Open Target Debiasing

Yue Cui, Qichen Tan, Jing Zhao

Abstract

[Contribution: Yue Cui] Fairness-aware machine learning aims to prevent demographic discriminative outcomes of machine learning models. Existing studies focus on debiasing *w.r.t.* a handful of pre-defined protected groups. However, this can be of concern in the wild for that 1) the bias of a model is hard to be known in advance, and 2) fairness of other demographic groups that are not known to the debiasing-target can be indeliberately harmed. This paper introduces open target debiasing (OTD), which imposes no specification on the biased groups and the model is expected to achieve fairness over all combined sensitive attributes. To solve the problem of OTD, we first extend the concept of demographic parity into OTD to quantify the fairness over all possible sensitive features. An open target fair representation learning framework, named OTIS, is proposed, where provable guarantees via the lens of mutual information are used to achieve parity and maintain the accuracy of downstream classification tasks. Practical instantiations of the objectives are also provided. The experiments on two public real-world datasets demonstrate OTIS achieves much better parity at similar level of classification F1 score compared with the baseline approaches.

1. Introduction

[Contribution: Qichen Tan] Despite the great convenience that machine learning algorithms provide, the unfairness and biases that already exist in real-world data may be involuntarily introduced to machine learning systems. Learned from data with ethical or legal implications, there is a piece of growing evidence that the algorithms may unintentionally provide discriminate predictions (Kearns et al., 2019; Sharifi-Malvajerdi et al., 2019; Lackner, 2020; Rios, 2020; Du et al., 2020). Therefore, debiasing in machine learning has drawn increasing awareness.

To address this issue, a variety of methods for fairness-aware machine learning have been proposed (Mehrabi et al., 2021). According to the phase of the actions taken, current solutions mainly falls into three categories: 1) pre-processing-based methods which process or re-weight the training data (Kamiran & Calders, 2012; Zemel et al., 2013); 2) in-processing-based methods, which adapt fairness constraints via techniques such as regularization, adversarial learning, or boosting, during model training, (Kamishima et al., 2012; Zafar et al., 2017b) and 3) post-processing methods (Kamiran et al., 2012; Dwork et al., 2012), which perturb only the model output without touching upon the inside.

Attributes	**Race**		Sex	
Training status	EO	DP	EO	DP
Start	0.2193	0.0917	0.0415	0.0114
Converged	**0.0242**	**0.0263**	**0.4063**	**0.1523**

Table 1. Performances of the 2-layer MLP with FairBatch (Roh et al., 2020b) debiasing on the attribute Race.

The majority of existing studies focus on debiasing *w.r.t.* pre-defined attributes, that is, the protected group is given beforehand. However, such a practice of defining the debiasing target can be problematic. This is because first, in most scenarios, the bias unintentionally learned by a model is hard to be known in advance or even be detected. Second, though fairness of the protected group may somehow be guaranteed, our observation that the bias on other groups can be aggravated indicates, however, the improvement of fairness is local and comes at a price of deepening model's discrimination on sensitive attributes that are not included in the debasing target. As an empirical example, we took a 2-layer multi-layer perception (MLP) (Friedman et al., 2001) as the classifier and adopted state-of-the-art method FairBatch (Roh et al., 2020b) to debias on the attribute **Race** of the AdultCensus dataset (Kohavi et al., 1996). We calculated the fairness metric Equalized Odds (EO) and Demographic Parity (DP) (the lower the fairer for both) (Roh et al., 2020b) for Race and another group Sex. As depicted in Table 1, the model tends to be biased more on Race at the beginning, but after training, there tends to be more bias on Sex. This might be resulted from that the learning algorithm tends to re-weight its feature extraction focus when fairness constraints are given.

In other words, in the scenarios where debiasing target is pre-defined, one may create new unfairness through debiasing, which is definitely not what is desired. Therefore, in this paper, we propose to explore a novel setting for

fairness-aware machine learning, named open target debiasing (OTD), where no prior knowledge *w.r.t.* bias is required before debiasing and the model is expected to achieve fairness *w.r.t.* all original sensitive attributes and their combinations (see Def. 2.2). Note that OTD can be seen as a generalization of the current setting of multi-attribute debiasing if the protected group should be specified. However, the novel problem of OTD brings non-trivial challenges.

[Contribution: Yue Cui] To overcome the problem that debiasing on one particular attribute could affect the fairness of the others, it is straightforward to consider all attributes together, *i.e.,* original singular attributes and their possible combinations, *e.g.,* Race, Sex, and Sex_Race for the above example. However, the consequence is, **the number of protected groups could be enormous**. Even in the simplest case, where there are m sensitive attributes in total and each is binary, there could be $C_m^1 + ... + C_m^m = 2^m - 1$ attributes and the i-th one $i \in [1, ..., m]$ has 2^i different values. Since fairness is usually achieved by adding constraints *w.r.t.* all protected groups (Roh et al., 2020b; Gupta et al., 2021; Kamishima et al., 2012), this can be problematic for previous methods even before debiasing. To formulate these constraints, it is an essential step to compute fairness metrics, the majority of which are defined *w.r.t.* group membership of instances, where one instance can be regarded as a member of a group if it has certain values of the attribute (see Eq. 1 for an example). Such a procedure is of reasonable computation cost in the traditional setting, since only a limited number of groups need to be considered, but can be problematic for OTD. In other words, the computation of fairness metric is of polymorphic complexity while that has not yet factor any debiasing operation. When it comes to debiasing, for models using group-wise weighted matrix (Zhu et al., 2018; Gupta et al., 2021), they would certainly face a rising number of parameters that is exponential to the number of original sensitive attributes.

Naive solutions do exist to bypass the trouble caused by adding a huge number of constraints. One may greedily adjust the debiasing target and monitor the fairness over all sensitive attributes until dynamic balance is reached. However, despite such an approach may intuitively make sense, **detecting and debiasing dynamically on an in-training decision algorithm might lead to divergence and instability**, due to constantly changing objectives. Empirical studies are conducted in the experimental section to verify this statement.

Last but not least, though it is often preferred and has been discussed widely in traditional debiasing setting (Gupta et al., 2021; Gitiaux & Rangwala, 2021), **provably guarantee is still a blank area in OTD**. Therefore, it is an essence to develop an efficient and effective approach to solving the problem of OTD.

With above new and significant challenges, in this paper, we propose an **O**pen **T**arget fa**I**r repre**S**entation learning framework, OTIS, including learning objectives concerning both classification accuracy and fairness, and instantiations of the objectives. Focusing on the fairness metric demographic parity, we first extend it to the OTD setting and prove that the parity can be achieved in a provable fashion via the lens of mutual information between the learned representation and sensitive attributes. To tackle the challenge of the massive number of protected groups, we prove that the fairness *w.r.t.* level-1 (see definition in Def. 2.2) and higher combined sensitive attributes can be bounded by that of level-1 ones, which indicates the OTD can be achieved by controlling the bias of only a reasonable number of sensitive attributes. Provable boundaries are adopted to estimate the mutual information terms. We instantiate the estimations by leveraging tools from KL divergence and deep neural networks.

Our contributions can be summarized as follows:

- We introduce a novel and practical setting for fairness-aware machine learning that imposes no specification on debiasing target, *i.e.,* no pre-defined protected groups, but aims to achieve fairness over all combined sensitive attributes.
- We evaluate existing debiasing methods under the OTD setting and proved that models tailored to specific debiasing targets do not work well in OTD. This emphasizes the necessity for an OTD model.
- An end-to-end and guaranteed solution is proposed via the lens of mutual information. Extensive experiments on two public real-world datasets verify that compared with baselines, OTIS can achieve much better parity at a similar level of classification F1 score.

2. Open Target Debiasing

[Contribution: Yue Cui] The problem of OTD can be formulated as follows:

Definition 2.1. Open Target Debiasing (OTD): For a dataset $\mathcal{D} = \{x_i, y_i, c_i\}_{i=1}^n$, where values of x_i, y_i, c_i are *i.i.d.* realizations of random variables $\mathbf{x}, \mathbf{y}, \mathbf{c}$ distributed by $p(\mathbf{x}, \mathbf{y}, \mathbf{c})$. $\mathbf{x}$ are features, $\mathbf{y}$ is the label, and $\mathbf{c}$ are demographic sensitive attributes, where $x_i \in \mathbb{R}^n, y_i \in \mathbb{R}^1, C_i \in \mathbb{R}^m$. $\mathbf{c}$ may all or partially be included in $\mathbf{x}$. Given a fairness criteria $\mathcal{M}$, any learning algorithm $\mathcal{A}$, the goal of open target debiasing (OTD) is to achieve fairness *w.r.t.* $\mathcal{M}$ over all original sensitive attributes and combined sensitive attributes (see Def. 2.2) on $\mathcal{A}$.

Examples of demographic sensitive attributes are {gender, race, age}. A large number of fairness criteria have been

proposed, addressing fairness from different perspectives, and we here mainly focus on demographic parity, which will be formally defined in the following context.

Denote the sensitive attributes in a given dataset as the original sensitive attributes, selecting any number of attributes from the original ones, one can form a combined sensitive attribute. Adopting the widely used assumption that each attribute can be simplified as categorical (Mehrabi et al., 2021), given m original sensitive attributes $\mathcal{C}_o = \{\mathbf{c}_1, ..., \mathbf{c}_m\}$, where $\mathbf{c}_i \in \{1, ..., N_i\}$, N_i the he total number of categories of $\mathbf{c}_i$, we define a combined sensitive attributes as follows.

Definition 2.2. Combined Sensitive Attributes: For an $L \leq m$, select any L attributes from $\mathcal{C}_o$, denoted as $\{\mathbf{c}_{i,1}, ..., \mathbf{c}_{i,L}\}$, the level-L combined sensitive attribute, denoted as $\mathbf{c}'$, , *i.e.,* $\mathbf{c}' \in \{1, ..., \prod_{j=1}^{L} N_{i,j}\}$, each of the new categorical value denotes a mixture of L values from every of the selected attributes.

For example, if the original sensitive attributes are $Gender = \{male, female\}$ and $Age = \{> 50, \leq 50\}$, then there will be one and only combined sensitive attribute: $Gender\_Age = \{(male, > 50), (female, > 50), (male, \leq 50), (female, \leq 50)\}$. The Level-1 combined sensitive attributes are the original sensitive attributes themselves. For m original attributes, there are in total, m levels of combined sensitive attributes. We denote the multivariate variable with all m levels of combined sensitive attributes included as $\mathbf{C}$. OTD also requires a model to be fair on all combined sensitive attributes.

The problem of OTD is different from multi-attribute debiasing in the scale of debiasing target and can be seen as an extension of the latter. This is because all possible combined sensitive attributes should be considered in OTD, *i.e.,* $\mathbf{C}$, while multi-attribute debiasing only deals with a fixed subset of $\mathbf{C}$.

Notations. Suppose we have an encoding network NN_{en}, which embeds the input features $\mathbf{x}$ into a representation vector $\mathbf{z}$. Denote that $\mathbf{z}, \mathbf{x} \sim p(\mathbf{z}|\mathbf{x})$. The i-th sensitive attribute is represented by $\mathbf{c}_i$, which corresponds to the i-th variable of $\mathbf{C}$. Note that from this section, c_i in light will be used to describe an instantiated $\mathbf{c}_i$ but not the i-th sample of $\mathbf{c}$ in dataset $\mathcal{D}$ as it is used in Def. 2.1. A decision algorithm $\mathcal{A}$ will act on $\mathbf{z}$ and produce $\hat{y}$, which is the prediction of the classification task. The mutual information between $\mathbf{z}$ and $\mathbf{c}_i$ is denoted as $I(\mathbf{z}; \mathbf{c}_i)$.

One of the primary things for OTD is to properly define a metric that measures fairness in OTD, which should act on all combined sensitive attributes $\mathbf{C}$. We here focus on demographic parity (DP). To define DP on $\mathbf{C}$, we first introduce the definition of attribute-wise demographic parity, which is widely used in the traditional debiasing settings (Gupta et al., 2021; Mehrabi et al., 2021).

Definition 2.3. Attribute-wise Demographic Parity: Given any learning algorithm $\mathcal{A}$, a sensitive attribute $\mathbf{c}_k$, the demographic parity *w.r.t.* $\mathbf{c}_k$ is defined as:

$$\Delta_{DP}(\mathcal{A}, \mathbf{c}_k) = \max_{i,j} |P(\hat{\mathbf{y}} = 1|\mathbf{c}_k = i) - P(\hat{\mathbf{y}} = 1|\mathbf{c}_k = j)| \tag{1}$$

where $\hat{\mathbf{y}}$ denotes outcome of $\mathcal{A}$, P is probability, $i, j \in [1, ..., N_k]$, and N_k is the total number of categories of $\mathbf{c}_k$.

Thus in the setting of OTD, we can define the following demographic parity.

Definition 2.4. Demographic Parity: Given all combined sensitive attributes, *i.e.,* $\mathbf{C}$, the demographic parity *w.r.t.* $\mathbf{C}$ can be defined as follows:

$$\begin{aligned}\Delta_{DP}(\mathcal{A}, \mathbf{C}) &= ||\mathbf{u}_{\Delta_{DP}(\mathcal{A},\mathbf{C})}||_2 \\ &= |||[\Delta_{DP}(\mathcal{A}, \mathbf{c}_i), ..., \Delta_{DP}(\mathcal{A}, \mathbf{c}_{|\mathbf{C}|})]||_2 \\ &= \sqrt{\sum\nolimits_{i=1}^{|\mathbf{C}|} \Delta_{DP}(\mathcal{A}, \mathbf{c}_i)^2}\end{aligned}$$

where $|\mathbf{C}|$ denotes the number of variables in $\mathbf{C}$, $\mathbf{u}_{\Delta_{DP}(\mathcal{A},\mathbf{C})} = [\Delta_{DP}(\mathcal{A}, \mathbf{c}_1), ..., \Delta_{DP}(\mathcal{A}, \mathbf{c}_{|\mathbf{C}|})] \in \mathbb{R}^{1\times|\mathbf{C}|}$ is a vector formed by attribute-wise Δ_{DP}, $||\cdot||_2$ is the L2-norm.

As in the traditional setting, the definition of fairness has been an active field of research and many are proposed to address different algorithmic biases (Mehrabi et al., 2021), and there is no golden standard. DP of OTD can also be defined in different ways. The advantages of Def. 2.4 are three-folds. First, it intuitively makes sense because it comprehensively measures the parity of all $\mathbf{c}_i$ in $\mathbf{C}$. Second, pairwise comparison between attributes can be avoided, *i.e.,* computing $|P(\hat{\mathbf{y}} = 1|\mathbf{c}_m = i) - P(\hat{\mathbf{y}} = 1|\mathbf{c}_n = j)|$, which is computationally expensive. More importantly, we prove in the next section that the defined parity can be controlled by limiting mutual information between the representations and the sensitive attributes.

3. Open Target Fair Representation Learning

Equipped with the definition of DP in OTD, we here introduce the open target fair representation learning framework (OTIS), which includes the learning objectives and their practical instantiations.

3.1. Deriving the Objective for OTIS

[Contribution: Yue Cui, Qichen Tan] In general, OTD has two goals: 1) maximizing the classification accuracy, and 2) minimizing the discrimination when performing the

downstream classification task. Interpreted in the principle of information (Tishby et al., 2000), since $\mathbf{z}$ is the trainable part of the framework, the first goal can be straightforwardly placed as: it requires the representation $\mathbf{z}$ to maximize the information to $\mathbf{y}$ (the prediction). However, it is unclear how parity, *i.e.,* $\Delta_{DP}(\mathcal{A}, \mathbf{C})$ can be related. That is to say, even though demographic parity can be clearly defined, it is a non-trivial task to model and optimize it directly.

The majority of early studies model the fairness criterion conceptually in an in-process fashion and evaluate related metrics after training, in which there could be a gap between the debiasing target and fairness measurement. Recently, provable controllable guarantees are proposed to bridge such gaps (McNamara et al., 2017), where mutual information has been proven to be promising to serve as a link between input features, intermediate representations and demographic parity (Gupta et al., 2021; Gitiaux & Rangwala, 2021). Inspired by these works, we here show that the sum of mutual information between $\mathbf{z}$ and $\mathbf{c}_i$ can be used to bound the parity over $\mathbf{C}$.

Theorem 3.1. *For $z, c_i \sim p(\mathbf{z}, \mathbf{c}_i)$, $z \in \mathbb{R}^d$, $\mathbf{c}_i \in \mathbf{C}$, $c_i \in \{1, ..., N_i\}$, and any decision algorithm $\mathcal{A}$ that acts on z, we have*

$$\sum\nolimits_{i=1}^{|\mathbf{C}|} I(\mathbf{z}; \mathbf{c}_i) \geq g(\gamma, \Delta_{DP}(\mathcal{A}, \mathbf{C}))$$

where $\gamma = \min_i \alpha_i$, $\alpha_i = \min_j \pi_{i,j}$, $\pi_{i,j} = P(\mathbf{c}_i = j)$.

The proof of Theorem 3.1 is delayed to Appendix A.1.

Proving DP can be limited by mutual information, we can result in the following objective:

$$\mathcal{O}' \triangleq \min_{p(\mathbf{z}|\mathbf{x}) \in \Omega} [-I(\mathbf{y}; \mathbf{z}) + \beta' \sum\nolimits_{i=1}^{|\mathbf{C}|} I(\mathbf{z}; \mathbf{c}_i)] \quad (2)$$

that is, maximizing the mutual information between $\mathbf{y}$ and $\mathbf{z}$ to achieve good performance on the classification task while minimizing the sum of mutual information between $\mathbf{z}$ and $\mathbf{c}_i$ to achieve good parity over $\mathbf{C}$. Ω denotes the search space of the conditional distribution of $\mathbf{z}$ given $\mathbf{x}$ and β' is a hyper-parameter to control the trade-off between the two sub-objectives.

Given that all combinations of sensitive attributes should be considered, for n single attributes, even if all are binary, there could be $C_n^1 + ... + C_n^n = 2^n - 1$ elements in $\mathbf{C}$ and the summation of the number of possible for values are $C_n^1 2^1 + ... + C_n^n 2^n$. This could be extremely expensive when parameterizing the distributions since it is usually implemented by values of attributes (Gupta et al., 2021). However, we show that this concern is unnecessary by introducing Theorem 3.2, which says that the mutual information between $\mathbf{z}$ and level-1 combined attributes bounds that of $\mathbf{z}$ and higher-level combined attributes. We defer the the proof of Theorem 3.2 to Appendix A.2.

To simplify the description, we now denote the combination of sensitive attributes $\mathbf{c}_i$ and $\mathbf{c}_j$ as $\mathbf{c}_{ij}$.

Theorem 3.2. *For a combined sensitive attribute $\mathbf{c}_{ij}$, which is formed by sensitive attribute $\mathbf{c}_i$ and $\mathbf{c}_j$, where $c_i \in \{1, .., N_i\}, c_j \in \{1, ..., N_j\}$, we have*

$$\min\{I(\mathbf{z}; \mathbf{c}_i), I(\mathbf{z}; \mathbf{c}_j)\} \geq I(\mathbf{z}; \mathbf{c}_{ij})$$

Based on this theorem, we can rewrite the objective as follows

$$\mathcal{O} \triangleq \min_{p(\mathbf{z}|\mathbf{x}) \in \Omega} [-I(\mathbf{y}; \mathbf{z}) + \beta \sum\nolimits_{i=1}^{|\mathbf{C}^1|} I(\mathbf{z}; \mathbf{c}_i)] \quad (3)$$

where $\mathbf{C}^1$ contains all level-1 combined sensitive attributes, *i.e.,* the original sensitive attributes.

3.2. Mutual Information Estimation

[Contribution: Jing Zhao] Despite being a pivotal tool linking parity and intermediate representation, mutual information has historically been difficult to compute (Paninski, 2003), especially when the probability distributions of data is unknown (Belghazi et al., 2018). As a common practice, we here adopt provable bounds and instantiate them using practical formalization to estimate the mutual information.

3.2.1. Estimate $I(\mathbf{y}; \mathbf{z})$

According to (Poole et al., 2019; Nguyen et al., 2010) for any random variables $\mathbf{a}$ and $\mathbf{b}$ and any function $f(\mathbf{a}, \mathbf{b}) \in \mathbb{R}$, we have

$$I(\mathbf{a}; \mathbf{b}) \geq \mathbb{E}_{p(\mathbf{a},\mathbf{b})}[f(a, b)] - \mathbb{E}_{p(\mathbf{a})p(\mathbf{b})}[exp(f(a, b) - 1)] \quad (4)$$

Proposition 3.3. ***The lower bound for*** *$I(\mathbf{y}; \mathbf{z})$: Our method to upper $I(\mathbf{y}; \mathbf{z})$ is similar to (Wu et al., 2020). Using the above formulation to $(\mathbf{y}, \mathbf{z})$ and set $f(\mathbf{y}, \mathbf{z}) = 1 + log\frac{p(y|z)}{p(y)p(y)}$, and then we have*

$$I(\mathbf{y}; \mathbf{z}) \geq 1 + \mathbb{E}_{p(\mathbf{y},\mathbf{z})}[log\frac{p(y|z)}{p(y)}] + \mathbb{E}_{p(\mathbf{y})p(\mathbf{z})}[\frac{p(y|z)}{p(y)}] \quad (5)$$

To characterize the above equation, we approximate $p(y|z)$ with $q(y|x) = NN_{de}(z)$, where NN_{de} denotes a parametrized neural network to decode z and get the prediction of y, and estimating $p(y)$ from the distribution of the data. We implement NN_{de} as a 1-layer MLP for training to get fair representations and replace it with several different decision algorithms $\mathcal{A}$ for test. To this end, ignoring constants, the RHS of Eq. 5 can be expressed as cross entropy of the classification task.

3.2.2. ESTIMATE $\sum_{i=1}^{|\mathbf{C}^1|} I(\mathbf{z};\mathbf{c}_i)$

Since it can be derived that for any three random variables $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$,

$$\begin{aligned} I(\mathbf{a};\mathbf{b},\mathbf{c}) &= I(\mathbf{a};\mathbf{b},\mathbf{c}) \\ I(\mathbf{a};\mathbf{b}) + I(\mathbf{a};\mathbf{c}|\mathbf{b}) &= I(\mathbf{a};\mathbf{c}) + I(\mathbf{a};\mathbf{b}|\mathbf{c}) \end{aligned} \tag{6}$$

as proposed in (Gupta et al., 2021), the second term of Eq. 3 can be re-written as:

$$\begin{aligned} \sum_{i=1}^{|\mathbf{C}^1|} I(\mathbf{z};\mathbf{c}_i) &= \sum_{i=1}^{|\mathbf{C}^1|} I(\mathbf{z};\mathbf{c}_i|\mathbf{x}) + I(\mathbf{z};\mathbf{x}) - I(\mathbf{z};\mathbf{x}|\mathbf{c}_i) \\ &= |\mathbf{C}^1| I(\mathbf{z};\mathbf{x}) - \sum_{i=1}^{|\mathbf{C}^1|} I(\mathbf{z};\mathbf{x}|\mathbf{c}_i) \end{aligned} \tag{7}$$

where $I(\mathbf{z};\mathbf{c}_i|\mathbf{x}) = 0$ because z is independent of $c_i, \forall c_i \in \mathbf{C}^1$. We can thus bound $I(\mathbf{z};\mathbf{c}_i)$ using the upper bound of $I(\mathbf{z};\mathbf{x})$, which is the information bottleneck (Tishby et al., 2000) term, and the lower bound of $I(\mathbf{z};\mathbf{x}|\mathbf{c}_i)$, aiming at maximizing the information between $\mathbf{z}$ and $\mathbf{x}$ without $\mathbf{c}_i$.

Estimate $I(\mathbf{z};\mathbf{x})$. Follow a similar fashion as (Gupta et al., 2021; Poole et al., 2019), we derive the upper bound of $I(\mathbf{z};\mathbf{x})$ as

$$\begin{aligned} I(\mathbf{z};\mathbf{x}) &= \mathbb{E}_{p(\mathbf{z},\mathbf{x})} log \frac{p(z|x)}{q(z)} - KL(p(\mathbf{z})\|q(\mathbf{z})) \\ &\leq \mathbb{E}_{p(\mathbf{z},\mathbf{x})} log \frac{p(z|x)}{q(z)} = \mathbb{E}_{p(\mathbf{x})} KL\big(p(\mathbf{z}|\mathbf{x})\|q(\mathbf{z})\big) \end{aligned} \tag{8}$$

where $p(\mathbf{z})$ is any distribution. To characterize Eq. 8, we may simply set $q(\mathbf{z})$ as $\mathcal{N}(0, I)$, and estimate $p(z|x)$ as $q(z|x) = NN_{en}(x)$, where NN_{en} is implemented as a 1-layer MLP. KL loss can thus be obtained to estimate Eq. 8.

Estimate $\sum_{i=1}^{|\mathbf{C}^1|} I(\mathbf{z};\mathbf{x}|\mathbf{c}_i)$. Applying the Nguyen, Wainright & Jordan's bound (Nguyen et al., 2010) to random variables $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$, we have

$$\begin{aligned} I(\mathbf{a};\mathbf{b}|\mathbf{c}) \geq\ & \mathbb{E}_{p(\mathbf{a},\mathbf{b},\mathbf{c})}[g(a,b,c)] \\ & - e^{-1}\mathbb{E}_{p(\mathbf{a}|\mathbf{c})p(\mathbf{b},\mathbf{c})}[\exp g(a,b,c)] \end{aligned} \tag{9}$$

Proposition 3.4. ***The lower bound for*** $I(\mathbf{z};\mathbf{x}|\mathbf{c}_i)$***:*** *Use Eq. 9 to* $(\mathbf{z},\mathbf{x},\mathbf{c}_i)$*, plugging in* $g(z,x,c_i) = 1 + log\frac{p(z,x,c_i)}{p(z|c_i)p(x,c_i)}$*, we can obtain that*

$$\begin{aligned} I(\mathbf{z};\mathbf{x}|\mathbf{c}_i) \geq\ & 1 + \mathbb{E}_{p(\mathbf{z},\mathbf{x},\mathbf{c}_i)}[log\frac{p(z,x,c_i)}{p(z|c_i)p(x,c_i)}] \\ & - \mathbb{E}_{p(\mathbf{z}|\mathbf{c}_i)p(\mathbf{x},\mathbf{c}_i)}[log\frac{p(z,x,c_i)}{p(z|c_i)p(x,c_i)}] \end{aligned} \tag{10}$$

Though $g(z,x,c_i) = 1 + log\frac{p(z,x,c_i)}{p(z|c_i)p(x,c_i)}$ is an optimal choice to make the bound tight, it can be challenging to compute the joint density. We here adopt generator-discriminator based method (Mukherjee et al., 2020; Molavipour et al., 2020) to empirically estimate the mutual information.

The key idea is to use trainable neural networks to discriminate samples from the joint distribution $p(\mathbf{z},\mathbf{x},\mathbf{c}_i)$ and product distribution $p(\mathbf{z}|\mathbf{c}_i)p(\mathbf{x},\mathbf{c}_i)$. In our implementation, we use an MLP to characterize the distribution of $p(\mathbf{z}|\mathbf{c}_i)$, *i.e.,* $p(\mathbf{z}|\mathbf{c}_i) = NN_G(\mathbf{c}_i)$, while $p(\mathbf{z},\mathbf{x},\mathbf{c}_i)$ and $p(\mathbf{x},\mathbf{c}_i)$ are estimated from the data. The sampled instances from the two distributions can be denoted as $\mathcal{B}_{joint}$ and $\mathcal{B}_{prod}$, with labels $\mathbf{1}$ and $\mathbf{0}$, respectively, where the bold numbers denote vectors. Then a discriminator, denoted as NN_D, is be trained on $\{\mathcal{B}_{joint}, \mathbf{1}\}$ and $\{\mathcal{B}_{prod}, \mathbf{0}\}$ with cross entropy loss. NN_D is also implemented as an MLP. Mathematically, suppose we sample K instances from $\{\mathcal{B}_{joint}, \mathbf{1}\}$ and $\{\mathcal{B}_{prod}, \mathbf{0}\}$ separately, the loss function can be formulated as:

$$\begin{aligned} L_{c_i} = -\frac{1}{2K}[& \sum_{(z,x,c_i)\in\mathcal{B}_{joint}} log\omega(z,x,c_i) + \\ & \sum_{(z,x,c_i)\in\mathcal{B}_{prod}} log\big(1-\omega(z,x,c_i)\big) \end{aligned} \tag{11}$$

where $\omega(z,x,c_i)$ is the prediction of NN_D *w.r.t.* the probability of whether the sample is from the joint distribution. In other words, the optimal $\omega(z,x,c_i)$ approximates $p(z,x,c_i)$ and minimizing Eq. 11 will help the parameterized distributions close to true ones. We then sum over L_{c_i} to obtain L_c, which serves as an estimation of conditional mutual information $\sum_{i=1}^{|\mathbf{C}^1|} I(\mathbf{z};\mathbf{x}|\mathbf{c}_i)$. We illustrated the process of computing L_c in Algorithm 1.

4. Experiment and Results

4.1. Experimental Settings

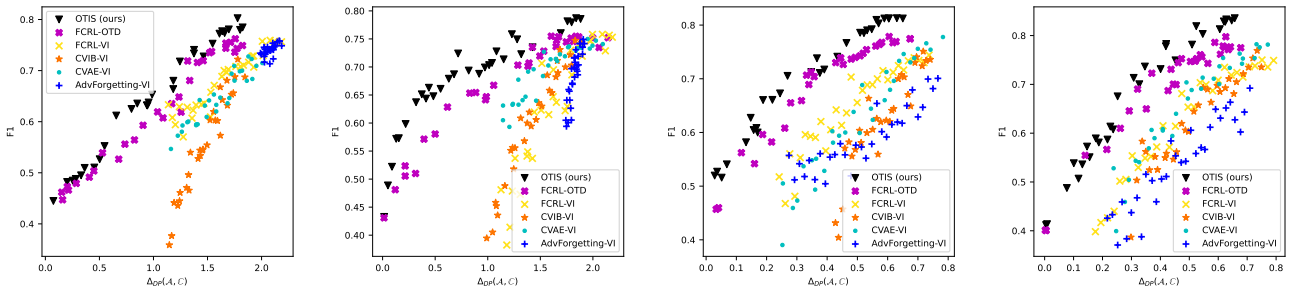
4.1.1. DATASETS AND METRICS

Dataset	# samples	# feat.	# original sensitive attr.
Heritage Health	55924	123	3
AdultCensus	45222	99	14

Table 2. Statistics of datasets

[Contribution: Qichen Tan, Yue Cui] We conduct experiments on two benchmark datasets Heritage Health [1] and AdultCensus (Kohavi et al., 1996). The Heritage Health dataset is collected by Heritage Provider Network, containing medical records of its member patients. The goal is to predict the Charleson Index for each patient, which indicates the ten-year mortality. The AdultCensus dataset is part of the 1994 census data of the US. The goal is to predict if a person makes more than 50K per year. Following the same procedure as (Gupta et al., 2021; Song et al., 2019;

[1] https://www.kaggle.com/c/hhp



(a) MLP - AdultCensus (b) AdaBoost - AdultCensus (c) MLP - Health Heritage (d) AdaBoost- Health Heritage

Figure 1. Trade-off between F1 and $\Delta_{DP}(\mathcal{A}, \mathbf{C})$ on the AdultCensus and Health Heritage dataset with MLP and AdaBoost.

Moyer et al., 2018), statistics of the pre-processed dataset are described in Table 2.

We compute the F1 score to measure the performance of the classification task and $\Delta_{DP}(\mathcal{A}, \mathbf{C})$ to measure OTD fairness.

4.1.2. Implementation Details

[Contribution: Jing Zhao] All experiments are implemented on one Intel (R) Xeon(R) Gold 6278C CPU @ 2.60GHz and one NVIDIA GeForce RTX 2080 GPU.

Each dataset is split into 6/2/2 for training, validation, and test. We fix the batch size to 128, following (Gupta et al., 2021). For representation learning, the weight decay rate is set as 1e-4. The embedding dimension of $\mathbf{z}$ is set as 8 for the Heritage Health dataset and 16 for the AdultCensus dataset. The depths of the MLP generator and discriminator in Sec. 3.2.2 are set as 1. We initialize the learning rate as 1e-3 and decay 0.01 every 10 epochs. In order to verify if the learned representation is model-agnostic, four decision algorithms are evaluated: Logistic Regression, 2-layer MLP, Random Forest, and AdaBoost. The maximum number of iterations for Logistic Regression and MLP is set as 1000, and apart from that other hyper-parameter settings are the same as they are default in scikit-learn [2].

We mainly consider two training paradigms to obtain fair representations: 1) pre-training based and 2) joint-training based. For the pre-training based approach, the network is optimized with the objective of minimizing classification loss only for the first few epochs before Eq. 3 taking over. For the joint-training-based approach, Eq. 3 will be applied from the very beginning of training. Unless otherwise specified, the default paradigm is joint training.

4.1.3. Comparison Baselines

[Contribution: Yue Cui] To the best of our knowledge, there is no previous model intentionally designed for OTD. Therefore, we re-implement existing methods with a modified training strategy to address OTD. Inspired by the thought experiment Veil of Ignorance (VI) (Rawls, 1971) (see App. C for more details), we consider adaptively and greedily choosing the debiasing target that benefits the least fair groups during training. For a model proposed for debiasing on a specified protected group, given $\mathbf{C}$, we initialize the protected attribute by arbitrarily choosing one from $\mathbf{C}$, denoted as $\mathbf{c}_t$, and start the representation learning. After a continuous number of iterations n after debiasing on $\mathbf{c}_t$, we measure $\Delta_{DP}(\mathcal{A}, \mathcal{C})$ of all attributes in $\mathbf{C}$, replace $\mathbf{c}_t$ with the most biased attribute. The training continues until the maximum number of iterations is reached or $\Delta_{DP}(\mathcal{A}, \mathcal{C})$ converges.

Five related models are used for comparison for their state-of-the-art performances, which include information theory based models: FCRL (Gupta et al., 2021), which uses contrastive information estimators, and CVAE (Gupta et al., 2021), which is a variational autoencoder implementation under the framework of FCRL, and CVIB (Moyer et al., 2018), which is information-theoretic optimization based; and state-of-the-art adversarial-based model Adversarial Forgetting (AdvForgetting) (Jaiswal et al., 2020). The structure and hyper-parameter settings of all the baselines are the same as (Gupta et al., 2021). n is set as 50. We denote the baseline under the proposed training strategy as *model*-VI. Since our work is closest to state-of-the-art work FCRL (Gupta et al., 2021), we also include a baseline FCRL-OTD by arbitrarily changing its objective function to Eq. 3 for comparison.

4.2. F1 *vs.* $\Delta_{DP}(\mathcal{A}, \mathbf{C})$

We measure the trade-off between F1-$\Delta_{DP}(\mathcal{A}, \mathbf{C})$ by tuning trade-off related hyper-parameters of the models, *i.e.,* β in the case of OTIS. The results on MLP and AdaBoostare are plotted in Fig. 1 and other decision algorithms can be found in App. D Fig. 5. It can be observed that OTIS surpasses all compared methods with a large margin. For near

[2] https://scikit-learn.org/stable/index.html

Method	MLP		Logistic Regression		AdaBoost		Random Forest	
	F1	DP	F1	DP	F1	DP	F1	DP
OTD	0.7821	1.7459	0.7815	1.7273	0.7774	1.6608	0.7704	1.7041
OTD w/o Iyz	0.7790	1.7549	0.7808	1.7480	0.7701	1.7052	0.7727	1.7400
OTD w/o Izc	0.7807	1.8178	0.7817	1.6654	0.7716	1.8722	0.7756	1.7966
OTD w/o Izc	0.7559	1.7633	0.7610	1.7278	0.7798	1.7804	0.7758	1.7436
OTD w/o Izxc	0.7857	1.8925	0.7829	1.8117	0.7815	1.8791	0.7783	1.7982

Table 3. The ablation study.

parity, OTIS achieves ∼4% higher F1 on the AdultCensus dataset and ∼3% on the Health Heritage dataset than the best baseline, and up to 0.2 lower disparity at a similar F1 score. This meets our expectations because the method is theoretically guaranteed. VI-baseline methods are shown to be not able to achieve high F1 and low parity, which is because the constantly varying objective makes it hard to find optimal but moving between saddle points. It is worth mentioning that the lead margin of OTIS is more evidential on the AdultCensus dataset. This is because when a larger number of sensitive attributes need to be considered, models proposed for debiasing on specified attributes suffer more from the instability of objectives. This empirically verifies the need for an OTD-friendly model and as well the effectiveness of OTIS. Due to limited space, more observations and discussions *w.r.t.* FCRL-OTD *vs.* OTIS and comparisons between baselines can be found in App. D.2.

4.3. Analysis on Training Paradigms

[Contribution: Qichen Tan] To further look into how does the proposed approach works on different training paradigms and test if it is fine-tuning friendly, we conduct experiments on pre-trained features. The pre-trained representations are obtained with classification task only, *i.e.,* without the second term in Eq. 3. To simulate representations of different expressiveness *w.r.t.* the downstream task, the number of [0, 100, 300, 500, 700, 900] iterations (batches) are implemented, where the expressiveness increases as the representation updates for more times. The parity and F1 *vs.* epoch plots on the training set of AdultCensus are reported in App. D Fig. 4. We can learn from plots that despite OTIS needing more iterations to debias on stronger representations, all tests convergent to the similar F1 and parity. This indicates that OTIS not only works well in the case of joint-training but can be capable of debiasing pre-trained feature representations, which implies a wider application scenario.

4.4. Ablation Study

[Contribution: Jing Zhao] To verify the effectiveness of each component, we conduct ablation studies with variants and test the F1 and the corresponding $\Delta_{DP}(\mathcal{A}, \mathbf{C})$ with β fixed as 0.05. The results on AdultCensus dataset with all four decision algorithm are demonstrated in Table 3, where DP denotes $\Delta_{DP}(\mathcal{A}, \mathbf{C})$. We observe similar trends on Health Heritage dataset so omit the results to avoid redundancy.

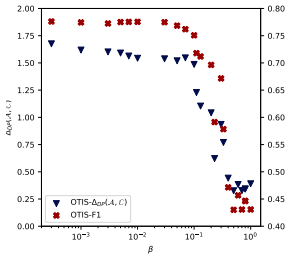


Figure 2. Effect of β.

Without fairness-related objectives, **OTIS w/o Izc** only serve as a simple classification-task pre-training encoder for the decision algorithm. Not surprisingly, it achieves high F1 but low $\Delta_{DP}(\mathcal{A}, \mathbf{C})$. **OTIS w/o Izx** denotes the variant of our approach that the information bottleneck term is removed. It can be found that the classification accuracy suffers a slight drop so as the fairness. Failing to obtain minimal sufficient information from data, during the process of debiasing, the variant is easier to be influenced by the fairness objective, which can be a conflict with the classification goal. As for **OTIS w/o Izxc**, it directly predicts the task label but with the information bottleneck principle equipped. With no fairness regulations and better feature extraction ability, it gets the best F1 but worst parity. There can be other objective functions for the classification task. For example, (Gupta et al., 2021) uses $I(\mathbf{y}; \mathbf{z}|\mathbf{c})$ to only infer $\mathbf{y}$ from $\mathbf{z}$ but with sensitive information eliminated. This could be an effective practice for a singular protected group, but we found from **OTIS w/o Iyz** that it fails to reach as high F1 and as low parity as the proposed one. This can be resulted from that 1) such an objective can be reached by minimizing the $I(\mathbf{z}; \mathbf{c})$ term, while itself, 2) giving an extra constrain on the classification task further reduces search space of $\mathbf{z}$.

4.5. Parameter Analysis

4.5.1. Effect of β

β serves as an trade-off parameter in in Eq. 3. To evaluate the effect of β, we vary it within the range of $[1e-4, 1]$. The F1 and $\Delta_{DP}(\mathcal{A}, \mathbf{C})$ *vs.* β plots are depicted in Fig 2. The x axis is plotted in log-scale. It can be found that with β increases, better $\Delta_{DP}(\mathcal{A}, \mathbf{C})$ can be obtained. Such a transition can be sensitive to β around 0.1.

4.5.2. Effect of Model Depth and Embedding Dimension

We now evaluate the influence of depth of the $I(\mathbf{z}; \mathbf{x}|\mathbf{c}_i)$ estimation network and embedding dimension of representation. With β fixed as 0.05, the results are shown in Appendix D Tab. D.3. **G-{1,2}-D-{1,2}** denotes the depth of generator and discriminator, and d denotes the embedding dimension

of representations, which is chosen from $\{16, 32, 64\}$. We can observe that different decision algorithms have a different preference on the model parameters and **G**-$\{\mathbf{1}\}$-**D**-$\{\mathbf{1}\}$ and d=16 is a relative fair choice that suits all.

4.6. Why OTIS Works: Case Study on A Toy Dataset

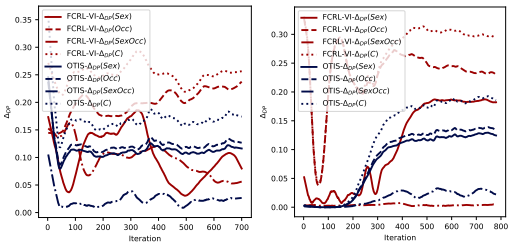
[Contribution: Yue Cui] To further investigate why OTIS works, we now look into the optimization trend over each individual sensitive attribute by visualizing corresponding parity *vs.* iteration. For simplicity, we evaluate the models on the toy setting that only attributes Sex and Occupation_Sales (Occ) of the AdultCensus dataset are considered as original sensitive attributes, *i.e.,* with the only combined sensitive attribute Sex-Occupation_Sales (SexOcc), there are three attributes in $\mathbf{C}$. We visualize the training parity *vs.* iteration for each attribute and $\mathbf{C}$ in Fig. 3. Baseline FCRL-VI is also included for comparison. Results on two training paradigms are reported. To make the results comparable, we tune hyper-parameters to ensure the methods reach similar F1 ~ 0.77 by the time of convergence.

Pre-training based: We omit the pre-training epochs here. As depicted in Fig. 3(a), generally, parity in all settings tends to go down as training iterations increase. With the VI-based training strategy, there is a fluctuation between parity of the three attributes, which can be problematic as it slows downs the training and leads to unsatisfying converged results. This is in line with the analysis in Introduction Section. Besides that the proposed method reaches the peaks at an early stage and achieves much better performance, it shows a much smoother curve. The trends of all attributes are also coherent in OTIS, without the phenomenon that one suffering from severe negative side-effect when the model is getting fairer towards another. We infer this is because with the proposed objective (Eq. 3), each backprop step of OTIS is taken with parities of all attributes in $\mathbf{C}$ considered.

Joint-training based: We can observe from Fig. 3(b) that the trend for joint-training is generally opposite to that of pre-training. Parity starts at relatively low values and goes up as training iterations increase. This is because the model initializes as a random guesser and tends to have no bias at the beginning. Similar fluctuation can be observed from FCRL-VI while the proposed approach OTIS still converges and achieves better performance earlier, which further verifies the effectiveness and efficiency of OTIS.

5. Related Works

[Contribution: Qichen Tan] Fairness-aware machine learning algorithms have attracted increasing attention. Based on the evaluation object, current researches mainly focus on three aspects: 1) individual fairness (Patro et al., 2020; Dwork et al., 2012; Kusner et al., 2017; Garcia-



(a) Pre-training based (b) Joint-training based

Figure 3. Parity *w.r.t.* individual attributes Sex, Occupation_Sales and their combination on the AdultCensus dataset.

Soriano & Bonchi, 2021; Dong et al., 2021), which gives similar prediction to similar individuals; 2) group fairness (Rios, 2020; Dwork et al., 2012; Kusner et al., 2017), which requires similar treatment for protected and unprotected groups; and 3) subgroup fairness (Kearns et al., 2019; Du et al., 2020), which intends to obtain the best properties of both the group and individual notions of fairness. Various methods have been proposed for debiasing, based on the process phase, there are: 1) prep-process based (Kamiran & Calders, 2012; Zemel et al., 2013; Feldman et al., 2015; Calmon et al., 2017; Choi et al., 2020; Jiang & Nachum, 2020), which transforms or re-weights the training data to remove underlying discrimination; 2) in-processing based(Kamishima et al., 2012; Zafar et al., 2017b;a; Agarwal et al., 2018; Zhang et al., 2018; Cotter et al., 2019; Roh et al., 2020a), which adopts model tuning or regularization methods to tailor model training; and 3) post-processing based (Kamiran et al., 2012; Dwork et al., 2012; Pleiss et al., 2017; Chzhen et al., 2019), which processes the model output with certain algorithms or accesses external datasets that are not used for training for fairness. While these methods are confined to debiasing *w.r.t.* pre-defined protected groups, they shed light on designing models that are fair to all sensitive attributes and their combinations. Our work falls into the categories of group fairness and in-processing.

6. Conclusion

[Contribution: Yue Cui] In this paper, we introduce a novel but practical open target debiasing (OTD) setting, where a learning algorithm is expected to achieve parity over all original sensitive attributes and their combinations. To solve the challenge of an exponential number of potential protected groups and provide stable solutions with grantees, we propose an open target fair representation learning framework, named OTIS. We show that via the lens of mutual information, demographic parity can be guaranteed in OTIS. A thorough evaluation on two public real-world datasets demonstrates that compared with baselines, OTIS can robustly achieve good parity with reasonable classification accuracy on a wide range of downstream classification models.

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. A reductions approach to fair classification. In *International Conference on Machine Learning*, pp. 60–69. PMLR, 2018.

Alajaji, F. and Chen, P.-N. *An Introduction to Single-User Information Theory*. Springer, 2018.

Belghazi, M. I., Baratin, A., Rajeshwar, S., Ozair, S., Bengio, Y., Courville, A., and Hjelm, D. Mutual information neural estimation. In *International Conference on Machine Learning*, pp. 531–540. PMLR, 2018.

Calmon, F. P., Wei, D., Vinzamuri, B., Ramamurthy, K. N., and Varshney, K. R. Optimized pre-processing for discrimination prevention. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 3995–4004, 2017.

Choi, K., Grover, A., Singh, T., Shu, R., and Ermon, S. Fair generative modeling via weak supervision. In *International Conference on Machine Learning*, pp. 1887–1898. PMLR, 2020.

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. Leveraging labeled and unlabeled data for consistent fair binary classification. *arXiv preprint arXiv:1906.05082*, 2019.

Cotter, A., Jiang, H., and Sridharan, K. Two-player games for efficient non-convex constrained optimization. In *Algorithmic Learning Theory*, pp. 300–332. PMLR, 2019.

Dong, Y., Kang, J., Tong, H., and Li, J. Individual fairness for graph neural networks: A ranking based approach. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pp. 300–310, 2021.

Du, X., Pei, Y., Duivesteijn, W., and Pechenizkiy, M. Fairness in network representation by latent structural heterogeneity in observational data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 3809–3816, 2020.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268, 2015.

Friedman, J., Hastie, T., Tibshirani, R., et al. *The elements of statistical learning*, volume 1. Springer series in statistics New York, 2001.

Garcia-Soriano, D. and Bonchi, F. Maxmin-fair ranking: Individual fairness under group-fairness constraints. *arXiv preprint arXiv:2106.08652*, 2021.

Gitiaux, X. and Rangwala, H. Fair representations by compression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 11506–11515, 2021.

Gupta, U., Ferber, A., Dilkina, B., and Ver Steeg, G. Controllable guarantees for fair outcomes via contrastive information estimation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 7610–7619, 2021.

Jaiswal, A., Moyer, D., Ver Steeg, G., AbdAlmageed, W., and Natarajan, P. Invariant representations through adversarial forgetting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 4272–4279, 2020.

Jiang, H. and Nachum, O. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 702–712. PMLR, 2020.

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

Kamiran, F., Karim, A., and Zhang, X. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*, pp. 924–929. IEEE, 2012.

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

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. An empirical study of rich subgroup fairness for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 100–109, 2019.

Kohavi, R. et al. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *KDD*, volume 96, pp. 202–207, 1996.

Kusner, M. J., Loftus, J. R., Russell, C., and Silva, R. Counterfactual fairness. *arXiv preprint arXiv:1703.06856*, 2017.

Lackner, M. Perpetual voting: Fairness in long-term decision making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 2103–2110, 2020.

McNamara, D., Ong, C. S., and Williamson, R. C. Provably fair representations. *arXiv preprint arXiv:1710.04394*, 2017.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

Molavipour, S., Bassi, G., and Skoglund, M. Conditional mutual information neural estimator. In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 5025–5029. IEEE, 2020.

Moyer, D., Gao, S., Brekelmans, R., Steeg, G. V., and Galstyan, A. Invariant representations without adversarial training. *arXiv preprint arXiv:1805.09458*, 2018.

Mukherjee, S., Asnani, H., and Kannan, S. Ccmi: Classifier based conditional mutual information estimation. In *Uncertainty in artificial intelligence*, pp. 1083–1093. PMLR, 2020.

Nguyen, X., Wainwright, M. J., and Jordan, M. I. Estimating divergence functionals and the likelihood ratio by convex risk minimization. *IEEE Transactions on Information Theory*, 56(11):5847–5861, 2010.

Paninski, L. Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253, 2003.

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

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., and Weinberger, K. Q. On fairness and calibration. *arXiv preprint arXiv:1709.02012*, 2017.

Poole, B., Ozair, S., Van Den Oord, A., Alemi, A., and Tucker, G. On variational bounds of mutual information. In *International Conference on Machine Learning*, pp. 5171–5180. PMLR, 2019.

Rawls, J. *A theory of justice*. 1971.

Rios, A. Fuzze: Fuzzy fairness evaluation of offensive language classifiers on african-american english. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pp. 881–889, 2020.

Roh, Y., Lee, K., Whang, S., and Suh, C. Fr-train: A mutual information-based approach to fair and robust training. In *International Conference on Machine Learning*, pp. 8147–8157. PMLR, 2020a.

Roh, Y., Lee, K., Whang, S. E., and Suh, C. Fairbatch: Batch selection for model fairness. *arXiv preprint arXiv:2012.01696*, 2020b.

Sharifi-Malvajerdi, S., Kearns, M., and Roth, A. Average individual fairness: Algorithms, generalization and experiments. *Advances in Neural Information Processing Systems*, 32:8242–8251, 2019.

Song, J., Kalluri, P., Grover, A., Zhao, S., and Ermon, S. Learning controllable fair representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 2164–2173. PMLR, 2019.

Tishby, N., Pereira, F. C., and Bialek, W. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

Vermeule, A. Veil of ignorance rules in constitutional law. *Yale LJ*, 111:399, 2001.

Wu, T., Ren, H., Li, P., and Leskovec, J. Graph information bottleneck. *arXiv preprint arXiv:2010.12811*, 2020.

Zafar, M. B., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th international conference on world wide web*, pp. 1171–1180, 2017a.

Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pp. 962–970. PMLR, 2017b.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning fair representations. In *International conference on machine learning*, pp. 325–333. PMLR, 2013.

Zhang, B. H., Lemoine, B., and Mitchell, M. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 335–340, 2018.

Zhu, Z., Hu, X., and Caverlee, J. Fairness-aware tensor-based recommendation. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pp. 1153–1162, 2018.

A. Proofs

A.1. Proof of Theorem 3.1

Theorem A.1. ***(Theorem 3.1 restated)*** *For $z, c_i \sim p(\mathbf{z}, \mathbf{c}_i)$, $z \in \mathbb{R}^d$, $\mathbf{c}_i \in \mathbf{C}$, $c_i \in \{1, ..., N_i\}$, and any decision algorithm $\mathcal{A}$ that acts on z, we have*

$$\sum_{i=1}^{|\mathbf{C}|} I(\mathbf{z}; \mathbf{c}_i) \geq g(\gamma, \Delta_{DP}(\mathcal{A}, \mathbf{C}))$$

where $\gamma = \min_i \alpha_i$, $\alpha_i = \min_j \pi_{i,j}$, $\pi_{i,j} = P(\mathbf{c}_i = j)$.

We prove Theorem 3.1 by first introducing the following Lemma:

Lemma A.2. *For some $z, c \sim p(\mathbf{z}, \mathbf{c})$, where c is any sensitive attribute $z \in \mathbb{R}^d, c \in \{1, ..., N\}$, and any decision algorithm $\mathcal{A}$ that acts on z, we have*

$$I(\mathbf{z}; \mathbf{c}) \geq \alpha^2 \Delta_{DP}(\mathcal{A}, \mathbf{c})\big)^2 \tag{12}$$

where $\alpha = \min \pi_i$, $\pi_i = P(\mathbf{c} = i)$.

Lemma A.2 is heavily inspired by (Gupta et al., 2021), which indicates that the mutual information bounds parity of any singular attribute.

The proof of Lemma A.2 is offered as follows. Note that Theorem 3.1 can not be straightforwardly obtained by respectively adding the LHS and RHS of Eq. 12 over i.

Proof of Lemma A.2. From the definition of $\Delta_{DP}(\mathcal{A}, \mathbf{c})$, we can get

$$\begin{aligned}
\Delta_{DP}(\mathcal{A}, \mathbf{c}) &= \max_{i,j} \Big| P(\hat{\mathbf{y}} = 1 | \mathbf{c} = i) - P(\hat{\mathbf{y}} = 1 | \mathbf{c} = j) \Big| \\
&= \max_{i,j} \left| \int_z dz P(\hat{\mathbf{y}} = 1 | \mathbf{z}) p(\mathbf{z} | \mathbf{c} = i) - P(\hat{\mathbf{y}} = 1 | \mathbf{z}) p(\mathbf{z} | \mathbf{c} = j) \right| \\
&\leq \max_{i,j} \int_z dz P(\hat{\mathbf{y}} = 1 | \mathbf{z}) \Big| p(\mathbf{z} | \mathbf{c} = i) - p(\mathbf{z} | \mathbf{c} = j) \Big| \\
&\leq \max_{i,j} \int_z dz \Big| p(\mathbf{z} | \mathbf{c} = i) - p(\mathbf{z} | \mathbf{c} = j) \Big| \\
&= \max_{i,j} \Big\| p(\mathbf{z} | \mathbf{c} = i) - p(\mathbf{z} | \mathbf{c} = j) \Big\| = \max_{i,j} V\big(p(\mathbf{z} | \mathbf{c} = i), p(\mathbf{z} | \mathbf{c} = j)\big)
\end{aligned} \tag{13}$$

where $V(p_1, p_2) = \|p_1 - p_2\|$ is the variational distance between distribution p_1 and p_2. We can also calculate the mutual information between $\mathbf{z}$ and $\mathbf{c}$ as

$$\begin{aligned}
I(\mathbf{z}; \mathbf{c}) &= \mathbb{E}_{\mathbf{z},\mathbf{c}} log \frac{p(\mathbf{z}, \mathbf{c})}{p(\mathbf{z}) p(\mathbf{c})} \\
&= \mathbb{E}_{\mathbf{z},\mathbf{c}} log \frac{p(\mathbf{z} | \mathbf{c})}{p(\mathbf{z})} \\
&= \sum_{i=1}^{N} \pi_i \mathbb{E}_{\mathbf{z}|\mathbf{c}=i} log \frac{p(\mathbf{z} | \mathbf{c} = i)}{p(\mathbf{z})} \\
&= \sum_{i=1}^{N} \pi_i KL\big(p(\mathbf{z} | \mathbf{c} = i) \| p(\mathbf{z})\big)
\end{aligned} \tag{14}$$

According to Pinsker's inequality (Alajaji & Chen, 2018), we know that

$$KL(p_1 \| p_2) \geq 2 \|p_1 - p_2\|^2 = f(V) \tag{15}$$

Thus, we can derive that

$$
\begin{aligned}
I(\mathbf{z};\mathbf{c}) &= \sum_{i=1}^{N} \pi_i KL\big(p(\mathbf{z}|\mathbf{c}=i)\|p(\mathbf{z})\big) \\
&\geq \sum_{i=1}^{N} \pi_i f\Big(V\big(p(\mathbf{z}|\mathbf{c}=i), p(\mathbf{z})\big)\Big) \\
&\geq f\Big(\sum_{i=1}^{N} \pi_i V\big(p(\mathbf{z}|\mathbf{c}=i), p(\mathbf{z})\big)\Big) \\
&\geq f\big(\max_{i,j} \pi_i \|p(\mathbf{z}|\mathbf{c}=i) - p(\mathbf{z})\| + \pi_j \|p(\mathbf{z}|\mathbf{c}=j) - p(\mathbf{z})\|\big) \\
&\geq f\big(\alpha \max_{i,j} \|p(\mathbf{z}|\mathbf{c}=i) - p(\mathbf{z}|\mathbf{c}=j)\|\big) \\
&\geq f\big(\alpha \max_{i,j} |P(\hat{\mathbf{y}}=1|\mathbf{c}=i) - P(\hat{\mathbf{y}}=1|\mathbf{c}=j)|\big) \\
&= f\big(\alpha \Delta_{DP}(\mathcal{A}, \mathbf{c})\big)
\end{aligned}
\tag{16}
$$

f is convex function and also strictly increasing, non-negative. This completes the proof. □

Equipped with Lemma A.2, we can then prove Theorem 3.1 as follows,

Proof of Theorem 3.1. We can obtain from Eq. 16 and the definition of $\Delta_{DP}(\mathcal{A}, \mathbf{C})$ that,

$$
\begin{aligned}
\sum_{i=1}^{|\mathbf{C}|} I(\mathbf{z};\mathbf{c}_i) &= \sum_{i=1}^{|\mathbf{C}|} \sum_{j=1}^{N_i} \pi_{i,j} KL\big(p(\mathbf{z}|\mathbf{c}_i=j)\|p(\mathbf{z})\big) \\
&\geq \sum_{i=1}^{|\mathbf{C}|} f\big(\alpha \Delta_{DP}(\mathcal{A}, \mathbf{c}_i)\big) \\
&= \sum_{i=1}^{|\mathbf{C}|} {\alpha_i}^2 \Delta_{DP}(\mathcal{A}, \mathbf{c}_i)^2 \\
&\geq \gamma \sum_{i=1}^{|\mathbf{C}|} \Delta_{DP}(\mathcal{A}, \mathbf{c}_i)^2 \\
&= g(\gamma, \Delta_{DP}(\mathcal{A}, \mathbf{C}))
\end{aligned}
\tag{17}
$$

where $\gamma = \min_i \alpha_i^2$. $g(x, y) = xy^2$, which is strictly increasing, non-negative, and convex. This completes the proof. □

A.2. Proof of Theorem 3.2

Theorem A.3. ***(Theorem 3.2 restated)*** *For a combined sensitive attribute $\mathbf{c}_{ij}$, which is formed by sensitive attribute $\mathbf{c}_i$ and $\mathbf{c}_j$, where $c_i \in \{1, .., N_i\}, c_j \in \{1, ..., N_j\}$, we have*

$$
\min\{I(\mathbf{z};\mathbf{c}_i), I(\mathbf{z};\mathbf{c}_j)\} \geq I(\mathbf{z};\mathbf{c}_{ij})
$$

Proof.

$$
\begin{aligned}
I(\mathbf{z};\mathbf{c}_{ij}) &= \mathbb{E}_{\mathbf{z},\mathbf{c}_{ij}} log \frac{p(\mathbf{z},\mathbf{c}_{ij})}{p(\mathbf{z})p(\mathbf{c}_{ij})} \\
&= \mathbb{E}_{\mathbf{z},\mathbf{c}_{ij}} log \frac{p(\mathbf{z}|\mathbf{c}_{ij})}{p(\mathbf{z})} \\
&= \sum_{x \in \mathbf{N}_i, y \in \mathbf{N}_j} P(\mathbf{c}_i = x, \mathbf{c}_j = y) \mathbb{E}_{\mathbf{z}|\mathbf{c}_i=x,\mathbf{c}_j=y} log \frac{p(\mathbf{z}|\mathbf{c}_i = x, \mathbf{c}_j = y)}{p(\mathbf{z})}
\end{aligned}
\tag{18}
$$

where $\mathbf{N}_i = \{1, ..., N_i\}, \mathbf{N}_j = \{1, ..., N_j\}$. Similarly, we also have,

$$\begin{aligned} I(\mathbf{z}; \mathbf{c}_i) &= \sum_{x \in \mathbf{N}_i} P(\mathbf{c}_i = x) \mathbb{E}_{\mathbf{z}|\mathbf{c}_i = x} log \frac{p(\mathbf{z}|\mathbf{c}_i = x)}{p(\mathbf{z})} \\ &= \sum_{x \in \mathbf{N}_i} \big(\sum_{y \in \mathbf{N}_j} P(\mathbf{c}_i = x, \mathbf{c}_j = y) \big) \mathbb{E}_{\mathbf{z}|\mathbf{c}_i = x} log \frac{p(\mathbf{z}|\mathbf{c}_i = x)}{p(\mathbf{z})} \\ &\geq \sum_{x \in \mathbf{N}_i} \sum_{y \in \mathbf{N}_j} P(\mathbf{c}_i = x, \mathbf{c}_j = y) \mathbb{E}_{\mathbf{z}|\mathbf{c}_i = x, \mathbf{c}_j = y} log \frac{p(\mathbf{z}|\mathbf{c}_i = x, \mathbf{c}_j = y)}{p(\mathbf{z})} \\ &= I(\mathbf{z}; \mathbf{c}_{ij}) \end{aligned} \tag{19}$$

Similarly, we can also get

$$I(\mathbf{z}; \mathbf{c}_j) \geq I(\mathbf{z}; \mathbf{c}_{ij}) \tag{20}$$

That is to say

$$\min\{I(\mathbf{z}; \mathbf{c}_i), I(\mathbf{z}; \mathbf{c}_j)\} \geq I(\mathbf{z}; \mathbf{c}_{ij}) \tag{21}$$

Generalizing the equality to higher-order interacted sensitive features is straightforward and this completes the proof.

□

B. Algorithm to Compute L_c

Algorithm 1 Compute L_c

Input:
/* Assume representation $\mathbf{z}$ has been obtained in previous steps.*/
Batch data $\mathcal{B} = (\mathbf{z}, \mathbf{x}, \mathbf{c})$, inner iteration T;
Output: L_c;
for $t \in \{1, ..., T\}$ **do**
  Randomly split $\mathcal{B}$ equally to two parts $\mathcal{B}_{joint} = \{z_j, x_j, c_j\}_{j=0}^{n/2}$ and $\mathcal{B}_{gen} = \{z_j, x_j, c_j\}_{j=n/2}^{n}$;
  Compute $z'_j = NN_G(c_j) \in \mathbb{R}^{|\mathbf{C}_0|}, \forall c_j \in \mathcal{B}_{joint}$ and obtain $\mathcal{B}_{prod} = \{z'_j, x_j, c_j\}_{j=0}^{n/2}$;
  Compute $L_{c_i,t}$ with $\{\mathcal{B}_{prod}, \mathbf{1}\}$, $\{\mathcal{B}_{prod}, \mathbf{0}\}$
end for
Monte Carlo average $L_c = \sum_i Avg_t(L_{c_i,t})$

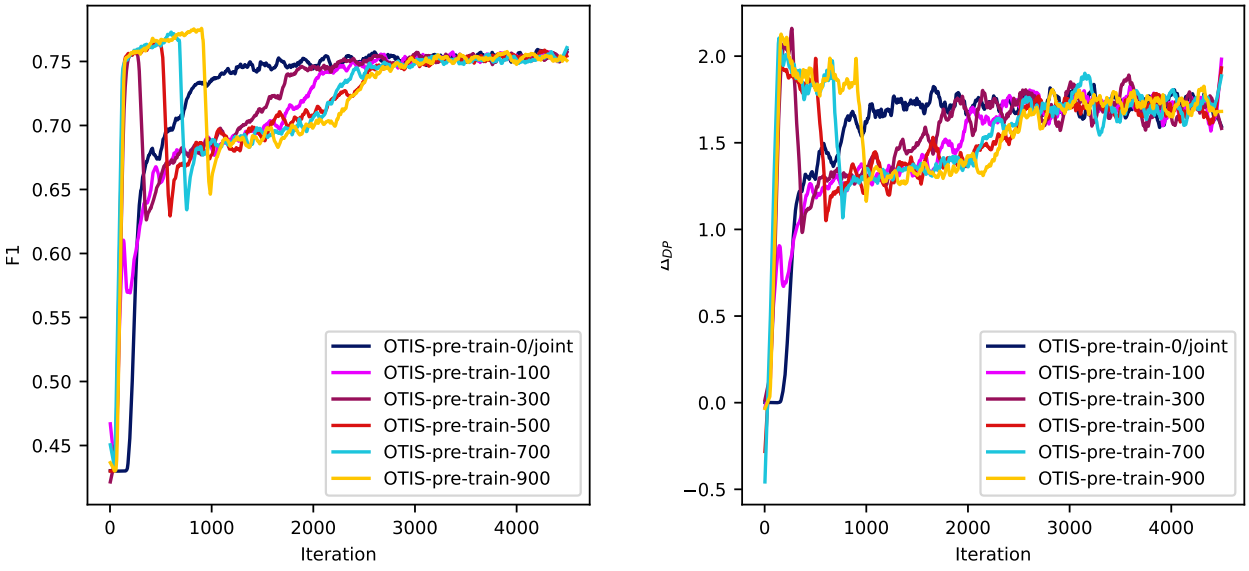
C. Veil of Ignorance

The veil of ignorance (VI), *aka.* the original position, is proposed by (Rawls, 1971) as a thought experiment. The idea of VI is to avoid self-interest behavior of decision-makers by subjecting the decision-maker to uncertainty about the distribution of benefits and burdens that will result from a decision (Vermeule, 2001). Specifically, (Rawls, 1971) assumes that citizens making decisions about the basic structure of society are required to sit behind a "veil of ignorance", where they do not know the status, gender, race, abilities, or occupation that they will have in the society. Without knowing if a rule will affect their own cases, they tend to choose fair rules that will at least benefit the most disadvantaged group.

D. Other Results

D.1. Results for Training Paradigms Analysis

The parity and F1 *vs.* iteration plots on the training set of AdultCensus can be found in Fig. 4.



(a) F1 (b) $\Delta_{DP}(\mathcal{A}, \mathbf{C})$

Figure 4. OTIS on AdultCensus dataset with different number of pre-training epochs.

D.2. Results and Discussions on The F1 - $\Delta_{DP}(\mathcal{A}, \mathbf{C})$ Trade-off

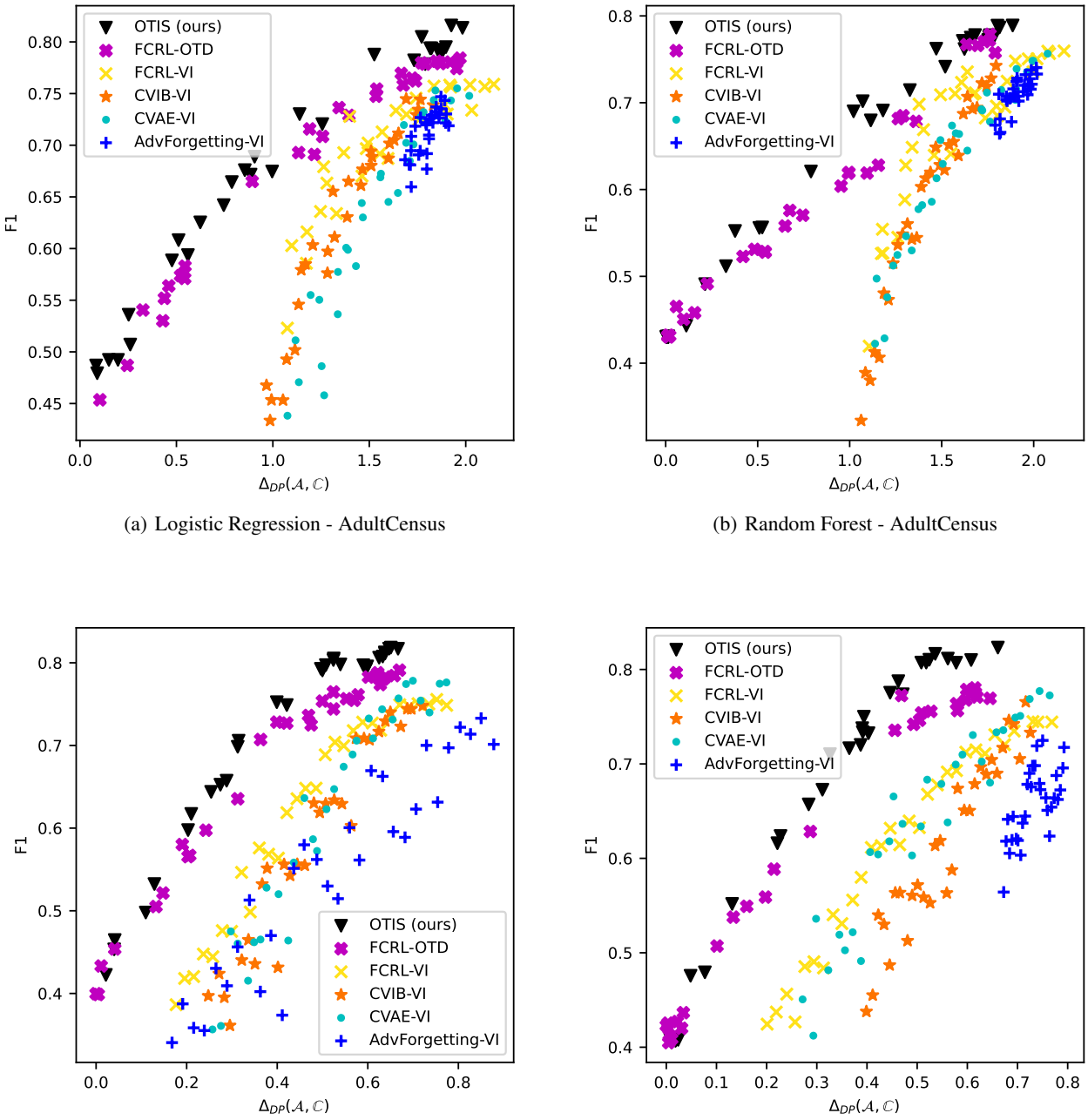
We here present further observations and discussions on the F1 - $\Delta_{DP}(\mathcal{A}, \mathbf{C})$ trade-off experiment. It can observe from Fig. 1 and Fig. 5 that the variant of FCRL-OTD, which is the FCRL with the same fairness constraints as OTIS, fails to achieve good parity. This is probably because it uses $I(\mathbf{y}; \mathbf{z}|\mathbf{c})$ as the objective of decoder for representation learning. Since in OTD $\mathbf{c}$ contains much more attributes and can be more likely to be explicitly (included) or implicitly (correlated) related to $\mathbf{x}$, and therefore $\mathbf{z}$. Using such an objective reduces information for an expressive-enough representation unless making a compromise on the fairness objective. Despite that OTIS takes the lead on all downstream decision algorithms and all datasets, which indicates the proposed method is model-agnostic, we find that CVAE-VI outperforms FCRL-VI significantly on AdaBoost and obtains comparable performance on Random Forest. This may indicate the advantage of using reconstruction-based method to achieve fairness on ensemble learning algorithms.

D.3. Results for Model Depth and Embedding Dimension Analysis

Tab. D.3 compares different depth of the $I(\mathbf{z}; \mathbf{x}|\mathbf{c}_i)$ estimation network and embedding dimension of representation on the AdultCensus dataset.

Method		MLP		Logistirc Regression		AdaBoost		Random Forest	
Depth	d	F1	DP	F1	DP	F1	DP	F1	DP
G-1-C-1	16	0.7788	1.6657	0.7833	1.9117	0.7805	1.9234	0.7772	1.8379
	32	0.7733	1.6299	0.7854	1.8363	0.7842	1.8874	0.7834	1.8160
	64	0.7490	1.5115	0.7845	1.7719	0.7807	1.7839	0.7811	1.7878
G-2-C-1	16	0.7733	1.6861	0.7811	1.8587	0.7767	1.8169	0.7706	1.8254
	32	0.7669	1.5894	0.7826	1.7130	0.7835	1.7599	0.7792	1.7685
	64	0.7574	1.5910	0.7841	1.7426	0.7786	1.7277	0.7806	1.7166
G-1-C-2	16	0.7783	1.7510	0.7814	1.8665	0.7816	1.8623	0.7796	1.8627
	32	0.7597	1.5777	0.7808	1.8360	0.7775	1.8492	0.7772	1.8198
	64	0.7554	1.7128	0.7822	1.8390	0.7781	1.8640	0.7830	1.8793
G-2-C-2	16	0.7709	1.7062	0.7818	1.7661	0.7797	1.7807	0.7775	1.7739
	32	0.7637	1.5290	0.7823	1.8077	0.7811	1.7742	0.7799	1.7937
	64	0.7508	1.6850	0.7843	1.7857	0.7784	1.8255	0.7713	1.7848

Table 4. Effect of model depth and embedding dimension.



(a) Logistic Regression - AdultCensus

(b) Random Forest - AdultCensus

(c) Logistic Regression - Health Heritage

(d) Random Forest- Health Heritage

Figure 5. Trade-off between F1 and $\Delta_{DP}(\mathcal{A}, \mathbf{C})$ on the AdultCensus and Health Heritage dataset with Logistic Regression and Random Forest.