to as the *redlining attributes* denoted by $\mathbf{R}$. We denote the sensitive attribute by C associated with two domain values c^- (e.g., female) and c^+ (e.g., male); denote the decision by E associated with two domain values e^- (i.e., negative decision) and e^+ (i.e., positive decision). For simplifying representation, we also make two reasonable assumptions: (1) C has no parent in the causal graph $\mathcal{G}$; (2) E has no child in the causal graph $\mathcal{G}$. The first one is due to the fact that the sensitive attribute is usually an inherent nature of an individual, and second one is because that the decision E is usually the output of a decision-making system. We assume that a causal graph $\mathcal{G}$ can be built to correctly represent the causal structure of dataset $\mathcal{D}$. Many algorithms have been proposed to learn the causal graph from data [57–60].

We consider discrimination is the causal effect of the sensitive attribute C on the decision attribute E. As we have discussed, the total causal effect of C on E is a combination of direct/indirect discriminatory effects and the explainable effects. To distinguish the different effects, we model them as the causal effects transmitted along different paths. For

direct discrimination, we consider the causal effect transmitted along the direct edge from C to E, i.e., $C \to E$. Define π_d as the path set that contains only $C \to E$. Then, the above causal effect that is caused by the change of C from c^- to c^+ is given by the π_d-specific effect $PSE_{\pi_d}(c^+, c^-)$. For a better understanding, the physical meaning of $PSE_{\pi_d}(c^+, c^-)$ can be explained as the expected change in decisions of individuals from protected group c^-, if the decision makers are told that these individuals were from the other group c^+. When applied to the example in Figure 4.1, it means the expected change in loan approval of the disadvantage group (e.g., black), if the bank was instructed to treat these applicants as from the advantage group (e.g., white). We can see that the π_d-specific effect perfectly follows the definition of direct discrimination in law and hence is an appropriate measure for direct discrimination.

Similarly, for indirect discrimination, we consider the causal effect transmitted along the indirect paths from C to E that contain the redlining attributes. Given the set of redlining attributes $\mathbf{R}$, we define π_i as the path set that contains all the causal paths from C to E which pass through $\mathbf{R}$, i.e., each of the paths includes at least one node in $\mathbf{R}$. Thus, the above causal effect is given by the π_i-specific effect $PSE_{\pi_i}(c^+, c^-)$. The physical meaning of $PSE_{\pi_i}(c^+, c^-)$ is the expected change in decisions of individuals from protected group c^-, if the values of the redlining attributes in the profiles of these individuals were changed as if they were from the other group c^+. When applied to the example in Figure 4.1, it means the expected change in loan approval of the disadvantage group if they had the same racial makeups shown in the ZIP code as the advantage group. As can be seen, the π_i-specific effect also follows the definition of indirect discrimination and is appropriate for measuring indirect discrimination.

Therefore, we have the following claim.

Claim 1. *The effect of direct discrimination is captured by the π_d-specific effect $PSE_{\pi_d}(c^+, c^-)$, and the effect of indirect discrimination is captured by the π_i-specific effect $PSE_{\pi_i}(c^+, c^-)$.*

Based on the above path-specific effect metrics, we propose the criterion for identifying

direct and indirect discrimination. We define that direct discrimination against protected group c^- exists if $PSE_{\pi_d}(c^+, c^-) > \tau$, where $\tau > 0$ is a user-defined threshold for discrimination depending on the law. For instance, the 1975 British legislation for sex discrimination sets $\tau = 0.05$, namely a 5% difference. Similarly, given the redlining attributes $\mathbf{R}$, we define that indirect discrimination against protected group c^- exists if $PSE_{\pi_i}(c^+, c^-) > \tau$. To avoid reverse discrimination, we do not specify which group is the protected group. As a result, we give the following criterion.

Theorem 3. *Given the sensitive attribute C, the decision E, and redlining attributes $\mathbf{R}$, direct discrimination exists if either $PSE_{\pi_d}(c^+, c^-) > \tau$ or $PSE_{\pi_d}(c^-, c^+) > \tau$ holds, and indirect discrimination exists if either $PSE_{\pi_i}(c^+, c^-) > \tau$ or $PSE_{\pi_i}(c^-, c^+) > \tau$ holds.*

The following theorem shows how to compute $PSE_{\pi_d}(c^+, c^-)$ and $PSE_{\pi_i}(c^+, c^-)$ from the observational data by using Theorem 2.

Theorem 4. *The π_d-specific effect $PSE_{\pi_d}(c^+, c^-)$ is given by*

$$PSE_{\pi_d}(c^+, c^-) = \sum_{\mathbf{Q}} \big(P(e^+|c^+, \mathbf{q})P(\mathbf{q}|c^-)\big) - P(e^+|c^-), \tag{4.1}$$

where $\mathbf{Q}$ is the parents of E except C, i.e., $\mathbf{Q} = \mathbf{Pa}_E \backslash \{C\}$. For the π_i-specific effect $PSE_{\pi_i}(c^+, c^-)$, divide C's children other than E into $\mathbf{S}_{\pi_i}$ and $\bar{\mathbf{S}}_{\pi_i}$ whose definitions are the same as those in Theorem 2. If $\mathbf{S}_{\pi_i} \cap \bar{\mathbf{S}}_{\pi_i} = \emptyset$, then $PSE_{\pi_i}(c^+, c^-)$ is given by

$$\begin{aligned} PSE_{\pi_i}(c^+, c^-) = \sum_{\mathbf{v}'} \Bigg(& P(e^+|c^-, \mathbf{q}) \prod_{G \in \mathbf{S}_{\pi_i}} P(g|c^+, \mathbf{pa}_G \backslash \{C\}) \\ & \times \prod_{H \in \bar{\mathbf{S}}_{\pi_i} \backslash \{E\}} P(h|c^-, \mathbf{pa}_H \backslash \{C\}) \prod_{O \in \mathbf{V} \backslash \mathbf{Ch}_C} P(o|\mathbf{pa}_O) \Bigg) - P(e^+|c^-), \end{aligned} \tag{4.2}$$

where $\mathbf{V}' = \mathbf{V} \backslash \{C, E\}$. It can be simplified to

$$PSE_{\pi_i}(c^+, c^-) = \sum_{\mathbf{q}} \big(P(e^+|c^-, \mathbf{q})P(\mathbf{q}|c^+)\big) - P(e^+|c^-), \tag{4.3}$$

if π_i contains all causal paths from C to E except direct edge $C \to E$.

Proof. According to the definition of $PSE_{\pi_d}(c^+, c^-)$, we have

$$PSE_{\pi_d}(c^+, c^-) = P(e^+ \mid do(c^+|_{\pi_d}, c^-|_{\bar{\pi}_d})) - P(e^+ \mid do(c^-)).$$

Since C has no parent, it is straightforward that $P(e^+ \mid do(c^-)) = P(e^+|c^-)$. For $P(e^+ \mid do(c^+|_{\pi_d}, c^-|_{\bar{\pi}_d}))$, following Theorem 2, we express $P(e^+|c^-)$ as the truncated factorization formula, given by

$$P(e^+|c^-) = \sum_{\mathbf{v}'} \left(P(e^+|c^-, \mathbf{q}) \prod_{V \in \mathbf{V}'} P(v \mid \mathbf{pa}_V) \right), \tag{4.4}$$

where $\mathbf{V}' = \mathbf{V} \backslash \{C, E\}$. It can be shown that $\prod_{V \in \mathbf{V}'} P(v \mid \mathbf{pa}_V) = P(\mathbf{v}'|c^-)$. In fact, if we sort all nodes in $\mathbf{V}'$ according to the topological ordering as $\{V_1, \cdots, V_j, \cdots\}$, we can see that all parents of each node V_j are before it in the ordering. In addition, since C has no parent, it must be V_j's non-descendant; since E has no child, it cannot be V_j's parent. Thus, based on the local Markov condition, we have $P(v_j \mid \mathbf{pa}_{V_j}) = P(v_j \mid c^-, v_1, \cdots, v_{j-1})$. According to the chain rule we obtain $P(\mathbf{v}'|c^-)$. Therefore, it follows that

$$P(e^+|c^-) = \sum_{\mathbf{q}} \left(P(e^+|c^-, \mathbf{q}) P(\mathbf{q}|c^-) \right).$$

Then, we divide the children of C into $\mathbf{S}_{\pi_d}$ and $\bar{\mathbf{S}}_{\pi_d}$, and replace c^- with c^+ for the terms corresponding to nodes in $\mathbf{S}_{\pi_d}$. Note that $\mathbf{S}_{\pi_d}$ contains only one node E. As a result, we have

$$P(e^+ \mid do(c^+|_{\pi_d}, c^-|_{\bar{\pi}_i})) = \sum_{\mathbf{q}} \left(P(e^+|c^+, \mathbf{q}) P(\mathbf{q}|c^-) \right),$$

which leads to Eq. (4.1).

For the indirect discrimination, by definition we have

$$PSE_{\pi_i}(c^+, c^-) = P(e^+ \mid do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i})) - P(e^+|do(c^-)).$$

To compute the first term, we also express $P(e^+|c^-)$ as Eq. (4.4), and divide the children of C into $\mathbf{S}_{\pi_i}$ and $\bar{\mathbf{S}}_{\pi_i}$. Then, node set $\mathbf{V}'$ can be divided into three disjoint subsets: $\mathbf{S}_{\pi_i}$, $\bar{\mathbf{S}}_{\pi_i}$ and $\mathbf{V}'\backslash\mathbf{Ch}_C$. We replace c^- with c^+ only for the terms corresponding to nodes in $\mathbf{S}_{\pi_i}$. As a result, we obtain Eq. (4.2).

If π_i contains all causal paths from C to E except $C \to E$, it means that $\mathbf{S}_{\pi_i} = \mathbf{Ch}_C\backslash\{E\}$ and $\bar{\mathbf{S}}_{\pi_i} = \emptyset$. Note that

$$\prod_{G\in\mathbf{Ch}_C\backslash\{E\}} P(g \mid c^+, \mathbf{pa}_G\backslash\{C\}) \prod_{O\in\mathbf{V}'\backslash\mathbf{Ch}_C} P(o \mid \mathbf{pa}_O) = \prod_{V\in\mathbf{V}'} P(v \mid \mathbf{pa}_V),$$

which can be similarly shown to equal to $P(\mathbf{v}'|c^+)$. As a result we obtain Eq. (4.3). □

Theorem 4 shows that $PSE_{\pi_d}(c^+, c^-)$ can always be computed from the observational data but $PSE_{\pi_i}(c^+, c^-)$ may not[1]. This is because the recanting witness criterion for the π_d-specific effect is guaranteed to be not satisfied, but the recanting witness criterion for the π_i-specific effect might be satisfied. The situation where $PSE_{\pi_i}(c^+, c^-)$ cannot be computed is referred to as the unidentifiable situation. How to deal with the unidentifiable situation will be discussed later in the next section.

The following two propositions further show two properties of the path-specific effect metrics.

Proposition 1. *If path set π contains all causal paths from C to E, then we have*

$$PSE_{\pi}(c^+, c^-) = TE(c^+, c^-) = P(e^+|c^+) - P(e^+|c^-).$$

[1]Note that Eq. (4.3) can still be computed from the observational data since $\bar{\mathbf{S}}_{\pi_i} = \emptyset$ when π_i contains all causal paths from C to E except $C \to E$.

The proof can be directly obtained from Definition 4, Definition 3 and Eq. (3.3). $P(e^+|c^+) - P(e^+|c^-)$ is known as the *risk difference* [56] widely used for discrimination measurement in the anti-discrimination literature. Therefore, the path-specific effect metrics can be considered as a significant extension to the risk difference for explicitly distinguishing the discriminatory effects of direct and indirect discrimination from the total causal effect.

Proposition 2. *For any path sets π_d and π_i, we do not necessarily have $PSE_{\pi_d}(c^+, c^-) + PSE_{\pi_i}(c^+, c^-) = PSE_{\pi_d \cup \pi_i}(c^+, c^-)$.*

The proof can be obtained from Definition 4 and Theorem 2. In fact, as shown in [61], the above equality holds if all functions in $\mathbf{F}$ of the causal model are linear, and π_i contains all causal paths from C to E other than $C \to E$. Thus, Proposition 2 implies that if the causal relationship is not linear, then a linear connection between direct and indirect discrimination also does not exist.

4.4.2 Discovery Algorithm

We propose a Path-Specific based Discrimination Discovery (*PSE-DD*) algorithm based on Theorem 3. It first builds the causal graph from the historical dataset, and then computes $PSE_{\pi_d}(\cdot)$ and $PSE_{\pi_i}(\cdot)$ according to Eq. (4.1) and (4.2). The procedure of the algorithm is shown in Algorithm 1.

The complexity of line 6 depends on how to identify $\mathbf{S}_{\pi_i}$ and $\bar{\mathbf{S}}_{\pi_i}$. A straightforward method is to find all paths in π_i, and for C's each child S check whether $C \to S$ is contained in any path in π_i. However, finding all paths between two nodes in a DAG has an exponential complexity. In our algorithm, we examine the existence of a path from S to E passing through $\mathbf{R}$. It can be easily observed that, a node S belongs to $\mathbf{S}_{\pi_i}$ if and only if there exists a path from S to E passing through $\mathbf{R}$ (a path from S to E passing through $\mathbf{R}$ also includes the path where S itself belongs to $\mathbf{R}$). Similarly, S belongs to $\bar{\mathbf{S}}_{\pi_i}$ if and only if there does not exist a path from S to E passing through $\mathbf{R}$. The subroutine of finding $\mathbf{S}_{\pi_i}$ and $\bar{\mathbf{S}}_{\pi_i}$ is presented in Algorithm 2, which checks whether there exists a node $R \in \mathbf{R}$ so that R is S's

Algorithm 1: *PSE-DD*

Input : A historical dataset $\mathcal{D}$, the sensitive attribute C, the decision attribute E, redlining attributes $\mathbf{R}$, a threshold τ.

Output : Direct/indirect discrimination $judge_d$, $judge_i$.

$\mathcal{G} = buildCausalNetwork(\mathcal{D})$;
$judge_d = judge_i = false$;
Compute $PSE_{\pi_d}(\cdot)$ according to Eq. (4.1);
if $PSE_{\pi_d}(c^+, c^-) > \tau \parallel PSE_{\pi_d}(c^-, c^+) > \tau$ **then**
    $judge_d = true$;
Call subroutine $[\mathbf{S}_{\pi_i}, \bar{\mathbf{S}}_{\pi_i}] = DivideChildren(\mathcal{G}, C, E, \mathbf{R})$;
if $\mathbf{S}_{\pi_i} \cap \bar{\mathbf{S}}_{\pi_i} \neq \emptyset$ **then**
    $judge_i = unknown$;
    return $[judge_d, judge_i]$;
Compute $PSE_{\pi_i}(\cdot)$ according to Eq. (4.2);
if $PSE_{\pi_i}(c^+, c^-) > \tau \parallel PSE_{\pi_i}(c^-, c^+) > \tau$ **then**
    $judge_i = true$;
return $[judge_d, judge_i]$;

descendant and E is R's descendant. Since the descendants of all the nodes involved in the algorithm can be obtained by traversing the network starting from C within the time of $O(|\mathcal{E}|)$, the computational complexity of the subroutine is given by $O(|\mathbf{V}|^2 + |\mathcal{E}|)$.

Algorithm 2: subroutine *DivideChildren*

Input : The causal graph $\mathcal{G}$, the sensitive attribute C, the decision attribute E, redlining attributes $\mathbf{R}$.

Output : $\mathbf{S}_{\pi_i}$ and $\bar{\mathbf{S}}_{\pi_i}$.

$\mathbf{S}_{\pi_i} = \emptyset$, $\bar{\mathbf{S}}_{\pi_i} = \emptyset$;
foreach $S \in \mathbf{Ch}_C \backslash \{E\}$ **do**
    foreach $R \in \mathbf{R}$ **do**
        if $R \in \mathbf{De}_S \cup \{S\}$ && $E \in \mathbf{De}_R$ **then**
            $\mathbf{S}_{\pi_i} = \mathbf{S}_{\pi_i} \cup \{S\}$;
        else
            $\bar{\mathbf{S}}_{\pi_i} = \bar{\mathbf{S}}_{\pi_i} \cup \{S\}$;
return $[\mathbf{S}_{\pi_i}, \bar{\mathbf{S}}_{\pi_i}]$;

The computational complexity of *PSE-DD* also depends on the complexities of building the causal graph and computing the path-specific effect according to Eq. (4.1) or (4.2). Many researches have been devoted to improving the performance of network construction [60,62,63]

and probabilistic inference in causal graphs [64, 65]. The complexity analysis can be found in these related literature.

4.4.3 Removal Algorithm

When direct or indirect discrimination is discovered for a dataset, the discriminatory effects need to be removed before the dataset is released for predictive analysis. A naive approach would be simply deleting the sensitive attribute from the dataset, which often incurs significant utility loss. In addition, this approach can eliminate direct discrimination, but indirect discrimination still presents.

We propose a Path-Specific Effect based Discrimination Removal (*PSE-DR*) algorithm to remove both direct and indirect discrimination. The general idea is to modify the causal graph and then use it to generate a new dataset. Specifically, we modify the CPT of E, i.e., $P(e|\mathbf{pa}_E)$, to obtain a new CPT $P'(e|\mathbf{pa}_E)$, so that the direct and indirect discriminatory effects are below the threshold τ. To maximize the utility of the modified dataset, we minimize the Euclidean distance between the joint distribution of the original causal graph (denoted by $P(\mathbf{v})$) and the joint distribution of the modified causal graph (denoted by $P'(\mathbf{v})$). As a result, we obtain the following quadratic programming problem with $P'(e|\mathbf{pa}_E)$ as the variables.

$$
\begin{aligned}
\text{minimize} \quad & \sum_{\mathbf{v}} \Big(P'(\mathbf{v}) - P(\mathbf{v}) \Big)^2 \\
\text{subject to} \quad & PSE_{\pi_d}(c^+, c^-) \leq \tau, \quad PSE_{\pi_d}(c^-, c^+) \leq \tau, \\
& PSE_{\pi_i}(c^+, c^-) \leq \tau, \quad PSE_{\pi_i}(c^-, c^+) \leq \tau, \\
& \forall \mathbf{pa}_E, \quad P'(e^+ \mid \mathbf{pa}_E) + P'(e^- \mid \mathbf{pa}_E) = 1, \\
& \forall \mathbf{pa}_E, e, \quad P'(e \mid \mathbf{pa}_E) \geq 0,
\end{aligned}
$$

where $P'(\mathbf{v})$ and $P(\mathbf{v})$ are computed according to Eq. (3.1) using $P'(e|\mathbf{pa}_E)$ and $P(e|\mathbf{pa}_E)$ respectively, and $PSE_{\pi_d}(\cdot)$ and $PSE_{\pi_i}(\cdot)$ are computed according to Eq. (4.1) and (4.2) respectively using $P'(e|\mathbf{pa}_E)$. The optimal solution is obtained by solving the quadratic

programming problem. After that, the joint distribution of the modified causal graph is computed using Eq. (3.1), and the new dataset is generated based on the joint distribution. The procedure of *PSE-DR* is shown in Algorithm 3

Algorithm 3: *PSE-DR*

Input : The historical dataset $\mathcal{D}$, the sensitive attribute C, the decision attribute E, redlining attributes $\mathbf{R}$, a threshold τ.
Output : Modified dataset $\mathcal{D}^*$.

1 $[judge_d, judge_i] = PSE\text{-}DD(\mathcal{D}, C, E, \mathbf{R}, \tau)$;
2 **if** $[judge_d, judge_i] == [false, false]$ **then**
3 ⌊ **return** $\mathcal{D}$;
4 $\mathcal{G} = buildCausalNetwork(\mathcal{D})$;
5 **if** $judge_i == unknown$ **then**
6 ⌊ Call subroutine *GraphPreprocess*;
7 Obtain the modified CPT of E by solving the quadratic programming problem;
8 Calculate $P^*(\mathbf{v})$ according to Eq. (3.1) using the modified CPTs;
9 Generate $\mathcal{D}^*$ based on $P^*(\mathbf{v})$;
10 **return** $\mathcal{D}^*$;

As stated in Theorems 1 and 4, when the recanting witness criterion is satisfied, the π_i-specific effect cannot be estimated from the observational data. However, the "kite" pattern implies potential indirect discrimination as there exist causal paths from C to E passing through the redlining attributes. Although the indirect discriminatory effect cannot be accurately measured, from a practical perspective, it is still meaningful to ensure non-discrimination while preserving reasonable data utility. As a straightforward method, we can first modify the causal graph to remove the "kite" pattern, and then obtain the modified CPT of E by solving the quadratic programming problem similar to the identifiable situation. To remove the "kite" pattern, for each node $S \in \mathbf{S}_{\pi_i} \cap \bar{\mathbf{S}}_{\pi_i}$, we cut off all the causal paths from S to E that pass through $\mathbf{R}$, so that S would not belong to $\mathbf{S}_{\pi_i}$ any more. Then, we must have $\mathbf{S}_{\pi_i} \cap \bar{\mathbf{S}}_{\pi_i} = \emptyset$ after the modification. When cutting off the paths, we focus on the edge from E's each parent Q, i.e., $Q \to E$. If there exists a path from S to Q passing through $\mathbf{R}$, then edge $Q \to E$ is removed from the network. The pseudo-code of this procedure called *GraphPreprocess* is shown below, which is added as a subroutine in line 5 of *PSE-DR*.

Algorithm 4: subroutine *GraphPreprocess*

Input : The causal graph $\mathcal{G}$, the sensitive attribute C, the decision attribute E, redlining attributes $\mathbf{R}$.

```
foreach S ∈ S_{π_i} ∩ S̄_{π_i} do
    foreach Q ∈ Pa_E do
        foreach R ∈ R do
            if R ∈ De_S && Q ∈ De_R then
                Remove edge Q → E from G;
                Break;
```

The computational complexity of *PSE-DR* depends on the complexity of solving the quadratic programming problem. It can be easily shown that, the coefficients of the quadratic terms in the objective function form a positive definite matrix. According to [66], the quadratic programming can be solved in polynomial time. Finally, it is also worth noting that our approach can be easily extended to handle the situation where either direct or indirect discrimination needs to be removed.

4.5 Dealing with Unidentifiable Situation

Under the unidentifiable situation where the recanting witness criterion is satisfied, *PSE-DD* and *PSE-DR* provide workable but crude solutions to the discrimination discovery and removal. In this section, we develop the refined discrimination discovery and removal algorithms by deriving upper and lower bounds for the unidentifiable indirect discrimination. Compared to the presence of the "kite" pattern, the bounds can be used as better indicators for discovering indirect discrimination, i.e., the upper bound smaller than τ indicates no indirect discrimination, while the lower bound larger than τ indicates its existence. We also prove that the refined removal algorithm is at least as good as *PSE-DR* in term of preserving the data utility. We start by giving several necessary preliminaries in addition to those presented in Section 4.3.

4.5.1 Preliminaries

In Section 4.3, we have shown that variables $\mathbf{Y}$ under an intervention $do(\mathbf{x})$ is still a set of random variables, whose distribution $P(\mathbf{y} \mid do(\mathbf{x}))$ is different from the observational distribution of $\mathbf{Y}$. We denote $\mathbf{Y}$ under intervention $do(\mathbf{x})$ by $\mathbf{Y_x}$, i.e., we define

$$P(\mathbf{y_x}) \triangleq P(\mathbf{Y_x} = \mathbf{y}) \triangleq P(\mathbf{y} \mid do(\mathbf{x})).$$

We can interpret $\mathbf{Y_x}$ as a counterfactual statement, which represents "the value that $\mathbf{Y}$ would have obtained, had $\mathbf{X}$ been $\mathbf{x}$". From the definition of the causal model we can observe that, if all the exogenous variables $\mathbf{U}$ are given, then $\mathbf{Y_x}$ are no longer random variables but are fixed values. We denote the $\mathbf{Y_x}$ under the context of $\mathbf{U} = \mathbf{u}$ by $\mathbf{Y_x}(\mathbf{u})$. In the following we present several properties regarding the counterfactual statement, which are proved to be held in the context of Markovian model [54].

Property 1. *For any variable Y, $Y_{\mathbf{Pa}_Y}$ is independent of the counterfactual statements of all Y's non-descendants.*

Property 2. *For any variable Y, we have*

$$P(y_{\mathbf{pa}_Y}) = P(y \mid \mathbf{pa}_Y).$$

Property 3. *For any set of endogenous variables $\mathbf{Y}$ and any set of endogenous variables $\mathbf{X}$ disjoint of $\{\mathbf{Y}, \mathbf{Pa_Y}\}$, we have*

$$P(\mathbf{y}_{\mathbf{pa_Y},\mathbf{x}}) = P(\mathbf{y}_{\mathbf{pa_Y}}).$$

Property 4. *For any three sets of endogenous variables $\mathbf{X}, \mathbf{Y}, \mathbf{Z}$,*

$$\mathbf{Z_x}(\mathbf{u}) = \mathbf{z} \quad \Longrightarrow \quad \mathbf{Y_x}(\mathbf{u}) = \mathbf{Y}_{\mathbf{x},\mathbf{z}}(\mathbf{u}).$$

Property 1 reflects the local Markov condition. Property 2 renders every parent set $\mathbf{Pa}_Y$ exogenous relative to its child Y. Property 3 reflects the insensitivity of Y to any intervention once its direct causes are held constant. Property 4 states that, if we know the values that $\mathbf{Z}$ would have in certain situation, then the values of any other variables $\mathbf{Y}$ are equivalent to that if we perform an intervention to force $\mathbf{Z}$ to $\mathbf{z}$.

Next, we introduce an essential concept regarding to the unidentifiability of the path-specific effect by using the notion of counterfactual statement. Straightforwardly, by $\mathbf{Y}_\mathbf{x}(\mathbf{u})$ and $P(\mathbf{u})$, we can represent $P(\mathbf{y}_\mathbf{x})$ as

$$P(\mathbf{y}_\mathbf{x}) = \sum_{\mathbf{u}:\mathbf{Y}_\mathbf{x}(\mathbf{u})=\mathbf{y}} P(\mathbf{u}). \tag{4.5}$$

In the same way, we can define the joint distribution of multiple counterfactual statements (which cannot be defined by using the *do*-operator), i.e., $P(\mathbf{Y}_\mathbf{x} = \mathbf{y}, \mathbf{Y}_{\mathbf{x}'} = \mathbf{y}')$ or $P(\mathbf{y}_\mathbf{x}, \mathbf{y}'_{\mathbf{x}'})$, which represents the probability to "$\mathbf{Y}$ would be $\mathbf{y}$ if $\mathbf{X} = \mathbf{x}$ and $\mathbf{Y}$ would be $\mathbf{y}'$ if $\mathbf{X} = \mathbf{x}'$", given as

$$P(\mathbf{y}_\mathbf{x}, \mathbf{y}'_{\mathbf{x}'}) = \sum_{\{\mathbf{u}:\mathbf{Y}_\mathbf{x}(\mathbf{u})=\mathbf{y}, \mathbf{Y}_{\mathbf{x}'}(\mathbf{u})=\mathbf{y}'\}} P(\mathbf{u}).$$

When $\mathbf{x} \neq \mathbf{x}'$, $\mathbf{Y}_\mathbf{x}$ and $\mathbf{Y}_{\mathbf{x}'}$ cannot be measured simultaneously. In fact, it is known that $P(\mathbf{y}_\mathbf{x}, \mathbf{y}'_{\mathbf{x}'})$ is unidentifiable from the observational data even in the Markovian model [67]. We will show that the unidentifiability of the $P(\mathbf{y}_\mathbf{x}, \mathbf{y}'_{\mathbf{x}'})$ is the source of the unidentifiability of the path-specific effect satisfying the recanting witness criterion. However, $P(\mathbf{y}_\mathbf{x}, \mathbf{y}'_{\mathbf{x}'})$ is certainly bounded by the following condition:

$$\sum_{\mathbf{y}'} P(\mathbf{y}_\mathbf{x}, \mathbf{y}'_{\mathbf{x}'}) = P(\mathbf{y}_\mathbf{x}). \tag{4.6}$$

4.5.2 Bounding Indirect Discrimination

Recalling the definition of the path-specific effect (Definition 4), in the π_i-specific effect, $P(e^+ \mid do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i}))$ represents the probability of $E = e^+$ after the intervention of

changing C from c^- to c^+ with the effect transmitted along π_i. By using the notation of the counterfactual statement, we can similarly denote the value of E after the intervention by E_{c^+}. However, keep in mind that different from the original counterfactual statement, here for E_{c^+} the effect of the intervention on C is transmitted only along π_i.

For any variable Y other than C, E, we can also denote their values that would be obtained after the intervention as counterfactual statement Y_{c^+}. Similar to E, the value of Y_{c^+} depends on whether it belongs to a path in π_i. If Y belongs to any path in π_i, then the value of Y_{c^+} will be affected by the intervention. If Y does not belong to any path in π_i, then the value of Y_{c^+} will not be affected by the intervention and remain the same as if $C = c^-$. Based on the causal effect transmission, to obtain Y_{c^+}, we need to know the value of Y's each ancestor W affected by the intervention if there exists a path from W to Y that is a segment of a path in π_i; or we need to know the value of W not affected by the intervention if there exists a path from W to Y that is not a segment of any path in π_i. As can be seen, if W has two emanating edges where one belongs to a path in π_i and the other one does not belong to any path in π_i, we need to simultaneously know the value of W affected by the intervention as well as the one not affected by the intervention. To distinguish these two counterfactual situations, we denote the former by W_{c^+} and the latter by W_{c^-}. According to the definition of the recanting witness criterion (Definition 5), it can be easily shown that W is a node where both W_{c^+} and W_{c^-} are needed if and only if W is a witness for the recanting witness criterion. Here we call such node W a *witness variable/node*.

The above analysis shows that, for each witness variable W, we need to consider two sets of realizations, one obtained by W_{c^+} (denoted as w^+), and the other obtained by W_{c^-} (denoted as w^-). For each variable Y that is not a witness variable, we only consider one set of realizations obtained by Y_{c^+}.

In the following, we derive a general expression of $SE_{\pi_i}(c^+, c^-)$ and then develop its upper and lower bounds when subject to the recanting witness criterion. We first provide a property and a proposition that are needed for the derivation.

Similar to Property 4, in the path-specific effect, if we know the two realizations that witness variables $\mathbf{W}$ would have in both counterfactual situations, then the values of any other variable Y are equivalent to that if we perform an intervention to force $\mathbf{W}$ to these realizations. Thus, we obtain the following property that is directly extended from Property 4.

Property 5. *For endogenous variables X, Y, W, assume that W is a witness variable, x, x' are two realizations of X, and w, w' are two realizations of W. For any π-specific effect of X we have*

$$W_x(\mathbf{u}) = w,\ W_{x'}(\mathbf{u}) = w' \quad \Longrightarrow \quad Y_x(\mathbf{u}) = Y_{x,w^*}(\mathbf{u}),$$

where w^* means that its value is specified by w if there exists a path from W to Y that is a segment of a path in π, and specified by w' otherwise.

Based on Properties 3, 4 and 5, we can prove the following proposition.

Proposition 3. *In π_i-specific effect $PSE_{\pi_i}(c^+, c^-)$, for any endogenous variable Y, use $\mathbf{pa}_Y^+$ to denote the realization of Y's parents meaning that if $\mathbf{Pa}_Y$ contains any witness node W or C, its value is specified by w^+ or c^+ if edge $W \to Y$ belongs to a path in π_i, and specified by w^- or c^- otherwise; and use $\mathbf{pa}_Y^-$ to denote the realization of Y's parents meaning that if $\mathbf{Pa}_Y$ contains any witness node W or C, its value is specified by w^- or c^-. If Y is not a witness variable, we have*

$$P(y_{c^+}, \cdots) = \begin{cases} P(y_{\mathbf{pa}_Y^+}, \cdots) & \text{if } Y \text{ belongs to any path in } \pi_i, \\ P(y_{\mathbf{pa}_Y^-}, \cdots) & \text{otherwise,} \end{cases} \tag{4.7}$$

and if Y is a witness variable, we have

$$P(y_{c^+}, \cdots) = P(y_{\mathbf{pa}_Y^+}, \cdots) \text{ and } P(y_{c^-}, \cdots) = P(y_{\mathbf{pa}_Y^-}, \cdots), \tag{4.8}$$

where $\cdots$ represents all other variables.

Proof. To prove Eq. (4.7), denote Y's parents by $\mathbf{Z}$, i.e., $\mathbf{X} = \mathbf{Pa}_Y$. Assume that $\mathbf{X}$ contains no witness node or C. Then $P(y_{c^+}, \cdots)$ can be written as $P(y_{c^+}, \mathbf{x}_{c^+}, \cdots)$. According to Eq. (4.5), we have

$$P(y_{c^+}, \mathbf{x}_{c^+}, \cdots) = \sum_{\{\mathbf{u}: Y_{c^+}(\mathbf{u})=y, \mathbf{X}_{c^+}(\mathbf{u})=\mathbf{x}, \cdots\}} P(\mathbf{u}).$$

Based on Property 4, we have

$$\mathbf{X}_{c^+}(\mathbf{u}) = \mathbf{x} \quad \Longrightarrow \quad Y_{c^+}(\mathbf{u}) = Y_{c^+,\mathbf{x}}(\mathbf{u}).$$

Since $\mathbf{X} = \mathbf{Pa}_Y$, according to Property 3 we have

$$Y_{c^+,\mathbf{x}}(\mathbf{u}) = Y_{\mathbf{x}}(\mathbf{u}).$$

Therefore, it follows that

$$P(y_{c^+}, \mathbf{x}_{c^+}, \cdots) = \sum_{\{\mathbf{u}: Y_{\mathbf{x}}(\mathbf{u})=y, \mathbf{X}_{c^+}(\mathbf{u})=\mathbf{x}, \cdots\}} P(\mathbf{u}) = P(y_{\mathbf{x}}, \mathbf{x}_{c^+}, \cdots),$$

which can be re-written as $P(y_{\mathbf{pa}_Y^+}, \cdots)$ according to the definition of $\mathbf{pa}_Y^+$.

Assume that $\mathbf{X}$ contains any witness node W or C. Then by applying Property 5, we can similarly obtain $P(y_{c^+}, \cdots) = P(y_{\mathbf{x}^*}, \cdots)$, where $\mathbf{x}^*$ means that if any witness node W or C connects Y with a segment of a path in π_i then its value is specified by w^+ or c^+, and specified by w^- or c^- otherwise. According to the definition of $\mathbf{pa}_Y^+$ and $\mathbf{pa}_Y^-$, $P(y_{\mathbf{x}^*}, \cdots)$ can be re-written as $P(y_{\mathbf{pa}_Y^+}, \cdots)$ if Y belongs to any path in π_i, and $P(y_{\mathbf{pa}_Y^-}, \cdots)$ otherwise.

If Y is a witness node, then the first case and second case of Eq. (4.8) can be proved similarly to the first case and second case of Eq. (4.7) respectively. □

For ease of representation, we divide all nodes on the causal paths from C to E (except C and E) into three disjoint subsets: the subset of witness nodes (denoted by $\mathbf{W}$), the subset of nodes not in $\mathbf{W}$ that belong to paths in π_i (denoted by $\mathbf{A}$), and the subset of nodes not in

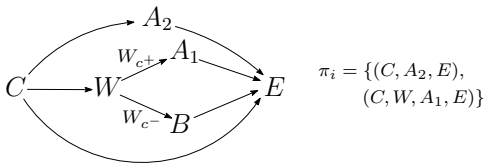


Figure 4.4: π_i-specific effect satisfying recanting witness criterion.

$\mathbf{W}$ that do not belong to any path in π_i (denoted by $\mathbf{B}$)[2]. An example is shown in Figure 4.4 where $\mathbf{W} = \{W\}$, $\mathbf{A} = \{A_1, A_2\}$, and $\mathbf{B} = \{B\}$. The notations on the edges represent the specification of the values of each node's parents.

In Theorem 5 we give the general expression of $PSE_{\pi_i}(c^+, c^-)$. Since by definition we have $PSE_{\pi_i}(c^+, c^-) = P(e^+ \mid do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i})) - P(e^+|c^-)$ where the second term is trivial, we focus on the general expression of $P(e^+ \mid do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i}))$.

Theorem 5. *When subject to the recanting witness criterion,* $P(e^+ \mid do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i}))$ *is given by*

$$\begin{aligned} &P(e^+ \mid do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i})) \\ &= \sum_{\mathbf{a},\mathbf{b},\mathbf{w}^+,\mathbf{w}^-} P(e^+|c^-, \mathbf{q}) \prod_{A\in\mathbf{A}} P(a|\mathbf{pa}_A^+) \prod_{B\in\mathbf{B}} P(b|\mathbf{pa}_B^-) \prod_{W\in\mathbf{W}} P(w^+_{\mathbf{pa}_W^+}, w^-_{\mathbf{pa}_W^-}). \end{aligned} \tag{4.9}$$

Proof. For simplicity and without loss of generality, assume that all nodes are along the causal paths from C to E. We can re-write distribution $P(e^+ \mid do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i}))$ as the sum of the joint distribution as follows.

$$\begin{aligned} &P(e^+ \mid do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i})) \triangleq P(E_{c^+} = e^+) \\ &= \sum_{\mathbf{a},\mathbf{b},\mathbf{w}^+,\mathbf{w}^-} P(E_{c^+} = e^+, \mathbf{A}_{c^+} = \mathbf{a}, \mathbf{B}_{c^+} = \mathbf{b}, \mathbf{W}_{c^+} = \mathbf{w}^+, \mathbf{W}_{c^-} = \mathbf{w}^-) \\ &\triangleq \sum_{\mathbf{a},\mathbf{b},\mathbf{w}^+,\mathbf{w}^-} P(e^+_{c^+}, \underbrace{a_{c^+}, \cdots}_{A\in\mathbf{A}}, \underbrace{b_{c^+}, \cdots}_{B\in\mathbf{B}}, \underbrace{w^+_{c^+}, w^-_{c^-}, \cdots}_{W\in\mathbf{W}}). \end{aligned}$$

[2]Redlining attributes can be contained in $\mathbf{W}$ and $\mathbf{A}$ but cannot be contained in $\mathbf{B}$.

By using Proposition 3, it follows that

$$P(e^+|do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i})) = \sum_{\mathbf{a},\mathbf{b},\mathbf{w}^+,\mathbf{w}^-} P(e^+_{c^-,\mathbf{q}}, \underbrace{a_{\mathbf{pa}^+_A}, \cdots}_{A\in\mathbf{A}}, \underbrace{b_{\mathbf{pa}^-_B}, \cdots}_{B\in\mathbf{B}}, \underbrace{w^+_{\mathbf{pa}^+_W}, w^-_{\mathbf{pa}^-_W}, \cdots}_{W\in\mathbf{W}}).$$

According to Property 1, the counterfactual statement of each variable is independent of all its non-descendants. Thus, we have

$$P(e^+ \mid do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i})) = \sum_{\mathbf{a},\mathbf{b},\mathbf{w}^+,\mathbf{w}^-} P(e^+_{c^-,\mathbf{q}}) \prod_{A\in\mathbf{A}} P(a_{\mathbf{pa}^+_A}) \prod_{B\in\mathbf{B}} P(b_{\mathbf{pa}^-_B}) \prod_{W\in\mathbf{W}} P(w^+_{\mathbf{pa}^+_W}, w^-_{\mathbf{pa}^-_W}).$$

According to Property 2, it follows that

$$\begin{aligned} &P(e^+ \mid do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i})) \\ &\sum_{\mathbf{a},\mathbf{b},\mathbf{w}^+,\mathbf{w}^-} P(e^+|c^-,\mathbf{q}) \prod_{A\in\mathbf{A}} P(a|\mathbf{pa}^+_A) \prod_{B\in\mathbf{B}} P(b|\mathbf{pa}^-_B) \prod_{W\in\mathbf{W}} P(w^+_{\mathbf{pa}^+_W}, w^-_{\mathbf{pa}^-_W}). \end{aligned} \tag{4.10}$$

Hence the theorem is proven. □

We can see that Eq. (4.10) contains the joint distribution of counterfactual statements $P(w^+_{\mathbf{pa}^+_W}, w^-_{\mathbf{pa}^-_W})$ which is unidentifiable from the observational data, making $P(e^+ \mid do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i}))$ and hence the π_i-specific effect $PSE_{\pi_i}(c^+, c^-)$ unidentifiable.

Next, we show how to bound $P(e^+ \mid do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i}))$ by scaling up and down certain terms in Eq. (4.10) and then eliminating $P(w^+_{\mathbf{pa}^+_W}, w^-_{\mathbf{pa}^-_W})$ using Eq. (4.6). For ease of representation, we further divide $\mathbf{A}$ into two disjoint subsets: (1) the set of nodes that are involved in the "kite" pattern, i.e., it is contained in a path in π_i that also contains any node in $\mathbf{W}$, denoted by $\mathbf{A}_1$; (2) the complementary set, i.e., those not involved in the "kite" pattern, denoted by $\mathbf{A}_2$. Then, we give the upper and lower bounds of $P(e^+ \mid do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i}))$ as shown in Theorem 6.

Theorem 6. *The upper bound of* $P(e^+ \mid do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i}))$ *is given by*

$$\sum_{\mathbf{a}_2,\mathbf{b},\mathbf{w}^-} \max_{\mathbf{a}_1,\mathbf{w}^+} \{P(e^+|c^-,\mathbf{q})\} \prod_{A\in\mathbf{A}_2} P(a|\mathbf{pa}_A^+) \prod_{B\in\mathbf{B}} P(b|\mathbf{pa}_B^-) \prod_{W\in\mathbf{W}} P(w^-|\mathbf{pa}_W^-), \tag{4.11}$$

and the lower bound of $P(e^+ \mid do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i}))$ *is given by*

$$\sum_{\mathbf{a}_2,\mathbf{b},\mathbf{w}^-} \min_{\mathbf{a}_1,\mathbf{w}^+} \{P(e^+|c^-,\mathbf{q})\} \prod_{A\in\mathbf{A}_2} P(a|\mathbf{pa}_A^+) \prod_{B\in\mathbf{B}} P(b|\mathbf{pa}_B^-) \prod_{W\in\mathbf{W}} P(w^-|\mathbf{pa}_W^-). \tag{4.12}$$

Proof. It is straightforward that

$$P(e^+|c^-,\mathbf{q}) \leq \max_{\mathbf{a}_1,\mathbf{w}^+} \{P(e^+|c^-,\mathbf{q})\}.$$

Thus, from Eq. (4.9) we have

$$\begin{aligned} P(e^+ \mid do(c^+|_{\pi_i}, c^-|_{\bar{\pi}_i})) \leq & \sum_{\mathbf{a}_1,\mathbf{a}_2,\mathbf{b},\mathbf{w}^+,\mathbf{w}^-} \max_{\mathbf{a}_1,\mathbf{w}^+} \{P(e^+|c^-,\mathbf{q})\} \prod_{A\in\mathbf{A}_1} P(a|\mathbf{pa}_A^+) \\ & \prod_{A\in\mathbf{A}_2} P(a|\mathbf{pa}_A^+) \prod_{B\in\mathbf{B}} P(b|\mathbf{pa}_B^-) \prod_{W\in\mathbf{W}} P(w^+_{\mathbf{pa}_W^+}, w^-_{\mathbf{pa}_W^-}). \end{aligned}$$

We can identify three properties for any node $A \in \mathbf{A}_1$: (1) A cannot be the parent of any node A' in $\mathbf{A}_2$. If not so, we have a path that contains C, A, A', E and any node $W \in \mathbf{W}$. This path must belong to π_i, otherwise A is contained in both a path in π_i and a path not in π_1, making A a witness node. Thus, A' is also involved in the "kite" pattern. (2) A cannot be the parent of any node in $\mathbf{B}$. Otherwise, A belongs to a path in π_i and also a path not in π_i, making A a witness node. (3) A cannot be the parent of any node in $\mathbf{W}$, otherwise A also becomes a witness node. Based on the three properties, the RHS of above inequality

equals to

$$\sum_{\mathbf{a}_2,\mathbf{b},\mathbf{w}^+,\mathbf{w}^-} \left[\begin{array}{c} \max_{\mathbf{a}_1,\mathbf{w}^+}\{P(e^+|c^-,\mathbf{q})\} \prod_{A\in\mathbf{A}_2} P(a|\mathbf{pa}_A^+) \prod_{B\in\mathbf{B}} P(b|\mathbf{pa}_B^-) \\ \prod_{W\in\mathbf{W}} P(w^+_{\mathbf{pa}_W^+}, w^-_{\mathbf{pa}_W^-}) \sum_{\mathbf{a}_1} \prod_{A\in\mathbf{A}_1} P(a|\mathbf{pa}_A^+) \end{array} \right]$$
$$= \sum_{\mathbf{a}_2,\mathbf{b},\mathbf{w}^+,\mathbf{w}^-} \max_{\mathbf{a}_1,\mathbf{w}^+}\{P(e^+|c^-,\mathbf{q})\} \prod_{A\in\mathbf{A}_2} P(a|\mathbf{pa}_A^+) \prod_{B\in\mathbf{B}} P(b|\mathbf{pa}_B^-) \prod_{W\in\mathbf{W}} P(w^+_{\mathbf{pa}_W^+}, w^-_{\mathbf{pa}_W^-}).$$

Then, we can similarly identify two properties for any node $W \in \mathbf{W}$ and its realization w^+: (1) w^+ cannot be involved in $\mathbf{pa}_A^+$ for any $A \in \mathbf{A}_2$, otherwise there exists a path in π_i that contains W, A, making A be involved in the "kite" pattern; (2) w^+ cannot be involved in $\mathbf{pa}_B^-$ for any $B \in \mathbf{B}$, which is by the definition of $\mathbf{B}$. Thus, the above expression further becomes

$$\sum_{\mathbf{a}_2,\mathbf{b},\mathbf{w}^-} \max_{\mathbf{a}_1,\mathbf{w}^+}\{P(e^+|c^-,\mathbf{q})\} \prod_{A\in\mathbf{A}_2} P(a|\mathbf{pa}_A^+) \prod_{B\in\mathbf{B}} P(b|\mathbf{pa}_B^-) \sum_{\mathbf{w}^+} \prod_{W\in\mathbf{W}} P(w^+_{\mathbf{pa}_W^+}, w^-_{\mathbf{pa}_W^-})$$
$$= \sum_{\mathbf{a}_2,\mathbf{b},\mathbf{w}^-} \max_{\mathbf{a}_1,\mathbf{w}^+}\{P(e^+|c^-,\mathbf{q})\} \prod_{A\in\mathbf{A}_2} P(a|\mathbf{pa}_A^+) \prod_{B\in\mathbf{B}} P(b|\mathbf{pa}_B^-) \prod_{W\in\mathbf{W}} P(w^-|\mathbf{pa}_W^-).$$

By using $P(e^+|c^-,\mathbf{q}) \geq \min_{\mathbf{a}_1,\mathbf{w}^+}\{P(e^+|c^-,\mathbf{q})\}$, similarly we can prove the lower bound. □

From Theorem 6 we can directly obtain the upper bound $ub(PSE_{\pi_i}(c^+,c^-))$ and lower bound $lb(PSE_{\pi_i}(c^+,c^-))$ of $PSE_{\pi_i}(c^+,c^-)$.

4.5.3 Algorithms for Unidentifiable Situation

Based on the derived bounds of the indirect discrimination, we can refine the proposed discovery algorithm *PSE-DD* to better deal with the unidentifiable situation, as shown in *PSE-DD** (Algorithm 5). On the other hand, we can also refine the proposed removal algorithm *PSE-DR* by replacing $SE_{\pi_i}(c^+,c^-)$ and $SE_{\pi_i}(c^-,c^+)$ in the constraints of the quadratic programming with $ub(SE_{\pi_i}(c^+,c^-))$ and $ub(SE_{\pi_i}(c^-,c^+))$. We refer to this new quadratic programming as the adjusted quadratic programming problem. The refined removal

Algorithm *PSE-DR** is shown in Algorithm 6.

Algorithm 5: *PSE-DD**

Input : The historical dataset $\mathcal{D}$, the sensitive attribute C, the decision attribute E, redlining attributes $\mathbf{R}$, a threshold τ.

Output : Direct/indirect discrimination $judge_d$, $judge_i$.

$\mathcal{G} = buildCausalNetwork(\mathcal{D})$;
$judge_d = judge_i = false$;
Compute $SE_{\pi_d}(\cdot)$ according to Eq. (4.1);
if $SE_{\pi_d}(c^+, c^-) > \tau \parallel SE_{\pi_d}(c^-, c^+) > \tau$ **then**
  $judge_d = true$;
Call subroutine $[\mathbf{S}_{\pi_i}, \bar{\mathbf{S}}_{\pi_i}] = DivideChildren(\mathcal{G}, C, E, \mathbf{R})$;
if $\mathbf{S}_{\pi_i} \cap \bar{\mathbf{S}}_{\pi_i} \neq \emptyset$ **then**
  Compute $ub(SE_{\pi_i}(c^+, c^-))$, $lb(SE_{\pi_i}(c^+, c^-))$, $lb(SE_{\pi_i}(c^-, c^+))$, $ub(SE_{\pi_i}(c^-, c^+))$ according to Eq. (4.11), (4.12);
  if $ub(SE_{\pi_i}(c^+, c^-)) \leq \tau$ & $ub(SE_{\pi_i}(c^-, c^+)) \leq \tau$ **then**
    $judge_i = false$;
  else if $lb(SE_{\pi_i}(c^+, c^-)) > \tau \parallel lb(SE_{\pi_i}(c^-, c^+)) > \tau$ **then**
    $judge_i = true$;
  else
    $judge_i = unknown$;
  return $[judge_d, judge_i]$;
Compute $SE_{\pi_i}(\cdot)$ according to Eq. (4.2);
if $SE_{\pi_i}(c^+, c^-) > \tau \parallel SE_{\pi_i}(c^-, c^+) > \tau$ **then**
  $judge_i = true$;
return $[judge_d, judge_i]$;

The following proposition shows that, the adjusted quadratic programming will at least produce an equivalently good solution as the quadratic programming after performing subroutine *GraphPreprocess*. This implies that *PSE-DR** performs at least as good as *PSE-DR* in term of the data utility preserving. Our experiments in Section 4.6 show that *PSE-DR** outperforms *PSE-DR* in the practical situations.

Proposition 4. *The modified CPT of E obtained from the quadratic programming after performing* GraphPreprocess *is a feasible solution of the adjusted quadratic programming problem.*

Proof. Firstly consider algorithm *PSE-DR*. Denote by $\mathcal{G}'$ the causal graph obtained after the

Algorithm 6: *PSE-DR**

Input : The historical dataset $\mathcal{D}$, the sensitive attribute C, the decision attribute E, redlining attributes $\mathbf{R}$, a threshold τ.

Output : Modified dataset $\mathcal{D}^*$.

```
[judge_d, judge_i] = PSE-DD*(D, C, E, R, τ);
if [judge_d, judge_i] == [false, false] then
    return D;
G = buildCausalNetwork(D);
if judge_i == unkonwn then
    Obtain the modified CPT of E by solving the adjusted quadratic
     programming problem;
else
    Obtain the modified CPT of E by solving the original quadratic programming
     problem;
Calculate P*(v) using the modified CPTs and generate D*;
return D*;
```

GraphPreprocess subroutine, denote by $\mathbf{Q}^*$ ($\mathbf{Q}^* \subseteq \mathbf{Q}$) the parents of E in $\mathcal{G}'$, and denote by $P^*(e|c, \mathbf{q}^*)$ the modified CPT of E obtained by solving the quadratic programming problem. Note that in $\mathcal{G}'$, based on the local Markov condition, $P^*(e|c, \mathbf{q}^*) = P^*(e|c, \mathbf{q})$ for all $\mathbf{q}$ that $\mathbf{q}^* \subseteq \mathbf{q}$. According to the constraints in the quadratic programming, the indirect discrimination based on the modified CPT of E is bounded by τ.

Now consider the original causal graph $\mathcal{G}$ with E's CPT $P^*(e|c, \mathbf{q}) = P^*(e|c, \mathbf{q}^*)$ for all $\mathbf{q}$ that $\mathbf{q}^* \subseteq \mathbf{q}$. We can see that causal graph $\mathcal{G}$ is actually equivalent to causal graph $\mathcal{G}'$, hence the indirect discrimination measured should also be the same[3]. In the following, we show that the indirect discrimination measured in $\mathcal{G}$ based on $P^*(e|c, \mathbf{q})$ equals to its upper bound given in Theorem 6, which means that $P^*(e|c, \mathbf{q})$ satisfies the constraints of the adjusted quadratic programming, and hence is a feasible solution of the adjusted quadratic programming problem.

[3]In fact, it can be easily shown that the indirect discrimination measured in $\mathcal{G}'$ based on Eq. (4.2) is equivalent to the indirect discrimination measured in $\mathcal{G}$ based on Eq. (4.10).

As shown in Theorem 5, the first term in Eq. (4.9) is given by

$$\sum_{\mathbf{a},\mathbf{b},\mathbf{w}^+,\mathbf{w}^-} \left(P^*(e^+|c^-,\mathbf{q}) \prod_{A\in\mathbf{A}} P(a|\mathbf{pa}_A^+) \prod_{B\in\mathbf{B}} P(b|\mathbf{pa}_B^-) \prod_{W\in\mathbf{W}} P(w^+_{\mathbf{pa}_W^+}, w^-_{\mathbf{pa}_W^-}) \right)$$
$$= \sum_{\mathbf{a},\mathbf{b},\mathbf{w}^+,\mathbf{w}^-} \left(P^*(e^+|c^-,\mathbf{q}^*) \prod_{A\in\mathbf{A}} P(a|\mathbf{pa}_A^+) \prod_{B\in\mathbf{B}} P(b|\mathbf{pa}_B^-) \prod_{W\in\mathbf{W}} P(w^+_{\mathbf{pa}_W^+}, w^-_{\mathbf{pa}_W^-}) \right).$$

Similar to Theorem 6, set $\mathbf{A}$ can be divided into two subsets $\mathbf{A}_1$ and $\mathbf{A}_2$. In addition to the properties shown in the proof of Theorem 6, we further identify two properties that appear after executing *GraphPreprocess*: (1) any node $A \in \mathbf{A}_1$ cannot belong to $\mathbf{Q}^*$, otherwise the "kite" pattern still exists, contradicting to that *GraphPreprocess* removes the "kite" pattern; (2) for similar reason w^+ of any $W \in \mathbf{W}$ cannot be involved in $\mathbf{q}^*$. Thus, the above expression becomes

$$\sum_{\mathbf{a}_2,\mathbf{b},\mathbf{w}^-} P^*(e^+|c^-,\mathbf{q}^*) \prod_{A\in\mathbf{A}_2} P(a|\mathbf{pa}_A^+) \prod_{B\in\mathbf{B}} P(b|\mathbf{pa}_B^-) \prod_{W\in\mathbf{W}} P(w^-|\mathbf{pa}_W^-). \tag{4.13}$$

Now back to the upper bound. Consider the first term of Eq. (4.11), which is given by

$$\sum_{\mathbf{a}_2,\mathbf{b},\mathbf{w}^-} \max_{\mathbf{a}_1,\mathbf{w}^+} \{P(e^+|c^-,\mathbf{q}^*)\} \prod_{A\in\mathbf{A}_2} P(a|\mathbf{pa}_A^+) \prod_{B\in\mathbf{B}} P(b|\mathbf{pa}_B^-) \prod_{W\in\mathbf{W}} P(w^-|\mathbf{pa}_W^-). \tag{4.14}$$

As stated, $\mathbf{a}_1$ and $\mathbf{w}^+$ cannot be involved in $\mathbf{q}^*$. Thus, the maximization operation on $P(e^+|c^-,\mathbf{q}^*)$ has no effect, making Eq. (4.13) and (4.14) equivalent. Hence, the the proposition is proved. □

4.6 Experiments

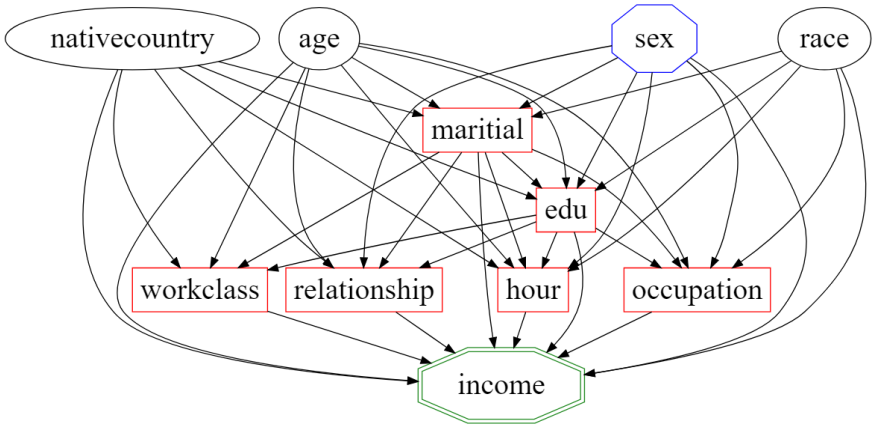
In this section, we conduct experiments using two real datasets: the *Adult* dataset [68] and the *Dutch Census of 2001* [18]. The description of the two datasets can be founded in Appendix A.1. We evaluate our discovery and removal algorithms under both identifiable and unidentifiable situations. For comparison, we involve the local massaging (*LMSG*) and

local preferential sampling (LPS) algorithms proposed in [18] and disparate impact removal algorithm (DI) proposed in [37, 69]. The causal graphs are constructed and presented by utilizing Tetrad [70]. We employ the original PC algorithm [57] and set the significance threshold 0.01 for conditional independence testing in causal graph construction. The quadratic programming is solved using CVXOPT [71]. By default, the discrimination threshold τ is set as 0.05. The preprocessed data and algorithm implementations are available at http://tiny.cc/pse-fairness.

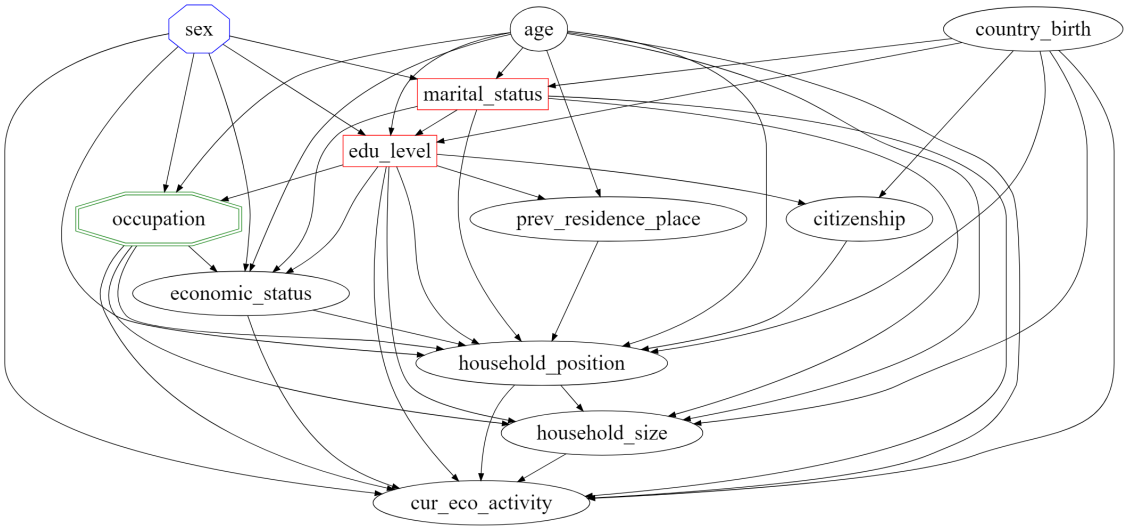
4.6.1 Discrimination Discovery

For the *Adult* dataset, due to the sparse data issue and the convention in collecting features by social-platforms [72], we binarize each attribute's domain values into two classes to reduce the domain sizes. We use three tiers in the partial order for temporal priority: `sex`, `age`, `native_country`, `race` are defined in the first tier, `edu_level` and `marital_status` are defined in the second tier, and all other attributes are defined in the third tier. The constructed causal graph is shown in Figure 4.5a. We treat `sex` as the sensitive attribute, `income` as the decision, and `marital_status` as the redlining attribute. Then set π_d contains the edge pointing from `sex` to `income`, and set π_i contains all the causal paths from `sex` to `income` that pass through `marital_status`. As can be seen, the π_i-specific effect does not satisfy the recanting witness criterion. By computing the path-specific effects, we obtain that $SE_{\pi_d}(c^+, c^-) = 0.025$ and $SE_{\pi_i}(c^+, c^-) = 0.175$. By setting $\tau = 0.05$, the results indicate no direct discrimination but significant indirect discrimination against females according to our criterion. In [18], it has been shown that each of the attributes `relationship`, `age` and `working_hours` can explain some of the discrimination. However, no conclusion regarding direct/indirect discrimination is drawn.

For the *Dutch Census of 2001* dataset, similarly, we binarize the domain values of attribute `age` due to its large domain size. Three tiers are used in the partial order for temporal priority: `sex`, `age`, `country_birth` are in the first tire, `edu` is in the second tire, and



(a) The *Adult* dataset.



(b) The *Dutch census of 2001* dataset.

Figure 4.5: The constructed causal graphs: the blue octagon node represents the sensitive attribute, the green double-octagon node represents the decision, and the red rectangle nodes represent represent the (potential) redlining attributes.

all other attributes are in the third tire. The constructed causal graph is shown in Figure 4.5b. We treat `sex` as the sensitive attribute, `occupation` as the decision, and `marital_status` as the redlining attribute. In this case, the recanting witness criterion is also not satisfied. For this dataset, we obtain $SE_{\pi_d}(c^+, c^-) = 0.220$ and $SE_{\pi_i}(c^+, c^-) = 0.001$, indicating significant direct discrimination but no indirect discrimination against females.

Table 4.1: Discrimination in the modified data ($\tau = 0.05$), and comparison of utility with varied τ values for the *Adult* dataset.

	Remove Algorithm				τ			
	PSE-DR	*DI*	*LMSG*	*LPS*	0.025	0.05	0.075	0.1
Direct	0.013	0.001	-0.142	-0.142	0.008	0.012	0.019	0.024
Indirect	0.049	0.050	0.288	0.174	0.024	0.049	0.074	0.100
$\chi^2(\times 10^4)$	1.038	4.964	1.924	1.292	1.247	1.038	1.029	0.819

4.6.2 Discrimination Removal

We run the removal algorithm *PSE-DR* to remove discrimination from both datasets, and then run the discovery algorithm *PSE-DD* to further examine whether discrimination is truly removed in the modified dataset. For comparison, we include removal algorithms from previous works: *LMSG*, *LPS* and *DI*. The discriminatory effects of the modified dataset are shown in Table 4.1 (left) for the *Adult* dataset, and in Table 4.2 (left) for the *Dutch Census of 2001* dataset. As can be seen, our method *PSE-DR* completely removes direct and indirect discrimination from both datasets. In addition, *PSE-DR* produces relatively small data utility loss in term of χ^2. For *LMSG* and *LPS*, indirect discrimination is not removed from the *Adult* dataset, and in both datasets direct discrimination seems to be over removed. The *DI* algorithm provides a parameter λ to indicate the amount of discrimination to be removed, where $\lambda = 0$ represents no modification and $\lambda = 1$ represents full discrimination removal. However, λ has no direct connection with the threshold τ. In our experiments, we execute *DI* multiple times with different λ values and report the one that is closest to achieve $\tau = 0.05$. Although *DI* indeed removes direct and indirect discrimination, its data utility is far more worse than *PSE-DR*, implying that it removes many information unrelated to discrimination.

We then examine how the data utility in term of χ^2 varies with different thresholds τ for *PSE-DR*. We change the value of τ from 0.025 to 0.1. From Tables 4.1 and 4.2 (right) we can see that less utility loss is incurred when larger τ value is used. This observation is consistent with our analysis since the larger the value of τ, the more relaxed the constraints

Table 4.2: Discrimination in the modified data ($\tau = 0.05$), and comparison of utility with varied τ values for the *Dutch Census of 2001* dataset.

	Remove Algorithm				τ			
	PSE-DR	*DI*	*LMSG*	*LPS*	0.025	0.05	0.075	0.1
Direct	0.049	0.000	-0.081	-0.100	0.022	0.049	0.073	0.099
Indirect	0.001	-0.001	0.001	0.001	0.001	0.001	0.001	0.001
$\chi^2(\times 10^4)$	1.104	4.604	4.084	1.742	1.279	1.104	1.099	0.934

in *PSE-DR*.

We also examine whether the predictive models built from the data modified by *PSE-DR* incur discrimination in decision making. We divide the original dataset into the training and testing datasets, and remove discrimination from the training dataset to obtain the modified training dataset. Then, we build the predictive models from the modified training dataset, and use them to make predictive decisions over the testing data. Four classifiers, logistic regression (*LR*), decision tree (*DT*), random forest (*RF*) and *SVM*, are used for prediction with five-fold cross-validation. Finally, we run *PSE-DD* to examine whether the predictions for the testing data contain discrimination. The prediction accuracy using both original and modified training dataset are reported as well. The results are shown in Tables 4.3 and 4.4. As can be seen, for the *Adult* dataset, the predictions of all classifiers do not incur direct or indirect discrimination, with the accuracy only slightly decreased. However, for the *Dutch Census of 2001* dataset, the predictions contain direct discrimination, which is smaller than that in the original data yet significant. Some recent works imply that, even if discrimination is removed from the training data, it can still appear in the predictions of classifiers [12, 73]. How to ensure non-discrimination in the prediction is a future direction of our work.

4.6.3 Unidentifiable Situation

In this subsection, we examine the proposed methods for handling the unidentifiable situation when measuring and removing the indirect discrimination. We consider each of

Table 4.3: Discrimination in prediction for the *Adult* dataset.

		LR	*DT*	*RF*	*SVM*
Direct		0.045	0.023	0.022	0.023
Indirect		0.047	0.042	0.050	0.041
Accuracy(%)	Original	81.70	81.77	81.81	81.78
	Modified	81.30	80.55	80.56	80.54

Table 4.4: Discrimination in prediction for the *Dutch Census of 2001* dataset.

		LR	*DT*	*RF*	*SVM*
Direct		0.059	0.103	0.098	0.099
Indirect		0.001	0.001	0.001	0.001
Accuracy(%)	Original	83.45	82.46	83.12	83.70
	Modified	81.93	81.36	81.57	82.10

attribute other than `marital_status` that is on the causal paths from the sensitive attribute to the decision as the redlining attribute and see whether the recanting witness criterion is satisfied, i.e., π_i forms the "kite" pattern. For the *Adult* dataset, these attributes include `edu_level`, `occupation`, `hours_per_week`, `workclass` and `relationship`, each of which creates the "kite" pattern if it is treated as the redlining attribute. For the *Dutch Census of 2001* dataset, only `edu_level` is on the causal paths from the sensitive attribute to the decision, and treating it as the redlining attribute will not create the "kite" pattern. Thus, the remaining of this subsection focus on the *Adult* dataset.

Upon selecting the redlining attribute, we execute algorithm *PSE-DD** to compute the π_d-specific effect $SE_{\pi_d}(c^+, c^-)$ as well as the upper and lower bounds of the π_i-specific effect $ub(SE_{\pi_i}(c^+, c^-))$ and $lb(SE_{\pi_i}(c^+, c^-))$. The results are shown in Table 4.5. As can be seen, for all attributes the π_d-specific effect is the same. This is reasonable since treating different

Table 4.5: Discrimination measured and bounded under the unidentifiable situation for the *Adult* dataset.

		`edu`	`occupation`	`hours`	`workclass`	`relationship`
Direct				0.025		
Indirect	*lb*	-0.114	-0.069	-0.027	-0.014	-0.086
	ub	0.361	0.039	0.072	0.016	0.015

attribute as the redlining attribute should not affect the direct discrimination. On the other hand, the upper and lower bounds imply that we can ensure no indirect discrimination if either `occupation`, `workclass` or `relationship` onsidered as the redlining attribute, and we are uncertain about indirect discrimination if either treating `edu_level` or `hours_per_week` as the redlining attribute.

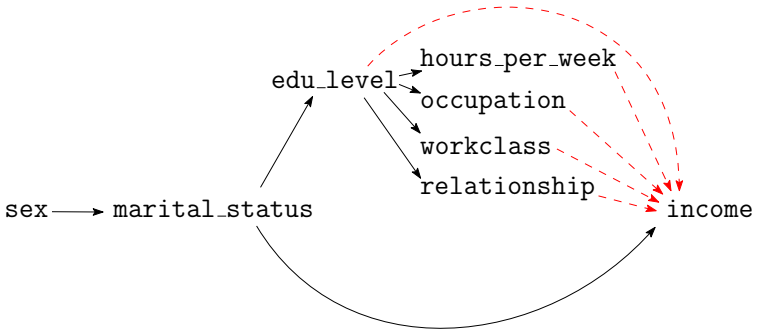


Figure 4.6: The "kite" pattern when treating `edu_leve` as redlining. Red dashed edges are to be deleted by *GraphPreprocess*.

We use `edu_level` as an example to show the results of discrimination removal. The subgraph shown in Figure 4.6 presents the "kite" pattern formed when treating `edu_leve` as the redlining attribute. The π_i-specific effect satisfies the recanting witness criterion with `marital_status` as the witness. We evaluate the two removal algorithms: *PSD-DR* and *PSD-DR**. For *PSD-DR*, subroutine *GraphPreprocess* needs to cut off all causal paths passing through the redlining attribute in order to remove the "kite" pattern, which means that it should delete all the edges highlighted by the red dashed edges. The discrimination in the modified data is shown in Table 4.6. As can be seen, both algorithms guarantee no direct discrimination as well as no indirect discrimination based on its upper bound. However, the utility of the modified data produced by *PSE-DR** is better than that produced *PSE-DR*, which is consistent with our theoretical result. A more straightforward explanation for this example can be that, since all the causal paths in π_i are involved in the "kite" pattern, *GraphPreprocess* must cut off all these paths, resulting a total elimination of all indirect discriminatory effect. However, *PSE-DR** can utilize the threshold $\tau = 0.05$, achieving a better balance between non-discrimination and utility preserving.

Table 4.6: Discrimination in the modified data when treating `edu_level` as redlining.

	PSE-DR	*PSE-DR**
Direct	0.038	0.033
Indirect (ub)	0	0.050
$\chi^2(\times 10^4)$	1.499	1.106

4.7 Summary

In this chapter, we studied the problem of discovering both direct/indirect discrimination from historical data, and removing them before performing predictive analysis. We made use of the causal graph to capture the causal structure of the data, and modeled direct and indirect discrimination as different path-specific effects. Based on that, we proposed the discovery algorithm *PSE-DD* to discover both direct and indirect discrimination, and the removal algorithm *PSE-DR* to remove them. For the situation where indirect discrimination cannot be exactly measured due to the unidentifiability of the path-specific effects, we derived the upper and lower bounds for the unidentifiable indirect discrimination, and developed the refined discovery algorithm *PSE-DD** and removal algorithm *PSE-DR**. The experiments using the real dataset showed that, our approach can ensure that the modified data dose not contain any type of discrimination while incurring small utility loss. Under the unidentifiable situation, the refined algorithms *PSE-DR** produces smaller utility loss than *PSE-DR* that directly deletes edges to remove the unidentifiability. The early versions of this work have been published in IJCAI 2017 [4] and TKDE 2019 [5].

5 Discrimination Discovery and Removal from Ranked Data

5.1 Introduction

Fairness in the classification models has been widely studied by the research community [19,41,44,45,52,74,75]. In this chapter, we investigate discrimination in ranking models, which are another widely used machine learning models adopted by search engines, recommendation systems, and auction systems, etc. To be more specific, we study the discrimination discovery and removal from the ranked data. A ranked dataset is a combination of the candidate profiles with the permutation of the candidates as the decision. Fairness concerns are raised for the ranking models since biases and discrimination can also be introduced into the ranking.

ID	u1	u2	u3	u4	u5	u6	u7	u8	u9	u10
Race	1	1	1	1	1	0	0	0	0	0
Zip Code	1	1	1	1	1	1	0	0	1	0
Interview	1	2	2	4	2	5	4	4	3	2
Edu	1	2	1	2	4	5	4	5	3	5

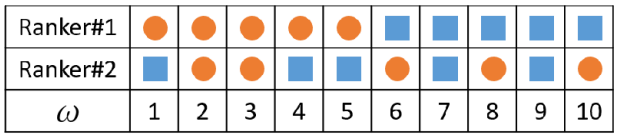


Figure 5.1: A toy example of dataset and ranking results produced by two rankers. Blue squares represent the favorable group and red circles represent unfavorable group.

Existing methods [76, 77] for studying the discrimination discovery and removal from ranked data are mainly based on statistical parity, which means that the demographics of individuals in any prefix of the ranking are identical to the demographics of the whole population. However, it has already been shown in classification that statistical parity does not take into account the fact that part of discrimination is explainable by some non-sensitive attributes and hence cannot accurately measure discrimination [46]. We believe that this observation also holds in the ranked data. Let us consider a toy example of ranked data for a company recruiting system shown in Figure 5.1. The data contains four profile attributes: `race` (C), `zip code` (Z), `education` (E), `interview result` (I), where `race` is the sensitive

attribute with a favorable group ($C = 1$) and an unfavorable group ($C = 0$), and `education` and `interview result` are the objective requirements of getting the job. Assume that there are two rankers, both of which compute the qualification scores to produce the rankings. The first ranker, denoted by Ranker#1, produces qualification scores as an equal-weighted linear combination of two attributes `education` and `interview result`. Intuitively, Ranker#1 produces a fair ranking since it purely depends on two objective attributes. However, as can be seen, the ranking results do not satisfy statistical parity. On the basis of Ranker#1, the second ranker, Ranker#2, further gives a bonus score of 2 for the favorable group (i.e., $C = 1$). The usage of the sensitive attribute explicitly results in the unequal treatment to the unfavorable group ($C = 0$). Nevertheless, the ranking results satisfy statistical parity as two race groups are well-mixed in equal proportion. This example shows that statistical parity may produce misleading conclusions regarding discrimination.

To address the limitation of the statistical parity-based methods, the causal graph-based discrimination detection and removal methods have been recently proposed by Zhang et al. [4]. It shows that the correlation between the sensitive attribute and the decision is a nonlinear combination of the direct discrimination, the indirect discrimination, as well as the explainable effect. The path-specific effect technique has been used to capture the causal effects passing through different paths. However, this work focuses on binary classification. In ranking systems, the decisions are given in term of a permutation of a series of unique, concatenating integers which cannot be treated as regular random variables. This means that causal graphs cannot be built in traditional ways. Thus, the methods in [4] cannot be applied directly to deal with ranked data.

In this chapter, we employ the causal graph to solve the fair ranking problem by adopting a continuous variable called `score` instead of the ranking positions to represent the qualifications of individuals in the rank. We use the Bradley-Terry model [78] to obtain a reasonable mapping from ranking positions to scores. We then construct the causal graph from the individuals' profiles and scores, a mix of categorical and continuous data. Traditional

causal graph construction and inference are limited to the single data-type situations where the variables are all discrete (e.g., causal Bayesian networks) or all continuous (e.g., linear Gaussian models). To address this challenge, we associate the score with a set of Conditional Gaussian (CG) distributions instead of the Conditional Probability Table (CPT). Then, we extend the path-specific effect technique to our mixed-variable causal graph for capturing direct and indirect discrimination. We also derive a relationship between the path-specific effects for the ranked data and those for the binary decision, assuming that binary decision is obtained based on certain cut-off point imposed on the ranking. Algorithms for detecting discrimination in the causal graph, as well as for removing rank biases from the data are developed. Finally, we conduct experiments using the real-world dataset to show the effectiveness of our methods. The results show that our methods can correctly detect and remove both direct and indirect discrimination with relatively small data utility loss, while the statistical parity based methods neither correctly identify discrimination nor successfully mitigate discrimination.

5.2 Related Work

In Chapter 2, we discuss the related work for fairness-aware machine learning, which mainly focus on classification. Fair ranking is an emerging topic in fairness-aware learning. Current works in fair ranking are mainly based on the statistical parity. In [76], it is required that a preset proportion of sensitive individuals that must be maintained in each prefix of the ranking for the rank to be fair. However, many existing works (e.g., [4]) have shown that statistical parity alone is insufficient as a general notion of fairness.

In the previous chapter, we develop a causality-based framework to capture direct and indirect discrimination in classification. Similarly, many researchers incorporate causal graphs into the fairness-aware machine learning. [4, 5, 26, 28, 29, 73, 79]. The limitation of these works is that they focus on the classification problems and cannot be applied directly to the fair ranking problem. This is because in their models, the decision of each individual is

treated as an independent random variable, but the ranking positions of different individuals are correlated. In this chapter we address above limitations and develop the causal-based fair ranking algorithms.

In the field of data science, it is well-studied in data mining how to model a ranking using a continuous score space [80]. Several models, such as the Plackett-Luce model [81, 82], the Mallows model [83] and the Bradley-Terry model [78], are widely used in this field. In this chapter, we adopt the Bradley-Terry model to characterize the ranked data and obtain the continuous scores from the ranks.

5.3 Preliminaries

Follow the notations presented in Chapter 3, an attribute is denoted by an uppercase letter, sets of attributes by a bold uppercase letter, a value by a lowercase letter, and a set of values of a set attributes by a bold lowercase letter. The domain space of an attribute is denoted by $\mathfrak{X}_X$. The domain space of an attribute set is a Cartesian product of the domain spaces of its elements, denoted by $\mathfrak{X}_{\mathbf{X}} = \prod_{X \in \mathbf{X}} \mathfrak{X}_X$.

In this chapter, we leverage causal inference techniques, e.g., the Structural Causal Model, the total causal effect, and the path-specific effect. The introduction to essential causal inference techniques can be found at Chapter 3.

A Bradley-Terry model $\mathcal{M}$ assigns each individual i a score s_i ($s_i \in \mathbb{R}$) to indicate the qualification preference of individual. Generally, a larger score represents a better qualification. The difference between the scores of two individuals i, j corresponds to the log-odds of the probability $p_{i,j}$ that individual i is ranked before individual j in the rank, i.e.,

$$s_i - s_j = \log \frac{p_{ij}}{1 - p_{ij}}.$$

Equivalently, solving for p_{ij} yields

$$p_{ij} = \frac{e^{s_i}}{e^{s_i} + e^{s_j}}.$$

On the other hand, the probability of any rank permutation ω given a Bradley-Terry model $\mathcal{M}$ is proportional to the product of the probability $p_{i,j}$ of all preference pairs subject to ω, i.e.,

$$P(\omega|\mathcal{M}) \propto \prod_{(i,j):\omega_i<\omega_j} p_{ij},$$

where ω_i, ω_j are the ranking positions of individuals i, j. Thus, the logarithm likelihood of the Bradley-Terry model $\mathcal{M}$ given the observed rank permutation ω is given by $\mathcal{L}(\mathcal{M}|\omega) = -\log P(\omega|\mathcal{M})$. As a result, the optimal Bradley-Terry model that best fits the observed rank permutation ω can be obtained by minimizing $\mathcal{L}(\mathcal{M}|\omega)$ as the loss function. Wu et al. [84] proved that the loss function is convex and could be efficiently optimized with gradient descent.

5.4 Modeling Direct and Indirect Discrimination in Ranked Data

In this section, we study how to model direct and indirect discrimination in a ranked dataset as the causal effect. We consider a ranked dataset $\mathcal{D}$ consisting of N individuals with a sensitive attribute C, several non-sensitive attributes $\mathbf{Z} = \{Z_1, \cdots, Z_j, \cdots\}$, and a rank permutation π as the decision. There is a subset of attributes $\mathbf{R} \subseteq \mathbf{Z}$ that may cause indirect discrimination, referred to as the *redlining attributes*. We assume all attributes are categorical. We further make two reasonable assumptions: 1) the sensitive attribute C has no parent; and 2) the score S has no child. The two assumptions are to make our theoretical results more concise and can be easily relaxed.

5.4.1 Building Causal Graph for Ranked Data

A rank permutation is a series of unique, concatenating integers that cannot be treated as normal categorical random variables. In data science, a number of models [80] are proposed to map the ranking positions in a ranked dataset to the continuous scores. In this chapter we use the Bradley-Terry model [78] but the logic also applies to other models. The comparison of the performance of different models is beyond the scope of the chapter and is left for future work.

After obtaining the score S using the Bradley-Terry model, we build a causal graph for variables C, $\mathbf{Z}$ and S. We first adopt the PC-algorithm for learning the structure of the causal graph. Since there exist both discrete and continuous variables, different conditional independence testing methods can be adopted, such as chi-square test for discrete variables, partial correlation matrix for continuous variables, and conditional Gaussian likelihood ratio test for mixed variables. Then, for parameterizing the causal graph, we treat discrete and continuous variables in different ways. For discrete variables C and $\mathbf{Z}$ (we can extend our method to the situation where some profile attributes are continuous), each of them is associated with a Conditional Probability Table (CPT). The conditional probabilities can be estimated from data using standard statistical estimation techniques (like the maximum likelihood estimation). For continuous score S, it is associated with the Conditional Gaussian (CG) distributions instead of the CPT. Let $\mathbf{Q} = Pa(S)\backslash\{C\}$. For each value assignment $c, \mathbf{q}$ of parents of S, there is a CG distribution whose mean and variance are based on $c, \mathbf{q}$. Thus, the CG distribution of S is given by

$$P(s|c, \mathbf{q}) = \mathcal{N}(\mu_{c,\mathbf{q}}, \sigma^2_{c,\mathbf{q}}).$$

Finally, we fit each CG distribution $\mathcal{N}(\mu_{c,\mathbf{q}}, \sigma^2_{c,\mathbf{q}})$ to the scores of all candidates with $C = c$ and $\mathbf{Q} = \mathbf{q}$ using standard statistical estimation techniques.

As an example, Figure 5.2 shows a causal graph of the toy example presented in

the Introduction. Each of C, Z, E, I is associated with a CPT representing the conditional probability given the parents, and S is associated with a set of CG distribution where the mean and the variance are based on its parents, the other four variables.

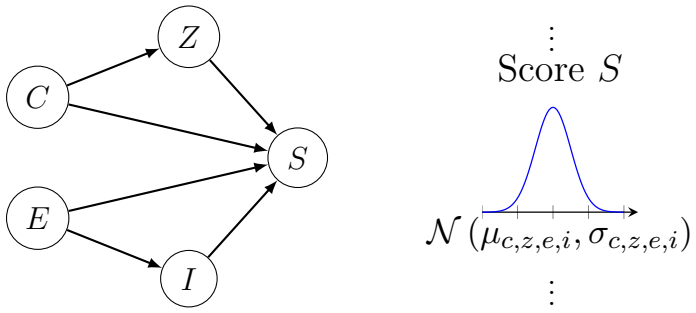


Figure 5.2: The causal graph of the toy example involving: `race` (C), `zip code` (Z), `education` (E), `interview result` (I), and `score` (S).

5.4.2 Quantitative Measurement

Now we show how direct and indirect discrimination in a ranked dataset can be quantitatively measured based on the causal graph we build. It is known that discrimination is a causal effect of the sensitive attribute on the decision. We first give the quantitative measure of the total causal effect of sensitive attribute C on score S as shown in Theorem 7.

Theorem 7. *The total causal effect is given by*

$$TE(c^+, c^-) = \sum_{\mathbf{q} \in \mathfrak{X}_{\mathbf{Q}}} \left(\mu_{c^+,\mathbf{q}} P(\mathbf{q}|c^+) - \mu_{c^-,\mathbf{q}} P(\mathbf{q}|c^-) \right) \tag{5.1}$$

Proof. According to Definition 3, total causal effect is given by

$$\begin{aligned} TE(c^+, c^-) &= \mathbb{E}\left[S|do(c^+)\right] - \mathbb{E}\left[S|do(c^-)\right] \\ &= \int s \cdot P(s|do(c^+))ds - \int s \cdot P(s|do(c^-))ds. \end{aligned}$$

According to Eq. (3.2), we have

$$\begin{aligned} P(s|do(c^+)) &= \sum_{\mathbf{z}\in\mathfrak{X}_\mathbf{Z}} P(s,\mathbf{z}|do(c^+)) \\ &= \sum_{\mathbf{z}\in\mathfrak{X}_\mathbf{Z}} P(s|c^+,\mathbf{q}) \prod_{Z_j\in\mathbf{Z}} P(z_j|Pa(Z_j))\delta_{C=c^+}. \end{aligned}$$

It can be shown that

$$\prod_{Z_j\in\mathbf{Z}} P(z_j|Pa(Z_j))\delta_{C=c^+} = P(\mathbf{z}|c^+). \tag{5.2}$$

In fact, if we sort all nodes in $\mathbf{Z}$ according to the topological ordering as $\{Z_1,\cdots,Z_j,\cdots\}$, we can see that all parents of each node Z_j are before it in the ordering. In addition, since C has no parent, it must be Z_j's non-descendant; since E has no child, it cannot be Z_j's parent. Thus, based on the local Markov condition, we have $P(z_j|Pa(Z_j)) = P(z_j|c^+,z_1,\cdots,z_{j-1})$. According to the chain rule we obtain $P(\mathbf{z}|c^+)$. Thus, it follows that

$$\begin{aligned} P(s|do(c^+)) &= \sum_{\mathbf{z}\in\mathfrak{X}_\mathbf{Z}} P(s|c^+,\mathbf{q})P(\mathbf{z}|c^+) \\ &= \sum_{\mathbf{q}\in\mathfrak{X}_\mathbf{Q}} P(s|c^+,\mathbf{q}) \sum_{\mathbf{Z}\backslash\mathbf{Q}} P(\mathbf{z}|c^+) \\ &= \sum_{\mathbf{q}\in\mathfrak{X}_\mathbf{Q}} P(s|c^+,\mathbf{q})P(\mathbf{q}|c^+). \end{aligned}$$

As a result, we have

$$\begin{aligned} \int s\cdot P(s|do(c^+))ds &= \int s\cdot \sum_{\mathbf{q}\in\mathfrak{X}_\mathbf{Q}} P(s|c^+,\mathbf{q})P(\mathbf{q}|c^+)ds \\ &= \sum_{\mathbf{q}\in\mathfrak{X}_\mathbf{Q}} P(\mathbf{q}|c^+)\int sP(s|c^+,\mathbf{q})ds \\ &= \sum_{\mathbf{q}\in\mathfrak{X}_\mathbf{Q}} \mu_{c^+,\mathbf{q}}P(\mathbf{q}|c^+). \end{aligned}$$

Hence, the theorem is proven. □

In [4], the authors show that in the single-type causal graph, total causal effect generally cannot correctly measure either direct discrimination or indirect discrimination, which should be modeled as the path-specific effects. By adopting similar strategy, we capture direct discrimination by the causal effect transmitted via the direct edge from C to S, and capture indirect discrimination by the causal effect transmitted via the paths that pass through redlining attributes. Formally, define π_d as the path set that contains only $C \to S$, and define π_i as the path set that contains all causal paths which are from C to S and pass through $\mathbf{R}$. Then, direct discrimination can be captured by the π_d-specific effect $PSE_{\pi_d}(\cdot)$, and indirect discrimination can be captured by the π_i-specific effect $PSE_{\pi_i}(\cdot)$. We extend the method in [4] for computing the path-specific effect from data to our mixed-variable causal graph for computing $PSE_{\pi_d}(\cdot)$ and $PSE_{\pi_i}(\cdot)$. The results are shown in Theorem 8.

Theorem 8. *The π_d-specific effect $PSE_{\pi_d}(c^+, c^-)$ is given by*

$$PSE_{\pi_d}(c^+, c^-) = \sum_{\mathbf{q} \in \mathfrak{X}_{\mathbf{Q}}} \left(\mu_{c^+,\mathbf{q}} - \mu_{c^-,\mathbf{q}}\right) P(\mathbf{q}|c^-), \tag{5.3}$$

The π_i-specific effect $PSE_{\pi_i}(c^+, c^-)$ is given by

$$\begin{aligned} PSE_{\pi_i}(c^+, c^-) = \sum_{\mathbf{z} \in \mathfrak{X}_{\mathbf{Z}}} \Big(\mu_{c^-,\mathbf{q}} \prod_{G \in \mathbf{V}_{\pi_i}} P(g|c^+, Pa(G)\backslash\{C\}) \prod_{H \in \bar{\mathbf{V}}_{\pi_i}} P(g|c^-, Pa(G)\backslash\{C\}) \\ \times \prod_{O \in \mathbf{Z}\backslash Ch(C)} P(o|Pa(O)) \Big) - \sum_{\mathbf{q} \in \mathfrak{X}_{\mathbf{Q}}} \left(\mu_{c^-,\mathbf{q}} P(\mathbf{q}|c^-) \right), \end{aligned} \tag{5.4}$$

where $\mathbf{V}_{\pi_i}$ and $\bar{\mathbf{V}}_{\pi_i}$ is obtained by dividing C's children except S based on the above method. Eq. (5.4) *can be simplified to*

$$PSE_{\pi_i}(c^+, c^-) = \sum_{\mathbf{q} \in \mathfrak{X}_{\mathbf{Q}}} \mu_{c^-,\mathbf{q}} \left(P(\mathbf{q}|c^+) - P(\mathbf{q}|c^-) \right) \tag{5.5}$$

if π_i contains all causal paths from C to S except the direct edge $C \to S$.

Proof. For the π_d-specific effect, according to Definition 4, we have

$$\begin{aligned} PSE_{\pi_d} &= \mathbb{E}\left[S|do(\mathbf{c}^+|_{\pi_d})\right] - \mathbb{E}\left[S|do(\mathbf{c}^-)\right] \\ &= \int s \cdot P(s|do(c^+|_{\pi_d}))ds - \int s \cdot P(s|do(c^-))ds. \end{aligned}$$

In the above equation, $P(s|do(c^-))$ can be computed according to the truncated factorization formula (3.2). To compute $P(s|do(c^+|_{\pi_d}))$, we follow the steps in [3]. First, express $P(s|do(c^+|_{\pi_d}))$ as the truncated factorization formula. Then, divide the children of C into two disjoint sets $\mathbf{V}_{\pi_d}$ and $\bar{\mathbf{V}}_{\pi_d}$. Let $\mathbf{V}_{\pi_d}$ contains C's each child V where edge $C \to V$ is a segment of a path in π_d; let $\bar{\mathbf{V}}_{\pi_d}$ contains C's each child V where either V is not included in any path from C to S, or edge $C \to V$ is a segment of a path not in π_d. Finally, replace values of C with c^+ for the terms corresponding to nodes in $\mathbf{V}_\pi$, and replace values of C with c^- for the terms corresponding to nodes in $\bar{\mathbf{V}}_{\pi_d}$.

Following the above procedure, we obtain

$$P(s|do(c^+|_{\pi_d})) = \sum_{\mathbf{z}\in\mathfrak{X}_\mathbf{Z}} P(s|c^+,\mathbf{q}) \prod_{Z_i\in\mathbf{Z}} P(z_i|Pa(Z_i))\delta_{C=c^-}.$$

By using Eq. (5.2), it follows that

$$P(s|do(c^+|_{\pi_d})) = \sum_{\mathbf{q}\in\mathfrak{X}_\mathbf{Q}} P(s|c^+,\mathbf{q})P(\mathbf{q}|c^-),$$

which leads to Eq. (5.3) in the theorem.

For the π_i-specific effect, following the above procedure similarly we can obtain

$$\begin{aligned} P(s|do(c^+|_{\pi_i})) = \sum_{\mathbf{z}\in\mathfrak{X}_\mathbf{Z}} \Big(& P(s|c^-,\mathbf{q}) \prod_{G\in\mathbf{V}_{\pi_i}} P(g|c^+,Qa(G)) \\ & \times \prod_{H\in\bar{\mathbf{V}}_{\pi_i}} P(g|c^-,Pa(G)\backslash\{C\}) \prod_{O\in\mathbf{Z}\backslash Ch(C)} P(o|Pa(O))\Big), \end{aligned}$$

which leads to Eq. (5.4). If π_i contains all causal paths from C to S except the direct edge,

it means that $\mathbf{V}_{\pi_i} = Ch(C)\backslash\{S\}$, and $\bar{\mathbf{V}}_{\pi_i} = \emptyset$. Thus, it follows that

$$\begin{aligned} P(s|do(c^+|_{\pi_i})) &= \sum_{\mathbf{z}\in\mathfrak{X}_{\mathbf{Z}}} P(s|c^-,\mathbf{q}) \prod_{Z\in\mathbf{Z}} P(z|Pa(Z)\backslash\{C\})\delta_{C=c^+} \\ &= \sum_{\mathbf{q}\in\mathfrak{X}_{\mathbf{Q}}} P(s|c^-,\mathbf{q})P(\mathbf{q}|c^+), \end{aligned}$$

which leads to Eq. (5.5). Hence, the theorem is proven. □

Theorems 7 and 8 present the quantitative measurement of the total causal effect as well as the π_d and π_i-specific effects. The following proposition reveals the relationship among $TE(\cdot)$, $PSE_{\pi_d}(\cdot)$ and $PSE_{\pi_i}(\cdot)$. It shows that the indirect (discriminatory) effect is equal to the total causal effect plus the "reversed" direct (discriminatory) effect.

Proposition 5. *If π_i contains all causal paths from C to S except the direct edge $C \to S$, we have*

$$PSE_{\pi_i}(c^+, c^-) = TE(c^+, c^-) + PSE_{\pi_d}(c^-, c^+).$$

Proof. The proof can be directly obtained from Eq. (5.1) and (5.5). □

5.4.3 Relationship between Ranking and Binary Decision

In the earlier work [4], we have derived the π_d and π_i-specific effects of the sensitive attribute C on a binary decision attribute E with positive decision e^+ and negative decision e^- (denoted by $PSE^E_{\pi_d}(\cdot)$ and $PSE^E_{\pi_i}(\cdot)$ for distinguishing with the path-specific effects derived for ranked data in this chapter). Assume that the decision is made based on a cut-off point θ of the score. Then an interesting question is to ask, given a discrimination-free rank, whether a binary decision made based on the cut-off point θ is also discrimination free. Answering this question needs to derive a relationship between $PSE_{\pi}(\cdot)$ and $PSE^E_{\pi}(\cdot)$. In this subsection, we derive such relationships under the condition that $\forall\mathbf{q}$, $\theta \geq \mu_{c^+,\mathbf{q}} \geq \mu_{c^-,\mathbf{q}}$ and $\sigma_{c^+,\mathbf{q}} = \sigma_{c^-,\mathbf{q}} = \sigma$. We first obtain the formulas of $PSE^E_{\pi_d}(\cdot)$ and $PSE^E_{\pi_i}(\cdot)$ using the cut-off point θ.

Lemma 1. *Given the causal graph based on score S, and a cut-off point θ for determining a binary decision E, we have*

$$PSE^E_{\pi_d}(c^+, c^-) = \sum_{\mathbf{q}\in\mathfrak{X}_\mathbf{Q}} \frac{1}{2}\left(\text{erf}(\frac{\theta-\mu_{c^-,\mathbf{q}}}{\sqrt{2}\sigma}) - \text{erf}(\frac{\theta-\mu_{c^+,\mathbf{q}}}{\sqrt{2}\sigma})\right) P(\mathbf{q}|c^-), \tag{5.6}$$

$$PSE^E_{\pi_i}(c^+, c^-) = \sum_{\mathbf{q}\in\mathfrak{X}_\mathbf{Q}} \frac{1-\text{erf}(\frac{\theta-\mu_{c^-,\mathbf{q}}}{\sqrt{2}\sigma})}{2}\Delta_\mathbf{q}. \tag{5.7}$$

Proof. Since θ is a cut-off point, we have $P(e^+|c^+,\mathbf{q}) = P(s \geq \theta|c^+,\mathbf{q})$ and $P(e^+|c^-,\mathbf{q}) = P(s \geq \theta|c^-,\mathbf{q})$. According to the CDF of the Gaussian distribution, we have

$$P(e^+|c^+,\mathbf{q}) = \frac{1-\text{erf}(\frac{\theta-\mu_{c^+,\mathbf{q}}}{\sqrt{2}\sigma})}{2}, \quad P(e^+|c^-,\mathbf{q}) = \frac{1-\text{erf}(\frac{\theta-\mu_{c^-,\mathbf{q}}}{\sqrt{2}\sigma})}{2}.$$

The lemma is proven by substituting $P(e^+|c^+,\mathbf{q})$ and $P(e^+|c^-,\mathbf{q})$ in the formulas of $PSE^E_{\pi_d}$ and $PSE^E_{\pi_i}$ in [4] with the above expressions. □

Then we present two lemmas to show the properties of $\text{erf}(\cdot)$.

Lemma 2. *For any $x_1 \geq x_2 \geq 0$, we have*

$$\frac{1}{2}\left(\text{erf}(x_1) - \text{erf}(x_2)\right) \leq \text{erf}(\frac{x_1-x_2}{2}).$$

Proof. Since $\text{erf}(x)$ $(x \geq 0)$ is concave and $\text{erf}(0) = 0$, we have

$$\frac{\text{erf}(x_2)}{x_2} \geq \frac{\text{erf}(x_1)}{x_1} \quad \Longrightarrow \quad \frac{x_2}{2x_1}\text{erf}(x_1) \leq \frac{1}{2}\text{erf}(x_2)$$

which follows that

$$\left(\frac{1}{2} - \frac{x_2}{2x_1}\right)\text{erf}(x_1) \geq \frac{1}{2}\text{erf}(x_1) - \frac{1}{2}\text{erf}(x_2).$$

Again, since $\text{erf}(x)$ $(x \geq 0)$ is concave and $\text{erf}(0) = 0$, we have

$$\left(\frac{1}{2} - \frac{x_2}{2x_1}\right)\text{erf}(x_1) \leq \text{erf}(\frac{x_1}{2} - \frac{x_2}{2}).$$

Combining the above two inequalities, the lemma is proven. □

Lemma 3. *For any $t \geq 0$, when $0 \leq x \leq t$, we have*

$$\alpha_t x \leq \text{erf}(x) \leq \alpha_t x + \beta_t,$$

where

$$\alpha_t = \frac{\text{erf}(t)}{t}, \quad \beta_t = \text{erf}(\sqrt{\ln \frac{2t}{\sqrt{\pi}\,\text{erf}(t)}}) - \frac{\text{erf}(t)}{t}\sqrt{\ln \frac{2t}{\sqrt{\pi}\,\text{erf}(t)}}.$$

Proof. It is obvious that $\text{erf}(x) \geq \alpha_t x$ $(0 \leq x \leq t)$. Then, β_t is obtained by calculating the tangent line with the slope α_t of $\text{erf}(x)$. □

Based on the above results, the following two theorems characterize the relationship between PSE_π and PSE_π^E.

Theorem 9. *Given the causal graph based on score S and an arbitrary cut-off point θ, if for the ranking derived from the score we have*

$$PSE_{\pi_d}(c^+, c^-) \leq \frac{2\sqrt{2}(\tau - \beta_t)\sigma}{\alpha_t},$$

for the binary decision derived from the score we must have $PSE_{\pi_i}^E(c^+, c^-) \leq \tau$, where

$$t = \max_{\mathbf{q}} \left\{ \frac{\mu_{c^+,\mathbf{q}} - \mu_{c^-,\mathbf{q}}}{2\sqrt{2}\sigma} \right\}.$$

Proof. Let $x_1 = \frac{\theta - \mu_{c^-,\mathbf{q}}}{\sqrt{2}\sigma}$, $x_2 = \frac{\theta - \mu_{c^+,\mathbf{q}}}{\sqrt{2}\sigma}$, according to Lemma 2 we have

$$\frac{1}{2}\left(\text{erf}(x_1) - \text{erf}(x_2)\right) \leq \text{erf}(\frac{x_1 - x_2}{2}) = \text{erf}(\frac{\mu_{c^+,\mathbf{q}} - \mu_{c^-,\mathbf{q}}}{2\sqrt{2}\sigma}).$$

According to Lemma 3 it follows that

$$\text{erf}(\frac{\mu_{c^+,\mathbf{q}} - \mu_{c^-,\mathbf{q}}}{2\sqrt{2}\sigma}) \leq \alpha_t \frac{\mu_{c^+,\mathbf{q}} - \mu_{c^-,\mathbf{q}}}{2\sqrt{2}\sigma} + \beta_t.$$

Combining the above inequality with Eq. (5.6), we have

$$PSE^E_{\pi_d} \leq \sum_{\mathbf{q}\in\mathfrak{X}_\mathbf{Q}} \left(\alpha_t \frac{\mu_{c^+,\mathbf{q}} - \mu_{c^-,\mathbf{q}}}{2\sqrt{2}\sigma} + \beta_t \right) P(\mathbf{q}|c^-) = \frac{\alpha_t}{2\sqrt{2}\sigma} PSE_{\pi_d} + \beta_t \leq \tau.$$

□

Theorem 10. *Given the causal graph based on score S and an arbitrary cut-off point θ, if for the ranking derived from the score we have*

$$PSE_{\pi_i}(c^+, c^-) \leq \frac{2\sqrt{2}(\tau - c)\sigma}{\alpha_t},$$

for the binary decision derived from the score we must have $PSE^E_{\pi_i}(c^+, c^-) \leq \tau$, where

$$t = \max_{\mathbf{q}} \left\{ \frac{\max\{s\} - \mu_{c^-,\mathbf{q}}}{\sqrt{2}\sigma} \right\},$$
$$c = \frac{1}{2} - \sum_{\mathbf{q}:\Delta_\mathbf{q}\geq 0} \left(\frac{\alpha_t \max_\mathbf{q}\{\mu_{c^+,\mathbf{q}}\}}{\sqrt{2}} \right) - \sum_{\mathbf{q}:\Delta_\mathbf{q}<0} \left(\frac{\alpha_t}{2\sqrt{2}} + \beta_t \right).$$

Proof. According to Lemma 3 we have

$$\text{erf}(\frac{\theta - \mu_{c^-,\mathbf{q}}}{\sqrt{2}\sigma}) \geq \text{erf}(\frac{\max_\mathbf{q}\{\mu_{c^+,\mathbf{q}}\} - \mu_{c^-,\mathbf{q}}}{\sqrt{2}\sigma}) \geq \alpha_t \frac{\max_\mathbf{q}\{\mu_{c^+,\mathbf{q}}\} - \mu_{c^-,\mathbf{q}}}{\sqrt{2}\sigma},$$

$$\text{erf}(\frac{\theta - \mu_{c^-,\mathbf{q}}}{\sqrt{2}\sigma}) \leq \text{erf}(\frac{\max\{s\} - \mu_{c^-,\mathbf{q}}}{\sqrt{2}\sigma}) \leq \alpha_t \frac{\max\{s\} - \mu_{c^-,\mathbf{q}}}{\sqrt{2}\sigma} + \beta_t.$$

Combining the above inequalities with Eq. (5.7), we have

$$\begin{aligned} PSE^{E}_{\pi_i} &= \sum_{\mathbf{q}:\Delta_{\mathbf{q}}\geq 0} \frac{1-\mathrm{erf}(\frac{\theta-\mu_{c^-,\mathbf{q}}}{\sqrt{2}\sigma})}{2}\Delta_{\mathbf{q}} + \sum_{\mathbf{q}:\Delta_{\mathbf{q}}<0} \frac{1-\mathrm{erf}(\frac{\theta-\mu_{c^-,\mathbf{q}}}{\sqrt{2}\sigma})}{2}\Delta_{\mathbf{q}} \\ &\leq \frac{1}{2} + \sum_{\mathbf{q}:\Delta_{\mathbf{q}}\geq 0} \alpha_t \frac{\mu_{c^-,\mathbf{q}} - \max_{\mathbf{q}}\{\mu_{c^+,\mathbf{q}}\}}{2\sqrt{2}\sigma}\Delta_{\mathbf{q}} + \sum_{\mathbf{q}:\Delta_{\mathbf{q}}<0} (\alpha_t \frac{\mu_{c^-,\mathbf{q}} - \max\{s\}}{2\sqrt{2}\sigma} - \beta_t)\Delta_{\mathbf{q}} \\ &= \frac{\alpha_t}{2\sqrt{2}\sigma} PSE_{\pi_i} + c \leq \tau. \end{aligned}$$

□

5.5 Discovery and Removal Algorithms

We develop the discrimination discovery and removal algorithms based on the derived π_d and π_i-specific effects. Since the values of $PSE_{\pi_d}(c^+, c^-)$ and $PSE_{\pi_i}(c^+, c^-)$ can be arbitrarily large, we give the criterion of direct and indirect discrimination in terms of relative difference. We require that the ratio of $PSE_{\pi_d}(c^+, c^-)$ and $PSE_{\pi_i}(c^+, c^-)$ over the expected score of the favorable group, i.e., $\mathbb{E}[S|c^+]$, is smaller than a given threshold τ. For example, the Equality and Human Rights Commission (EHRC) consider 0.05 as a significant threshold for the gender pay gap. By defining the discrimination measures

$$DE_d(c^+, c^-) = \frac{PSE_{\pi_d}(c^+, c^-)}{\mathbb{E}[S|c^+]}$$

and

$$DE_i(c^+, c^-) = \frac{PSE_{\pi_i}(c^+, c^-)}{\mathbb{E}[S|c^+]},$$

the criterion of discrimination is shown below. To avoid reverse discrimination, we also similarly define $DE_d(c^+, c^-)$ and $DE_i(c^+, c^-)$. Then, we give the criterion of discrimination as follows.

Criterion 1. *Given a user-defined threshold τ, direct discrimination exists if either $DE_d(c^+, c^-) > \tau$ or $DE_d(c^-, c^+) > \tau$ holds, and indirect discrimination exists if either $DE_i(c^+, c^-) > \tau$ or*

$DE_i(c^-, c^+) > \tau$ *holds.*

Based on the above analysis, we develop the algorithm for discovering discrimination in a rank, referred to as *FDetect*, as shown in Algorithm 7. Once direct or indirect discrimination is detected, the discriminatory effects need to be eliminated before the ranked data is used for training or sharing. We propose a path-specific-effect-based Fair Ranking (*FRank*) algorithm to remove both discrimination from the ranked data and reconstruct a fair ranking. We first modify the score distributions so that the causal graph contains no discrimination, and then reconstruct a fair ranking based on the modified causal graph. As shown in Theorem 8, the discriminatory effect only depends on the means of the score distributions. Hence we only need to modify the means of the score.

Algorithm 7: *FDetect*

Input : Ranked dataset $\mathcal{D}$, sensitive attribute C, user-defined parameter τ.
Output : Direct/indirect discrimination $judge_d$, $judge_i$.

```
judge_d = judge_i = false;
Derive the score S using the Bradley-Terry model;
Build the causal graph for S and attributes in D;
Compute DE_d(·) according to Theorem 8;
if DE_d(c+, c−) > τ || DE_d(c−, c+) > τ then
    judge_d = true;
Divide C's children except S into V_{π_i} and V̄_{π_i};
Compute DE_i(·) according to Theorem 8;
if DE_i(c+, c−) > τ || DE_i(c−, c+) > τ then
    judge_i = true;
return [judge_d, judge_i];
```

To maximize the utility during the modification process, we minimize the distance between the original score distributions and the modified score distributions, as measured by the Bhattacharyya distance [85]. Specifically, for each score distribution $\mathcal{N}(\mu_{c,\mathbf{q}}, \sigma^2_{c,\mathbf{q}})$, denote the modified distribution by $\mathcal{N}(\mu'_{c,\mathbf{q}}, \sigma^2_{c,\mathbf{q}})$. The Bhattacharyya distance between the two distributions is given by

$$D_B = -\ln \int \sqrt{\mathcal{N}(\mu_{c,\mathbf{q}}, \sigma^2_{c,\mathbf{q}})\mathcal{N}(\mu'_{c,\mathbf{q}}, \sigma^2_{c,\mathbf{q}})}ds = \frac{\left(\mu_{c,\mathbf{q}} - \mu'_{c,\mathbf{q}}\right)^2}{8\sigma^2_{c,\mathbf{q}}}.$$

We define the objective function as the sum of the Bhattacharyya distances for all score distributions. As a result, we obtain the following quadratic programming problem with $\mu_{c,\mathbf{q}}$ as the variables.

$$\begin{aligned} \text{minimize} \quad & \sum_{c \in \mathfrak{X}_C, \mathbf{q} \in \mathfrak{X}_\mathbf{Q}} \frac{\left(\mu_{c,\mathbf{q}} - \mu'_{c,\mathbf{q}}\right)^2}{\sigma^2_{c,\mathbf{q}}} \\ \text{subject to} \quad & DE_d(c^+, c^-) \leq \tau, \quad DE_d(c^-, c^+) \leq \tau, \\ & DE_i(c^+, c^-) \leq \tau, \quad DE_i(c^-, c^+) \leq \tau. \end{aligned}$$

After obtaining the modified score distribution by solving the quadratic programming problem, we reconstruct a fair ranking as follows. Consider the individuals with the same profile $c, \mathbf{q}$, i.e., $\forall i, c_i = c, \mathbf{q}_i = \mathbf{q}$. For each individual i, the new score s'_i is regenerated from the new CG distribution $\mathcal{N}(\mu'_{c,\mathbf{q}}, \sigma^2_{c,\mathbf{q}})$ at the same percentile as the score s_i in the original distribution. Specifically, since $s_i = \mu_{c,\mathbf{q}} + \rho\sigma_{c,\mathbf{q}}$ and $s'_i = \mu'_{c,\mathbf{q}} + \rho\sigma_{c,\mathbf{q}}$ where ρ is the value from the standard normal distribution for the percentile, we have $s'_i = s_i + (\mu'_{c,\mathbf{q}} - \mu_{c,\mathbf{q}})$. Finally, we re-rank all individuals according to the descending order of their new scores. Since the new scores contain no discrimination, so does the new rank. The procedure is shown in Algorithm 8, referred to as *FRank*.

Algorithm 8: *FRank*

Input : Ranked dataset $\mathcal{D}$, sensitive attribute C, user-defined parameter τ.
Output : Modified dataset $\mathcal{D}^*$.

```
if PSE-DD(D, C, τ) == [false, false] then
    return;
Obtain the modified distributions of S by solving the quadratic programming
  problem;
foreach c, q do
    foreach i : c_i = c, q_i = q do
        s'_i = s_i + (μ'_{c,q} − μ_{c,q});
Compute the new rank of each individual according to the descending order of S,
  and replace the rank in D with the new one to obtain D*;
return D*;
```

The computational complexity of our discovery and removal algorithms depends on how efficiently to derive the score S using Bradley-Terry model. Wu et al. [84] proved that the likelihood function is convex and the optimal solution can be efficiently obtained using gradient descent. The complexity also depends on the complexities of building the causal graph and computing the path-specific effect. Many researches have been devoted to improving the performance of network construction [60, 62, 63] and probabilistic inference in causal graphs [64, 65]. The complexity analysis can be found in these related literature.

5.6 Experiments

5.6.1 Experimental Setup

In the experiments, the causal graphs are then constructed using Tetrad [70] and parameterized as described in Section 5.4.1. The quadratic programming is solved using CVXOPT [71]. The discrimination threshold τ is set as 0.05 for both direct and indirect discrimination. The generated data and algorithm implementations are available at http://tiny.cc/fair-ranking.

Dataset. We use a real world dataset, the *German Credit* dataset [68], which is also used in previous works [76, 77]. The description of the *German Credit* dataset is given in Appendix A.1. Due to the small sample size, we only select 8 attributes in our experiments including `age`, `dependent`, `duration`, `housing`, `job`, `property`, `purpose`, `residence`. We treat `age` as the sensitive attribute, `housing` as the redlining attribute.

Based on the *German Credit* dataset, we generate three ranked datasets for experiments. We employ the weighted-sum ranking strategy proposed in [76, 77] to generate two ranked datasets, denoted by **D1** and **D2**. The weighted sum is computed through a weighted linear summation of certain attributes, and then all candidates are ranked according to the weighted sum. In **D1**, all attributes are summed up with equal weight, while in **D2**, the summation is for all attributes except `age`. We also use another ranked data **D** where the ranking is directly based on an original attribute `credit amount`. After that, we derive the continuous

qualification scores from each ranked dataset using the Bradley-Terry model and build the causal graph. As an example, the constructed causal graph for **D** is shown in Figure 5.3.

Baseline. We involve the statistical parity-based discrimination discovery and removal algorithms proposed by Yang et al. [77] and Zehlike et al. [76]. For discrimination discovery, Yang et al. [77] proposed three set-based discrimination measures called **rRD**, **rND**, and **rKL** to compute the difference between the favorable group and the whole dataset in terms of *risk difference*, *risk ratio*, and *Kullback-Leibler distance*. They compute the values of difference at several discrete points (e.g., top-10, top-20, $\cdots$) and sum up all values with the logarithmic discounts. All measures are normalized to 0-1 range (0 is the most fair value and 1 is the least fair value). Since they do not provide any criterion for discrimination discovery, we simply use 0.05 as the threshold for all three measures. Zehlike et al. [76] proposed an adjusted fairness condition (*FairCon*) that requires the minimum number of sensitive candidates in every prefix of the ranking list. For discrimination removal, Yang et al. [77] proposed a fair data generator (*FairGen*) that manipulates the permutation according to the user-defined preference f. For example, if $f = 0.05$, all the candidates are well mixed in equal proportion at every prefix; if $f = 1$, the candidates from the unfavorable group are ranked at the bottom. Zehlike et al. [76] proposed discrimination removal methods, *FA*IR*, to select the most qualified candidate from the corresponding unfavorable group at every prefix in order to satisfy the adjusted fairness condition.

To evaluate the data utility of all removal approaches, we adopt two widely used metrics, the Spearman's footrule distance (*SFD*) and the Kendall's tau distance (*KTD*) [86]. The Spearman's footrule distance (*SFD*) measures the total element-wise displacement between the modified permutation and the original one. The Kendall's tau distance (*KTD*) measures the total number of pairwise inversions between the two permutations. For both of the distance metrics, the larger values indicate more data utility loss.

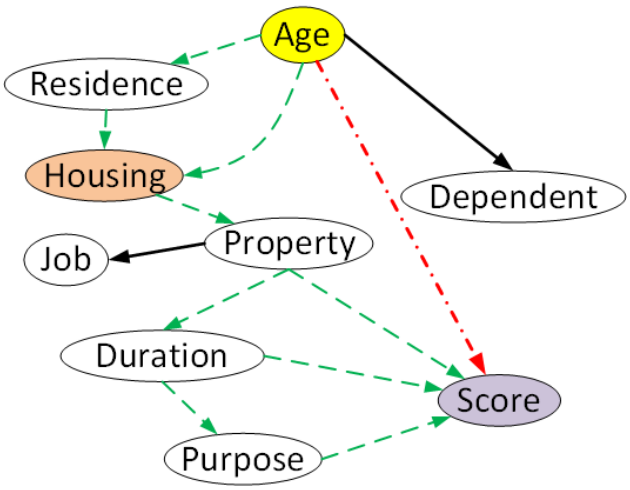


Figure 5.3: The causal graph of **D**. The yellow node Age is the sensitive attribute, the orange node Housing is the redlining attribute, and the purple node Score is the decision attribute. The red dash-dot line captures the direct discrimination from Age to Score, and the green dashed line captures the indirect discrimination through Housing.

5.6.2 Discrimination Discovery

We quantify the strength of direct and indirect discrimination using our method *FDetect* for all three ranked datasets. The results are shown in Table 5.1. For dataset **D1**, all attributes including the sensitive attribute are used for ranking directly. Thus, the ground-truth is that both direct and indirect discrimination occurs in this dataset. Our method obtains $DE_d(c^+, c^-) = 0.231$ and $DE_i(c^+, c^-) = 0.055$, showing that both direct and indirect discrimination are correctly identified. For dataset **D2**, since we use all the other attributes except the sensitive attribute in the ranking process, the ground-truth is that the indirect discrimination occurs but the direct discrimination does not. Our method shows $DE_d(c^+, c^-) = 0.026$ and $DE_i(c^+, c^-) = 0.061$, which is also consistent with the ground-truth. For dataset **D**, we do not have the ground-truth. Our method obtains that $DE_d(c^+, c^-) = 0.005$ and $DE_i(c^+, c^-) = 0.013$. The results imply that neither direct discrimination nor indirect discrimination exists in this dataset.

The statistical parity-based methods **rRD**, **rND**, **rKL** and *FairCon* cannot distinguish direct and indirect discrimination. We directly report the results produced by these methods

as shown in Table 5.1. For **D1**, the method proposed by Yang et al. shows that **rRD** = 0.590, **rND** = 0.440, and **rKL** = 0.204, while Zehlike's *FairCon* shows that the third position does not satisfy the minimum fair requirement. For **D2**, Yang's method shows that **rRD** = 0.160, **rND** = 0.102, and **rKL** = 0.022, while *FairCon* reports that the 5-th position does not satisfy the fair requirement. Most methods conclude discrimination for both dataset, which kind of match our conclusions. However, for **D**, Yang's method shows that **rRD** = 0.109, **rND** = 0.070, and **rKL** = 0.008, where three values make the contradictory conclusions: there is no discrimination according to **rRD** but **rND** and **rKL** report significant discrimination. *FairCon* shows that the ranking cannot satisfy the fair requirement at the 20-th position. All methods cannot obtain the results that are consistent with ours, implying that they may produce incorrect or misleading conclusions.

Table 5.1: Comparison of discrimination discovery methods. The second column represents the ground-truth for direct and indirect discrimination.

	Ground-Truth	DE_d	DE_i	*FairCon*	**rRD**	**rND**	**rKL**
D1	Y/Y	0.231	0.055	3rd	0.590	0.440	0.204
D2	N/Y	0.026	0.061	5th	0.160	0.102	0.022
D	-	0.005	0.013	20th	0.109	0.070	0.008

5.6.3 Discrimination Removal

We perform *FRank* to remove discrimination and reconstruct fairly ranked datasets with neither direct nor indirect discrimination. Our theoretical results guarantee that there is no discrimination after modification. For comparison, we also execute *FairGen* [77] and *FA*IR* [76]. After removing discrimination, we further apply *FDetect* to evaluate whether the newly-generated data achieves truly discrimination-free. The results of three removal methods are shown in Table 5.2. As can be seen, our method *FRank* removes both direct and indirect discrimination precisely. However, *FairGen* and *FA*IR* cannot achieve discrimination-free. *FairGen* removes neither direct nor indirect discrimination. It even introduces more discrimination to **D**. *FA*IR* can mitigate part of direct discrimination, but fails to remove

indirect discrimination.

Table 5.2: Discrimination and data utility measured on the new ranked data produced by *FairGen*, *FA*IR*, and our *FRank*. Values violating the discrimination criterion are marked in bold.

Data	Methods	DE_d	DE_i	*KTD*	*SFD*
D1	*FRank*	0.050	0.050	24602	72938
	FairGen	**0.234**	**0.064**	11150	44600
	*FA*IR*	**0.077**	**0.066**	13882	55528
D2	*FRank*	0.029	0.050	5851	18090
	FairGen	**0.246**	**0.060**	19483	77934
	*FA*IR*	0.022	**0.061**	231	924
D	*FRank*	0.005	0.013	0	0
	FairGen	**0.250**	0.012	20806	83226
	*FA*IR*	0.003	0.013	143	572

We adopt the Spearman's footrule distance (*SFD*) and the Kendall's tau distance (*KTD*) to evaluate the data utility loss when mitigating the discrimination. As can be seen from the last two columns of Table 5.2, our method *FRank* incurs relatively small data utility loss, but *FairGen* suffers large data utility loss while not achieving discrimination-free. Although *FA*IR* introduces quite a small data utility loss, it fails to mitigate indirect discrimination. It is worth pointing out that there is no direct or indirect discrimination in **D** so our *FRank* does not result in any distortion. On the contrary, *FairGen* leads to too much utility loss.

We also examine how the data utility varies with different values of the discrimination threshold τ. We perform *FRank* on **D1** and vary the threshold τ for *FRank* from 0.00 to 0.25 for evaluating how much data utility loss is incurred. In Table 5.3, we can see that both the Spearman's footrule distance (*SFD*) and the Kendall's tau distance (*KTD*) decrease with the increase of τ, which means that less utility loss is incurred with a larger threshold. This observation is consistent with our analysis since the larger τ, the more relaxed the constraints in *FRank*.

Table 5.3: Comparison of *FRank* with varied τ.

τ	0.00	0.05	0.10	0.15	0.20	0.25
$DE_d(c^+, c^-)$	0.000	0.050	0.100	0.150	0.200	0.231
$DE_i(c^+, c^-)$	0.000	0.050	0.055	0.055	0.055	0.055
SFD	43490	24602	14370	9041	3444	0
KTD	123626	72938	45636	28652	11054	0

5.7 Summary

In this chapter, we studied the problem of discovering discrimination in a rank and reconstructing a fair rank if discrimination is detected. We leveraged structural causal model to capture the bias in the rank as the causal effect. To address the limitation of the existing single data-type causal graph, we modeled the ranking positions using a continuous score, and built the causal graph for the profile attributes as well as the score. Then, we extended the path-specific effect technique to the mixed-variable causal graph, which is used to quantitatively measure direct and indirect discrimination in the ranked data. We also theoretically analyzed the relationship between the path-specific effects for the ranked data and those for the binary decision. Based on that, we developed an algorithm for discovering both direct and indirect discrimination, as well as an algorithm to reconstruct a fair rank from the causal graph. The experiments using the *German Credit* dataset showed that our methods correctly measure the discrimination in the rank and reconstruct a rank that does not contain either direct or indirect discrimination, while the statistical parity-based method may obtain incorrect and misleading results. This work has been published in KDD 2018 [6].

In Theorem 8 we assumed that the π_i-specific effect is identifiable from the data. In some cases, the π_i-specific effect is not able to be computed from the data due to the inherent unidentifiability of the path-specific effect [2]. In Chapter 4, we have discussed how to deal with this situation and developed lower and upper bounds for the unidentifiable path-specific effect. Similar ideas can be adopted to deal with unidentifiable situation for ranked data.

6 Counterfactual Fairness

6.1 Introduction

Recently, the research community has studied fairness-aware machine learning from the causal perspective [4, 5, 28–30, 73] using causal modeling [54]. In these works, fairness is generally formulated and quantified as the average causal effect of the sensitive attribute on the decision attribute. The effect is evaluated by the intervention through the post-interventional distributions. Different from above works, Kusner et al. [7] introduced counterfactual fairness, based on the counterfactual inference, which considers the causal effect within a particular individual/group specified by of observational profile attributes. The notion of counterfactual fairness is more general than the intervention-based notions where the set of profile attributes is empty. Consequently, the counterfactual inference is more challenging than the intervention. This is because measuring interventions only considers the post-interventional distributions, but counterfactual inference considers both the real world without the intervention and the counterfactual world with the intervention. Researchers have proved that the counterfactual quantity cannot be uniquely computed from the observational data in some situations, which are referred to as the unidentifiable situations [54].

The unidentifiable situations are big barriers to the application of counterfactual fairness. In [7], the authors proposed three methods to evade the unidentifiability issue: 1) only non-descendants of the sensitive attribute are used in classification, 2) the non-deterministic substitutions of the hidden variables are postulated and inferred based on domain knowledge, or 3) the complete causal model is postulated and estimated, e.g., , being treated as the additive noise model then estimating the errors. However, the sensitive attribute is usually an inherent nature of data hence many attributes are its descendants. If all descendants are forbidden, very few attributes are allowed for classifier training, weakening the resultant fair classifier dramatically. Also, it is over-simplified to postulate the substitutions and their

distributions, since the exogenous variables represent all possible sources of randomness; or presuppose that the causal model, which is supposed to represent the underlying mechanism of the world, is an additive model.

In this chapter, we address the problem of learning counterfactually fair classifiers by mathematically bounding the unidentifiable counterfactual quantity. We leverage the counterfactual graph proposed in [87] for depicting the independence relationships among variables in the real world and the counterfactual world which are of concern in the counterfactual quantity. Then, we adopt the c-component factorization to decompose the counterfactual quantity, and identify the terms that are the source of unidentification. We propose a graphical criterion for determining the identification of counterfactual fairness and develop the lower and upper bounds of counterfactual fairness in unidentifiable situations. Finally, we propose a post-processing method for reconstructing arbitrary classifiers in order to achieve counterfactual fairness. We formulate the reconstruction problem as a linear constrained optimization problem with the bounded counterfactual fairness criterion as the constraints.

In the experiments, we evaluate our methods and compare them with existing ones using real-world datasets and synthetic datasets where the ground-truth of counterfactual fairness can be precisely quantified. The results show that our method correctly achieves counterfactual fairness as expected according to our theorem, while obtaining high accuracy of prediction. On the contrary, the methods proposed in [7] either fail to achieve counterfactual fairness or suffer from low accuracy due to simplified assumptions.

6.2 Preliminaries

6.2.1 Counterfactual Inference and Unidentification

In Definition 3 of Chapter 3, the total causal effect is estimated using intervention where the post-intervention distribution concerns the counterfactual world represented by submodel $\mathcal{M}_x$ only. If we infer the post-intervention distribution while conditioning on certain individuals or groups specified by a subset of endogenous variables, the inferred

quantity will involve two worlds simultaneously, the real world represented by causal model $\mathcal{M}$, and the counterfactual world $\mathcal{M}_x$, hence cannot be resolved by *do*-calculus directly. Such causal inference problem is called the counterfactual inference, and the distribution of Y_x conditioning on the real world observation $\mathbf{O} = \mathbf{o}$ is denoted by $P(y_x|\mathbf{o})$. Note that Y_x is a variable in submodel $\mathcal{M}_x$, while $\mathbf{O}$ are variables in original causal model $\mathcal{M}$.

Apparently, inferring $P(y_x|\mathbf{o})$ requires to know the connection between the real world and the counterfactual world. This can be done if we have complete knowledge of the causal model. According to [54], the counterfactual inference can be exactly performed using three steps if the complete model, including all the structural equations, is known: **1. Abduction**: Update $P(\mathbf{u})$ by observation $\mathbf{O} = \mathbf{o}$ to obtain $P(\mathbf{u}|\mathbf{o})$. **2. Action**: Modify $\mathcal{M}$ by intervention $do(x)$ to obtain the submodel $\mathcal{M}_x$. **3. Prediction**: Use modified submodel $\langle \mathcal{M}_x, P(\mathbf{u}|\mathbf{o}) \rangle$ to compute the probability of Y_x,, i.e., the consequence of the counterfactual inference.

The above method is usually infeasible in practice due to the lack of the complete knowledge of the causal model. If we only have the causal graph and observational data, which is a common scenario in the literature, the counterfactual quantity might be evaluated by using the **IDC*** algorithm developed in [87]. However, in certain situations where the **IDC*** algorithm fails, the corresponding counterfactual quantity cannot be uniquely computed from the observational data in theory. These situations are referred to as the unidentifiable situations based on Definition 2. One typical unidentifiable situation [87] is shown in Lemma 4.

Lemma 4. *Let X, Y be two variables such that Y is a parent of X, then $P(Y = y, Y_x = y')$ is unidentifiable if $y \neq y'$.*

6.3 Quantifying and Bounding Counterfactual Fairness

Fairness-aware learning is widely studied using causal modeling to capture the causal connection between the sensitive attribute and the challenged decision [4, 7, 29–31, 88, 89]. We adopt the notion of counterfactual fairness proposed in [7], which formulates fairness as the equivalence of two counterfactual quantities. Although this notion captures the true intuition

behind fairness, it faces significant computational challenges due to the unidentifiability of counterfactual inference. In this section, we first give the formal definition of counterfactual fairness for predictive models and explain its physical meaning. Then, we show how to address above challenges by mathematically bounding the unidentifiable counterfactual quantity.

In our notations, $S \in \{s^+, s^-\}$ denotes the sensitive attribute, $Y \in \{y^+, y^-\}$ denotes the decision, and $\mathbf{X}$ denotes the set of other attributes. The historical dataset $\mathcal{D}$ drawn from a distribution $P(\mathbf{X}, S, Y)$ is used to train a classifier $f : \mathbf{X}, S \to \hat{Y}$. The underlying mechanism that determines a distribution $P(\mathbf{X}, S, \hat{Y})$ is represented by a causal model $\mathcal{M}$. The causal graph associated with the causal model is denoted by $\mathcal{G}$. Then, counterfactual fairness is defined as follows.

Definition 6. *(**Counterfactual Fairness**) Given a set of attributes $\mathbf{Z} \subseteq \mathbf{X}$, a classifier $f : \mathbf{X}, S \to \hat{Y}$ is counterfactually fair w.r.t. $\mathbf{Z}$, if under any observational condition $\mathbf{Z} = \mathbf{z}$ we have*

$$P(\hat{y}_{s'}|s', \mathbf{z}) = P(\hat{y}_s|s', \mathbf{z}), \text{ where } s', s \in \{s^+, s^-\}.$$

Recall that a lowercase letter with a subscript represents a value assignment to the corresponding variable in the submodel, e.g., $\hat{y}_s$ is a value of $\hat{Y}_s$ in the submodel $\mathcal{M}_s$.

The physical meaning of counterfactual fairness can be interpreted as follows. Consider candidates are applying for a job and a predictive model is used to make the decision $\hat{Y}$. We concern an individual from disadvantage group s^- who is specified by a profile $\mathbf{z}$. Straightforwardly, the probability of the individual to get the positive decision is $P(\hat{y}|s^-, \mathbf{z})$, which is equivalent to $P(\hat{y}_{s^-}|s^-, \mathbf{z})$ since the intervention makes no change to S's value of that individual. Now assume the value of S for this very individual had been changed from s^- to s^+. The probability of this individual to get the positive decision after the hypothetical change is given by $P(\hat{y}_{s^+}|s^-, \mathbf{z})$. Therefore, if two probabilities $P(\hat{y}_{s^-}|s^-, \mathbf{z})$ and $P(\hat{y}_{s^+}|s^-, \mathbf{z})$ are identical, we can claim the individual is treated fairly as if he/she had been from the other group.

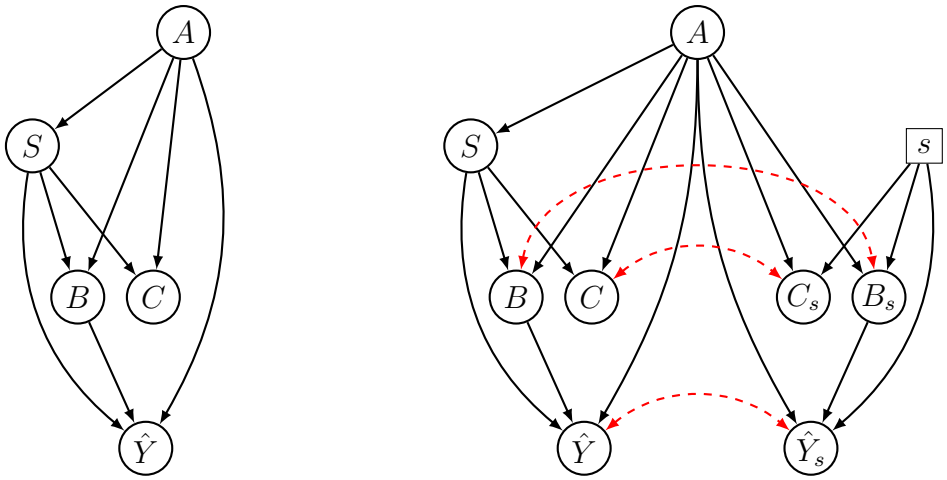


Figure 6.1: (a) Causal Graph $\mathcal{G}$. (b) Counterfactual Graph $\mathcal{G}'$ for $P(\hat{y}_s|s', \mathbf{z})$.

6.3.1 Identification of Counterfactual Quantity

In this section, we identify the source of unidentification for the counterfactual quantity and give a graphical criterion determining the identifiability of the counterfactual quantity. Our method is inspired by the **IDC*** algorithm and we further extend it to bound the unidentifiable quantity.

The analysis of $P(\hat{y}_s|s', \mathbf{z})$ concerns the connection between two causal models, $\mathcal{M}$ and $\mathcal{M}_s$. Thus, we apply the **make-cg** algorithm [87] to the causal graph $\mathcal{G}$ to construct a new graph $\mathcal{G}'$ that depicts the independence relationship among all variables in $\mathcal{M}$ and $\mathcal{M}_s$ that are of concern in the analysis. The **make-cg** algorithm first combines the two causal graphs and makes them share the same exogenous variables $\mathbf{U}$, corresponding to the shared causal context or background. Then, it removes the duplicated endogenous nodes which are also not affected by $do(s)$. The resultant graph is the so-called counterfactual graph. Next, we apply the c-component factorization [90] to decompose counterfactual graph $\mathcal{G}'$ into disjoint subgraphs called the c-components, such that any two nodes in the same c-component are connected by a bi-directed path[1]. After that, the joint distribution of all variables in the counterfactual graph can be factorized as the product of the conditional distribution of

[1]A bi-directed path is a path consisting of bi-directed edges only.

each c-component. Our theoretical analysis will show that if certain c-component has the unidentifiability issue that cannot be resolved by summation, the corresponding counterfactual quantity is unidentifiable. Without loss of generality, we first use an example to illustrate our idea. Consider the causal graph $\mathcal{G}$ shown in Figure 6.1 (a) where there are five attributes $A, B, C, S, \hat{Y}$: S is the sensitive attribute; $\hat{Y}$ is the prediction of the decision attribute obtained by any classifier; A is the ancestor of $\hat{Y}$ but not the descendant of S; B is the intersection between the ancestor of Y and the descendant of S; and C is the descendant of S but not the ancestor of $\hat{Y}$. We aim to study the identifiability of $P(\hat{y}_s|s', \mathbf{z})$, where $\mathbf{Z}$ is an arbitrary subset of $\{A, B, C\}$.

The counterfactual graph denoted by $\mathcal{G}'$ is shown in Figure 6.1 (b), where the bi-directed dash edge implies that the two nodes share the same exogenous variables. Note that A and A_s are merged as A since they are duplicated. Next, we apply the c-component factorization. In Figure 6.1 (b), there are five c-components: $\langle A \rangle$, $\langle S \rangle$, $\langle B, B_s \rangle$, $\langle C, C_s \rangle$, and $\langle \hat{Y}, \hat{Y}_s \rangle$. We can factorize $P(\hat{y}_s, s', \mathbf{z})$ as

$$P(\hat{y}_s, s', \mathbf{z}) = \sum_{\mathbf{x} \setminus \mathbf{z}, \hat{y}, b', c'} R(a)R(s')R(c, c'_s)R(b, b'_s)R(\hat{y}', \hat{y}_s),$$

where $R(\mathbf{w}) = P\left(\mathbf{w}|\mathsf{Pa}(\mathbf{W})_{\mathcal{G}'}\right)$ for any node set $\mathbf{W}$, $\mathbf{x} = \{a, b, c\}$, and $\mathbf{z}$ is any subset of $\mathbf{x}$. Then, we can derive that

$$P(\hat{y}_s|s', \mathbf{z}) = \frac{\sum_{\mathbf{x} \setminus \mathbf{z}, \hat{y}, b', c'} \left[P(a)P(s'|a)P(c, c'_s|s', a)P(b, b'_s|s', a)P(\hat{y}', \hat{y}_s|a, b, b'_s)\right]}{P(s', \mathbf{z})}.$$

Note that c'_s in $P(c, c'_s|s', a)$ and $\hat{y}'$ in $P(\hat{y}, \hat{y}_s|a, b, b'_s)$ can be canceled out by summation. By applying the m-separation, we can remove b from $P(\hat{y}_s|a, b, b'_s)$, as B is d-separated from $\hat{Y}_s$ conditioning on A and B_s. Thus, we obtain

$$P(\hat{y}_s|s', \mathbf{z}) = \frac{\sum_{\mathbf{x} \setminus \mathbf{z}, b'} \left[P(a)P(s'|a)P(c|s', a)P(b, b'_s|s', a)P(\hat{y}_s|a, b'_s)\right]}{P(s', \mathbf{z})}. \tag{6.1}$$

To further analyze Eq. (6.1), we consider two cases below.

Case 1 ($B \notin \mathbf{Z}$): In this case, we have b under the Σ of Eq. (6.1), hence b in $P(b, b'_s|s', a)$ can be canceled out by summation, resulting in $P(b'_s|s', a)$. Then, we can remove s' from $P(b'_s|s', a)$ as B_s is d-separated from S conditioning on A, resulting in $P(b'_s|a)$. We can further rewrite $P(\hat{y}_s|a, b'_s)$ as $P_s(\hat{y}|a, b')$, and rewrite $P(b'_s|a)$ as $P(b'_s|a)$. At last, we invoke *do*-calculus Rule 2 [54] to convert $P_s(\hat{y}|a, b')$ to $P(\hat{y}|a, b', s)$, and $P_s(b'|a)$ to $P(b'|a, s)$. Finally, we obtain

$$\begin{aligned} P(\hat{y}_s|s', \mathbf{z}) &= \frac{\sum_{\mathbf{x}\setminus\mathbf{z}\setminus\{b\},b'} P(s', a, c)P(b'|a, s)P(\hat{y}|a, b', s)}{P(s', \mathbf{z})} \\ &= \frac{\sum_{\mathbf{x}\setminus\mathbf{z}\setminus\{b\}} P(s', a, c)P(\hat{y}|a, s)}{P(s', \mathbf{z})}. \end{aligned} \tag{6.2}$$

Case 2 ($B \in \mathbf{Z}$): In this case, since we do not have b under the Σ, term $P(b, b'_s|s', a)$ cannot be reduced, resulting in

$$P(\hat{y}_s|s', \mathbf{z}) = \frac{\sum_{\mathbf{x}\setminus\mathbf{z},b'} P(s', a, c)P(b, b'_s|a, s')P(\hat{y}|a, b', s)}{P(s', \mathbf{z})}. \tag{6.3}$$

From above two cases we see that, $P(\hat{y}_s|s', \mathbf{z})$ in Case 1 is identifiable as all terms in Eq. (6.2) can be read from observational data. One can verify that this result is consistent with the **IDC*** algorithm. However in Case 2, since $P(b, b'_s|s', a)$ in Eq. (6.3) is unidentifiable according to Lemma 4, $P(\hat{y}_s|s', \mathbf{z})$ is also unidentifiable. In this example, the identifiability of $P(\hat{y}_s|s', \mathbf{z})$ depends on whether node B, the intersection of S's descendants and $\hat{Y}$'s ancestors, is in set $\mathbf{Z}$ or not. We summarize this result as follows.

Proposition 6. *For the causal graph in Figure 6.1 (a), $P(\hat{y}_s|s', \mathbf{z})$ is unidentifiable if and only if $B \in \mathbf{Z}$.*

6.3.2 Bounding Unidentifiable Counterfactual Quantity

In Eq. (6.3), we identify the source of unidentifiability. Next, we derive the lower and upper bounds for $P(\hat{y}_s|s', \mathbf{z})$ as shown in the following proposition, which works for both identifiable and unidentifiable situations.

Proposition 7. *For the causal graph in Figure 6.1 (a) we have*

$$P(\hat{y}_s|s', \mathbf{z}) \leq \frac{\sum_{\mathbf{x}\setminus\mathbf{z}} P(s', \mathbf{x}) \max_{\mathbf{m}'} \{P(\hat{y}|s, a, b')\}}{P(s', \mathbf{z})}, \tag{6.4}$$

$$P(\hat{y}_s|s', \mathbf{z}) \geq \frac{\sum_{\mathbf{x}\setminus\mathbf{z}} P(s', \mathbf{x}) \min_{\mathbf{m}'} \{P(\hat{y}|s, a, b')\}}{P(s', \mathbf{z})}, \tag{6.5}$$

where $\mathbf{x} = \{a, b, c\}$, $\mathbf{z}$ *is any subset of* $\mathbf{x}$, *and* $\mathbf{M} = \{B\} \cap \mathbf{Z}$.

Proof. Suppose $B \in \mathbf{Z}$, then $\mathbf{M} = \{B\}$. Obviously, we have

$$P(\hat{y}|s, a, b') \leq \max_{b'} \{P(\hat{y}|s, a, b')\}.$$

By applying this inequality to Eq. (6.3), we have

$$\begin{aligned} P(\hat{y}_s|s', \mathbf{z}) &\leq \frac{\sum_{\mathbf{x}\setminus\mathbf{z}} P(s', a, c) \max_{b'} \{P(\hat{y}|s, a, b')\} \sum_{b'} P(b, b'_s|s', a)}{P(s', \mathbf{z})} \\ &= \frac{\sum_{\mathbf{x}\setminus\mathbf{z}} P(s', a, c) P(b|s', a) \max_{b'} \{P(\hat{y}|s, a, b')\}}{P(s', \mathbf{z})} \\ &= \frac{\sum_{\mathbf{x}\setminus\mathbf{z}} P(s', \mathbf{x}) \max_{b'} \{P(\hat{y}|s, a, b')\}}{P(s', \mathbf{z})}. \end{aligned}$$

The second step is due to the condition $\sum_{b'} P(b, b'_s|s', a) = P(b|s', a)$, and the third step is due to $B \perp C|A, S$. Similarly, we can replace *max* with *min* to obtain Eq. (6.5).

If $B \notin \mathbf{Z}$, $\mathbf{M} = \emptyset$. Then, we have $\max_{\emptyset}\{P(\hat{y}|s, a, b')\} = \min_{\emptyset}\{P(\hat{y}|s, a, b')\} = P(\hat{y}|s, a)$ and both Eq. (6.4) and Eq. (6.5) become $\frac{\sum_{\mathbf{x}\setminus\mathbf{z}\setminus\{b\}} P(s', a, c) P(\hat{y}|s, a)}{P(s', \mathbf{z})}$, which is consistent with the identifiable situations (i.e., Eq. (6.2)). □

6.3.3 Extending to General Case

Above results can be extended to the general case. Let $\mathbf{A}$ denote the ancestors of $\hat{Y}$ which are not the descendants of S, $\mathbf{B}$ denote the intersection between the ancestors of $\hat{Y}$ and the descendants of S, $\mathbf{C}$ denote the descendants of S which are not the ancestors of $\hat{Y}$, i.e.,

$$\mathbf{A} = \mathsf{An}(\hat{Y})_{\mathcal{G}} \setminus \mathsf{De}(S)_{\mathcal{G}}, \qquad \mathbf{B} = \mathsf{An}(\hat{Y})_{\mathcal{G}} \cap \mathsf{De}(S)_{\mathcal{G}},$$
$$\mathbf{C} = \mathsf{De}(S)_{\mathcal{G}} \setminus \mathsf{An}(\hat{Y})_{\mathcal{G}}.$$

Note that $\mathbf{A}, \mathbf{B}, \mathbf{C}$ are disjoint and $\mathbf{X} = \mathbf{A} \cup \mathbf{B} \cup \mathbf{C}$. Now we are ready to extend Propositions 6 and 7 to the general case.

Theorem 11. ***(Identification of Counterfactual Quantity)*** *Given a causal graph $\mathcal{G}$ and the set of profile attributes $\mathbf{Z}$, the counterfactual quantity $P(\hat{y}_s|s', \mathbf{z})$ is unidentifiable if and only if $\mathbf{B} \cap \mathbf{Z} \neq \emptyset$.*

Theorem 12. ***(Bounds of Counterfactual Quantity)*** *Given a causal graph $\mathcal{G}$ and a set of profile attributes $\mathbf{Z}$, we have*

$$P(\hat{y}_s|s', \mathbf{z}) \leq \frac{\sum_{\mathbf{x}\setminus\mathbf{z}} \left[P(s', \mathbf{x}) \max_{\mathbf{m}'} \left\{ P(\hat{y}|s, \mathsf{pa}(\hat{\mathbf{Y}})_{\mathcal{G}} \cap \mathbf{m}', \mathsf{pa}(\hat{\mathbf{Y}})_{\mathcal{G}} \setminus \{s, \mathbf{m}'\} \right\} \right]}{P(s', \mathbf{z})},$$
$$P(\hat{y}_s|s', \mathbf{z}) \geq \frac{\sum_{\mathbf{x}\setminus\mathbf{z}} \left[P(s', \mathbf{x}) \min_{\mathbf{m}'} \left\{ P(\hat{y}|s, \mathsf{pa}(\hat{\mathbf{Y}})_{\mathcal{G}} \cap \mathbf{m}', \mathsf{pa}(\hat{\mathbf{Y}})_{\mathcal{G}} \setminus \{s, \mathbf{m}'\} \right\} \right]}{P(s', \mathbf{z})},$$

where we partition $\mathbf{B}$ to two disjoint sets: a set $\mathbf{M} \in \mathbf{Z}$ and a set $\mathbf{N} \notin \mathbf{Z}$ such that $\mathbf{M} = \mathbf{B} \cap \mathbf{Z}, \mathbf{N} = \mathbf{B} \setminus \mathbf{Z}$.

The proofs are similar to the previous ones.

6.4 Achieving Counterfactual Fairness in Classification

The derived bounds clear the path towards constructing counterfactually fair classifiers. In this section, we propose a post-processing method for reconstructing any classifier to

achieve counterfactual fairness. To this end, we first give a relaxed quantitative criterion of fairness based on Definition 6.

Definition 7 (τ-Counterfactual Fairness)**.** *Given a profile attribute set* $\mathbf{Z} \subseteq \mathbf{X}$ *and a threshold* τ*, a classifier* $f : \mathbf{X}, S \to \hat{Y}$ *is counterfactually fair if under any condition* $\mathbf{Z} = \mathbf{z}$*,*

$$\left|DE(\hat{y}_{s^- \to s^+}|\mathbf{z})\right| \leq \tau,$$

where $DE(\hat{y}_{s^- \to s^+}|\mathbf{z}) = P(\hat{y}_{s^+}|s^-, \mathbf{z}) - P(\hat{y}_{s^-}|s^-, \mathbf{z})$.

In above definition, $\left|DE(\hat{y}_{s^- \to s^+}|\mathbf{z})\right|$ captures the amount of unfairness or discrimination of a classifier in terms of the difference in the positive decision rate for a certain group of individuals (specified by $\mathbf{z}$) between the counterfactual world (where they had been changed to s^+) and the real world (where they are actually in s^-). If the amount of unfairness of a classifier is smaller than τ, we claim this classifier is (counterfactually) fair. Note that the first term $P(\hat{y}_{s^+}|s^-, \mathbf{z})$ has the identification issue, but the second term $P(\hat{y}_{s^-}|s^-, \mathbf{z})$ simply equals to $P(\hat{y}|s^-, \mathbf{z})$ since the intervention $do(s^-)$ makes no change to the value of S for this group. By denoting the upper and lower bounds of $P(\hat{y}_{s^+}|s^-, \mathbf{z})$ obtained in Theorem 12 as $ub(P(\hat{y}_{s^+}|s^-, \mathbf{z}))$ and $lb(P(\hat{y}_{s^+}|s^-, \mathbf{z}))$ respectively, we obtain the bounds of $DE(\hat{y}_{s^- \to s^+}|\mathbf{z})$ as follows.

Corollary 1. ***(Bounds of Counterfactual Fairness)*** *The upper and lower bounds of counterfactual fairness* $DE(\hat{y}_{s^- \to s^+}|\mathbf{z})$ *are given by*

$$ub\left(DE(\hat{y}_{s^- \to s^+}|\mathbf{z})\right) = ub\left(P(\hat{y}_{s^+}|s^-, \mathbf{z})\right) - P(\hat{y}|s^-, \mathbf{z}), \tag{6.6}$$

$$lb\left(DE(\hat{y}_{s^- \to s^+}|\mathbf{z})\right) = lb\left(P(\hat{y}_{s^+}|s^-, \mathbf{z})\right) - P(\hat{y}|s^-, \mathbf{z}). \tag{6.7}$$

Corollary 3 can facilitate the detection of unfairness from observational data. Specifically, if we have $ub\left(DE(\hat{y}_{s^- \to s^+}|\mathbf{z})\right) \leq \tau$ and $lb\left(DE(\hat{y}_{s^- \to s^+}|\mathbf{z})\right) \geq -\tau$, then it is guaranteed that τ-counterfactual fairness is satisfied. If we have $ub\left(DE(\hat{y}_{s^- \to s^+}|\mathbf{z})\right) \leq -\tau$ or

$lb\left(DE(\hat{y}_{s^-\to s^+}|\mathbf{z})\right) > \tau$, then it is guaranteed that τ-counterfactual fairness cannot be satisfied. Otherwise, it is uncertain and cannot be determined from data.

Based on Corollary 3, we then propose an efficient method for constructing counterfactually fair classifiers. Note that the bounds are consistent with identifiable situations, so the method works for both identifiable/unidentifiable situations.

We consider to construct a new decision variable $\tilde{Y}$ from $\hat{Y}$ in the causal model such that τ-counterfactual fairness regarding $\tilde{Y}$ is satisfied. The objective is to find an optimal probabilistic mapping function $P(\tilde{y}|\hat{y}, \mathsf{pa}(\hat{Y})_{\mathcal{G}})$ that minimizes the difference between Y and $\tilde{Y}$, measured by the empirical loss $\mathbb{E}_{\mathcal{D}}[\ell(Y, \tilde{Y})]$, meanwhile, the new decisions are counterfactually fair. The formulation of this optimization problem is given below.

Problem Formulation 1. *Given a dataset $\mathcal{D}$ with prediction $\hat{Y}$ made by an arbitrary classifier, we aim to learn a post-processing mapping function $P(\tilde{y}|\hat{y}, \mathsf{pa}(\hat{Y})_{\mathcal{G}})$ by solving the following optimization problem:*

$$
\begin{aligned}
&\min \qquad \mathbb{E}_{\mathcal{D}}[\ell(Y, \tilde{Y})] \\
&\textit{s.t. for any } \mathbf{z}: \\
&ub\left(DE(\tilde{y}_{s^-\to s^+}|\mathbf{z})\right) \le \tau, \quad lb\left(DE(\tilde{y}_{s^+\to s^-}|\mathbf{z})\right) \ge -\tau, \\
&\sum_{\tilde{y}} P(\tilde{y}|\hat{y}, \mathsf{pa}(\hat{Y})_{\mathcal{G}}) = 1, \quad 0 \le P(\tilde{y}|\hat{y}, \mathsf{pa}(\hat{Y})_{\mathcal{G}}) \le 1,
\end{aligned}
$$

where $\ell(Y, \tilde{Y})$ is the 0-1 loss function.

It is easy to show that Problem Formulation 1 is a linear programming problem with $P(\tilde{y}|\hat{y}, \mathsf{pa}(\hat{Y})_{\mathcal{G}})$ as variables. Note that distribution $P(\tilde{y}|\mathsf{pa}(\hat{Y})_{\mathcal{G}})$ can be obtained by $P(\tilde{y}|\mathsf{pa}(\hat{Y})_{\mathcal{G}}) = \sum_{\hat{y}} P(\hat{y}|\mathsf{pa}(\hat{Y})_{\mathcal{G}}) P(\tilde{y}|\hat{y}, \mathsf{pa}(\hat{Y})_{\mathcal{G}})$. Thus, all constraints are linear w.r.t. $P(\tilde{y}|\hat{y}, \mathsf{pa}(\hat{Y})_{\mathcal{G}})$. On the other hand, for the objective function we have

$$
\mathbb{E}_{\mathcal{D}}[\ell(Y, \tilde{Y})] = \sum_{y, \tilde{y} \in \{y^+, y^-\}} \ell(y, \tilde{y}) P(\tilde{y}, y) = 2P(\tilde{y} \neq y).
$$

And we also have

$$
\begin{aligned}
&P(\tilde{y} \neq y) = P(\hat{y} \neq y)P(\tilde{y} = \hat{y}) + P(\hat{y} = y)P(\tilde{y} \neq \hat{y}) \\
&= \sum_{\mathbf{x},s} P(\mathbf{x}, s) \left\{ P(\hat{y} \neq y|\mathbf{x}, s) \begin{bmatrix} P(\tilde{y} = y^-|\hat{y} = y^-, \mathbf{x}, s) \\ P(\hat{y} = y^-|\mathbf{x}, s) \end{bmatrix} + \begin{bmatrix} P(\tilde{y} = y^+|\hat{y} = y^+, \mathbf{x}, s) \\ P(\hat{y} = y^+|\mathbf{x}, s) \end{bmatrix} \right. \\
&\left. + P(\hat{y} = y|\mathbf{x}, s) \begin{bmatrix} P(\tilde{y} = y^+|\hat{y} = y^-, \mathbf{x}, s) \\ P(\hat{y} = y^-|\mathbf{x}, s) \end{bmatrix} + \begin{bmatrix} P(\tilde{y} = y^-|\hat{y} = y^+, \mathbf{x}, s) \\ P(\hat{y} = y^+|\mathbf{x}, s) \end{bmatrix} \right\}
\end{aligned}
$$

In the above expression, all probabilities except $P(\tilde{y}|\hat{y}, \mathbf{x}, s)$ are read from the training set $\mathcal{D}$, making it a linear expression of $P(\tilde{y}|\hat{y}, \mathbf{x}, s)$.

6.5 Experiments

We evaluate our method and compare it with previous methods on two datasets. To show the correctness of our method, we generate a synthetic dataset from a known causal model with complete knowledge in our evaluation. We also use the *Adult* dataset [68] to evaluate these methods in a real-world environment. We evaluate four methods for constructing classifiers: (1) the original learning algorithm without fairness constraints as the baseline (denoted by **BL**); (2) two methods (denoted by **A1** and **A3**) from [7] where **A1** uses non-descendants of S only for building classifiers, and **A3** presuppose the additive noise model for estimating the noise terms, which are then used for building classifiers; (3) our method (denoted as **CF**). By default, the discrimination threshold τ is set as 0.05. The datasets and implementations are available at http://tiny.cc/counterfactual-fairness.

6.5.1 Datasets

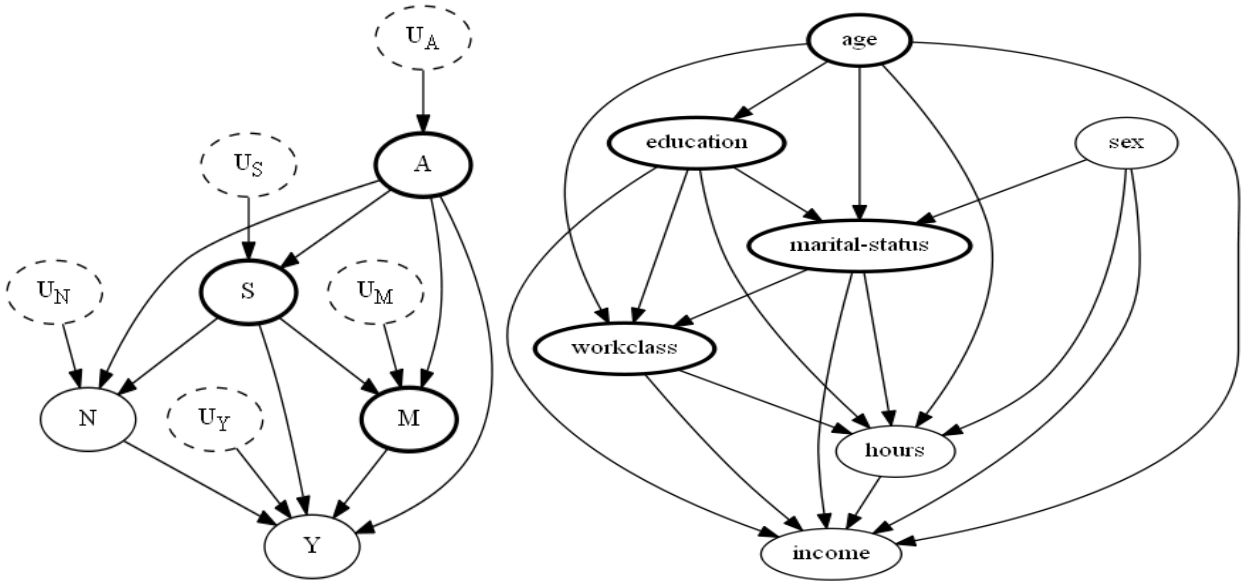
Synthetic Dataset. We manually build a causal model (where all variables are discrete) with complete knowledge of the exogenous variables and the functions (i.e., the contingency table) using Tetrad [70]. The corresponding causal graph is shown in Figure 6.2a.

This causal model consists of 5 endogenous variables, A, S, M, N, Y, and 5 independent exogenous variables, U_A, U_S, U_M, U_N, U_Y. For simplicity, all endogenous variables have two domain values and all exogenous variables have three domain values. The distributions of the exogenous variables and the deterministic functions of the endogenous variables are randomly assigned. Then, we generate 100,000 examples from this causal model and split the data into training and testing sets with a ratio of 80/20. We consider S as the sensitive attribute and Y as the decision attribute. The profile attribute set $\mathbf{Z}$ contains A, M.

***Adult* Dataset.** The *Adult* dataset is described in Appendix A.1. We select 7 attributes, binarize their domain values, and split the dataset into the training and testing sets, following the 80/20 ratio. We apply the PC algorithm implemented in Tetrad to build the causal graph while the significant threshold is set as 0.01 for conditional independence testing. We use three tiers in the partial order for temporal priority: `sex`, `age` in Tier 1, `education`, `marital-status` and `workclass` are defined in Tier 2, and `income` defined in Tier 3. The causal graph is shown in Figure 6.2b, where `sex` is considered as the sensitive attribute and `income` is the decision attribute. `age`, `education`, `marital-status`, and `workclass` are contained the profile attributes $\mathbf{Z}$.

6.5.2 Experiment on the Synthetic Dataset

Quantifying Counterfactual Fairness. According to Theorem 11, the counterfactual fairness quantity is unidentifiable in this dataset. We evaluate the bounds of counterfactual fairness using Theorem 12. The ground truth (i.e., the exact values of all counterfactual quantities) is computed by applying the Abduction-Action-Prediction method. The results are shown in Table 6.1, where the first column indicates the indices of $\mathbf{z}$'s value combinations. As can be seen, the exact values of $DE(\hat{y}_{s^- \to s^+}|\mathbf{z})$ fall into the range of our bounds for all value combinations of $\mathbf{Z}$, which validates our theorem.



(a) Causal Graph for the synthetic dataset.

(b) Causal graph for the *Adult* dataset.

Figure 6.2: Causal graphs for the synthetic dataset and the *Adult* dataset. Dashed nodes represent the exogenous variables. Bold nodes represent the profile attributes in **Z**.

# of **z**	$DE(\hat{y}_{s^- \to s^+} \mid \mathbf{z})$		
	ub	lb	$Truth$
1	0.399	0.105	0.328
2	0.471	0.177	0.467
3	0.147	-0.082	-0.038
4	0.374	0.145	0.145

Table 6.1: Bounds and ground truth of counterfactual fairness for all value combinations of **Z** using the synthetic dataset.

Building Counterfactually Fair Classifiers. We then evaluate the classifier learning methods. For the baseline method, we adopt the logistic regression (**LR**) and support vector machine (**SVM**). Then, we apply **A1**, **A3**, and **CF** on top of both classifiers. The counterfactual fairness is precisely evaluated and shown in Table 6.2 for all the methods using the Abduction-Action-Prediction method. The predictive accuracy is reported in Table 6.3. As expected, both **A1** and **CF** achieve fairness, but our method achieves higher accuracy than **A1**, implying that **A1** loses more information. On the other hand, we see that **BL** fails to achieve counterfactual fairness, because it ignores the fairness during the training. In

# of **z**	**LR**				**SVM**			
	BL	**A1**	**A3**	**CF**	**BL**	**A1**	**A3**	**CF**
1	0.000	0.000	**-0.233**	0.049	**0.114**	0.000	**0.174**	0.049
2	**1.000**	0.000	**1.000**	0.049	**0.762**	0.000	**0.648**	0.049
3	0.000	0.000	0.000	0.000	-0.021	0.000	-0.021	0.000
4	**1.000**	0.000	0.000	0.048	**1.000**	0.000	0.000	0.048

Table 6.2: Counterfactual fairness for prediction of the synthetic dataset. Values violating the threshold are highlighted in bold.

Accu. (%)	Data	**BL**	**A1**	**A3**	**CF**
LR	Train	60.103	55.760	59.433	61.987
	Test	60.421	56.563	59.713	62.512
SVM	Train	65.710	55.760	62.466	61.977
	Test	65.841	56.563	62.542	62.463

Table 6.3: Prediction accuracy for the synthetic dataset.

addition, **A3** also fails to achieve counterfactual fairness. This implies that assuming additive model may produce biased results when the underlying causal model is non-linear.

6.5.3 Experiment on the *Adult* Dataset

We evaluate the fair classifier learning methods using the *Adult* dataset. Since we do not have the ground truth, we report bounds of counterfactual fairness for different methods. Table 6.4 shows that only **A1** and **CF** can achieve counterfactual fairness for all value combinations of **Z**, but our **CF** consistently achieves higher accuracy than **A1** as shown in Table 6.5. This is as expected since **A1** is proved to be fair in [7] (and also identifiable according to Theorem 11), but will inevitably lead to lower accuracy as only S's non-descendants are used. For **BL** and **A3** in Table 6.4, either the lower bound is larger than τ or the upper bound is less than $-\tau$, indicating the τ-counterfactual fairness is not achieved.

6.6 Summary

We focused on the unidentifiability challenge when applying counterfactual fairness in practice. We decomposed the counterfactual quantities and identified the source of

	# of $\mathbf{z}$	**BL**		**A1**	**A3**		**CF**	
		ub	*lb*	*val*	*ub*	*lb*	*ub*	*lb*
LR	2	0.321	0.000	0.000	**-1.000**	**-1.000**	-0.007	-0.047
	4	0.523	0.000	0.000	**-1.000**	**-1.000**	0.038	-0.027
	13	**1.000**	**0.304**	0.000	0.000	-1.000	0.049	-0.016
	15	**1.000**	**0.398**	0.000	0.000	-1.000	0.050	-0.007
SVM	2	0.135	-0.186	0.000	**-0.186**	**-0.186**	-0.007	-0.047
	4	0.283	-0.240	0.000	**-0.240**	**-0.240**	0.038	-0.027
	13	**0.866**	**0.170**	0.000	**0.866**	**0.170**	0.049	-0.016
	15	**0.907**	**0.305**	0.000	**0.907**	**0.305**	0.050	-0.007

Table 6.4: Counterfactual fairness for prediction of the *Adult* dataset.

Accu. (%)	Data	**BL**	**A1**	**A3**	**CF**
LR	Train	77.728	67.624	74.845	70.433
	Test	77.200	66.934	73.867	69.451
SVM	Train	78.071	67.624	77.845	70.413
	Test	77.449	66.934	77.166	69.438

Table 6.5: Prediction accuracy for the *Adult* dataset.

unidentification by leveraging the counterfactual graph and c-component factorization from Pearl's framework. We then developed the criterion of identification and the upper/lower bounds for counterfactual fairness. Finally, we formulated counterfactually fair classification as a linear programming problem. Empirical evaluations showed our method is guaranteed to achieve counterfactual fairness in classification, while previous approaches either cannot achieve counterfactual fairness or suffer bad performance due to over-simplified assumptions. This work has appeared in IJCAI 2019 [8].

7 Path-specific Counterfactual Fairness

7.1 Introduction

Based on Pearl's structural causal models [54], a number of causality-based fairness notions have been proposed for capturing fairness in different situations, including total effect [4,5,30], direct/indirect discrimination [4,5,29,30], and counterfactual fairness [7,8,91,92].

One common challenge of all causality-based fairness notions is identifiability, i.e., whether they can be uniquely measured from observational data. As causality-based fairness notions are defined based on different types of causal effects, such as total effect on interventions, direct/indirect discrimination on path-specific effects, and counterfactual fairness on counterfactual effects, their identifiability depends on the identifiability of these causal effects. Unfortunately, in many situations these causal effects are in general unidentifiable, referred to as unidentifiable situations [87]. Identifiability is a critical barrier for the causality-based fairness to be applied to real applications. In previous works, simplifying assumptions are proposed to evade this problem [4, 7, 31]. However, these simplifications may severely damage the performance of predictive models. In [5] the authors propose a method to bound indirect discrimination as the path-specific effect in unidentifiable situations, and in [8] a method is proposed to bound counterfactual fairness. Nevertheless, the tightness of these methods is not analyzed. In addition, it is not clear whether these methods can be applied to other unidentifiable situations, and more importantly, a combination of multiple unidentifiable situations.

In this chapter, we propose a framework for handling different causality-based fairness notions. We first propose a general representation of all types of causal effects, i.e., the path-specific counterfactual effect, based on which we define a unified fairness notion that covers most previous causality-based fairness notions, namely the path-specific counterfactual fairness (PC fairness). We summarize all unidentifiable situations that are discovered in

the causal inference literature. Then, we develop a constrained optimization problem for bounding the PC fairness, which is motivated by the method proposed in [9] for bounding confounded causal effects. The key idea is to parameterize the causal model using so-called response-function variables, whose distribution captures all randomness encoded in the causal model, so that we can explicitly traverse all possible causal models to find the tightest possible bounds. In the experiments, we evaluate the proposed method and compare it with previous bounding methods using both synthetic and real-world datasets. The results show that our method is capable of bounding causal effects under any unidentifiable situation or combinations. When only path-specific effect or counterfactual effect is considered, our method provides tighter bounds than methods in [5] or [8]. The proposed framework settles a general theoretical foundation for causality-based fairness. We make no assumption about the hidden confounders so that hidden confounders are allowed to exist in the causal model. We also make no assumption about the data generating process and whether the observation data is generated by linear or non-linear functions would not introduce bias into our results. We only assume that the causal graph is given, which is a common assumption in structural causal models.

Relationship to other work. In [89], the author introduces the term "path-specific counterfactual fairness", which states that a decision is fair toward an individual if it coincides with the one that would have been taken in a counterfactual world in which the sensitive attribute along the unfair pathways were different. They develop a correction method called PSCF for eliminating the individual-level unfair information contained in the observations while retaining fair information. Compared to [89], we formally define a general fairness notion which, besides the individual-level fairness, is also applied to fairness in any sub-group of the population. In addition, we further consider the identifiability issue in causal inference that is inevitably brought by conditioning on the individual level. Unidentifiable situation means that there exist two causal models which exactly agree with the same observational distribution (hence cannot be distinguished using statistic methods such as

maximum likelihood), but lead to very different causal effects. In this chapter, we address various unidentifiable situations by developing a general bounding method. The authors in [93] study the conditional path-specific effect and develop a complete identification algorithm with the application to the problem of algorithmic fairness. Similar to our proposed notion, their notion is also quantified via conditional distributions over the interventional variant. However, the conditional path-specific effect generalizes the conditional causal effect, where the factual condition is assumed to be "non-contradictory" (such as age in measuring the effect of smoking on lung cancer) [87]. The path-specific counterfactual effect, on the other hand, generalizes the counterfactual effect, where the factual condition can be contradictory to the observation. Formally, in the conditional path-specific effect, the condition is performed on the pre-intervention distribution, but in the path-specific counterfactual effect, the condition is performed on the post-intervention distribution.

7.2 Preliminaries

In this chapter, we leverage the Structural Causal Model, identification of causal inference, the path-specific effect, and the counterfactual effect whose definitions can be found in Chapter 3, Chapter 4, and Chapter 6.

7.3 Path-specific Counterfactual Fairness

In this section, we define a unified fairness notion for representing different causality-based fairness notions. The key component of our notion is a general representation of causal effects. Consider an intervention on X which is transmitted along a subset of causal paths π to Y, conditioning on observation $\mathbf{O} = \mathbf{o}$. Based on that, we define path-specific counterfactual effect as follows.

Definition 8 (Path-specific Counterfactual Effect)**.** *Given a factual condition* $\mathbf{O} = \mathbf{o}$ *and a causal path set* π*, the path-specific counterfactual effect of the value change of* X *from* x_0 *to*

x_1 on $Y = y$ through π (with reference x_0) is given by

$$\text{PCE}_\pi(x_1, x_0|\mathbf{o}) = P(y_{x_1|\pi,x_0|\bar{\pi}}|\mathbf{o}) - P(y_{x_0}|\mathbf{o}).$$

In the context of fair machine learning, we use $S \in \{s^+, s^-\}$ to denote the protected attribute, $Y \in \{y^+, y^+\}$ to denote the decision, and $\mathbf{X}$ to denote a set of non-protected attributes. The underlying mechanism of the population over the space $S \times \mathbf{X} \times Y$ is represented by a causal model $\mathcal{M}$, which is associated with a causal graph $\mathcal{G}$. A historical dataset $\mathcal{D}$ is drawn from the population, which is used to construct a predictor $h : \mathbf{X}, S \to \hat{Y}$. The causal model for the population over space $S \times \mathbf{X} \times \hat{Y}$ can be considered the same as $\mathcal{M}$ except that function f_Y is replaced with a predictor h. We use Π to denote all causal paths from S to $\hat{Y}$ in the causal graph.

Then, we define the path-specific counterfactual fairness based on Definition 8.

Definition 9 (Path-specific Counterfactual Fairness (PC Fairness))**.** *Given a factual condition $\mathbf{O} = \mathbf{o}$ where $\mathbf{O} \subseteq \{S, \mathbf{X}, Y\}$ and a causal path set π, predictor $\hat{Y}$ achieves the PC fairness if $\text{PCE}_\pi(s_1, s_0|\mathbf{o}) = 0$ where $s_1, s_0 \in \{s^+, s^-\}$. We also say that $\hat{Y}$ achieves the τ-PC fairness if $\left|\text{PCE}_\pi(s_1, s_0|\mathbf{o})\right| \leq \tau$.*

We show that previous causality-based fairness notions can be expressed as special cases of the PC fairness. Their connections are summarised in Table 7.1, where π_d contains the direct edge from S to $\hat{Y}$, and π_i is a path set that contains all causal paths passing through any redlining attributes (i.e., a set of attributes in $\mathbf{X}$ that cannot be legally justified if used in decision-making). Based on whether $\mathbf{O}$ equals $\emptyset$ or not, the previous notions can be categorized into the ones that deal with the system level ($\mathbf{O} = \emptyset$) and the ones that have certain conditions ($\mathbf{O} \neq \emptyset$). Based on whether π equals Π or not, the previous notions can be categorized into the ones that deal with the total causal effect ($\pi = \Pi$), the ones that consider the direct discrimination ($\pi = \pi_d$), and the ones that consider the indirect discrimination ($\pi = \pi_i$).

Table 7.1: Connection between previous fairness notions and PC fairness

Description	References	Relating to PC fairness
Total effect	[4, 30]	$\mathbf{O} = \emptyset$ and $\pi = \Pi$
(System) Direct discrimination	[4, 29, 30]	$\mathbf{O} = \emptyset$ or $\{S\}$ and $\pi = \pi_d = \{S \to \hat{Y}\}$
(System) Indirect discrimination	[4, 29, 30]	$\mathbf{O} = \emptyset$ or $\{S\}$ and $\pi = \pi_i \subset \Pi$
Individual direct discrimination	[26]	$\mathbf{O} = \{S, \mathbf{X}\}$ and $\pi = \pi_d = \{S \to \hat{Y}\}$
Group direct discrimination	[28]	$\mathbf{O} = \mathbf{Q} = \mathsf{PA}_Y \backslash \{S\}$ and $\pi = \pi_d = \{S \to \hat{Y}\}$
Counterfactual fairness	[7, 8, 92]	$\mathbf{O} = \{S, \mathbf{X}\}$ and $\pi = \Pi$
Counterfactual error rate	[91]	$\mathbf{O} = \{S, Y\}$ and $\pi = \pi_d$ or π_i

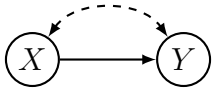


Figure 7.1: The "bow" graph.

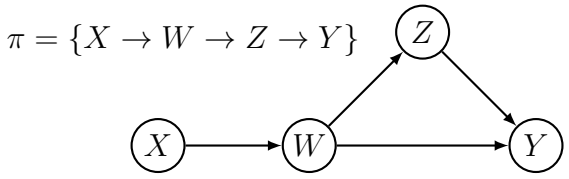


Figure 7.2: The "kite" graph.

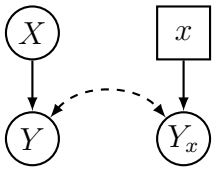


Figure 7.3: The "w" graph.

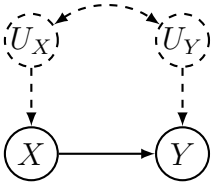


Figure 7.4: The causal graph for a semi-Markovian model.

In addition to unifying the existing notions, the notion of PC fairness also resolves new types of fairness that the previous notions cannot do. One example is individual indirect discrimination, which means discrimination along the indirect paths for a particular individual. Individual indirect discrimination has not been studied yet in the literature, probably due to the difficulty in definition and identification. However, it can be directly defined and analyzed using PC fairness by letting $\mathbf{O} = \{S, \mathbf{X}\}$ and $\pi = \pi_i$.

7.4 Measuring Path-specific Counterfactual Fairness

In this section, we develop a general method for bounding the path-specific counterfactual effect in any unidentifiable situation. In the causal inference field, researchers have studied the reasons for unidentifiability under different cases. When $\mathbf{O} = \emptyset$ and $\pi \subset \Pi$,

the reason for unidentifiability can be the existence of the "kite" graph (see Figure 7.2) in the causal graph [2]. When $\mathbf{O} \neq \emptyset$ and $\pi = \Pi$, the reason for unidentifiability can be the existence of the "w" graph (see Figure 7.3) [94]. In any situation, as long as there exists a "hedge" graph (where the simplest case is the "bow" graph as shown in Figure 7.1), then the causal effect is unidentifiable [87]. Obviously, all above unidentifiable situations can exist in the path-specific counterfactual effect.

Our method is motivated by [9] which formulates the bounding problem as a constrained optimization problem. The general idea is to parameterize the causal model and use the observational distribution $P(\mathbf{V})$ to impose constraints on the parameters. Then, the path-specific counterfactual effect of interest is formulated as an objective function of maximization or minimization for estimating its upper or lower bound. The bounds are guaranteed to be tight as we traverse all possible causal models when solving the optimization problem. Thus, a byproduct of the method is a unique estimation of the path-specific counterfactual effect in the identifiable situation.

For presenting our method, we first introduce a key concept called the response-function variable.

7.4.1 Response-function Variable

Response-function variables are proposed in [9] for parameterizing the causal model. Consider an arbitrary endogenous variable denoted by $V \in \mathbf{V}$, its endogenous parents denoted by PA_V, its exogenous parents denoted by U_V, and its associated structural function in the causal model denoted by $v = f_V(\mathsf{pa}_V, u_V)$. In general, U_V can be a variable of any type with any domain size, and f_V can be any function, making the causal model very difficult to be handled. However, we can note that, for each particular value u_V of U_V, the functional mapping from PA_V to V is a particular deterministic response function. Thus, we can map each value of U_V to a deterministic response function. Although the domain size of U_V is unknown which might be very large or even infinite, the number of different deterministic

response functions is known and limited, given the domain sizes of PA_V and V. This means that the domain of U_V can be divided into several equivalent regions, each corresponding to the same response function. As a result, we can transform the original non-parameterized structural function to a limited number of parameterized functions.

Formally, we represent equivalent regions of each endogenous variable V by the *response-function variable* $R_V = \{0, \cdots, N_V - 1\}$ where $N_V = |V|^{|\mathsf{PA}_V|}$ is the total number of different deterministic response functions mapping from PA_V to V ($N_V = |V|$ if V has no parent). Each value r_V represents a pre-defined response function. We also denote the mapping from U_V to R_V as $r_V = \ell_V(u_V)$. Then, for any $f_V(\mathsf{pa}_V, u_V)$, it can be re-formulated as

$$f_V(\mathsf{pa}_V, u_V) = f_V(\mathsf{pa}_V, \ell_V^{-1}(r_V)) = f_V \circ \ell_V^{-1}(\mathsf{pa}_V, r_V) = g_V(\mathsf{pa}_V, r_V),$$

where g_V is the composition of f_V and ℓ_V^{-1}, and denotes the response functions represented by r_V. We denote the set of all response-function variables by $\mathbf{R} = \{R_V : V \in \mathbf{V}\}$.

Next, we show how joint distribution $P(\mathbf{v})$ can be expressed as a linear function of $P(\mathbf{r})$. According to [67], $P(\mathbf{v})$ can be expressed as the summation over the probabilities of certain values $\mathbf{u}$ of $\mathbf{U}$ that satisfy following corresponding requirements: for each $V \in \mathbf{V}$, we must have $f_V(\mathsf{pa}_V, u_V) = v$ where v, pa_V are specified by $\mathbf{v}$ and u_V is specified by $\mathbf{u}$. In other words, denoting by $V(\mathbf{u})$ the value that V would obtain if $\mathbf{U} = \mathbf{u}$, we have $P(\mathbf{v}) = \sum_{\mathbf{u}:\mathbf{V}(\mathbf{u})=\mathbf{v}} P(\mathbf{u})$. Then, by mapping from $\mathbf{U}$ to $\mathbf{R}$, we accordingly obtain $P(\mathbf{v}) = \sum_{\mathbf{r}:\mathbf{V}(\mathbf{r})=\mathbf{v}} P(\mathbf{r})$, where for each $V \in \mathbf{V}$, $V(\mathbf{r}) = v$ means that $g_V(\mathsf{pa}_V, r_V) = v$. As a result, by defining an indicator function

$$\mathbb{I}(v; \mathsf{pa}_V, r_V) = \begin{cases} 1 & \text{if } g_V(\mathsf{pa}_V, r_V) = v, \\ 0 & \text{otherwise,} \end{cases}$$

we obtain

$$P(\mathbf{v}) = \sum_{\mathbf{r}} P(\mathbf{r}) \prod_{V \in \mathbf{V}} \mathbb{I}(v; \mathsf{pa}_V, r_V), \tag{7.1}$$

which is a linear expression of $P(\mathbf{r})$.

Example 1. *Consider the causal graph shown in Figure 7.4 with two endogenous variables X and Y, and two exogenous variables U_X and U_Y with unknown domains. Assume that both X and Y are binary, i.e., $X \in \{x_0, x_1\}$ and $Y \in \{y_0, y_1\}$, and denote their response variables as R_X and R_Y. For Y, since there are a total number of $2^2 = 4$ response functions, response-function variable R_Y and response function g_Y can be defined as follows:*

$$r_Y = \ell_Y(u_Y) = \begin{cases} 0 \text{ if } f_Y(x_0, u_Y) = y_0, f_Y(x_1, u_Y) = y_0; \\ 1 \text{ if } f_Y(x_0, u_Y) = y_0, f_Y(x_1, u_Y) = y_1; \\ 2 \text{ if } f_Y(x_0, u_Y) = y_1, f_Y(x_1, u_Y) = y_0; \\ 3 \text{ if } f_Y(x_0, u_Y) = y_1, f_Y(x_1, u_Y) = y_1. \end{cases} \quad g_Y(x, r_Y) = \begin{cases} y_0 \text{ if } r_Y = 0; \\ y_0 \text{ if } x = x_0, r_Y = 1; \\ y_1 \text{ if } x = x_1, r_Y = 1; \\ y_1 \text{ if } x = x_0, r_Y = 2; \\ y_0 \text{ if } x = x_1, r_Y = 2; \\ y_1 \text{ if } r_Y = 3. \end{cases}$$

Similarly, response-function variable R_X and response function g_X can be defined as

$$r_X = \ell_X(u_X) = \begin{cases} 0 \quad \text{if } f_X(u_X) = x_0; \\ 1 \quad \text{if } f_X(u_X) = x_1. \end{cases} \qquad g_X(r_X) = \begin{cases} x_0 \quad \text{if } r_X = 0; \\ x_1 \quad \text{if } r_X = 1. \end{cases}$$

As a result, the joint distribution over X, Y is given by

$$P(x, y) = \sum_{r_X, r_Y} P(r_X, r_Y)\mathbb{I}(x; r_X)\mathbb{I}(y; x, r_Y).$$

7.4.2 Expressing Path-specific Counterfactual Fairness

For bounding the path-specific counterfactual effect, i.e., $\text{PCE}_\pi(s_1, s_0|\mathbf{o}) = P(\hat{y}_{s_1|\pi, s_0|\bar{\pi}}|\mathbf{o}) - P(\hat{y}_{s_0}|\mathbf{o})$, we also apply response-function variables to express it. We focus on the expression of $P(\hat{y}_{s_1|\pi, s_0|\bar{\pi}}|\mathbf{o})$, and the expression of $P(\hat{y}_{s_0}|\mathbf{o})$ can be similarly obtained as a simpler case. Similar to the previous section, we first express $P(\hat{y}_{s_1|\pi, s_0|\bar{\pi}}|\mathbf{o})$ as the summation over the probabilities of certain values of $\mathbf{U}$ that satisfy corresponding requirements. However, as described below, the requirements are much more complicated than previous ones due to the

integration of intervention, path-specific effect, and counterfactual.

Firstly, since the path-specific counterfactual effect is under a factual condition $\mathbf{O} = \mathbf{o}$, values $\mathbf{u}$ must satisfy that $\mathbf{O}(\mathbf{u}) = \mathbf{o}$, i.e., for each $O \in \mathbf{O}$, we must have $f_O(\mathsf{pa}_O, u_O) = o$. Secondly, the path-specific counterfactual effect is transmitted only along some path set π. According to [5], for the variables of $\mathbf{X}$ that lie on both π and $\bar{\pi}$, referred to as *witness variables/nodes* [2], we need to consider two sets of values, one obtained by treating them on π and the other obtained by treating them on $\bar{\pi}$. Formally, non-protected attributes $\mathbf{X}$ are divided into three disjoint sets. We denote by $\mathbf{W}$ the set of witness variables, denote by $\mathbf{A}$ the set of non-witness variables on π, and denote by $\mathbf{B}$ the set of non-witness variables on $\bar{\pi}$. A simple example is given in Figure 7.5. We denote the interventional variant of $\mathbf{A}$ by $\mathbf{A}_{s_1|\pi}$, the interventional variant of $\mathbf{B}$ by $\mathbf{B}_{s_0|\bar{\pi}}$, the interventional variant of $\mathbf{W}$ treated on π by $\mathbf{W}_{s_1|\pi}$, and the interventional variant of $\mathbf{W}$ treated on $\bar{\pi}$ by $\mathbf{W}_{s_0|\bar{\pi}}$. Then, $P(\hat{y}_{s_1|\pi,s_0|\bar{\pi}}|\mathbf{o})$ can be written as

$$P(\hat{y}_{s_1|\pi,s_0|\bar{\pi}}|\mathbf{o}) = \sum_{\mathbf{a},\mathbf{b},\mathbf{w}_1,\mathbf{w}_0} P(\hat{Y}_{s_1|\pi,s_0|\bar{\pi}} = y, \mathbf{A}_{s_1|\pi} = \mathbf{a}, \mathbf{B}_{s_0|\bar{\pi}} = \mathbf{b}, \mathbf{W}_{s_1|\pi} = \mathbf{w}_1, \mathbf{W}_{s_0|\bar{\pi}} = \mathbf{w}_0 \mid \mathbf{o}).$$

To obtain the above joint distribution, in addition to $\mathbf{O}(\mathbf{u}) = \mathbf{o}$, values $\mathbf{u}$ must also satisfy that:

1. $\mathbf{A}_{s_1|\pi}(\mathbf{u}) = \mathbf{a}$, which means for each $A \in \mathbf{A}$, we must have $f_A(\mathsf{pa}_A^1, u_A) = a$, where pa_A^1 means that if PA_A contains S or any witness node W, its value is specified by s_1 or w_1 if edge $S/W \to Y$ belongs to a path in π, and specified by s_0 or w_0 otherwise;

2. $\mathbf{B}_{s_0|\bar{\pi}}(\mathbf{u}) = \mathbf{b}$, which means for each $B \in \mathbf{B}$, we must have $f_B(\mathsf{pa}_B^0, u_B) = b$, where pa_B^0 means that if PA_B contains S or any witness node W, its value is specified by s_0 or w_0;

3. $\mathbf{W}_{s_1|\pi}(\mathbf{u}) = \mathbf{w}_1$, which means for each $W \in \mathbf{W}$, we must have $f_W(\mathsf{pa}_W^1, u_W) = w_1$;

4. $\mathbf{W}_{s_0|\pi}(\mathbf{u}) = \mathbf{w}_0$, which means for each $W \in \mathbf{W}$, we must have $f_W(\mathsf{pa}_W^0, u_W) = w_0$.

Then, by mapping from $\mathbf{U}$ to $\mathbf{R}$, we can obtain the requirements for $\mathbf{R}$ accordingly. Finally, denoting the values of $\mathbf{R}$ that satisfy $\mathbf{O}(\mathbf{r}) = \mathbf{o}$ by $\mathbf{r_o}$, we obtain

$$P(\hat{y}_{s_1|\pi,s_0|\bar{\pi}}|\mathbf{o}) = \sum_{\substack{\mathbf{a},\mathbf{b},\mathbf{w}_1 \\ \mathbf{w}_0,\mathbf{r}\in\mathbf{r_o}}} \left[\begin{array}{c} \frac{P(\mathbf{r})}{P(\mathbf{o})}\mathbb{I}(\hat{y};\mathsf{pa}^1_{\hat{Y}},r_{\hat{Y}}) \prod\limits_{A\in\mathbf{A}} \mathbb{I}(a;\mathsf{pa}^1_A,r_A) \prod\limits_{B\in\mathbf{B}} \mathbb{I}(b;\mathsf{pa}^0_B,r_B) \\ \prod\limits_{W\in\mathbf{W}} \mathbb{I}(w_1;\mathsf{pa}^1_W,r_W)\mathbb{I}(w_0;\mathsf{pa}^0_W,r_W) \end{array} \right], \tag{7.2}$$

which is still a linear expression of $P(\mathbf{r})$.

Similarly, we can obtain

$$P(\hat{y}_{s_0}|\mathbf{o}) = \sum_{\mathbf{v}',\mathbf{r}\in\mathbf{r_o}} \frac{P(\mathbf{r})}{P(\mathbf{o})}\mathbb{I}(\hat{y};\mathsf{pa}_{\hat{Y}},r_{\hat{Y}}) \prod_{V\in\mathbf{V}'} \mathbb{I}(v;\mathsf{pa}_V,r_V), \tag{7.3}$$

where $\mathbf{V}' = \mathbf{V}\backslash\{S,Y\}$.

Example 2. *Consider causal graphs shown in Figures 7.1, 7.2, 7.3 and following unidentifiable causal effects: total causal effect* $\mathrm{TCE}(x_1,x_0)$ *in Figure 7.1, path-specific effect* $\mathrm{PE}_\pi(x_1,x_0)$ *in Figure 7.2, and counterfactual effect* $\mathrm{CE}(x_1,x_0|x_0,y_0)$ *in Figure 7.3. By similarly defining response functions as in Example 1, for Figure 7.1 with* $\mathbf{R} = \{R_X,R_Y\}$*, we have*

$$\mathrm{TCE}(x_1,x_0) = \sum_{r_X,r_Y} P(r_X,r_Y)\mathbb{I}(y;x_1,r_Y) - \sum_{r_X,r_Y} P(r_X,r_Y)\mathbb{I}(y;x_0,r_Y),$$

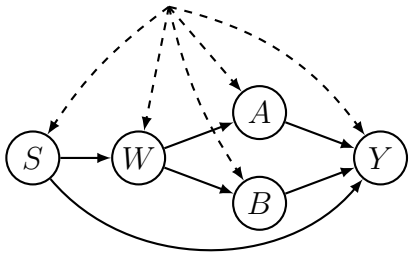
for Figure 7.2 with $\mathbf{R} = \{R_X,R_W,R_Z,R_Y\}$*, we have*

$$\begin{aligned}\mathrm{PE}_\pi(x_1,x_0) = &\sum_{z,w_1,w_0,\mathbf{r}} P(\mathbf{r})\mathbb{I}(y;z,w_0,r_Y)\mathbb{I}(z;w_1,r_Z)\mathbb{I}(w_1;x_1,r_W)\mathbb{I}(w_0;x_0,r_W) \\ &- \sum_{z,w,\mathbf{r}} P(\mathbf{r})\mathbb{I}(y;z,w,r_Y)\mathbb{I}(z;w,r_Z)\mathbb{I}(w;x_0,r_W),\end{aligned}$$

for Figure 7.3 with $\mathbf{R} = \{R_X,R_Y\}$*, we have*

$$\mathrm{CE}(x_1,x_0) = \sum_{r_X,r_Y\in\mathbf{r_o}} \frac{P(r_X,r_Y)}{P(x_0,y_0)}\mathbb{I}(y;x_1,r_Y) - \sum_{r_X,r_Y\in\mathbf{r_o}} \frac{P(r_X,r_Y)}{P(x_0,y_0)}\mathbb{I}(y;x_0,r_Y).$$

Note that in Figures 7.1, the total causal effect is identifiable if U_X and U_Y are independent. This is reflected in our formulation such that when R_X and R_Y are independent, we have $P(y_{x_1}) = \sum_{r_X, r_Y} P(r_X)P(r_Y)\mathbb{I}(y; x_1, r_Y) = P(y|x_1)$, which can be directly measured from observational data. Similar phenomenons can be observed in other identifiable situations.



$$\pi = \{S \to W \to A \to \hat{Y}, \\ S \to \hat{Y}\}$$

Figure 7.5: A causal graph with unidentifiable path-specific counterfactual fairness.

Example 3. *Consider a causal graph shown in Figure 7.5, and the path-specific counterfactual effect* $\text{PCE}_\pi(s_1, s_0|\mathbf{o})$ *where* $\pi = \{S \to \hat{Y}, S \to W \to A \to \hat{Y}\}$ *and* $\mathbf{o} = \{s_0, w', a', b'\}$. *Any pair of exogenous variables can be correlated. Response-function variables are given by* $\mathbf{R} = \{R_S, R_W, R_A, R_B, R_{\hat{Y}}\}$. *By similarly defining response functions as in Example 1, we can obtain*

$$P(\hat{y}_{s_1|\pi, s_0|\bar{\pi}}|\mathbf{o}) = \sum_{\substack{a,b,w_1,w_0 \\ \mathbf{r}\in\mathbf{r_o}}} \frac{P(\mathbf{r})}{P(\mathbf{o})}\mathbb{I}(\hat{y}; a, b, s_1, r_{\hat{Y}})\mathbb{I}(a; w_1, r_A)\mathbb{I}(b; w_0, r_B)\mathbb{I}(w_1; s_1, r_W)\mathbb{I}(w_0; s_0, r_W),$$

and

$$P(\hat{y}_{s_0}|\mathbf{o}) = \sum_{a,b,w,\mathbf{r}\in\mathbf{r_o}} \frac{P(\mathbf{r})}{P(\mathbf{o})}\mathbb{I}(\hat{y}; a, b, s_0, r_{\hat{Y}})\mathbb{I}(a; w, r_A)\mathbb{I}(b; w, r_A)\mathbb{I}(w; s_0, r_W).$$

7.4.3 Bounding Path-specific Counterfactual Fairness

In above two sections we express both joint distribution $P(\mathbf{v})$ and the path-specific counterfactual effect as linear functions of $P(\mathbf{r})$. All causal models (represented by different $P(\mathbf{r})$) that agree with the distribution of observational data $\mathcal{D}$ cannot be distinguished and should be considered in bounding PC fairness. Therefore, finding the lower or upper bound

of the path-specific counterfactual effect is equivalent to finding the $P(\mathbf{r})$ that minimizes or maximizes the path-specific counterfactual effect, subject to that the derived joint distribution $P(\mathbf{v})$ agrees with the observational distribution $P(\mathcal{D})$. This fact results in the following linear programming problem for deriving the lower/upper bound of path-specific counterfactual effect.

$$\begin{aligned} \min/\max \quad & P(\hat{y}_{s_1|\pi,s_0|\bar{\pi}}|\mathbf{o}) - P(\hat{y}_{s_0}|\mathbf{o}), \\ \text{s.t.} \quad & P(\mathbf{V}) = P(\mathcal{D}), \quad \sum_{\mathbf{r}} P(\mathbf{r}) = 1, \quad P(\mathbf{r}) \geq 0, \end{aligned} \tag{7.4}$$

where $P(\hat{y}_{s_1|\pi,s_0|\bar{\pi}}|\mathbf{o})$ is given by Eq. (7.2), $P(\hat{y}_{s_0}|\mathbf{o})$ is given by Eq. (7.3), and $P(\mathbf{v})$ is given by Equation (7.1).

The lower and upper bounds derived by solving the above optimization problem is guaranteed to be the tightest, since the response function is an equivalent mapping that covers all possible causal models thus we can explicitly traverse all possible causal models.

We use the derived bounds for examining τ-PC fairness: if the upper bound is less than τ and the lower bound is greater than $-\tau$, then τ-PC fairness must be satisfied; if the upper bound is less than $-\tau$ or the lower bound is greater than τ, τ-PC fairness must not be satisfied; otherwise, it is uncertain and cannot be determined from data.

7.5 Experiments

Datasets. For synthetic datasets, we manually build a causal model with complete knowledge of exogenous variables and equations using Tetrad [70] according to the causal graphs. The causal model consists of 4 endogenous variables, S, W, A, $\hat{Y}$, all of which have two domain values. Then, we consider two versions of the causal model: (1) we assume a shared exogenous variables, i.e., a hidden confounder, with 100 domain values (the causal graph is shown in Figure 7.6); (2) we assume all exogenous variables are mutually independent as shown in Figure 7.7. The distribution of exogenous variables and structural equations of

endogenous variables are randomly assigned. Finally, we generate two datasets from each version of the causal model, denoted by $\mathcal{D}_1$ and $\mathcal{D}_2$ respectively.

For the real-world dataset, we adopt the *Adult* dataset described in Appendix A.1, we select 7 attributes, binarize their values, and build the causal graph. Fairness threshold τ is set to 0.1. The datasets and implementation are available at http://tiny.cc/pc-fairness-code.

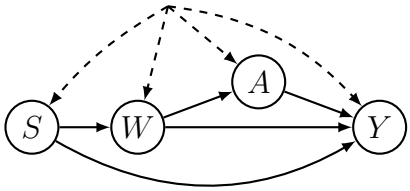


Figure 7.6: The causal graph for the synthetic dataset $\mathcal{D}_1$.

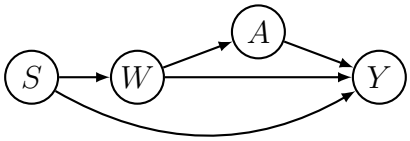


Figure 7.7: The causal graph for the synthetic dataset $\mathcal{D}_2$.

Bounding Path-specific Counterfactual Fairness. We use $\mathcal{D}_1$ to validate our method in Eq. (7.4) for bounding $\text{PCE}_\pi(s^+, s^-|\mathbf{o})$ where $\mathbf{O} = \{S, W, A\}$ and $\pi = \{S \to W \to A \to \hat{Y}, S \to \hat{Y}\}$. The ground truth can be computed by exactly executing the intervention under given conditions using the complete causal model. The results are shown in Table 7.2, where the first column indicates the indices of $\mathbf{o}$'s value combinations. As can be seen, the true values of $\text{PCE}_\pi(s^+, s^-|\mathbf{o})$ fall into the range of our bounds for all value combinations of $\mathbf{O}$, which validates our method.

Table 7.2: Bounds and ground truth of PC fairness on $\mathcal{D}_1$.

# of $\mathbf{o}$	$\text{PCE}_\pi(s^+, s^-\|\mathbf{o})$		
	lb	ub	$Truth$
1	-0.4548	0.5452	0.1507
2	-0.5565	0.4435	-0.0928
3	-0.5065	0.4935	0.0561
4	-0.4598	0.5402	0.0548

Comparing with previous bounding methods. We use $\mathcal{D}_2$ to compare with the previous methods [5, 8] which are derived under the Markovian assumption. We compare with [5] for bounding $\text{PE}_\pi(s^+, s^-)$ with $\pi = \{S \to W \to A \to \hat{Y}, S \to \hat{Y}\}$. We also compare with [8] for bounding $\text{CE}(s^+, s^-|\mathbf{o})$ with $\mathbf{O} = \{S, W, A\}$. The results are shown in Table 7.3

where the bold indicates that our method makes different judgments on discrimination detection due to the tighter bounds. As can be seen, our method achieves much tighter bounds than previous methods, which can be used to examine fairness more accurately. For example, when measuring indirect discrimination using $\text{PE}_\pi(s^+, s^-)$ (Row 1 in Table 7.3), it is uncertain for [5] since the lower and upper bounds are -0.2605 and 0.2656, but our method can guarantee that the decision is discriminatory as the lower bound 0.1772 is larger than $\tau = 0.1$. As another example, when measuring counterfactual fairness of the 2nd groups of $\mathbf{o}$ using $\text{CE}(s^+, s^-|\mathbf{o})$ (Row 3 in Table 7.3), the method in [8] is uncertain since the lower and upper bounds are $-0.4383, -0.0212$ but our method can guarantee that the decision is fair due to the range of $[-0.0783, -0.0212]$.

We also use the *Adult* datset to compare with the method in [8] for bounding $\text{CE}(s^+, s^-|\mathbf{o})$ with $\mathbf{O} = \{\texttt{age}, \texttt{edu}, \texttt{marital-status}\}$ and obtain similar results, which are shown in Table 7.4.

Table 7.3: Comparison with existing methods in [5, 8] on $\mathcal{D}_2$.

	$\mathbf{o}$	*Truth*	Previous methods		Our method	
			lb	*ub*	*lb*	*ub*
PSF	*N/A*	**0.1793**	**-0.2605**	**0.2656**	**0.1772**	**0.1836**
CF	1	0.3438	0.0878	0.5049	0.0878	0.5049
	2	**-0.0557**	**-0.4383**	**-0.0212**	**-0.0783**	**-0.0212**
	3	**0.2318**	**-0.1192**	**0.2979**	**0.1282**	**0.2847**
	4	0.0800	-0.2101	0.2070	0.0110	0.1499

Table 7.4: Comparison with the existing method in [8] on the *Adult* dataset.

# of $\mathbf{o}$	Method in [8]		Our Method	
	lb	*ub*	*lb*	*ub*
0	0.0541	0.2946	0.1498	0.1944
1	-0.1314	0.1091	-0.1314	0.1091
2	0.1878	0.3210	0.2507	0.2890
3	-0.0356	0.0976	-0.0356	0.0976
4	0.1676	0.5289	0.4419	0.5289
5	-0.1634	0.1979	-0.0731	0.1979
6	0.1290	0.4689	0.3942	0.4689
7	-0.1808	0.1591	0.0014	0.1591

7.6 Summary

In this chapter, we developed a general framework for measuring causality-based fairness. We proposed a unified definition that covers most of previous causality-based fairness notions, namely the path-specific counterfactual fairness (PC fairness). Then, we formulated a linear programming problem to bound PC fairness which can produce the tightest possible bounds. Experiments using synthetic and real-world datasets showed that, our method can bound causal effects under any unidentifiable situation or combinations and achieves tighter bounds than previous methods. This work has been published in NeurIPS 2019 [10].

8 Convexity and Bounds of Fairness-aware Classification

8.1 Introduction

Fairness-aware classification is receiving increasing attention in the machine learning fields. Since the classification models seek to maximize the predictive accuracy, individuals may get unwanted digital bias when the models are deployed for making predictions. As fairness becomes a more and more important requirement in machine learning, it is imperative to ensure that the learned classification models can strike a balance between accurate and fair predictions. Previous works on this topic can be mainly categorized into two groups: the in-processing methods which incorporate the fairness constraints into the classic classification models (e.g., [13, 42, 43, 95, 96]), and the pre/post-processing methods which modify the training data and/or derive fair predictions based on the potentially unfair predictions made by the classifier (e.g., [4, 12, 37, 73, 79]). In this work, we focus on the in-processing methods.

Very recently, several works have been proposed for formulating the fairness-aware classification as constrained optimization problems [13, 42, 43, 95–98]. Generally, they aim to minimize a loss function subject to certain fairness constraints, e.g., demographic parity (i.e., the difference of the positive predictions between the sensitive group and non-sensitive group) is less than some threshold. However, most quantitative fairness metrics such as demographic parity [19], mistreatment parity [13], etc., are non-convex due to the use of the indicator function, thus making the optimization problem intractable. A widely-used strategy to achieve convexity in optimization is to adopt surrogate functions for both loss function and constraints. In [43], the authors applied the linear surrogate functions to non-convex risk difference as the decision boundary fairness for margin-based classifiers. Similarly in [95], a convex constraint is derived from the risk difference. One challenge is that, when surrogate functions are used to convert non-convex functions to convex functions, estimation errors must exist due to the difference between the surrogate function and the original non-

convex function. Thus, achieving the fairness constraints represented by surrogate functions does not necessarily guarantee achieving the real fairness criterion. Hence, how to achieve fairness-aware classification via constrained optimization still remains an open problem.

In this chapter, we propose a general framework for fairness-aware classification which addresses the gap incurred by the estimation errors due to the surrogate function. The framework can formulate various commonly-used fairness metrics (risk difference [20], risk ratio [20], equal odds [12], etc.) as convex constraints that are then directly incorporated into classic classification models. Within the framework, we first present a constraint-free criterion (derived from the training data) which ensures that any classifier learned from the data will guarantee to be fair in terms of the specified fairness metric. Thus, when the criterion is satisfied, there is no need to add any fairness constraint into optimization for learning fair classifiers. When the criterion is not satisfied, we need to learn fair classifiers by solving the constrained optimization problems. To connect the surrogated fairness constraints to the original non-convex fairness metric, we further derive the lower and upper bounds of the real fairness measure based on the surrogate function, and develop the refined fairness constraints. This means that, if the refined constraints are satisfied, then it is guaranteed that the real fairness measure is also bounded within the given interval. The bounds work for any surrogate function that is convex and differentiable at zero with the derivative larger than zero. In the experiments, we evaluate our method and compare with existing works using the real-world datasets. The results demonstrate the correctness of the constraint-free criterion and the superiority of our method over existing ones in terms of achieving fairness and retaining prediction accuracy.

8.2 Fairness-aware Classification

In this section we present our fairness-aware classification framework. We first introduce the unconstrained optimization formulation for the classic classification models as proposed in [99], and then present our constrained optimization formulation for fairness-aware classification.

Throughout the chapter, we use the vector $\mathbf{X} \in \mathcal{X}$ to denote the features used in classification, and $Y \in \mathcal{Y} = \{-1, 1\}$ to denote the binary label. We denote the sensitive attribute by S, assuming that it is associated with two values: sensitive group s^- and non-sensitive group s^+. The training data $\mathbb{D} = \{(\mathbf{x}_i, s_i, y_i)\}_{i=1}^N$ is a sample drawn from an unknown but fixed distribution.

8.2.1 Classification Problem

The learning goal of classification is to find a classifier: $f\colon \mathcal{X} \mapsto \mathcal{Y}$ that minimizes the average of the classification loss (a.k.a the empirical loss), given by

$$\mathbb{L}(f) = \mathbb{E}_{\mathbf{X},Y}[\mathbb{1}_{f(\mathbf{x})\neq y}], \tag{8.1}$$

where $\mathbb{1}_{[\cdot]}$ is an indicator function. The classification problem can then be formulated as an optimization problem:

$$\min_{f\in\mathcal{F}} \mathbb{L}(f) = \min_{f\in\mathcal{F}} \mathbb{E}_{\mathbf{X},Y}[\mathbb{1}_{f(\mathbf{x})\neq y}].$$

Directly solving this optimization problem is intractable since the objective function is non-convex [99]. For efficient computation, another predictive function h is adopted which is performed in real number domain $\mathcal{R}$, i.e., $h\colon \mathcal{X} \mapsto \mathcal{R}$. By letting $f = \text{sign}(h)$, the empirical loss can be reformulated as

$$\mathbb{L}(f) = \mathbb{L}(h) = \mathbb{E}_{\mathbf{X},Y}\big[\mathbb{1}_{\text{sign}\big(h(\mathbf{x})\big)\neq y}\big] = \mathbb{E}_{\mathbf{X}}\big[Pr(Y=1|\mathbf{x})\mathbb{1}_{h(\mathbf{x})<0} + Pr(Y=-1|\mathbf{x})\mathbb{1}_{h(\mathbf{x})>0}\big]. \tag{8.2}$$

If we replace the indicator function (a.k.a 0-1 loss function) with a convex surrogate

function ϕ, the empirical loss can be rewritten as

$$\mathbb{L}_\phi(h) = \mathbb{E}_{\mathbf{X}}\big[Pr(Y=1|\mathbf{x})\phi\big(h(\mathbf{x})\big) + \big(1 - Pr(Y=1|\mathbf{x})\big)\phi\big(-h(\mathbf{x})\big)\big],$$

which is known as the ϕ-loss, and the optimization problem is reformulated as $\min_{h\in\mathcal{H}} \mathbb{L}_\phi(h)$. In the past decades, a number of surrogate loss functions have been proposed and well studied, such as the hinges loss, the square loss, the logistic loss, the exponential loss, etc.

8.2.2 Fairness-aware Classification Problem

The fairness-aware classification aims to find a classifier that minimizes the empirical loss while satisfying certain fairness constraints. Several fairness notions or definitions are proposed in the literature, such as demographic parity [19], mistreatment parity [13], etc.

Demographic parity is the most widely-used fairness notion in the fairness-aware learning field. It requires the decision made by the classifier is independent to the sensitive attribute, such as sex or race. Usually, demographic parity is quantified with regard to risk difference [20], i.e., the difference of the positive predictions between the sensitive group and non-sensitive group. For example, in the context of hiring, risk difference can be given by the probability difference of being predicted to be hired between male applicants and female applicants. Using the same language as that in the previous subsection, the risk difference produced by a classifier f is expressed as

$$\mathbb{RD}(f) = \mathbb{E}_{\mathbf{X}|S=s^+}[\mathbb{1}_{f(\mathbf{x})=1}] - \mathbb{E}_{\mathbf{X}|S=s^-}[\mathbb{1}_{f(\mathbf{x})=1}]. \tag{8.3}$$

As a quantitative metric, we say that classifier f is considered as fair if $|\mathbb{RD}(f)| \leq \tau$, where τ is the user-defined threshold. For instance, the 1975 British legislation for sex discrimination sets $\tau = 0.05$. By directly incorporating the risk difference into the optimization problem, we formulate the fair classification problem as follows.

Problem Formulation 2. *The goal of the fairness-aware classification is to find a classifier*

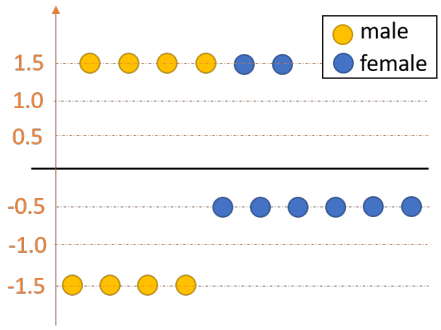
f that minimizes the loss $\mathbb{L}(f)$ while satisfying fairness constraint $|\mathbb{RD}(f)| \leq \tau$. It can be approached by solving the following constrained optimization problem

$$\begin{aligned} \min_{f \in \mathcal{F}} \quad & \mathbb{L}(f) \\ \textit{subject to} \quad & \mathbb{RD}(f) \leq \tau, \quad -\mathbb{RD}(f) \leq \tau, \end{aligned} \tag{8.4}$$

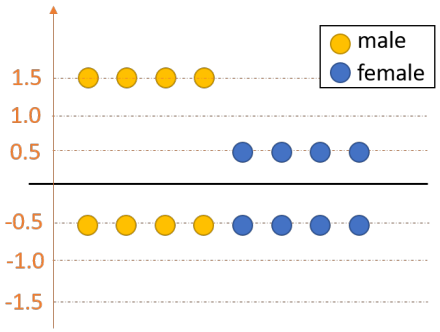
where $\mathbb{L}(f)$ and $\mathbb{RD}(f)$ are defined in Eq. (8.1) *and Eq.* (8.3).

Obviously, solving the above problem is computationally intractable, since both $\mathbb{L}(f)$ and $\mathbb{RD}(f)$ contain indicator functions.

The real-value function $h(\mathbf{x})$ and the surrogate functions have been proposed in the recent works [43, 96, 100, 101]. For example, Zafar et al. [100] have proposed the decision boundary covariance to quantify the fairness and serve as constraints, which is equivalent to applying the linear surrogate functions to Problem Formulation 2. They have set the constraint thresholds as c and $-c$, which specify the threshold for the covariance. However, solving the optimization problem with surrogated constraints does not necessarily result in a fair classifier in terms of the original non-convex fairness requirements, e.g., $-\tau \leq \mathbb{RD}(f) \leq \tau$. In fact, there is no any fairness guarantee on the produced classifier. We use an example to show this. Consider two margin-based classifiers where the surrogate functions are linear functions of the distance from the data point to the decision boundary. Therefore, the risk difference is computed by counting the number of data points above and below the decision boundary, and the surrogated risk difference (a.k.a the decision boundary covariance) is computed by measuring the average signed distance from the data points to the decision boundary. In the dataset shown in Figure 8.1a, we obtain that the surrogated risk difference is 0 but the real risk difference is 0.25. This means that a classier obtained by solving the constrained optimization problem actually can be very unfair. In the dataset shown in Figure 8.1b, the risk difference is 0 but the surrogated risk difference is 0.5, meaning that some fair classifiers cannot be obtained by solving the optimization problem with surrogated



(a) A classifier that meets the surrogated RD constraint makes unfair predictions.



(b) A classifier that does not meet the surrogated RD constraint makes fair predictions.

Figure 8.1: Two classifiers and their predictions.

constraints.

The use of the surrogate function inevitably produces estimation errors and leads to the mismatch between the surrogated constraints and the original non-convex fairness constraints. Some intuitive techniques have been introduced to tune the threshold of the surrogated constraints for learning fair classifiers. For example, Zafar et al. [100] have proposed to build an unconstrained classifier and consider its risk difference as the initial threshold, say c^*, then they heuristically select a factor $m \in [0, 1]$ and let the threshold $c = m \times c^*$. However, the relationship between the threshold c of the surrogated constraints and the hard threshold τ of the original metrics is unclear hence users have to repeatedly conduct experiments on the datasets.

8.3 Convex Fairness Classification Framework

In this section, we propose a general framework for fairness-aware classification which addresses the gap incurred by the estimation error due to the use of the surrogate function. Our framework can formulate various fairness metrics (e.g., risk difference, risk ration, equal odds, etc.) as convex constraints and incorporate them into classic classification model. In the following sections, we present our framework based on the risk difference. In the next section, we show how our framework can be easily extended to other fairness metrics, e.g., risk

ratio, equalized odds.

We first present a constraint-free criterion that is derived from the data. This criterion ensures that any classifier learned from the data are fair in terms of the specified fairness metric. Then when this criterion is satisfied, there is no need to incorporate any fairness constraints for learning fair classification. When this criterion is not met, we formulate the fairness-aware classification task as a convex optimization problem. To fill the gap between the surrogated constraints and the real fairness metrics, we derive the upper and lower bounds for the real fairness metrics and further develop refined convex constraints. If the refined constraints are satisfied, it is guaranteed that the original non-convex fairness requirements are satisfied, e.g., $-\tau \leq \mathbb{RD}(f) \leq \tau$.

8.3.1 Constraint-free Criterion

We propose a constraint-free criterion to determine whether the fairness constraints are necessary. As discussed in Section 2.1, the unconstrained classification problem is well studied and users can safely apply the classic methods for building a fair classifier.

We first define two special classifiers f_{max} and f_{min} which obtain the maximal and the minimal risk differences respectively.

Definition 10. *The maximal risk difference classifier $f_{\max}$ and the minimal risk difference classifier $f_{\min}$ are defined as:*

$$f_{\max}(\mathbf{x}) = \begin{cases} 1 & \text{if } \eta(\mathbf{x}) \geq p, \\ -1 & \text{otherwise,} \end{cases} \qquad f_{\min}(\mathbf{x}) = \begin{cases} -1 & \text{if } \eta(\mathbf{x}) \geq p, \\ 1 & \text{otherwise,} \end{cases}$$

where we denote $P(S = s^+|\mathbf{x})$ by $\eta(\mathbf{x})$ and $P(S = s^+)$ by p.

These two classifiers provide the maximum and minimum of risk difference among all classifiers f out of the model space $\mathcal{F}$:

Theorem 13. *For any classifier f, it always holds that $\mathbb{RD}^- \leq \mathbb{RD}(f) \leq \mathbb{RD}^+$, where*

$\mathbb{RD}^- = \mathbb{RD}(f_{\min})$ *and* $\mathbb{RD}^+ = \mathbb{RD}(f_{\max})$.

Proof. Following Eq. (8.3), the risk difference of the maximum risk difference classifier $f_{\max}$ is given by:

$$\mathbb{RD}(f_{\max}) = \mathbb{E}_{\mathbf{X}}\Big[\frac{\eta(\mathbf{x})}{p}\mathbb{1}_{f_{\max}(\mathbf{X})=1} + \frac{1-\eta(\mathbf{x})}{1-p}\mathbb{1}_{f_{\max}(\mathbf{X})=-1}\Big] - 1.$$

The difference between $\mathbb{RD}(f_{\max})$ and any deterministic classifier $\mathbb{RD}(f)$ is given as:

$$\mathbb{RD}(f_{\max}) - \mathbb{RD}(f) = \mathbb{E}_{\mathbf{X}}\Big[\frac{\eta(\mathbf{x})}{p}\big[\mathbb{1}_{f_{\max}(\mathbf{X})=1} - \mathbb{1}_{f(\mathbf{x})=1}\big] + \frac{1-\eta(\mathbf{x})}{1-p}\big[\mathbb{1}_{f_{\max}(\mathbf{X})=-1} - \mathbb{1}_{f(\mathbf{x})=-1}\big]\Big].$$

Let us consider the difference of the conditional risk difference:

$$DC(\mathbf{x}) = \frac{\eta(\mathbf{x})}{p}\big[\mathbb{1}_{f_{\max}(\mathbf{X})=1} - \mathbb{1}_{f(\mathbf{x})=1}\big] + \frac{1-\eta(\mathbf{x})}{1-p}\big[\mathbb{1}_{f_{\max}(\mathbf{X})=-1} - \mathbb{1}_{f(\mathbf{x})=-1}\big],$$

1. if $\eta(\mathbf{x}) \geq p$,
   - if $f(\mathbf{x}) = 1$, $DC(\mathbf{x}) = 0$;
   - if $f(\mathbf{x}) = -1$, $DC(\mathbf{x}) = \frac{\eta(\mathbf{x})}{p} - \frac{1-\eta(\mathbf{x})}{1-p} \propto \eta(\mathbf{x}) - p \geq 0$;
2. if $\eta(\mathbf{x}) < p$, $f_{\max}(\mathbf{x}) = -1$,
   - if $f(\mathbf{x}) = 1$, $DC(\mathbf{x}) = -\frac{\eta(\mathbf{x})}{p} + \frac{1-\eta(\mathbf{x})}{1-p} \propto -\eta(\mathbf{x}) + p > 0$;
   - if $f(\mathbf{x}) = -1$, $DC(\mathbf{x}) = 0$.

We can find the difference of the conditional risk difference $DC(\mathbf{x})$ is always non-negative. Thus, the difference $\mathbb{RD}(f_{\max}) - \mathbb{RD}(f)$, the weighted average of $DC(\mathbf{x})$, is also non-negative. So $\mathbb{RD}(f_{\max}) \geq \mathbb{RD}(f)$ is proved. Similarly we can prove that $\mathbb{RD}(f_{\min}) \leq \mathbb{RD}(f)$. □

From Theorem 13, we directly obtain Corollary 2.

Corollary 2. *Given the threshold τ, for a training data if we have $\mathbb{RD}^+ \leq \tau$ and $\mathbb{RD}^- \geq -\tau$, then any classifier learned from this dataset is fair in terms of risk difference.*

Given a dataset, we can always build two classifiers f_{max} and f_{min}, then compute $\mathbb{RD}^+$ and $\mathbb{RD}^-$. If Corollary 2 is satisfied, users can safely apply any classification models to build classifiers without any fairness concern.

8.3.2 Convex Fairness-aware Classification

When the constraint-free criterion is not satisfied, it is required to incorporate fairness constraints when learning classifiers, e.g., solving Problem Formulation 2. To this end, we adopt two different surrogate functions for converting the original problem into a convex optimization. We firstly adopt a real-value predictive function h and let $f = \text{sign}(h)$, then rewrite the risk difference as

$$\begin{aligned}\mathbb{RD}(f) &= \mathbb{RD}(h) \\ &= \mathbb{E}_{\mathbf{X}|S=s^+}\Big[\mathbb{1}\big[\text{sign}\big(h(\mathbf{x})\big) = 1\big]\Big] - \mathbb{E}_{\mathbf{X}|S=s^-}\Big[\mathbb{1}\big[\text{sign}\big(h(\mathbf{x})\big) = 1\big]\Big] \\ &= \mathbb{E}_{\mathbf{X}|S=s^+}[\mathbb{1}_{h(\mathbf{x})>0}] + \mathbb{E}_{\mathbf{X}|S=s^-}[\mathbb{1}_{h(\mathbf{x})<0}] - 1.\end{aligned}$$

It follows that

$$\begin{aligned}\mathbb{RD}(f) &= \mathbb{E}_{\mathbf{X}}\left[\frac{P(S=s^+|\mathbf{x})}{P(S=s^+)}\mathbb{1}_{h(\mathbf{x})>0} + \frac{P(S=s^-|\mathbf{x})}{P(S=s^-)}\mathbb{1}_{h(\mathbf{x})<0} - 1\right] \\ &= \mathbb{E}_{\mathbf{X}}\left[\frac{\eta(\mathbf{x})}{p}\mathbb{1}_{h(\mathbf{x})>0} + \frac{1-\eta(\mathbf{x})}{1-p}\mathbb{1}_{h(\mathbf{x})<0} - 1\right],\end{aligned} \tag{8.5}$$

where we denote $P(S = s^+|\mathbf{x})$ by $\eta(\mathbf{x})$ and $P(S = s^+)$ by p for simplicity, thus $P(S = s^-|\mathbf{x}) = 1 - \eta(\mathbf{x})$ and $P(S = s^-) = 1 - p$.

It is intuitive that the indicator function in above formula can be replaced with the surrogate function. The challenge here is, two constraints $\mathbb{RD}(f) \leq \tau$ and $-\mathbb{RD}(f) \leq \tau$ are

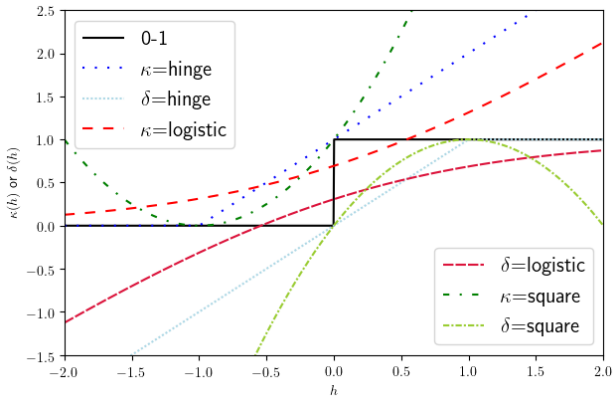


Figure 8.2: Curves of examples for $\kappa(\cdot)$ and $\delta(\cdot)$.

opposite to each other. Thus, replacing all indicator functions with a single surrogate function will result in a convex-concave problem, where only heuristic solutions for finding the local optima are known to exist. Therefore, we adopt two surrogate functions, a convex one $\kappa(\cdot)$ and a concave one $\delta(\cdot)$, each of which replaces the indicator function for one constraint. As a result, the formulated constrained optimization problem is convex and can be efficiently solved. We call the risk difference represented by $\kappa(\cdot)$ and $\delta(\cdot)$ as the κ, δ-risk difference, denoted by $\mathbb{RD}_\kappa(h)$ and $\mathbb{RD}_\delta(h)$. Almost all commonly-used surrogate functions can be adopted for $\kappa(\cdot)$ and $\delta(\cdot)$, by performing some shift or flip. Curves of some examples for $\kappa(\cdot)$ and $\delta(\cdot)$ are shown in Figure 8.2.

As a result, we obtain the following convex optimization formulation for learning fair classifiers.

Problem Formulation 3. *The fairness-aware classification is converted into a convex optimization problem. The optimal solution h^* can be obtained by solving*

$$\begin{aligned} \min_{h \in \mathcal{H}} \quad & \mathbb{L}_\phi(h) \\ \textit{subject to} \quad & \mathbb{RD}_\kappa(h) \leq c_1, \quad -\mathbb{RD}_\delta(h) \leq c_2, \end{aligned}$$

where $\kappa(\cdot)$ is a convex surrogate function, $\delta(\cdot)$ is a concave surrogate function, c_1, c_2 are the

thresholds of the κ, δ-risk difference, and

$$\begin{aligned}
\mathbb{L}_\phi(h) &= \mathbb{E}_{\mathbf{X}}\big[Pr(Y=1|\mathbf{x})\phi\big(h(\mathbf{x})\big) + \big(1-Pr(Y=1|\mathbf{x})\big)\phi\big(-h(\mathbf{x})\big)\big], \\
\mathbb{RD}_\kappa(h) &= \mathbb{E}_{\mathbf{X}}\big[\frac{\eta(\mathbf{x})}{p}\kappa\big(h(\mathbf{x})\big) + \frac{1-\eta(\mathbf{x})}{1-p}\kappa\big(-h(\mathbf{x})\big) - 1\big], \\
\mathbb{RD}_\delta(h) &= \mathbb{E}_{\mathbf{X}}\big[\frac{\eta(\mathbf{x})}{p}\delta\big(h(\mathbf{x})\big) + \frac{1-\eta(\mathbf{x})}{1-p}\delta\big(-h(\mathbf{x})\big) - 1\big].
\end{aligned}$$

After obtaining h^*, we build the fair classifier by letting $f^* = \text{sign}(h^*)$ and f^* is the final fair classifier. We emphasize that in Problem Formulation 3, the constraint thresholds are rewritten as c_1 and c_2 due to the difference between the surrogated constraints and the original non-convex constraints.

8.3.3 Refined Fairness-aware Classification

In this section, we develop the upper and lower bounds of the risk difference $\mathbb{RD}(h)$ with the κ, δ-risk difference $\mathbb{RD}_\kappa(h)$ and $\mathbb{RD}_\kappa(h)$. Based on the bounds, we present the method to derive c_1, c_2 for $\mathbb{RD}_\kappa(h), \mathbb{RD}_\delta(h)$, which provides a fairness guarantee that the solution $f^* = \text{sign}(h^*)$ to Problem Formulation 3 satisfies the fairness requirements, e.g., $-\tau \leq \mathbb{RD}(f^*) \leq \tau$. The method works for various types of surrogate functions (e.g., hinge, square, logistic, exponential, etc.).

We begin with defining the conditional risk difference $C^\eta\big(h(\mathbf{x})\big)$ for a specific subpopulation $\mathbf{x}$:

$$C^\eta\big(h(\mathbf{x})\big) = \frac{\eta(\mathbf{x})}{p}\mathbb{1}_{h(\mathbf{x})>0} + \frac{1-\eta(\mathbf{x})}{1-p}\mathbb{1}_{h(\mathbf{x})<0} - 1,$$

where η is the abbreviation of $\eta(\mathbf{x})$.

Then, according to Eq. (8.5), we have $\mathbb{RD}(f) = \mathbb{E}_{\mathbf{X}}[C^\eta\big(h(\mathbf{x})\big)]$. When surrogate function $\kappa(\cdot)$ (resp. $\delta(\cdot)$) is adopted, we similarly define the conditional κ-risk difference

$$C^\eta_\kappa\big(h(\mathbf{x})\big) = \frac{\eta(\mathbf{x})}{p}\kappa\big(h(\mathbf{x})\big) + \frac{1-\eta(\mathbf{x})}{1-p}\kappa\big(-h(\mathbf{x})\big) - 1,$$

and we have $\mathbb{RD}_\kappa(h) = \mathbb{E}_{\mathbf{X}}\big[C^\eta_\kappa\big(h(\mathbf{x})\big)\big]$.

Note that the values of $C^\eta\big(h(\mathbf{x})\big)$ and $C^\eta_\kappa\big(h(\mathbf{x})\big)$ depend on $\eta(\mathbf{x})$ and $h(\mathbf{x})$, which are determined by the subpopulation of the data specified by $\mathbf{x}$, as well as predictive function h. In order to study the general situations for any specific subpopulation and any possible predictive function, we denote $h(\mathbf{x})$ as α and define the generic conditional risk difference $C^\eta(\alpha)$ and the generic conditional κ-risk difference $C^\eta_\kappa(\alpha)$:

$$C^\eta(\alpha) = \frac{\eta}{p}\mathbb{1}_{\alpha>0} + \frac{1-\eta}{1-p}\mathbb{1}_{\alpha<0} - 1, \quad C^\eta_\kappa(\alpha) = \frac{\eta}{p}\kappa(\alpha) + \frac{1-\eta}{1-p}\kappa(-\alpha) - 1,$$

for any $\eta \in [0,1]$ and $\alpha \in \mathcal{R}$. Then, the minimal conditional risk difference $H^-(\eta)$ and the minimal conditional κ-risk difference $H^-_\kappa(\eta)$ for any arbitrary subpopulation and any possible predictive function are given by

$$H^-(\eta) = \min_{\alpha\in R} C^\eta(\alpha) = \min_{\alpha\in R}\big[\frac{\eta}{p}\mathbb{1}_{\alpha>0} + \frac{1-\eta}{1-p}\mathbb{1}_{\alpha<0} - 1\big],$$

$$H^-_\kappa(\eta) = \min_{\alpha\in R} C^\eta_\kappa(\alpha) = \min_{\alpha\in R}\big[\frac{\eta}{p}\kappa(\alpha) + \frac{1-\eta}{1-p}\kappa(-\alpha) - 1\big]. \tag{8.6}$$

It is straightforward that the minimal risk difference $\mathbb{RD}^-$ is equivalent to the expectation of $H^-(\eta(\mathbf{x}))$ since for any possible $\mathbf{x}$, $H^-(\eta(\mathbf{x}))$ provides the minimal conditional risk difference. Similarly, the minimal κ-risk difference achieved by any predictive function (denoted by $\mathbb{RD}^-_\kappa$) is the expectation of $H^-_\kappa(\eta(\mathbf{x}))$, as given by

$$\mathbb{RD}^-_\kappa = \mathbb{E}_{\mathbf{X}}\big[H^-_\kappa(\eta(\mathbf{x}))\big].$$

Finally, we define the minimal conditional κ-risk difference within interval α s.t. $\alpha(\eta - p) \geq 0$:

$$H^\circ_\kappa(\eta) = \min_{\alpha:\alpha(\eta-p)\geq 0} C^\eta_\kappa(\alpha). \tag{8.7}$$

We similarly define $H^+(\eta)$ the maximal conditional risk difference, $H^+_\delta(\eta)$ the maximal

conditional δ-risk difference, $\mathbb{RD}_\delta^+$ the maximal δ-risk difference, as well as $H_\delta^\circ(\eta)$ the minimal conditional δ-risk difference within interval α s.t. $\alpha(\eta - p) \geq 0$.

Now, we are able to present our results, which are given in Theorem 14 and Corollary 3.

Theorem 14. *If $\kappa(\cdot)$ is convex and differentiable at zero with $\kappa'(0) > 0$, $\delta(\cdot)$ is concave and differentiable at zero with $\delta'(0) > 0$, then for any predictive function h, we have*

$$\begin{aligned} \psi_\kappa(\mathbb{RD}(h) - \mathbb{RD}^-) &\leq \mathbb{RD}_\kappa(h) - \mathbb{RD}_\kappa^-, \\ \psi_\delta(\mathbb{RD}^+ - \mathbb{RD}(h)) &\leq \mathbb{RD}_\delta^+ - \mathbb{RD}_\delta(h), \end{aligned} \tag{8.8}$$

where

$$\begin{aligned} \psi_\kappa(\mu) &= H_\kappa^\circ\big(p(1-p)\mu + p\big) - H_\kappa^-\big(p(1-p)\mu + p\big), \\ \psi_\delta(\mu) &= H_\delta^+\big(p(1-p)\mu + p\big) - H_\delta^\circ\big(p(1-p)\mu + p\big). \end{aligned}$$

Proof. Let us firstly verify that ψ_κ is convex.

Because κ is convex and $\kappa'(0) > 0$, we have

$$\begin{aligned} H_\kappa^\circ(\eta) &= \min_{\alpha:\alpha(\eta-p)\geq 0} \frac{\eta}{p}\kappa(\alpha) + \frac{1-\eta}{1-p}\kappa(-\alpha) \\ &= \min_{\alpha:\alpha(\eta-p)\geq 0} \big(\frac{\eta}{p} + \frac{1-\eta}{1-p}\big)\Big[\frac{\frac{\eta}{p}}{\frac{\eta}{p} + \frac{1-\eta}{1-p}}\kappa(\alpha) + \frac{\frac{1-\eta}{1-p}}{\frac{\eta}{p} + \frac{1-\eta}{1-p}}\kappa(-\alpha)\Big]. \end{aligned}$$

Let $\nu = \frac{\eta}{p} + \frac{1-\eta}{1-p}$, the above can be reformulated as

$$H_\kappa^\circ(\eta) = \min_{\alpha:\alpha(\eta-p)\geq 0} \nu \times \Big[\frac{\eta}{p\nu}\kappa(\alpha) + \frac{1-\eta}{(1-p)\nu}\kappa(-\alpha)\Big].$$

Since κ is convex and according to Jensen's inequality, we can derive

$$\begin{aligned} H_{\phi}^{\circ}(\eta) &\geq \min_{\alpha:\alpha(\eta-p)\geq 0} \nu \times \kappa\big(\frac{\eta}{p\nu}\alpha - \frac{1-\eta}{(1-p)\nu}\alpha)\big) \\ &= \min_{\alpha:\alpha(\eta-p)\geq 0} \nu \times \kappa\big(\frac{\alpha(\eta-p)}{\nu * p(1-p)}\big) \geq \nu\kappa(0). \end{aligned}$$

The equality is achieved when $\alpha(\eta - p) = 0$, so that

$$H_{\kappa}^{\circ}(\eta) = \big(\frac{\eta}{p} + \frac{1-\eta}{1-p}\big)\kappa(0).$$

So it follows that

$$H_{\kappa}^{\circ}\big(p(1-p)\mu + p\big) = (\mu - 2p\mu + 2)\kappa(0).$$

Since H_{ϕ}° and H_{ϕ}^{-} are convex (H_{κ}^{-} is a point-wise minimum over linear functions and H_{κ}° is a linear function of μ), we conclude that $\psi_{\kappa}(\mu) = H_{\kappa}^{\circ}(p(1-p)\mu + p) - H_{\kappa}^{-}(p(1-p)\mu + p)$ is convex.

Let us move back to Eq. (8.8) whose argument could be rewrite as

$$\begin{aligned} \mathbb{RD}(h) - \mathbb{RD}^{-} &= \mathbb{E}_{\mathbf{x}}\big[C^{\eta}(h,\mathbf{x})\big] - \min_{h\in\mathcal{H}} \mathbb{E}_{\mathbf{x}}\big[C^{\eta}(h,\mathbf{x})\big] \\ &= \mathbb{E}_{\mathbf{x}}\big[C^{\eta}(h,\mathbf{x}) - \min_{h\in\mathcal{H}} C^{\eta}(h,\mathbf{x})\big] \\ &= \mathbb{E}_{\mathbf{x}}\Big[\mathbb{1}_{(\eta-p)h(\mathbf{x})<0} \times \big[\frac{\mid \eta - p \mid}{p(1-p)}\big]\Big] \\ &= \mathbb{E}_{\mathbf{x}}\big[g(\mathbf{x})\big]. \end{aligned}$$

By Jensen's inequality, if ψ_κ is convex, then we have

$$\begin{aligned}
\psi_\kappa\big(\mathbb{RD}(h) - RD^-\big) &= \psi_\kappa\Big(\mathbb{E}_{\mathbf{x}}\big[g(\mathbf{x})\big]\Big) \leq \mathbb{E}_{\mathbf{x}}\Big[\psi_\kappa\big(g(\mathbf{x})\big)\Big] \\
&\leq \mathbb{E}_{\mathbf{x}}\Big[\psi_\kappa\Big(\mathbb{1}_{(\eta-p)h(\mathbf{x})>0}\big[\frac{\mid \eta - p \mid}{p(1-p)}\big]\Big)\Big] \\
&= \mathbb{E}_{\mathbf{x}}\Big[\mathbb{1}_{(\eta-p)h(\mathbf{x})>0} \times \psi_\kappa\big(\frac{\mid \eta - p \mid}{p(1-p)}\big)\Big] \\
&= \mathbb{E}_{\mathbf{x}}\Big[\mathbb{1}_{(\eta-p)h(\mathbf{x})>0} \times \big[H_\kappa^\circ(\eta) - H_\kappa^-(\eta)\big]\Big] \\
&= \mathbb{1}_{(\eta-p)h(\mathbf{x})>0} \times \mathbb{E}_{\mathbf{x}}\big[H_\kappa^\circ(\eta) - H_\kappa^-(\eta)\big].
\end{aligned}$$

Note that if $(\eta - p)h(\mathbf{x}) \geq 0$, we always have $C_\kappa^\eta\big(h(\mathbf{x})\big) \geq H_\kappa^\circ(\eta)$ because of the definition of H_κ°. Otherwise, we always have $C_\kappa^\eta\big(h(\mathbf{x})\big) \geq H_\kappa^-(\eta)$ because of the definition of H_ϕ^-. Thus,

$$\begin{aligned}
\psi_\kappa\big(\mathbb{RD}(h) - \mathbb{RD}^-\big) &\leq \mathbb{1}_{(\eta-p)h(\mathbf{x})>0}\mathbb{E}_{\mathbf{x}}\big[H_\kappa^\circ(\eta) - H_\kappa^-(\eta)\big] + \mathbb{1}_{(\eta-p)h(\mathbf{x})\leq 0} \times 0 \\
&\leq \mathbb{1}_{(\eta-p)h(\mathbf{x})>0}\mathbb{E}_{\mathbf{x}}\Big[C_\kappa^\eta\big(h(\mathbf{x})\big) - H_\kappa^-\big(\eta\big)\Big] + \mathbb{1}_{(\eta-p)h(\mathbf{x})\leq 0}\mathbb{E}_{\mathbf{x}}\Big[C_\kappa^\eta\big(h(\mathbf{x})\big) - H_\kappa^-\big(\eta\big)\Big] \\
&= \mathbb{E}_{\mathbf{x}}\Big[C_\kappa^\eta\big(h(\mathbf{x})\big) - H_\kappa^-\big(\eta\big)\Big] \\
&= \mathbb{RD}_\kappa(h) - \mathbb{RD}_\kappa^-.
\end{aligned}$$

Similarly, we can prove $\psi_\delta\big(\mathbb{RD}^+ - \mathbb{RD}(h)\big) \leq \mathbb{RD}_\delta^+ - \mathbb{RD}_\delta(h)$. Thus, Theorem 3 is proved. □

In Theorem 14, $\psi_\kappa(\mu)$ and $\psi_\delta(\mu)$ are directly derived from the surrogate function κ and δ. Some commonly-used surrogate functions κ, δ and their corresponding ψ_κ, ψ_δ functions are listed in Table 8.1. The inequalities in Theorem 14 bound the difference between $\mathbb{RD}(h)$ and $\mathbb{RD}^+, \mathbb{RD}^-$ by the differences $\mathbb{RD}_\kappa(h) - \mathbb{RD}_\kappa^-$ and $\mathbb{RD}_\delta^+ - \mathbb{RD}_\delta(h)$. Since $\mathbb{RD}^-, \mathbb{RD}^+, \mathbb{RD}_\kappa^-, \mathbb{RD}_\delta^+$ can be computed from the dataset, we connect the original non-convex constraints and the surrogated convex constraints.

We reformulate Theorem 14 and explicitly give the upper and lower bounds of $\mathbb{RD}(h)$

in Corollary 3.

Corollary 3. *For any predictive function h, let classifier $f = sign(h)$, if $\kappa(\cdot)$ is convex and differentiable at zero with $\kappa'(0) > 0$, $\delta(\cdot)$ is concave and differentiable at zero with $\delta'(0) > 0$, then risk difference $\mathbb{RD}(f)$ is bounded by following inequalities:*

$$\mathbb{RD}(f) \leq \mathbb{RD}^- + {\psi_\kappa}^{-1}\big(\mathbb{RD}_\kappa(h) - \mathbb{RD}_\kappa^-\big),$$
$$\mathbb{RD}(f) \geq \mathbb{RD}^+ - {\psi_\delta}^{-1}\big(\mathbb{RD}_\delta^+ - \mathbb{RD}_\delta(h)\big).$$

Based on the upper and lower bounds of $\mathbb{RD}(f)$, we can derive the thresholds c_1, c_2 for the surrogated constraints in Problem Formulation 3. For example, if we aim to obtain a classifier f such that $-\tau \leq \mathbb{RD}(f) \leq \tau$, we only require the upper bound of $\mathbb{RD}(f)$ is smaller than τ and the lower bound is larger than $-\tau$. That is:

$$\mathbb{RD}^- + {\psi_\kappa}^{-1}\big(\mathbb{RD}_\kappa(h) - \mathbb{RD}_\kappa^-\big) \leq \tau,$$
$$\mathbb{RD}^+ - {\psi_\delta}^{-1}\big(\mathbb{RD}_\delta^+ - \mathbb{RD}_\delta(h) \geq -\tau.$$

Thus, we obtain the refined constraints and if the refined constraints are satisfied, the original risk difference requirements are guaranteed to be satisfied.

We modify Problem Formulation 3 to obtain Problem Formulation 4 with refined fairness constraints which guarantee the real non-convex fairness requirement.

Problem Formulation 4. *A classifier $f^* = sign(h^*)$ that achieves fairness guarantee $-\tau \leq \mathbb{RD}(f) \leq \tau$ can be obtained by solving the following constrained optimization*

$$\begin{aligned} \min_{h\in\mathcal{H}} \quad & \mathbb{L}_\phi(h) \\ \textit{subject to} \quad & \mathbb{RD}_\kappa(h) \leq \psi_\kappa(\tau - \mathbb{RD}^-) + \mathbb{RD}_\kappa^-, \\ & -\mathbb{RD}_\delta(h) \leq \psi_\delta(-\tau + \mathbb{RD}^+) + \mathbb{RD}_\delta^+. \end{aligned} \tag{8.9}$$

Note that the right-hand sides of above two inequalities are constants for a given dataset. Therefore, the constrained optimization problem is still convex. We can optimally solve this problem and the solution $f^* = sign(h^*)$ is guaranteed to satisfy $-\tau \leq \mathbb{RD}(f^*) \leq \tau$.

Table 8.1: Some common surrogate functions for κ-δ and the corresponding $\psi_\kappa(\mu)$ and $\psi_\delta(\mu)$.

Name of κ-δ	$\kappa(\alpha)$ for $\alpha \in \mathbf{R}$	$\delta(\alpha)$ for $\alpha \in \mathbf{R}$	$\psi_\kappa(\mu)$ or $\psi_\delta(\mu)$ for $\mu \in (0, 1/p]$
Hinge	$\max\{\alpha+1, 0\}$	$\min\{\alpha, 1\}$	μ
Square	$(\alpha+1)^2$	$1-(1-\alpha)^2$	μ^2
Exponential	$\exp(\alpha)$	$1-\exp(-\alpha)$	$(\sqrt{(1-p)\mu+1} - \sqrt{1-p\mu})^2$

8.4 Extension to Other Fairness Notions

In the above sections, we present our framework based on the risk difference. We show how our framework can be easily extended to other fairness metrics, e.g., risk ratio, equalized odds.

Risk ratio is a common fairness notion [20, 56]. It also requires the decision is independent with the protected attribute. Different with the risk difference, the unfairness is quantified by the ratio of the positive decisions between the non-protected group and the protected group. Let us formalize the risk ratio $\mathbb{RR}(h)$ of classifier h:

$$\mathbb{RR}(h) = \frac{\mathbb{E}_{\mathbf{X}|S=s^+}\left[\mathbb{1}_{h(\mathbf{x})>0}\right]}{\mathbb{E}_{\mathbf{X}|S=s^-}\left[\mathbb{1}_{h(\mathbf{x})>0}\right]}.$$

The fairness constraints with regards to risk ratio could be expressed as

$$\mathbb{RR}(h) = \frac{\mathbb{E}_{\mathbf{X}|S=s^+}\left[\mathbb{1}_{h(\mathbf{x})>0}\right]}{\mathbb{E}_{\mathbf{X}|S=s^-}\left[\mathbb{1}_{h(\mathbf{x})>0}\right]} \leq \tau.$$

Similar to Eq. (8.5), we express the constraints as

$$\mathbb{E}_{\mathbf{X}}\left[\frac{\eta}{p}\mathbb{1}_{h(\mathbf{x})>0} + \tau\frac{1-\eta(\mathbf{x})}{1-p}\mathbb{1}_{h(\mathbf{x})>0}\right] - \tau \leq 0. \tag{8.10}$$

Equalized odds and equalized opportunity are proposed by Hardt et al. [12]. Equalized odds requires the protected attribute and the predicted label are independent conditional on the truth label. To quantify the strength of equalized odds, we simply propose the prediction difference between two groups conditional on the truth label. So, the equalized odds is

$$\mathbb{EO}(h) = \mathbb{E}_{\mathbf{X}|S=s^+,Y}[\mathbb{1}_{h(\mathbf{x})>0}] - \mathbb{E}_{\mathbf{X}|S=s^-,Y}[\mathbb{1}_{h(\mathbf{x})>0}].$$

Similarly, a classifier h is considered as fair with regard to equalized odds if $\mathbb{EO}(h) \leq \tau$.

Let us reformulate the equalized odds constraints:

$$\begin{aligned}\mathbb{EO}(h) &= \mathbb{E}_{\mathbf{X}|S=s^+,Y}[\mathbb{1}_{h(\mathbf{x})>0}] + \mathbb{E}_{\mathbf{X}|S=s^-,Y}[\mathbb{1}_{h(\mathbf{x})<0}] - 1 \\ &= \mathbb{E}_{\mathbf{X}|Y}\Big[\frac{P(S=s^+|\mathbf{x},y)}{P(S=s^+|y)}\mathbb{1}_{h(\mathbf{x})>0} + \frac{1-P(S=s^+|\mathbf{x},y)}{1-P(S=s^+|y)}\mathbb{1}_{h(\mathbf{x})<0}\Big] - 1 \leq \tau. \end{aligned} \tag{8.11}$$

Equalized opportunity is a relaxation of equalized odds where only the positive group ($Y = 1$) is taken into account:

$$\mathbb{EOP}(h) = \mathbb{E}_{\mathbf{X}|Y=1}\Big[\frac{P(S=s^+|\mathbf{x},Y=1)}{P(S=s^+|Y=1)}\mathbb{1}_{h(\mathbf{x})>0} + \frac{1-P(S=s^+|\mathbf{x},Y=1)}{1-P(S=s^+|Y=1)}\mathbb{1}_{h(\mathbf{x})<0}\Big] - 1 \leq \tau. \tag{8.12}$$

By simply replacing the indicator functions with surrogate functions, we can readily extend our framework to the constraints (8.10), (8.11), (8.12) with regard to the three notions. Our criterion and bounds are also extensible to the three notions.

8.5 Experiments

8.5.1 Experimental Setup

Dataset. In the experiments we use two datasets: *Adult* and *Dutch Census of 2001*. The description of the *Adult* dataset is given in Appendix A.1. We consider `sex` as the

sensitive attribute with two values, `male` and `female`. Then, we consider `income` as the class label. For the *Dutch Census of 2001* dataset, details are also given in Appendix A.1. Similarly, we use `sex` as the sensitive attribute, and `occupation` as the class label.

Baseline. We compare our method with two related works, referred to as Zafar-1 [43] and Zafar-2 [13], both of which formulate the fairness-aware classification problem as constrained optimization problems. In [43], the authors quantify fairness using the covariance between the users' sensitive attribute and the signed distance from the feature vectors to the decision boundary. The fairness constraint is formulated as covariance $\leq m \times c^*$, where c^* is the measured fairness of the unconstrained optimal classifier and m is a multiplication factor $\in [0, 1]$. In [13], the fairness is quantified similarly with the distance function being replaced with a convex non-linear function. As a result, the obtained problem is a convex-concave optimization problem. In the experiments, we adopt the Disciplined Convex-Concave Programming (DCCP) [102] as proposed in [13] for solving the convex-concave optimization problem. For our method and Zafar-1, the convex optimization problem is solved using CVXPY [103]. The datasets and implementation are available at http://tiny.cc/fair-classification.

8.5.2 Constraint-free Criterion of Ensuring Fairness

To demonstrate the sufficiency criterion of learning fair classifiers, we build the maximal/minimal risk difference classifiers $f_{\min}, f_{\max}$ for both *Adult* and *Dutch Census of 2001* datasets, and measure the risk differences they produce, i.e., $\mathbb{RD}^-$ and $\mathbb{RD}^+$. The results are shown in the first two rows in Table 8.2. As can be seen, in both datasets we have large maximal and minimal risk differences. In order to evaluate a situation with small a risk difference, we also create a variant of the *Adult* dataset, referred to as *Adult**, where all attributes are binarized and the sensitive attribute `sex` is shuffled to incur a small risk difference. Then, we build a number of classifiers including Linear Regression (LR), Support Vector Machine (SVM) with linear kernel, Decision Tree (DT), and Naive Bayes (NB), using

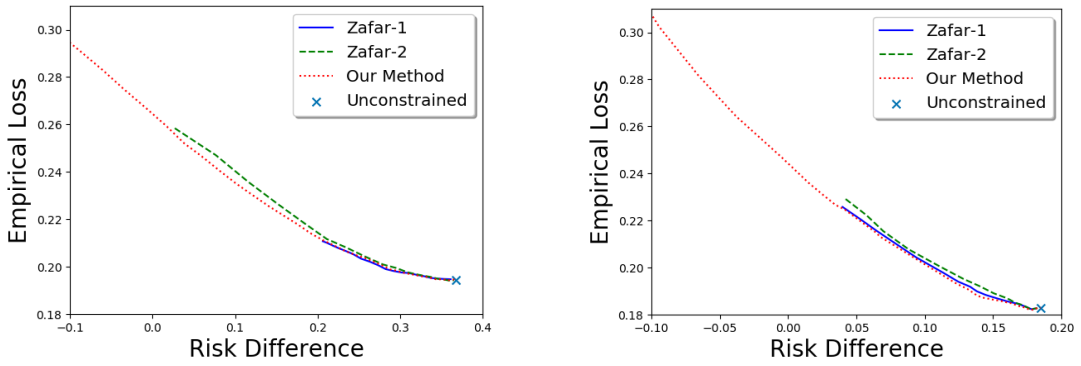
Table 8.2: $\mathbb{RD}^+, \mathbb{RD}^-$ and risk differences of Linear Regression (LR), Support Vector Machine (SVM), Decision Tree (DT), and Naive Bayes (NB).

$\mathbb{RD}(\cdot)$	*Adult*	*Dutch Census of 2001*	*Adult**
$\mathbb{RD}^+$	0.967	0.516	0.046
$\mathbb{RD}^-$	-0.967	-0.516	-0.046
LR	0.371	0.185	0.000
SVM	0.434	0.156	0.001
DT	0.316	0.184	0.001
NB	0.447	0.144	0.001

the three datasets as the training data with with 5-fold cross-validation. After that, their risk differences are quantified on the testing data, as shown in the last four rows in Table 8.2. We can see that all values are within $\mathbb{RD}^-, \mathbb{RD}^+$ which are consistent with our constraint-free criterion.

8.5.3 Learning Fair Classifiers

We build our fair classifiers on both *Adult* and *Dutch Census of 2001* datasets by solving the optimization problem defined in Problem Formulation 3. For surrogate functions, we use the logistic function for $\phi(\cdot)$, and the hinge function for $\kappa(\cdot)$ and $\delta(\cdot)$. We also compare our methods with Zafar-1 and Zafar-2. The results are shown in Figure 8.3, which depict the relationship between the obtained risk difference and empirical loss. For our method, different risk differences are obtained by adjusting relax terms c_1 and c_2, while for Zafar-1 and Zafar-2 different risk differences are obtained by adjusting the multiplication factor m. As can be seen, our method can achieve much smaller risk difference than Zafar-1 and Zafar-2. This may be because Zafar-1 linear functions to formulate the fairness constraints, which may incur large estimation errors; while Zafar-2 formulates a convex-concave optimization problem, where only the local optima can be reached. For the same reason, we can observe that our method produces better empirical loss than Zafar-2 given any same risk difference.



(a) The *Adult* dataset.

(b) The *Dutch Census of 2001* dataset.

Figure 8.3: Comparison of fair classifiers on two datasets.

8.6 Summary

In this chapter, we studied the fairness-aware classification problem and formulated it as the constrained optimization problem. We proposed a general framework which addresses all limitations of previous works in terms of: (1) various fairness metrics can be incorporated into classic classification models as constraints; (2) the formulated constrained optimization problem is convex and can be solved efficiently; and (3) the lower and upper bounds of real fairness measures are established using surrogate functions, which provide a fairness guarantee for our framework. Within the framework, we proposed a constraint-free criterion under which the learned classifier is guaranteed to be fair in terms of the specified fairness metric, as well as developed the method for learning fair classifiers if the constraint-free criterion fails to satisfy. The results demonstrated the correctness of the constraint-free criterion and the superiority of our method over existing ones in terms of achieving fairness and retaining prediction accuracy. This work has been published in The Web Conference 2019 (formerly known as WWW) [11].

9 Conclusions and Future Work

Fairness is an active research topic in the machine learning community. Researchers have developed novel and efficient association-based methods to quantify discrimination in data or models, then to debias data and models. However, there are inevitable gaps between association and causation in machine learning. In this dissertation, we investigated several problems in the fairness-aware machine learning field from the causal perspective. The conclusions and future work are summarized as follows.

9.1 Conclusions

Discrimination is divided into direct and indirect discrimination in the legal field. Due to the limitation of association-based fairness notions, existing notions, e.g., *demographic parity*, fail to distinguish direct and indirect discrimination. Thus, the crude removal methods simply remove all connections between the sensitive attribute and the decision attribute. Consequently, the resultant training data or predictions may be still biased or suffer from significant utility loss. To fill the gap between association and causation in the fairness-aware machine learning field, we formulated the direct and indirect discrimination from the causality perspective, and developed efficient algorithms for mitigating discrimination before performing predictive analysis in Chapter 4. We leveraged the structural causal model and defined discrimination as the causal effects of the sensitive attribute on the decision that are carried by the causal paths. Direct discrimination is modeled as the causal effect transmitted along the direct path from the sensitive attribute to the decision. Indirect discrimination is modeled as the causal effect transmitted along other indirect causal paths. We employed the *path-specific effect* technique [2] and formulated direct and indirect discrimination as path-specific effects. In theory and practice, the path-specific effect has the unidentification issue. In our research, we showed that direct discrimination is always identifiable, but indirect discrimination may be unidentifiable. To address this challenge, we provided a bounding

method for unidentifiable indirect discrimination. Based on the path-specific effect and the theoretical bounding results, we further proposed effective algorithms that can deal with both identifiable and unidentifiable situations, including quantifying direct/indirect discrimination, as well as mitigating any type of discrimination. In the experimental parts, we evaluated the proposed methods on real-world datasets and showed the effectiveness of our methods.

Apart from classification, ranked data analysis is a common task in machine learning where the decisions are a series of unique, concatenating integers that cannot be treated as normal categorical random variables. Thus, existing methods designed for classification are not applicable to ranked data analysis. The concern of discrimination in the ranked data is receiving increasing attention. In Chapter 5, we focused on quantifying and removing discrimination in ranked data. We employed the structural causal model and the causal graph to the ranked data by adopting a continuous substituted variable for the ranking permutation to represent the qualifications of individuals. We deployed the Bradley-Terry model [78] to map the ranking permutations into a new substituted variable called `score`. Thus, we constructed a causal graph from the original data as well as the `score` and modeled the mixed data using the conditional Gaussian distributions. We extended the path-specific effect technique into the mixed data and captured the direct and indirect discrimination in ranked data. We also built the theoretical relationship between the direct/indirect discrimination in classification and these in ranked data. Finally, the experimental results showed the effectiveness of the proposed methods on the real-world dataset.

Counterfactual fairness [7] is a new fairness notion based on counterfactual inference. The notion is evaluated by performing interventions on some individuals specified by some observed attributes. So two worlds, the factual world and the counterfactual world, are involved in this notion. However, researchers have shown that counterfactual inference cannot be uniquely computed from the observed data in some situations, referred to as unidentifiable situations. These situations pose big barriers to the applications of counterfactual fairness. In Chapter 6, we addressed the problem of learning counterfactually fair classifiers in

the unidentifiable situations by theoretically bounding the counterfactual quantities. We proposed a graphical criterion for determining the identification of counterfactual fairness and further derived the upper and lower bounds for counterfactual fairness. Finally, we proposed an efficient post-processing method for re-constructing arbitrary classifiers to achieve counterfactual fairness. This re-construction was formulated as a linear optimization problem where the bounded counterfactual fairness is constraints. In the experiments, we evaluated our method as well as existing ones using synthetic and real datasets. The results showed our method correctly achieved counterfactual fairness as expected while obtaining higher prediction accuracy. On the contrary, the comparison methods either failed to achieve counterfactual fairness or suffered from low accuracy.

The path-specific fairness and counterfactual fairness have been separately investigated in Chapter 4 and Chapter 6. There is a lack of a unified framework that can deal with them simultaneously. More importantly, one common and inevitable challenge of all causality-based fairness notions is identification. The bounding methods proposed in Chapter 4 and Chapter 6 are not applicable to the general situations. In Chapter 7, we developed a framework for handling various causality-based fairness notions and their combinations. We proposed a general representation of all types of causal effects, known as the path-specific counterfactual effect. Then we defined a unified fairness notion, path-specific counterfactual fairness, to cover most previous causality-based fairness notions. Inspired by the Response Variable method [9], we developed a constrained optimization problem for bounding the unidentifiable path-specific counterfactual fairness. The obtained bounds were guaranteed to be the tightest in theory. We evaluated the proposed method using synthetic and real datasets and compared with existing methods. The comparison showed our method was capable of bounding causal effects under any unidentifiable situations or combinations. Our method provided tighter bounds than existing methods.

Fairness-aware classification is increasingly important. Since the classification algorithms are designed to maximize predictive accuracy, the obtained classifiers may treat

individuals with undesired bias. Recently, researchers have proposed to formulate fairness-aware classification as constrained optimization. Nevertheless, the common fairness notions, e.g., risk difference, risk ratio, are non-convex. It is common to adopt surrogated functions in the optimization of non-convex functions. The consequent challenge is that there must be estimation errors due to the adoption of surrogated functions. Thus, the obtained classifiers are not guaranteed to satisfy the desired fairness requirements. In Chapter 8, we proposed a general framework for fairness-aware classification which addressed the gap incurred by the estimation errors. The framework formulated the common fairness notions as convex constraints which could be directly integrated with classification models. We also theoretically connected the surrogated fairness constraints with the original fairness requirements, then proposed refined constraints with fairness guarantees. If the refined constraints are satisfied, it is guaranteed that the original fairness measure is bounded with a given interval. The experiments demonstrated the effectiveness of the proposed method and superiority over the existing methods.

9.2 Future Work

We showed our proposed fairness notions and corresponding mitigation approaches in this dissertation. There are several interesting directions that deserve further exploration and investigation.

The proposed fairness notions are based on the structural causal model where the conditional probabilities are required. The size of the conditional probability tables are exponential to the domain spaces of variables and the number of parents. As a consequence, scalability is a key issue when we evaluate the fairness notions, including the path-specific fairness in Chapter 4, the counterfactual fairness in Chapter 6, and the PC fairness in Chapter 7. One possible solution is to leverage the deep learning techniques, e.g., deep neural networks, to encode the conditional distributions with parameters. Louizos et al. [104] designed a method based on Variational Autoencoder (VAE) to model the hidden variables

and estimate the causal effect. It is a potential direction to quantify and mitigate causal-based discrimination leveraging this type of architecture.

In Chapter 7, we derived a bounding method to estimate PC fairness. How to construct fair predictive models based on the derived bounds is an open problem and deserves investigation. One possible method would be to incorporate the bounding formulation into a post-processing method. The new formulation will be a min-max optimization problem, where the optimization variables will include response variables $P(\mathbf{r})$ as well as a post-processing mapping $P(\tilde{y}|\hat{y}, \mathsf{pa}_Y)$. The inner optimization is to maximize the path-specific counterfactual effect to find the upper bound, and the outer optimization is to minimize both the loss function and the upper bound.

More broadly, addressing discrimination issues in machine learning is still open and deserves further exploration. We elaborate our future research plan for fairness from three aspects: developing causal inference theories and technologies for discrimination detection, addressing the fairness issues in various machine learning tasks, and applying fairness-aware learning theories and techniques to real-world applications.

The development of causal inference has significant benefits for establishing principles of fairness-aware learning. However, there remain great theoretical and conceptual challenges that are worthy of further exploration in the causal inference and fairness fields. Firstly, most causality-based fairness notions and methods are on the basis of the causal graphs. Nevertheless, it is difficult to construct causal graphs from the observational data and domain knowledge. Commonly, only an equivalence class of causal models can be inferred from observational data. Some existing research [105–108] proposed the identification criteria for the equivalence class of the causal models (a.k.a. Markov equivalence class). Whereas, the identifiability of the path-specific effect and counterfactual effect is not clear until now. It deserves further investment on how to achieve fairness in the Markov equivalence class. Secondly, our previous research focuses on the static data generated by the underlying static causal models. However, many underlying causal models are dynamic and they are not readily

modeled by the existing structural causal model. The state-of-the-art technique for modeling the dynamic causal relationship is the Granger causality. How to integrate the Granger causality with fairness-aware learning is a challenging problem. Thirdly, most previous works focus on the structured data. It is well known that the unstructured data, e.g., image, text, social networks, play a key role in the Big Data Era. It is an open problem to identify causal effects from the unstructured data and achieve causal fairness in those unstructured data.

Most existing works in the fairness-aware learning literature target classification, one of the best-studied tasks in machine learning. There are many more machine learning tasks beyond classification, where the concerns have been raised about the potential impacts of discrimination. Usually, the existing methods designed for classification cannot be directly extended to other machine learning tasks, e.g., ranking, recommendation, text mining, reinforcement learning. It is urgent to develop new theorems, measurements, and algorithms for achieving fairness in other tasks.

It is important to apply the fairness-aware learning techniques to real-world applications in order to make a broader impact. For example, Correctional Offender Management Profiling for Alternative Sanctions (a.k.a. COMPAS) [1] is a risk assessment instrument that is adopted across the United States to help judges assess whether defendants should be detained or released while awaiting trial. COMPAS assigns each defendant a risk score ranging from 1 to 10 which indicates how likely the defendant to commit a crime, with 10 being the highest risk. The risk score is calculated based on more than 100 factors, including age, gender, and criminal history, but excluding race. However, COMPAS suffers a great deal of criticism since statistically black defendants are more likely to be classified as high risk. I believe our research can help analyze practical decision algorithms such as COMPAS to find out whether it is subject to discrimination, and if so help in designing fair algorithms.

Bibliography

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner, "Machine Bias: There's Software Used Across the Country to Predict Future Criminals. And it's Biased Against Blacks. - ProPublica," https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, 2016.

[2] C. Avin, I. Shpitser, and J. Pearl, "Identifiability of path-specific effects," in *IJCAI-05, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence, Edinburgh, Scotland, UK, July 30 - August 5, 2005*, vol. 2, 2005, pp. 357–363.

[3] I. Shpitser, "Counterfactual Graphical Models for Longitudinal Mediation Analysis With Unobserved Confounding," *Cognitive Science*, vol. 37, no. 6, pp. 1011–1035, 2013.

[4] L. Zhang, Y. Wu, and X. Wu, "A causal framework for discovering and removing direct and indirect discrimination," in *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI 2017, Melbourne, Australia, August 19-25, 2017*, 2017, pp. 3929–3935.

[5] ——, "Causal modeling-based discrimination discovery and removal: Criteria, bounds, and algorithms," *IEEE Trans. Knowl. Data Eng.*, vol. 31, no. 11, pp. 2035–2050, 2019.

[6] Y. Wu, L. Zhang, and X. Wu, "On Discrimination Discovery and Removal in Ranked Data using Causal Graph," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, KDD 2018, London, UK, August 19-23, 2018*, 2018, pp. 2536–2544.

[7] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, Dec. 2017, pp. 4066–4076.

[8] Y. Wu, L. Zhang, and X. Wu, "Counterfactual fairness: Unidentification, bound and algorithm," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, 2019, pp. 1438–1444.

[9] A. Balke and J. Pearl, "Counterfactual probabilities: Computational methods, bounds and applications," in *UAI '94: Proceedings of the Tenth Annual Conference on Uncertainty in Artificial Intelligence, Seattle, Washington, USA, July 29-31, 1994*, 1994, pp. 46–54.

[10] Y. Wu, L. Zhang, X. Wu, and H. Tong, "PC-Fairness: A Unified Framework for Measuring Causality-based Fairness," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, December 8-14, 2019, Vancouver, Canada, 2019*, Dec. 2019, pp. 3399–3409.

[11] Y. Wu, L. Zhang, and X. Wu, "On Convexity and Bounds of Fairness-aware Classification," in *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, 2019, pp. 3356–3362.

[12] M. Hardt, E. Price, None, and N. Srebro, "Equality of Opportunity in Supervised Learning," in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 2016, pp. 3315–3323.

[13] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi, "Fairness Beyond Disparate Treatment & Disparate Impact," in *Proceedings of the 26th International Conference on World Wide Web (WWW 2017)*. New York, New York, USA: ACM Press, 2017, pp. 1171–1180.

[14] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, "Algorithmic Decision Making and the Cost of Fairness," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '17*. New York, New York, USA: ACM Press, 2017, pp. 797–806.

[15] J. Kleinberg, S. Mullainathan, and M. Raghavan, "Inherent Trade-Offs in the Fair Determination of Risk Scores," in *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA*, 2017, pp. 43:1–43:23.

[16] R. S. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning Fair Representations," in *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, 2013.

[17] B. T. Luong, S. Ruggieri, and F. Turini, "K-NN as an implementation of situation testing for discrimination discovery and prevention," in *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '11*. New York, New York, USA: ACM Press, 2011, p. 502.

[18] I. Zliobaite, F. Kamiran, and T. Calders, "Handling conditional discrimination," in *11th IEEE International Conference on Data Mining, ICDM 2011, Vancouver, BC, Canada, December 11-14, 2011*, 2011, pp. 992–1001.

[19] D. Pedreshi, S. Ruggieri, and F. Turini, "Discrimination-aware data mining," in *Proceeding of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD 08*. New York, New York, USA: ACM Press, 2008, p. 560.

[20] D. Pedreschi, S. Ruggieri, and F. Turini, "Measuring Discrimination in Socially-Sensitive Decision Records," in *Proceedings of the 2009 SIAM International Conference on Data Mining*. Philadelphia, PA: Society for Industrial and Applied Mathematics, Apr. 2009, pp. 581–592.

[21] S. Ruggieri, D. Pedreschi, and F. Turini, "Data Mining for Discrimination Discovery," *ACM Trans. Knowl. Discov. Data*, vol. 4, no. 2, pp. 9:1–9:40, May 2010.

[22] K. Mancuhan and C. Clifton, "Combating discrimination using Bayesian networks," *Artificial Intelligence and Law*, vol. 22, no. 2, pp. 211–238, Jun. 2014.

[23] S. Hajian and J. Domingo-Ferrer, "A Methodology for Direct and Indirect Discrimination Prevention in Data Mining," *IEEE Transactions on Knowledge and Data Engineering*, vol. 25, no. 7, pp. 1445–1459, Jul. 2013.

[24] Y. Wu and X. Wu, "Using Loglinear Model for Discrimination Discovery and Prevention," in *2016 IEEE International Conference on Data Science and Advanced Analytics, DSAA 2016, Montreal, QC, Canada, October 17-19, 2016*. IEEE, Oct. 2016, pp. 110–119.

[25] F. Bonchi, S. Hajian, B. Mishra, and D. Ramazzotti, "Exposing the probabilistic causal structure of discrimination," *International Journal of Data Science and Analytics*, vol. 3, no. 1, pp. 1–21, Feb. 2017.

[26] L. Zhang, Y. Wu, and X. Wu, "Situation Testing-Based Discrimination Discovery: A Causal Inference Approach," in *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, IJCAI 2016, New York, NY, USA, 9-15 July 2016*, vol. 2016-Janua. IJCAI/AAAI Press, 2016, pp. 2718–2724.

[27] ——, "On Discrimination Discovery Using Causal Networks," in *Social, Cultural, and Behavioral Modeling, 9th International Conference, SBP-BRiMS 2016, Washington, DC, USA, June 28 - July 1, 2016, Proceedings*, vol. 9708. Springer, 2016, pp. 83–93.

[28] ——, "Achieving Non-Discrimination in Data Release," in *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*. New York, New York, USA: ACM Press, Nov. 2017, pp. 1335–1344.

[29] R. Nabi and I. Shpitser, "Fair inference on outcomes," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, Feb. 2018, pp. 1931–1940.

[30] J. Zhang and E. Bareinboim, "Fairness in decision-making - the causal explanation formula," in *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th Innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, Feb. 2018, pp. 2037–2045.

[31] N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf, "Avoiding discrimination through causal reasoning," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, 2017, pp. 656–666.

[32] W. Huang, Y. Wu, L. Zhang, and X. Wu, "Fairness through equality of effort," in *Companion Proceedings of the Web Conference 2020*. New York, NY, USA: Association for Computing Machinery, 2020, pp. 743–751.

[33] W. Huang, Y. Wu, and X. Wu, "Multi-cause discrimination analysis using potential outcomes," in *Social, Cultural, and Behavioral Modeling, 13rd International Conference, SBP-BRiMS 2020, Washington, DC, USA, October 18 - 21, 2020, Proceedings*. Springer, 2020.

[34] A. Khademi, S. Lee, D. Foley, and V. Honavar, "Fairness in Algorithmic Decision Making: An Excursion Through the Lens of Causality," in *The World Wide Web Conference, WWW 2019, San Francisco, CA, USA, May 13-17, 2019*, May 2019, pp. 2907–2914.

[35] N. Kilbertus, P. J. Ball, M. J. Kusner, A. Weller, and R. Silva, "The Sensitivity of Counterfactual Fairness to Unmeasured Confounding," in *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019*, Jul. 2019, p. 213.

[36] F. Kamiran and T. Calders, "Data preprocessing techniques for classification without discrimination," *Knowledge and Information Systems*, vol. 33, no. 1, pp. 1–33, Oct. 2012.

[37] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and Removing Disparate Impact," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '15*. New York, New York, USA: ACM Press, 2015, pp. 259–268.

[38] H. Edwards and A. Storkey, "Censoring Representations with an Adversary," in *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016*, 2016.

[39] D. Madras, E. Creager, T. Pitassi, and R. S. Zemel, "Learning adversarially fair and transferable representations," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, vol. 80, 2018, pp. 3381–3390.

[40] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA, February 02-03, 2018*, 2018, pp. 335–340.

[41] T. Calders and S. Verwer, "Three naive Bayes approaches for discrimination-free classification," *Data Mining and Knowledge Discovery*, vol. 21, no. 2, pp. 277–292, Sep. 2010.

[42] T. Kamishima, S. Akaho, and J. Sakuma, "Fairness-aware Learning through Regularization Approach," in *2011 IEEE 11th International Conference on Data Mining Workshops*. IEEE, Dec. 2011, pp. 643–650.

[43] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi, "Fairness Constraints: Mechanisms for Fair Classification," in *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, Jul. 2017, pp. 962–970.

[44] F. Kamiran, T. Calders, and M. Pechenizkiy, "Discrimination aware decision tree learning," in *Proceedings - IEEE International Conference on Data Mining, ICDM*. IEEE, Dec. 2010, pp. 869–874.

[45] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-Aware Classifier with Prejudice Remover Regularizer," in *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 2012, vol. 7524 LNAI, pp. 35–50.

[46] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference on - ITCS '12*. New York, New York, USA: ACM Press, 2012, pp. 214–226.

[47] F. Kamiran, A. Karim, and X. Zhang, "Decision theory for discrimination-aware classification," in *Proceedings of the 12nd IEEE International Conference on Data Mining (ICDM 2012)*. IEEE, Dec. 2012, pp. 924–929.

[48] D. Xu, S. Yuan, L. Zhang, and X. Wu, "FairGAN: Fairness-aware generative adversarial networks," in *IEEE International Conference on Big Data, Big Data 2018, Seattle, WA, USA, December 10-13, 2018*. IEEE, 2018, pp. 570–575.

[49] P. Sattigeri, S. C. Hoffman, V. Chenthamarakshan, and K. R. Varshney, "Fairness GAN: Generating datasets with fairness properties using a generative adversarial network," *IBM Journal of Research and Development*, vol. 63, pp. 3:1–3:9, 2019.

[50] D. Xu, S. Yuan, L. Zhang, and X. Wu, "FairGAN$^+$: Achieving fair data generation and classification through generative adversarial nets," in *2019 IEEE International Conference on Big Data (Big Data), Los Angeles, CA, USA, December 9-12, 2019*. IEEE, 2019, pp. 1401–1406.

[51] D. Xu, Y. Wu, S. Yuan, L. Zhang, and X. Wu, "Achieving Causal Fairness through Generative Adversarial Networks," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, Aug. 2019, pp. 1452–1458.

[52] T. Calders, F. Kamiran, and M. Pechenizkiy, "Building Classifiers with Independency Constraints," in *2009 IEEE International Conference on Data Mining Workshops*. IEEE, Dec. 2009, pp. 13–18.

[53] L. Zhang, Y. Wu, and X. Wu. (2018, Aug.) Anti-discrimination learning: A causal modeling-based framework. [Online]. Available: http://csce.uark.edu/~xintaowu/kdd18-tutorial/

[54] J. Pearl, *Causality: Models, Reasoning and Inference*, 2nd ed. New York, NY, USA: Cambridge University Press, 2009.

[55] D. Koller and N. Friedman, *Probabilistic Graphical Models: Principles and Techniques*. The MIT Press, 2009.

[56] A. Romei and S. Ruggieri, "A multidisciplinary survey on discrimination analysis," *The Knowledge Engineering Review*, vol. 29, no. 05, pp. 582–638, Nov. 2014.

[57] P. Spirtes, C. Glymour, and R. Scheines, *Causation, Prediction, and Search*. MIT Press, 2000.

[58] R. E. Neapolitan, *Learning Bayesian Networks*. Pearson Prentice Hall Upper Saddle River, NJ, 2004, vol. 38.

[59] D. Colombo and M. H. Maathuis, "Order-Independent Constraint-Based Causal Structure Learning," *Journal of Machine Learning Research*, vol. 15, pp. 3921–3962, 2014.

[60] M. Kalisch and P. Buhlmann, "Estimating High-Dimensional Directed Acyclic Graphs with the PC-Algorithm," *Journal of Machine Learning Research*, vol. 8, pp. 613–636, 2007.

[61] J. Pearl, "Direct and indirect effects," in *UAI '01: Proceedings of the 17th Conference in Uncertainty in Artificial Intelligence, University of Washington, Seattle, Washington, USA, August 2-5, 2001*, 2001, pp. 411–420.

[62] I. Tsamardinos, C. F. Aliferis, A. Statnikov, and L. E. Brown, "Scaling-up bayesian network learning to thousands of variables using local learning techniques," *Vanderbilt University DSL TR-03-02*, 2003.

[63] C. F. Aliferis, A. R. Statnikov, I. Tsamardinos, S. Mani, and X. D. Koutsoukos, "Local Causal and Markov Blanket Induction for Causal Discovery and Feature Selection for Classification Part I: Algorithms and Empirical Evaluation," *Journal of Machine Learning Research*, vol. 11, pp. 171–234, 2010.

[64] D. Heckerman and J. S. Breese, "A New Look at Causal Independence," in *UAI '94: Proceedings of the Tenth Annual Conference on Uncertainty in Artificial Intelligence, Seattle, Washington, USA, July 29-31, 1994*, 1994, pp. 286–292.

[65] ——, "Causal independence for probability assessment and inference using Bayesian networks," *IEEE Trans. Systems, Man, and Cybernetics, Part A*, vol. 26, no. 6, pp. 826–831, 1996.

[66] M. K. Kozlov, S. P. Tarasov, and L. G. Khachiyan, "The polynomial solvability of convex quadratic programming," *USSR Computational Mathematics and Mathematical Physics*, vol. 20, no. 5, pp. 223–228, 1980.

[67] J. Tian and J. Pearl, "Probabilities of causation: Bounds and identification," in *UAI '00: Proceedings of the 16th Conference in Uncertainty in Artificial Intelligence, Stanford University, Stanford, California, USA, June 30 - July 3, 2000*, 2000, pp. 589–598.

[68] M. Lichman, "UCI Machine Learning Repository," http://archive.ics.uci.edu/ml, 2013.

[69] P. Adler, C. Falk, S. A. Friedler, G. Rybeck, C. Scheidegger, B. Smith, and S. Venkatasubramanian, "Auditing Black-Box Models for Indirect Influence," in *2016 IEEE 16th International Conference on Data Mining (ICDM)*. IEEE, Dec. 2016, pp. 1–10.

[70] R. Scheines, P. Spirtes, C. Glymour, C. Meek, and T. Richardson, "The TETRAD Project: Constraint Based Aids to Causal Model Specification," *Multivariate Behavioral Research*, vol. 33, no. 1, pp. 65–117, Jan. 1998.

[71] J. Dahl and L. Vandenberghe, "Cvxopt: A python package for convex optimization," in *Proc. Eur. Conf. Op. Res*, vol. 2, 2006, p. 3.

[72] T. Speicher, M. Ali, G. Venkatadri, F. N. Ribeiro, G. Arvanitakis, F. Benevenuto, K. P. Gummadi, P. Loiseau, and A. Mislove, "Potential for discrimination in online targeted advertising," in *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, vol. 81, 2018, pp. 5–19.

[73] L. Zhang, Y. Wu, and X. Wu, "Achieving Non-Discrimination in Prediction," in *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, Jul. 2018, pp. 3097–3103.

[74] F. Kamiran and T. Calders, "Classifying without discriminating," in *2009 2nd International Conference on Computer, Control and Communication*. IEEE, Feb. 2009, pp. 1–6.

[75] ——, "Classification with no discrimination by preferential sampling," in *Proc. 19th Machine Learning Conf. Belgium and The Netherlands*. Citeseer, 2010, pp. 1–6.

[76] M. Zehlike, F. Bonchi, C. Castillo, S. Hajian, M. Megahed, and R. Baeza-Yates, "FA*IR: A Fair Top-k Ranking Algorithm," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management (CIKM '17)*. New York, New York, USA: ACM Press, 2017, pp. 1569–1578.

[77] K. Yang and J. Stoyanovich, "Measuring Fairness in Ranked Outputs," in *Proceedings of the 29th International Conference on Scientific and Statistical Database Management - SSDBM '17*, vol. format. New York, New York, USA: ACM Press, 2017, pp. 1–6.

[78] R. A. Bradley, "Rank Analysis of Incomplete Block Designs," *Biometrika*, vol. 41, no. 3-4, pp. 502–537, 1954.

[79] L. Zhang and X. Wu, "Anti-discrimination learning: A causal modeling-based framework," *International Journal of Data Science and Analytics*, vol. 4, no. 1, pp. 1–16, Aug. 2017.

[80] J. I. Marden, *Analyzing and Modeling Rank Data*. CRC Press, 1996.

[81] R. L. Plackett, "The Analysis of Permutations," *Applied Statistics*, vol. 24, no. 2, pp. 193–202, 1975.

[82] R. D. Luce, *Individual Choice Behavior: A Theoretical Analysis*. Wiley, 1959.

[83] C. L. Mallows, "Non-null ranking models. I," *Biometrika*, vol. 44, no. 1/2, pp. 114–130, 1957.

[84] F. Wu, J. Xu, H. Li, and X. Jiang, "Ranking Optimization with Constraints," *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management - CIKM '14*, pp. 1049–1058, 2014.

[85] A. Bhattacharyya, "On a measure of divergence between two multinomial populations," *Sankhyā: the indian journal of statistics*, pp. 401–406, 1946.

[86] R. Kumar and S. Vassilvitskii, "Generalized distances between rankings," in *Proceedings of the 19th International Conference on World Wide Web - WWW '10*. New York, New York, USA: ACM Press, 2010, p. 571.

[87] I. Shpitser and J. Pearl, "Complete identification methods for the causal hierarchy," *J. Mach. Learn. Res.*, vol. 9, pp. 1941–1979, 2008.

[88] J. Li, J. Liu, L. Liu, T. D. Le, S. Ma, and Y. Han, "Discrimination detection by causal effect estimation," in *2017 IEEE International Conference on Big Data (Big Data)*. IEEE, Dec. 2017, pp. 1087–1094.

[89] S. Chiappa, "Path-specific counterfactual fairness," in *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019.*, 2019, pp. 7801–7808.

[90] J. Tian and J. Pearl, "A general identification condition for causal effects," in *Proceedings of the Eighteenth National Conference on Artificial Intelligence and Fourteenth Conference on Innovative Applications of Artificial Intelligence, July 28 - August 1, 2002, Edmonton, Alberta, Canada*, 2002, pp. 567–573.

[91] J. Zhang and E. Bareinboim, "Equality of opportunity in classification: A causal approach," in *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, 3-8 December 2018, Montréal, Canada*, 2018, pp. 3675–3685.

[92] C. Russell, M. J. Kusner, J. R. Loftus, and R. Silva, "When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, Dec. 2017, pp. 6414–6423.

[93] D. Malinsky, I. Shpitser, and T. S. Richardson, "A potential outcomes calculus for identifying conditional path-specific effects," in *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, 2019, pp. 3080–3088.

[94] I. Shpitser and J. Pearl, "What Counterfactuals Can Be Tested," in *UAI 2007, Proceedings of the Twenty-Third Conference on Uncertainty in Artificial Intelligence, Vancouver, BC, Canada, July 19-22, 2007*, 2007, pp. 352–359.

[95] G. Goh, A. Cotter, M. Gupta, and M. Friedlander, "Satisfying Real-world Goals with Dataset Constraints," in *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, 2016, pp. 2415–2423.

[96] A. K. Menon and R. C. Williamson, "The cost of fairness in binary classification," in *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, 2018, pp. 107–118.

[97] M. Olfat and A. Aswani, "Spectral algorithms for computing fair support vector machines," in *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, vol. 84, 2018, pp. 1933–1942.

[98] B. Woodworth, S. Gunasekar, M. I. Ohannessian, and N. Srebro, "Learning Non-Discriminatory Predictors," in *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, vol. 65, Feb. 2017, pp. 1920–1953.

[99] P. L. Bartlett, M. I. Jordan, and J. D. McAuliffe, "Convexity, Classification, and Risk Bounds," *Journal of the American Statistical Association*, vol. 101, no. 473, pp. 138–156, Mar. 2006.

[100] M. B. Zafar, I. Valera, M. Rodriguez, K. Gummadi, and A. Weller, "From Parity to Preference-based Notions of Fairness in Classification," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, 2017, pp. 228–238.

[101] A. Agarwal, A. Beygelzimer, M. Dudik, J. Langford, and H. Wallach, "A Reductions Approach to Fair Classification," in *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, Mar. 2018, pp. 60–69.

[102] X. Shen, S. Diamond, Y. Gu, and S. P. Boyd, "Disciplined convex-concave programming," in *55th IEEE Conference on Decision and Control, CDC 2016, Las Vegas, NV, USA, December 12-14, 2016*. IEEE, 2016, pp. 1009–1014.

[103] S. Diamond and S. Boyd, "CVXPY: A Python-Embedded Modeling Language for Convex Optimization," *Journal of Machine Learning Research*, vol. 17, no. 83, pp. 1–5, 2016.

[104] C. Louizos, U. Shalit, J. M. Mooij, D. Sontag, R. Zemel, and M. Welling, "Causal Effect Inference with Deep Latent-Variable Models," in *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, 2017, pp. 6449–6459.

[105] A. Jaber, J. Zhang, and E. Bareinboim, "Causal identification under markov equivalence," in *Proceedings of the Thirty-Fourth Conference on Uncertainty in Artificial Intelligence, UAI 2018, Monterey, California, USA, August 6-10, 2018*, 2018, pp. 978–987.

[106] ——, "On Causal Identification under Markov Equivalence," in *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI-19*. International Joint Conferences on Artificial Intelligence Organization, Aug. 2019, pp. 6181–6185.

[107] ——, "Causal identification under markov equivalence: Completeness results," in *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, vol. 97, 2019, pp. 2981–2989.

[108] ——, "Identification of conditional causal effects under markov equivalence," in *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, 8-14 December 2019, Vancouver, BC, Canada*, 2019, pp. 11 512–11 520.

[109] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

A Appendix

A.1 Datasets

To evaluate the methods proposed in this dissertation, multiple real-world datasets are used, including the *Adult* dataset, the *Dutch Census of 2001* dataset, and the *German Credit* dataset. The sources of the datasets are listed in Table A.1. The detailed description is given as below.

A.1.1 The *Adult* Dataset

The *Adult* dataset was extracted from the *1994 Census database*, consisting of 48,842 tuples (32,561 training examples and 16,281 testing examples) with 11 attributes including `age`, `education`, `sex`, `occupation`, `income`, `marital_status` etc. This dataset is originally designed for classification and the decision attribute is `income`, i.e., whether the income is larger than `50k`.

A.1.2 The *Dutch Census of 2001* Dataset

The *Dutch Census of 2001* dataset was used and preprocessed in [18]. It consists of 60,420 tuples with 12 attributes, including `sex`, `age`, `household_position`, `household_size`, `prev_residence_place`, `occupation`, etc. This dataset is formulated for a binary classification task where individuals are classified into `high income` and `low income` professions. The decision attribute is `occupation` with two values.

Table A.1: Download links of datasets.

Dataset Name	Link
Adult	http://archive.ics.uci.edu/ml/datasets/Adult
Dutch Census of 2001	https://sites.google.com/site/conditionaldiscrimination/
German Credit	http://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)

Table A.2: Links to the implementations of proposed methods.

Chapter	Link
Chapter 4	http://tiny.cc/pse-fairness
Chapter 5	http://tiny.cc/fair-ranking
Chapter 6	http://tiny.cc/counterfactual-fairness
Chapter 7	http://tiny.cc/pc-fairness
Chapter 8	http://tiny.cc/fair-classification

A.1.3 The *German Credit* Dataset

The *German Credit* dataset consists of 1000 individuals with 20 attributes applying for loans. Among 20 attributes, 7 are numerical and 13 are categorical. This dataset is designed for binary classification to predict whether an individual is a good or bad customer. The decision attribute is the customer label where the value `1` means "good" and the value `2` means "bad".

A.2 Software

Throughout this dissertation, the causal graphs are constructed using the PC algorithm implemented by the open source software **Tetrad** [70]. This software is available at http://www.phil.cmu.edu/projects/tetrad/. The classic classification algorithms are implemented by **scikit-learn** [109]. The optimization algorithms are implemented based on **CVXOPT** [71] and **CVXPY** [103].

To promote research reproducibility, the implementations of all proposed methods in this dissertation are publicly available. The download links are given in Table A.2.