Enhancing Fairness in Machine Learning through Reweighting

Dissertation

der Mathematisch-Naturwissenschaftlichen Fakultät

der Eberhard Karls Universität Tübingen

zur Erlangung des Grades eines

Doktors der Naturwissenschaften

(Dr. rer. nat.)

vorgelegt von

M.Sc. Xuan Zhao

aus Suizhou, China

Tübingen

2024

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

Tag der mündlichen Qualifikation:	30.07.2025
Dekan:	Prof. Dr. Thilo Stehle
1. Berichterstatter:	Prof. Dr. Gjergji Kasneci
2. Berichterstatter:	Prof. Dr. Hendrik Lensch

Abstract

This cumulative PhD thesis explores innovative approaches to enhancing fairness in machine learning through advanced reweighting techniques. It addresses critical issues in the fairness of predictive models by proposing methods to mitigate bias and ensure equitable treatment of minority groups in datasets.

To tackle bias issues, the research first introduces a novel adversarial reweighting method designed to address the disparate impact that minority groups often face in biased datasets. Traditional machine learning models typically optimize for predictive utility and fairness metrics, but the under-representation of minorities makes it challenging to address these biases effectively. The proposed approach utilizes the Wasserstein distance to identify and preferentially sample majority group data points that are more similar to the minority group, thereby balancing the data distribution and enhancing fairness. Theoretical analyses confirm the method's effectiveness, and empirical results on both image and tabular benchmark datasets demonstrate significant mitigation of disparate impact without sacrificing classification accuracy. This approach outperforms related state-of-the-art methods, highlighting its practical utility in real-world applications.

Building on the importance of achieving fairness from a causal perspective, the thesis leverages Pearl's causal framework to propose a reweighting scheme that integrates causal relationships among variables into the data reweighting process. This method employs two neural networks that mirror the structures of a causal graph and an interventional graph, respectively. These networks approximate the causal model of the data and the effects of interventions, guiding a discriminator-based reweighting process to achieve various fairness notions. Experiments conducted on real-world datasets demonstrate that this approach effectively achieves causal fairness while preserving the integrity of the data for downstream tasks. This method represents a significant step forward in addressing biases that stem from underlying causal relationships in the data.

Furthermore, the research enhances the empirical risk minimization (ERM) process in model training through a refined reweighting scheme aimed at improving fairness. This approach adheres to the sufficiency rule in fairness by ensuring that optimal predictors are consistent across different sub-groups. It introduces a bilevel formulation to explore sample reweighting strategies, shifting the focus from the size of the model to the space of sample weights. To enhance training efficiency, the method discretizes these weights. Empirical validations reveal that this approach consistently improves the balance between prediction performance and fairness metrics, demonstrating its effectiveness and robustness across various experimental settings. This innovative method offers a practical solution for integrating fairness considerations into the core of the model training

process.

Collectively, the studies in this thesis contribute significant advancements in the field of fair machine learning by introducing novel reweighting techniques that address multiple facets of bias and fairness. The methods developed not only improve fairness without compromising predictive utility but also provide robust frameworks for incorporating causal considerations and optimizing fairness during the ERM process. The findings have broad implications for the development of fair and equitable machine learning models across diverse applications, paving the way for more inclusive and just AI systems.

Kurzfassung

Diese kumulative Dissertation untersucht innovative Ansätze zur Verbesserung der Fairness im maschinellen Lernen durch fortschrittliche Neugewichts-Techniken. Sie adressiert zentrale Probleme bei der Fairness von Vorhersagemodellen, indem sie Methoden zur Minderung von Verzerrungen und zur Sicherstellung einer gerechten Behandlung von Minderheitsgruppen in Datensätzen vorschlägt.

Um Bias-Probleme zu bekämpfen, wird zunächst eine neuartige adversariale Neugewichts-Methode vorgestellt, die darauf abzielt, die unterschiedliche Auswirkung auf Minderheitsgruppen in voreingenommenen Datensätzen zu adressieren. Traditionelle maschinelle Lernmodelle optimieren typischerweise für Vorhersagegenauigkeit und Fairness-Kriterien, aber die Unterrepräsentation von Minderheiten erschwert es, diese Verzerrungen effektiv zu behandeln. Der vorgeschlagene Ansatz nutzt die Wasserstein-Distanz, um Datenpunkte der Mehrheit zu identifizieren und bevorzugt zu gewichten, die den Datenpunkten der Minderheit ähnlicher sind, wodurch die Verteilung der Daten ausgeglichen und die Fairness erhöht wird. Theoretische Analysen bestätigen die Wirksamkeit der Methode, und empirische Ergebnisse auf Bild- und tabellarischen Benchmark-Datensätzen zeigen eine signifikante Minderung der unterschiedlichen Auswirkungen ohne Beeinträchtigung der Klassifikationsgenauigkeit. Dieser Ansatz übertrifft bestehende Methoden und hebt seine praktische Nützlichkeit in realen Anwendungen hervor.

Aufbauend auf der Bedeutung, Fairness aus einer kausalen Perspektive zu erreichen, nutzt die Dissertation Pearls kausalen Rahmen, um ein Neugewichts-Schema vorzuschlagen, das kausale Beziehungen zwischen Variablen in den Neugewichtsprozess integriert. Diese Methode verwendet zwei neuronale Netzwerke, die die Strukturen eines kausalen Graphen und eines Intervention Graphen widerspiegeln. Diese Netzwerke approximieren das kausale Modell der Daten und die Auswirkungen von Interventionen und leiten einen discriminator-basierten Neugewichtsprozess, um verschiedene Fairness-Kriterien zu erreichen. Experimente mit realen Datensätzen zeigen, dass dieser Ansatz effektiv kausale Fairness erreicht, während die Integrität der Daten für nachgelagerte Aufgaben erhalten bleibt. Diese Methode stellt einen bedeutenden Fortschritt bei der Behandlung von Verzerrungen dar, die aus zugrunde liegenden kausalen Beziehungen in den Daten resultieren.

Darüber hinaus verbessert die Forschung den Prozess der empirischen Risiko-Minimierung (ERM) im Modelltraining durch ein verfeinertes Neugewichts-Schema, das auf die Verbesserung der Fairness abzielt. Dieser Ansatz hält sich an die Suffizienzregel der Fairness, indem sichergestellt wird, dass optimale Prädiktoren konsistent über verschiedene Untergruppen hinweg bleiben. Er führt eine Bilevel-Formulierung ein, um Neugewichts-

Strategien zu untersuchen und verlagert den Fokus von der Modellgröße auf den Raum der Gewichtungen der Proben. Um die Trainingseffizienz zu verbessern, werden diese Gewichtungen diskretisiert. Empirische Validierungen zeigen, dass dieser Ansatz konstant das Gleichgewicht zwischen Vorhersageleistung und Fairness-Kriterien verbessert und seine Effektivität und Robustheit in verschiedenen experimentellen Einstellungen demonstriert. Diese innovative Methode bietet eine praktische Lösung für die Integration von Fairness-Überlegungen in den Kernprozess des Modelltrainings.

Insgesamt leisten die Studien dieser Dissertation wesentliche Fortschritte im Bereich des fairen maschinellen Lernens, indem sie neuartige Neugewichts-Techniken einführen, die mehrere Aspekte von Verzerrungen und Fairness adressieren. Die entwickelten Methoden verbessern nicht nur die Fairness, ohne die Vorhersagegenauigkeit zu beeinträchtigen, sondern bieten auch robuste Rahmenbedingungen zur Integration kausaler Überlegungen und zur Optimierung der Fairness im ERM-Prozess. Die Ergebnisse haben weitreichende Auswirkungen auf die Entwicklung fairer und gerechter maschineller Lernmodelle in verschiedenen Anwendungen und ebnen den Weg für inklusivere und gerechtere KI-Systeme.

Acknowledgments

First and foremost, I would like to express my deepest gratitude to my primary supervisor, Prof. Dr. Gjergji Kasneci, and my second supervisor, Dr. Klaus Broelemann. Their invaluable guidance, unwavering support, and insightful feedback have been pivotal to the completion of this thesis. Their expertise and encouragement have not only enhanced my research but also profoundly shaped my academic journey. Their dedication to fostering a rigorous yet supportive research environment has been instrumental in my development as a researcher.

I am also deeply indebted to my parents for their unconditional love, support, and encouragement throughout this journey. Their belief in me has been a constant source of motivation and strength. I would like to express my heartfelt thanks to Huan, a loyal and loving dog, for still remembering me after my long absences from home during the pandemic, and for recognizing me in less than one second when we met again afterward.

My sincere thanks go to Prof. Dr. Salvatore Ruggieri and Prof. Dr. Steffen Staab from the Marie Curie Training Network NOBIAS. Their expert advice, generous assistance, and constructive critiques have significantly contributed to the development and refinement of my research. The opportunities provided by the NOBIAS network have been invaluable in expanding my academic horizons and fostering collaborations that have enriched my work. I am grateful for the support and friendship of my fellow Early Stage Researchers (ESRs) from NOBIAS. The camaraderie and collaborative spirit within our group have been truly inspiring and have made this journey both enriching and enjoyable. The shared experiences, lively discussions, and mutual support have been an essential part of this journey, and I am thankful for the bonds we have formed.

Thank you.

Contents

Chapter 1

Introduction

1.1 Bias in Automated Decision-Making

Machine learning (ML) is increasingly applied to automated decision-making (ADM) in critical areas such as hiring and lending. However, because these models are often trained on historical data that is typically biased, they tend to not only perpetuate but also uncover new unfair and potentially discriminatory patterns. Prominent examples include the COMPAS tool (Mattu *et al.*, 2016) and systems used for screening job applications (Raghavan *et al.*, 2020; van den Broek *et al.*, 2019), where ML models have demonstrated potentially racist, sexist, or xenophobic tendencies in ADM for impactful decisions.

The emerging interdisciplinary field of Fair ML aims to identify, mitigate, and improve biased models by applying fairness principles, which are largely focused on equality. This section will first analyze sources of bias in AI, then provide an overview of central fairness concepts from recent research, followed by an examination of various bias mitigation techniques. Finally, we will offer our reflections on the current state of the field.

1.1.1 Sources of Bias in ADM

Many sources of bias have been identified in the literature (Caton and Haas, 2024) and we discuss some of them hereafter. On the one side, dataset bias (Tommasi *et al.*, 2017; Torralba and Efros, 2011) has been studied extensively, and can emerge for several reasons. An ML pipeline typically consists of multiple stages, including problem formulation, data collection, model training, deployment, and monitoring. At any stage in this process, biases and unfairness can be introduced. Ensuring fairness and responsibility in ML systems necessitates thoughtful attention to fairness at every step of the pipeline.We then provide a brief overview of a typical ML lifecycle to help identify potential sources of unfairness at each stage. This background will be valuable as we later explore strategies for promoting fairness in ML and discuss the main challenges in putting these strategies into practice.

Defining the Problem and Collecting Data

The first step in an ML pipeline often involves transforming a real-life challenge into a predictive task and assembling the appropriate training data. This includes defining the population from which data samples will be drawn and choosing the attributes and target labels to gather. Each of these choices has the potential to introduce bias into the training data. Since ML models understand the world based on this data, any existing biases or problematic correlations could be reproduced or even amplified by the model.

Another notable example is capture bias, which stems from the data collection methods. For instance, two images may differ based on the type of camera used, rather than their content, due to variations in resolution or exposure. Sampling bias and representation bias happen when the dataset distribution fails to accurately mirror the actual population because of unrepresentative sampling.

Example: A study by (Buolamwini and Gebru, 2018) investigated face recognition systems and discovered that, despite achieving roughly 90% accuracy overall for a classification task, these systems exhibited notable performance disparities among demographic groups. Specifically, accuracy was lower for female subjects compared to male subjects, and dark-skinned individuals were recognized with less precision than light-skinned individuals. Among these, dark-skinned females—an intersectional group—faced the largest accuracy gap, with differences reaching up to 34%.

Model Training and Assessment

Machine learning models are usually trained and evaluated with overall performance metrics, like maximizing average accuracy for classification tasks or minimizing mean squared error in regression tasks. However, focusing on average loss can create issues when the training data does not represent all groups within the population evenly. Likewise, using aggregate metrics, such as average accuracy, for evaluation can obscure disparities, as model performance can differ significantly across groups. In fact, these metrics may hide substantial variations in the model's effectiveness for different population segments. Additionally, feature distributions and relationships to the target variable may differ between groups, meaning that optimizing for average accuracy can result in the model favoring the majority group (often with more data), leading to higher error rates for underrepresented groups.

Biases can also be introduced through the selection of features or the methods used to measure them, leading to issues like omitted variable bias or measurement bias. Even if data collection is unbiased, real-world biases—such as those stemming from historical discrimination—may still appear in the data, creating what is known as historical bias. Further, the learning algorithm itself can introduce bias by prioritizing the majority group to optimize average accuracy, a form of bias known as algorithmic bias. This stems from how the algorithm is designed. If the model is optimized with unsuitable metrics or benchmarks, it can also lead to evaluation bias.

Example: In the U.S., graduate program admissions often rely on SAT scores to gauge a student's potential for success. Students are permitted to take the SAT multiple times, typically submitting only their highest score to enhance their application. Additionally, many students benefit from SAT tutoring, though both tutoring and retaking the exam involve financial expenses. Because of historical disparities and various social factors, African-American students, on average, score lower on the SAT than their white counterparts (Brooks, 1991). If these patterns are not carefully addressed, training machine learning models on such data could result in biased predictions against African-American applicants. This example highlights how biased data collection can undermine fairness in predictive models.

Deployment and Monitoring

During model development, it is commonly assumed that the training data accurately represents the target population for future use. In reality, however, this assumption often falls short. Models are frequently deployed in different settings or with populations (such as other geographic regions or age demographics) that vary from the training data. Additionally, models may be applied in ways that diverge from their initial intended purpose. These discrepancies between training and deployment contexts can lead to unforeseen model failures. Worryingly, these failures often occur unnoticed, even when the model generates predictions with high confidence.

Example: (Zech *et al.*, 2018) found that machine learning models designed to automatically detect pneumonia from chest X-ray images showed strong performance on data from a single hospital. However, when these models were tested on X-ray images from a different hospital, their diagnostic accuracy declined notably.

1.1.2 Notions of Fairness

Various legal frameworks address fairness concerns in decision-making. Notable examples include the U.S. Equal Employment Opportunity Commission and Title VII of the Civil Rights Act of 1964 (Barocas and Selbst, 2014), as well as the European Union's General Data Protection Regulation (Voigt and Bussche, 2017) and the AI Act[1]. Within these frameworks, two primary types of discrimination in machine learning predictions are identified (Hajian and Domingo-Ferrer, 2013; Kilbertus *et al.*, 2017; Aghaei *et al.*, 2019). Disparate treatment, or direct discrimination, involves treating individuals differently based explicitly on sensitive characteristics. This can be mitigated by omitting sensitive attributes from the inference process; however, because these attributes can often be inferred from non-sensitive data, avoiding disparate treatment alone is usually insufficient (Kilbertus *et al.*, 2018). The second type, disparate impact or indirect discrimination, occurs when decision-making practices don't overtly use sensitive attributes

[1] https://artificialintelligenceact.eu/

yet still lead to unequal outcomes for certain groups based on those attributes. As noted in Zafar *et al.* (2017a), a conflict often exists between the goals of preventing disparate treatment (i.e., not using sensitive attributes in decisions) and minimizing disparate impact (which may require using sensitive attributes to train a fair model). Given the multiple sources of bias in machine learning systems, and the potential harm to individuals or groups, fair decision-making is both a legal and ethical imperative. Various strategies to address these issues are discussed in the following subsection.

We outline the primary fairness definitions relevant to this work by focusing on the joint probability distribution $P(Y,\hat{Y},X,S)$ (and its variations), which is based on the ML model $\hat{f}$. For simplicity, we assume a single neutral attribute and a single protected attribute. Let $Y = 1$ represent a positive outcome (e.g., loan approval) and S membership in a protected group (e.g., female). While these definitions apply to this basic scenario, they can also extend to more complex cases. For illustration, consider a hiring example where S represents an applicant's gender, X their academic performance, $\hat{Y}$ the model's recommendation to interview, and Y a measure of long-term success, such as staying with the company for five years or becoming a manager.

In this section, we introduce the primary correlation-based and causality-based fairness definitions and outline essential characteristics of these definitions before delving deeper. This section is not comprehensive, as new fairness definitions continue to emerge. For a concise survey of leading fairness definitions, we recommend Verma and Rubin (2018). Additionally, works such as Binns (2018) and Hutchinson and Mitchell (2019) provide context on how Fair ML definitions compare with fairness concepts in other fields. The definitions here are aimed primarily at classification problems, which are prevalent in automated decision-making (ADM) contexts. However, these definitions have also been extended to other ML tasks; for an overview of fairness definitions in ranking problems, see Zehlike *et al.* (2023a,b).

Correlation-Based

Group Fariness Statistical fairness, often referred to as group fairness, aims to ensure that a specific statistical measure maintains similar values across different protected groups, defined by one or more sensitive attributes in the set S. The main principle here is that sensitive attributes, such as race or gender, should not unduly influence the model's predictions. Formally, these fairness metrics seek to ensure that any probabilistic differences in the chosen measure between protected groups, over the entire data distribution, do not exceed a predefined threshold for fairness. A common strategy is to constrain the observed differences in training data. Several metrics have been proposed, depending on which values are being equalized across groups. Here, we focus on four widely adopted metrics: Statistical Parity (Dwork *et al.*, 2012) (SP), Predictive Equality (Chouldechova, 2017) (PE) — which is based on the false positive rate, Equal Opportunity (Hardt *et al.*, 2016) (EOpp) — based on the true positive rate, and Equalized Odds (Hardt *et al.*, 2016) (EO). Each of these metrics seeks to balance specific values from the classifier's confu-

sion matrix across different protected groups. For instance, Equal Opportunity ensures that the true positive rates between groups remain within a certain tolerance.

Metrics like Predictive Equality, Equal Opportunity, and Equalized Odds are tied closely to model accuracy; theoretically, a perfectly accurate model would also achieve perfect fairness under these metrics as they preserve existing dataset biases (Wachter *et al.*, 2021). However, they address algorithmic bias rather than dataset bias, meaning any pre-existing bias within the dataset can still affect model predictions. By contrast, Statistical Parity is a bias-altering metric, as it bypasses true labels to target dataset bias directly. Consequently, datasets with greater inherent bias may create a stronger trade-off between Statistical Parity and model accuracy.

In the following, we provide an in-depth discussion of correlation-based (or data-driven) fairness definitions. These definitions are classified as individual or group fairness, depending on whether fairness is sought at the individual level (e.g., a specific female applicant) or the group level (e.g., all female applicants). These fairness definitions rely on probabilities, as the model $\hat{f}$ approximates the conditional distribution $P(Y|X,S)$, representing how outcomes change based on the given attributes.

Demographic Parity

Demographic Parity (DP) requires that a model's predictions remain uninfluenced by the values of sensitive attributes, such as A. For a binary sensitive attribute S, this criterion is formally stated as:

$$P(\hat{Y} = y|S = 1) = P(\hat{Y} = y|S = 0) \tag{1.1}$$

This condition, represented as $\hat{Y} \perp\!\!\!\perp S$ and known as the independence criterion, ensures that the predictions do not rely on sensitive attributes. DP is also commonly called statistical parity.

Although DP aims to eliminate bias, it can sometimes permit or even necessitate positive discrimination (also known as reverse discrimination), which may be deemed illegal in certain jurisdictions depending on local laws (Romei and Ruggieri, 2014). Concepts related to DP have been analyzed in various studies, such as Kamiran and Calders (2009) and Feldman *et al.* (2015). DP is legally related to the concept of disparate impact, or, in the European Union, to direct discrimination as outlined in anti-discrimination regulations.

It's noteworthy that Equation (1.1) does not consider the true label Y. This feature makes DP particularly useful in cases where the model operates on unlabeled data, a frequent scenario in machine learning applications. However, it also introduces a potential conflict between fairness and the model's predictive accuracy regarding the true outcome. Despite this trade-off, DP is one of the most popular and widely used fairness definitions.

Equalized Odds Equalized Odds (EO) mandates that the model achieves consistent true positive and false positive rates for the favorable outcome $Y = 1$ across varying levels of the sensitive attribute S.

$$P(\hat{Y} = 1|S = 1, Y = y) = P(\hat{Y} = 1|S = 0, Y = y) \tag{1.2}$$

This criterion implies that the model's prediction is conditionally independent of the sensitive attribute S when given the actual outcome Y, represented as $\hat{Y} \perp\!\!\!\perp S|Y$, also known as the separation criterion for non-discrimination (Barocas *et al.*, 2023). Equalized Odds (EO) was introduced by Hardt *et al.* (2016) to address Demographic Parity's (DP) limitations, specifically its lack of consideration for fairness in relation to both the predicted outcome $\hat{Y}$ and the true outcome Y.

While EO requires access to the true outcome Y, usually available during training, its practical application is often limited unless strong assumptions hold about the stability of the training and future data. Nonetheless, EO offers a way to evaluate fairness in both predictions and decisions, as discussed in Kilbertus *et al.* (2020). An extension of EO introduced by Hardt *et al.* (2016), known as equal opportunity, considers only the true positive rate.

$$P(\hat{Y} = 1|S = 1, Y = 1) = P(\hat{Y} = 1|S = 0, Y = 1) \tag{1.3}$$

Calibration A model is considered calibrated (CA) if, when it predicts that an applicant has label Y, the probability of the applicant truly having this label is consistent across all values of S:

$$P(Y = y|S = 1, \hat{Y} = y) = P(Y = y|S = 0, \hat{Y} = y) \tag{1.4}$$

This criterion, denoted as $Y \perp\!\!\!\perp S|\hat{Y}$, is known as the sufficiency criterion for non-discrimination (Barocas *et al.*, 2023). Calibration (CA), as introduced by Chouldechova (Chouldechova, 2017), is essentially the inverse of Equalized Odds (EO). Despite their similarities, CA and EO are fundamentally incompatible; achieving compatibility would require either zero prediction error or the independence of Y from S, which are generally unrealistic in practice (Kleinberg *et al.*, 2017). Estimating EO and CA involves calculating the confusion matrix for the model, which includes estimating:

- True Positives (TP): the total count of instances where $P(\hat{Y} = 1|Y = 1)$;
- True Negatives (TN): the total count of instances where $P(\hat{Y} = 0|Y = 0)$;
- the false positives (FP), or type-I error: the total count of instances where $P(\hat{Y} = 1|Y = 0)$; and
- the false negatives (FN), or type-II error: the total count of instances where $P(\hat{Y} = 0|Y = 1)$.

Using these four values, we can calculate metrics such as the true positive rate (TPR), given by $TRP = TP/(TP + FN)$, and the true negative rate (TNR), given by $TNR =$

$TN/(TN+FP)$, among others. For additional details, see Verma and Rubin (Verma and Rubin, 2018).

Estimating demographic parity (DP) offers flexibility, essentially reflecting the balance in representation within $\hat{Y}$. This can be assessed, for example, by examining the distance between the distributions $P(\hat{Y}|S=1)$ and $P(\hat{Y}|S=0)$. These definitions can also be applied conditionally by adjusting for certain neutral attributes.

Individual Fairness Individual fairness concepts are rooted in the principle that similar individuals should be treated similarly. This was initially proposed in Dwork *et al.* (2012), where the authors noted that while statistical fairness is often desirable, it can lead to outcomes that seem unfair on an individual level (for instance, situations where more qualified applicants from a majority group are denied in favor of less qualified applicants from a minority group). Furthermore, fairness through unawareness—refraining from using sensitive attributes in decisions—was shown to be inadequate due to proxy features (non-sensitive attributes correlated with sensitive ones). The authors, therefore, propose fairness through awareness, defined as a Lipschitz condition on the classifier that relies on a distance metric reflecting individual similarity for a given task. The challenge lies in defining this distance function, which is context-specific and best crafted by experts and policymakers.

Some methods lessen the need to define this metric explicitly. For instance, Madras *et al.* (2018) learn a mapping of individuals to clusters, each linked to "prototypes" that guide decisions, ensuring that individuals within the same cluster are treated similarly. Similarly, Lahoti *et al.* (2019) create an individually fair representation by mapping individuals to prototypes via clustering but with a task-agnostic objective function. When neither manual nor automatic metric definition is possible, Jung *et al.* (2021) involve a stakeholder panel to provide pairwise constraints on whether outcomes for specific individual pairs should be similar, ordered, or unrestricted, integrating these constraints into the learning process. In prior work, Ilvento (2020) propose learning a distance metric using a limited number of expert responses. Alternatively, the approach by Joseph *et al.* (2016) avoids any similarity metric, ensuring instead that no less-qualified applicant is favored over a more-qualified one, such that an individual's probability of a positive outcome reflects the true probability of that outcome.

Individual Fairness (IF) formalizes the principle that individuals who are similar should receive similar treatment. Mathematically, this implies that for two distinct but comparable profiles, i and j:

$$P(\hat{Y}=y_i|X=x_i,S=1) \approx P(\hat{Y}=y_j|X=x_j,S=0) \tag{1.5}$$

where similarity is determined by the metric d , such that $d(x_i,x_j) \approx 0$. This can also be expressed in terms of an allowable ε-deviation, meaning the model is regarded as individually fair as long as:

$$|P(\hat{Y} = y_i|X = x_i, S = 1) - P(\hat{Y} = y_j|X = x_j, S = 0)| \leq \varepsilon \tag{1.6}$$

which enables consideration of a spectrum of Individual Fairness (IF) concepts.

The formalization presented in (1.5) represents one of several approaches to defining Individual Fairness (IF), as it relies on how similarity is defined. Learning or determining a specific metric d is challenging because labeling two individuals as similar is inherently subjective. This emphasis on individual-level assessment distinguishes IF from the previously discussed fairness definitions. While IF lacks a specific non-discrimination criterion, it underpins non-discrimination laws, which advocate for treating similar individuals alike—a principle that dates back to Aristotle and continues to influence Western perspectives on non-discrimination (Wachter *et al.*, 2021). Currently, the dominant narrative in Fair Machine Learning (Fair ML) categorizes fairness definitions into individual and group fairness. This distinction highlights an inherent tension; for example, under Demographic Parity (DP), a model might achieve group fairness between male and female applicants while favoring a less qualified female candidate over her similarly qualified male counterpart, thus violating IF. However, this issue is not straightforward, given the stringent similarity requirements necessary for implementing individual fairness. Perspectives on similarity can vary; one person may view the poorly qualified female applicant and her male equivalent as similar, while another may not. This variability illustrates that similarity is a normative concept, which complicates discussions in Fair ML and poses challenges in substantiating discrimination claims (Weerts *et al.*, 2023).

Causality Based

Structural Causal Models The foundational conceptual framework of Chapter 3 in this thesis revolves around causality. The process of discovering, articulating, and assessing causal claims, such as $X \rightarrow Y$ or "X causes Y", has been fundamental in enhancing our understanding of the world. Researchers generally agree on what does not constitute causation, often encapsulated in the phrase "correlation does not imply causation," a well-known adage in introductory statistics courses. However, the definition of causation itself is less straightforward. Perspectives on causality can vary significantly among philosophers, economists, and other scholars. Therefore, when discussing causation, it is essential to specify which interpretation we are referencing.

Causality suggests an underlying structure. In a causal world, everything that occurs is determined by the laws governing that world along with its initial conditions. Events follow a natural order, driven by cause-and-effect processes that map how information moves within this framework. Studying causality, therefore, requires a certain level of assumption (to varying extents) that such structures indeed exist. Despite ongoing philosophical discussions, causality remains a valuable approach across many fields and has recently gained traction in Machine Learning (Schölkopf, 2022; Wang and Jordan, 2021).

This thesis adopts structural causal models, as advanced by Pearl (Shanmugam, 2001), to explore causality. Structural causal models are probabilistic graphical tools that integrate Bayesian networks with structural equation models, enabling us to describe the data-generating process (DGM) of a probability distribution. Additionally, as discussed later in this section, these models allow us to manipulate the DGM to produce new probability distributions that answer observational (what is), interventional (what if), and counterfactual (what might have been) questions (Pearl *et al.*, 2016). This capacity to create new probabilistic representations is one of the key reasons structural causal models are often preferred over other causal frameworks, such as potential outcomes (Briggs, 2012), making them particularly suitable for Machine Learning. The main motivations for selecting structural causal models as the framework here include their prominence in Machine Learning and their unique capabilities.

For example, many definitions of causal fairness (e.g., Chiappa (2019); Kilbertus *et al.* (2017); Kusner *et al.* (2017)) rely on structural causal models. Secondly, these models align with human reasoning processes (Fenton *et al.*, 2020), enabling us to address concepts, like discrimination, that are often framed causally. Lastly, structural causal models provide an effective structure for organizing assumptions about the world, which supports stakeholder involvement in discussions on complex topics like fairness. By leveraging causality, researchers and practitioners can move beyond merely detecting bias to addressing it in a principled way, laying a strong foundation for achieving causal fairness. This approach ensures that models are not only technically accurate but also ethically aligned with fairness standards.

A structural causal model (SCM) is defined as a tuple $\mathcal{M} = \langle U, V, F \rangle$ that represents the data-generating process, mapping a set of p exogenous latent random variables $U \sim P_U$ to a set of p endogenous observed random variables V. This transformation occurs through a set of structural equations F, such that:

$$P_U = P(U_1, ..., U_p) \quad V_j \coloneqq f_j(V_{pa(j)}, U_j) \quad \text{for } j \text{ in } 1, ..., p \tag{1.7}$$

Here, $U_j \in U$, $V_j \in V$, and $f_j \in F$, where each function f_j maps the exogenous variable U_j to the endogenous variable V_j. This mapping is based on the subset of endogenous variables that directly influence V_j, known as its causal parents, denoted by $V_{pa(j)}$.

Note that we use the operator $\coloneqq$ rather than $=$ in (1.7), indicating an assignment rather than equality and equivalent to $\leftarrow$. Unlike the equality operator $=$, this assignment operator represents a directional flow of information, showing cause-effect or parent-child relationships instead of simple equivalence. Conceptually, an SCM describes the data-generating process for a dataset (or context) in terms of causal relationships. For example, it allows us to represent causal relationships between X and Y, enabling causal reasoning about these variables and the problem at hand. In (1.7), we assume causal sufficiency, meaning that there are no hidden common causes or confounders in the model. This assumption leads to independence among the exogenous latent random variables in

U:

$$P(U_1,...,U_j) = P(U_1) \times \cdots \times P(U_j) \tag{1.8}$$

which enables us to break down P_U into its individual parts. While it is challenging to assume and verify causal sufficiency, this assumption—though not essential—is common because it simplifies the generation of counterfactuals. The structural causal model (SCM) $\mathcal{M}$ generates a related causal graph $\mathcal{G}$, where each node represents a random variable and each directed edge signifies a causal relationship. For example, an edge $V_i \to V_j$ exists in the causal graph if $i \in pa(j)$. Essentially, the causal graph visually represents the causal dependencies within the SCM $\mathcal{M}$. By definition, the causal graph $\mathcal{G}$ is directed and, as we assume it to be acyclic, it contains no feedback loops, making it a causal directed acyclic graph (DAG). In the rest of this section, we provide foundational concepts needed to understand SCMs; explore the interventionist view of causality that underlies SCMs; and conclude with a discussion on generating counterfactuals within this framework.

In this section, we outline fairness definitions that depend on additional knowledge provided by a structural causal model $\mathcal{M}$, categorizing them as causality-based fairness definitions. This area is expanding within Fair ML, as causal inference has already been employed by social scientists to assess cases of discrimination (Heidari *et al.*, 2019). This approach is relevant because discrimination concerns whether an outcome is caused, either directly or indirectly, by a protected attribute. Here, we focus on the most pertinent definitions for this study. For a comprehensive overview of these definitions, refer to causal fairness surveys by Loftus *et al.* (2018) and Makhlouf *et al.* (2020).

Total Effect The total effect infers the causal effect of S on Y through all possible causal paths from S to Y. The total effect of the difference of s^- to s^+ on Y is given by $TE(s^+,s^-) = P(Y_{s^+}) - P(Y_{s^-})$, where $P(\cdot)$ here refers to the interventional distribution probability. Total fairness is satisfied if $|TE(s^+,s^-)| < \varepsilon$ (ε is the fairness threshold). Note that statistical parity is similar to total effect but is fundamentally different. Statistical parity measures the conditional distributions of Y change of the sensitive attribute from s^- to s^+.

Path-specific fairness The path-specific effect is a fine-grained assessment of causal effects, that is, it can evaluate the causal effect transmitted along certain paths. It is used to distinguish among direct discrimination, indirect discrimination, and explainable bias. It infers the causal effect of S on Y through a subset of causal paths from S to Y, which is referred to as the π-specific effect denoting the subset of causal paths as π. The specific effect of a path set π on Y, caused by changing the value of S from s^- to s^+ with reference to s^-, is given by the difference of the interventional distributions: $SE_\pi(s^+,s^-) = P(Y_{s^+|\pi,s^-|\overline{\pi}}) - P(Y_{s^-})$, where $P(Y_{s^+|\pi,s^-|\overline{\pi}})$ represents the distribution resulting from intervening $do(s^+)$ only along the paths in π while $s-$ is used as a reference through other paths $\overline{\pi}$. If π contains all direct edge from S to Y, $SE_\pi(s^+,s^-)$ measures the direct discrimination. If π contains all indirect paths from S to Y that pass through proxy attributes, $SE_\pi(s^+,s^-)$ evaluates the indirect discrimination. Path-specific fairness

is met if $|SE_\pi(s^+, s^-)| < \varepsilon$.

A predictor $\hat{Y}$ of Y is considered counterfactually fair with respect to the protected attribute $O = o$, an unobserved (latent) variable U, and any observed variables X if:

$$P(\hat{Y}_{O\leftarrow o}(U) = y | X = x, O = o) = P(\hat{Y}_{O\leftarrow o'}(U) = y | X = x, O = o) \tag{1.9}$$

for all y and $o' \neq o$. This concept was initially introduced by Kusner *et al.* (2017).

The idea behind (1.9) is straightforward: a decision is counterfactually fair if it remains unchanged had the individual belonged to a different group in O. In terms of notation, $\hat{Y}_{O\leftarrow o'}(U)$ in (1.9) signifies "what the outcome $\hat{Y}$ would have been" under the latent variable U if O had been o'. Given a structural causal model (SCM) $\mathcal{M}$, each individual counterfactual quantity is derived through causality's abduction, action, and prediction steps, based on individual values for X and O. Since its initial introduction, this concept has been expanded in various studies. For instance, Chiappa (2019) examines counterfactual fairness along particular paths in the SCM $\mathcal{M}$, while Kilbertus *et al.* (2019) investigates the robustness of counterfactual assertions in the presence of hidden confounders within $\mathcal{M}$. Additionally, Kusner *et al.* (2017), a companion to Kusner *et al.* (2017), explores counterfactual fairness across multiple causal graphs, introducing the concept of approximate counterfactual fairness. Here, the model is considered approximately counterfactually fair within an ε difference across various worldviews, meaning fairness claims hold across multiple causal frameworks.

A key challenge with causal definitions, particularly for counterfactual questions in Fair ML, lies in the fact that we can only observe the factual outcome Y, not the hypothetical counterfactual outcome Y^{CF}. Addressing this issue generally boils down to either a matching approach, widely used outside of ML, or a representation learning approach, which is more common within ML. At its core, counterfactual reasoning aims to answer questions about a particular individual or unit where we observe only one outcome (what actually occurred) and imagine an alternative outcome (what might have occurred). Ideally, we would have access to both the factual and counterfactual versions of each unit or individual, but this isn't possible in reality. The next best option is either to find a comparable individual that closely resembles the counterfactual scenario or to generate a representation that serves as the desired counterfactual.

In causal fairness definitions, the line between individual and group-level fairness is often unclear. Since many of these definitions rely on comparing individual profiles, they tend to emphasize individual fairness. This focus is even stronger in cases involving counterfactual reasoning, which is inherently individual-based (e.g., Kusner *et al.* (2017)). However, causality also implies a structure that applies universally to individuals (e.g.,). If a causal effect holds for one person within a group, it should logically extend to other members of that group. Although rarely discussed, this distinction matters—particularly as discrimination claims are made on an individual basis yet often argued at the group level. We now examine Kilbertus *et al.* (2017), which uses causal reasoning with an SCM $\mathcal{M}$ to formalize concepts of discrimination. This approach, as

detailed in the definitions below, relies on the structure of the causal graph. The starting assumption is that any path, direct or indirect, from the protected attribute O is a potential issue. This "skeptical" viewpoint is then relaxed by recognizing that some descendants of O are less concerning than others. In this context, Kilbertus *et al.* (2017) introduce resolving and proxy variables to better describe discrimination. A resolving variable is any variable in the causal graph "that is influenced by O in a way deemed nondiscriminatory," while a proxy variable is any descendant of O "that is closely correlated with O but, in principle, should not influence the prediction."

1.1.3 Fairness-Enhancing Methods

Fairness-enhancing methods can be categorized into three main types, based on the stage of the machine learning pipeline where they are applied (Bellamy *et al.*, 2019; Friedler *et al.*, 2019). Numerous methods have been proposed for each category, and recent surveys offer comprehensive overviews of these approaches (Caton and Haas, 2024). Below, we outline the core principles of these three categories, discuss their inherent benefits and limitations, and provide brief examples of methods within each.

Pre-processing Pre-processing methods focus on removing unwanted correlations from training data before applying standard learning techniques to the modified data. A primary advantage of these methods is their flexibility; because they are independent of the hypothesis class and learning algorithm, a single sanitized dataset can support multiple downstream tasks. However, these approaches do not guarantee that the final model will be fair, as they only aim to reduce bias within the data itself, without addressing potential biases introduced by the learning process. Additionally, pre-processing methods may lead to notable utility losses due to the adjustments made to the training data.

Examples include Feldman *et al.* (2015), which adjusts the non-sensitive attributes in the training data so that each attribute's distribution remains consistent across groups with different sensitive attributes. Similarly, Kamiran and Calders (2011) propose several strategies to mitigate discrimination in the training data, such as altering class labels, re-weighting instances, suppressing certain attributes, or re-sampling.

Jiang and Nachum (2020) note that altering instance attributes or labels may raise legal issues, as it could be viewed as training on falsified data. They propose a re-weighting approach to adjust instance distribution and satisfy fairness constraints without modifying the data directly. Zemel *et al.* (2013) present a framework for Learning Fair Representations (LFR), which seeks to obscure sensitive attributes while retaining essential information. The resulting representations address both individual and group fairness. Similarly, Calmon *et al.* (2017) aim to transform the data to reduce discrimination, maintain utility, and limit distortion of individual instances by framing it as a convex optimization problem under specific conditions.

Other methods, such as those in (Aïvodji *et al.*, 2021), use local adversarial debiasing,

employing a discriminator to detect sensitive attribute information in representations and a sanitizer to obscure it. This process builds a sanitizer that can be applied before new data is released, targeting fairness by preventing inference of sensitive attributes from non-sensitive ones.

In-processing In-processing techniques, also known as algorithmic modification methods, directly modify the learning process to create models that are inherently fair. For example, Kamishima *et al.* (2012) incorporate a fairness-aware regularizer into the objective function of a learning algorithm, integrating it into logistic regression models. Additionally, Raff *et al.* (2018) develop fair decision trees and random forests by altering how information gain is computed during greedy tree induction. Their modified splitting criterion penalizes any splits that correlate with the values of sensitive attributes.

The Exponentiated Gradient method (Agarwal *et al.*, 2018) addresses fairness-constrained learning by utilizing Lagrangian relaxation to approximate the problem, framing it as a two-player min-max game. In this setup, the learner aims to minimize an objective function, while an auditor, who possesses information about the sensitive attributes, seeks to maximize it by identifying the most significant fairness violations. The learner then iteratively adjusts the model's parameters by solving a cost-sensitive classification problem, where costs are dictated by the unfairness violation coefficients established by the auditor.

Zafar *et al.* (2017a) introduce a convex relaxation of conventional fairness constraints termed decision boundary unfairness, which is measured as the covariance between sensitive attributes and their signed distance from the decision boundary. This concept is integrated into the training of convex margin-based classifiers like logistic regression and support vector machines.

Adversarial training techniques can also be employed to prevent the inference of sensitive attributes. For instance, in Beutel *et al.* (2017), an adversarial classifier is linked to the latent representation of a deep learning model, aiming to extract sensitive attribute information, while the main network tries to obscure it. Similarly, Zhang *et al.* (2018a) adopt an analogous strategy but focus on preventing correlations between the output and sensitive attributes by connecting the adversarial component to the output layer of the model.

In-processing methods are the most extensively researched category in the literature (Friedler *et al.*, 2019), with many approaches proposed within this realm. They typically yield the best trade-offs between fairness and utility, as they utilize all available information in the learning process to find optimal balances between the two objectives. However, their main disadvantage is that they necessitate the design of specific algorithms, which can increase complexity and complicate training processes (Jiang and Nachum, 2020).

Post-processing Post-processing methods adjust the outputs of an already trained classifier to ensure fairness. One such approach is the Threshold Optimizer (Hardt *et al.*, 2016), which utilizes Linear Programming to calculate the probabilities of modifying the original model's predictions for each protected group. This results in a randomized classifier that meets statistical fairness criteria (such as equalized odds) on average. In subsequent research, Pleiss *et al.* (2017) apply similar strategies, but with a focus on outputs from calibrated classifiers.

Another method, proposed by Kamiran *et al.* (2012), seeks to restore fairness by assigning unfavorable outcomes to individuals from privileged groups while attributing favorable outcomes to those from unprivileged groups, particularly targeting examples that are low in confidence and close to the decision boundary. Additionally, Lohia *et al.* (2019) adjust the predictions for certain examples within protected groups to enhance specific statistical fairness metrics. This method identifies examples likely to produce poor individual fairness scores, thereby improving both individual and statistical fairness simultaneously.

There are also model-specific techniques, such as the one proposed by Kamiran *et al.* (2010), which alters the labels in the leaves of a trained decision tree to comply with fairness criteria. However, most post-processing approaches are model-agnostic, making them particularly effective in scenarios where an unfair model has already been trained. A common limitation of these methods is that they typically require access to sensitive attributes during inference, which can be considered a form of disparate treatment and may be restricted in many contexts (Barocas and Selbst, 2014; Zafar *et al.*, 2017a). Moreover, because these fairness adjustments occur after the training phase, the resulting trade-offs may be far from optimal (Woodworth *et al.*, 2017).

1.1.4 Further Considerations in Fair Machine Learning

Compatibility and Applicability of Fairness Notions

The literature highlights various challenges related to the implementation of the fairness metrics mentioned earlier. For instance, recent studies indicate that certain fairness concepts may be mutually exclusive. Additionally, these definitions have primarily been developed by the computer science field to measure and quantify fairness, leading to skepticism about whether they genuinely reflect meaningful fairness. Moreover, practical applications of these metrics present additional concerns. The following sections will delve deeper into these issues.

(In)compatibility of individual and statistical fairness. According to Friedler *et al.* (2016), achieving both statistical and individual fairness at the same time is often impossible, and they present results to support this claim. Empirical evidence has demonstrated that these two concepts can be in conflict with each other. Nevertheless, some studies have attempted to tackle both fairness notions together.

Joint enforcement of both concepts. Various frameworks have been developed to ad-

dress both individual and statistical fairness concurrently. For example, Zemel *et al.* (2013) incorporate statistical fairness constraints into their method for learning fair representations, ensuring that an individual's likelihood of being included in a prototype's cluster is not affected by their belonging to a protected group. Additionally, Lohia *et al.* (2019) introduce a post-processing technique that adjusts the predictions for certain examples within protected groups to enhance statistical fairness. By focusing on examples that have low individual fairness scores, this approach effectively optimizes both individual and statistical fairness at the same time.

Alignment of both concepts. According to Dwork *et al.* (2012), individual fairness, defined as a Lipschitz condition, suggests that statistical parity can be achieved if members of different protected groups are sufficiently similar. They illustrate this by showing that with a carefully selected distance metric for individuals, the violation of statistical parity can be bounded by the original distance between the distributions of protected groups and the enforced Lipschitz condition for individual fairness. When the distributions of the protected groups differ significantly, this theoretical upper bound becomes loose, necessitating affirmative action (i.e., preferential treatment) to maintain fairness. In such instances, individual and statistical fairness may strongly conflict, potentially resulting in trivial classifiers. The individual fairness condition requires that similar examples yield similar outcomes, while the statistical fairness condition could require dissimilar examples to have comparable outcomes, which can lead to a situation where all examples must have similar predictions. One proposed solution is to enforce individual fairness only within protected groups rather than between them, while still upholding statistical fairness. Another approach involves adjusting the metric so that the Lipschitz condition naturally leads to statistical parity. Lahoti *et al.* (2019) noted that their method for ensuring individual fairness also positively affected group fairness metrics. However, they acknowledged, consistent with the findings of Dwork *et al.* (2012), that unless the distributions of features and labels are consistent across protected groups, jointly enforcing individual and group fairness will involve some trade-offs.

Conceptual compatibility of both concepts. Binns (2020) argues that the perceived conflict between individual and group fairness stems from the technical frameworks used rather than being a fundamental issue with the concepts themselves. Individual fairness relates to the principle of consistency in justice, as articulated by Aristotle, which states that similar cases should yield similar judgments. Conversely, statistical fairness embodies the principle of luck egalitarianism, which asserts that inequalities should not arise from attributes or circumstances beyond an individual's control. The principles of consistency and egalitarianism are not inherently at odds and may even support each other. Furthermore, the author suggests that individual and statistical fairness can be applied interchangeably by choosing appropriate metrics.

On the relationship with non-technical fairness notions. An important finding by Binns (2020) is that neither individual fairness nor statistical fairness effectively captures the concept of individualized justice as articulated by Aristotle. Fair models, like all machine learning systems, depend on prior examples and generalization to produce

new predictions, which contradicts the principle of completely individualized treatment. Selbst *et al.* (2019) point out that current fairness metrics do not adequately address the societal context that is critical to fairness applications. Fairness tends to be specific to particular problems, which can be at odds with the proposed mathematical frameworks. Additionally, a fair model should be viewed as a part of a larger system. Optimizing the model without considering its broader context and anticipating its impact on that context are significant factors that current formulations often overlook. These studies highlight fundamental shortcomings in existing fairness definitions: they may not resonate with broader, non-technical interpretations of fairness (Datta *et al.*, 2023), and they might also fall short of fulfilling legal obligations, which typically encompass more than a single statistical metric (Watkins and Chen, 2024). Moreover, fairness-enhancing strategies can lead to cascading effects, where addressing unfairness at one stage of the machine learning pipeline may introduce new biases at later stages (Krco *et al.*, 2023). Static fairness measures may also produce unintended consequences over time (Liu *et al.*, 2019a). This underscores the importance of thoroughly understanding the long-term impacts of fairness interventions throughout the entire development and deployment processes.

On the applicability of current fairness metrics. As previously discussed, a variety of fairness concepts and corresponding metrics have been proposed. According to Ignatiev *et al.* (2020), a set of desirable properties for a fairness metric has been identified, and it has been shown that only fairness through unawareness meets these criteria. However, this method has proven ineffective when proxy features—non-sensitive attributes correlated with sensitive ones—are present (Dwork *et al.*, 2012). For example, it has been demonstrated that geographic location can be a reliable predictor of racial identity (Fiscella and Fremont, 2006; Long and Albert, 2021). This illustrates that no fairness metric is without flaws. Additionally, several studies have theoretically established impossibility results, indicating that many well-known group fairness metrics are not compatible, making it impractical to apply more than two or three simultaneously (Defrance and Bie, 2023). Nevertheless, certain metrics may be better suited for specific contexts. Consequently, when confronted with a decision-making scenario, it is essential to determine which fairness metric is most appropriate. The work of Makhlouf *et al.* (2021) addresses this issue by first identifying the critical characteristics of the problem, such as the presence of explaining variables (proxy features that can justifiably account for an unfair outcome), the likelihood of intersectionality, and the significance of false/true positives/negatives in the context. They ultimately develop a decision diagram to guide the selection of a fairness concept based on these characteristics. Another limitation is that much of the fairness literature focuses on binary classification tasks. The definition and implementation of fairness in other contexts, such as multi-class classification, regression, or even reinforcement learning, are still underdeveloped. In the following subsection, we will explore the definitions and applications of fairness concepts in multi-class classification settings.

Long-Term Impact of Fairness A typical ML pipeline in the real-world does not end after model learning or model deployment. Instead, the ML life-cycle continues with post-deployment tasks such as further data collection to perform model updates, i.e., re-train future versions of the ML system. The current version of the model determines which data/labels are collected for future training, thus inadvertently introducing feedback loops into the system. For example, in a credit lending system possible loan defaults (target class label) can only be observed if the current lending systems grants a loan in the first place (Liu *et al.*, 2019a). Similar feedback loops have been observed in predictive policing, recommender systems, pretrial detention and employment (Ferraro *et al.*, 2021; Ensign *et al.*, 2018; Lum and Isaac, 2016). Similarly, Liu *et al.* (2019a) and Zhang *et al.* (2018a) investigate long-term impact of fairness interventions in ML systems and show that short term fairness goals can introduce unexpected and counter-intuitive harms to the protected group. For instance, enforcing fairness criteria such as statistical parity in the predictive outcomes may have undesirable effects such as increased loan default rate amongst the members of the protected group, which can in turn affect future loan opportunities for the protected group members. D'Amour *et al.* (2020) built on these results and developed a simulation framework to study long term dynamics in fair ML systems.

In real-world applications, an ML pipeline does not conclude with model training or deployment. Instead, the lifecycle extends to post-deployment tasks, such as collecting new data for updating or retraining future versions of the model. The current model version influences which data and labels are collected for future training, unintentionally introducing feedback loops. For instance, in a credit lending system, possible loan defaults (the target label) can only be observed if loans are initially approved by the current system (Liu *et al.*, 2019a). Similar feedback loops have been documented in areas like predictive policing, recommendation engines, pretrial detention, and hiring (Ferraro *et al.*, 2021; Ensign *et al.*, 2018; Lum and Isaac, 2016). Additionally, Liu *et al.* (2019a) and Zhang *et al.* (2018a) explored the long-term effects of fairness interventions, revealing that short-term fairness goals can sometimes lead to unintended harms for protected groups. For example, enforcing fairness criteria such as statistical parity in predictions may inadvertently raise loan default rates within the protected group, potentially limiting future loan access for its members. Building on these findings, D'Amour *et al.* (2020) developed a simulation framework to analyze long-term dynamics in fair ML systems.

1.1.5 Importance reweighting in Machine learning

Importance weighting is a key technique in statistics and machine learning that adjusts the objective function or probability distribution according to the significance of each instance. Its straightforward yet effective nature has resulted in numerous applications across various domains. For instance, in supervised learning, when there is a known difference between training and test distributions—referred to as distribution shift—importance weighting can ensure statistically favorable outcomes by utilizing the

density ratio. This concept has been successfully adapted in many areas of machine learning in recent years. Often termed importance weighting, this approach has been shown to be beneficial in a range of applied tasks. It has proven useful in several subfields of machine learning, including Distribution Shift Adaptation, Active Learning, Model Calibration, Positive-Unlabeled Learning, Label Noise Correction, and Subgroup Fairness.

1.2 Publications

I list below the journal and conference papers published by the time of writing:

- Jose M Alvarez, Alejandra Bringas Colmenarejo, Alaa Elobaid, Simone Fabbrizzi, Miriam Fahimi, Antonio Ferrara, Siamak Ghodsi, Carlos Mougan, Ioanna Papageorgiou, Paula Reyero, Mayra Russo, Kristen M Scott, Laura State, Xuan Zhao and Salvatore Ruggieri. Policy advice and best practices on bias and fairness in ai. *Ethics and Information Technology*, 26 (2):31, 2024.

- Xuan Zhao, Klaus Broelemann, Salvatore Ruggieri and Gjergji Kasneci. Enhancing Fairness through Reweighting: A Path to Attain the Sufficiency Rule. *Accepted at ECAI 2024*.

- Xuan Zhao, Klaus Broelemann, Salvatore Ruggieri and Gjergji Kasneci. Causal Fairness-Guided Dataset Reweighting using Neural Networks. *2023 IEEE International Conference on Big Data (BigData)*, 2023, pp. 1386-1394.

- Simone Fabbrizzi, Xuan Zhao, Emmanouil Krasanakis, Symeon Papadopoulos and Eirini Ntoutsi. Studying Bias in Visual Features through the Lens of Optimal Transport. *Data Mining and Knowledge Discovery*, vol. 38, no. 1, 2024, pp. 281-312.

- Xuan Zhao, Klaus Broelemann and Gjergji Kasneci. Counterfactual Explanation for Regression via Disentanglement in Latent Space. *2023 IEEE International Conference on Data Mining Workshops (ICDMW)*, 2023, pp. 976-984.

- Xuan Zhao, Klaus Broelemann and Gjergji Kasneci. Counterfactual Explanation via Search in Gaussian Mixture Distributed Latent Space. *IJCAI 2023 XAI Workshop*.

- Alejandra Bringas Colmenarejo, Luca Nannini, Alisa Rieger, Kristen M Scott, Xuan Zhao, Gourab K Patro, Gjergji Kasneci and Katharina Kinder-Kurlanda. Fairness in Agreement with European Values: An Interdisciplinary Perspective on AI Regulation. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics*, and Society, 2022, pp. 107-118.

I also include a list of papers under submission at the time of writing:

- Xuan Zhao, Simone Fabbrizzi, Paula Reyero Lobo, Siamak Ghodsi, Klaus Broelemann, Steffen Staab and Gjergji Kasneci. Adversarial Reweighting Guided by Wasserstein Distance for Bias Mitigation.

Not all of the publications listed above are included in this thesis. I will highlight when a chapter is based on a published work.

Chapter 2

Adversarial Reweighting Guided by Wasserstein Distance for Bias Mitigation

This chapter is based on the paper: Zhao, X., et al. Adversarial Reweighting Guided by Wasserstein Distance for Bias Mitigation. It is currently under submission.

Machine learning models trained on personal data may discriminate against groups with sensitive attributes. Broadly speaking, there are three major paradigms to address this problem. The first paradigm assumes fairness can be measured. Then, the minimization of unfairness metrics is integrated in the empirical risk minimization as a multi-objective optimization problem (Aghaei *et al.*, 2019; Berk *et al.*, 2017). The second paradigm assumes that discrimination arises from the use of protected (i.e., sensitive) attributes and those correlated to them. Removing sensitive information from the input data can support learning fair models (Creager *et al.*, 2019). The third paradigm builds on the assumption that discrimination arises from biased labeling processes (e.g., through biased domain knowledge or biased human feedback). Corresponding approaches aim at identifying and correcting label bias (Jiang and Nachum, 2020; Krasanakis *et al.*, 2018). These paradigms do not deal directly with the issue that, by definition, minority groups are smaller than the majority. The effects of under-represented data samples in the learning process are 'overridden' by the prevalence of data samples from the majority group. The under-representation negatively can hide undesirable correlations between attributes in the minority group. That is, leading to a *representation bias*.

We propose a reweighting scheme to mitigate predictive quality issues arising from the imbalance among sensitive groups. We transform the data into a latent space and use the critic of a Wasserstein Generative Adversarial Network (WGAN) with gradient penalty (Gulrajani *et al.*, 2017) to approximate distances between samples from the minority and reweighted majority groups. We train the model to ensure that the data distribution is non-discriminatory with respect to sensitive groups, thereby mitigating disparate impact. Simultaneously, the empirical risk for the classification task is minimized. The rationale for our method is that if subgroups are sufficiently represented in a non-discriminatory way, *bias in prediction* would be substantially reduced, if not elimi-

nated (Chai and Wang, 2022). Our method is different from existing adversarial methods (Adel *et al.*, 2019; Madras *et al.*, 2018; Stanczuk *et al.*, 2021) in exploiting the competition between the reweighing component and the discriminator of the GAN framework. We perform experiments on different datasets and compare with four state-of-the-art fairness-aware methods. Our method outperforms its competitors in mitigating disparate impact while maintaining high prediction quality, as demonstrated by the experimental evaluation of image and tabular benchmark datasets.

We summarize our contribution as follows: (1) We formulate a novel data transformation and sample-based reweighting method for mitigating representation bias related to sensitive groups in classification tasks. (2) We show theoretically that by closing the Wasserstein distance gap between sensitive groups in the latent space during training, our reweighting approach leads to predictions that adhere to demographic parity. (3) We provide a thorough evaluation of the proposed technique on image and tabular benchmark datasets and show the viability of our approach with respect to robustness to fairness, accuracy and label noise.

2.1 Problem Setting

The disproportionate presence of various demographic groups within a sample population can result in the marginalization of minority communities when machine learning algorithms are used for automated decision-making. The inherent lack of adequate representation of minority groups in the dataset obscures the disparities in treatment among different subpopulations, posing challenges in addressing them effectively during the learning process. The inherent nature of minority groups being smaller in size compared to the majority is a recognized issue. In the learning process, the impact of under-represented data samples is overshadowed by the abundance of data from the majority group. This under-representation detrimentally impacts the effectiveness of fairness metrics, potentially masking problematic correlations among attributes within the minority group. This phenomenon is commonly referred to as *representation bias*.

In this study, we analyze binary classifiers that generate predictions $\hat{y} \in \{0,1\}$ for a given dataset $\mathcal{D} = \{(x_1, y_1), \ldots, (x_n, y_n) | x_i \in X \subseteq \mathbb{R}^d, y_i \in Y = \{0,1\}\}$, where x_i denotes attribute vectors, and y_i represents the target label of data instance i. The first component of x_i, denoted as $s_i = x_{i,1} \in \{0,1\}$, describes the sensitive attribute. The values of the sensitive attributes s_i distinguish between the majority group, comprising n_p samples, and the minority (i.e., sensitive or under-represented) group, comprising n_u samples. Without loss of generality, we assume that $\forall i : 1 \leqslant i \leqslant n_p \Rightarrow s_i = 1. \forall i : n_p + 1 \leqslant i \leqslant n \Rightarrow s_i = 0$.

2.2 Our Adversarial Reweighting Approach

2.2.1 Problem formulation for representation bias

We introduce an innovative adversarial reweighting technique aimed at mitigating the effects of *representation bias*. Our method aims to equalize the distribution of data between the majority and minority groups by reducing the emphasis on samples from the majority group. To optimize empirical risk, our approach prioritizes samples from the majority group that exhibit proximity to the minority group, as assessed through the Wasserstein distance.

Consider a feature extractor $F_\phi : X \to Z \subseteq \mathbb{R}^k$ that converts raw data from the dataset $\mathcal{D}$ into a latent feature space. This transformation function acts as an embedding component, facilitating more efficient comparison of instances in the latent space. A binary classifier $C_\theta : Z \to Y$ with parameters θ maps the outputs of the transformation $F_\phi(x)$ to a binary label $\hat{y} \in Y = \{0,1\}$. For simplicity, we illustrate our approach in the context of a binary sensitive attribute. However, extending our method to handle a multi-categorical sensitive attribute or multiple sensitive attributes is straightforward. This extension involves designating one subgroup as a reference group and iteratively reweighting other subgroups to achieve demographic parity.

As part of our problem formulation, we specify the training objective as minimizing the weighted empirical risk, which is defined as:

$$\min_{\theta} \sum_{i=1}^{n} w_i \mathcal{L}(y_i, (C_\theta \circ F_\phi)(x_i)), \text{ with } w_i \geqslant 0 \tag{2.1}$$

with $\mathcal{L}$ denoting the cross-entropy loss, the overarching pipeline we aim to construct is depicted in Figure 2.1. The inclusion of a feature extractor is discretionary and might be unnecessary given the low dimensionality of the data.

When all training samples are assigned equal weights, the classifier tends to prioritize the majority group, thus introducing representation bias. We aim to preserve the weights for samples from the minority group (i.e., $\forall i : s_i = 0 \Rightarrow w_i = 1$) while reducing the weights for samples from the majority group. This adjustment ensures that the total weights remain consistent across both groups:

$$\sum_{i=1}^{n_p} w_i = n_u \tag{2.2}$$

In order to prevent information loss caused by assigning zero weights to certain samples from the majority group, we introduce a regularization constraint into our risk minimization term:

$$\sum_{i=1}^{n_p} (w_i - \frac{n_u}{n_p})^2 \leqslant T n_u \tag{2.3}$$

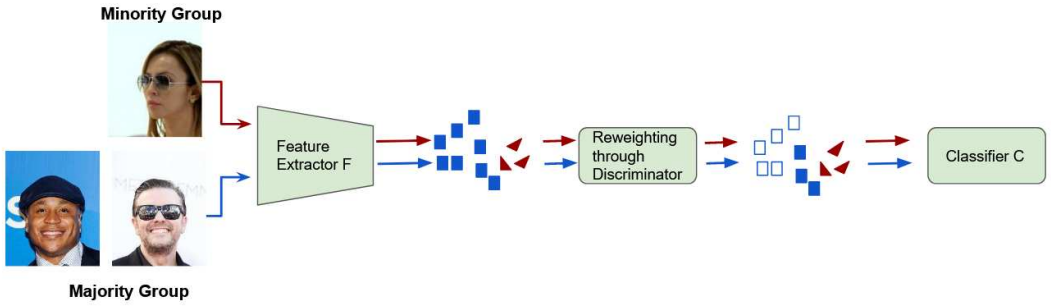


Figure 2.1: Architecture of our approach. The arrows show the computational flow for the minority (resp. majority) group in the classification task (e.g., predicting whether a person in the image is wearing a hat). Representation bias is indicated by blue and red triangles. Both minority and majority groups are mapped onto a latent space by the feature extractor. Then, majority group instances are reweighted to match the minority group distribution, aiming to decrease the distance with respect to the sensitive attribute.

The sum is minimal (namely zero) if $\forall i : w_i = \frac{n_u}{n_p}$. By manipulating the value of T, we can achieve a balance between the similarity and dissimilarity of the weights assigned to samples from the majority group.

Collectively, Equations (2.1), (2.2), and (2.3) delineate the problem space. However, individually, they do not comprehensively address differences within and between sample groups, thus not consistently enhancing group-based fairness metrics. The current challenge is to devise a weighting scheme that satisfies Equations (2.1), (2.2), and (2.3) while effectively mitigating representation bias in a robust manner.

2.2.2 Adversarial reweighting for demorgraphic parity

In the following, we first show that enforcing a small Wasserstein distance in the latent space ensures small distance in the prediction space. Then, we discuss the detailed adversarial reweighting model.

Our weighting scheme aims to determine weights such that the distributions of the majority and minority groups become more similar, thereby reducing the likelihood of biased and unfair predictions by the classifier. Through adversarial learning of data weights in the majority group, our objective is to focus more on samples within the majority group that closely resemble those in the minority group during training, while retaining valuable information from other samples in the majority group. We gauge the similarity of weighted distributions using the Wasserstein distance in the latent space, as approximating this distance in a low-dimensional space is computationally more efficient.

In the subsequent discussion, we initially demonstrate that enforcing a small Wasserstein distance in the latent space correlates with a small distance in the prediction space. Subsequently, we delve into the intricacies of our adversarial reweighting model.

Theoretical proof of enforcing demographic parity We demonstrate that ensuring a small Wasserstein distance, denoted as $W(\cdot,\cdot)$, in the latent space also leads to a small Wasserstein distance in the prediction space.

Proposition 1. *Given two measures μ and ν over a metric space (Z,d_Z) and a K-Lipschitz function $C:(Z,d_Z)\to(Y,d_Y)$, we have that*

$$W(C_{\#}\mu,C_{\#}\nu)\leq K\cdot W(\mu,\nu)$$

Where $C_{\#}\mu$ is the push-forward measure along the function C.

Proof. The proof goes as follows:

$$\begin{aligned}
W(C_{\#}\mu,C_{\#}\nu) &= \sup_{f\in Lip_1(Y)} \int_Y f dC_{\#}\mu - \int_Y f dC_{\#}\nu \\
&\text{(Kantorovich duality)} \\
&= \sup_{f\in Lip_1(Y)} \int_Z f\circ C d\mu - \int_Z f\circ C d\nu \\
&\text{(Property of the push-forward)} \\
&= \sup_{f\in Lip_1(Y)} K\cdot\left(\int_Z \frac{f\circ C}{K} d\mu - \int_Z \frac{f\circ C}{K} d\nu\right) \\
&\leq \sup_{h\in Lip_1(Z)} K\cdot\left(\int_Z h d\mu - \int_Z h d\nu\right) \\
&(\frac{f\circ C}{K} \text{ is 1-Lipschitz}) \\
&= K\cdot W(\mu,\nu)
\end{aligned} \tag{4}$$

Where $Lip_1(Y)$ indicates the set of 1-Lipschitz functions $f:Y\to\mathbb{R}$. □

Please note that when dealing with classifiers over a finite dataset, the K-Lipschitz condition for a binary classifier $C:Z\to\{0,1\}$ implies that for every pair of points z and z' such that $C(z)\neq C(z')$, the distance between them, denoted as $d_Z(z,z')$, satisfies $\frac{1}{K}\leq d_Z(z,z')$. This condition holds because the set $\{0,1\}$ is equipped with the discrete metric. Given that we are working with finite datasets, we can always find a suitable K. It's essential to note that this result is specific to a particular dataset and doesn't generalize unless we assume that the condition $\frac{1}{K}\leq d_Z(z,z')$ holds true for every z and z' such that $C(z)\neq C(z')$ also applies to new data.

Therefore, if the Wasserstein distance is close to 0 in the latent space, it will also be close to 0 in the prediction space. Here, $W(C_{\#}\mu,C_{\#}\nu)=0$ implies that $C_{\#}\mu=C_{\#}\nu$, indicating demographic parity. In essence, defining the distribution $\zeta:=\frac{1}{2}\mu+\frac{1}{2}\nu$ represents the probability of being sampled from either the majority or minority groups. As both μ

and ν are discrete, we have $C_{\#}\zeta = \frac{1}{2}C_{\#}\mu + \frac{1}{2}C_{\#}\nu$. Thus, $C_{\#}\mu = C_{\#}\nu$ implies that the probability of $C(z) = 1$ is independent of whether z is sampled from the majority or minority groups.

Adversarial reweighting model

We estimate the Wasserstein distance computationally by employing a neural network discriminator D and implementing the gradient penalty technique, as introduced in WGAN-GP (Gulrajani *et al.*, 2017):

$$W(\mu,\nu) \approx \max_{\theta_D} \left(\mathbb{E}_{z\sim\mu}[D(z;\theta_D)] - \mathbb{E}_{z'\sim\nu}[D(z';\theta_D)]\right) \tag{2.5}$$

We define the (weighted) empirical distributions of the minority group $\mathcal{P}_U$ and the majority group $\mathcal{P}_P(\mathbf{w})$ using the Dirac delta function $\delta(\cdot)$ as:

$$\mathcal{P}_U = \frac{1}{n_u}\sum_{i=n_p+1}^{n} \delta(F(x_i)), \tag{2.6}$$

$$\mathcal{P}_P(\mathbf{w}) = \frac{1}{n_u}\sum_{i=1}^{n_p} w_i\delta(F(x_i)), \text{ with } \sum_{i=1}^{n_p} w_i = n_u \tag{2.7}$$

Then, we optimize the weights by minimizing the Wasserstein distance between the minority and reweighted majority distributions, whereby Equations (2.2) and (2.3) define the solution space for the weights $\mathcal{W} = \{\mathbf{w} : \mathbf{w} = (w_1, w_2, ..., w_{n_p})^T, w_i \geqslant 0, \sum_{i=1}^{n_p} w_i = n_u, \sum_{i=1}^{n_p}(w_i - \frac{n_u}{n_p})^2 \leqslant Tn_u\}$:

$$\min_{\mathbf{w}\in\mathcal{W}} W(\mathcal{P}_U, \mathcal{P}_P(\mathbf{w})) \tag{2.8}$$

Because of Proposition 1, we know that such minimization contributes to reducing the disparity between majority and minority groups.

If f is a measurable function and $\mu = \sum \alpha_i\delta(x_i)$ a discrete distribution, we have that $f_{\#}\mu = \sum \alpha_i\delta(f(x_i))$. Hence, combining Equations (2.5) and (2.8) results into a minmax problem, yields:

$$\min_{\mathbf{w}\in\mathcal{W}}\max_{\theta_D}\left(\sum_{i=1}^{n_p} w_i D(z_i^p;\theta_D) - \sum_{i=1}^{n_u} D(z_i^u;\theta_D)\right) \tag{2.9}$$

In Equation (2.9), the discriminator is trained to maximize the average of its outputs on both the minority and majority groups. Adversarially, the weights assigned to samples from the majority group are adjusted to minimize the (reweighted) average of the discriminator's outputs. Consequently, samples from the majority group with lower discriminator outputs (indicating proximity to the minority group) receive higher weights. Thus, formulating the reweighted cross-entropy loss based on the (reweighted) data distribution in Equation (2.1) helps alleviate representation bias concerning the minority

groups.

2.2.3 Training algorithm

To train the feature extractor F_ϕ and the classifier network C_θ, we update the network parameters (ϕ, θ) and learn the weights w with the discriminator D while keeping others fixed. This training process involves alternating between the following two steps:

Step 1: Updating ϕ and θ while fixing w and θ_D. We update ϕ and θ to minimize the loss defined in Equation (2.1) for S steps in batches, while keeping w and θ_D fixed.

Step 2: Updating w and θ_D while fixing ϕ and θ. We obtain embeddings of training data from both majority and minority groups using the feature extractor F, while keeping ϕ and θ fixed. The weights w in Equation (2.9) are learned through a min-max optimization problem. Specifically, we optimize the weights $\mathbf{w}$ and the parameters θ_D of the discriminator alternatively. Initially, we set $w_i = \frac{n_u}{n_p}$ for all i and optimize θ_D to maximize the objective function in Equation (2.9) using the gradient penalty technique, similar to WGAN-GP (Gulrajani *et al.*, 2017). Then, with the discriminator fixed, we optimize $\mathbf{w}$. Here, we denote $d_i = D(F_\theta(x_i); \theta_D)$ and $\mathbf{d} = (d_1, d_2, ...d_{n_p})^T$. The optimization for $\mathbf{w}$ becomes a constrained least squares problem:

$$\min_{\mathbf{w}} \mathbf{d}^T \mathbf{w}, s.t. w_i \geqslant 0, \sum_{i=1}^{n_p} w_i = n_u, \sum_{i=1}^{n_p} (w_i - \frac{n_u}{n_p})^2 \leqslant T n_u \tag{2.10}$$

2.3 Experiments

We assess the performance of our reweighting approach on three benchmark datasets, comparing it to eight methods using four different metrics: Accuracy, Disparate Impact, Disparate FPR, and Disparate FNR. Inspired by Chai and Wang (2022) we compare against (1) **Baseline** (Neural Network (NN) based classification without fairness constraints); (2) Simple **Reweighing**: NN classification with assigning same balancing weights to samples the majority group; (3) **Undersampling** forms the training dataset by balancing group sizes via undersampling from the majority group; (4) **Oversampling** balances group sizes by repeating sampling from the minority group.

We choose four further competing methods mentioned in earlier sections: (5) Adaptive sensitive reweighting (**ASR**) reweights samples to balance target class occurrences. (6) Wasserstein fair classification (**WFC**) matches quantiles of the predictive distribution of the sensitive group to the all-group Wasserstein barycenter. (7) Fair Adversarial Discriminative (**FAD**) model (Adel *et al.*, 2019) decorrelates the sensitive information from the embeddings by adjusting the encoding/feature extraction process using adversarial training. (8) Fairness with Adaptive Weights (**FAW**) constrains the sum of weights. They (i) are designed to address bias, (ii) follow conceptually similar strategies, and (iii) can also be flexibly applied to different modalities (tabular and images).

(a) male samples assigned with the lowest weights

(b) male samples assigned with the highest weights

Figure 2.2: Samples from the male group with the lowest and highest weights. Samples with the lowest weights tend to wear suits and have short hair, while samples with the highest weights tend to have longer hair.

We trained our networks on an Intel(r) Core(TM) i7-8700 CPU. The networks used in our experiments were constructed using PyTorch (Paszke *et al.*, 2019), and the optimization described in Equation (2.10) was carried out with the CVXPY Python package (Diamond and Boyd, 2016).

2.3.1 Data and training details

Image dataset

We create a dataset derived from CelebA (Liu *et al.*, 2015), which consists of 90% male images and 10% female images to highlight the issue of under-representation. The classification task is to determine whether the individual in the image is wearing a hat. For the feature extractor F, we use the ResNet-18 (He *et al.*, 2016) architecture, pre-trained on ImageNet (Deng *et al.*, 2009), but without the final fully-connected layer for simplicity. The feature extractor F_ϕ and classifier C_θ are updated using the stochastic gradient descent (SGD) algorithm (Shamir and Zhang, 2013) with a momentum of 0.9. For the

discriminator D, we employ a similar architecture to Gulrajani *et al.* (2017), comprising three fully connected layers with 512, 256, 128, 64, and 1 node, respectively, and excluding the final sigmoid function. The Adam algorithm (Kingma and Ba, 2015) is used to update θ_D with a learning rate of 0.0001. Following Gulrajani *et al.* (2017), we adjust the learning rate η using $\eta = \frac{0.01}{(1+10p)^{-0.75}}$, where p linearly increases from 0 to 1 as training progresses. The batch sizes are set to n_p = 900 and n_u = 100. We update ϕ and θ for four steps before updating θ_D for one step. A larger batch size is chosen to accurately estimate the Wasserstein distance between distributions. Additional training details, such as the split between training and testing datasets, are provided in the Appendix A.2.

Tabular dataset

For experiments involving tabular data, we use the Adult dataset (Kohavi, 1996) and the UCI German Credit Risk dataset (Dua and Graff, 2017). More details about these datasets can be found in the Appendix A.1. It is important to note that tabular datasets typically require more preprocessing than image datasets (Borisov *et al.*, 2022). Although we recognize that gradient boosting would be more suitable for tabular data, we were unable to find any related methods for mitigating representation bias based on boosting. Additional training details are provided in the Appendix A.2.

2.3.2 Analysis results

Performance comparison

From Table 2.1 to Table 2.3, our approach demonstrates no loss in accuracy. Concurrently, it effectively mitigates Disparate Impact regarding the sensitive attribute, highlighting a significant advantage of our optimization over related methods. Our method is both theoretically designed and empirically validated to reduce Disparate Impact. However, there are cases where its performance may not meet expectations or may unintentionally worsen Disparate Mistreatment. To gain a clearer understanding of our method's performance, we present the accuracy for male and female groups separately in Table A.4 in the Appendix. Here, we primarily analyze Accuracy and Disparate Impact: WFC did not perform as well in terms of accuracy. We believe this is due to its adjustment of prediction results by aligning the Wasserstein distance between predictions over sensitive groups, which can decrease accuracy. While ASR is a strong competitor, it requires multiple training iterations to achieve neural network convergence, making it more resource-intensive than other approaches. The same applies to FAW. FAD struggles with imbalance during decorrelation (which also tends to focus more on the majority group) but maintains a high accuracy rate.

Figure 2.2 illustrates samples from the male group that are assigned the lowest and highest weights. We observe that male samples which are more distant from the female distribution receive lower weights, thereby balancing and harmonizing the male and fe-

Table 2.1: Experimental results of classifier (Wearing Hat) on CelabA

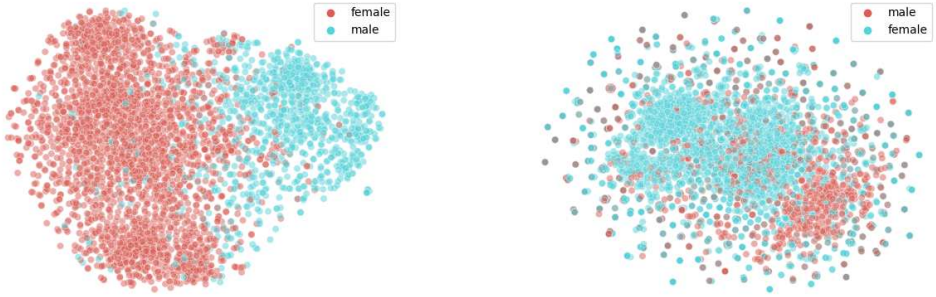
	simple methods				state-of-the-art methods				
	baseline	reweighting	undersampling	oversampling	ASR	WFC	FAD	FAW	Ours
Accuracy rate (%)	95.0±0.5	93.7±0.7	94.2±0.8	94.1±1.4	94.5±0.5	92.3±0.4	94.6±0.7	92.5±0.7	**95.3±0.2**
Disparate impact (%)	1.9±0.6	3.3±0.3	2.1±0.3	1.5±0.3	2.0±0.4	7.9±2.6	27.2±1.0	1.9±0.5	**0.2±0.3**
Disparate FPR (%)	-33.0±1.3	-27.2±5.1	-29.2±1.1	-35.5±2.1	-23.0±3.4	**17.7±1.3**	-30.8±1.7	-18.7±1.7	-22.8±1.0
Disparate FNR (%)	30.0±1.2	36.6±4.2	30.0±1.3	29.8±2.4	26.8±0.9	**22.2±1.8**	26.9±1.8	26.0±1.3	26.2±2.4

Table 2.2: Experimental results of classifier on Adult dataset

	simple methods				state-of-the-art methods				
	baseline	reweighting	undersampling	oversampling	ASR	WFC	FAD	FAW	Ours
Accuracy rate (%)	83.1±0.4	82.5±0.3	82.1±0.3	82.1±0.9	81.6±0.3	81.8±0.5	82.4±0.5	81.2±0.6	**83.0±0.1**
Disparate impact (%)	17.8±0.3	21.0±0.4	18.7±0.5	18.6±0.4	**0.4±0.2**	2.5±1.0	5.7±1.4	1.7±0.4	0.8±0.5
Disparate FPR (%)	17.0±1.0	**2.3±4.7**	9.2±1.3	8.4±1.6	27.2±4.5	-9.8±0.7	-8.7±1.8	-8.5±2.4	-10.5±1.1
Disparate FNR (%)	6.1±0.6	12.1±3.5	4.2±0.8	12.7±0.7	**2.3±1.2**	22.4±1.3	3.2±0.7	4.0±1.7	7.2±0.7

Table 2.3: Experimental results of the classifier on German Credit dataset

	simple methods				state-of-the-art methods				
	baseline	reweighting	undersampling	oversampling	ASR	WFC	FAD	FAW	Ours
Accuracy rate (%)	70.1±1.4	69.2±1.3	65.5±1.1	67.3±2.9	69.1±2.4	67.0±0.5	68.9±0.5	69.5±0.5	**70.0±1.4**
Disparate impact (%)	15.4±9.3	8.9±3.4	11.9±6.5	9.2±4.7	6.7±2.5	3.2±1.0	3.5±0.4	3.8±0.5	**1**.1±1.2
Disparate FPR (%)	7.5±4.8	19.0±3.7	13.4±7.3	12.5±8.5	**1.5±0.9**	7.2±0.8	**1.5±0.2**	9.2±2.8	5.3±1.7
Disparate FNR (%)	6.3±3.2	9.0±3.2	15.7±8.8	11.8±8.7	**0.9±0.4**	5.2±2.3	1.1±0.3	4.9±2.5	6.7±0.9



(a) t-SNE before reweighting in the latent space (b) t-SNE after reweighting in the latent space

Figure 2.3: t-SNE of extracted embeddings before and after reweighting

male distributions. Conversely, male samples that are closer to the female distribution are given relatively higher weights, providing additional information for the classification task.

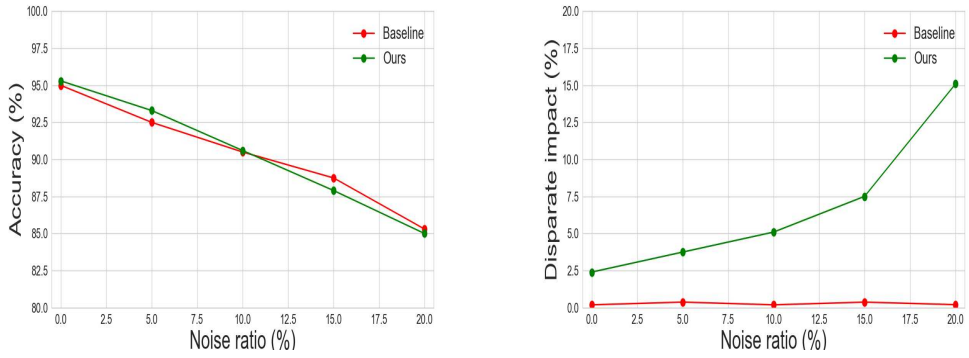
Embeddings and reweighting visualization

We visualize the learned weights of the majority group versus the minority group in the CelebA dataset by showing the t-SNE embeddings of the original and reweighted embeddings in Figure 2.3. In Figure 2.3a on the left, the male and female groups are not well-aligned, leading to discrimination against the female group as discussed earlier. Our proposed reweighting method aligns the embeddings of the female group with those of the male group before the classification step, as shown in Figure 2.3b. These visualizations partially illustrate the effectiveness of our approach in addressing representation bias related to a sensitive attribute. Additionally, the original Wasserstein distance between the two distributions is 15.87, which reduces to 0.23 after reweighting. For more details, please refer to Appendix A.2.

Classification with noisy label

Sensitivity to the choice of hyper-parameters

We have also examined the sensitivity of our method to the hyper-parameter T as shown in Figure A.1 in the Appendix. The plots indicate that our adversarial reweighting scheme's performance exhibits low sensitivity to the choice of this hyper-parameter. In our experiments, we set T to 5. For an analysis involving other datasets, refer to Figure A.2 in the Appendix.



(a) Accuracy vs. Noise ratio (b) Disparate Impact vs. Noise ratio

Figure 2.4: Change of accuracy and disparate impact under different noise ratios on CelebA. Note that we enforce Demographic Parity which is selection rate parity (which our method can implicitly mitigate) and we believe this is the reason why our method is so robust to noisy labeling.

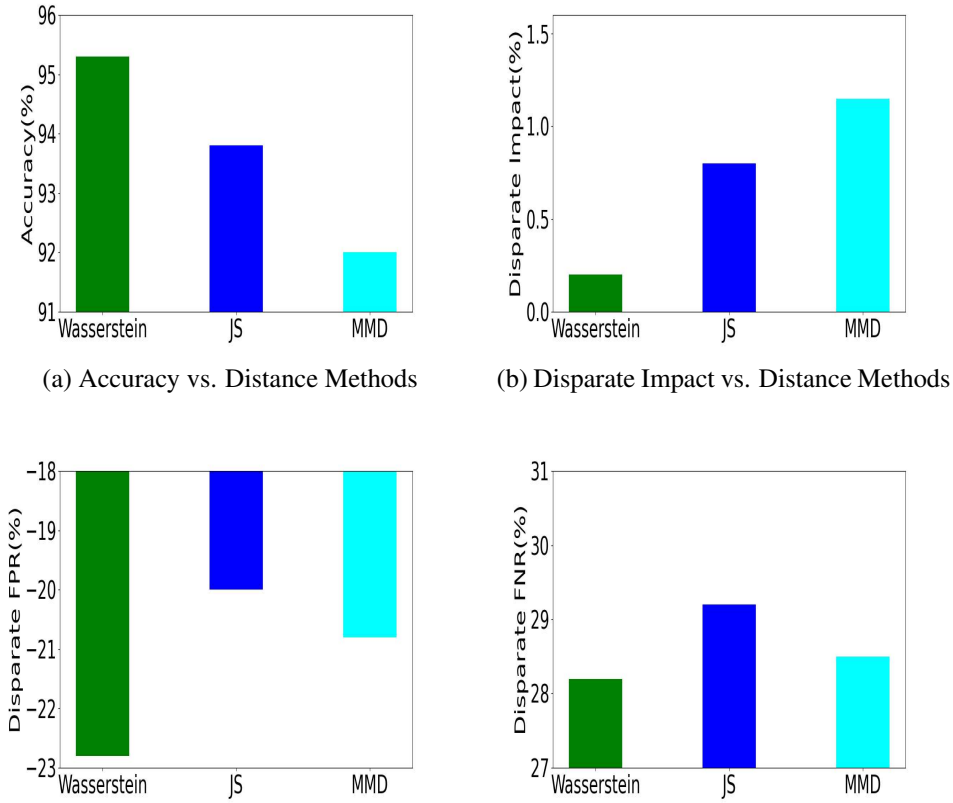
Ablation tests

Ablation test for MMD and JS-divergence dissimilarity measures.

We also performed an ablation test using Jensen-Shannon divergence (JS) and maximum mean discrepancy (MMD) instead of Wasserstein distance to learn the weights in our framework on the CelebA dataset. As shown in Figures 2.5 and A.5 (see Appendix), our method's performance using the Wasserstein distance surpasses that of JS and MMD. The Wasserstein distance may be more effective for measuring distances between distributions that are more disjoint. MMD with kernels may struggle to capture very complex distances in high-dimensional spaces compared to the Wasserstein distance. While the Wasserstein distance is superior for accuracy and disparate impact, it is not necessarily better for Disparate FPR and FNR.

Ablation test for assigning weights to both groups or only to the minority group.

Our method is specifically designed to reweight samples from the majority group in order to narrow the gap of the Wasserstein distance between the two sensitive groups. To investigate the assignment of weights to both groups, we conducted an ablation test where we alternately assigned weights to one group while keeping the other group with fixed weights until convergence. The Wasserstein distance between the two groups was measured at 0.17 for CelebA. Additionally, we tested assigning weights only to the minority group, resulting in a Wasserstein distance after reweighting of 1.29. As mentioned earlier, the Wasserstein distance achieved by our method (assigning weights only to the majority group) after reweighting is 0.23. This suggests that the difference between reweighting both groups and only the majority group is minimal, while the disparity between reweighting only the majority group and the minority group is substantial. Hence, we opt to reweight the majority group in our method. For further details, please consult



(a) Accuracy vs. Distance Methods

(b) Disparate Impact vs. Distance Methods

(c) Disparate FPR vs. Distance Methods

(d) Disparate FNR vs. Distance Methods

Figure 2.5: Sensitivity of different distance measures on CelebA dataset

Appendix A.2.

Convergence

Please refer to Figure A.4 in the Appendix for the convergence of test error in our method. It can be observed that our method achieves convergence in a time frame that is approximately similar to the baseline (NN-based classification without fairness constraint), albeit with a slightly higher test error.

2.4 Related Work

2.4.1 Adversarial methods

Existing adversarial fairness methods (Adel *et al.*, 2019; Madras *et al.*, 2018; Stanczuk *et al.*, 2021) utilize a discriminator in an in-processing manner to decorrelate embeddings

from the sensitive attribute. The authors of Choi *et al.* (2019) and Kim *et al.* (2019) propose minimizing mutual information between biased labels and embeddings through adversarial training. These approaches aim to disentangle the sensitive attribute in the latent space, but they do not address the under-representation of sensitive groups. In contrast, our work tackles representation bias in the decorrelation process by reweighting to align the distributions of sensitive groups, rather than solely adjusting the encoder.

2.4.2 Reweighting methods

The Fairness with Adaptive Weights approach (Chai and Wang, 2022) enforces equal sums of weights among sensitive groups, assigning weights to samples according to their likelihood of misclassification. Meanwhile, Adaptive Sensitive Reweighting to mitigate bias (Krasanakis *et al.*, 2018) assigns weights based on how well samples align with unobserved true labeling. For instance, Adversarial Reweighting for domain adaptation (Gu *et al.*, 2021) aims to align distributions of source and target domains, addressing the domain adaptation problem. In our work, we extend reweighting based on Wasserstein distance to the realm of fairness.

2.4.3 Imbalanced classification

There are two primary approaches to imbalanced classification: resampling and cost-sensitive learning. Resampling methods aim to achieve balance between class groups by either oversampling the group with a small size (typically the minority group in fairness settings) or undersampling the group with a large size (usually the majority group in fairness settings), or a combination of both. For example, Bao *et al.* (2020) conducts classification by utilizing clustering centers in the latent space to balance among the groups, effectively equivalent to undersampling all groups.

Cost-sensitive learning involves assigning higher weights to samples from groups with smaller sizes during training, thereby increasing the cost of misclassifying these samples compared to those from larger groups. Various methods exist for determining these weighting schemes. For instance, in Huang *et al.* (2020), the authors balance group representations by imposing constraints on the embedding to maintain inter-cluster margins within and between classes. It's important to note that these methods address class imbalance, while our work specifically targets imbalance related to sensitive attributes.

2.4.4 Fairness notions

Disparate treatment (Zafar *et al.*, 2017b) arises when the classifier produces different predictions for individuals with identical input features but from different sensitive groups. To mitigate this, the classifier should be calibrated across sensitive groups: $P(\hat{y}|x,s) = P(\hat{y}|x)$.

Disparate impact (Kamiran and Calders, 2011) assesses the difference in positive outcome rates between groups and is eradicated when the predictive outcome $\hat{y}$ is independent of s: $P(\hat{y}|s=0) = P(\hat{y}|s=1)$. However, eliminating disparate impact doesn't guarantee a fair classifier. Uneven sample distribution among sensitive groups may cause the classifier to prioritize the majority group while overlooking decisions regarding the minority group. Moreover, achieving zero disparate impact may come at the expense of classifier performance since statistical characteristics of different sensitive groups often differ.

Disparate mistreatment (Hardt *et al.*, 2016) arises when misclassification rates, including false positives and false negatives, vary across different sensitive groups. Assessing disparate mistreatment requires labeled data.

Previous studies, such as Chouldechova (2017), have noted that there is typically a trade-off among the criteria for disparate mistreatment. Disparate False Positive Rate (FPR) and Disparate False Negative Rate (FNR) are commonly utilized to mitigate disparate mistreatment: $P(\hat{y} \neq y|y=1,s) = P(\hat{y} \neq y|y=1)$ and $P(\hat{y} \neq y|y=0,s) = P(\hat{y} \neq y|y=0)$.

2.4.5 Wasserstein distance methods

Definition The Wasserstein distance between two distributions μ and ν is defined by $W(\mu,\nu) = \min_{\pi\in\Pi}\mathbb{E}_{(x,x')\sim\pi}[\|x-x'\|^p]$, where Π is the set of couplings of μ and ν, i.e., $\Pi = \{\pi|\int\pi(x,x')dx' = \mu(x), \int\pi(x,x')dx = \nu(x')\}$, and $p \geqslant 1$. In the later sections of this work, $p=2$. Following the Kantorovich-Rubinstein duality, we have the dual form of the Wasserstein distance of $W(\mu,\nu) = \max_{\|f\|_L\leqslant 1}\mathbb{E}_{x\sim\mu}[f(x)] - \mathbb{E}_{x'\sim\nu}[f(x')]$, where the maximization is over all 1-Lipschitz functions f: $\mathbb{R}^d \to \mathbb{R}$.

2.4.6 Fairness-aware classification

The Wasserstein distance, also referred to as Optimal Transport (OT) distance, is a metric within the space of measures with finite moments, providing a means to assess the dissimilarity between two distributions. A significant application of this metric's properties is the computation of the barycenter of two distributions. This technique has been utilized in fairness-aware classification methods (Zehlike *et al.*, 2020; Jiang *et al.*, 2019) to enforce statistical parity. Notably, Wasserstein Fair Classification (WFC) (Jiang *et al.*, 2019) matches the predictions of the sensitive group quantiles to the predictions of the barycenter of all groups.

Fairness with Continuous Optimal Transport (Chiappa and Pacchiano, 2021) presents a stochastic-gradient fairness approach that relies on a dual formulation of continuous Optimal Transport (OT) instead of discrete OT, aiming to enhance performance.

2.4.7 Generative methods

Wasserstein Generative Adversarial Networks (WGANs) (Gulrajani *et al.*, 2017) aim to minimize the Wasserstein distance between a real and a generated distribution by employing weight clipping to impose a Lipschitz constraint on the critic, thereby enhancing the performance of conventional GANs. WGANs with Gradient Penalty (WGAN-GP) (Gulrajani *et al.*, 2017) offer a more relaxed version of the Lipschitz constraint, stipulating that functions are 1-Lipschitz if their gradients have a norm of at most 1 everywhere.

2.5 Discussion and Conclusion

Our approach represents a conceptual departure from previous efforts in fairness-aware machine learning. We focus on balancing and harmonizing protected groups defined by sensitive attributes, minimizing empirical risk, and achieving competitive predictive accuracy, fairness, and robustness. Theoretical literature has extensively explored the inherent balance between fairness and utility, and empirical experiments have consistently demonstrated trade-offs in practice. However, these discussions often revolve around a fixed distribution that may not align with real-world scenarios. We contend that an ideal distribution exists where fairness and utility coexist harmoniously. Through our combined approach of data reweighing and classifier training, we aim to transcend biased distributions and alleviate trade-offs. Our experiments indicate that our method can mitigate the need to discard sensitive attributes or impose specific fairness constraints, thus circumventing challenges associated with determining critical hyperparameters such as regularization factors. One limitation of our work is that WGAN-GP may struggle to accurately approximate the Wasserstein distance (Mallasto *et al.*, 2019; Stanczuk *et al.*, 2021).

Chapter 3

Causal Fairness-Guided Dataset Reweighting using Neural Networks

This chapter is based on the conference paper: Zhao, X., K. Broelemann, S. Ruggieri, and G. Kasneci. Causal Fairness-Guided Dataset Reweighting using Neural Networks. *2023 IEEE International Conference on Big Data (BigData)*, 2023, pp. 1386-1394.

Pre-processing data to meet fairness criteria is a significant research area in machine learning. Models trained on biased datasets can internalize these biases and apply them, resulting in discriminatory decisions against socially sensitive groups based on gender, race, age, or other protected categories (Pedreschi *et al.*, 2008; Zliobaite *et al.*, 2011; Hardt *et al.*, 2016; Zhang *et al.*, 2017, 2018c). Many methods have been proposed to adjust the training data in order to reduce biases and meet specific fairness criteria (Zhang *et al.*, 2017, 2019; Edwards and Storkey, 2016; Xie *et al.*, 2017; Madras *et al.*, 2018; Zhang *et al.*, 2018b).

For reliable and effective treatment, especially in legal contexts, discrimination claims typically need to demonstrate causal links between sensitive attributes and questionable decisions (or predictions) rather than just associations or correlations. Unlike fairness notions based on correlation, causality-based fairness notions and methods incorporate an understanding of the causal structure of the problem. This understanding often uncovers the data generation mechanism, aiding in the interpretation of how sensitive attributes affect decision outcomes. Causal fairness aims to address the underlying causes of disparities rather than merely attempting to eliminate them in a post-hoc manner.

We base our research framework on the ideas and concepts presented in CFGAN (Xu *et al.*, 2019). However, instead of generating a fair dataset as in CFGAN, we propose a method that reweights samples to meet fairness criteria. This method employs two neural networks to represent causal and interventional graphs, along with a discriminator to guide the reweighting process. The general requirement for modifying datasets is to maintain data utility for downstream tasks as much as possible. The intuition behind our reweighting scheme is that within a dataset, some individuals are treated more fairly within the causal mechanism. By assigning higher weights to these individuals, we can slightly adjust the underlying causal mechanism to achieve fairness without significantly impacting the performance of downstream tasks. This approach aims to mitigate histor-

ical bias and, by examining the high and low weights assigned to samples, provides a high-level understanding of these biases.

The experiments (Section 3.3) demonstrate that reweighted data surpass generated data in terms of utility. Within the classification of bias mitigation methods into pre-processing, in-processing, and post-processing (Roh *et al.*, 2021; Calmon *et al.*, 2017; Aghaei *et al.*, 2019; Berk *et al.*, 2017; Hardt *et al.*, 2016), our method is categorized as pre-processing, as it modifies the dataset before it is used by the downstream learning algorithm. Consequently, like other pre-processing methods, our approach is model-agnostic.

Our contributions can be summarized as follows: (1) We introduce a novel, sample-based reweighting method to mitigate various causal biases related to sensitive groups. (2) We demonstrate that by simulating the underlying causal model that represents the causal relationships in real data, and the causal model post-intervention, with the assistance of a discriminator, our reweighting method results in fair reweighted data. (3) We thoroughly evaluate the proposed technique using benchmark datasets and validate the effectiveness of our approach.

3.1 Preliminary

Throughout this paper, we consider a structural causal model $\mathcal{M} = \langle U, V, F \rangle$, that is learned from a dataset $\mathcal{D} = \{(s_i, x_i, y_i)\}_{i=1}^{n}$ where $s_i \in S = \{0,1\}, x_i \in X \subseteq \mathbb{R}^d, y_i \in Y = \{0,1\}$.

1) U denotes exogenous variables that cannot be observed but constitute the background knowledge behind the model. $P(U)$ is a joint probability distribution of the variables in U.

2) V denotes endogenous variables that can be observed. In our work, we set $V = \{S, X, Y\}$. S represents the sensitive attribute, Y represents the outcome attribute, and X represents all other attributes. Additionally, s^+ is used to denote $S = 1$ and s^- to denote $S = 0$.

3) F denotes the deterministic functions. For each $V_j \in V$, there is a corresponding function f_{V_j} that maps from domains of the variables in $Pa_{V_j} \cup U_{V_j}$ to V_i, namely $V_j = f_{V_j}(Pa_{V_j}, U_{V_j})$. Here, $Pa_{V_j} \subseteq V \backslash V_j$ represents the parents of V_j, and U_{V_j} also represents the parents (exogenous variables) of V_j, $U_{V_j} \subseteq U$.

We denote by $\mathcal{G}$ the causal graph $\mathcal{G}$ associate with $\mathcal{M}$, and assume it is a Directed Acyclic Graph (DAG).

3.1.1 Causal Fairness Criteria

To understand causal effects in the causal model $\mathcal{M}$, we can use the do-operator (Pearl, 2009), which represents a physical intervention that sets a variable $S \in V$ to a constant value s. By performing an intervention $do(S = s)$, we replace the original function $S =$

$f_S(Pa_S, U_S)$ with $S = s$. This results in a change in the distribution of all variables that are descendants of S in the causal graph. $\mathcal{M}_s$ is the interventional causal model and its corresponding graph $\mathcal{G}_s$ the interventional graph. In $\mathcal{G}_s$, edges to S are deleted according to the definition of intervention and S is replaced with constant s. The interventional distribution for Y is denoted by $P(Y|do(S = s))$. Using the do-operator, we can compare the interventional distributions under different interventions to infer the causal effect of S on Y. In this paper, we focus on the following causal causal fairness notions:

Total effect The total effect infers the causal effect of S on Y through all possible causal paths from S to Y. The total effect of the difference of s^- to s^+ on Y is given by $TE(s^+, s^-) = P(Y_{s^+}) - P(Y_{s^-})$, where $P(\cdot)$ here refers to the interventional distribution probability. Total fairness is satisfied if $|TE(s^+, s^-)| < \varepsilon$ (ε is the fairness threshold). Note that statistical parity is similar to total effect but is fundamentally different. Statistical parity measures the conditional distributions of Y change of the sensitive attribute from s^- to s^+.

Path-specific fairness The path-specific effect is a fine-grained assessment of causal effects, that is, it can evaluate the causal effect transmitted along certain paths. It is used to distinguish among direct discrimination, indirect discrimination, and explainable bias. It infers the causal effect of S on Y through a subset of causal paths from S to Y, which is referred to as the π-specific effect denoting the subset of causal paths as π. The specific effect of a path set π on Y, caused by changing the value of S from s^- to s^+ with reference to s^-, is given by the difference of the interventional distributions: $SE_\pi(s^+, s^-) = P(Y_{s^+|\pi, s^-|\overline{\pi}}) - P(Y_{s^-})$, where $P(Y_{s^+|\pi, s^-|\overline{\pi}})$ represents the distribution resulting from intervening $do(s^+)$ only along the paths in π while $s-$ is used as a reference through other paths $\overline{\pi}$. If π contains all direct edge from S to Y, $SE_\pi(s^+, s^-)$ measures the direct discrimination. If π contains all indirect paths from S to Y that pass through proxy attributes, $SE_\pi(s^+, s^-)$ evaluates the indirect discrimination. Path-specific fairness is met if $|SE_\pi(s^+, s^-)| < \varepsilon$.

Counterfactual fairness The counterfactual effect of changing S from s^- to s^+ on Y under certain conditions $O = o$ (where O is a subset of observed attributes $O \subseteq X$) for an individual with features o is given by the difference between the interventional distributions $P(Y_{s^+}|o)$ and $P(Y_{s^-}|o)$: $CE(s^+, s^-|o) = P(Y_{s^+}|o) - P(Y_{s^-}|o)$. Counterfactual fairness is met if $|CE(s^+, s^-|o)| < \varepsilon$. Any context $O = o$ represents a certain sub-group of the population, specifically, when $O = X$, it represents specific individual(s).

3.1.2 Causal Discovery

Methods for extracting a causal graph from given data (causal discovery) can be broadly classified into constraint-based and score-based approaches (Spirtes and Zhang, 2016;

Glymour *et al.*, 2019). Constraint-based methods, such as those by Spirtes *et al.* (2013), Spirtes and Glymour (1991) and Colombo *et al.* (2012), use conditional independence tests under specific assumptions to identify the Markov equivalence class of causal graphs. Score-based methods, exemplified by Vowels *et al.* (2023), assess candidate graphs with a predefined score function and search for the optimal graph within the space of Directed Acyclic Graphs (DAGs). This approach is framed as a combinatorial optimization problem.

$$\begin{aligned} &\min_{\mathcal{G}} Score(\mathcal{G};V) = \mathcal{L}(\mathcal{G};V) + \lambda \mathcal{R}_{sparse}(\mathcal{G}), \\ &s.t.\ \mathcal{G} \in DAG \end{aligned} \tag{3.1}$$

In the realm of causal discovery, the problem can be divided into two components, which constrain the score function $Score(\mathcal{G};V)$ and $\mathcal{G} \in DAG$. The score function is comprised of: (1) the goodness-of-fit $\mathcal{L}(\mathcal{G};V) = \frac{1}{n}\sum_{i=1}^{n} l(v_i, F(v_i))$ is the loss of fitting observation of v_i; F denotes the deterministic functions as defined earlier in Section 3.1 (2) the sparsity $\mathcal{R}_{sparse}(\mathcal{G})$ which regulates the number of edges in $\mathcal{G}$. λ serves as a hyperparameter that controls the regularization strengths.

In this work, we assume that the given causal graph $\mathcal{G}$ is learned from a score-based causal discovery, so $\mathcal{G}$ should have goodness-of-fit and sparsity.

3.1.3 Intervention through Controlled Neural Networks

In CausalGAN (Kocaoglu *et al.*, 2018), a noise vector Z is partitioned into $\{Z_{V_1}, Z_{V_2}, ..., Z_{V_{|V|}}\}$ to mimic the exogenous variables U in the structural causal model $\mathcal{M}$ described in Section 3.1. The generator $G(Z)$ contains $|V|$ sub-neural networks $\{G_{V_1}, G_{V_2}, ..., G_{V_{|V|}}\}$ to generate the values of each node V_i in the graph. The input of G_{V_j} is the output of $G_{Pa_{V_j}}$ combined with Z_{V_j}. Here, G_{V_j} is trying to approximate the corresponding function $f_{V_j}(Pa_{V_j}, U_{V_j})$ in the causal model $\mathcal{M}$. The adversarial game is played to ensure that the generated observational distribution is not differentiable from the real observational distribution. In the work of CFGAN, two generators are used to simulate the causal model $\mathcal{M}$ and the interventional model $\mathcal{M}_s$, while two discriminators try to maintain that: (1) the generated data is close to the orginal distribution, and (2) the causal effect is mitigated. In our work, we also use a similar design but we do not model the noise Z since our goal is not to generate fairness-aware data, but to reweigh the given data.

3.2 A Reweighting Approach for Different Causal Fairness Critiria

3.2.1 Problem Formulation

As mentioned in Section 3.1, the notation used in our work is based on the conventional approach. We are given a causal graph $\mathcal{G}$ and a dataset $\mathcal{D}$ with m i.i.d. samples drawn from $P(V)$. We assume that $\mathcal{G}$ is sufficient to describe the causal relationships between the variables V. In this paper, we build our method on a causal graph of observational data, so we do not specifically model U. The problem we are facing is that from the given causal graph $\mathcal{G}$, S has a causal effect on Y. Our method aims to achieve two objectives: (1) preserve the goodness-of-fit (mentioned in Section 3.1.2) by maintaining the empirical reweighted data distribution close to the original data distribution for utility of the downstream tasks; and (2) ensure that S cannot be used to discriminate when predicting Y based on various causal criteria in the interventional model $\mathcal{M}_s$. We treat S and Y as binary variables in this paper. However, this can be easily extended to multi-categorical or numerical cases. Also, we focus on the causal effect of S on Y, but the model can deal with causal effects among multiple variables. We try to reach the following causal fairness notions mentioned in Section 3.1.1, including total fairness (Zhang and Bareinboim, 2018), path-specific fairness (elimination of indirect discrimination) (Zhang *et al.*, 2017), and counterfactual fairness (Kusner *et al.*, 2017).

3.2.2 Reweighting For Causal Fairness

We propose a reweighting scheme which consists of neural networks (F^1, F^2) and one discriminator (D). Figure 3.1 shows the framework of our method.

As shown in Section 3.2.1, causal fairness notions measures the difference between the interventional distributions. To guarantee these notions, our method adopts two neural networks to approximate the causal relations. One neural network F^1 simulates the causal model $\mathcal{M}$, while the other neural network F^2 approximates the interventional model $\mathcal{M}_s$ according to which kind of causal effect is measured. F^1 aims to force the reweighted data close to the given causal graph, and F^2 aims to drive the interventional distributions to satisfy the specific notion defined in Section 3.2.1. To represent the connections between the two causal models, the two neural networks share certain structures and parameters, while they differ in sub-neural networks to indicate the intervention (the edges to S in the interventional graph is deleted). Then, our method adopts a discriminator D trying to distinguish the two interventional distributions (reweighted) $P(Y_{s^+})$ and $P(Y_{s^-})$. Finally, the discriminator and reweighting play an adversarial game to produce weights for individuals in the dataset.

To better illustrate our design, we divide X into $\{A,B\}$ and $V = \{S,A,B,Y\}$ based on the positions of the nodes in the causal graph – variables in A are direct causes of S and

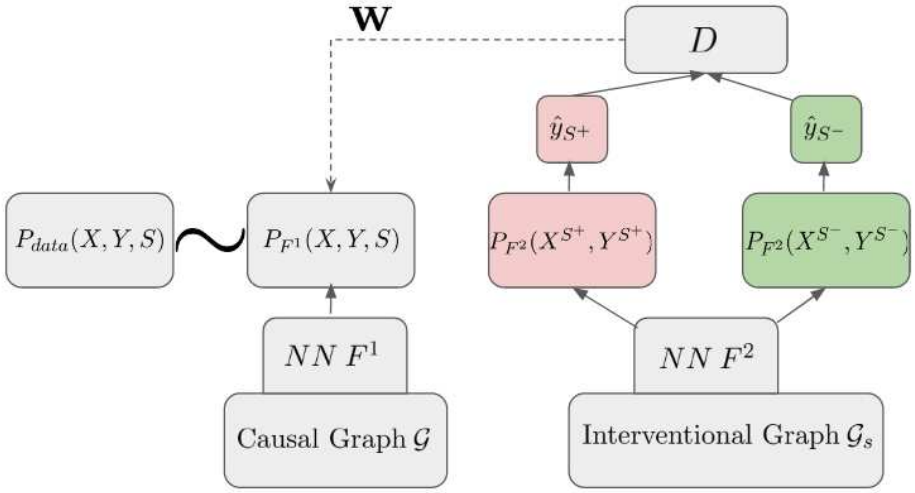


Figure 3.1: The framework of reweighting: the structure of NN (neural network) F^1 reflects the original causal graph $\mathcal{G}$; the structure of NN F^2 refelects the interventional causal graph $\mathcal{G}_s$; the discriminator D tells if a $\hat{y}$ estimated by F^2 is from the group S^+ or the group S^-. An adverserial game is played between the reweighting on the data samples and D to reach a situation where D is not capable of differentiating whether y is from S^+ or S^- and a specific causal fairness is reached. The weights of samples are also forwarded to F^1 to make sure that the reweighted empirical data distribution is close to the original data distribution from which the causal graph $\mathcal{G}$ is learned.

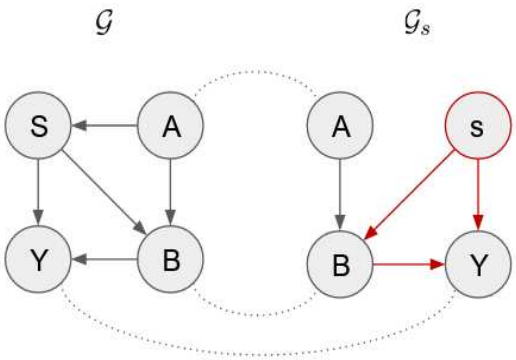
variables in B are descendants of S and A.

Reweighting for Total Fairness

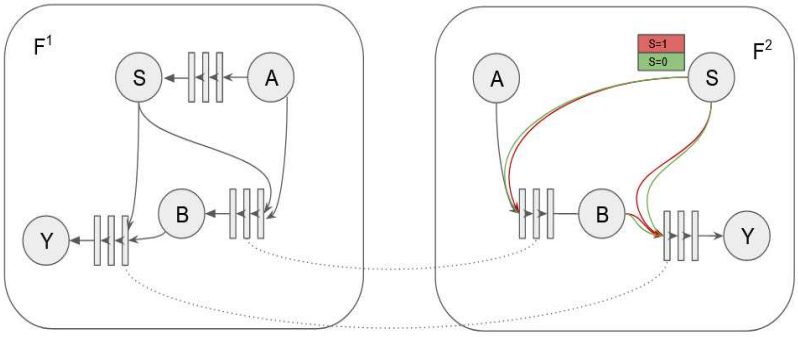
The causal graph $\mathcal{G}$ is shown in Figure 3.2a. We also show the interventional graph $\mathcal{G}_s$ with the intervention $do(S = s)$ and the edge from A to S is deleted in $\mathcal{G}_s$, which is also altered in F^2. The pair of nodes connected by dashed lines indicate that they share the same function (structures and parameters of the corresponding sub-neural networks) as shown in Figure 3.2b. For parallel nodes in the two graphs, the corresponding sub-neural networks are synchronized during the training process.

We first show our method to achieve total fairness by describing each components of our design. As mentioned in Section 3.1.1 , $|TE(s^+,s^-)| < \varepsilon$ must hold for all possible paths from S to Y shown in Figure 3.2a.

Neural Networks F^1 and F^2 The feed-forward Neural Network F^1 is constructed to correspond with the causal graph $\mathcal{G}$. It consists of $|V| - r$ sub-neural networks (r is the total number of the root nodes in $\mathcal{G}$), with each corresponding to a node in V (expect



(a) Causal Graphs $\mathcal{G}$ and $\mathcal{G}_s$



(b) Neural Networks F^1 and F^2

Figure 3.2: The Neural Networks F^1 and F^2 on total effect. S is 1 or 0 for the interventional joint distributions $P_{F^2}(s^+)$ (red path) and $P_{F^2}(s^-)$ (green path), respectively. The pair of nodes connected by dashed lines indicate that they share the same function (structures and parameters of the corresponding sub-neural networks).

for the root nodes). Similar to what is described as the design of CFGAN in Section 3.1.3, each sub-neural network $F^1_{V_j}$ is trying to approximate the corresponding function $f_{V_j}(Pa_{V_j})$ in the causal model $\mathcal{M}$ of the given causal graph $\mathcal{G}$. When F^1 is properly trained, the causal model $\mathcal{M}$ is learned. Then, $F^1_{V_j}$ outputs the estimated values of V_j, i.e., $\hat{v_j}$. The other neural network F^2 is constructed to align with the interventional graph $\mathcal{G}_s$, where all the incoming edges to S are removed under the intervention $do(S = s)$. The layout of F^2 is analogous to F^1, but with the exception that the sub-neural network F^2_S

is designated as $F_S^2 \equiv 1$ if $s = s^+$, and $F_S^2 \equiv 0$ if $s = s^-$. To synchronize the two neural networks F_1 and F_2, they share the identical set of structures and parameters for every corresponding pair of sub-neural networks, i.e., $F_{V_j}^1$ and $F_{V_j}^2$ for each V_j except for S. When F_2 is properly trained, the interventional model $\mathcal{M}_s$ is learned. With $\mathcal{M}$ and $\mathcal{M}_s$ learned, we could manipulate the interventional distributions to reach our goal of causal fairness.

Discriminator D is used to differentiate between the two interventional distributions $\hat{y}_{s^+} \sim P_{F^2}(Y_{s^+})$ and $\hat{y}_{s^-} \sim P_{F^2}(Y_{s^-})$. The aim of the discriminator D to minimize the bias by penalizing differences between both groups.

Weights Assuming the to-reach-causal-fairness-importance of each individual in the given dataset is known, we can assign importance to different individuals in $\mathcal{M}_s$ to improve causal fairness for any downstream task. $\mathbf{w} = (w_1, ..., w_n)$ is a sample reweighting vector with length n, where w_i indicates the importance of the i-th observed sample (s_i, x_i, y_i). We want to reach a balance of goodness-of-fit to the known causal graph $\mathcal{G}$ which is learned from $\mathcal{D}$ and reweighting for causal fairness.

Recall that here we assume that the known causal graph $\mathcal{G}$ is learned from a causal discovery which means it achieves goodness-of-fit. We do not want the reweighted data to drift too far from the original causal graph. We use hatted variables to represent the output of the neural networks of the graphs. To reach this objective, we have:

$$S_{F^1}(\mathcal{G}) = \min_{F^1} \sum_{i=1}^{n} w_i l((s_i, x_i, y_i), (s_i, \hat{x}_i, \hat{y}_i)) \tag{3.2}$$

where $l((s_i, x_i, y_i), (s_i, \hat{x}_i, \hat{y}_i))$ represents the loss of fitting observation (x_i, y_i, s_i). In the experiment, we use weighted MSE loss for the continuous variables and weighted cross entropy loss for the categorical variables. The problem then becomes how to learn appropriate the sample reweighting vector $\mathbf{w}$ for the objective of causal fairness. We formulate our objective as a minmax problem to reweight with $\mathcal{M}_s$:

$$\min_{\mathbf{w}} \max_{D} \sum_{i=1}^{n} w_i (D(\hat{y}_i^{s^+}) - D(\hat{y}_i^{s^-})), \tag{3.3}$$

To avoid information loss by assigning close to zero weights to some samples from the group of S^+, we introduce a regularization constraint to the minimization term:

$$\sum_{i=1}^{m} (w_i - 1)^2 \leqslant Tn \tag{3.4}$$

Thus, by adjusting the value of T, we can balance between similarity and dissimilarity of the weights of samples.

Samples easily fitted with fairness constraint should contribute more to $\mathcal{G}_s$: these are the samples with less difference of discriminator outputs from $do(S = s^-)$ to $do(S = s^+)$. We therefore use downweighting on the not-hold-fairness samples, and upweighting on the hold-fairness samples. This could be achieved by assigning weights to samples based on the discriminator D outputs. When the neural networks are properly trained, the discriminator should not be able to tell if the sample is from the group of S^+ or S^- which could achieve total fairness as we describe in Section 3.2.1.

Reweighting for Path-Specific Fairness

The notions of direct and indirect discrimination are connected to effects specific to certain paths. We concentrate on indirect discrimination, even though fulfilling the criterion for direct discrimination is comparable. As mentioned in Section 3.1.1 , $|SE_\pi(s^+, s^-)| = |P(Y_{s^+|\pi_C, s^-|\overline{\pi}_C}) - P(Y_{s^-})| < \varepsilon$ must hold for a path set π_C that includes paths passing through certain attributes, shown in Figure B.4a (in Appendix). F^1 for indirect discrimination is similar to that in Section 3.2.2. However, the design of F^2 is altered because it needs to adapt to the situation where the intervention is transferred only through π_C, shown in Figure B.4b (in Appendix). We examine two possible states for the sub-neural network F^2_S: the reference state and the interventional state. Under the reference state, F^2_S is constantly set to 0. On the other hand, under the interventional state, F^2_S is set to 1 if $s = s^+$, and 0 if $s = s^-$. For other sub-neural networks, there are also two possible values: the reference state and the interventional state, according to the state of F^2_S. If a sub-neural network corresponds to a node that is not present on any path in π_C, it only accepts reference states as input and generates reference states as output. However, for any other sub-neural network $F^2_{V_j}$ that exists on at least one path in π_C, it may accept both reference and interventional states as input and generate both types of states as output.

Reweighting for Counterfactual Fairness

In the context of counterfactual fairness, interventions are made based on a subset of variables $O = o$. Both F^1 and F^2 have similar structures to those in Section 3.2.2. However, we only use samples in F^2 as interventional samples if they satisfy the condition $O = o$. This means that the interventional distribution from F^2 is conditioned on $O = o$ as $P_{F^2}(X_s, Y_s|o)$. The discriminator D is designed to distinguish between $\hat{y}_{s^+}|o \sim P_{F^2}(Y_{s^+}|o)$ and $\hat{y}_{s^-}|o \sim P_{F^2}(Y_{s^-}|o)$, and aims to reach $P_{F^2}(Y_{s^+}|o) = P_{F^2}(Y_{s^-}|o)$. During training, the value of m should be adjusted based on the number of samples that are involved in the intervention.

3.2.3 Training Algorithm

To train the network F^1 to minimize the loss in Equation (3.2), we alternately optimize the network parameters of F^1 and D and learn the weights $\mathbf{w}$ by fixing others as known.

Updating parameters of F^1 with fixed w Fixing **w**, we update F^1 to minimize the loss in Equation (3.2) for M steps, using the mini-batch stochastic gradient descent algorithm.

Updating w with fixed F^2 (synchronized with F^1) Fixing parameters of F^2, we control the training data into two groups (S^+ and S^-) for intervention, and learn **w** in Equation (3.3). Since Equation (3.3) is a min-max optimization problem, we can alternately optimize the weights **w** and the parameters of D of the discriminator by fixing the other one as known. Therefore, we first fix $w_i = 1$ for all i and optimize D to maximize the objective function in Equation (3.3) using the gradient penalty technique, as in WGAN with Gradient Penalty (Gulrajani *et al.*, 2017). Note that when $w_i = 1$ for all i, Equation (3.2) is equivalent to the situation when there is no reweighting applied. Then, fixing the discriminator D, we optimize **w**.

We denote $d_i = D(\hat{y}_i^{S^+}) - D(\hat{y}_i^{S^-})$ and $\mathbf{d} = (d_1, d_2, ...d_n)^T$. The optimization problem for **w** becomes a constrained least squares problem:

$$\min_{\mathbf{w}} \mathbf{d}^T \mathbf{w}, s.t. w_i \geqslant 0, \sum_{i=1}^{n} w_i = n, \sum_{i=1}^{n} (w_i - 1)^2 \leqslant Tn \tag{3.5}$$

3.3 Experimental Evaluation

We conduct experiments on two benchmark datasets (ADULT (Kohavi, 1996) and COMPAS (Mattu *et al.*, 2016)) to evaluate our reweighting approach and compare it with state-of-the-art methods: FairGAN (Xu *et al.*, 2018), CFGAN (Xu *et al.*, 2019), and Causal Inference for Social Discrimination Reasoning (CISD) (Qureshi *et al.*, 2020) for total effect and indirect discrimination (see Appendix (B.1) for more details about the datasets). CISD (Qureshi *et al.*, 2020) introduces a technique for identifying causal discrimination using propensity score analysis. It mitigates the influence of confounding variables by reweighting samples based on propensity scores calculated from logistic regression. However, this approach is purely statistical and does not utilize causal knowledge. We also compare our method with CFGAN and two methods from Kusner *et al.* (2017) (referred to as CE_1 and CE_3 in our paper) for counterfactual effect. CE_1 uses only non-descendants of S for classification, while CE_3 is similar but assumes an additive U.

We choose these methods because FairGAN addresses statistical parity, and CFGAN addresses causal fairness using adversarial methods to mitigate bias, similar to our design. CISD approaches causal fairness with a weighting scheme. We then compare the performance of our method with the aforementioned methods on total effect, indirect discrimination, and counterfactual fairness using four different downstream classifiers: decision tree (DT) (Wu *et al.*, 2008), logistic regression (LR) (Cox, 1958), support vector machine (SVM) (Cortes and Vapnik, 1995), and random forest (RF) (Ho, 1998). We assess the accuracy of the downstream tasks to determine if the data maintains good utility, with higher accuracy indicating better utility. Additionally, for the utility of the

downstream task, we compute the Wasserstein distance between the manipulated data and the original data, where a smaller Wasserstein distance indicates closer distributions and better utility for the downstream tasks.

3.3.1 The datasets and setup

Please refer to Appendix for the details of datasets and training.

3.3.2 Analysis

Total Effect

In Table 3.1, we present the total effect (TE) calculated for the original dataset and the datasets processed using various methods. The original ADULT dataset has a total effect of 0.1854 and COMPAS 0.2389, while applying FairGAN to achieve demographic parity yields almost no total effect. As mentioned in Section 3.1.1, total effect is very similar to demographic parity. However, FairGAN is limited by its focus on statistical fairness, rather than causal fairness, and does not perform well on Wasserstein distance or downstream tasks. It is quite intuitive that if total fairness is met, total fairness should be achieved too on the condition that the causal graph is sufficient. We test it on our two datasets and the result is acceptable. CFGAN produces no total effect, but it performs worse than our method on Wasserstein distance, possibly because reweighted data could manage to stay closer to the original data distribution. Our method also outperforms CISD, which may be due to the use of a neural network instead of logistic regression to calculate weights, allowing for greater flexibility in capturing the dataset.

A Closer Look at the Weights After ranking the weights of samples in the Adult dataset, we observed that older individuals from Europe or Asia (e.g., Germany and India) tend to have the highest weights, while younger black individuals from Caribbean countries (e.g., Jamaica and Haiti) tend to have lower weights. This suggests that when sex is intervened from female to male, the former group is less influenced by the change, while the latter group is more influenced in terms of income. White, middle-aged individuals born in the US are assigned medium weights. To visualize it, we build a decision tree to classify top 10% individuals with highest weights and bottom 10% individuals with lowest weights using the three root nodes $\{race, native\_country, age\}$, shown in Figure 3.3.

Indirect Discrimination

To address indirect discrimination (SE), we identify all possible paths except the direct one $\{S \to Y\}$ as the path π_C and evaluated the results in Table 3.1. Similar to total effect, FairGAN removes indirect discrimination but at the cost of significant utility loss. In contrast, both CISD and our method can effectively remove indirect discrimination

Table 3.1: The total effect (TE) and indirect discrimination (SE) on Adult and COMPAS datasets

ADULT								COMPAS							
	total effect	indirect discrimination	Wasserstein distance	classifier accuracy (%)					total effect	indirect discrimination	Wasserstein distance	classifier accuracy (%)			
				SVM	DT	LR	RF					SVM	DT	LR	RF
original data	0.1854 (0.0301)	0.1773 (0.0489)	0	81.78 (1.45)	81.77 (1.75)	81.70 (1.63)	81.78 (1.76)	orginal data	0.2389 (0.0245)	0.2137 (0.0985)	0	65.24 (2.34)	65.15 (1.46)	65.10 (2.19)	65.27 (1.09)
Ours (TE)	**0.0017 (0.0009)**	**0.0012 (0.0007)**	**0.71 (0.19)**	81.12 (1.72)	81.20 (1.86)	81.60 (2.03)	81.14 (1.05)	Ours (TE)	**0.0037 (0.0018)**	0.0017 (0.0009)	**1.21 (0.32)**	65.09 (2.75)	65.13 (1.76)	65.06 (2.08)	65.11 (1.02)
Ours (SE)			**0.69 (0.23)**	81.14 (1.58)	80.97 (2.01)	81.65 (1.96)	81.17 (1.92)	Ours (SE)			**0.72 (0.35)**	65.11 (1.98)	65.14 (2.06)	65.02 (1.12)	65.09 (1.95)
FairGAN	0.0021 (0.0007)	0.0148 (0.0075)	5.21 (0.78)	79.88 (1.47)	79.81 (1.89)	80.36 (1.32)	80.82 (1.65)	FairGAN	0.0075 (0.0056)	0.0341 (0.0075)	3.24 (1.45)	64.24 (1.77)	64.15 (2.01)	64.50 (2.75)	64.26 (2.34)
CFGAN (TE)	0.0106 (0.0008)	0.0034 (0.0012)	1.78 (0.65)	80.34 (2.56)	80.15 (1.52)	80.07 (1.65)	80.39 (1.32)	CFGAN (TE)	0.0364 (0.0175)	**0.0016 (0.0025)**	2.76 (1.65)	64.59 (2.65)	65.13 (2.73)	65.02 (2.03)	65.01 (2.45)
CFGAN (SE)			1.89 (0.29)	80.37 (1.56)	80.49 (2.05)	80.04 (1.67)	80.24 (1.09)	CFGAN (SE)			2.64 (0.91)	64.21 (2.45)	64.25 (1.75)	64.80 (1.97)	64.87 (1.54)
CISD (TE)	0.0206 (0.0074)	0.0098 (0.0045)	2.57 (0.18)	80.73 (1.42)	80.74 (1.75)	81.15 (1.82)	81.27 (1.47)	CISD (TE)	0.0356 (0.0246)	0.0175 (0.0231)	2.57 (1.61)	65.04 (1.76)	65.17 (1.54)	65.04 (2.47)	65.05 (1.75)
CISD (SE)			2.82 (0.23)	80.75 (1.28)	80.72 (1.58)	80.77 (1.96)	81.32 (1.95)	CISD (SE)			2.65 (1.56)	64.01 (1.56)	65.02 (1.49)	64.09 (2.45)	64.11 (1.32)

Table 3.2: The counterfactual effect (CE) on Adult and COMPAS datasets

ADULT								COMPAS							
	counterfactual effect		Wasserstein distance	classifier accuracy (%)					counterfactual effect		Wasserstein distance	classifier accuracy (%)			
	o_1	o_2		SVM	DT	LR	RF		o_1	o_2		SVM	DT	LR	RF
original data	0.1254 (0.0107)	0.1265 (0.0489)	0	81.78	81.77	81.70	81.78	original data	0.2079 (0.0303)	0.1973 (0.0484)	0	65.24 (2.35)	65.15 (1.32)	65.10 (1.97)	65.27 (1.35)
Ours	**0.0017 (0.0009)**	0.0030 (0.0007)	**0.98 (0.10)**	81.14 (1.32)	81.25 (1.37)	81.67 (1.36)	81.29 (1.42)	Ours	**0.0027 (0.0009)**	0.0037 (0.0109)	**1.43 (1.68)**	65.11 (2.01)	65.14 (1.35)	65.07 (1.89)	65.13 (1.68)
CE_1	0.0021 (0.0157)	0.0148 (0.0025)	4.95 (0.45)	77.88 (1.65)	78.81 (1.32)	76.36 (2.06)	77.82 (1.78)	CE_1	0.0034 (0.0007)	0.0048 (0.0075)	4.53 (1.56)	63.24 (1.76)	63.16 (1.43)	62.19 (1.46)	62.27 (2.21)
CE_3	0.1123 (0.0017)	0.1046 (0.0057)	3.11 (0.65)	80.75 (1.76)	81.59 (2.15)	81.62 (2.01)	81.43 (1.37)	CE_3	0.1021 (0.0981)	0.0145 (0.0025)	1.76 (0.89)	64.07 (1.96)	65.03 (1.95)	64.01 (1.45)	65.10 (2.89)
CFGAN (CE)	0.0021 (0.0039)	**0.0027 (0.0064)**	2.99 (0.78)	81.03 (1.35)	81.14 (2.19)	81.11 (2.17)	81.15 (1.95)	CFGAN(CE)	0.0034 (0.0024)	**0.0031 (0.0045)**	1.74 (1.06)	65.04 (1.72)	65.11 (1.74)	64.08 (1.19)	64.29 (1.32)

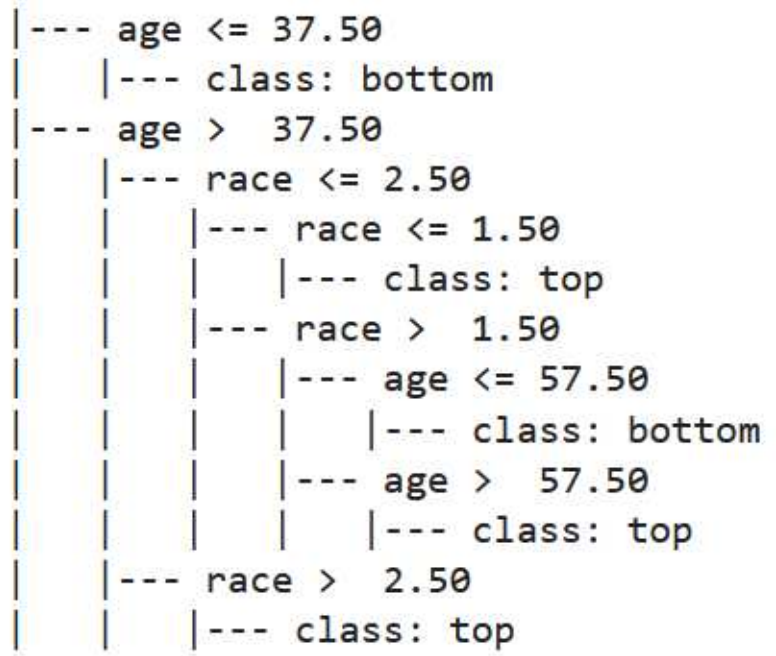


Figure 3.3: The visualize of the decision tree trying to classify individuals with low or high weights. we see that age and race are the most important attributes to build the tree. The mapping of label encoder for *race* is $\{'Amer-Indian-Eskimo': 0, 'Asian-Pac-Islander': 1, 'Black': 2, 'Other': 3, 'White': 4\}$

while maintaining better data utility than FairGAN. Although CFGAN and CISD perform similarly using different techniques, our method outperforms both methods in terms of Wasserstein distance, indicating the best overall utility among these approaches.

Counterfactual Fairness

To evaluate counterfactual effect (CE), we consider the conditions on two variables - race and native country (binarized) for ADULT, and sex and age (binarized) for COMPAS - resulting in four value combinations. Table 3.2 presents the results for two selections (see Appendix (B.2) for more details). We find biases in the original data regarding counterfactual fairness in these two selections. CE_1 is counterfactually fair, but the classifier accuracy is poor because it solely employs non-descendants of the sensitive attributes for outcome attributes. CE_3 cannot achieve counterfactual fairness, probably due to the strong assumptions while introducing U. In contrast, our method performs well on both dimensions due to its flexibility. Although CFGAN performs well in some aspects, our method outperforms it in Wasserstein distance, likely because reweighting better preserves the original distribution than generation methods.

Summary We find out that in general neural nets-based methods outperform due to

the flexibility of neural networks to capture any function, while reweighting outperforms generation. We could see from the experiment results above, imposing strong assumptions on the U and F could cause unwanted problems, and we argue that is why neural nets should be explored more in causal fairness problem settings. Fairness related methods usually formalize the problem as an optimization trade-off between utility and specific fairness objectives. Nevertheless, these discussions are often based on a fixed distribution that does not align with our current situation. We think that an ideal distribution might exist where fairness and utility are in harmony. To include the reweighting scheme into the downstream tasks could be an very interesting future direction to locate this harmonious distribution.

3.4 Conclusion, Limitation and Future Work

We propose a novel approach for achieving causal fairness by dataset reweighting. Our method considers different causal fairness objectives, such as total fairness, path-specific fairness and counterfactual fairness. It consists of two feed-forward neural networks F^1 and F^2 and a discriminator D. The structures of F^1 and F^2 are designed based on the original causal graph $\mathcal{G}$ and interventional graph $\mathcal{G}_s$, and the discriminator D is used to ensure causal fairness combined with a reweighting scheme. Our experiments on two datasets show that the approach improves over state-of-the-art approaches for the considered causal fairness notions achieving minimal loss of utility. Moreover, by analyzing the sample weights assigned by the approach, the user can gain an understanding of the distribution of the biases in the original dataset. Future work involve analyzing the sample weights further, e.g., by using methods from the eXplainable in AI research area. As another relevant research direction, since practitioners often lack sufficient causal graphs when working with a dataset (Binkyte-Sadauskiene *et al.*, 2022), an extension of our work could involve causal discovery as an integral part of the approach.

Chapter 4

Enhancing Fairness through Reweighting: A Path to Attain the Sufficiency Rule

This chapter is based on the conference paper: Zhao, X., K. Broelemann, S. Ruggieri, and G. Kasneci. Enhancing Fairness through Reweighting: A Path to Attain the Sufficiency Rule. *Accepted at ECAI 2024*.

We introduce an innovative approach to enhancing the empirical risk minimization (ERM) process in model training through a refined reweighting scheme of the training data to enhance fairness. This scheme aims to uphold the sufficiency rule in fairness, ensuring that optimal predictors maintain consistency across diverse sub-groups. We employ a bilevel formulation to address this challenge, wherein we explore sample reweighting strategies. Unlike conventional methods that hinge on model size, our formulation bases generalization complexity on the space of sample weights. We discretize the weights to improve training speed. Empirical validation of our method showcases its effectiveness and robustness, revealing a consistent improvement in the balance between prediction performance and fairness metrics across various experiments.

Machine learning has found extensive application in real-world decision-making processes, including areas such as health care systems (Ahmad *et al.*, 2020). However, concerns have arisen over the propensity of learning algorithms to exhibit bias towards certain population groups, as exemplified by predictions of crime likelihood based on ethnicity, gender, or age (Hardt *et al.*, 2016). In response, algorithmic fairness has garnered significant attention as a means to mitigate predictive bias linked to protected features like gender. Consequently, numerous fairness notions catering to diverse objectives have been proposed. While many existing approaches in classification or regression adhere to independence or separation rules (refer to Section 4.1 and related references) (Madras *et al.*, 2018; Song *et al.*, 2019; Chzhen *et al.*, 2020), it's worth noting that these rules may not always be suitable in various applications. In such cases, alternative fairness notions, such as the sufficiency rule (Chouldechova, 2017), are favored. In simple terms, the sufficiency rule, detailed in Section 4.1, ensures that the conditional expectation of $\mathbb{E}[Y|\hat{Y}]$ remains consistent across different sub-groups, providing a more nuanced

approach to fairness.

In practical terms, neglecting the sufficiency rule can result in significant biases within intelligent healthcare systems. For instance, many health systems utilize algorithms to identify and support patients with complex health requirements. These algorithms generate a score indicating the level of healthcare needs, with higher scores suggesting greater sickness and the need for more care. Notably, a study by Obermeyer *et al.* (2019) uncovers a widely used industry algorithm affecting millions of patients, which exhibits pronounced racial bias. It was found that for a given predicted score $\hat{Y} = c$, black patients tend to be considerably sicker than white patients ($\mathbb{E}_{black}[Y|\hat{Y} = c] > \mathbb{E}_{white}[Y|\hat{Y} = c]$). Moreover, the study highlights that rectifying this disparity could significantly increase the percentage of black patients receiving additional care from 17.7% to 46.5%. From an algorithmic perspective, the sufficiency rule is generally incompatible with concepts such as independence or separation, as demonstrated in Section 4.1. This suggests that existing fair algorithms designed for independence or separation may not enhance or could even exacerbate issues related to the sufficiency rule. Notably, recent work on Invariant Risk Minimization (IRM) proposed by Arjovsky *et al.* (2019); Bühlmann (2018) has potential to address this challenge since the criteria of the sufficiency rule and IRM are intrinsically consistent (see more details in Section 4.2.1). IRM seeks to maintain invariant correlations between the embedding (or representation) and the true label by incorporating regularization techniques into Deep Neural Network (DNN) training; the idea being that if these correlations remain robust and unaffected by specific sub-groups, the resulting representation can be considered fair. IRM aims to acquire an invariant representation by disregarding spurious features, while fair methods aim to eliminate the dependency from sensitive features. IRM approaches have attracted attention due to their promising performance on modest models and datasets (Arjovsky *et al.*, 2019) and their simplicity in facilitating end-to-end training. Nevertheless, recent studies have indicated diminished effectiveness of the regularization terms when applied to DNNs (Cherepanova *et al.*, 2021; Lin *et al.*, 2022). For instance, CelebA comprises only 200k training data, whereas ResNet-18 boasts 11.4 million parameters. Overparameterized DNNs can easily diminish the regularization term of IRM to zero during training while still depending on spurious features. In such scenarios, applying IRM methods directly for fairness to uphold the sufficiency rule in relatively larger models is deemed inappropriate.

We introduce a novel approach to address the aforementioned limitation by proposing a model-agnostic sample reweighting method. Our method transforms the parameter search space of the model into one of sample weights by formalizing the learning of sample reweighting as a bilevel optimization problem. Within the **inner loop**, we train DNN on the weighted training samples. In the **outer loop**, we employ the IRM criterion as the outer objective to guide the learning process of the sample weights, thereby enforcing the sufficiency rule. We iteratively alternate between the inner and outer loops, ultimately obtaining a set of weights w with an advantageous characteristic: utilizing only learned sample weights on training samples, we can conduct weighted ERM training to achieve superior fairness.

Our contributions are:

1. We introduce a model-agnostic sample reweighting approach rooted in bilevel optimization for IRM learning to promote fairness. This method offers notable advantages, particularly in transforming the optimization problem from the parameter space of DNNs to the space of sample weights. This shift effectively mitigates the overfitting issues commonly encountered by IRM regularization-based methods.

2. Our method does not impose specific fairness constraints, thus avoiding the issue of determining critical hyperparameters for fairness regularization. We substantiate the superior performance of our approach through empirical evaluations across diverse tasks, showcasing its effectiveness compared to state-of-the-art methods.

4.1 Preliminaries and Related Work

4.1.1 Sufficiency Rule in Fairness

We denote the predictive features as $X \in \mathcal{X}$, the ground truth label as $Y \in \mathcal{Y}$, and the algorithm's output as $\hat{Y} \in \mathcal{Y}$. We consider a binary protected feature or two sub-groups $\mathcal{D}_0$ and $\mathcal{D}_1$. Then, in accordance with Liu *et al.* (2019b), the sufficiency rule is defined as follows:

$$\mathbb{E}_{\mathcal{D}_0}[Y|\hat{Y}=c] = \mathbb{E}_{\mathcal{D}_1}[Y|\hat{Y}=c], \forall c \in \mathcal{Y} \tag{4.1}$$

Sufficiency Gap

Eq. (4.1) indicates that the conditional expectation of the ground truth label Y is consistent across both $\mathcal{D}_0$ and $\mathcal{D}_1$, given the same prediction output c. In Shui *et al.* (2022), the sufficiency gap is proposed as a metric for fairness measurement. In binary classification, the sufficiency gap is naturally defined as follows:

Relation to Other Fairness Notions

We briefly contrast the Sufficiency rule with the commonly employed Independence and Separation rules in binary classification. For comprehensive justifications and comparisons, please consult Appendix C.1.

The *Independence rule* is:

$$\mathbb{E}_{\mathcal{D}_0}[\hat{Y}] = \mathbb{E}_{\mathcal{D}_1}[\hat{Y}] \tag{4.2}$$

In binary classification, the Independence rule is often referred to as demographic parity (DP) (Zemel *et al.*, 2013). Furthermore, it can be argued that if $P_{\mathcal{D}_0}(Y=y) \neq P_{\mathcal{D}_1}(Y=y)$

(indicating distinct label distributions in the sub-groups), it is impossible for both the Sufficiency and Independence rules to hold simultaneously (Castelnovo *et al.*, 2022).

Separation Rule is:

$$\mathbb{E}_{\mathcal{D}_0}[\hat{Y}|Y=c] = \mathbb{E}_{\mathcal{D}_1}[\hat{Y}|Y=c], \forall c \in Y \tag{4.3}$$

In binary classification, the Separation rule is also referred to as Equalized Odds (EO) Hardt *et al.* (2016). Additionally, Barocas *et al.* (2023) have further illustrated that if $P_{\mathcal{D}_0}(Y=y) \neq P_{\mathcal{D}_1}(Y=y)$ and the joint distribution of $(Y,\hat{Y})$ has a positive probability in $\mathcal{D}_0$ and $\mathcal{D}_1$, then it is impossible for both the Sufficiency and Separation rules to coexist Castelnovo *et al.* (2022) (please refer to Appendix C.1 for further details).

4.1.2 Invariant Risk Minimization

IRM operates under the assumption that there are multiple environments $\mathcal{E} := \{e_1, e_2, ..., e_E\}$ within the sample space $\mathcal{X} \times \mathcal{Y}$, each characterized by distinct joint distributions. Furthermore, it assumes that the correlation between the spurious features and labels varies inconsistently across these environments. The predictor $f(\cdot;\theta)$ in IRM is expressed as a composite function of a representation $\phi(\cdot;\Phi)$ and a classifier $h(\cdot;v)$, formulated as $f(\cdot;\theta) = h(\phi(\cdot;\Phi);v)$, where $\theta = \{v,\Phi\}$ represents the trainable parameters. The fundamental idea is that if a predictor $f(\cdot;\theta)$ performs effectively across all environments, it suggests that the correlation between the spurious features and labels is not accurately captured due to its instability (Peters *et al.*, 2015; Arjovsky *et al.*, 2019). Consequently, IRM aims to minimize a specific IRM risk to identify such a robust predictor. Several approaches have been proposed to enhance IRM: Krueger *et al.* (2021); Xie *et al.* (2020) advocate for penalizing the variance of risks across different environments, while Chang *et al.* (2020); Xu and Jaakkola (2021) attempt to estimate the violation of invariance by training neural networks. Moreover, theoretical guarantees for IRM on linear models with adequate training environments are provided by Arjovsky *et al.* (2019); Rosenfeld *et al.* (2021); Chen *et al.* (2022).

Two popular risks are:

$$\mathcal{R}^{\text{IRMv1}}(\mathcal{D},\theta) := \sum_e \mathcal{L}(\mathcal{D}^e,\theta) + \lambda \|\nabla_v \mathcal{L}(\mathcal{D}^e,\theta)\|_2^2 \tag{4.4}$$

$$\mathcal{R}^{\text{REx}}(\mathcal{D},\theta) := \sum_e \mathcal{L}(\mathcal{D}^e,\theta) + \lambda \mathbb{V}_e[\mathcal{L}(\mathcal{D}^e,\theta)] \tag{4.5}$$

where $\mathcal{D} = \bigcup^e \mathcal{D}^e$ denotes the data drawn from all environments, where $\mathcal{D}^e$ represents the data from environment e. The expression $\mathbb{V}_e[\mathcal{L}(\mathcal{D}^e,\theta)]$ signifies the variance of the loss across various environments.

However, it has been observed that IRM exhibits diminished efficacy when applied to overparameterized neural networks (Gulrajani and Lopez-Paz, 2021; Choe *et al.*, 2020).

Lin *et al.* (2022) elucidates that this limitation can largely be attributed to the problem of overfitting. Consequently, utilizing these methods directly for addressing fairness concerns is not straightforward.

4.1.3 Reweighting

Sample reweighting constitutes a classical approach for addressing various tasks such as distribution shifts, imbalanced classification, and fairness concerns. Here, we specifically delve into reweighting methodologies associated with fairness considerations. Fairness with Adaptive Weights (Chai and Wang, 2022) imposes constraints on the sum of weights across sensitive groups to ensure equality, assigning weights to each sample based on its likelihood of misclassification. Adaptive Sensitive Reweighting to Mitigate Bias (Krasanakis *et al.*, 2018) assigns weights to samples based on their alignment with the unobserved true labeling. Li and Liu (2022) intricately models the impact of each training sample on fairness-related metrics and predictive utility. Additionally, Zhao *et al.* (2023b) utilizes Neural Networks to reweigh samples, aiming to achieve causal fairness. To the best of our knowledge, no reweighting method has been specifically applied to achieve the sufficiency rule in fairness. Furthermore, our method stands apart from heuristic reweighting methods, as it does not necessitate complex hyper-parameter selection processes.

4.2 Reweighting to Achieve Sufficiency Rule

4.2.1 Bilevel Formulation of Reweighting

Regarding a dataset $\mathcal{D}$ constituted as a set $\{(x_i, y_i)\}_{i=1}^n$, where each (x_i, y_i) is drawn from $\mathcal{X} \times \mathcal{Y}$, the weighted empirical loss is defined as $\mathcal{L}(\mathcal{D}, \theta; w) := \frac{1}{n}\sum_{i=1}^n w_i l(f(x_i;\theta), y_i)$, with $f(\cdot;\theta)$ representing a neural network parameterized by θ, $l(\cdot,\cdot)$ indicating the loss function (e.g., cross-entropy or least squares loss), and $w_i \in \mathbb{R}^+$ denoting the non-negative weight assigned to each sample.

We formulate the objective of learning sample weights to mitigate reliance on sensitive features as the subsequent bilevel optimization problem:

$$\begin{aligned} &\min_{w \in \mathcal{W}} \mathcal{R}(\mathcal{D}, \theta^*(w)), \\ &\text{s.t. } \theta^*(w) \in \arg\min_{\theta} \mathcal{L}(\mathcal{D}, \theta; w) \end{aligned} \tag{4.6}$$

Here, w denotes a vector of sample weights with a length of n, indicating the importance of each training sample, where each component w_i of w satisfies $w_i \geq 0$. Any IRM Risk $\mathcal{R}(\mathcal{D}, \theta)$ discussed in Section 4.1 can function as the outer objective. In our

subsequent experiments, we employ the risk (4.4), denoted as IRMv1. Within the inner loop, we minimize the weighted ERM loss on the training samples to derive a model $\theta^*(w)$, while within the outer loop, we evaluate the learned model's reliance on sensitive features through IRM Risk and adjust the sample weights accordingly. By iteratively alternating between the inner and outer loops, the sample weights gradually adjust to a state where they can yield satisfactory IRM performance via straightforward ERM training. It's worth noting that instead of environments, we have distinct sensitive groups, such as $\mathcal{D}_0$ and $\mathcal{D}_1$ as depicted in Section 4.1. Although we showcase our approach within the context of binary sensitive groups in this section, it can be readily extended to scenarios involving multi-categorical sensitive groups (refer to the experimental details on the toxic comments dataset and COMPAS dataset in Section 4.3).

Our approach provides the following benefits: 1) by establishing an implicit mapping from the sample weight space to the model parameter space in the outer loop, where the former consistently remains smaller than the latter in deep learning tasks (as detailed in earlier section), we effectively address overfitting issues typically associated with IRM regularization-based methods (the objective of the outer loop); 2) our approach avoids the need to impose specific fairness constraints, thereby circumventing the challenge of determining critical hyperparameters for fairness regularization to achieve a better trade-off between fairness and accuracy.

$$\Delta\text{Suf} = \frac{1}{2} \sum_{y \in \{0,1\}} |P_{\mathcal{D}_0}(Y = y|\hat{Y} = y) - P_{\mathcal{D}_1}(Y = y|\hat{Y} = y)| \tag{4.7}$$

The sufficiency gap $\Delta\text{Suf} \in [0,1]$. A value close to 0 indicates equality between two sub-groups, which have close Positive Predictive Values (PPV) and Negative Predictive Values (NPV). To grasp the significance of this metric, consider a healthcare system that only outputs binary scores: High Risk or Low Risk. As highlighted in Obermeyer *et al.* (2019), if $P_{\mathcal{D}_{\text{black}}}(Y = \text{High Risk}|\hat{Y} = \text{Low Risk}) \gg P_{\mathcal{D}_{\text{white}}}(Y = \text{High Risk}|\hat{Y} = \text{Low Risk})$, then the severity of illness is underestimated more for black patients than for white patients. Therefore, a small value of ΔSuf indicates that racial discrimination is being addressed.

Connection to the Sufficiency Rule:

We elaborate the connection between the outer loop of our bilevel objective and the Sufficiency Rule (Shui *et al.*, 2022).

Proposition 2. *In a classification task where the prediction loss is designated as the logistic regression loss, defined as* $\log(1+\exp(-yh(z)))$ *with* $Y = \{0,1\}$*, minimizing the loss in the outer loop is tantamount to:*

$$\mathbb{E}_{\mathcal{D}_0}[Y|Z=z] = \mathbb{E}_{\mathcal{D}_1}[Y|Z=z], \tag{4.8}$$
$$\mathbb{E}_{\mathcal{D}_0}[Y|\hat{Y}=h^*(z)] = \mathbb{E}_{\mathcal{D}_1}[Y|\hat{Y}=h^*(z)] \tag{4.9}$$

where $h^* = h_1^* = h_0^*$ *and* $z = \phi(x)$.

Proposition 2 illustrates that the objective of the outer loop loss aligns with the sufficiency rule in binary classification. The theoretical backbone of the paper is Proposition 2, from which the method is derived through relaxations. The proof of Proposition 2 is straightforward, since by Definition 3 from the original paper of IRM (Arjovsky *et al.*, 2019) it follows that an optimal predictor performs effectively across all environments. The data representation function ϕ elicits an invariant predictor across environments $\mathcal{E}$ if and only if for all latent z in the intersection of the supports of $\phi(X^e)$. We have $\mathbb{E}[Y^e|\phi(X^e)=z] = \mathbb{E}[Y^{e'}|\phi(X^{e'})=z]$, for all $e, e' \in \mathcal{E}$. The optimal predictor is the same across all environments, so we easily have the $\mathbb{E}[Y^e|\hat{Y}^e] = \mathbb{E}[Y^{e'}|\hat{Y}^{e'}]$ which is the same as the sufficiency rule. To better understand this concept, IRM addresses the challenge of ensuring that *a cow is correctly classified as a cow in a picture, regardless of whether the background is Grass or Desert*. On the other hand, the sufficiency rule aims to ensure that *a patient predicted to be high-risk is truly high-risk, regardless of whether the patient is Black or White*.

4.2.2 Enhance Reweighting by Sparsity

We discretize the optimization method (Zhou *et al.*, 2022) here,

$$\begin{aligned} &\min_{m\in\mathcal{C}} \mathcal{R}(\mathcal{D}, \theta^*(m)), \\ \text{s.t. } &\theta^*(m) \in \arg\min_{\theta} \mathcal{L}(\mathcal{D}, \theta; m) \end{aligned} \tag{4.10}$$

where the mask $m \in \{0,1\}_n$ represents a binary vector, and $m_i = 1$ denotes that sample i is included in the training set, otherwise it is excluded. K is a positive integer that determines the size of the selected set, and $\mathcal{C} = \{m : m_i \in \{0,1\}, \|m\|_0 \leq K\}$ denotes the feasible region of m. Essentially, the inner loop trains the network to converge on the selected set to obtain the model $\theta^*(m)$, while the outer loop assesses the loss of $\theta^*(m)$ on the entire set and optimizes it to guide the learning of m.

The distinction between our discrete bilevel formulation (4.10) and the original bilevel formulation (4.6) lies in the absence of individual weights w_i for each sample in the sparse formulation (4.10). We opt for this sparse formulation for several reasons: 1) empirical results demonstrate satisfactory performance even without these weights; 2) it simplifies the development of an efficient training algorithm; 3) excluding noisy data enhances the robustness of the model.

Given the discrete nature of the mask m, directly solving the bilevel optimization problem (4.10) is intractable due to its NP-hard nature. Hence, we adopt a continualization approach (Zhou *et al.*, 2022) via probabilistic reparameterization to render gradient-based optimization feasible. We treat each mask m_i as an independent binary random variable and transform the problem into the continuous probability space. Specifically, we reparameterize m_i as a Bernoulli random variable with probability s_i for being 1 and $1-s_i$ for being 0, i.e., $m_i \sim \text{Bern}(s_i)$, where $s_i \in [0,1]$. Assuming independence among the variables m_i, the distribution function of m becomes $p(m|s) = \prod_{i=1}^{n}(s_i)^{m_i}(1-s_i)^{(1-m_i)}$. Thus, we control the selected size through the sum of probabilities s_i, since $\mathbb{E}_{m\sim p(m|s)}[\|m\|_0] = \sum_{i=1}^{n} s_i$. Consequently, $\mathcal{C}$ can be relaxed into $\tilde{\mathcal{C}} = \{s_i : 0 \le s_i \le 1, \|s\|_1 \le K\}$. Finally, problem (4.10) naturally relaxes into the following:

$$\begin{aligned} \min_{s\in\tilde{\mathcal{C}}} \Psi(s) &= \mathbb{E}_{p(m|s)}[\mathcal{R}(\mathcal{D}, \theta^*(m))], \\ \text{s.t. } \theta^*(m) &\in \arg\min_{\theta} \mathcal{L}(\mathcal{D}, \theta; m) \end{aligned} \tag{4.11}$$

where $\tilde{\mathcal{C}} = \{s_i : 0 \le s_i \le 1, \|s\|_1 \le K\}$ is the domain.

Current bilevel optimization algorithms (Pedregosa, 2016; Grazzi *et al.*, 2020) typically incur high computational costs owing to the resource-intensive implicit differentiation inherent in their chain-rule-based gradient estimator. Specifically, if employed in our context, they commonly approximate the gradient in the following manner:

$$\nabla_s \Psi(s) \approx \nabla_s \theta^*(m) \nabla_\theta \mathcal{R}(\mathcal{D}, \theta^*(m)) \tag{4.12}$$

Therefore, these methods require computing the implicit differentiation of the inner loop optimum, denoted as $\nabla_s \theta^*(m)$, which is computationally expensive due to the necessity of calculating the inverse of a large Hessian matrix or unrolling backward propagation for numerous steps. Despite the introduction of some efficient bilevel optimization algorithms aimed at mitigating the computational load (such as Lorraine *et al.* (2020), which utilized the Neumann series to approximate the Hessian inverse), the approximation process remains time-consuming. This is also why we opt for the discretization of w.

Several beneficial aspects of our formulation (4.11) include:

1. Our formulation serves as a close relaxation (though not equivalent) of Problem (4.10). This is evident for the following reasons:
   a) It is apparent that $\min_{s\in\tilde{\mathcal{C}}} \Psi(s) \le \min_{m\in\tilde{\mathcal{C}}} \Psi(m)$ since any deterministic binary mask m can be represented as a specific stochastic one by setting s_i to either 0 or 1.
   b) Our constraint $\tilde{\mathcal{C}}$ induces sparsity on s through the l_1-norm and the range $[0,1]$, resulting in most components of the optimal s being either 0 or 1.

Therefore, our eventually learned stochastic weight is nearly deterministic.

2. Due to the sparsity constraint, the size of the selected set in the inner loop, denoted as $|m|$, remains small, which greatly enhances the efficiency of optimizing θ^* (refer to details in Appendix C.3.2).

3. As indicated in Eq. (4.13), our outer objective $\Psi(s)$ is differentiable, enabling the utilization of general gradient-based methods for optimization.

The probabilistic formulation of the bilevel problem allows us to circumvent these costly computations by computing the gradient using forward propagation instead of backward propagation. This can be illustrated by the following equations:

$$\begin{aligned}
\nabla_s \Psi(s) &= \nabla_s \mathbb{E}_{p(m|s)}[\mathcal{R}(\mathcal{D}, \theta^*(m))] \\
&= \nabla_s \int \mathcal{R}(\mathcal{D}, \theta^*(m)) p(m|s) dm \\
&= \int \mathcal{R}(\mathcal{D}, \theta^*(m)) \frac{\nabla_s p(m|s)}{p(m|s)} p(m|s) dm \\
&= \int \mathcal{R}(\mathcal{D}, \theta^*(m)) \nabla_s \ln p(m|s) p(m|s) dm \\
&= \mathbb{E}_{p(m|s)}[\mathcal{R}(\mathcal{D}, \theta^*(m)) \nabla_s \ln p(m|s)]
\end{aligned} \tag{4.13}$$

This indicates that $\mathcal{R}(\mathcal{D}, \theta^(m))\nabla_s \ln p(m|s)$ serves as an unbiased stochastic gradient of $\nabla_s \Psi(s)$. Consequently, with the inner loop optimum $\theta^(m)$ at hand, we can update s (probability) via projected stochastic gradient descent:

$$s \leftarrow \mathcal{P}_{\tilde{\mathcal{C}}}(s - \eta \mathcal{R}(\mathcal{D}, \theta^*(m)) \nabla_s \ln p(m|s)) \tag{4.14}$$

It's evident that this approach does not entail any implicit differentiation, and its component $\mathcal{R}(\mathcal{D}, \theta^*(m))$ can be computed through forward propagation. Additionally, $\ln p(m|s)$ exhibits a straightforward form, and the projection possesses a closed-form solution (Zhou *et al.*, 2022) given the simplicity of the constraint $\tilde{\mathcal{C}}$. Consequently, we can efficiently update s.

Thus, we can tackle our bilevel optimization problem (4.11) by alternately: 1) sampling m, i.e., a selected set, from $p(m|s)$ for the inner loop and training the model on this selected set to obtain $\theta^*(m)$; 2) updating the probability s. The detailed steps are shown in Algorithm 1.

4.3 Experiments

Table 4.1: Accuracy and ΔSuf in Toxic comments (left) and CelebA datasets (right)

Toxic comments	Accuracy(↑)	ΔSuf(↓)	CelebA	Accuracy(↑)	ΔSuf(↓)
ERM(I)	0.768±0.004	0.173±0.008	ERM(I)	0.956±0.005	0.210 ±0.094
NUF(II)	0.762±0.007	0.190±0.008	NUF(II)	0.947±0.007	0.104±0.004
IPA(III)	0.745±0.007	0.091±0.012	IPA(III)	0.938±0.103	0.092±0.161
AR(IV)	0.756±0.006	0.128±0.097	AR(IV)	0.950±0.012	0.197±0.007
Ours(V)	**0.763±0.004**	**0.028±0.004**	Ours(V)	**0.953±0.094**	**0.045±0.004**
IRMv1(VI)	0.753±0.004	0.068±0.008	IRMv1(VI)	0.946±0.009	0.088±0.007

Table 4.2: Accuracy and ΔSuf in Adult (left) and COMPAS datasets (right)

Adult	Accuracy(↑)	ΔSuf(↓)
ERM(I)	0.831±0.014	0.160±0.007
NUF(II)	0.815±0.017	0.068±0.015
IPA(III)	0.810±0.004	0.058±0.024
AR(IV)	0.820±0.023	0.230±0.014
Ours(V)	**0.827±0.016**	0.036±0.007
IRMv1(VI)	0.825±0.018	**0.032±0.012**

COMPAS	Accuracy(↑)	ΔSuf(↓)
ERM(I)	0.652±0.024	0.276±0.094
NUF(II)	0.633±0.032	0.156±0.008
IPA(III)	0.647±0.017	0.097±0.009
AR(IV)	**0.659±0.019**	0.285±0.018
Ours(V)	0.647±0.004	**0.068±0.015**
IRMv1(VI)	0.645±0.008	0.078±0.017

We adopt the aforementioned sufficiency gap as the fair metric and accuracy as the metric for utility. Our neural network models are trained on an Intel(r) Core(TM) i7-8700 CPU. The networks in our experiments are built using the Pytorch package (Paszke *et al.*, 2019).

4.3.1 Baselines

We compare our method with (I) Empirical Risk Minimization (**ERM**) which trains the model without considering fairness; (II) No Utility-Cost Fairness via Data Reweighing (**NUF**) (Li and Liu, 2022); (III) Fair Representation Learning through Implicit Path Alignment (**IPA**) (Shui *et al.*, 2022), an approach in the fair representation learning to achieve also the sufficiency rule; (IV) Adversarial Reweighting Guided by Wasserstein Distance for Bias Mitigation (**AR**) (Zhao *et al.*, 2023a). Notably, the baseline (IV) is grounded in Demographic Parity (DP), illustrating their general incompatibility with addressing the sufficiency rule. Additionally, we include the original Invariant Risk Minimization regularization (Arjovsky *et al.*, 2019), denoted as IRMv1, which incorporates a gradient penalty to encourage invariance across different groups. Even though it is designed for another purpose, as shown earlier in Section 4.2, it has potential to address

Algorithm 1 Reweighting for the Sufficiency Rule

Require: a neural network parameterized by θ, a dataset $\mathcal{D}$, and a selected set size K.

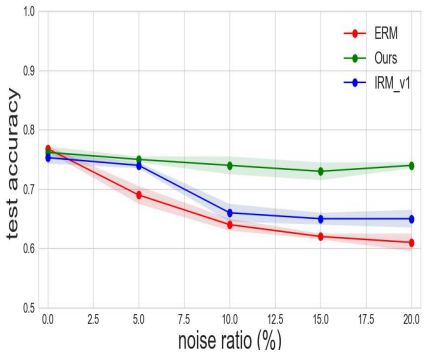
1: Initiate probabilities s^1 as $\frac{K}{|\mathcal{D}|}\mathbf{1}$.
2: **for** iteration t of training, where t is from 1 to T. **do**
3: Sample m based on the probability vector s^t.
4: Continue training the inner loop until convergence achieved:
$\theta^*(m) \leftarrow \underset{\theta}{\text{argmin}}\hat{\mathcal{L}}(\theta; m)$
5: Sample a mini-batch $\mathcal{K}$ from the dataset $\mathcal{D}$:
$\mathcal{K} = \{(x_1, y_1), ..., (x_\mathcal{K}, y_\mathcal{K})\}$
6: Update s according to $\theta^*(m)$ and $\mathcal{K}$.
$s^{t+1} \leftarrow \mathcal{P}_{\tilde{\mathcal{C}}}(s^t - \eta\mathcal{R}_\mathcal{K}(\mathcal{D}, \theta^*(m))\nabla_s \ln p(m|s^t))$
7: **end for**
8: Output: The selected set $\{(x_i, y_i) : m_i = 1 \text{ and } (x_i, y_i) \in \mathcal{D}\}$ where m is sampled from $p(m|s^{T+1})$.

fairness to reach the sufficiency. Results are averaged over five repetitions. Further experimental results are provided in Appendix C.3.2.

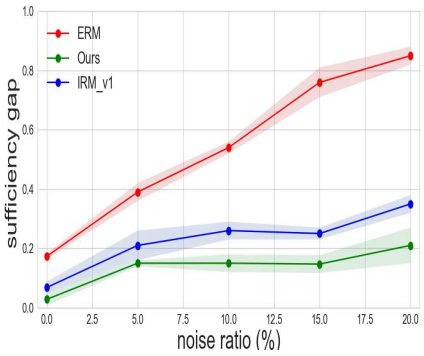
4.3.2 Datasets and Experiment Setups

The **toxic comments dataset** Jigsaw (2018) presents a binary classification challenge in natural language processing (NLP), aiming to determine whether a comment exhibits toxicity. Originally, the labeling process for this dataset is not binary due to involvement from multiple annotators, leading to potential discrepancies. To address this, we adopt a straightforward strategy where a comment is classified as toxic if at least one annotator marks it as such, similar to the approach in Shui *et al.* (2022). Notably, some comments in this dataset are annotated with identity attributes such as gender and race. It has been observed that the race attribute correlates with the toxicity label, posing a risk of predictive discrimination. Therefore, we designate race as the protected feature and specifically focus on two sub-groups: Black and Asian. For computational efficiency, we begin by leveraging a pre-trained BERT model (Devlin *et al.*, 2019) to extract word embeddings, resulting in vectors of 748 dimensions.

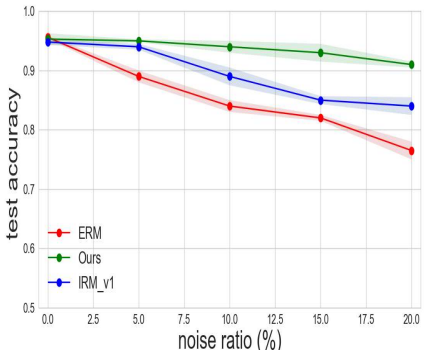
The **CelebA dataset** Liu *et al.* (2015) comprises approximately 200K images featuring celebrity faces, each associated with 40 human-annotated binary attributes such as gender, hair color, and age. For our experiment, we randomly partitioned the dataset, selecting approximately 82K images for training and 18K for validation. We employed the ResNet-18 architecture (He *et al.*, 2016), pre-trained on ImageNet (Deng *et al.*, 2009), omitting the final fully-connected layer to obtain embeddings of 512 dimensions for simplicity. Within the CelebA dataset, our specific task involves predicting hair color ({blond, dark}) based on the image input. Notably, the gender attribute ({male, female})



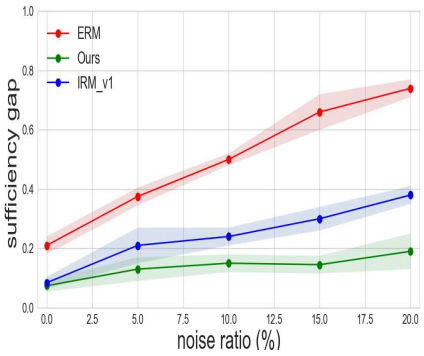
(a) Accuracy vs. Noise ratio on Toxic comments



(b) Sufficiency gap vs. Noise ratio on toxic comments



(c) Accuracy vs. Noise ratio on CelabA



(d) Sufficiency gap vs. Noise ratio on CelabA

Figure 4.1: Change of accuracy and sufficiency gap under different noise ratios on Toxic comments and CelebA datasets, which shows that our method is robust when the data label is noisy.

is correlated with hair color.

For experiments on tabular data, we use the **Adult dataset** (Kohavi, 1996) and the **COMPAS dataset** (Mattu *et al.*, 2016) (For more details of the datasets, please refer to Appendix C.3.1). Adult dataset used personal information such as education level and working hours per week to predict whether an individual earns more or less than $50,000 per year. We use gender as the sensitive feature in Adult dataset. COMPAS dataset is a popular commercial algorithm used by judges and parole officers for scoring criminal defendant's likelihood of reoffending. We use race as the sensitive feature. Here we report the sufficiency gap between two sub-groups of African American and Caucasian even though that ethnicity group is a multi-categorical feature. Please refer to Appendix for more data and training details.

4.3.3 Analysis

Performance Comparison

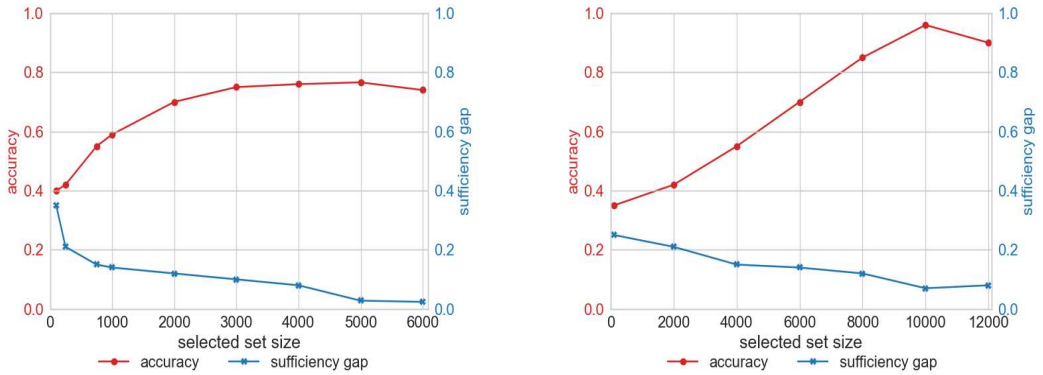
In Table 4.1 and Table 4.2, we present the accuracy and sufficiency gap metrics. Notably, we observe that the Demographic Parity (DP) based fair method (IV) is incompatible with the sufficiency rule, as evidenced by its tendency to increase ΔSuf even surpassing that of ERM. On the other hand, baselines (III, VI), which aim to track the sufficiency rule, exhibit improved sufficiency gap ΔSuf with comparable accuracy, albeit inferior to our approach in Table 4.1. This discrepancy may stem from an overparameterization issue, as previously discussed. Our method consistently demonstrates a superior Accuracy-Fairness trade-off, significantly enhancing sufficiency without substantial accuracy loss. We observe a similar performance pattern on tabular datasets (Table 4.2). However, the performance drop of baselines (III, VI) is less pronounced. This discrepancy may be attributed to the comparatively smaller DNN models utilized in training on tabular datasets, which are less susceptible to overparameterization compared to the toxic comments dataset and CelebA dataset.

Robustness with Noisy Data

We extend our experimentation to scenarios where the dataset incorporates corrupted labels, aiming to demonstrate the robustness of our approach. Following the model configuration outlined in Section 4.3.2, we introduce symmetric noise (Song *et al.*, 2023) into the dataset. Notably, as illustrated in Figure 2.4, our method exhibits robustness towards variations in the dataset's label quality, as evidenced by consistent performance in both accuracy and sufficiency gap metrics. This robustness can be attributed to the comprehensive information assimilated through iterative sampling, leading to the construction of the final weight vector w. Essentially, the sparsity induced by our method facilitates the elimination of noisy data samples, thus preserving the model's effectiveness.

Sensitivity to Choices of K

The selected sizes for the Toxic comments and CelebA experiments are 5000 and 10000, respectively, as shown in Figure 4.2. As the selected sizes increase, we observe an improvement in both accuracy and sufficiency gap performance. Yet, beyond a certain threshold, this improvement plateaus, aligning with the corset concept (Mirzasoleiman *et al.*, 2020). Corset theory suggests that there exists a small subset capable of summarizing the larger dataset effectively. Training exclusively on this condensed set yields competitive performance compared to training on the entire dataset.



(a) size choice for Toxic comments dataset

(b) size choice for CelebA dataset

Figure 4.2: Choices of selected set size K. The size is set to 5000 for Toxic comments and 10000 for CelebA.

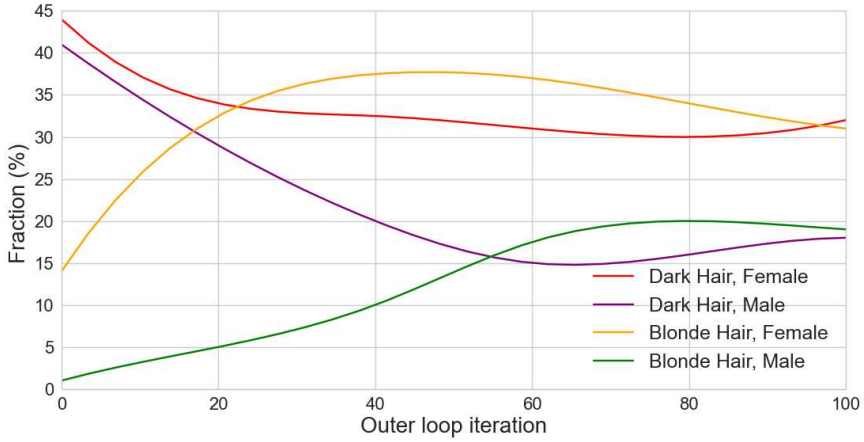


Figure 4.3: The fluctuation in the distribution of group weight fractions for ResNet-18 on the CelebA dataset is notable. Specifically, there's a shift to 20% for both the (Male, Blond Hair) and (Male, Dark Hair) gourps. Similarly, the (Female, Dark Hair) and (Female, Blond Hair) groups see their fractions adjusted to 30%. These observations suggest that our methodology is capable of autonomously adapting weight fractions across various (sub)-groups.

Convergence of Probabilities during Search

A simplified approach is taken by selecting 1000 samples from a larger pool of 10000 training data instances (CelebA). Figure 4.4 illustrates the evolution of probability distributions throughout the search process. Initially, all sample probabilities are uniformly

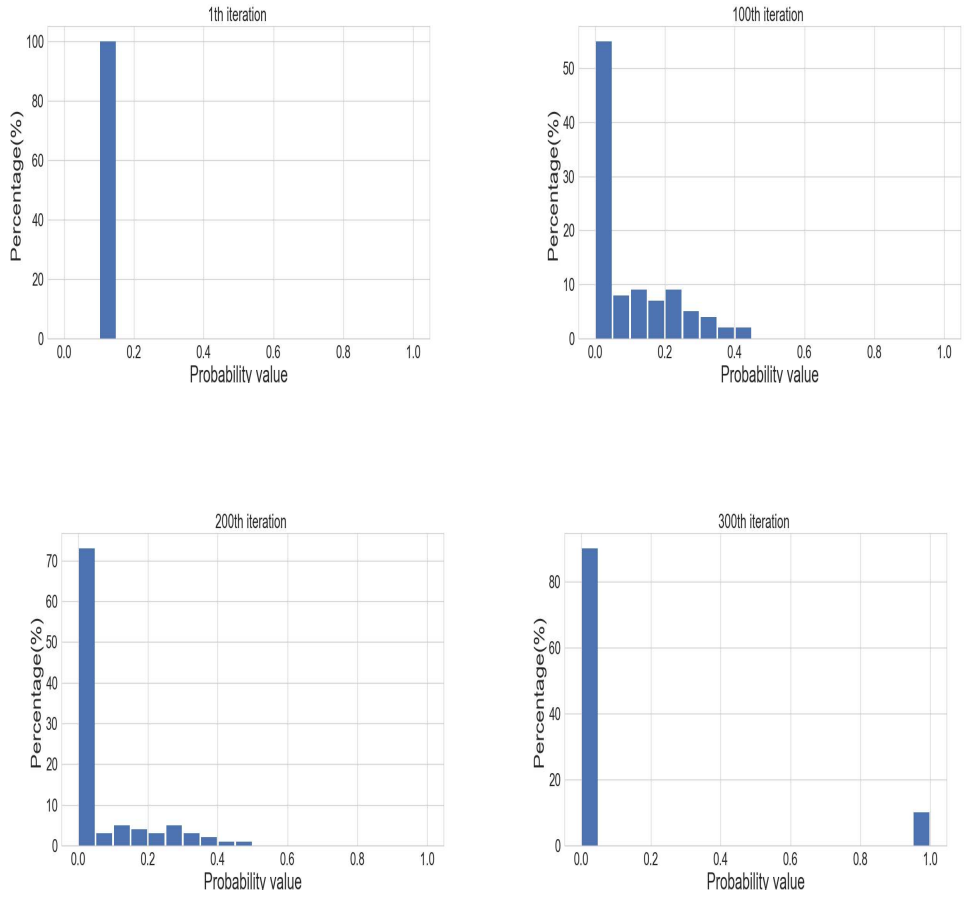


Figure 4.4: The evolution of probability score distribution during the search process reveals a trend where most probabilities tend to converge towards either 0 or 1. This convergence ultimately leads to deterministic weights and convergence of the algorithm.

distributed at 0.1. Over the course of the search, most of these probabilities tend to converge towards either 0 or 1, indicative of diminishing uncertainty. Consequently, a sparse mask with minimal variance is formed, reflecting a nearly deterministic pattern in weight assignment. This trend ultimately leads to the establishment of deterministic weights, signifying algorithmic convergence.

Gradual Change of Group Weights

In Figure 4.3, we depict the training dynamics of sample weight fractions from our CelebA experiment. Initially, all sample weights are uniformly set to 1. The weight fraction of the (Male, Blond Hair) group begins at a mere 0.085%. Following 100 iterations of updates, this fraction gradually increases to approximately 20%. Simultaneously, the

weight fraction of the (Male, Dark Hair) group decreases to approximately 20%, while both the (Female, Dark Hair) and (Female, Blond Hair) groups stabilize at approximately 30%. Interestingly, in Chai and Wang (2022), the assumption is that bias is introduced due to under-representation of the minority groups, hence, they upweight/downweight sensitive groups to the same importance level. Figure 4.3 demonstrates that even though we do not constrain on the group level importance, somehow, our method possesses the ability to dynamically adjust the weight fraction of (sub)-groups automatically.

4.4 Discussion and Conclusion

We presented a model agnostic sample reweighing method to achieve the sufficiency rule of fairness. We formulated this problem as a bilevel optimization to learn sample weights. We further enhance our method with sparsity constraints to improve training speed. Then, we analyzed the sufficiency gap and prediction accuracy of the reweighting algorithm, demonstrating its superior performance over state-of-the-art approaches. The empirical results also show that our method is robust towards noisy labels. One drawback of our framework is that the general performance with respect to IRMv1 consistently improves when the difference between training environments – and specifically sensitive groups – becomes larger. Although this behavior hinders comparability across datasets, it is also in line with our intuition that in fairness-critical scenarios, the performance with respect to the sufficiency rule should quickly improve when there is a bigger gap between sensitive groups.

Appendix A

Adversarial Reweighting Guided by Wasserstein Distance for Bias Mitigation

A.1 Dataset Details

CelebA

CelebA contains 202,600 face images, each endowed with 40 attributes. When we try to construct the datasets from CelebA based on our needs, we maintain the class imbalance in the three datasets constant, which is 70% not wearing hats and 30% wearing hats, since class imbalance is not our priority in this paper. We randomly select samples from CelebA dataset to form our own chosen distribution of male and female groups. We form two datasets based on CelebA which contain 70% male images vs. 30% female images and 80% male images vs. 20% female images and show the experiment results in Table 4.

Adult dataset

The Adult dataset was drawn from the 1994 United States Census Bureau data. It used personal information such as education level and working hours per week to predict whether an individual earns more or less than $50,000 per year. The dataset is imbalanced – the instances made less than $50,000 constitute 25% of the dataset, and the instances made more than $50,000 include 75% of the dataset. As for gender, it is also imbalanced. We use age, years of education, capital gain, capital loss, hours-per-week, etc., as continuous features, and education level, gender, etc., as categorical features. The sensitive attribute is gender. We normalize the continuous features and use one-hot encoding to deal with the categorical features. We train the model for 50 epochs with batch-sizes of 1000 and 500 for the male and female samples.

UCI German Credit Risk dataset

This dataset contains 1000 entries with 20 categorial/symbolic attributes. In this dataset, each entry represents a person who takes credit from a bank. Each person is classified as having good or bad credit risks according to their attributes. The sensitive attribute is sex.

A.2 Training Details and Results

Details of WGAN-GP adaptation for our method

In the original design of WGAN-GP of the training for one batch, the sizes of generated and original samples are equal for the Gradient Penalty as a regularizer to be applied. Here we need to make some changes: we control the sum of the majority group by the weights in one batch by the Wasserstein distance to let it be equal to the sample size of the minority group in one batch. Then we send them for further computation of the regularizer.

Repetition

We repeat experiments on each dataset five times. Before each repetition, we randomly split data into training data and test data for the computation of the standard errors of the metrics.

CelebA training

For the CelebA dataset, since the original data is highly dimensional image data, we use ResNet18 and remove the last layer as a feature extractor. The dimension of the latent space is 512. Note that we use relatively large batch sizes during the training, and we control the sizes of the majority and minority constant during each batch. Papers mention that the large batch size could cause the potential failure of the approximation using neural networks to evaluate the distributions. Our training dataset has 10000 samples, and the test dataset has 2000 samples for all three datasets. From Figure A.3, we can see that our method has its limitation regarding Disparate FPR and Disparate FNR.

Convergence of the training loss

We try to show the stability of training of our method. Figure A.4 shows our method's and baseline's convergence for the CelebA dataset with 90% male and 10% female.

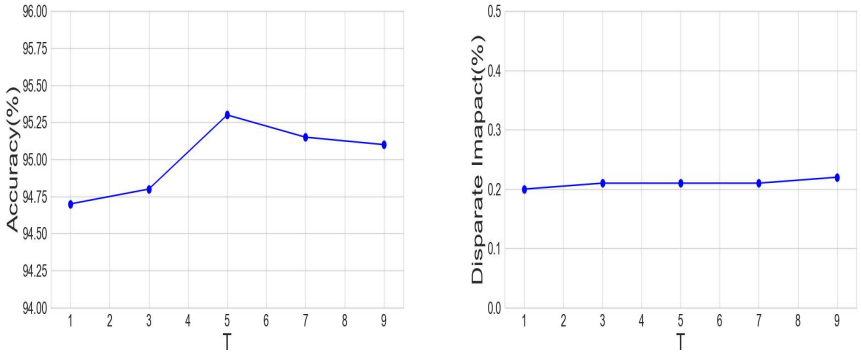
Table A.1: Experiment Results on CelebA

(a) Experimental results of classifier (Wearing Hat) on dataset (30% female and 70% male)

methods	simple methods				state-of-the-art methods				ours
	baseline	reweighing	undersampling	oversampling	ASR	WFC	FAD	FAW	AR
Accuracy rate (%)	95.1 (0.7)	94.9 (0.6)	**95.3 (0.4)**	94.7 (0.9)	93.5 (0.6)	93.6 (0.5)	**95.0 (0.7)**	93.7 (0.6)	94.7 (0.7)
Disparate Impact (%)	6.0 (0.8)	6.0 (0.4)	5.7 (0.2)	4.9 (0.7)	5.1 (0.7)	5.3 (2.4)	5.5 (2.4)	**4.7 (0.4)**	**0.8 (0.5)**
Disparate FPR (%)	-29.1 (1.4)	-31.3 (9.2)	-31.8 (8.1)	-24.7 (7.5)	-26.2 (8.1)	**4.5 (1.1)**	-25.7 (5.8)	-13.7 (2.1)	**-17.0 (5.1)**
Disparate FNR (%)	7.3 (2.9)	8.1 (4.2)	7.1 (3.6)	7.9 (1.9)	8.2 (1.8)	7.6 (3.9)	**6.9 (1.2)**	10.0 (1.1)	**6.5 (1.9)**

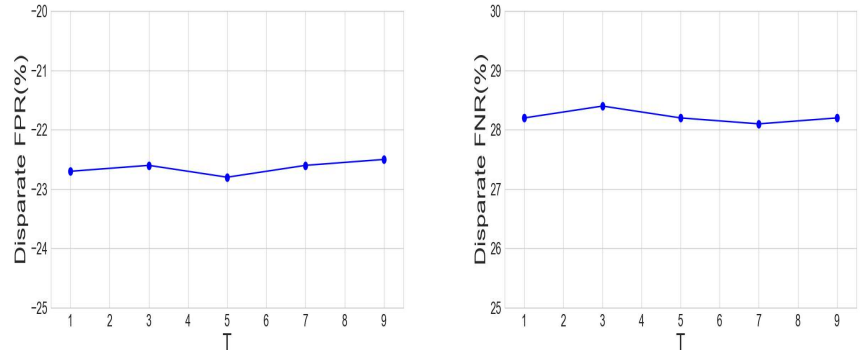
(b) Experimental results of classifier (Wearing Hat) on the dataset (20% female and 80% male)

methods	simple methods				state-of-the-art methods				ours
	baseline	reweighing	undersampling	oversampling	ASR	WFC	FAD	FAW	AR
Accuracy rate (%)	95.3 (0.9)	93.7 (0.5)	95.0 (0.4)	94.9 (1.6)	93.1 (0.7)	93.3 (0.8)	**95.0 (0.5)**	93.6 (0.8)	**95.3 (0.9)**
Disparate Impact (%)	3.7 (0.7)	3.9 (0.9)	**3.1 (0.3)**	3.4 (0.3)	4.3 (0.9)	5.1 (1.4)	18.2 (1.4)	3.8 (0.4)	**0.7 (0.4)**
Disparate FPR (%)	-16.0 (1.7)	-32.1 (7.3)	-28.9 (4.1)	-17.4 (5.0)	**-2.0 (4.4)**	**4.0 (1.7)**	-21.0 (2.6)	-19.7 (1.1)	-22.2 (5.3)
Disparate FNR (%)	10.3 (2.1)	16.6 (5.4)	11.9 (3.7)	10.9 (3.7)	**8.1 (0.8)**	**-7.2 (2.1)**	10.5 (1.4)	10.0 (1.1)	9.0 (3.7)



(a) Accuracy vs. T

(b) Disparate Impact vs. T



(c) Disparate FPR vs. T

(d) Disparate FNR vs. T

Figure A.1: Sensitivity of metrics on the change of T on CelebA 90% male and 10% female.

Details to approximate Wasserstein distance with and without reweighting

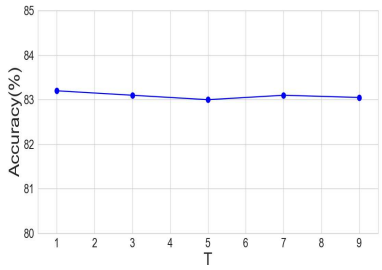
To approximate the Wasserstein distance before reweighting, we train the feature extractor and discriminator without the weights assigned to any samples. When the NNs are trained, we use the discriminator to approximate the Wasserstein distance between the two groups.

Breakdown of accuracy on sensitive groups

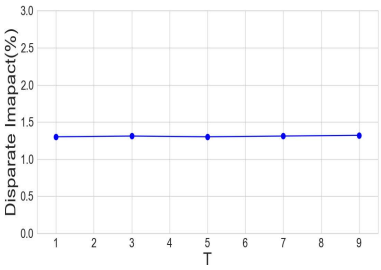
We could see that the method sacrifices the accuracy of the majority group for the accuracy of the minority group in Table A.2.

Ablation test for assigning weights

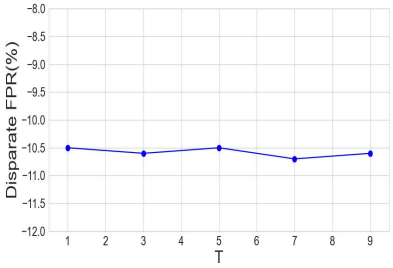
We try to assign weights only to the minority group. We could not close the Wasserstein distance gap and assign weights only to the majority group. Assigning weights to



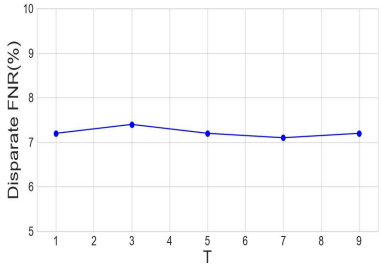
(a) Accuracy vs. T



(b) Disparate Impact vs. T

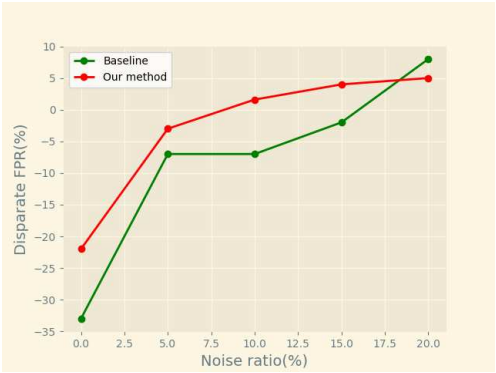


(c) Disparate FPR vs. T

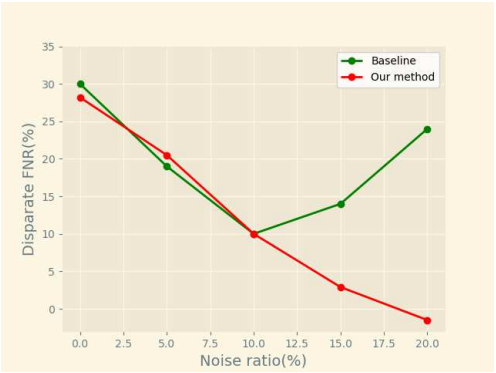


(d) Disparate FNR vs. T

Figure A.2: Sensitivity of metrics to the change of T on Adult dataset



(a) Disparate FPR vs. Noise ratio



(b) Disparate FNR vs. Noise ratio

Figure A.3: Change of disparate FPR and disparate FNR under different noise ratios on CelebA 90% male and 10% female.

both groups could achieve similar results as our method. However, we might need an additional statistical test to claim so.

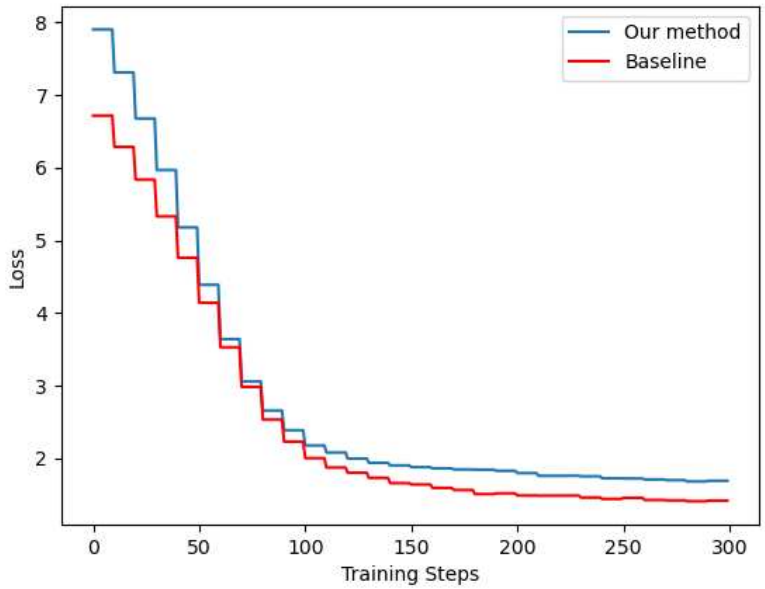


Figure A.4: convergence of training loss

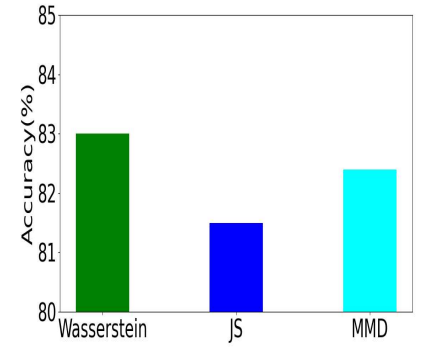
methods	accuracy (%)		
	male group	female group	total
baseline	95.6	92.6	95.0
our method	95.1	94.1	95.3

Table A.2: breakdown of accuracy on 90% male and 10% female CelebA dataset

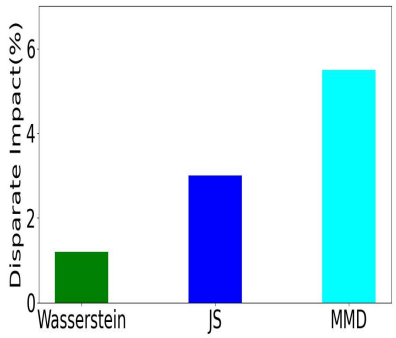
Training on Adult Dataset and German Credit Dataset

Since the tabular datasets have relatively lower dimensions than the image datasets, we could avoid using feature extractors. However, we use one-hot encoding to deal with the categorical features in the latent space. We could have a better continuity for the Wasserstein distance approximation. From Figure A.2, we can see that the metrics are not sensitive to the change of T for the Adult dataset. Figure A.5 shows the sensitivity of different distance measures on the Adult dataset.

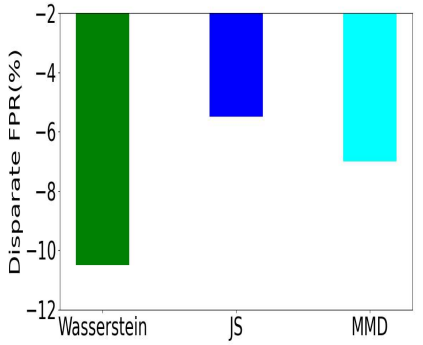
For the feature extractor F_ϕ and the classifier C_θ, we also apply fully connected layers. For the discriminator D, we use the same architecture in Gulrajani *et al.* (2017), without the last sigmoid function. We apply SGD algorithm with a momentum of 0.9 to update ϕ and θ. The learning rate of θ is ten times that of ϕ. θ_D is updated by Adam algorithm with a learning rate 0.0001. Following Gulrajani *et al.* (2017), we adjust the learning rate η of θ by $\eta = \frac{0.01}{(1+10p)^{-0.75}}$, where p is the training progress linearly changing from 0 to



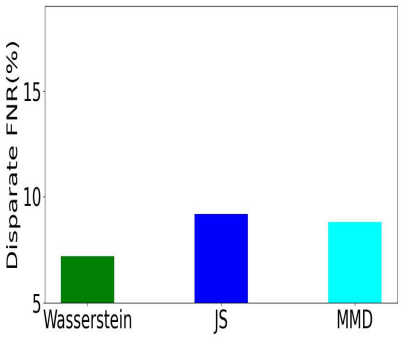
(a) Accuracy vs. Distance Methods



(b) Disparate Impact vs. Distance Methods



(c) Disparate FPR vs. Distance Methods



(d) Disparate FNR vs. Distance Methods

Figure A.5: sensitivity of different distance measures on Adult dataset

1. We update ϕ and θ for 2 steps then update θ_D for 1 step.

Multi-categorical sensitive attribute situation

It is straight-forward to extend our method to handle a multi-categorical sensitive attribute or multi-sensitive attributes by using one subgroup as reference group and reweighing other subgroups alternatively (and in turn) to reach a state of demographic parity. We also use Adult dataset to demonstrate this. However, we choose *race* here as the sensitive attribute. *race* is $\{'Amer-Indian-Eskimo': 0, 'Asian-Pac-Islander': 1, 'Black': 2, 'Other': 3, 'White': 4\}$ in the dataset. We use $'Asian-Pac-Islander'$ as the reference subgroup and reweighs samples from other subgroups. We report the disparate impact between the subgroup $'White'$ and $'Black'$. Before and after applying our method, the disparate impact is 15.1% and 1.7% and the accuracy is 83.1% and 82.8%.

Appendix B

Causal Fairness-Guided Dataset Reweighting using Neural Networks

B.1 Dataset and Training Details

The causal graph (Chae *et al.*, 2018) for ADULT is shown in Figure B.1, and for COMPAS (Plecko *et al.*, 2021) in Figure B.2. Note that the causal graphs here are sourced from existing literature.

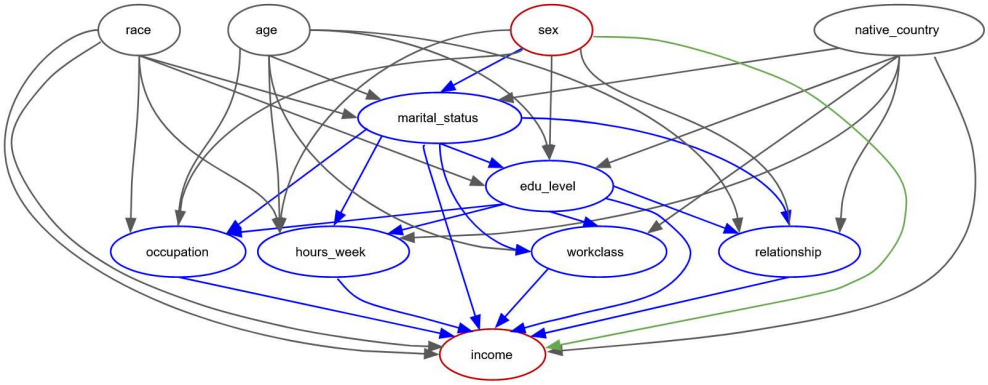


Figure B.1: The causal graph of the Adult dataset depicts the indirect path set with blue paths, while the direct path is represented by the green path.

Adult Dataset

The Adult dataset was drawn from the 1994 United States Census Bureau data. It contains 65,123 samples with 11 variables. It used personal information such as education level and working hours per week to predict whether an individual earns more or less than $50,000 per year. The dataset is imbalanced – the instances made less than $50,000 constitute 25% of the dataset, and the instances made more than $50,000 constitute 75%

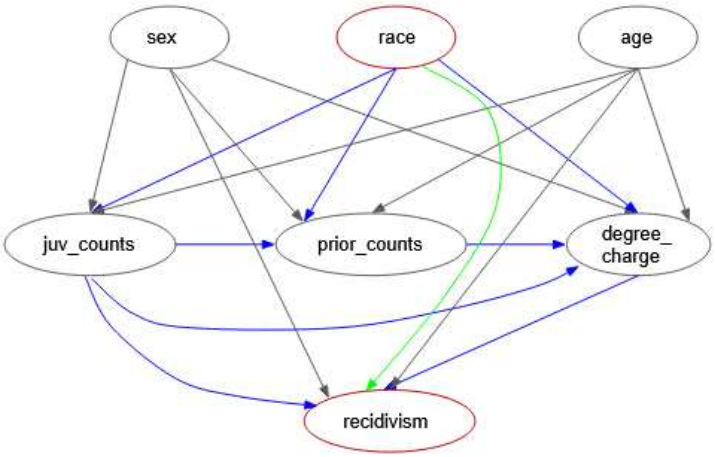


Figure B.2: The causal graph of the COMPAS dataset depicts the indirect path set with blue paths, while the direct path is represented by the green path

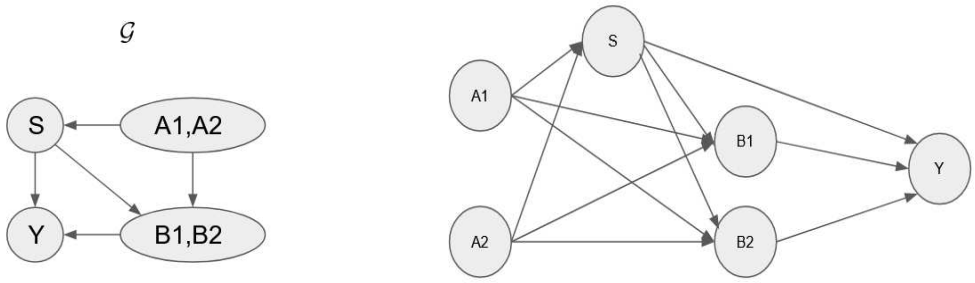
of the dataset. As for gender, it is also imbalanced. We use age, years of education, capital gain, capital loss, hours-per-week, etc., as continuous features, and education level, gender, etc., as categorical features. We set the batch size at 640 and train 30 epochs for convergence. We set the learning rate η at 0.001 according to the experiment result.

COMPAS Dataset

COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is a popular commercial algorithm used by judges and parole officers for scoring criminal defendant's likelihood of reoffending (recidivism). The COMPAS dataset includes the processed COMPAS data between 2013-2014. The data cleaning process followed the guidance in the original COMPAS repo. It Contains 6172 observations and 14 features. In our causal graph, we use 7 features. Due to the limited size of COMPAS dataset, it does not perform so well on NN based tasks.

B.2 Training Details

For ADULT and COMPAS datasets, some pre-processing is performed. We normalize the continuous features and use one-hot encoding to deal with the categorical features for the input of F^1 and F^2. We use sex and race as the sensitive variable S in ADULT and COMPAS respectively, income and two year recidivism as the outcome variable Y.



(a) Causal graph $\mathcal{G}$ (b) Neural Network F^1

Figure B.3: details of the connection of the neural nets of a given $\mathcal{G}$. In Figure B.3b, each nodes are either input or output of a sub-neural nets or both. Note that we do not show the inner layers here for simplicity.

For F^1 and F^2, we apply fully connected layers. For the discriminator D, we use the same architecture proposed in Gulrajani *et al.* (2017). We apply SGD algorithm with a momentum of 0.9 to update F^1 and F^2. D is updated by the Adam algorithm with a learning rate 0.0001. Following Gulrajani *et al.* (2017), we adjust the learning rate η by $\eta = \frac{0.01}{(1+10p)^{-0.75}}$, where p is the training progress linearly changing from 0 to 1. We update F^1 and F^2 for 2 steps then update D for 1 step. We then evaluate the performance of our method of reweighting to achieve different types of causal fairness and utility.

Our test are run on an Intel(r) Core(TM) i7-8700 CPU. The networks in the experiments are built based on Pytorch (Paszke *et al.*, 2019), the optimization in Equation (2.10) is performed with the Python package CVXPY (Diamond and Boyd, 2016).

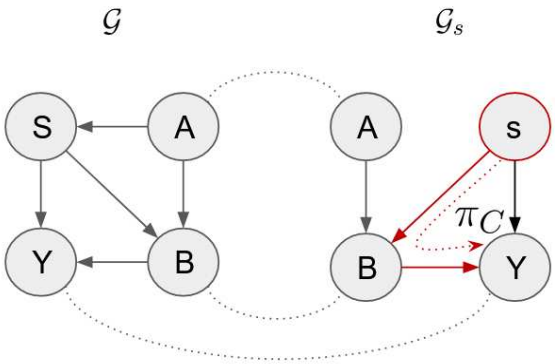
Details of architectures of the feed-forward Neural Networks F^1 and F^2 with sub-neural networks

To simplify our demonstration, we consider a causal graph $\mathcal{G}$ with 6 attributes $\{S, A_1, A_2, B_1, B_2, Y\}$ as shown in Figure B.3a. And Figure B.3b shows the joint neural network of it.

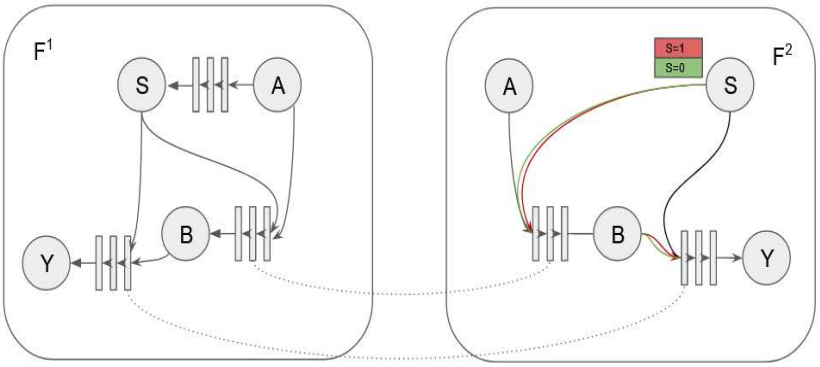
Details of WGAN-GP adaptation for our method

In our design, we adopt the discriminator from WGAN-GP: in the original work, the discriminator is used to differentiate between the generated and real data while we are trying to differentiate between S^+ and S^-. The difference between orginal GAN and WGAN-GP is that WGAN-GP introduces a gradient penalty term in the training objective to guarantee Wasserstein distance. Wasserstein distance itself has been used a lot in fairness realted topic to help detect or mitigate bias. Note that we choose relatively larger

batch size since to approximate Wasserstein distance between two distributions requires relatively larger batch size.



(a) Causal graph $\mathcal{G}$ and interventional graph $\mathcal{G}_s$ with the indirect interventional path π_C

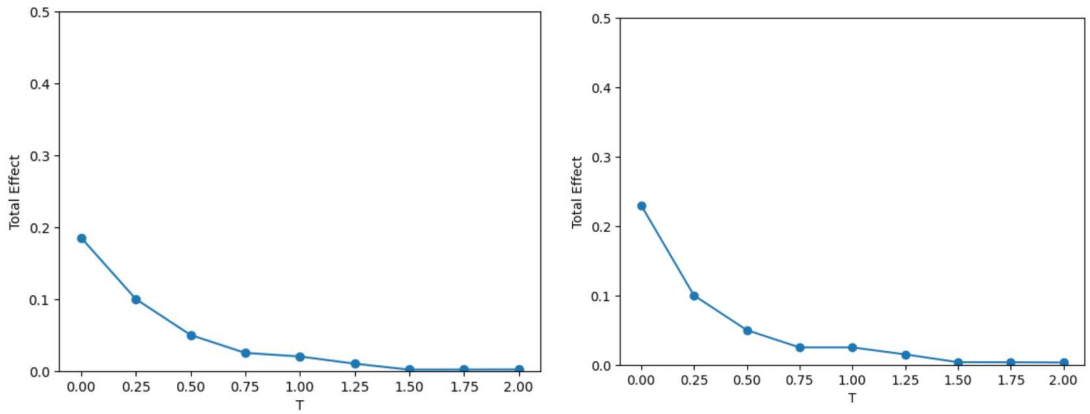


(b) Neural Networks F^1 and F^2

Figure B.4: the Neural Networks F^1 and F^2 based on indrect discrimination. S is 1 or 0 and the intervention is only along $\pi_C = \{S \to B \to Y\}$ for the interventional distributions $P_{F^2}(s^+)$ (red) and $P_{F^2}(s^-)$ (green) respectively. Compared with Figure 3.2, we could see that the intervention is not transferred directly from S to Y ($\{S \to Y\}$) in Figure B.4.

Sensitivity to the Choice of Hyper-Parameters

We conduct an analysis of the sensitivity of our method to the hyper-parameters discussed in Section 3.2, and the results are shown in Figure B.5a. The figures demon-



(a) ADULT dataset. (b) COMPAS dataset.

Figure B.5: Sensitivity of total effect on the change of T.

strate that our adversarial reweighting scheme's performance has low sensitivity to hyper-parameter choice when T is above 1. Therefore, we set T at 1.5.

We could see from the Figure B.5b that the trend of T on total effect on COMPAS dataset is similar to what is shown earlier on ADULT dataset.

Details of calculating different causal effects

We have a discriminator in our design, to calculate different causal effect, we just send the samples from different groups (S^+ and S^-), normalize the output of D within groups, then get the difference as causal effect.

Repetition

We repeat experiments on each dataset five times. Before each repetition, we randomly split data into training data (80%) and test data (20%) for the computation of the standard errors of the metrics. We train 30 epochs for convergence.

Choices of Counterfactual Effect

Due to space limit, we only show two combinations of the counterfactual effect on individual features. For ADULT, we use $o_1 = \{white, us\}$, $o_2 = \{non\_white, us\}$. For COMPAS, we use $o_1 = \{male, under\_25\}$, $o_2 = \{male, above\_25\}$

Details to approximate Wasserstein distance

To approximate the Wasserstein distance, we WGAN-GP discriminator between the orginal data and data for evaluation. When the NNs are trained, we use the discriminator to approximate the Wasserstein distance between the two datasets.

Appendix C

Enhancing Fairness through Reweighting: A Path to Attain the Sufficiency Rule

C.1 Sufficiency Rule

The Sufficiency principle adopts a perspective where individuals receiving the same model decision are expected to experience similar outcomes regardless of their sensitive attributes. In this context, S represents the sensitive attribute.

On the other hand, Separation deals with error rates concerning the proportion of errors relative to the ground truth. For instance, it considers the number of individuals whose loan application would have been approved but were actually denied. Sufficiency, however, considers the number of individuals who would default on their loan among those who were granted it.

From a mathematical standpoint, this distinction resembles that between recall (or true positive rate) and precision, denoted as $P[\hat{Y} = 1|Y = 1]$ and $P[Y = 1|\hat{Y} = 1]$, respectively. A fairness principle that focuses on this type of error rate is known as Predictive Parity, also termed as the outcome test:

$$P[Y = 1|S = a, \hat{Y} = 1] = P[Y = 1|S = b, \hat{Y} = 1], \forall a, b \in \mathcal{S} \tag{C.1}$$

In other words, the model's precision should remain consistent across different sensitive groups. If we extend this requirement to apply to the scenario where $Y = 0$, we obtain the following statement of conditional independence:

$$Y \perp S|\hat{Y} \tag{C.2}$$

This concept is known as sufficiency.

Predictive Parity, including its broader form of sufficiency, focuses on ensuring parity in errors among individuals receiving the same decision. Predictive Parity, in particular, considers the viewpoint of the decision-maker, as they categorize individuals based on decisions rather than actual outcomes. For instance, in the context of credit lending,

sufficiency is more under the control of the decision-maker than separation. This is because achieving parity given a decision is directly observable, whereas parity given truth is only known afterward. Furthermore, the group of individuals granted a loan ($\hat{Y} = 1$) is less susceptible to selection bias compared to the group of loan repayments ($Y = 1$). We have information only on repayment for the $\hat{Y} = 1$ group, while nothing is known about the others ($\hat{Y} = 0$).

Similarly, other group metrics can be defined, such as Equality of Accuracy across groups: $P[\hat{Y} = Y|S = a] = P[\hat{Y} = Y|S = b]$, for all $a, b \in \mathcal{S}$, focusing on unconditional errors, among others.

Incompatibility

In most cases, even in classification setting, the actual output of a model is not a binary value, but rather a score $C \in \mathbb{R}$.

1. if $Y \not\perp S$, then it's impossible for sufficiency and independence to coexist. Consequently, if there's an imbalance in base rates among groups identified by A, enforcing both sufficiency and independence simultaneously is unfeasible.

2. if $Y \not\perp S$ and the distribution (S, C, Y) is strictly positive, then separation and sufficiency cannot both be achieved simultaneously. This implies that separation and sufficiency can coexist only under two conditions: when there's no imbalance in sensitive groups (indicating independence of the target from sensitive attributes), or when the joint probability (S, C, Y) is degenerate. In the case of binary targets, this degeneracy occurs when there are specific values of S and C for which only $Y = 1$ (or $Y = 0$) holds. In other words, separation and sufficiency coincide when the score perfectly resolves the uncertainty in the target. For instance, the ideal classifier where $C = Y$ effortlessly satisfies both sufficiency and separation.

C.2 Invariant Risk Minimization

Minimizing training error leads machines into recklessly absorbing all the correlations found in training data. Understanding which patterns are useful has been previously studied as a correlation-versus-causation dilemma, since spurious correlations stemming from data biases are unrelated to the causal explanation of interest. Following this line, IRM leverage tools from causation to develop the mathematics of spurious and invariant correlations, in order to alleviate the excessive reliance of machine learning systems on data biases, allowing them to generalize to new test distributions.

C.2.1 Invariance, causality and generalization

IRM principle promotes low error and invariance across training environments. When do these conditions imply invariance across all environments? More importantly, when do these conditions lead to low error across all enviroments, and consequently out-of-distribution generalization? And at a more fundamental level, how does statistical invariance and out-of-distribution generalization relate to concepts from the theory of causation? So far, we have omitted how different environments should relate to enable out-of-distribution generalization. The answer to this question is rooted in the theory of causation. We begin by assuming that the data from all the environments share the same underlying Structural Equation Model.

By running the structural equations of a SEM according to the topological ordering of its causal graph, we can draw samples from the observational distribution $P(X)$. In addition, we can manipulate (intervene) an unique SEM in different ways, indexed by e, to obtain different but related SEMs. Similarly, by running the structural equations of the intervened SEM, we can draw samples from the interventional distribution $P(X^e)$. Admitting a slight abuse of notation, each intervention e generates a new environment e with interventional distribution $P(X^e, Y^e)$. Valid interventions e, those that do not destroy too much information about the target variable Y , form the set of all environments.

Prior work (Peters *et al.*, 2015) considered valid interventions as those that do not change the structural equation of Y , since arbitrary interventions on this equation render prediction impossible. In this work, we also allow changes in the noise variance of Y, since varying noise levels appear in real problems, and these do not affect the optimal prediction rule. We formalize this as follows.

Consider a SEM governing the random vector $(X_1, ..., X_d, Y)$, and the learning goal of predicting Y from X. Then, the set of all environments indexes all the interventional distributions $P(X^e, Y^e)$ obtainable by valid interventions e. An intervention e is valid as long as (i) the causal graph remains acyclic, (ii) $\mathbb{E}[Y^e|Pa(Y)] = \mathbb{E}[Y|Pa(Y)]$, and (iii) $\mathbb{V}[Y^e|Pa(Y)]$ remains within a finite range. Condition (iii) can be waived if one takes into account environment specific baselines into the definition of R^{OOD}, similar to those appearing in the robust learning objective R_{rob}. We leave additional quantifications of out-of-distribution generalization for future work. The previous definitions establish fundamental links between causation and invariance. Moreover, one can show that a predictor $v : X \to Y$ is invariant across all enviroments if and only if it attains optimal R^{OOD}, and if and only if it uses only the direct causal parents of Y to predict, that is, $v(x) = \mathbb{E}_{N_Y}[f_Y(Pa(Y), N_Y)]$. The rest of this section follows on these ideas to showcase how invariance across training environments can enable out-of-distribution generalization across all environments.

C.2.2 Risks of Invariant Risk Minimization

Various works on invariant prediction (Muandet *et al.*, 2013; Ghassami *et al.*, 2017; Heinze-Deml *et al.*, 2018; Rojas-Carulla *et al.*, 2018; Subbaswamy *et al.*, 2019; Christiansen *et al.*, 2022) consider regression in both the linear and non-linear setting, but they exclusively focus on learning with fully or partially observed covariates or some other source of information. Under such a condition, results from causal inference (Peters *et al.*, 2017) allow for formal guarantees of the identification of the invariant features, or at least a strict subset of them. With the rise of deep learning, more recent literature has developed objectives for learning invariant representations when the data are a non-linear function of unobserved latent factors, a common assumption when working with complex, high-dimensional data such as images. Causal discovery and inference with unobserved confounders or latents is a much harder problem (Peters *et al.*, 2017), so while empirical results seem encouraging, these objectives are presented with few formal guarantees. IRM is one such objective for invariant representation learning. The goal of IRM is to learn a feature embedder such that the optimal linear predictor on top of these features is the same for every environment—the idea being that only the invariant features will have an optimal predictor that is invariant. Recent works have pointed to shortcomings of IRM and have suggested modifications which they claim prevent these failures. However, these alternatives are compared in broad strokes, with little in the way of theory.

We promote invariance as the main feature of causation. Unsurprisingly, we are not pioneers in doing so. To predict the outcome of an intervention, we rely on (i) the properties of our intervention and (ii) the properties assumed invariant after the intervention. Pearl's do-calculus (Shanmugam, 2001) on causal graphs is a framework that tells which conditionals remain invariant after an intervention. Rubin's ignorability (Rubin, 1974) plays the same role. What's often described as autonomy of causal mechanisms (Aldrich, 1989; Haavelmo *et al.*, 1995) is a specification of invariance under intervention. A large body of philosophical work (Mitchell, 2000; Cartwright, 2003; Woodward, 2003) studies the close link between invariance and causation. Some works in machine learning (Schölkopf *et al.*, 2012; Ghassami *et al.*, 2017) pursue similar questions.

Another motivation supporting the invariance view of causation are the problems studied in machine learning. For instance, consider the task of image classification. Here, the observed variables are hundreds of thousands of correlated pixels. What is the causal graph governing them? It is reasonable to assume that causation does not happen between pixels, but between the real-world concepts captured by the camera. In these cases, invariant correlations in images are a proxy into the causation at play in the real world. To find those invariant correlations, we need methods which can disentangle the observed pixels into latent variables closer to the realm of causation, such as IRM. In rare occasions we are truly interested in the entire causal graph governing all the variables in our learning problem. Rather, our focus is often on the causal invariances improving generalization across novel distributions of examples.

During training, a classifier will learn to leverage correlations between features and labels in the training data to make its predictions. If a correlation varies with the environment, it may not be present in future test distributions—worse yet, it may be reversed—harming the classifier's predictive ability. IRM (Arjovsky *et al.*, 2019) is a recently proposed approach to learning environmentally invariant representations to facilitate invariant prediction. The IRM objective. IRM posits the existence of a feature embedder such that the optimal classifier on top of these features is the same for every environment. The authors argue that such a function will use only invariant features, since non-invariant features will have different joint distributions with the label and therefore a fixed classifier on top of them won't be optimal in all environments. To learn this embedder, the IRM objective is the following constrained optimization problem.

This bilevel program is highly non-convex and difficult to solve. To find an approximate solution, the authors consider a Langrangian form, whereby the sub-optimality with respect to the constraint is expressed as the squared norm of the gradients of each of the inner optimization problems.

Alternative objectives. IRM is motivated by the existence of a featurizer such that $\mathbb{E}[y|\phi(x)]$ is invariant. Follow-up works have proposed variations on this objective, based instead on the strictly stronger desideratum of the invariance of $p(y|\phi(x))$. Krueger *et al.* (2021) suggest penalizing the variance of the risks, while Xie *et al.* (2020) give the same objective but taking the square root of the variance. Many papers have suggested similar alternatives (Jin *et al.*, 2020; Mahajan *et al.*, 2021). These objectives are compelling—indeed, it is easy to show that the optimal invariant predictor constitutes a stationary point of each of these objective.

C.3 Experiment Details and Results

C.3.1 Datasets

Adult dataset The Adult dataset was drawn from the 1994 United States Census Bureau data. It used personal information such as education level and working hours per week to predict whether an individual earns more or less than $50,000 per year. The dataset is imbalanced – the instances made less than $50,000 constitute 25% of the dataset, and the instances made more than $50,000 include 75% of the dataset. As for gender, it is also imbalanced. We use age, years of education, capital gain, capital loss, hours-per-week, etc., as continuous features, and education level, gender, etc., as categorical features.

COMPAS dataset COMPAS (Correctional Offender Management Profiling for Alternative Sanctions) is a popular commercial algorithm used by judges and parole officers for scoring criminal defendant's likelihood of reoffending (recidivism). The COMPAS dataset includes the processed COMPAS data between 2013-2014. The data cleaning

process followed the guidance in the original COMPAS repo. It Contains 6172 observations and 14 features. In our causal graph, we use 7 features. Due to the limited size of COMPAS dataset, it does not perform so well on NN based tasks. Note that we need more pre-processing on the tabular datasets. We normalize the continuous features and use one-hot encoding to deal with the categorical features.

C.3.2 Training Details

In our experiments, we set the following hyperparameters for optimization. In the inner loop, the model undergoes training for 100 epochs using Stochastic Gradient Descent (SGD) with a learning rate of 0.1 and momentum of 0.9. In the outer loop, the probabilities s are optimized using Adam with a learning rate of 2.5 and a cosine scheduler. The outer loop is updated iteratively for 500 to 2000 times. It's worth mentioning that we employ architectures with fully connected layers for the classifier.

Toxic comments We split the training, validation and testing set as 70%, 10% and 20%. The mini-batch-size is set as 500.

CelebA The training/validation/test set are around 82K, 18K and 18K. The batch-size is set as 1000.

Effectiveness of Sparsity in Promoting Training Speed An experiment is conducted on the CelebA dataset to compare the training speed between our method with and without a sparsity constraint on sample sizes. The inclusion of the constraint significantly reduces the inner loop computation time, decreasing it from 9.24 to 5.72 hours.

Repetition We repeat experiments on each dataset five times. Before each repetition, we randomly split data into training data and test data for the computation of the standard errors of the metrics.

Bibliography

Adel, T., Valera, I., Ghahramani, Z., and Weller, A. (2019). One-network adversarial fairness. In *AAAI*, pages 2412–2420. AAAI Press.

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. M. (2018). A reductions approach to fair classification. In J. G. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69. PMLR.

Aghaei, S., Azizi, M. J., and Vayanos, P. (2019). Learning optimal and fair decision trees for non-discriminative decision-making. In *AAAI*, pages 1418–1426. AAAI Press.

Ahmad, M. A., Patel, A., Eckert, C., Kumar, V., and Teredesai, A. (2020). Fairness in machine learning for healthcare. In *KDD*, pages 3529–3530. ACM.

Aïvodji, U., Bidet, F., Gambs, S., Ngueveu, R. C., and Tapp, A. (2021). Local data debiasing for fairness based on generative adversarial training. *Algorithms*, **14**(3), 87.

Aldrich, J. (1989). Autonomy. *Oxford Economic Papers*, **41**(1), 15–34.

Arjovsky, M., Bottou, L., Gulrajani, I., and Lopez-Paz, D. (2019). Invariant risk minimization. *CoRR*, **abs/1907.02893**.

Bao, F., Deng, Y., Kong, Y., Ren, Z., Suo, J., and Dai, Q. (2020). Learning deep landmarks for imbalanced classification. *IEEE Trans. Neural Networks Learn. Syst.*, **31**(8), 2691–2704.

Barocas, S. and Selbst, A. D. (2014). Big Data's Disparate Impact. *SSRN eLibrary*.

Barocas, S., Hardt, M., and Narayanan, A. (2023). *Fairness and Machine Learning: Limitations and Opportunities*. MIT Press.

Bellamy, R. K. E., Dey, K., Hind, M., Hoffman, S. C., Houde, S., Kannan, K., Lohia, P., Martino, J., Mehta, S., Mojsilovic, A., Nagar, S., Ramamurthy, K. N., Richards, J. T., Saha, D., Sattigeri, P., Singh, M., Varshney, K. R., and Zhang, Y. (2019). AI fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM J. Res. Dev.*, **63**(4/5), 4:1–4:15.

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M. J., Morgenstern, J., Neel, S., and Roth, A. (2017). A convex framework for fair regression. *CoRR*, **abs/1706.02409**.

Beutel, A., Chen, J., Zhao, Z., and Chi, E. H. (2017). Data decisions and theoretical implications when adversarially learning fair representations. *CoRR*, **abs/1707.00075**.

Binkyte-Sadauskiene, R., Makhlouf, K., Pinzón, C., Zhioua, S., and Palamidessi, C. (2022). Causal discovery for fairness. *CoRR*, **abs/2206.06685**.

Binns, R. (2018). Fairness in machine learning: Lessons from political philosophy. In *FAT*, volume 81 of *Proceedings of Machine Learning Research*, pages 149–159. PMLR.

Binns, R. (2020). On the apparent conflict between individual and group fairness. In M. Hildebrandt, C. Castillo, L. E. Celis, S. Ruggieri, L. Taylor, and G. Zanfir-Fortuna, editors, *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pages 514–524. ACM.

Borisov, V., Leemann, T., Seßler, K., Haug, J., Pawelczyk, M., and Kasneci, G. (2022). Deep neural networks and tabular data: A survey. *IEEE Transactions on Neural Networks and Learning Systems*.

Briggs, R. (2012). Interventionist counterfactuals. *Philosophical Studies*, **160**.

Brooks, R. L. (1991). *Rethinking the American Race Problem*. University of California Press, Berkeley.

Buolamwini, J. and Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. In S. A. Friedler and C. Wilson, editors, *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pages 77–91. PMLR.

Bühlmann, P. (2018). Invariance, causality and robustness.

Calmon, F. P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., and Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In *NIPS*, pages 3992–4001.

Cartwright, N. (2003). Two theorems on invariance and causality. *Philosophy of Science*, **70**(1), 203–224.

Castelnovo, A., Crupi, R., Greco, G., Regoli, D., Penco, I. G., and Cosentini, A. C. (2022). A clarification of the nuances in the fairness metrics landscape. *Scientific Reports*, **12**.

Caton, S. and Haas, C. (2024). Fairness in machine learning: A survey. *ACM Comput. Surv.*, **56**(7), 166:1–166:38.

Chae, D., Kang, J., Kim, S., and Lee, J. (2018). CFGAN: A generic collaborative filtering framework based on generative adversarial networks. In *CIKM*, pages 137–146. ACM.

Chai, J. and Wang, X. (2022). Fairness with Adaptive Weights. In *Proceedings of the 39th International Conference on Machine Learning*, pages 2853–2866. PMLR.

Chang, S., Zhang, Y., Yu, M., and Jaakkola, T. S. (2020). Invariant rationalization. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 1448–1458. PMLR.

Chen, Y., Rosenfeld, E., Sellke, M., Ma, T., and Risteski, A. (2022). Iterative feature matching: Toward provable domain generalization with logarithmic environments. In *NeurIPS*.

Cherepanova, V., Nanda, V., Goldblum, M., Dickerson, J. P., and Goldstein, T. (2021). Technical challenges for training fair neural networks. *CoRR*, **abs/2102.06764**.

Chiappa, S. (2019). Path-specific counterfactual fairness. In *AAAI*, pages 7801–7808. AAAI Press.

Chiappa, S. and Pacchiano, A. (2021). Fairness with continuous optimal transport. *CoRR*, **abs/2101.02084**.

Choe, Y. J., Ham, J., and Park, K. (2020). An empirical study of invariant risk minimization.

Choi, J., Gao, C., Messou, J. C. E., and Huang, J. (2019). Why can't I dance in the mall? learning to mitigate scene bias in action recognition. In *NeurIPS*, pages 851–863.

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

Christiansen, R., Pfister, N., Jakobsen, M. E., Gnecco, N., and Peters, J. (2022). A causal framework for distribution generalization. *IEEE Trans. Pattern Anal. Mach. Intell.*, **44**(10), 6614–6630.

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. (2020). Fair regression with wasserstein barycenters. In *NeurIPS*.

Colombo, D., Maathuis, M. H., Kalisch, M., and Richardson, T. S. (2012). Learning high-dimensional directed acyclic graphs with latent and selection variables. *The Annals of Statistics*, **40**(1), 294–321.

Cortes, C. and Vapnik, V. (1995). Support-vector networks. *Machine learning*, **20**(3), 273–297.

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

Creager, E., Madras, D., Jacobsen, J., Weis, M. A., Swersky, K., Pitassi, T., and Zemel, R. S. (2019). Flexibly fair representation learning by disentanglement. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 1436–1445. PMLR.

D'Amour, A., Srinivasan, H., Atwood, J., Baljekar, P., Sculley, D., and Halpern, Y. (2020). Fairness is not static: deeper understanding of long term fairness via simulation studies. In M. Hildebrandt, C. Castillo, L. E. Celis, S. Ruggieri, L. Taylor, and G. Zanfir-Fortuna, editors, *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pages 525–534. ACM.

Datta, T., Nissani, D., Cembalest, M., Khanna, A., Massa, H., and Dickerson, J. (2023). Tensions between the proxies of human values in AI. In *2023 IEEE Conference on Secure and Trustworthy Machine Learning, SaTML 2023, Raleigh, NC, USA, February 8-10, 2023*, pages 678–689. IEEE.

Defrance, M. and Bie, T. D. (2023). Maximal fairness. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2023, Chicago, IL, USA, June 12-15, 2023*, pages 851–880. ACM.

Deng, J., Dong, W., Socher, R., Li, L., Li, K., and Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *CVPR*, pages 248–255. IEEE Computer Society.

Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2019). BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186. Association for Computational Linguistics.

Diamond, S. and Boyd, S. (2016). CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, **17**(83), 1–5.

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

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. S. (2012). Fairness through awareness. In *ITCS*, pages 214–226. ACM.

Edwards, H. and Storkey, A. J. (2016). Censoring representations with an adversary. In *ICLR (Poster)*.

Ensign, D., Friedler, S. A., Neville, S., Scheidegger, C. E., and Venkatasubramanian, S. (2018). Decision making with limited feedback. In F. Janoos, M. Mohri, and K. Sridharan, editors, *Algorithmic Learning Theory, ALT 2018, 7-9 April 2018, Lanzarote, Canary Islands, Spain*, volume 83 of *Proceedings of Machine Learning Research*, pages 359–367. PMLR.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *KDD*, pages 259–268. ACM.

Fenton, N. E., Neil, M., and Constantinou, A. C. (2020). The book of why: The new science of cause and effect, judea pearl, dana mackenzie. basic books (2018). *Artif. Intell.*, **284**, 103286.

Ferraro, A., Serra, X., and Bauer, C. (2021). Break the loop: Gender imbalance in music recommenders. In F. Scholer, P. Thomas, D. Elsweiler, H. Joho, N. Kando, and C. Smith, editors, *CHIIR '21: ACM SIGIR Conference on Human Information Interaction and Retrieval, Canberra, ACT, Australia, March 14-19, 2021*, pages 249–254. ACM.

Fiscella, K. and Fremont, A. (2006). Use of geocoding and surname analysis to estimate race and ethnicity. *Health services research*, **41**, 1482–500.

Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. (2016). On the (im)possibility of fairness. *CoRR*, **abs/1609.07236**.

Friedler, S. A., Scheidegger, C., Venkatasubramanian, S., Choudhary, S., Hamilton, E. P., and Roth, D. (2019). A comparative study of fairness-enhancing interventions in machine learning. In danah boyd and J. H. Morgenstern, editors, *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 329–338. ACM.

Ghassami, A., Salehkaleybar, S., Kiyavash, N., and Zhang, K. (2017). Learning causal structures using regression invariance. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 3011–3021.

Glymour, C., Zhang, K., and Spirtes, P. (2019). Review of Causal Discovery Methods Based on Graphical Models. *Frontiers in Genetics*, **10**.

Grazzi, R., Franceschi, L., Pontil, M., and Salzo, S. (2020). On the iteration complexity of hypergradient computation. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 3748–3758. PMLR.

Gu, X., Yu, X., Yang, Y., Sun, J., and Xu, Z. (2021). Adversarial reweighting for partial domain adaptation. In *NeurIPS*, pages 14860–14872.

Gulrajani, I. and Lopez-Paz, D. (2021). In search of lost domain generalization. In *ICLR*. OpenReview.net.

Gulrajani, I., Ahmed, F., Arjovsky, M., Dumoulin, V., and Courville, A. C. (2017). Improved training of wasserstein gans. In *NIPS*, pages 5767–5777.

Haavelmo, T., Hendry, D. F., and Morgan, M. S. (1995). *The Probability Approach in Econometrics (Supplement to Econometrica, vol. 12, 1944, pp. iii–iv, 1–11, 49–52, 114–15)*, page 477–490. Cambridge University Press.

Hajian, S. and Domingo-Ferrer, J. (2013). A methodology for direct and indirect discrimination prevention in data mining. *IEEE Trans. Knowl. Data Eng.*, **25**(7), 1445–1459.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In *NIPS*, pages 3315–3323.

He, K., Zhang, X., Ren, S., and Sun, J. (2016). Deep residual learning for image recognition. In *CVPR*, pages 770–778. IEEE Computer Society.

Heidari, H., Loi, M., Gummadi, K. P., and Krause, A. (2019). A moral framework for understanding fair ML through economic models of equality of opportunity. In danah boyd and J. H. Morgenstern, editors, *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 181–190. ACM.

Heinze-Deml, C., Peters, J., and Meinshausen, N. (2018). Invariant causal prediction for nonlinear models.

Ho, T. K. (1998). The random subspace method for constructing decision forests. *IEEE Trans. Pattern Anal. Mach. Intell.*, **20**(8), 832–844.

Huang, C., Li, Y., Loy, C. C., and Tang, X. (2020). Deep imbalanced learning for face recognition and attribute prediction. *IEEE Trans. Pattern Anal. Mach. Intell.*, **42**(11), 2781–2794.

Hutchinson, B. and Mitchell, M. (2019). 50 years of test (un)fairness: Lessons for machine learning. In *FAT*, pages 49–58. ACM.

Ignatiev, A., Cooper, M. C., Siala, M., Hebrard, E., and Marques-Silva, J. (2020). Towards formal fairness in machine learning. In H. Simonis, editor, *Principles and Practice of Constraint Programming - 26th International Conference, CP 2020, Louvain-la-Neuve, Belgium, September 7-11, 2020, Proceedings*, volume 12333 of *Lecture Notes in Computer Science*, pages 846–867. Springer.

Ilvento, C. (2020). Metric learning for individual fairness. In A. Roth, editor, *1st Symposium on Foundations of Responsible Computing, FORC 2020, June 1-3, 2020, Harvard University, Cambridge, MA, USA (virtual conference)*, volume 156 of *LIPIcs*, pages 2:1–2:11. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Jiang, H. and Nachum, O. (2020). Identifying and correcting label bias in machine learning. In S. Chiappa and R. Calandra, editors, *The 23rd International Conference on Artificial Intelligence and Statistics, AISTATS 2020, 26-28 August 2020, Online [Palermo, Sicily, Italy]*, volume 108 of *Proceedings of Machine Learning Research*, pages 702–712. PMLR.

Jiang, R., Pacchiano, A., Stepleton, T., Jiang, H., and Chiappa, S. (2019). Wasserstein fair classification. In *UAI*, volume 115 of *Proceedings of Machine Learning Research*, pages 862–872. AUAI Press.

Jigsaw (2018). Toxic comment classification challenge.

Jin, W., Barzilay, R., and Jaakkola, T. S. (2020). Domain extrapolation via regret minimization. *CoRR*, **abs/2006.03908**.

Joseph, M., Kearns, M. J., Morgenstern, J., and Roth, A. (2016). Fairness in learning: Classic and contextual bandits. In D. D. Lee, M. Sugiyama, U. von Luxburg, I. Guyon, and R. Garnett, editors, *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 325–333.

Jung, C., Kearns, M., Neel, S., Roth, A., Stapleton, L., and Wu, Z. S. (2021). An algorithmic framework for fairness elicitation. In K. Ligett and S. Gupta, editors, *2nd Symposium on Foundations of Responsible Computing, FORC 2021, June 9-11, 2021, Virtual Conference*, volume 192 of *LIPIcs*, pages 2:1–2:19. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Kamiran, F. and Calders, T. (2009). Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*, pages 1–6.

Kamiran, F. and Calders, T. (2011). Data preprocessing techniques for classification without discrimination. *Knowl. Inf. Syst.*, **33**(1), 1–33.

Kamiran, F., Calders, T., and Pechenizkiy, M. (2010). Discrimination aware decision tree learning. In G. I. Webb, B. Liu, C. Zhang, D. Gunopulos, and X. Wu, editors, *ICDM 2010, The 10th IEEE International Conference on Data Mining, Sydney, Australia, 14-17 December 2010*, pages 869–874. IEEE Computer Society.

Kamiran, F., Karim, A., and Zhang, X. (2012). Decision theory for discrimination-aware classification. In M. J. Zaki, A. Siebes, J. X. Yu, B. Goethals, G. I. Webb, and X. Wu, editors, *12th IEEE International Conference on Data Mining, ICDM 2012, Brussels, Belgium, December 10-13, 2012*, pages 924–929. IEEE Computer Society.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. (2012). Fairness-aware classifier with prejudice remover regularizer. In P. A. Flach, T. D. Bie, and N. Cristianini,

editors, *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II*, volume 7524 of *Lecture Notes in Computer Science*, pages 35–50. Springer.

Kilbertus, N., Rojas-Carulla, M., Parascandolo, G., Hardt, M., Janzing, D., and Schölkopf, B. (2017). Avoiding discrimination through causal reasoning. In *NIPS*, pages 656–666.

Kilbertus, N., Gascón, A., Kusner, M. J., Veale, M., Gummadi, K. P., and Weller, A. (2018). Blind justice: Fairness with encrypted sensitive attributes. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 2635–2644. PMLR.

Kilbertus, N., Ball, P. J., Kusner, M. J., Weller, A., and Silva, R. (2019). The sensitivity of counterfactual fairness to unmeasured confounding. In A. Globerson and R. Silva, editors, *Proceedings of the Thirty-Fifth Conference on Uncertainty in Artificial Intelligence, UAI 2019, Tel Aviv, Israel, July 22-25, 2019*, volume 115 of *Proceedings of Machine Learning Research*, pages 616–626. AUAI Press.

Kilbertus, N., Rodriguez, M. G., Schölkopf, B., Muandet, K., and Valera, I. (2020). Fair decisions despite imperfect predictions. In *AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pages 277–287. PMLR.

Kim, B., Kim, H., Kim, K., Kim, S., and Kim, J. (2019). Learning not to learn: Training deep neural networks with biased data. In *CVPR*, pages 9012–9020. Computer Vision Foundation / IEEE.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *ICLR (Poster)*.

Kleinberg, J. M., Mullainathan, S., and Raghavan, M. (2017). Inherent trade-offs in the fair determination of risk scores. In *ITCS*, volume 67 of *LIPIcs*, pages 43:1–43:23. Schloss Dagstuhl - Leibniz-Zentrum für Informatik.

Kocaoglu, M., Snyder, C., Dimakis, A. G., and Vishwanath, S. (2018). CausalGAN: Learning causal implicit generative models with adversarial training. In *ICLR (Poster)*. OpenReview.net.

Kohavi, R. (1996). Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *KDD*, pages 202–207. AAAI Press.

Krasanakis, E., Xioufis, E. S., Papadopoulos, S., and Kompatsiaris, Y. (2018). Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *WWW*, pages 853–862. ACM.

Krco, N., Laugel, T., Loubes, J., and Detyniecki, M. (2023). When mitigating bias is unfair: A comprehensive study on the impact of bias mitigation algorithms. *CoRR*, **abs/2302.07185**.

Krueger, D., Caballero, E., Jacobsen, J., Zhang, A., Binas, J., Zhang, D., Priol, R. L., and Courville, A. C. (2021). Out-of-distribution generalization via risk extrapolation (rex). In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pages 5815–5826. PMLR.

Kusner, M. J., Loftus, J. R., Russell, C., and Silva, R. (2017). Counterfactual fairness. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4066–4076.

Lahoti, P., Gummadi, K. P., and Weikum, G. (2019). ifair: Learning individually fair data representations for algorithmic decision making. In *35th IEEE International Conference on Data Engineering, ICDE 2019, Macao, China, April 8-11, 2019*, pages 1334–1345. IEEE.

Li, P. and Liu, H. (2022). Achieving fairness at no utility cost via data reweighing with influence. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 12917–12930. PMLR.

Lin, Y., Dong, H., Wang, H., and Zhang, T. (2022). Bayesian invariant risk minimization. In *CVPR*, pages 16000–16009. IEEE.

Liu, L. T., Dean, S., Rolf, E., Simchowitz, M., and Hardt, M. (2019a). Delayed impact of fair machine learning. In S. Kraus, editor, *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*, pages 6196–6200. ijcai.org.

Liu, L. T., Simchowitz, M., and Hardt, M. (2019b). The implicit fairness criterion of unconstrained learning. In *ICML*, volume 97 of *Proceedings of Machine Learning Research*, pages 4051–4060. PMLR.

Liu, Z., Luo, P., Wang, X., and Tang, X. (2015). Deep learning face attributes in the wild. In *ICCV*, pages 3730–3738. IEEE Computer Society.

Loftus, J. R., Russell, C., Kusner, M. J., and Silva, R. (2018). Causal reasoning for algorithmic fairness. *CoRR*, **abs/1805.05859**.

Lohia, P. K., Ramamurthy, K. N., Bhide, M., Saha, D., Varshney, K. R., and Puri, R. (2019). Bias mitigation post-processing for individual and group fairness. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 2847–2851. IEEE.

Long, K. and Albert, S. (2021). Use of zip code based aggregate indicators to assess race disparities in covid-19. *Ethnicity and Disease*, **31**, 399–406.

Lorraine, J., Vicol, P., and Duvenaud, D. (2020). Optimizing millions of hyperparameters by implicit differentiation. In *AISTATS*, volume 108 of *Proceedings of Machine Learning Research*, pages 1540–1552. PMLR.

Lum, K. and Isaac, W. (2016). To predict and serve? *Significance*, **13**, 14–19.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. S. (2018). Learning adversarially fair and transferable representations. In *ICML*, volume 80 of *Proceedings of Machine Learning Research*, pages 3381–3390. PMLR.

Mahajan, D., Tople, S., and Sharma, A. (2021). Domain generalization using causal matching. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 7313–7324. PMLR.

Makhlouf, K., Zhioua, S., and Palamidessi, C. (2020). Survey on causal-based machine learning fairness notions. *CoRR*, **abs/2010.09553**.

Makhlouf, K., Zhioua, S., and Palamidessi, C. (2021). On the applicability of machine learning fairness notions. *SIGKDD Explor.*, **23**(1), 14–23.

Mallasto, A., Montúfar, G., and Gerolin, A. (2019). How well do wgans estimate the wasserstein metric? *CoRR*, **abs/1910.03875**.

Mattu, Angwin, J., Kirchner, L., Surya, and Larson, J. (2016). How We Analyzed the COMPAS Recidivism Algorithm.

Mirzasoleiman, B., Bilmes, J. A., and Leskovec, J. (2020). Coresets for data-efficient training of machine learning models. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pages 6950–6960. PMLR.

Mitchell, S. D. (2000). Dimensions of scientific law. *Philosophy of Science*, **67**(2), 242–265.

Muandet, K., Balduzzi, D., and Schölkopf, B. (2013). Domain generalization via invariant feature representation. In *Proceedings of the 30th International Conference on Machine Learning, ICML 2013, Atlanta, GA, USA, 16-21 June 2013*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 10–18. JMLR.org.

Obermeyer, Z., Powers, B., Vogeli, C., and Mullainathan, S. (2019). Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, **366**, 447–453.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Köpf, A., Yang, E. Z., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., Bai, J., and Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*, pages 8024–8035.

Pearl, J. (2009). *Causality: Models, Reasoning and Inference*. Cambridge University Press, 2 edition.

Pearl, J., Glymour, M., and Jewell, N. (2016). *Causal Inference in Statistics: A Primer*. Wiley.

Pedregosa, F. (2016). Hyperparameter optimization with approximate gradient. In *ICML*, volume 48 of *JMLR Workshop and Conference Proceedings*, pages 737–746. JMLR.org.

Pedreschi, D., Ruggieri, S., and Turini, F. (2008). Discrimination-aware data mining. In *KDD*, pages 560–568. ACM.

Peters, J., Bühlmann, P., and Meinshausen, N. (2015). Causal inference using invariant prediction: identification and confidence intervals.

Peters, J., Janzing, D., and Schölkopf, B. (2017). *Elements of Causal Inference: Foundations and Learning Algorithms*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA.

Plecko, D., Bennett, N., and Meinshausen, N. (2021). fairadapt: Causal reasoning for fair data pre-processing. *CoRR*, **abs/2110.10200**.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J. M., and Weinberger, K. Q. (2017). On fairness and calibration. In I. Guyon, U. von Luxburg, S. Bengio, H. M. Wallach, R. Fergus, S. V. N. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5680–5689.

Qureshi, B., Kamiran, F., Karim, A., Ruggieri, S., and Pedreschi, D. (2020). Causal inference for social discrimination reasoning. *J. Intell. Inf. Syst.*, **54**(2), 425–437.

Raff, E., Sylvester, J., and Mills, S. (2018). Fair forests: Regularized tree induction to minimize model bias. In J. Furman, G. E. Marchant, H. Price, and F. Rossi, editors, *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA, February 02-03, 2018*, pages 243–250. ACM.

Raghavan, M., Barocas, S., Kleinberg, J. M., and Levy, K. (2020). Mitigating bias in algorithmic hiring: evaluating claims and practices. In *FAT**, pages 469–481. ACM.

Roh, Y., Lee, K., Whang, S., and Suh, C. (2021). Sample selection for fair and robust training. In *NeurIPS*, pages 815–827.

Rojas-Carulla, M., Schölkopf, B., Turner, R. E., and Peters, J. (2018). Invariant models for causal transfer learning. *J. Mach. Learn. Res.*, **19**, 36:1–36:34.

Romei, A. and Ruggieri, S. (2014). A multidisciplinary survey on discrimination analysis. *Knowl. Eng. Rev.*, **29**(5), 582–638.

Rosenfeld, E., Ravikumar, P. K., and Risteski, A. (2021). The risks of invariant risk minimization. In *ICLR*. OpenReview.net.

Rubin, D. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, **66**.

Schölkopf, B. (2022). Causality for machine learning. In H. Geffner, R. Dechter, and J. Y. Halpern, editors, *Probabilistic and Causal Inference: The Works of Judea Pearl*, volume 36 of *ACM Books*, pages 765–804. ACM.

Schölkopf, B., Janzing, D., Peters, J., Sgouritsa, E., Zhang, K., and Mooij, J. M. (2012). On causal and anticausal learning. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*. icml.cc / Omnipress.

Selbst, A. D., danah boyd, Friedler, S. A., Venkatasubramanian, S., and Vertesi, J. (2019). Fairness and abstraction in sociotechnical systems. In danah boyd and J. H. Morgenstern, editors, *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* 2019, Atlanta, GA, USA, January 29-31, 2019*, pages 59–68. ACM.

Shamir, O. and Zhang, T. (2013). Stochastic gradient descent for non-smooth optimization: Convergence results and optimal averaging schemes. In *ICML (1)*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 71–79. JMLR.org.

Shanmugam, R. (2001). Causality: Models, reasoning, and inference : Judea pearl; cambridge university press, cambridge, uk, 2000, pp 384, ISBN 0-521-77362-8. *Neurocomputing*, **41**(1-4), 189–190.

Shui, C., Chen, Q., Li, J., Wang, B., and Gagné, C. (2022). Fair representation learning through implicit path alignment. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 20156–20175. PMLR.

Song, H., Kim, M., Park, D., Shin, Y., and Lee, J.-G. (2023). Learning from noisy labels with deep neural networks: A survey. *IEEE Transactions on Neural Networks and Learning Systems*, **34**(11), 8135–8153.

Song, J., Kalluri, P., Grover, A., Zhao, S., and Ermon, S. (2019). Learning controllable fair representations. In *AISTATS*, volume 89 of *Proceedings of Machine Learning Research*, pages 2164–2173. PMLR.

Spirtes, P. and Glymour, C. (1991). An Algorithm for Fast Recovery of Sparse Causal Graphs. *Social Science Computer Review*, **9**(1), 62–72.

Spirtes, P. and Zhang, K. (2016). Causal discovery and inference: Concepts and recent methodological advances. *Applied Informatics*, **3**(1), 3.

Spirtes, P., Meek, C., and Richardson, T. S. (2013). Causal inference in the presence of latent variables and selection bias. *CoRR*, **abs/1302.4983**.

Stanczuk, J., Etmann, C., Kreusser, L. M., and Schönlieb, C. (2021). Wasserstein gans work because they fail (to approximate the wasserstein distance). *CoRR*, **abs/2103.01678**.

Subbaswamy, A., Schulam, P., and Saria, S. (2019). Preventing failures due to dataset shift: Learning predictive models that transport. In K. Chaudhuri and M. Sugiyama, editors, *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pages 3118–3127. PMLR.

Tommasi, T., Patricia, N., Caputo, B., and Tuytelaars, T. (2017). A deeper look at dataset bias. In *Domain Adaptation in Computer Vision Applications*, Advances in Computer Vision and Pattern Recognition, pages 37–55. Springer.

Torralba, A. and Efros, A. A. (2011). Unbiased look at dataset bias. In *CVPR*, pages 1521–1528. IEEE Computer Society.

van den Broek, E., Sergeeva, A. V., and Huysman, M. (2019). Hiring algorithms: An ethnography of fairness in practice. In *ICIS*. Association for Information Systems.

Verma, S. and Rubin, J. (2018). Fairness definitions explained. In *FairWare@ICSE*, pages 1–7. ACM.

Voigt, P. and Bussche, A. (2017). *The EU General Data Protection Regulation (GDPR): A Practical Guide*. Springer International Publishing.

Vowels, M. J., Camgöz, N. C., and Bowden, R. (2023). D'ya like dags? A survey on structure learning and causal discovery. *ACM Comput. Surv.*, **55**(4), 82:1–82:36.

Wachter, S., Mittelstadt, B., and Russell, C. (2021). Bias preservation in machine learning: The legality of fairness metrics under eu non-discrimination law. *SSRN Electronic Journal*.

Wang, Y. and Jordan, M. I. (2021). Desiderata for representation learning: A causal perspective. *CoRR*, **abs/2109.03795**.

Watkins, E. A. and Chen, J. (2024). The four-fifths rule is not disparate impact: A woeful tale of epistemic trespassing in algorithmic fairness. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency, FAccT 2024, Rio de Janeiro, Brazil, June 3-6, 2024*, pages 764–775. ACM.

Weerts, H. J. P., Xenidis, R., Tarissan, F., Olsen, H. P., and Pechenizkiy, M. (2023). Algorithmic unfairness through the lens of EU non-discrimination law: Or why the law is not a decision tree. In *FAccT*, pages 805–816. ACM.

Woodward, J. F. (2003). *Making Things Happen: A Theory of Causal Explanation*. Oxford University Press, New York.

Woodworth, B. E., Gunasekar, S., Ohannessian, M. I., and Srebro, N. (2017). Learning non-discriminatory predictors. In S. Kale and O. Shamir, editors, *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, volume 65 of *Proceedings of Machine Learning Research*, pages 1920–1953. PMLR.

Wu, X., Kumar, V., Quinlan, J. R., Ghosh, J., Yang, Q., Motoda, H., McLachlan, G. J., Ng, A., Liu, B., Philip, S. Y., *et al.* (2008). Top 10 algorithms in data mining. *Knowledge and information systems*, **14**(1), 1–37.

Xie, C., Chen, F., Liu, Y., and Li, Z. (2020). Risk variance penalization: From distributional robustness to causality. *CoRR*, **abs/2006.07544**.

Xie, Q., Dai, Z., Du, Y., Hovy, E. H., and Neubig, G. (2017). Controllable invariance through adversarial feature learning. In *NIPS*, pages 585–596.

Xu, D., Yuan, S., Zhang, L., and Wu, X. (2018). Fairgan: Fairness-aware generative adversarial networks. In *IEEE BigData*, pages 570–575. IEEE.

Xu, D., Wu, Y., Yuan, S., Zhang, L., and Wu, X. (2019). Achieving causal fairness through generative adversarial networks. In *IJCAI*, pages 1452–1458. ijcai.org.

Xu, Y. and Jaakkola, T. S. (2021). Learning representations that support robust transfer of predictors. *CoRR*, **abs/2110.09940**.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. P. (2017a). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In R. Barrett, R. Cummings, E. Agichtein, and E. Gabrilovich, editors, *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, pages 1171–1180. ACM.

Zafar, M. B., Valera, I., Gomez-Rodriguez, M., and Gummadi, K. P. (2017b). Fairness constraints: Mechanisms for fair classification. In *AISTATS*, volume 54 of *Proceedings of Machine Learning Research*, pages 962–970. PMLR.

Zech, J., Badgeley, M., Liu, M., Costa, A., Titano, J., and Oermann, E. (2018). Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: A cross-sectional study. *PLOS Medicine*, **15**, e1002683.

Zehlike, M., Hacker, P., and Wiedemann, E. (2020). Matching code and law: achieving algorithmic fairness with optimal transport. *Data Min. Knowl. Discov.*, **34**(1), 163–200.

Zehlike, M., Yang, K., and Stoyanovich, J. (2023a). Fairness in ranking, part I: score-based ranking. *ACM Comput. Surv.*, **55**(6), 118:1–118:36.

Zehlike, M., Yang, K., and Stoyanovich, J. (2023b). Fairness in ranking, part II: learning-to-rank and recommender systems. *ACM Comput. Surv.*, **55**(6), 117:1–117:41.

Zemel, R. S., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning fair representations. In *ICML (3)*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 325–333. JMLR.org.

Zhang, B. H., Lemoine, B., and Mitchell, M. (2018a). Mitigating unwanted biases with adversarial learning. In J. Furman, G. E. Marchant, H. Price, and F. Rossi, editors, *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2018, New Orleans, LA, USA, February 02-03, 2018*, pages 335–340. ACM.

Zhang, B. H., Lemoine, B., and Mitchell, M. (2018b). Mitigating unwanted biases with adversarial learning. In *AIES*, pages 335–340. ACM.

Zhang, J. and Bareinboim, E. (2018). Fairness in decision-making - the causal explanation formula. In *AAAI*, pages 2037–2045. AAAI Press.

Zhang, L., Wu, Y., and Wu, X. (2017). A causal framework for discovering and removing direct and indirect discrimination. In *IJCAI*, pages 3929–3935. ijcai.org.

Zhang, L., Wu, Y., and Wu, X. (2018c). Achieving non-discrimination in prediction. In *IJCAI*, pages 3097–3103. ijcai.org.

Zhang, L., Wu, Y., and Wu, X. (2019). Causal modeling-based discrimination discovery and removal: Criteria, bounds, and algorithms. *IEEE Trans. Knowl. Data Eng.*, **31**(11), 2035–2050.

Zhao, X., Fabbrizzi, S., Lobo, P. R., Ghodsi, S., Broelemann, K., Staab, S., and Kasneci, G. (2023a). Adversarial reweighting guided by wasserstein distance for bias mitigation.

Zhao, X., Broelemann, K., Ruggieri, S., and Kasneci, G. (2023b). Causal fairness-guided dataset reweighting using neural networks. In *IEEE Big Data*, pages 1386–1394. IEEE.

Zhou, X., Pi, R., Zhang, W., Lin, Y., Chen, Z., and Zhang, T. (2022). Probabilistic bilevel coreset selection. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pages 27287–27302. PMLR.

Zliobaite, I., Kamiran, F., and Calders, T. (2011). Handling conditional discrimination. In *ICDM*, pages 992–1001. IEEE Computer Society.