Algorithmic Decision Making with Conditional Fairness

Renzhe Xu
xrz19@mails.tsinghua.edu.cn
Tsinghua University

Peng Cui
cuip@tsinghua.edu.cn
Tsinghua University

Kun Kuang
kunkuang@zju.edu.cn
Zhejiang University

Bo Li
libo@sem.tsinghua.edu.cn
Tsinghua University

Linjun Zhou
zhoulj16@mails.tsinghua.edu.cn
Tsinghua University

Zheyan Shen
shenzy17@mails.tsinghua.edu.cn
Tsinghua University

Wei Cui
cuiwei@songshuai.com
Squirrel AI Learning

ABSTRACT

Nowadays fairness issues have raised great concerns in decision-making systems. Various fairness notions have been proposed to measure the degree to which an algorithm is unfair. In practice, there frequently exist a certain set of variables we term as fair variables, which are pre-decision covariates such as users' choices. The effects of fair variables are irrelevant in assessing the fairness of the decision support algorithm. We thus define conditional fairness as a more sound fairness metric by conditioning on the fairness variables. Given different prior knowledge of fair variables, we demonstrate that traditional fairness notations, such as demographic parity and equalized odds, are special cases of our conditional fairness notations. Moreover, we propose a Derivable Conditional Fairness Regularizer (DCFR), which can be integrated into any decision-making model, to track the trade-off between precision and fairness of algorithmic decision making. Specifically, an adversarial representation based conditional independence loss is proposed in our DCFR to measure the degree of unfairness. With extensive experiments on three real-world datasets, we demonstrate the advantages of our conditional fairness notation and DCFR.

CCS CONCEPTS

• **Computing methodologies → Machine learning**.

KEYWORDS

fairness; conditional independence

ACM Reference Format:
Renzhe Xu, Peng Cui, Kun Kuang, Bo Li, Linjun Zhou, Zheyan Shen, and Wei Cui. 2020. Algorithmic Decision Making with Conditional Fairness. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '20), August 23–27, 2020, Virtual Event, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3394486.3403263

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than ACM must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
KDD '20, August 23–27, 2020, Virtual Event, CA, USA
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7998-4/20/08…$15.00
https://doi.org/10.1145/3394486.3403263

1 INTRODUCTION

Nowadays fairness issues have raised great concerns in decision-making systems such as loan applications[24], hiring processes[28], and criminal justice[22]. Poorly designed algorithms tend to amplify the bias existed in data, resulting in discriminations towards specific groups of individuals based on their inherent characteristics, which are often named as sensitive attributes in fairness problems. For example, race is a sensitive attribute for crime judgment. ProPublica[22] found it is unfair that African Americans were more likely to be incorrectly labeled as higher risk compared with Caucasians in the COMPAS system. However, what is fair and how to develop fair algorithms for algorithmic decision making are of paramount importance for both academic research and practical applications.

Recently, many works defined their fairness and proposed corresponding fair algorithms, from which, the definition of fairness can be divided into three types: individual fairness[11], group fairness[15, 20], and causality-base fairness notions[6, 19, 21, 25, 29, 34]. Individual fairness requires similar individuals should have similar outcomes. However, it is difficult to define the similarity between individuals. Group fairness requires equity among different groups but they only use sensitive attributes and outcomes as measuring features. As a result, these notions may fail to distinguish between fair and unfair parts in the problem. For example, Pearl [26] studied the case of Berkeley's alleged sex bias in graduate admission[4] and found that data showed a higher rate of admission for male applicants overall but the result is different when looking into the department choice. The bias caused by department choice should be considered fair but traditional group fairness notions fail to judge fairness since they do not take the department choice into account. Inspired by this, causality-based fairness notions arise. In these papers, the authors firstly assumed a causal graph between the features, and afterward, they could define the unfair causal effect from the sensitive attribute to the outcome as a metric. However, these fairness notions need very strong assumptions and they are not scalable.

In practice, there frequently exist a certain set of variables we term as fair variables, which are pre-decision covariates such as the department choice in Berkeley's graduate admission problem. The effects of fair variables are irrelevant in assessing the fairness of the decision support algorithm. We thus define conditional fairness

as a more sound fairness metric by conditioning on the fairness variables. In detail, outcome variables should be independent of sensitive attributes conditional on these fair variables.

The definition of conditional fairness has several advantages. Firstly, the fair variables can be any variables determined by the decision-makers or decision-making system inspectors, which provides a more flexible judging method. Secondly, conditional fairness can be viewed as a more general fairness notion since it can be easily reduced to demographic parity and equalized odds. If we believe that none of the features are fair variables, the conditional independence constraint becomes a normal independence constraint. If we choose fair variables as the outcome, the constraint is transformed into the target in equalized odds. Thirdly, the definition does not need strong assumptions in causality-based fairness definitions, which makes it much easier to be applied to real problems with a large amount of data and features.

The main challenge of this definition is to formulate the conditional independence constraint into a derivable loss function, which makes it impossible to be applied to commonly used gradient-based methods. Inspired by the conditional independence testing techniques in causality structure discovery literature[8, 32, 38], we formulate the conditional independence constraint into a computationally amenable form, which serves as a regularizer and is subsequently appended to the prediction loss function. We call the working regularizer as the Derivable Conditional Fairness Regularizer (DCFR).

As fair representation learning is a common in-processing framework to deal with fairness issues, we apply the regularizer and handle the conditional fairness constraint in this framework. With the regularizer, the target optimization problem can be cast as a minimax problem, which can be solved via adversarial learning. We further show that our method is also a general method that can be used in common group fairness notions (demographic parity and equalized odds). Most interestingly, when we believe none of the variables are fair, the target problem is reduced to demographic parity and our method also becomes the same as that proposed by Madras et al. [23] when dealing with the same problem.

Finally, we apply our method to real datasets and plot accuracy-fairness curves for three targets (demographic parity, equalized odds and, conditional fairness). We show that our method performs better on the conditional fairness task while having similar results with the state-of-the-art on the other two tasks.

In summary, our contributions are highlighted as follows:

- By exploring the fair variables, we propose conditional fairness, which is more general than previous definitions on fairness.
- We propose a novel Derivable Conditional Fairness Regularizer (DCFR) to optimize conditional fairness by learning conditional independent representation. Our DCFR can be easily integrated into decision-making models to improve fairness.
- We demonstrate the effectiveness of our proposed algorithm on fair decision making with three real-world datasets. Furthermore, DCFR performs especially better than baselines when increasing the number of values that fair variables can take.

2 RELATED WORKS

There are several common types of fairness notions including individual fairness, group fairness, and causality-based fairness. The most commonly used individual fairness notion is fairness through awareness[11] which requires that similar individuals should be treated similarly. However, it is difficult to define the similarity function between different individuals. Therefore, individual fairness still lacks further research up to today. Group fairness notions require the algorithm should treat different groups of individuals equally. The most commonly used group fairness notions include demographic parity[11], equal opportunity[15], equalized odds[15] and calibration[20]. These fairness notions are easy to understand and implement in real machine learning problems. However, they only use sensitive attributes and outcomes as measuring features. As a result, these notions may fail to distinguish between fair and unfair parts in the problem. To define fairness more elaborately, causality-based fairness notions are proposed recently. In these causality-based fairness notions such as counterfactual fairness[21], path-specific counterfactual fairness[6, 25, 34], the authors first define a causal graph among the features, and afterward, they can distinguish the unfair causal effect from the sensitive attribute to the outcome. However, these fair notions need very strong assumptions and they are not scalable.

Kamiran et al. [18] proposed the most similar fairness notions as us. In [18], the authors defined the variables as explanatory variables and proposed algorithms to mitigate the illegal discrimination they defined. However, their method is limited as it may do great harm to accuracy and it cannot be applied in practice as it cannot provide a tunable tradeoff between fairness and utility. Recently, Dutta et al. [10] also proposed an information-theoretic decomposition of the total discrimination.

Methods that mitigate biases in the algorithms fall under three categories: pre-processing[13, 17, 33], in-processing[16, 35], and post-processing[15] algorithms. Representation learning is a common in-processing method which is first proposed by Zemel et al. [36]. The authors try to mitigate individual unfairness and demographic discrimination simultaneously. Recently, learning representations via adversary has become the state-of-the-art method. Edwards and Storkey [12] first proposed this kind of method and they provided a framework to mitigate demographic discrimination. Several works followed this framework such as [1, 3, 23, 37, 39]. In particular, Madras et al. [23] proposed to use different adversarial loss function when faced with different fair notions. Zhao et al. [39] redesigned the loss functions to mitigate the gap of demographic parity and equalized odds simultaneously, which is proved to be difficult in [20]. However, these works all focus on the most commonly used group fairness notions. Therefore they cannot be applied to the general conditional fairness target. Agarwal et al. [2] proposed a general method to mitigate any fairness notions that can be written as linear inequalities on conditional moments. But they still require the categorical fair variables which makes it difficult to be extended to more general form.

Conditional independence tests have been popularly used in causal structure discovery problems[31]. In order to deal with more flexible distributions, several novel conditional independence tests have been proposed[14, 27, 30, 32]. However, these methods cannot

be mixed with gradient-based machine learning algorithms, since they usually calculate a statistic first and estimate a p-value with random methods. Our method is based on an equivalent relation of conditional independence[8] and is tractable in common machine learning algorithms.

3 PRELIMINARY

3.1 Notations

We suppose the dataset consists of a tuple $D = (S, X, Y)$, where S represents sensitive attributes such as gender and race, X represents features, and Y represents the outcome. Furthermore, we divide features X into two parts $X = (F, O)$, where F represents fair variables and O represents other features. We use m_X, m_F, m_O to denote the dimension of the features and we have $m_X = m_F + m_O$. We use calligraphic fonts to represent the range of corresponding random variables. For example $\mathcal{X}$ represents the space of X and $\mathcal{X} \subset \mathbb{R}^{m_X}$. Similarly, we have $\mathcal{F} \subset \mathbb{R}^{m_F}$. To simplify, we suppose the sensitive attribute and the outcome are binary, which means $\mathcal{Y}, \mathcal{S} = \{0, 1\}$. We set $S = 1$ as the privileged group and $Y = 1$ as the favored outcome.

We suppose there are N samples in total and we use S_i, X_i, Y_i, F_i, O_i to represent the features of i-th sample. In addition, for a condition E, we use $D(E)$ to represent the samples that satisfy the condition and $|D(E)|$ to represent the number of these samples. For example, $D(Y = 1)$ means the samples that satisfy $Y_i = 1$ and $|D(Y = 1)|$ is the total number of such samples.

A fair machine learning problem is to design a fair predictor $\hat{Y}$ with parameters $\boldsymbol{\theta} : \mathcal{X} \times \mathcal{S} \to \mathcal{Y}$, which maximizes the likelihood $P(Y, X, S|\boldsymbol{\theta})$ while satisfying some specific fair constraints, which we will introduce in the next section.

3.2 Fairness Notions

We first introduce some well-known fair notions in machine learning problems.

Definition 3.1 (demographic parity (DP)). Given the joint distribution D, the classifier $\hat{Y}$ satisfies demographic parity with respect to sensitive attribute S if $\hat{Y}$ is independent of S, i.e.

$$\hat{Y} \perp S. \tag{1}$$

The definition of DP is clear and concise, representing that S has no predictive power to $\hat{Y}$, but in practice we are also interested in some evaluation metric to reveal how fair the system is. Thus the following equivalent form ΔDP is proposed to measure the degree of fairness.

$$\Delta DP \triangleq |P(\hat{Y} = 1|S = 1) - P(\hat{Y} = 1|S = 0)|. \tag{2}$$

Easy to show that $\hat{Y} \perp S$ if and only if $\Delta DP = 0$.

One of the drawbacks of ΔDP is that when the base rate differs significantly among two groups, *i.e.*, $P(Y = 1|S = 0) \neq P(Y = 1|S = 1)$, the utility could be limited. Hardt et al. [15] further proposed another notion Equalized Odds to avoid this problem.

Definition 3.2 (Equalized odds (EO)). Given the joint distribution D, the classifier $\hat{Y}$ satisfies equalized odds with respect to sensitive attribute S if $\hat{Y}$ is independent of S conditional on Y, i.e.

$$\hat{Y} \perp S \mid Y. \tag{3}$$

Similarly, the metric ΔEO is defined as the expectation of the absolute difference of true positive rate and false positive rate across two groups.

$$\begin{aligned}\Delta EO &\triangleq \mathbb{E}_y \left[\left|P(\hat{Y} = 1|S = 1, Y = y) - P(\hat{Y} = 1|S = 0, Y = y)\right|\right] \\ &= P(Y = 0)\left|P(\hat{Y} = 1|S = 1, Y = 0) - P(\hat{Y} = 1|S = 0, Y = 0)\right| \\ &+ P(Y = 1)\left|P(\hat{Y} = 1|S = 1, Y = 1) - P(\hat{Y} = 1|S = 0, Y = 1)\right|.\end{aligned} \tag{4}$$

It is also easy to show that $\hat{Y} \perp S \mid Y$ if and only if $\Delta EO = 0$.

None of the fairness notions above take fair variables into account, inspired by Kamiran et al. [18] and Corbett-Davies et al. [7], we denote conditional fairness as

Definition 3.3 (Conditional fairness (CF)). Given the joint distribution D, the classifier $\hat{Y}$ satisfies conditional fairness with respect to sensitive attribute S and fair variables F if $\hat{Y}$ is independent of S conditional on F, i.e.

$$\hat{Y} \perp S \mid F. \tag{5}$$

In addition, similar to ΔEO, we define a metric ΔCF as:

$$\Delta CF \triangleq \mathbb{E}_f \left[\left|P(\hat{Y} = 1|S = 1, F = f) - P(\hat{Y} = 1|S = 0, F = f)\right|\right] \tag{6}$$

Specifically, when fair variables are continuous, Equation 6 becomes:

$$\Delta CF = \int_{f \in \mathcal{F}} \left|P(\hat{Y} = 1|S = 1, F = f) - P(\hat{Y} = 1|S = 0, F = f)\right| \mathrm{d}\mathbb{P}(f). \tag{7}$$

and when fair variables are categorical, ΔCF becomes

$$\Delta CF = \sum_{f \in \mathcal{F}} \left|P(\hat{Y} = 1|S = 1, F = f) - P(\hat{Y} = 1|S = 0, F = f)\right| P(F = f). \tag{8}$$

ΔCF aims to calculate the mean of the absolute difference between two groups among all potential values of the fair variables. Similarly, we have $\hat{Y} \perp S \mid F$ if and only if $\Delta CF = 0$.

Compare CF with DP and EO. On the one hand, conditional fairness can take more complex situations into account. On the other hand, conditional fairness is more general and it can be easily reduced to DP and EO.

Consider the data-generating graph for a toy example of a college admission case in figure 1. Because qualification requirements usually differ among various departments, it is fair to determine outcomes according to department choices and qualifications. Hence any predictors with form $\hat{Y} = f(Q, D)$ can be considered fair in practice. It is easy to show that when setting department choice D as the fair variable, $\hat{Y}$ is conditional fair. However, DP and EO fail to judge the fairness of $\hat{Y}$ as equation (1) and (3) may not be satisfied.

In addition, conditional fairness is a more flexible fairness notion as:

- If we believe none of the features X is fair, which means $F = \emptyset$, the conditional independence target is reduced to the independence condition as shown in equation (1) and conditional fairness is reduced to DP.

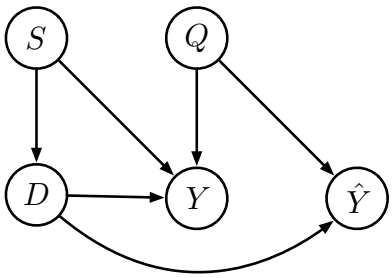

Figure 1: The data-generating graph for a toy example of a college admission case. S, D, Q, Y represent gender, department choice, qualification, and historical admission decision, respectively. $\hat{Y}$ represents a conditional fair decision-making system.

- If we set F as Y, the conditional independence target is reduced to the conditional independence as shown in equation (3) and conditional fairness is reduced to EO.

Compare CF with causality-based fairness notions. Generally speaking, conditional fairness requires much fewer assumptions than causality-based fairness notions, which makes CF practical in real problems.

Under some circumstances, a conditional fair decision-making system can satisfy causality-based fairness notions. Consider path-specific fairness[6] in the example shown in figure 1. The directed path $S \to Y$ can be viewed as an unfair path while $S \to D \to Y$ and $Q \to Y$ are fair paths. Hence, the historical decisions Y is not path-specific fair for the existence of unfair path $S \to Y$. However, the conditional fair decision-making system $\hat{Y} = f(Q, D)$ successfully satisfies the requirement as the unfair path $S \to \hat{Y}$ does not exist. As for deeper connections between conditional fairness and causality-based fairness notions, we remain as future works.

3.3 Problem Formulation

Next we will apply our definition of conditional fairness into real fair problems. In general, the goal of a fairness problem is to achieve a balance between fairness and algorithm performance. Formally, we need to design a loss function on prediction $L_{\text{pred}}(\hat{Y}, Y)$ and another loss function on fairness $L_{\text{fair}}(\hat{Y}, S, F)$. The optimization goal of a fairness problem can be formulated as:

$$\boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta}} L(\hat{Y}) = \arg\min_{\boldsymbol{\theta}} L_{\text{pred}}(\hat{Y}, Y) + \lambda \cdot L_{\text{fair}}(\hat{Y}, S, F), \quad (9)$$

where the hyper-parameter λ provides a trade-off between fairness and performance. When λ is large, the target tends to make L_{fair} small which can ensure fairness while doing harm to performance, and the result is opposite when λ is small.

As for the prediction loss, any form of traditional loss functions are suitable such as cross-entropy or L1 loss. While the fairness loss targeted for conditional fairness is difficult to design relatively. When fair variables are categorical, we can use the ΔCF metric as a loss function. However, in practice, the fair variables may contain many different values or they may be continuous. Under this circumstance, the metric can no longer be a suitable loss function for optimization. Inspired by this issue, we will propose a new derivable loss function that can deal with these situations in the next section.

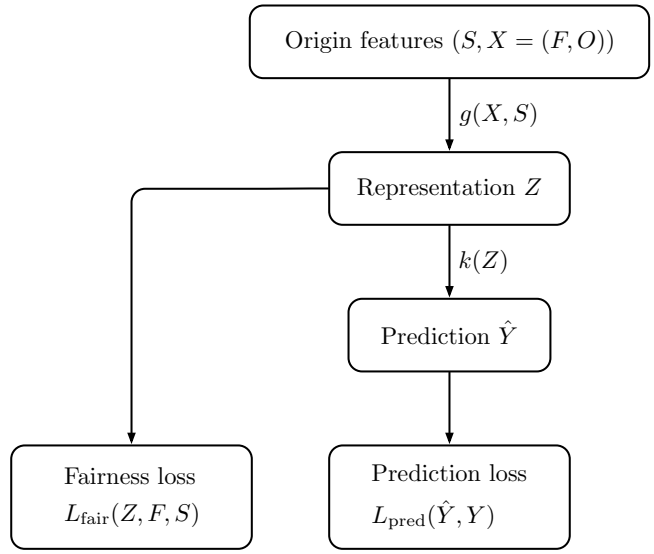

Figure 2: The framework of our method. The variables include sensitive attribute S, features X, outcome Y, representation Z, and prediction $\hat{Y}$. X is divided into fair variables F and other variables O. The function g maps the original features into the representation space, and function k maps the representations into the outcome space. There are two loss functions that measure utility and fairness respectively.

4 PROPOSED METHOD

An intuitive way to deal with this conditional independence is to divide the whole training samples into different groups with respect to the value of fair variables and then deal with these groups separately using traditional methods handling naive independence problems. The main drawback is that this method assumes that fair variables are categorical and $|\mathcal{F}|$ is small. Meanwhile, when $|\mathcal{F}|$ becomes very large, representing that fair variables can take many different values, this naive method requires exactly $|\mathcal{F}|$ different models to deal with different subgroups, which has the potential of overfitting due to lack of training data in each subgroup. Furthermore, when fair variables are continuous, it becomes impossible to group by fair variables directly. Hence we need a more general framework to ensure the model's scalability.

Our solution to this problem is to learn a latent representation Z, which satisfies condition (5). Suppose the representation has m_Z dimensions, $g : \mathbb{R}^{m_X} \times \{0, 1\} \to \mathbb{R}^{m_Z}$ is the function from the space of X and S to representation space. The prediction function $k : \mathbb{R}^{m_Z} \to [0, 1]$ yields the probability of the sample in the positive class. The framework of our model is shown in figure 2. We now rewrite the equation (9) under this representation learning framework as:

$$\boldsymbol{\theta} = \arg\min_{\boldsymbol{\theta}} L_{\text{pred}}(k(g(X, S)), Y) + \lambda \cdot L_{\text{fair}}(g(X, S), F, S). \quad (10)$$

Table 1: Important Math Notations

Notation	Explanation
S	Sensitive attributes
$X = (F, O)$	Features (fair Variables, others)
Z	Latent features
$Y, \hat{Y}$	True outcome, predicted outcome
$h(Z, F)$	Adversary function
$Q(h)$	Adversary loss w.r.t. h
L^2_{ZF}	Function space of $h(Z, F)$, see Eqn. (13)
$\mathcal{E}_{ZF}$	Function space of $\tilde{h}(Z, F)$, see Eqn. (14)
H_{ZF}	Function space of $h(Z, F)$, see Eqn. (19)

4.1 Conditional Independence

In this section, we first introduce a conditional independence theorem proposed by Daudin [8]. Afterward we will transform it into the form that can be applied to fairness problems. Finally we will give a regularizer to measure conditional independence.

Lemma 4.1 (Characterization of conditional independence [8]). *The random variables Z, S are independent conditional on F ($Z \perp S \mid F$) if and only if, for any function $u \in L^2_S$, $\tilde{h} \in \mathcal{E}_{ZF}$,*

$$\mathbb{E}[u(S) \cdot \tilde{h}(Z, F)] = 0, \tag{11}$$

where

$$L^2_S = \left\{u(S) \mid \mathbb{E}[u^2] < \infty\right\}, \tag{12}$$

$$L^2_{ZF} = \left\{h(Z, F) \mid \mathbb{E}[h^2] < \infty\right\}, \tag{13}$$

$$\mathcal{E}_{ZF} = \left\{\tilde{h}(Z, F) \in L^2_{ZF} \mid \mathbb{E}[\tilde{h}|F] = 0\right\}. \tag{14}$$

Lemma 4.1 is designed for general cases and can be simplified in fairness issues. Considering cases of one single binary sensitive attribute, which means S is binary, the condition can be transformed into the following form.

Proposition 4.2. *If random variable S is binary and $S \in \{0, 1\}$, the random variables Z, S are independent conditional on F ($Z \perp S \mid F$) if and only if, for any $\tilde{h} \in \mathcal{E}_{ZF}$,*

$$\mathbb{E}[\mathbb{I}(S = 1) \cdot \tilde{h}(Z, F)] = 0, \tag{15}$$

where $\mathcal{E}_{ZF}$ is shown in equation (14) and $\mathbb{I}(S = 1)$ is the indicative function defined as follow:

$$\mathbb{I}(S = 1) = \begin{cases} 1, & \text{if} \quad S = 1, \\ 0, & \text{if} \quad S = 0. \end{cases} \tag{16}$$

However, proposition 4.2 can hardly be applied to practice directly because of the complexity of function space $\mathcal{E}_{ZF}$ shown in equation (14). Therefore, we need to transform equation (15) to another form which depends on a simpler function space.

Theorem 4.3 (Characterization with binary variable). *If random variable S is binary and $S \in \{0, 1\}$, the random variables Z, S are independent conditional on F ($Z \perp S \mid F$) if and only if, for any $h \in L^2_{ZF}$,*

$$\begin{aligned} Q(h) \triangleq\ & \mathbb{E}\left[\mathbb{I}(S = 1)P(S = 0|F)h(Z, F)\right] \\ & - \mathbb{E}\left[\mathbb{I}(S = 0)P(S = 1|F)h(Z, F)\right] = 0. \end{aligned} \tag{17}$$

Compared with function space $\mathcal{E}_{ZF}$ in proposition 4.2, L_{ZF} space in theorem 4.3 is much simpler. Based on this theorem, we propose the following regularizer:

Definition 4.4 (Derivable Conditional Fairness Regularizer). Z, F, S are random variables. S is binary and $S \in \{0, 1\}$. $Q(h)$ is defined in equation (17). Define the regularizer $L_{\text{fair}}(Z, F, S)$

$$L_{\text{fair}}(Z, F, S) \triangleq \sup_{h \in H_{ZF}} |Q(h)|, \tag{18}$$

where

$$H_{ZF} = \left\{h \in L^2_{ZF} | 0 \le h(Z, F) \le 1\right\}. \tag{19}$$

The motivations of regularizer are that, firstly, notice that if there exists a function $h \in L^2_{ZF}$ so that $Q(h) \neq 0$, then $\sup_{h \in L^2_{ZF}} |Q(h)|$ can be arbitrarily large. And the L^2_{ZF} space is too large for further analysis. Therefore we first bound the range of h into $[0, 1]$, which produces the H_{ZF} space. Secondly, when $L_{\text{fair}}(Z, F, S) = 0$, according to theorem 4.3, the random variables Z and S are independent conditional on F.

Furthermore, we can simplify equation (18) by the following theorem.

Theorem 4.5. *$L_{fair}(Z, F, S)$, H_{ZF}, and $Q(h)$ are defined in theorem 4.3 and definition 4.4. Then*

$$L_{fair}(Z, F, S) = \sup_{h \in H_{ZF}} |Q(h)| = \sup_{h \in H_{ZF}} Q(h). \tag{20}$$

Theorem 4.5 provides a computationally amenable form, which serves as a regularizer and is applied to the prediction loss function.

To better understand $Q(h)$, we transform the equation (17) as a weighted L1 loss when using $h(Z, F)$ to predict S.

$$\begin{aligned} & Q(h) \\ =& \mathbb{E}[\mathbb{I}(S = 1)P(S = 0|F)h(Z, F)] - \mathbb{E}[\mathbb{I}(S = 0)P(S = 1|F)h(Z, F)] \\ =& C - \left[\mathbb{E}[\mathbb{I}(S = 1)P(S = 0|F)(1 - h)] + \mathbb{E}[\mathbb{I}(S = 0)P(S = 1|F)h]\right], \end{aligned} \tag{21}$$

where

$$C = \mathbb{E}[\mathbb{I}(S = 1)P(S = 0|F)] \tag{22}$$

is a constant.

The term in the bracket is a weighted L1 loss and conditional probability $P(S = \cdot|F)$ is the weight. Actually, this weight can be learned from finite samples with any non-parametric or parametric algorithms such as regression methods. Therefore our model can be applied to real large datasets with continuous fair variables.

In addition, we give the formulation of $Q(h)$ when fair variables are categorical in order to apply it to simpler circumstances, such as demographic parity or equalized odds. In detail, we can use $|D(S = 1 - S_i, F = F_i)|/|D(F = F_i)|$ to estimate the weight for the i-th sample, therefore the equation (21) can be written as

$$Q(h) \approx C - \frac{1}{N}\left[\sum_{i=1}^{N} \frac{|D(S = 1 - S_i, F = F_i)|}{|D(F = F_i)|} |h(F_i, Z_i) - S_i|\right]. \tag{23}$$

4.2 Adversarial Learning

Now we combine the total loss function as shown in equation (10) and the conditional independence in equation (20) and get:

$$\begin{aligned}\boldsymbol{\theta} =& \arg\min_{g,k} L_{\text{pred}}(k(g(X,S)),Y) + \lambda \cdot \sup_{h} Q(h) \\ =& \arg\min_{g,k} \sup_{h} L_{\text{pred}}(k(g(X,S)),Y) + \lambda \cdot Q(h).\end{aligned} \tag{24}$$

As $Q(h)$ is actually a weighted L1 loss, the loss function above can be optimized with the method of adversarial learning by setting the $Q(h)$ as the adversarial loss. There are several works that use adversarial learning to solve fairness notions. While the frameworks among these works are similar, the main difference lies in the design of loss functions.

Our method is most closed to LAFTR[23]. And actually, when $F = \emptyset$, which means the conditional independence constraint $\hat{Y} \perp S \mid F$ is reduced to $\hat{Y} \perp S$, our method is exactly the same as theirs.

Consider $Q(h)$ with finite samples as shown in equation (23), when $F = \emptyset$, the weight of sample i is actually $|D(S = 1 - S_i)|/N$. Multiple equation (23) with $a = N^2/(|D(S = 0)| \cdot |D(S = 1)|)$ and we get

$$Q_{\text{DP}}(h) \triangleq a \cdot Q(h) \approx C' - \sum_{i=1}^{N} \frac{1}{|D(S = S_i)|} |h(Z_i) - S_i|, \tag{25}$$

which becomes the same as the adversarial loss function provided by [23].

When facing with equalized odds task, we can replace the fair variables F in equation (23) with Y and get:

$$Q_{\text{EO}}(h) \triangleq C - \frac{1}{N}\left[\sum_{i=1}^{N} \frac{|D(S = 1 - S_i, Y = Y_i)|}{|D(Y = Y_i)|} |h(Y_i, Z_i) - S_i|\right]. \tag{26}$$

With equation (25) and (26), we can apply our method into demographic parity and equalized odds target.

4.3 Practical Implementation

In practice, we cannot enumerate all the functions in H_{ZF}, we use an MLP with sigmoid as an estimation. Furthermore, we find it difficult to optimize with L1 loss as we use sigmoid functions to bound the h into $[0, 1]$ and this can result in vanishing gradient problem in practice. Instead, We define the L2 loss function $Q'(h)$ as the surrogate of $Q(h)$ and the corresponding conditional independence regularizer L'_{fair}.

$$\begin{aligned}Q'(h) \triangleq C - \Big[&\mathbb{E}\left[\mathbb{I}(S = 1)P(S = 0|F) \cdot (1 - h)^2\right] \\ &+ \mathbb{E}\left[\mathbb{I}(S = 0)P(S = 1|F) \cdot h^2\right]\Big],\end{aligned} \tag{27}$$

where C is the constant defined in equation (22). The corresponding regularizer is

$$L'_{\text{fair}} \triangleq \sup_{h \in H_{ZF}} Q'(h). \tag{28}$$

To show that directly optimizing L2 loss could also reach our target, we give the following theorem.

Theorem 4.6. *L'_{fair} provides an upper bound of L_{fair}, i.e.*

$$L'_{fair} \geq L_{fair}. \tag{29}$$

Algorithm 1 *Derivable Conditional Fairness Regularizer (DCFR)*

Input: Dataset $D = (X, Y, S)$, $X = (F, O)$, EPOCH, BATCH_SIZE, ADV_STEPS.
Output: g, k, h as in equation (24)

```
/* Step I */
for epoch_i ← 1 to EPOCH do
    Random mini-batch D' = (X' = (F', O'), Y', S') from D.
    Freeze h. Unfreeze g, k.
    Optimize g, k with gradient descent according to D'.
    Freeze g, k. Unfreeze h.
    for adv_step ← 1 to ADV_STEPS do
        Optimize h with gradient descent according to D'.
    end for
end for

/* Step II */
Freeze g, h. Unfreeze k.
for epoch_i ← 1 to EPOCH do
    Random mini-batch D' = (X' = (F', O'), Y', S') from D.
    Optimize k with gradient descent according to D'.
    if accuracy on validation set does not increase for continuous
    20 epochs then
        Break.
    end if
end for
return g, k, h.
```

With this theorem, it makes sense to directly optimize with L2 loss, as L1 loss will decrease synchronously with L2 loss during the optimization process. Using L2 loss instead of L1 loss makes the algorithm much easier to converge in real experiments.

Our algorithm has two steps. We train the model g, h, k adversarially firstly, and afterward we fine-tune the function k for better performance. During the training step, for each sampled mini-batch, we train predictor part g and k once and train adversarial part h for several times. The number of adversarial steps is also a hyper-parameter. During the fine-tuning step, we run the models with early stop when the accuracy on the validation set does not increase for continuous 20 epochs. The pseudo-code is shown in algorithm 1.

5 EXPERIMENTS

In this section, we provide the experimental settings and verify the effectiveness of our method in multiple real datasets.

5.1 Datasets

We perform experiments on three real-world datasets that are widely used in fair machine learning problems, including the Adult dataset[9], the Dutch census dataset[5], and the COMPAS dataset[22].

- **Adult**: The goal of the Adult dataset is to predict whether a person makes more than $50k per year or not. Each instance contains 112 attributes including sex, gender, education level, occupation, etc. In our experiments, we set gender as the sensitive attribute, and consider occupation (with 14 possible categorical values) as the fair variable. The target variable

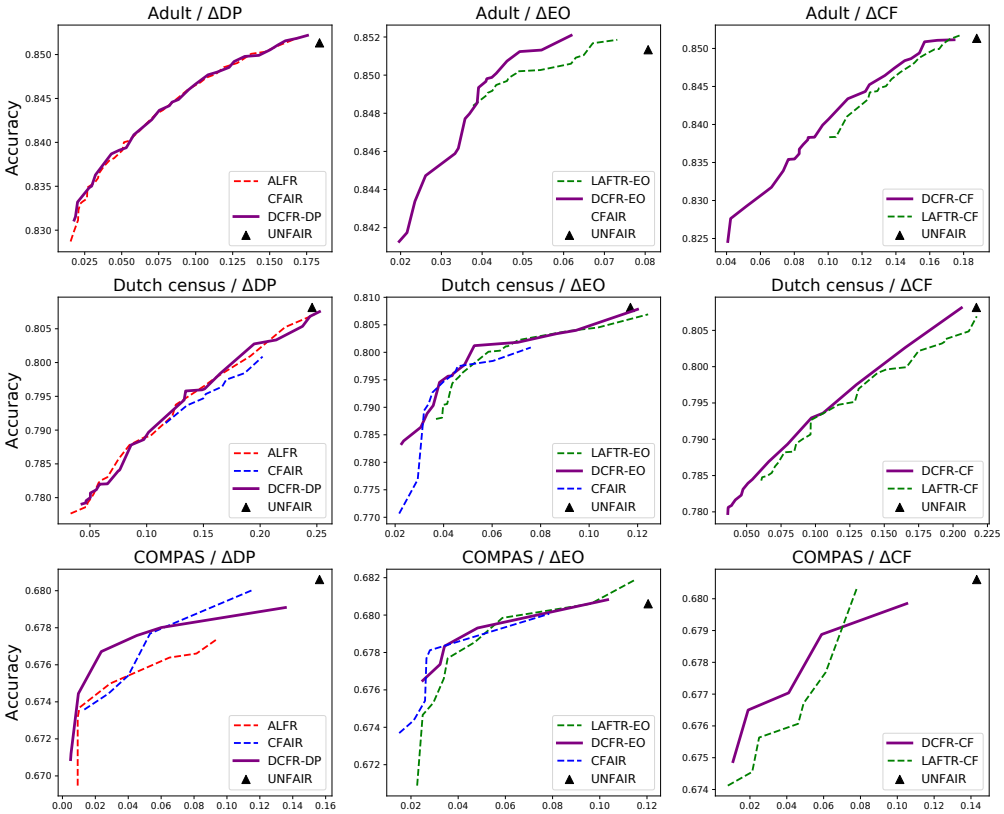

Figure 3: The accuracy-fairness trade-off curves for different fairness metrics (ΔDP, ΔEO, ΔCF from left to right) on various datasets (Adult, Dutch census, COMPAS dataset from top to bottom, with $|\mathcal{F}| = 14, 7, 2$ respectively). The upper-left corner is preferred. Our method is shown in bolded lines. The UNFAIR algorithm is a triangle mark while other baselines are in dashed lines. We take different values of λ from 0.1 to 20, get the mean of accuracy and fairness metric across 5 runs for each model, and plot the Pareto front on the test dataset. While our model performs similarly on ΔDP and ΔEO task with baselines, with the increase of $|\mathcal{F}|$, our method performs much better than baselines on ΔCF task. Note that we do not plot the curve of CFAIR in the Adult dataset because the curve goes beyond the axis range.

(income) is binary and we set "≥ $50k per year" as the favored outcome.

- **Dutch census**: This dataset is sampled from the Dutch census dataset, which is conducted by Statistics Netherlands to predict whether a person has a prestigious occupation. Each instance contains 35 attributes including age, gender, marital status, etc. In our experiments, we set gender as the sensitive attribute and level of educational attainment (with 7 possible categorical values) as the fair variable. The target variable is binary and we set "having a prestigious occupation" as the favored outcome.
- **COMPAS**: The COMPAS dataset aims to predict whether a criminal defendant will recidivate within two years or not. Each instance contains 11 attributes including age, race, gender, number of prior crimes, etc. In our experiments, we set race as the sensitive attribute and set the charge degree (with 2 possible categorical values) as the fair variable. The target variable (recidivism or not) is binary and we define "not recidivism" as the favored outcome.

As a summary, the basic statistics of the datasets are listed in table 2.

Table 2: Basic statistics of the datasets

Dataset	Train/Test	$P(S=0)$	$P(Y=1)$	$\lvert\mathcal{F}\rvert$
Adult	30,162/15,060	0.325	0.248	14
Dutch census	27,060/11,595	0.492	0.521	7
COMPAS	4,321/1,851	0.659	0.455	2

5.2 Baselines

As adversarial representation learning has become a prominent solver for fairness-related constrained optimization problems, we also adopt it to solve our target problem. For a fair comparison, we mainly select the following state-of-the-art fairness optimization algorithms that are also solved by adversarial representation learning as baseline methods.

- **UNFAIR**: We design a baseline predictive model without any fairness constraint by setting λ to be 0 in equation (24).
- **ALFR**[12]: ALFR is specifically designed for demographic parity problems.
- **CFAIR**[39]: CFAIR aims to mitigate the gap of demographic parity and equalized odds simultaneously.
- **LAFTR**[23]: LAFTR consists of two different loss functions, which target demographic parity and equalized odds respectively. Therefore, we implement two variants **LAFTR-DP** and **LAFTR-EO**.

For general conditional fairness, since none of the methods above propose the method to handle this situation explicitly, we extend the method of LAFTR-EO by replacing the conditional target Y as F in adversarial loss, namely **LAFTR-CF**. In detail, the original target adversarial loss function in LAFTR is

$$L_{Adv}^{EO}(h) = 2 - \sum_{i=1}^{N} \frac{1}{|D(Y = Y_i, S = S_i)|} |h(Z_i) - S_i|. \tag{30}$$

We transform it into conditional fairness setting as

$$L_{Adv}^{CF}(h) = |\mathcal{F}| - \sum_{i=1}^{N} \frac{1}{|D(F = F_i, S = S_i)|} |h(Z_i) - S_i|. \tag{31}$$

The differences between the equation above and our methods lie on the input of function h and the sample weight. Equation (31) tends to assign relatively high weights to minority groups with the same F compared with majority groups, which may lose stability while our method treats different groups divided by F equally.

As conditional fairness is a general notion that encompasses the demographic parity and equalized odds, we implement the following three variants of our method:

- **DCFR-DP**: We transform our method by setting $\mathcal{F}$ to be a null set and optimize it by equation (25), so that it can be used directly to solve demographic parity problems.
- **DCFR-EO**: We transform our method by setting $\mathcal{F}$ to be $\mathcal{Y} = \{0, 1\}$ and optimize it by equation (26), so that it can be used directly to solve equalized odds problems.
- **DCFR-CF**: We use the general form of our method to solve general conditional fairness problems.

For demographic parity, we compare our DCFR-DP with ALFR, CFAIR, and UNFAIR. For equalized odds, we compare our DCFR-EO with LAFTR-EO, CFAIR, and UNFAIR, while ALFR cannot handle this fairness target. For conditional fairness, we mainly compare our DCFR-CF with LAFTR-CF. Note that CFAIR method can hardly be applied to conditional fairness target as it requires $|\mathcal{F}|$ different adversarial predictors when calculating adversarial loss, which is impractical in real problems. For the sake of fair comparison and easier convergence, we replace L1 loss with L2 loss function for adversary losses in LAFTR model and our model DCFR. In addition, we use cross-entropy loss as the prediction loss function in our model DCFR.

As fair variables are categorical in these experiments, we use ΔDP, ΔEO, ΔCF as evaluation metrics, and smaller values of these metrics mean higher fairness. More experimental details are shown in the appendix.

5.3 Results

The results are shown in Figure 3. The columns show the accuracy-fairness trade-off curves for demographic parity, equalized odds, and conditional fairness respectively, and the rows correspond to different datasets.

For the tradeoff curves, there are two observation points.

(1) If a curve is closer to the left-top point than other curves in the majority range of an evaluation metric, the corresponding method is better. Because it means that given a certain degree of fairness, the method can achieve higher prediction accuracy, while given a certain prediction accuracy, the method can achieve better fairness.
(2) As a fair algorithm, it is important to evaluate how much fairness can be achieved, which is indicated by the left-end point of a curve.

From figure 3, we can get the following observations.

- For the conditional fairness task, it is obvious that our DCFR-CF is more advantageous than LAFTR-CF in the sense of both two observation points. In the COMPAS dataset, both methods can reach similar fairness ranges, while our DCFR-CF can achieve better trade-off performance. In Adult and Dutch census datasets, the two curves are close, while DCFR-CF can reach a higher fairness region than LAFTR-CF, which is more obvious in the Adult dataset. The plausible reason is that a larger $|\mathcal{F}|$ in Adult will make the limitation of LAFTR-CF more obvious as it is designed in the context of one single binary conditional variable.
- For the demographic parity and equalized odds tasks, the degenerated variants of our method produce comparable performances with state-of-the-art baselines that are specifically designed for these tasks. In some datasets, our method reports even better results, for example, in the Adult dataset of ΔEO setting and COMPAS dataset in ΔDP setting. We attribute this to the strong expressive ability of DCFR.
- The overall performance of CFAIR is not satisfactory, especially in the Adult dataset where the curve of CFAIR goes beyond the axis range. We notice that, as shown in table 2, Y is seriously biased in the Adult dataset. In CFAIR, however, the balanced error rate is used in optimization.

6 CONCLUSIONS

In this work, we propose the conditional fairness concerning fair variables and show that it is a general fairness notion with several practical reductions. However, conditional fairness is difficult to optimize directly as it cannot be written as a derivable loss function straightforwardly especially when fair variables are continuous or contain many categorical values. Inspired by conditional independence test methods, we derive an equivalent condition of conditional independence under fairness settings. Based on the equivalent condition, we propose a conditional independence regularizer that can be integrated into gradient-based methods, namely Derivable Conditional Fairness Regularizer (DCFR). We apply the regularizer into the representation learning framework and solve it with adversarial learning. We validate the effectiveness of our method on real datasets and achieve good performance on conditional fairness targets. It is worth mentioning that our method

becomes much better than baselines when the number of potential values of fair variable increases.

Potential future work is to apply our method into unsupervised settings as the conditional fairness notion does not rely on Y under most circumstances. With the information of F, we can ideally get a more elaborate representation compared with demographic parity. Besides, we find it difficult to measure the performance between different models. On the one hand, the target fairness notions of various models are usually different, which makes it impossible to compare with each other. On the other hand, even for the same fairness target, the most common practice is to plot the fairness-utility trade-off curve, which cannot become an accurate metric. This issue remains open and we believe it is worthwhile to study on.

ACKNOWLEDGMENTS

This work was supported in part by National Key R&D Program of China (No. 2018AAA0102004, No. 2018AAA0101900), National Natural Science Foundation of China (No. U1936219, 61772304, U1611461, 71490723, 71432004), Beijing Academy of Artificial Intelligence (BAAI), the Fundamental Research Funds for the Central Universities, Tsinghua University Initiative Scientific Research Grant (No. 2019THZWJC11), Science Foundation of Ministry of Education of China (No. 16JJD630006), and a grant from the Institute for Guo Qiang, Tsinghua University. All opinions, findings, conclusions and recommendations in this paper are those of the authors and do not necessarily reflect the views of the funding agencies.

REFERENCES

[1] Tameem Adel, Isabel Valera, Zoubin Ghahramani, and Adrian Weller. 2019. One-network adversarial fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 2412–2420.

[2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. 2018. A Reductions Approach to Fair Classification. In *International Conference on Machine Learning*. 60–69.

[3] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed H Chi. 2017. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075* (2017).

[4] Peter J Bickel, Eugene A Hammel, and J William O'Connell. 1975. Sex bias in graduate admissions: Data from Berkeley. *Science* 187, 4175 (1975), 398–404.

[5] Minnesota Population Center. 2019. Integrated public use microdata series: Version 7.2 [dataset]. *Minneapolis, MN: IPUMS* (2019).

[6] Silvia Chiappa. 2019. Path-specific counterfactual fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 33. 7801–7808.

[7] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. 797–806.

[8] JJ Daudin. 1980. Partial association measures and an application to qualitative regression. *Biometrika* 67, 3 (1980), 581–590.

[9] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[10] Sanghamitra Dutta, Praveen Venkatesh, Piotr Mardziel, Anupam Datta, and Pulkit Grover. 2020. An Information-Theoretic Quantification of Discrimination with Exempt Features.. In *AAAI*. 3825–3833.

[11] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[12] Harrison Edwards and Amos Storkey. 2015. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897* (2015).

[13] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*. 259–268.

[14] Kenji Fukumizu, Arthur Gretton, Xiaohai Sun, and Bernhard Schölkopf. 2008. Kernel measures of conditional dependence. In *Advances in neural information processing systems*. 489–496.

[15] Moritz Hardt, Eric Price, and Nati Srebro. 2016. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*. 3315–3323.

[16] Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. 2018. Fairness Without Demographics in Repeated Loss Minimization. In *International Conference on Machine Learning*. 1929–1938.

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

[18] Faisal Kamiran, Indrė Žliobaitė, and Toon Calders. 2013. Quantifying explainable discrimination and removing illegal discrimination in automated decision making. *Knowledge and information systems* 35, 3 (2013), 613–644.

[19] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. 2017. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems*. 656–666.

[20] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807* (2016).

[21] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual fairness. In *Advances in Neural Information Processing Systems*. 4066–4076.

[22] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. 2016. How we analyzed the COMPAS recidivism algorithm. *ProPublica (5 2016)* 9 (2016).

[23] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning Adversarially Fair and Transferable Representations. In *International Conference on Machine Learning*. 3384–3393.

[24] Amitabha Mukerjee, Rita Biswas, Kalyanmoy Deb, and Amrit P Mathur. 2002. Multi–objective evolutionary algorithms for the risk–return trade–off in bank loan management. *International Transactions in operational research* 9, 5 (2002), 583–597.

[25] Razieh Nabi and Ilya Shpitser. 2018. Fair inference on outcomes. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

[26] Judea Pearl. 2009. *Causality*. Cambridge university press.

[27] Joseph D Ramsey. 2014. A scalable conditional independence test for nonlinear, non-gaussian data. *arXiv preprint arXiv:1401.5031* (2014).

[28] Lauren A Rivera. 2012. Hiring as cultural matching: The case of elite professional service firms. *American sociological review* 77, 6 (2012), 999–1022.

[29] Chris Russell, Matt J Kusner, Joshua Loftus, and Ricardo Silva. 2017. When worlds collide: integrating different counterfactual assumptions in fairness. In *Advances in Neural Information Processing Systems*. 6414–6423.

[30] Dino Sejdinovic, Bharath Sriperumbudur, Arthur Gretton, and Kenji Fukumizu. 2013. Equivalence of distance-based and RKHS-based statistics in hypothesis testing. *The Annals of Statistics* (2013), 2263–2291.

[31] Peter Spirtes, Clark N Glymour, Richard Scheines, and David Heckerman. 2000. *Causation, prediction, and search*. MIT press.

[32] Eric V Strobl, Kun Zhang, and Shyam Visweswaran. 2019. Approximate kernel-based conditional independence tests for fast non-parametric causal discovery. *Journal of Causal Inference* 7, 1 (2019).

[33] Hao Wang, Berk Ustun, and Flavio Calmon. 2019. Repairing without Retraining: Avoiding Disparate Impact with Counterfactual Distributions. In *International Conference on Machine Learning*. 6618–6627.

[34] Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. 2019. PC-Fairness: A Unified Framework for Measuring Causality-based Fairness. In *Advances in Neural Information Processing Systems*. 3399–3409.

[35] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*. 1171–1180.

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

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

[38] Kun Zhang, Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2012. Kernel-based conditional independence test and application in causal discovery. *arXiv preprint arXiv:1202.3775* (2012).

[39] Han Zhao, Amanda Coston, Tameem Adel, and Geoffrey J. Gordon. 2020. Conditional Learning of Fair Representations. In *International Conference on Learning Representations*.

A PROOFS

A.1 Proof of Proposition 4.2

Proposition. *If random variable S is binary and $S \in \{0,1\}$, the random variables Z, S are independent conditional on F ($Z \perp S \mid F$) if and only if, for any $\tilde{h} \in \mathcal{E}_{ZF}$,*

$$\mathbb{E}[\mathbb{I}(S=1) \cdot \tilde{h}(Z,F)] = 0,$$

where $\mathcal{E}_{ZF}$ is shown in the equation (14) and $\mathbb{I}(S=1)$ is the indicative function defined as follow:

$$\mathbb{I}(S=1) = \begin{cases} 1, & \text{if} \quad S=1, \\ 0, & \text{if} \quad S=0. \end{cases}$$

Proof. On the one hand, $\mathbb{I}(S=1) \in L_S^2$. Thus when Z, S are independent conditional on F, for any $\tilde{h} \in \mathcal{E}_{ZF}$, according to lemma 4.1, $\mathbb{E}[\mathbb{I}(S=1) \cdot \tilde{h}] = 0$.

On the other hand, $S \in \{0,1\}$. Therefore for any $u \in L_S^2$, function u can be expressed as:

$$\begin{aligned} u(S) &= a \cdot \mathbb{I}(S=0) + b \cdot \mathbb{I}(S=1) \\ &= a + (b-a) \cdot \mathbb{I}(S=1), \end{aligned}$$

where $a, b \in \mathbb{R}$. Thus for any $\tilde{h} \in \mathcal{E}_{ZF}$, when equation (15) is satisfied,

$$\begin{aligned} \mathbb{E}[u \cdot \tilde{h}] &= \mathbb{E}\left[(a + (b-a) \cdot \mathbb{I}(S=1))\, \tilde{h}\right] \\ &= a \cdot \mathbb{E}[\tilde{h}] + (b-a) \cdot \mathbb{E}[\mathbb{I}(S=1) \cdot \tilde{h}] \\ &= a \cdot \mathbb{E}[\mathbb{E}[\tilde{h}|F]] \\ &= 0. \end{aligned}$$

As a result, according to lemma 4.1, Z and S are independent conditional on F. □

A.2 Proof of Theorem 4.3

Theorem. *If random variable S is binary and $S \in \{0,1\}$, the random variables Z, S are independent conditional on F ($Z \perp S \mid F$) if and only if, for any $h \in L_{ZF}^2$,*

$$\begin{aligned} Q(h) \triangleq\ & \mathbb{E}\left[\mathbb{I}(S=1)P(S=0|F)h(Z,F)\right] \\ & - \mathbb{E}\left[\mathbb{I}(S=0)P(S=1|F)h(Z,F)\right] = 0. \end{aligned}$$

Proof. On the one hand, if $Z \perp S \mid F$, according to proposition 4.2, for any $\tilde{h} \in \mathcal{E}_{ZF}$, $\mathbb{E}[\mathbb{I}(S=1) \cdot \tilde{h}] = 0$. Then for any function $h \in L_{ZF}$, define the corresponding $\tilde{h}$ as

$$\tilde{h}(Z,F) = h(Z,F) - \mathbb{E}[h|F].$$

Because

$$\mathbb{E}[\tilde{h}|F] = \mathbb{E}\left[[h(Z,F) - \mathbb{E}[h|F]]|F\right] = \mathbb{E}[h|F] - \mathbb{E}[h|F] = 0,$$

we can know that $\tilde{h} \in \mathcal{E}_{ZF}$. Hence, we have

$$\begin{aligned} Q(h) =& \mathbb{E}[\mathbb{I}(S=1)P(S=0|F) \cdot h] - \mathbb{E}[\mathbb{I}(S=0)P(S=1|F) \cdot h] \\ =& \mathbb{E}[\mathbb{I}(S=1) \cdot h] - \mathbb{E}[P(S=1|F) \cdot h] \\ =& \mathbb{E}[\mathbb{I}(S=1) \cdot h] - \mathbb{E}\left[\mathbb{E}[P(S=1|F) \cdot h|F]\right] \\ =& \mathbb{E}[\mathbb{I}(S=1) \cdot h] - \mathbb{E}\left[P(S=1|F) \cdot \mathbb{E}[h|F]\right] \\ =& \mathbb{E}[\mathbb{I}(S=1) \cdot h] - \mathbb{E}\left[\mathbb{E}[\mathbb{I}(S=1)|F] \cdot \mathbb{E}[h|F]\right] \\ =& \mathbb{E}[\mathbb{I}(S=1) \cdot h] - \mathbb{E}\left[\mathbb{E}\left[\mathbb{I}(S=1) \cdot \mathbb{E}[h|F]|F\right]\right] \\ =& \mathbb{E}\left[\mathbb{I}(S=1) \cdot h - \mathbb{I}(S=1) \cdot \mathbb{E}[h|F]\right] \\ =& \mathbb{E}[\mathbb{I}(S=1) \cdot \tilde{h}] = 0. \end{aligned}$$

On the other hand, if for any $h \in L_{ZF}$, $Q(h) = 0$, then consider any function $\tilde{h} \in \mathcal{E}_{ZF}$. Similarly, we can get

$$\mathbb{E}[\mathbb{I}(S=1) \cdot \tilde{h}] = Q(\tilde{h}) = 0.$$

According to proposition 4.2, Z and S are independent conditional on F. □

A.3 Proof of Theorem 4.5

Theorem. *$L_{fair}(Z,F,S)$, H_{ZF}, and $Q(h)$ are defined in theorem 4.3 and definition 4.4. Then*

$$L_{fair}(Z,F,S) = \sup_{h \in H_{ZF}} |Q(h)| = \sup_{h \in H_{ZF}} Q(h).$$

Proof. Because

$$\begin{aligned} & Q(h) + Q(1-h) \\ =& \mathbb{E}[\mathbb{I}(S=1)P(S=0|F)] - \mathbb{E}[\mathbb{I}(S=0)P(S=1|F)] \\ =& \mathbb{E}\left[\mathbb{E}[\mathbb{I}(S=1)P(S=0|F)|F]\right] - \mathbb{E}\left[\mathbb{E}[\mathbb{I}(S=0)P(S=1|F)|F]\right] \\ =& \mathbb{E}\left[P(S=0|F)\mathbb{E}[\mathbb{I}(S=1)|F]\right] - \mathbb{E}\left[P(S=1|F)\mathbb{E}[\mathbb{I}(S=0)|F]\right] \\ =& \mathbb{E}\left[P(S=0|F)P(S=1|F)\right] - \mathbb{E}\left[P(S=0|F)P(S=1|F)\right] = 0, \end{aligned}$$

we have

$$\sup_{h \in H_{ZF}} Q(h) = \inf_{h \in H_{ZF}} Q(h)$$

Hence,

$$\sup_{h \in H_{ZF}} |Q(h)| = \max\left\{\sup_{h \in H_{ZF}} Q(h), \inf_{h \in H_{ZF}} Q(h)\right\} = \sup_{h \in H_{ZF}} Q(h).$$

□

A.4 Proof of Theorem 4.6

Theorem. *L'_{fair} provides an upper bound of L_{fair}, i.e.*

$$L'_{fair} \geq L_{fair}.$$

Proof. Since $h \in H_{ZF}$, we have $0 \leq h \leq 1$. Therefore $h \geq h^2$ and $1-h \geq (1-h)^2$. As a result, for any $h \in H_{ZF}$,

$$\begin{aligned} & Q'(h) \\ =& C - \left[\mathbb{E}[\mathbb{I}(S=1)P(S=0|F) \cdot (1-h)^2] + \mathbb{E}[\mathbb{I}(S=0)P(S=1|F) \cdot h^2]\right] \\ \geq& C - \left[\mathbb{E}[\mathbb{I}(S=1)P(S=0|F) \cdot (1-h)] + \mathbb{E}[\mathbb{I}(S=0)P(S=1|F) \cdot h]\right] \\ =& Q(h). \end{aligned}$$

Finally we get

$$L'_{\text{fair}} = \sup_{h \in H_{ZF}} Q'(h) \geq \sup_{h \in H_{ZF}} Q(h) = L_{\text{fair}}.$$

□

B EXPERIMENTAL DETAILS

Table 3: Hyper-parameters in the experiments

	Adult	COMPAS	Dutch
# of hidden units in prediction	60	8	35
# of hidden units in adversary	50	8	20
# of adversarial steps	10	5	5
Batch size	512	256	512
Epoch	400	400	400
Learning algorithm	Adadelta	Adadelta	Adadelta
Learning rate	1.0	1.0	1.0

We fix the baseline network architectures so that they are shared among different methods. In detail, we set UNFAIR as a single hidden layer MLP with ReLU as activation function and logistic regression as the outcome function. For the adversary part of CFAIR, LAFTR, and ALFR, we also use a single hidden layer MLP. Its input is the hidden layer in UNFAIR and we apply logistic regression to the outcome. As for our method, we add another $|\mathcal{F}|$ units to the input of the adversarial network. The information about the hyper-parameters is shown in table 3.

To get the accuracy-fairness trade-off curve, we sweep across the coefficient λ in equation (24) from 0.1 to 20. For each coefficient and each model, we train and fine-tune it for 5 times and get the mean of accuracy and fairness metric on the test set. Finally, we calculate the Pareto front of these results as commonly used in literatures[2, 23].