INFOFAIR: Information-Theoretic Intersectional Fairness

Jian Kang[1], Tiankai Xie[2], Xintao Wu[3], Ross Maciejewski[2], and Hanghang Tong[1]

[1]University of Illinois Urbana-Champaign, {jiank2, htong}@illinois.edu
[2]Arizona State University, {txie21, rmacieje}@asu.edu
[3]University of Arkansas, xintaowu@uark.edu

***Abstract*—Algorithmic fairness is becoming increasingly important in data mining and machine learning. Among others, a foundational notation is *group fairness*. The vast majority of the existing works on group fairness, with a few exceptions, primarily focus on debiasing with respect to a single sensitive attribute, despite the fact that the co-existence of multiple sensitive attributes (e.g., gender, race, marital status, etc.) in the real-world is commonplace. As such, methods that can ensure a fair learning outcome with respect to all sensitive attributes of concern simultaneously need to be developed. In this paper, we study the problem of information-theoretic intersectional fairness (INFOFAIR), where statistical parity, a representative group fairness measure, is guaranteed among demographic groups formed by multiple sensitive attributes of interest. We formulate it as a mutual information minimization problem and propose a generic end-to-end algorithmic framework to solve it. The key idea is to leverage a variational representation of mutual information, which considers the variational distribution between learning outcomes and sensitive attributes, as well as the density ratio between the variational and the original distributions. Our proposed framework is generalizable to many different settings, including other statistical notions of fairness, and could handle any type of learning task equipped with a gradient-based optimizer. Empirical evaluations in the fair classification task on three real-world datasets demonstrate that our proposed framework can effectively debias the classification results with minimal impact to the classification accuracy.**

***Index Terms*—Group fairness, mutual information, intersectional fairness**

I. INTRODUCTION

The increasing amount of data and computational power have empowered machine learning algorithms to play crucial roles in automated decision-making for a variety of real-world applications, including credit scoring [1], criminal justice [2] and healthcare analysis [3]. As the application landscape of machine learning continues to broaden and deepen, so does the concern regarding the potential, often unintentional, bias it could introduce or amplify. For example, recent media coverage has revealed that a well-trained image generator could turn a low-resolution picture of a black man into a high-resolution image of a white man due to the skewed data distribution that causes the model to disfavor the minority group,[1] and another article highlighted an automated credit card application system assigning a dramatically higher credit limit to a man than to his female partner, even though his partner has a better credit history.[2]

As such, algorithmic fairness, which aims to mitigate unintentional bias caused by automated learning algorithms, has become increasingly important. To date, researchers have proposed a variety of fairness notions [4], [5]. Among them, one of the most fundamental notions is *group fairness*.[3] Generally speaking, to ensure group fairness, the first step is to partition the entire population into a few demographic groups based on a pre-defined sensitive attribute (e.g., gender). Then the fair learning algorithm will enforce parity of a certain statistical measure among those demographic groups. Group fairness can be instantiated with many statistical notions of fairness. Statistical parity [6] enforces the learned classifier to accept equal proportion of population from the pre-defined majority group and minority group. Likewise, disparate impact [5] ensures the acceptance rate for the minority group should be no less than four-fifth of that for the group with the highest acceptance rate, which is analogous to the famous 'four-fifth' rule in the legal support area [7]. In addition, equalized odds and equal opportunity [8] are used to enforce the classification accuracies to be equal across all demographic groups conditioned on ground-truth outcomes or positively labeled populations, respectively. The vast majority of the existing works in group fairness primarily focus on debiasing with respect to a single sensitive attribute. However, it is quite common for multiple sensitive attributes (e.g., gender, race, marital status, etc.) to co-exist in a real-world application. We ask: *would a debiasing algorithm designed to ensure the group fairness for a particular sensitive attribute (e.g., marital status) unintentionally amplify the group bias with respect to another sensitive attribute (e.g., gender)? If so, how can we ensure a fair learning outcome with respect to all sensitive attributes of concern simultaneously?*

Existing works for answering these questions [5], [6], [9], [10] have two major limitations. The first limitation is that some existing works could only debias multiple *distinct*

[1]https://www.theverge.com/21298762/face-depixelizer-ai-machine-learning-tool-pulse-stylegan-obama-bias

[2]https://www.nytimes.com/2019/11/10/business/Apple-credit-card-investigation.html

[3]An orthogonal work in algorithmic fairness is individual fairness. Although it promises fairness by 'treating similar individuals similarly' in principle, it is often hard to be operationalized in practice due to its strong assumption on distance metrics and data distributions.

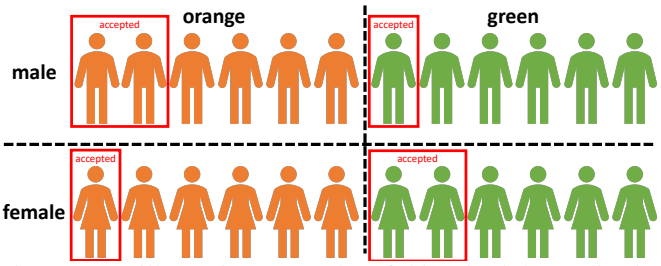

Fig. 1: An illustrative example of bias in job application classification when considering multiple sensitive attributes. Rows indicate gender (e.g., male vs. female) and columns indicate race (e.g., orange vs. green).[4]Boxed individuals receive job offers. If we consider gender or race alone, statistical parity is enforced due to the equal acceptance rate. However, when considering gender and race (i.e., forming finer-grained gender-race groups), the classification result is biased in the fine-grained gender-race groups. This is because, the acceptance rates in two fine-grained groups (i.e., male-green group and female-orange) are lower than that of the two other fine-grained groups (i.e., male-orange and female-green).

sensitive attributes [10], which fails to mitigate bias on the fine-grained groups formed by multiple sensitive attributes. Figure 1 provides an illustrative example of the difference between fairness with respect to multiple distinct sensitive attributes and fairness among fine-grained groups of multiple sensitive attributes. The second limitation is that the optimization problems behind some existing works are often subject to surrogate constraints of statistical parity [5], [6], [9] instead of directly optimizing statistical parity itself, resulting in unstable performance on bias mitigation unless the learned models could perfectly model the relationship between the training data and the ground-truth outcome.

In this paper, we tackle these two limitations by studying the problem of *information-theoretic intersectional fairness* (INFOFAIR), which aims to directly enforce statistical parity on multiple sensitive attributes simultaneously. Though our focused fairness notion is statistical parity, the proposed method can be generalized to other statistical fairness notions (e.g., equalized odds and equal opportunity) with minor modifications. The key idea in solving the INFOFAIR problem is to consider all sensitive attributes of interest as a vectorized sensitive attribute in order to partition the demographic groups and then minimize the dependence between learning outcomes and this vectorized attribute. More specifically, we measure the dependence using mutual information originated in information theory [11]. Building upon it, we formulate the INFOFAIR problem as an optimization problem regularized on mutual information minimization.

The main contributions of this paper are as follows.

- **Problem Definition.** We formally define the problem of information-theoretic intersectional fairness and formulate it as an optimization problem, where the key idea is to minimize both the task-specific loss function (e.g., cross-entropy loss in classification) and mutual information between learning outcomes and the vectorized sensitive attribute.
- **End-to-End Algorithmic Framework.** We propose a novel end-to-end bias mitigation framework, named INFOFAIR, by optimizing a variational representation of mutual information. The proposed framework is extensible and capable of solving any learning task with a gradient-based optimizer.
- **Empirical Evaluations.** We perform empirical evaluations in the fair classification task on three real-world datasets. The evaluation results demonstrate that our proposed framework can effectively mitigate bias with little sacrifice in the classification accuracy.

[4]We use imaginary race groups to avoid potential offenses.

II. PROBLEM DEFINITION

In this section, we present a table of the main symbols used in this paper. Then, we briefly review the concepts of statistical parity and mutual information, as well as their relationships. Finally, we formally define the problem of information-theoretic intersectional fairness.

TABLE I: Table of symbols.

Symbols	Definitions
$\mathcal{D}$	a set
$\mathbf{W}$	a matrix
$\mathbf{h}$	a vector
$\mathbf{h}[i]$	the i-th element in $\mathbf{h}$
$\Pr(\cdot)$	the probability of an event happening
$p_{\cdot,\cdot}$	joint distribution of two random variables
$p_{\cdot}$	marginal distribution of a random variable
$H(\cdot)$	entropy
$H(\cdot\|\cdot)$	conditional entropy
$I(\cdot,\cdot)$	mutual information

In this paper, matrices are denoted by bold uppercase letters (e.g., $\mathbf{X}$), vectors are denoted by bold lowercase letters (e.g., $\mathbf{y}$), scalars are denoted by italic lowercase letters (e.g., c) and sets are denoted by calligraphic letters (e.g., $\mathcal{D}$). We use superscript ^{T} to denote transpose (e.g., $\mathbf{h}^T$ is the transpose of $\mathbf{h}$) and superscript $\mathcal{C}$ to denote the complement of a set (e.g., set $\mathcal{D}^{\mathcal{C}}$ is the complement of set $\mathcal{D}$). We use a convention similar to NumPy for vector indexing (e.g., $\mathbf{h}[i]$ is the i-th element in vector $\mathbf{h}$).

A. Preliminaries

Statistical Parity is one of the most intuitive and widely-used group fairness notions. Given a set of data points $\mathcal{X}$, their corresponding labels $\mathbf{y}$ and a sensitive attribute s, classification with statistical parity aims to learn a classifier to predict outcomes that (1) are as accurate as possible with respect to $\mathbf{y}$ and (2) do not favor one group over another with respect to s. Mathematically, statistical parity is defined as follows.

Definition 1: (Statistical Parity [6]). Suppose we have (1) a population $\mathcal{X}$, (2) a hypothesis $h : \mathcal{X} \to \{0, 1\}$ which assigns a binary label to individual x drawn from $\mathcal{X}$ and (3) a sensitive attribute which splits the population $\mathcal{X}$ into majority group $\mathcal{M}$ and minority group $\mathcal{M}^{\mathcal{C}}$ (i.e., $\mathcal{X} = \mathcal{M} \cup \mathcal{M}^{\mathcal{C}}$). An individual x is accepted if $h(x) = 1$ and rejected if $h(x) = 0$. The hypothesis $h : \mathcal{X} \to \{0, 1\}$ is said to have statistical parity on the population $\mathcal{X}$ as long as

$$\Pr[h(x) = 1 | x \in \mathcal{M}] = \Pr[h(x) = 1 | x \in \mathcal{M}^{\mathcal{C}}] \quad (1)$$

where $\Pr[\cdot]$ denotes the probability of an event happening.

Many methods have been proposed to achieve statistical parity. For example, Zemel et al. [12] learn fair representation by regularizing the difference in expected positive rate for majority and minority groups. Zhang et al. [13] propose an adversarial learning-based framework for fair classification, in which the output of the predictor is used to predict the sensitive attribute by the adversary. Kearns et al. [9] propose a learner-auditor framework to enforce subgroup fairness through fictitious play strategy.

Mutual Information was first introduced in the 1940s [11]. Given two random variables, mutual information measures the dependence between them by quantifying the amount of information in bits obtained on one random variable through observing the other one. Let (x, y) be a pair of random variables x and y. Suppose their joint distribution is $p_{x,y}$ and the marginal distributions are p_x and p_y. The mutual information between x and y is defined as

$$I(x;y) = H(x) - H(x|y) = \int_x \int_y p_{x,y} \log \frac{p_{x,y}}{p_x p_y} dxdy \tag{2}$$

where $H(x) = -\int_x p_x \log p_x dx$ is the entropy of x and $H(x|y) = -\int_x \int_y p_{x,y} \log p_{x|y} dxdy$ is the conditional entropy of x given y. Unlike correlation coefficients (e.g., Pearson's correlation coefficient) which could only capture the linear dependence between two random variables, mutual information is more general in capturing both the linear and nonlinear dependence between two random variables. We have $I(x;y) = 0$ if and only if two random variables x and y are independent to each other.

According to Lemma 1, there is an equivalence between statistical parity and zero mutual information.

Lemma 1: (Equivalence between statistical parity and zero mutual information [12], [14]). Statistical parity requires a sensitive attribute to be statistically independent to the learning results, which is equivalent to zero mutual information. Mathematically, given a learning outcome $\tilde{\mathbf{y}}$ and the sensitive attribute s, we have

$$\underbrace{p_{\tilde{\mathbf{y}}|s} = p_{\tilde{\mathbf{y}}}}_{\text{statistical parity}} \Leftrightarrow p_{\tilde{\mathbf{y}},s} = p_{\tilde{\mathbf{y}}} p_s \Leftrightarrow \underbrace{I(\tilde{\mathbf{y}}; s) = 0}_{\text{zero mutual information}} \tag{3}$$

Proof: Omitted for brevity. ■

B. Information-Theoretic Intersectional Fairness

In order to generalize Lemma 1 from a single sensitive attribute to a set of sensitive attributes $\mathcal{S} = \{s^{(1)}, \ldots, s^{(k)}\}$, we first introduce the concept of vectorized sensitive attribute $\mathbf{s}$ given $\mathcal{S}$. We define the vectorized sensitive attribute $\mathbf{s} = [s^{(1)}, \ldots, s^{(k)}]$ as a multi-dimensional random variable where each element of $\mathbf{s}$ represents the corresponding sensitive attribute in $\mathcal{S}$ (e.g., $\mathbf{s}[i] = s^{(i)}$ is the i-th sensitive attribute). Based on that, we have the following equivalence. For notational simplicity, we denote $I(\tilde{\mathbf{y}}; s^{(1)}, \ldots, s^{(k)})$, $p_{\tilde{\mathbf{y}}, s^{(1)}, \ldots, s^{(k)}}$ and $p_{s^{(1)}, \ldots, s^{(k)}}$ with $I(\tilde{\mathbf{y}}; \mathbf{s})$, $p_{\tilde{\mathbf{y}},\mathbf{s}}$ and $p_{\mathbf{s}}$, respectively.

$$p_{\tilde{\mathbf{y}}|\mathbf{s}} = p_{\tilde{\mathbf{y}}} \Leftrightarrow p_{\tilde{\mathbf{y}},\mathbf{s}} = p_{\tilde{\mathbf{y}}} p_{\mathbf{s}} \Leftrightarrow I(\tilde{\mathbf{y}}; \mathbf{s}) = 0 \tag{4}$$

Based on Eq. (4), we formally define the problem of information-theoretic intersectional fairness as a mutual information minimization problem.

Problem 1: INFOFAIR: Information-Theoretic Intersectional Fairness

Input: (1) a set of k sensitive attributes $\mathcal{S} = \{s^{(1)}, \ldots, s^{(k)}\}$; (2) a set of n data points $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{s}_i, y_i) | i = 1, \ldots, n\}$ where $\mathbf{x}_i$ is the feature vector of the i-th data point, y_i is its label and $\mathbf{s}_i = [s_i^{(1)}, \ldots, s_i^{(k)}]$ describes the vectorized sensitive attributes on $\mathcal{S}$ of the i-th data point (with $s_i^{(j)}$ being the corresponding attribute value of the j-th sensitive attribute $s^{(j)}$); and (3) a learning algorithm represented by $l(\mathbf{x}; \mathbf{s}; y; \tilde{\mathbf{y}}; \theta)$, where l is the loss function, $\tilde{\mathbf{y}}^* = \text{argmin}_{\tilde{\mathbf{y}}} l(\mathbf{x}; \mathbf{s}; y; \tilde{\mathbf{y}}; \theta)$ is the optimal learning outcome on the input data with θ being model parameters.

Output: a set of revised learning outcomes $\{\tilde{\mathbf{y}}^*\}$ which minimizes (1) the empirical risk $\mathbb{E}_{(\mathbf{x},\mathbf{s},y)\sim\mathcal{D}}[l(\mathbf{x}; \mathbf{s}; y; \tilde{\mathbf{y}}; \theta)]$ and (2) the expectation of mutual information between the learning outcomes and the sensitive attributes $\mathbb{E}_{(\mathbf{x},\mathbf{s},y)\sim\mathcal{D}}\big[I(\tilde{\mathbf{y}}; \mathbf{s})\big]$.

Remark: a byproduct of INFOFAIR is that the statistical parity can also be achieved on any subset of sensitive attributes included in $\mathcal{S}$, which is summarized in Lemma 2. This could be particularly useful in that the algorithm administrator does not need to re-train the model in order to obtain fair learning results if s/he is only interested in a subset of available sensitive attributes.

Lemma 2: Consider statistical parity as the fairness notion. Given a learning outcome $\tilde{\mathbf{y}}$, a set of k sensitive attributes $\mathcal{S} = \{s^{(1)}, \ldots, s^{(k)}\}$ and the vectorized sensitive attribute $\mathbf{s} = [s^{(1)}, \ldots, s^{(k)}]$. If $\tilde{\mathbf{y}}$ is fair with respect to $\mathbf{s}$, then $\tilde{\mathbf{y}}$ is fair with respect to any vectorized sensitive attribute $\mathbf{s}_{\text{sub}}$ induced from the subset of sensitive attributes $\mathcal{S}_{\text{sub}} \subseteq \mathcal{S} = \{s^{(1)}, \ldots, s^{(k)}\}$.

Proof: Omitted for brevity. ■

III. PROPOSED METHOD

In this section, we present a generic end-to-end algorithmic framework, named INFOFAIR, for information-theoretic intersectional fairness. We first formulate the problem as a mutual information minimization problem, and then present a variational representation of mutual information. Based on that, we present the INFOFAIR framework to solve the optimization problem, followed by discussions on generalizations and variants of our proposed framework.

A. Objective Function

Given a dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{s}_i, y_i) | i = 1, \ldots, n\}$, the INFOFAIR problem (Problem 1) can be naturally formulated as minimizing the following objective function,

$$J = \mathbb{E}_{(\mathbf{x},\mathbf{s},y)\sim\mathcal{D}}\big[l(\mathbf{x}; \mathbf{s}; y; \tilde{\mathbf{y}}; \theta) + \alpha I(\tilde{\mathbf{y}}; \mathbf{s})\big] \tag{5}$$

where l is a task-specific loss function for a learning task, θ is the model parameter, $\tilde{\mathbf{y}}$ is the learning outcome and $\alpha > 0$ is the regularization hyperparameter. An example of loss function l is the negative log likelihood

$$l(\mathbf{x}; \mathbf{s}; y; \tilde{\mathbf{y}}; \theta) = -\log \tilde{\mathbf{y}}[y] \tag{6}$$

where y is the class label and $\tilde{\mathbf{y}}$ denotes the probabilities of being classfied into the corresponding class.

To optimize the above objective function, a key challenge lies in optimizing the mutual information between the learning outcome and the vectorized sensitive feature $I(\tilde{\mathbf{y}}; \mathbf{s})$. Inspired by the seminal work of Belghazi et al. [15], a natural choice

would be to apply off-the-shelf mutual information estimation methods for high-dimensional data. Examples include MINE [15], Deep Infomax [16] and CCMI [17], which estimate mutual information by parameterizing neural networks to maximize tight lower bounds of mutual information. However, in a mutual information minimization problem like Eq. (5), it is often counter-intuitive to maximize a lower bound of mutual information. Though one could still maximize the objective function of these estimators to estimate the mutual information and use such estimation to guide the optimization of Eq. (5) as a minimax game, it is hindered by two hurdles. First, it requires learning a well-trained estimator to estimate the mutual information during each epoch of optimizing Eq. (5). Second, if the estimator is not initialized with proper parameter settings, mutual information may be poorly estimated, which could further result in failing to find a good saddle point in such a minimax game.

B. Variational Representation of Mutual Information

In this paper, we take a different strategy from MINE and other similar methods by deriving a variational representation of mutual information $I(\tilde{\mathbf{y}};\mathbf{s})$. Our variational representation leverages a variational distribution of the vectorized sensitive feature $\mathbf{s}$ given the learning outcome $\tilde{\mathbf{y}}$ (Lemma 3).

Lemma 3: Suppose the joint distribution of the learning outcome $\tilde{\mathbf{y}}$ and the vectorized sensitive feature $\mathbf{s}$ is $p_{\tilde{\mathbf{y}},\mathbf{s}}$ and the marginal distributions of $\tilde{\mathbf{y}}$ and $\mathbf{s}$ are $p_{\tilde{\mathbf{y}}}$ and $p_{\mathbf{s}}$, respectively. Mutual information $I(\tilde{\mathbf{y}},\mathbf{s})$ between $\tilde{\mathbf{y}}$ and $\mathbf{s}$ is as follows.

$$I(\tilde{\mathbf{y}};\mathbf{s}) = H(\mathbf{s}) + \mathbb{E}_{(\tilde{\mathbf{y}},\mathbf{s})\sim p_{\tilde{\mathbf{y}},\mathbf{s}}}\big[\log q_{\mathbf{s}|\tilde{\mathbf{y}}}\big] + \mathbb{E}_{(\tilde{\mathbf{y}},\mathbf{s})\sim p_{\tilde{\mathbf{y}},\mathbf{s}}}\left[\log \frac{p_{\tilde{\mathbf{y}},\mathbf{s}}}{p_{\tilde{\mathbf{y}}} q_{\mathbf{s}|\tilde{\mathbf{y}}}}\right] \tag{7}$$

where $q_{\mathbf{s}|\tilde{\mathbf{y}}}$ is the variational distribution of $\mathbf{s}$ given $\tilde{\mathbf{y}}$.

Proof: Omitted for brevity. ■

Next, we minimize the variational representation shown in Lemma 3, which contains three terms: (1) the entropy $H(\mathbf{s})$, (2) the expectation of log likelihood $\mathbb{E}_{(\tilde{\mathbf{y}},\mathbf{s})\sim p_{\tilde{\mathbf{y}},\mathbf{s}}}\big[\log q_{\mathbf{s}|\tilde{\mathbf{y}}}\big]$ and (3) the expectation of log density ratio $\mathbb{E}_{(\tilde{\mathbf{y}},\mathbf{s})\sim p_{\tilde{\mathbf{y}},\mathbf{s}}}\left[\log \frac{p_{\tilde{\mathbf{y}},\mathbf{s}}}{p_{\tilde{\mathbf{y}}} q_{\mathbf{s}|\tilde{\mathbf{y}}}}\right]$. For the first term $H(\mathbf{s})$, we assume it to be a constant term, which can be ignored in the optimization stage. The rationale behind our assumption is that, in most (if not all) use cases, the vectorized sensitive feature $\mathbf{s}$ relates to the demographic information of an individual (e.g., gender, race, marital status, etc.), which should remain unchanged during the learning process. Then the remaining key challenges lie in (C1) calculating $\log q_{\mathbf{s}|\tilde{\mathbf{y}}}$ and (C2) estimating $\log \frac{p_{\tilde{\mathbf{y}},\mathbf{s}}}{p_{\tilde{\mathbf{y}}} q_{\mathbf{s}|\tilde{\mathbf{y}}}}$. The intuition of C1 and C2 is that we strive to find a learning outcome $\tilde{\mathbf{y}}$ such that (1) $\tilde{\mathbf{y}}$ fails to predict the vectorized sensitive feature $\mathbf{s}$ (refer to C1), while (2) making it hard to distinguish if the vectorized sensitive feature $\mathbf{s}$ is generated from the variational distribution or sampled from the original distribution (refer to C2).

C1 – Calculating $\log q_{\mathbf{s}|\tilde{\mathbf{y}}}$. It can be naturally formulated as a prediction problem, where the input is the learning outcome $\tilde{\mathbf{y}}$ and the output is the probability of $\mathbf{s}$ being predicted. To solve it, we parameterize a decoder $f(\tilde{\mathbf{y}};\mathbf{s};\mathbf{W})$ (e.g., a neural network) as a sensitive feature predictor to 'reconstruct' $\mathbf{s}$, where $\mathbf{W}$ is the learnable parameters in the decoder.

$$\log q_{\mathbf{s}|\tilde{\mathbf{y}}} = \log f(\tilde{\mathbf{y}};\mathbf{s};\mathbf{W}) \tag{8}$$

For categorical sensitive attribute, $\log q_{\mathbf{s}|\tilde{\mathbf{y}}}$ refers to the log likelihood of classifying $\tilde{\mathbf{y}}$ into label $\mathbf{s}$, which can be interpreted as the negative of cross-entropy loss of the decoder $f(\tilde{\mathbf{y}};\mathbf{s};\mathbf{W})$. Moreover, if $\mathbf{s}$ contains multiple categorical sensitive attributes, solving Eq. (8) requires solving a multi-label classification problem, which itself is not trivial to solve. In this case, we further reduce it to a single-label problem by applying a mapping function *map*() to map the multi-hot encoding $\mathbf{s}$ into a one-hot encoding $\hat{\mathbf{s}}$ (i.e., $\hat{\mathbf{s}} = \textit{map}(\mathbf{s})$).

C2 – Estimating $\log \frac{p_{\tilde{\mathbf{y}},\mathbf{s}}}{p_{\tilde{\mathbf{y}}} q_{\mathbf{s}|\tilde{\mathbf{y}}}}$. In practice, calculating $p_{\tilde{\mathbf{y}},\mathbf{s}}$ and $p_{\tilde{\mathbf{y}}} q_{\mathbf{s}|\tilde{\mathbf{y}}}$ individually is hard since the underlying distributions $p_{\tilde{\mathbf{y}},\mathbf{s}}$ and $p_{\tilde{\mathbf{y}}}$ are often unknown. Recall that our goal is to estimate the log of the ratio between these two joint distributions. Therefore, we estimate it through *density ratio estimation*, where the numerator $p_{\tilde{\mathbf{y}},\mathbf{s}}$ denotes the original joint distribution of the learning outcome $\tilde{\mathbf{y}}$ and ground-truth vectorized sensitive feature $\mathbf{s}$, and the denominator $p_{\tilde{\mathbf{y}}} q_{\mathbf{s}|\tilde{\mathbf{y}}}$ denotes the joint distribution of the learning outcome $\tilde{\mathbf{y}}$ and the vectorized sensitive feature $\tilde{\mathbf{s}}$ generated from the learning outcome using the aforementioned decoder.

We further reduce this density ratio estimation problem to a class probability estimation problem, which was originally developed in [18] for solving a different problem (i.e., the classification problem with the input distribution and the test distribution differing arbitrarily). The core idea is that, given a pair of learning outcome and vectorized sensitive feature, we want to predict whether it is drawn from the original joint distribution or from the joint distribution inferred by the decoder. We label each pair of learning outcome and ground-truth vectorized sensitive feature $(\tilde{\mathbf{y}},\mathbf{s})$ with a positive label ($c = 1$) and each pair of learning outcome and generated vectorized sensitive feature $(\tilde{\mathbf{y}},\tilde{\mathbf{s}})$ with a negative label ($c = -1$). After that, we rewrite the probability densities as $p_{\tilde{\mathbf{y}},\mathbf{s}} = \Pr[c = 1|\tilde{\mathbf{y}},\mathbf{s}]$ and $p_{\tilde{\mathbf{y}}} q_{\mathbf{s}|\tilde{\mathbf{y}}} = \Pr[c = -1|\tilde{\mathbf{y}},\mathbf{s}]$. Then the density ratio can be further rewritten as

$$\log \frac{p_{\tilde{\mathbf{y}},\mathbf{s}}}{p_{\tilde{\mathbf{y}}} q_{\mathbf{s}|\tilde{\mathbf{y}}}} = \log \frac{\Pr[c = 1|\tilde{\mathbf{y}},\mathbf{s}]}{\Pr[c = -1|\tilde{\mathbf{y}},\mathbf{s}]} = \textit{logit}(\Pr[c = 1|\tilde{\mathbf{y}},\mathbf{s}]) \tag{9}$$

Furthermore, if we model $\Pr[c = 1|\tilde{\mathbf{y}},\mathbf{s}]$ using logistic regression (i.e., $\Pr[c = 1|\tilde{\mathbf{y}},\mathbf{s}] = \textit{logistic}(\tilde{\mathbf{y}},\mathbf{s})$), Eq. (9) is reduced to a simple linear function as

$$\log \frac{p_{\tilde{\mathbf{y}},\mathbf{s}}}{p_{\tilde{\mathbf{y}}} q_{\mathbf{s}|\tilde{\mathbf{y}}}} = \textit{logit}(\textit{logistic}(\tilde{\mathbf{y}},\mathbf{s})) = \mathbf{w}_1^T\tilde{\mathbf{y}} + \mathbf{w}_2^T\mathbf{s} \tag{10}$$

where both $\mathbf{w}_1$ and $\mathbf{w}_2$ are learnable parameters. Putting everything together, we rewrite Eq. (5) as

$$\begin{aligned} J = {} & \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\big[l(\mathbf{x};\mathbf{s};y;\tilde{\mathbf{y}};\theta) + \alpha \log q_{\mathbf{s}|\tilde{\mathbf{y}}}\big] \\ & + \alpha\mathbb{E}_{\{(\tilde{\mathbf{y}},\mathbf{s})\sim p_{\tilde{\mathbf{y}},\mathbf{s}}\}\cup\{(\tilde{\mathbf{y}},\mathbf{s})\sim p_{\tilde{\mathbf{y}}} q_{\mathbf{s}|\tilde{\mathbf{y}}}\}}\big[\mathbf{w}_1^T\tilde{\mathbf{y}} + \mathbf{w}_2^T\mathbf{s}\big] \end{aligned} \tag{11}$$

where $p_{\tilde{\mathbf{y}},\mathbf{s}}$ is the joint distribution of the learning outcome $\tilde{\mathbf{y}}$ and ground-truth vectorized sensitive feature $\mathbf{s}$, $p_{\tilde{\mathbf{y}}} q_{\mathbf{s}|\tilde{\mathbf{y}}}$ is the joint distribution of the learning outcome $\tilde{\mathbf{y}}$ and predicted vectorized sensitive feature $\mathbf{s}$.

C. INFOFAIR: Overall Framework

Based on the objective function (Eq. (11)), we propose a generic end-to-end framework to solve the information-theoretic intersectional fairness problem. A general overview

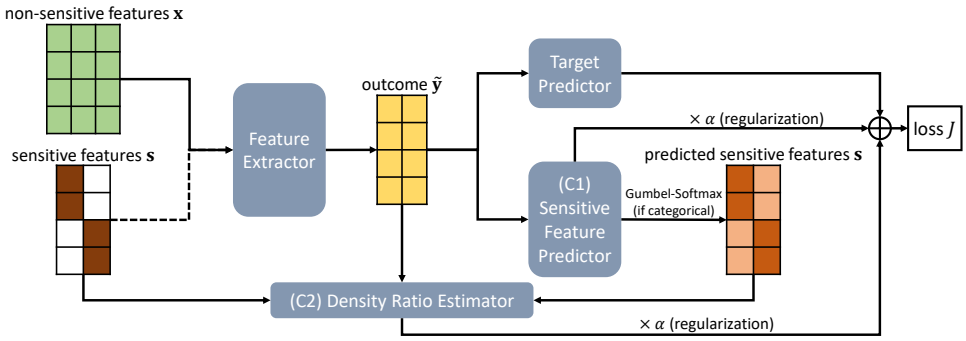

Fig. 2: A General overview of our proposed INFOFAIR framework. The dashed line between sensitive feature $\mathbf{s}$ and feature extractor means that sensitive features can be optionally passed into feature extractor as the input.

of the model architecture is shown in Fig. 2. Our proposed model contains four main modules, including (1) feature extractor, (2) target predictor, (3) sensitive feature predictor and (4) a density ratio estimator. In principle, as long as each module is differentiable, the proposed framework can be optimized by any gradient-based optimizer.

The general workflow of INFOFAIR is as follows.

1. The non-sensitive features and sensitive features (optional) are passed into a feature extractor to extract the learning outcomes;
2. The learning outcomes will be fed into a target predictor to predict the targets for a certain downstream task (i.e., $l(\mathbf{x}; \mathbf{s}; y; \tilde{\mathbf{y}}; \theta)$ in Eq. (11));
3. The learning outcomes will be passed into the sensitive feature predictor to 'reconstruct' the vectorized sensitive features (i.e., $\log q_{\mathbf{s}|\tilde{\mathbf{y}}}$ in Eq. (11));
4. Together with the learning outcomes and the ground-truth vectorized sensitive features, the predicted vectorized sensitive features will be used to estimate the density ratio between the original distribution and the variational distribution (i.e., $\mathbf{w}_1^T\tilde{\mathbf{y}} + \mathbf{w}_2^T\mathbf{s}$ in Eq. (11)).

Given a data point with categorical sensitive attribute(s), the predicted vectorized sensitive feature $\mathbf{s}$ is usually denoted as a one-hot vector. However, learning a one-hot vector is a difficult problem due to the discrete nature of vector elements, which makes the computation non-differentiable. To address this issue, we approximate such one-hot encoding by Gumbel-Softmax [19], which can be calculated as $\mathbf{s}[i] = \frac{\exp(\left[\log(\mathbf{o}_\mathbf{s}[i]) + g_i\right]/\tau)}{\sum_{j=1}^{n_\mathbf{s}} \exp(\left[\log(\mathbf{o}_\mathbf{s}[j]) + g_j\right]/\tau)}$, where $\mathbf{o}_\mathbf{s}$ is the output of the sensitive feature predictor, $n_\mathbf{s}$ is the dimension of $\mathbf{s}$, $g_1, \ldots, g_{n_\mathbf{s}}$ are i.i.d. points drawn from Gumbel$(0, 1)$ distribution, and τ is the softmax temperature. As $\tau \to \infty$, the Gumbel-Softmax samples are uniformly distributed; while as $\tau \to 0$, the Gumbel-Softmax distribution converges to a one-hot categorical distribution. In INFOFAIR, we start with a high temperature and then anneal it during epochs of training.

D. INFOFAIR: *Generalizations and Variants*

The proposed INFOFAIR is able to be generalized in multiple aspects. Due to the space limitation, we only give some brief descriptions, each of which could be a future direction in applying our proposed framework.

A – INFOFAIR with equal opportunity. Our INFOFAIR framework is generalizable to enforce equal opportunity [8], another widely-used group fairness notions. We leave for future work to explore the potential of INFOFAIR in enforcing equal opportunity.

Equal opportunity ensures equality across demographic groups for a preferred label (i.e., the label that benefits an individual). Mathematically, it is defined as follows.

Definition 2: (Equal Opportunity [8]). Following the settings of Definition 1, if equal opportunity is enforced, the hypothesis $h : \mathcal{X} \to \{0, 1\}$ satisfies

$$\Pr[h(x) = 1 | x \in \mathcal{M}, y = 1] = \Pr[h(x) = 1 | x \in \mathcal{M}^\mathcal{C}, y = 1] \tag{12}$$

where $\Pr[\cdot]$ denotes the probability of an event happening.

Analogous to the relationship between mutual information and statistical parity, ensuring equal opportunity is essentially a conditional mutual information minimization problem.

$$\underbrace{p_{\tilde{\mathbf{y}}|\mathbf{s},y=1} = p_{\tilde{\mathbf{y}}|y=1}}_{\text{equal opportunity}} \Leftrightarrow \quad \underbrace{I(\tilde{\mathbf{y}}; \mathbf{s} | y = 1) = 0}_{\text{zero conditional mutual information}} \tag{13}$$

By the definition of conditional mutual information, we have $I(\tilde{\mathbf{y}}; \mathbf{s}|y = 1) = H(\mathbf{s}|y = 1) - H(\mathbf{s}|\tilde{\mathbf{y}}, y = 1)$. For $H(\mathbf{s}|y = 1)$, we assume it as a constant term by the similar rationale of statistical parity. Similarly, we can rewrite $H(\mathbf{s}|\tilde{\mathbf{y}}, y = 1)$ as

$$\begin{aligned} H(\mathbf{s}|\tilde{\mathbf{y}}, y = 1) =& \mathbb{E}_{(\tilde{\mathbf{y}},\mathbf{s})\sim p_{\tilde{\mathbf{y}},\mathbf{s}|y=1}}\left[-\log q_{\mathbf{s}|\tilde{\mathbf{y}},y=1}\right] \\ & - \mathbb{E}_{(\tilde{\mathbf{y}},\mathbf{s})\sim p_{\tilde{\mathbf{y}},\mathbf{s}|y=1}}\left[\log \frac{p_{\tilde{\mathbf{y}},\mathbf{s}|y=1}}{p_{\tilde{\mathbf{y}}|y=1} q_{\mathbf{s}|\tilde{\mathbf{y}},y=1}}\right] \end{aligned} \tag{14}$$

Then, to compute $\log q_{\mathbf{s}|\tilde{\mathbf{y}},y=1}$, we could adopt similar strategy as computing $\log q_{\mathbf{s}|\tilde{\mathbf{y}}}$ in Section III-B by constructing a decoder $f(\tilde{\mathbf{y}}, \mathbf{s}, \mathbf{W})$ to 'reconstruct' $\mathbf{s}$ for *positive training samples*. Similarly, estimating the density ratio can be achieved by applying Eq. (10) on *positive training samples*. Thus, INFOFAIR is able to enforce equal opportunity by minimizing

$$\begin{aligned} J =& \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[l(\mathbf{x}; \mathbf{s}; y; \tilde{\mathbf{y}}; \theta) + \alpha \log q_{\mathbf{s}|\tilde{\mathbf{y}}}\right] \\ & + \alpha \mathbb{E}_{\{(\tilde{\mathbf{y}},\mathbf{s})\sim p_{\tilde{\mathbf{y}},\mathbf{s}|y=1}\}\cup\{(\tilde{\mathbf{y}},\mathbf{s})\sim p_{\tilde{\mathbf{y}}|y=1} q_{\mathbf{s}|\tilde{\mathbf{y}},y=1}\}}\left[\mathbf{w}_1^T\tilde{\mathbf{y}} + \mathbf{w}_2^T\mathbf{s}\right] \end{aligned} \tag{15}$$

B – Relationship to adversarial debiasing. Adversarial debiasing framework [13] consists of (1) a predictor that predicts the class membership probabilities using given data and (2) an adversary that takes the output of the predictor to predict the sensitive attribute of given data. The framework is

optimized to minimize the loss function of the predictor while maximizing the loss function of the adversary. If we merge feature extractor and target predictor to one single module and remove the density ratio estimator, INFOFAIR will degenerate to the adversarial debiasing method.

C – Relationship to Information Bottleneck. If we set the loss function l in Eq. (5) as the negative mutual information $-I(\tilde{\mathbf{y}}; y)$, Eq. (5) becomes the information bottleneck method [20]. Then the goal becomes to learn $\tilde{\mathbf{y}}$ that depends on the vectorized sensitive attribute $\mathbf{s}$ minimally and ground truth y maximally.

D – Fairness for continuous-valued sensitive features. Most existing works in fair machine learning only consider categorical sensitive attribute (e.g., gender, race). Our proposed INFOFAIR framework could be generalized to continuous-valued features as mutual information supports continuous-valued random variables. This advantage could empower our framework to work in even more application scenarios. For example, in image classification, we can classify images without the impact of certain image patches (e.g., patches that relate to individual's skin color). However, a major difficulty lies in modeling the variational distribution of sensitive attribute given the learning outcomes extracted from feature extractor. A potential solution could be utilizing a generative model (e.g., VAEs [21]) as the sensitive feature predictor.

E – Fairness for non-i.i.d. graph data. For fair graph mining, given a graph $G = (\mathbf{A}, \mathbf{X})$ where $\mathbf{A}$ is the adjacency matrix and $\mathbf{X}$ is the node feature matrix, we can use graph convolutional layer(s) as a feature extractor with the weight of the last layer to be identity matrix $\mathbf{I}$ and no nonlinear activation in the last graph convolution layer, in order to extract node representations. The reason for such a specific architecture in the last graph convolution layer is as follows. In general, a graph convolutional layer consists of two operations: feature aggregation $\mathbf{Z} = f_{\text{aggregate}}(\mathbf{A}; \mathbf{X}) = \mathbf{AX}$ and feature transformation $\mathbf{H} = f_{\text{transform}}(\mathbf{Z}; \mathbf{W}) = \sigma(\mathbf{ZW})$ where $\mathbf{W}$ is learnable parameters and σ is usually a nonlinear activation. The last layer in GCN [22] is simply $softmax(\mathbf{AXW})$, which can be viewed as a general multi-class logistic regression on the aggregated feature $\mathbf{Z} = \mathbf{AX}$ (i.e., $softmax(\mathbf{ZW})$).

F – Fairness beyond classification. Note that INFOFAIR does not have specific restrictions on the architecture of the feature extractor, target predictor or sensitive target predictor, which empowers it to handle many different types of tasks by selecting the proper architecture for each module. For example, if an analyst aims to learn fair representations with respect to gender for recommendation, s/he can set the feature extractor to be a multi-layer perceptron (MLP) for learning outcome extraction, the target predictor layer to be an MLP that predicts a rating and minimizes the mean squared error (MSE) between the predicted rating and ground-truth rating, and the sensitive target predictor to be another MLP with softmax to predict the gender based on extracted embedding.

IV. EXPERIMENTAL EVALUATION

In this section, we conduct experimental evaluations. All experiments are designed to answer the following questions:

RQ1. How does the fairness impact the learning performance?
RQ2. How effective is INFOFAIR in mitigating bias?

A. Experimental Settings

A – Datasets. We test the proposed method on three commonly-used datasets in fair machine learning research. The statistics of these datasets are summarized in Table II.

TABLE II: Statistics of datasets.

Datasets	# Samples	# Attributes	# Classes
COMPAS	6,172	52	2
Adult Income	45,222	14	2
Dutch Census	60,420	11	2

B – Baseline Methods. We compare INFOFAIR with several baseline methods, including *Learning Fair Representations (LFR)* [12], *MinDiff* [23], *Generalized Demographic Parity (GDP)* [24], *Adversarial Debiasing (Adversarial)* [13], *Fair Classification with Fairness Constraints (FCFC)* [6], *GerryFair* [9] and *Disparate Impact (DI)* [5].

C – Metrics. To answer **RQ1**, we measure the performance of classification using micro F1 and macro F1 (Micro/Macro F1). To answer **RQ2**, we measure to what extent the bias is reduced by the average statistical imparity (Imparity) and the relative bias reduction (Reduction) on average statistical imparity. The average statistical imparity (Imparity) is defined as $Imparity = avg(|\Pr(\hat{y} = c|\mathbf{x} \in g_1) - \Pr(\hat{y} = c|\mathbf{x} \in g_2)|)$ for any class label c and any pair of two different demographic groups g_1 and g_2. The relative bias reduction measures the relative decrease of the imparity of the debiased outcomes $Imparity_{debiased}$ to the imparity of vanilla outcomes (i.e., outcomes without fairness consideration) $Imparity_{vanilla}$. It is computed mathematically as $Reduction = 1 - \frac{Imparity_{debiased}}{Imparity_{vanilla}}$. Note that relative bias reduction defined above can be negative if the debiased learning outcome contains more biases than the vanilla learning outcome.

More experimental settings regarding reproducibility are provided in Appendix.

B. Main Results

We test our proposed framework, as well as baseline methods, in three different settings: debiasing binary sensitive attribute (i.e., gender for all three datasets), debiasing non-binary sensitive attribute (i.e., race for *COMPAS* and *Adult Income*, marital status for *Dutch Census*) and debiasing multiple sensitive attributes (i.e., gender & race for *COMPAS* and *Adult Income*, gender & marital status for *Dutch Census*). For each dataset and each setting, we train each model on training set, then select the trained model with best bias mitigation performance on validation set and report its performance on the test set. For the vanilla model (without any fairness consideration), we report the model with the highest micro and macro F1 scores. This is because the algorithm administrators are often more concerned with maximizing the utility of classification algorithms. The results of *LFR* and *MinDiff* in debiasing non-binary sensitive attribute and multiple sensitive attributes are absent since they only handle binary sensitive attribute by design.

A – Effectiveness results. The effectiveness results of INFOFAIR and baseline methods on *COMPAS*, *Adult Income* and *Dutch Census* datasets are shown in Table III. We provide

TABLE III: Debiasing results on all datasets. Lower is better for the gray column (Imparity). Higher is better for all others.

Debiasing results on *COMPAS* dataset									
Method	**gender**			**race**			**gender & race**		
	Micro/Macro F1	**Imparity**	**Reduction**	**Micro/Macro F1**	**Imparity**	**Reduction**	**Micro/Macro F1**	**Imparity**	**Reduction**
Vanilla	0.972/0.972	0.050	0.000%	0.972/0.972	0.181	0.000%	0.972/0.972	0.234	0.000%
LFR	0.554/0.357	0.000	100.0%	N/A	N/A	N/A	N/A	N/A	N/A
MinDiff	0.972/0.972	0.050	0.000%	N/A	N/A	N/A	N/A	N/A	N/A
DI	0.972/0.972	0.050	0.000%	0.972/0.972	0.181	0.000%	0.972/0.972	0.234	0.000%
Adversarial	0.554/0.357	0.000	100.0%	0.554/0.357	0.000	100.0%	0.554/0.357	0.000	100.0%
FCFC	0.446/0.308	0.000	100.0%	0.446/0.308	0.000	100.0%	0.446/0.308	0.000	100.0%
GerryFair	0.972/0.972	0.050	0.000%	0.972/0.972	0.181	0.000%	0.972/0.972	0.234	0.000%
GDP	0.972/0.972	0.050	0.000%	0.972/0.972	0.181	0.000%	0.972/0.972	0.234	0.000%
INFOFAIR	0.924/0.923	0.038	23.15%	0.815/0.803	0.179	1.010%	0.877/0.872	0.231	1.350%
Debiasing results on *Adult Income* dataset									
Method	**gender**			**race**			**gender & race**		
	Micro/Macro F1	**Imparity**	**Reduction**	**Micro/Macro F1**	**Imparity**	**Reduction**	**Micro/Macro F1**	**Imparity**	**Reduction**
Vanilla	0.830/0.762	0.066	0.000%	0.830/0.762	0.062	0.000%	0.830/0.762	0.083	0.000%
LFR	0.743/0.426	0.000	100.0%	N/A	N/A	N/A	N/A	N/A	N/A
MinDiff	0.828/0.746	0.058	12.06%	N/A	N/A	N/A	N/A	N/A	N/A
DI	0.823/0.730	0.053	19.85%	0.825/0.743	0.056	10.62%	0.823/0.736	0.081	2.276%
Adversarial	0.743/0.426	0.000	100.0%	0.743/0.426	0.000	100.0%	0.743/0.426	0.000	100.0%
FCFC	0.257/0.204	0.000	100.0%	0.257/0.204	0.000	100.0%	0.257/0.204	0.000	100.0%
GerryFair	0.833/0.752	0.056	15.70%	0.833/0.752	0.067	−7.664%	0.797/0.710	0.215	−158.3%
GDP	0.825/0.744	0.055	16.73%	0.827/0.749	0.059	6.351%	0.824/0.740	0.075	9.246%
INFOFAIR	0.816/0.721	0.047	29.24%	0.810/0.686	0.042	32.11%	0.818/0.714	0.082	1.532%
Debiasing results on *Dutch Census* dataset									
Method	**gender**			**marital status**			**gender & marital status**		
	Micro/Macro F1	**Imparity**	**Reduction**	**Micro/Macro F1**	**Imparity**	**Reduction**	**Micro/Macro F1**	**Imparity**	**Reduction**
Vanilla	0.832/0.831	0.119	0.000%	0.832/0.831	0.079	0.000%	0.832/0.831	0.172	0.000%
LFR	0.521/0.342	0.000	100.0%	N/A	N/A	N/A	N/A	N/A	N/A
MinDiff	0.831/0.830	0.107	10.16%	N/A	N/A	N/A	N/A	N/A	N/A
DI	0.825/0.824	0.104	12.43%	0.830/0.830	0.080	−1.156%	0.814/0.811	0.127	26.65%
Adversarial	0.521/0.342	0.000	100.0%	0.521/0.342	0.000	100.0%	0.521/0.342	0.000	100.0%
FCFC	0.479/0.324	0.000	100.0%	0.479/0.324	0.000	100.0%	0.479/0.324	0.000	100.0%
GerryFair	0.826/0.823	0.078	34.29%	0.826/0.823	0.070	11.70%	0.826/0.823	0.125	27.53%
GDP	0.828/0.826	0.097	18.31%	0.827/0.826	0.086	−9.056%	0.827/0.825	0.131	23.80%
INFOFAIR	0.817/0.813	0.068	43.08%	0.815/0.811	0.077	2.017%	0.819/0.817	0.128	25.65%

additional results on visualizing the trade-off between micro F1 score and average statistical imparity in Appendix. From the tables, we have the following observations. First, Our method is the only method that can mitigate bias (i.e., Imparity and Reduction) effectively and consistently with a small degree of sacrifice to the vanilla classification performance (i.e., Micro/Macro F1) for all datasets and all settings. Second, though *LFR*, *Adversarial Debiasing* and *FCFC* achieves the perfect bias reduction, their classification performance is severely reduced by predicting all data samples with the same label (i.e., negative sample for *LFR* and *Adversarial Debiasing*, positive sample for *FCFC*). Though, in a few settings, *DI*, *GerryFair* and *GDP* mitigate more bias than INFOFAIR, they either *amplify* the bias or fail to outperform INFOFAIR in the other settings. All in all, INFOFAIR achieves the best balance in reducing the bias and maintaining the classification accuracy in most cases.

B – Trade-off between micro F1 and average statistical imparity. Figure 3 shows the results of trade-off between micro F1 (Micro F1) and average statistical imparity (Imparity). From the figure, we can observe that INFOFAIR achieves the best trade-off between accuracy and fairness (i.e., being closer to the bottom right corner in Figure 3) in most cases.

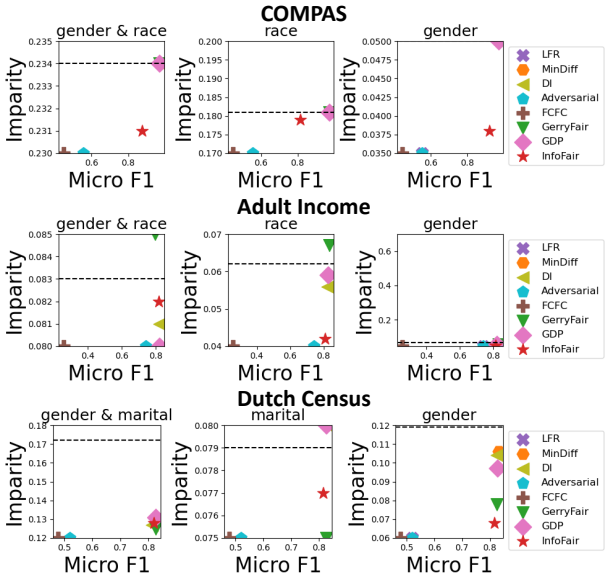

Fig. 3: Trade-off between micro F1 score and average statistical imparity. Best viewed in color. Red star represents INFOFAIR. The closer to bottom right, the better trade-off between micro F1 score and average statistical imparity. Bias is amplified by a method if its corresponding point is above the dashed line (which denotes the imparity of Vanilla).

C. Ablation Study

Let $T = \mathbb{E}\big[l(\mathbf{x}; \mathbf{s}; y; \tilde{\mathbf{y}}; \theta)\big]$ be the empirical loss of target predictor, $S = \alpha\mathbb{E}\big[\log q_{\mathbf{s}|\tilde{\mathbf{y}}}\big]$ be the empirical loss of sensitive feature predictor and $D = \alpha\mathbb{E}\big[\mathbf{w}_1^T\tilde{\mathbf{y}} + \mathbf{w}_2^T\mathbf{s}\big]$ be the empirical loss of density ratio estimator, objective function of INFOFAIR (Eq. (11)) can be written as $J = T + S + D$. To evaluate the effectiveness on optimizing the proposed variational representation of mutual information, we compare with two variants of objective function, i.e., $T + S$ and $T + D$, on the

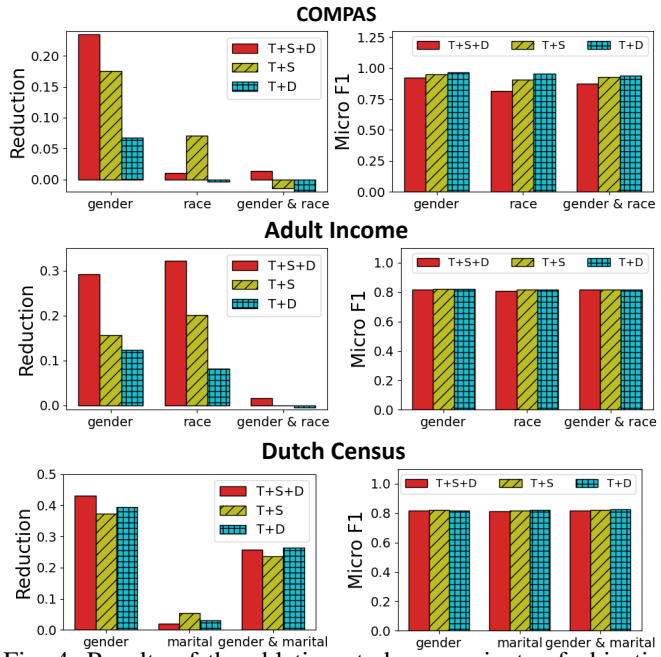

Fig. 4: Results of the ablation study on variants of objective function. Best viewed in color. Higher is better.

same datasets and the same set of sensitive attributes as in Section IV-B. Experimental protocols and parameter settings are kept the same among all compared objective functions (i.e., $T+S+D$, $T+S$ and $T+D$). The results of the ablation study are shown in Figure 4. From the figure, we observe that our objective function (i.e., $T+S+D$) can mitigate more bias than the other two variants (i.e., $T+S$ and $T+D$) in most cases. This implies that our proposed variational representation can better model the dependence between the learning outcomes and the vectorized sensitive features.

When debiasing race on COMPAS dataset and debiasing marital status on Dutch Census dataset, we observe that the cardinalities of the demographic groups are more imbalanced. Recall the goal of sensitive feature predictor is to reduce the accuracy of predicting sensitive feature using the extracted embedding. When the demographic groups are more imbalanced, it tends to learn an embedding that contains information about a wrong demographic group to reduce its accuracy. Thus, though $T+S$ may achieve lower imparity, the statistical dependence between the extracted embeddings and the sensitive feature may not be reduced, meaning that the extracted embeddings are merely shifted to correlate with wrong demographic groups. However, with the addition of density ratio estimator, we ensure that (1) not only the sensitive feature predictor makes wrong prediction (2) but also the distribution of sensitive attribute and the extracted embeddings modeled by S ($p_{\tilde{\mathbf{y}}}q_{\mathbf{s}|\tilde{\mathbf{y}}}$) are similar to its corresponding original distribution (i.e., $p_{\tilde{\mathbf{y}},\mathbf{s}}$).

V. Related Work

A – Group fairness aims to ensure statistical-based fairness notions across the entire populations. It has been extensively studied in many application domains, including credit scoring [5], recidivism [25], healthcare [12], recommender systems [26] and natural language processing [27]. Kamishima et al. [28] estimate mutual information between the learning outcome and sensitive attribute by marginalizing the output of a probabilistic discriminative model. Zemel et al. [12] use a regularized approach to learn fair embeddings for group fairness and individual fairness. Feldman et al. [5] debias input data distribution by linear interpolation of original data distribution and fair data distribution. Zhang et al. [13] propose an adversarial debiasing framework. Bose et al. [10] learn fair node representations by adversarial learning. However, their proposed framework could only debias multiple distinct sensitive attributes instead of multiple sensitive attributes simultaneously. Kearns et al. [9] further consider subgroup fairness from game-theoretic perspective. Different from [9], InfoFair directly optimizes statistical parity through mutual information minimization instead of optimizing the self-defined surrogate 'fairness violation' functions using game-theoretic method. Zafar et al. [6] ensure statistical parity by minimizing the covariance between the sensitive attribute of each data sample and its distance to the decision boundary of a convex margin-based classifier. Adeli et al. [29] remove statistical dependence by minimizing Pearson's correlation for a convolutional neural network. Nevertheless, [6], [29] only remove the linear dependence whereas our proposed InfoFair removes both linear and nonlinear dependence directly. In addition to statistical parity and disparate impact, Hardt et al. [8] propose another widely-used fairness notion named equal opportunity. Prost et al. [23] achieve equal false positive rate through maximum mean discrepancy (MMD) minimization. However, it can only debias with respect to binary sensitive attribute by design. Jiang et al. [24] propose generalized demographic parity for tractable calculation demographic parity with respect to continuous-valued sensitive attributes. In terms of intersectional fairness, Kim et al. [30] propose Multiaccuracy Boost to ensure low classification error for each intersectional demographic group. Foulds et al. [31] propose ϵ-differential fairness, which ensures pairwise equal acceptance rate. They further estimate ϵ-differential fairness and its corresponding uncertainty [32]. Morina et al. [33] develop the equivalence between minimizing ϵ-differential fairness and minimizing a linear combination of false positive rate and false negative rate in a binary classification problem. Ramos et al. [34] ensure intersectional fairness in reputation-based ranking systems by minimizing the difference among the average reputations of a user from different demographic groups. Different from [34], our proposed InfoFair ensures intersectional fairness from information-theoretic perspective, and is applicable to various learning tasks as shown in Section III-D.

B – Mutual information estimation for high-dimensional data has been made possible in recent decades by analyzing variational bounds of mutual information with machine learning techniques. Regarding variational upper bound of mutual information, Kingma et al. [21] and Rezende et al. [35] almost concurrently propose Variational Auto-Encoders (VAEs) which optimizes a variational upper bound of mutual informa-

tion conceptually. Variational lower bounds of mutual information have been extensively studied recently. Barber et al. [36] propose a variational lower bound of mutual information and maximize the mutual information through moment matching. Belghazi et al. [15] propose Mutual Information Neural Estimation (MINE) which maximizes Donsker-Varadhan representation of Kullback-Leibler (KL) divergence [37]. In [15], MINE-f, a variant of MINE, is proposed to maximize the variational estimation of f-divergence introduced by Nguyen et al. [38]. The same variational representation of f-divergence has been applied to other generative models like f-GAN [39]. Mukherjee et al. [17] propose a classifier-based neural estimator for conditional mutual information named CCMI. In addition, van den Oord et al. [40] propose infoNCE based on noise contrastive estimation (NCE) [41]. Hjelm et al. [16] propose Deep Infomax (DIM) to maximize the mutual information between global representation and local regions of the input, which is further generalized to graphs [42].

VI. Conclusion

In this paper, we study information-theoretic intersectional fairness, where we aim to simultaneously debias the learning results with respect to multiple sensitive attributes. We formally define the information-theoretic intersectional fairness problem by measuring the dependence between the learning results and multiple sensitive attributes as the mutual information between learning results and a joint attribute formed by these sensitive attributes. Based on that, we formulate it as an optimization problem and further propose a generic end-to-end framework, which effectively minimizes mutual information between the learning results and the joint attribute through its variational representation. We perform fair classification on three real-world datasets with the consideration of categorical sensitive attributes. The empirical evaluation results demonstrate that our proposed framework can effectively debias the classification results with respect to one or more sensitive attribute(s) with little sacrifice to the classification accuracy. Our framework is generalizable to different settings beyond the scope of fair classification with categorical sensitive attributes in our experimental evaluation. In the future, we will investigate our framework in other learning tasks (e.g., recommendation) and its effectiveness in mitigating bias for continuous-valued sensitive attributes (e.g., age, income).

Acknowledgement

This work is supported by NSF (1947135, 2134079 and 2147375), the NSF Program on Fairness in AI in collaboration with Amazon (1939725), DARPA (HR001121C0165), NIFA (2020-67021-32799), ARO (W911NF2110088), and DHS (2017-ST-061-QA0001 and 17STQAC00001-03-03). The content of the information in this document does not necessarily reflect the position or the policy of the Government or Amazon, and no official endorsement should be inferred. The U.S. Government is authorized to reproduce and distribute reprints for Government purposes notwithstanding any copyright notation here on.

References

[1] C. Luo, D. Wu, and D. Wu, "A deep learning approach for credit scoring using credit default swaps," *Engineering Applications of Artificial Intelligence*, 2017.

[2] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, "Fairness in criminal justice risk assessments: The state of the art," *arXiv preprint arXiv:1703.09207*, 2017.

[3] M. A. Ahmad, C. Eckert, and A. Teredesai, "Interpretable machine learning in healthcare," in *BCB*, 2018.

[4] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *ITCS*, 2012.

[5] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *KDD*, 2015.

[6] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *AISTATS*, 2017.

[7] S. B. Morris and R. E. Lobsenz, "Significance tests and confidence intervals for the adverse impact ratio," *Personnel Psychology*, 2000.

[8] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in *NIPS*, 2016.

[9] M. Kearns, S. Neel, A. Roth, and Z. S. Wu, "Preventing fairness gerrymandering: Auditing and learning for subgroup fairness," in *ICML*, 2018.

[10] A. Bose and W. Hamilton, "Compositional fairness constraints for graph embeddings," in *ICML*, 2019.

[11] C. E. Shannon, "A mathematical theory of communication," *The Bell System Technical Journal*, 1948.

[12] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *ICML*, 2013.

[13] B. H. Zhang, B. Lemoine, and M. Mitchell, "Mitigating unwanted biases with adversarial learning," in *AIES*, 2018.

[14] A. Ghassami, S. Khodadadian, and N. Kiyavash, "Fairness in supervised learning: An information theoretic approach," in *ISIT*, 2018.

[15] M. I. Belghazi, A. Baratin, S. Rajeshwar, S. Ozair, Y. Bengio, A. Courville, and D. Hjelm, "Mutual information neural estimation," in *ICML*, 2018.

[16] R. D. Hjelm, A. Fedorov, S. Lavoie-Marchildon, K. Grewal, P. Bachman, A. Trischler, and Y. Bengio, "Learning deep representations by mutual information estimation and maximization," in *ICLR*, 2018.

[17] S. Mukherjee, H. Asnani, and S. Kannan, "Ccmi: Classifier based conditional mutual information estimation," in *UAI*, 2020.

[18] S. Bickel, M. Brückner, and T. Scheffer, "Discriminative learning under covariate shift." *Journal of Machine Learning Research*, 2009.

[19] E. Jang, S. Gu, and B. Poole, "Categorical reparameterization with gumbel-softmax," in *ICLR*, 2017.

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

[21] D. P. Kingma and M. Welling, "Auto-encoding variational bayes," in *ICLR*, 2014.

[22] T. N. Kipf and M. Welling, "Semi-supervised classification with graph convolutional networks," in *ICLR*, 2017.

[23] F. Prost, H. Qian, Q. Chen, E. H. Chi, J. Chen, and A. Beutel, "Toward a better trade-off between performance and fairness with kernel-based distribution matching," *arXiv preprint arXiv:1910.11779*, 2019.

[24] Z. Jiang, X. Han, C. Fan, F. Yang, A. Mostafavi, and X. Hu, "Generalized demographic parity for group fairness," in *ICLR*, 2022.

[25] J. Dressel and H. Farid, "The accuracy, fairness, and limits of predicting recidivism," *Science Advances*, 2018.

[26] S. Yao and B. Huang, "Beyond parity: Fairness objectives for collaborative filtering," in *NIPS*, 2017.

[27] J. Zhao, T. Wang, M. Yatskar, V. Ordonez, and K.-W. Chang, "Men also like shopping: Reducing gender bias amplification using corpus-level constraints," in *EMNLP*, 2017.

[28] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma, "Fairness-aware classifier with prejudice remover regularizer," in *ECML/PKDD*, 2012, pp. 35–50.

[29] E. Adeli, Q. Zhao, A. Pfefferbaum, E. V. Sullivan, L. Fei-Fei, J. C. Niebles, and K. M. Pohl, "Representation learning with statistical independence to mitigate bias," *arXiv preprint arXiv:1910.03676*, 2019.

[30] M. P. Kim, A. Ghorbani, and J. Zou, "Multiaccuracy: Black-box post-processing for fairness in classification," in *AIES*, 2019, pp. 247–254.

[31] J. R. Foulds, R. Islam, K. N. Keya, and S. Pan, "An intersectional definition of fairness," in *ICDE*. IEEE, 2020, pp. 1918–1921.

[32] ——, "Bayesian modeling of intersectional fairness: The variance of bias," in *SDM*. SIAM, 2020, pp. 424–432.
[33] G. Morina, V. Oliinyk, J. Waton, I. Marusic, and K. Georgatzis, "Auditing and achieving intersectional fairness in classification problems," *arXiv preprint arXiv:1911.01468*, 2019.
[34] G. Ramos, L. Boratto, and M. Marras, "Reputation equity in ranking systems," in *CIKM*, 2021, pp. 3378–3382.
[35] D. J. Rezende, S. Mohamed, and D. Wierstra, "Stochastic backpropagation and approximate inference in deep generative models," in *ICML*, 2014.
[36] D. Barber and F. V. Agakov, "The im algorithm: A variational approach to information maximization," in *NIPS*, 2003.
[37] M. D. Donsker and S. S. Varadhan, "Asymptotic evaluation of certain markov process expectations for large time. iv," *Communications on Pure and Applied Mathematics*, 1983.
[38] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *IEEE Transactions on Information Theory*, 2010.
[39] S. Nowozin, B. Cseke, and R. Tomioka, "f-gan: Training generative neural samplers using variational divergence minimization," in *NIPS*, 2016.
[40] A. v. d. Oord, Y. Li, and O. Vinyals, "Representation learning with contrastive predictive coding," *arXiv preprint arXiv:1807.03748*, 2018.
[41] M. Gutmann and A. Hyvärinen, "Noise-contrastive estimation: A new estimation principle for unnormalized statistical models," in *AISTATS*, 2010.
[42] P. Veličković, W. Fedus, W. L. Hamilton, P. Liò, Y. Bengio, and R. D. Hjelm, "Deep graph infomax," in *ICLR*, 2018.
[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," in *ICLR*, 2015.

APPENDIX

A. Descriptions of Baseline Methods

Learning Fair Representations (LFR) [12] learns a set of fair prototype representations. Each data sample is first mapped to a prototype, which is used to predict fair outcome. We use the implementation by IBM AIF360 and the same grid search strategy for hyperparameters in [12].

MinDiff [23] ensures equal false positive rate by minimizing the maximum mean discrepancy (MMD) between the two demographic groups with negative samples only. We implement our own version of MinDiff with the Gaussian kernel. The hyperparameters for the Gaussian kernel is set to be consistent with [23]. For fair comparison, we set the regularization hyperparameter to 0.1, which is the same with the corresponding setting for INFOFAIR.

Disparate Impact (DI) [5] ensures disparate impact by interpolating the original data distribution with an unbiased distribution. For fair comparison, we set the linear interpolation coefficient, which is referred to as λ in [5], such that the interpolation ratios of [5] and ours are the same, i.e., $\frac{1-\lambda}{\lambda} = \frac{1}{\alpha}$.

Adversarial Debiasing (Adversarial) [13] uses an adversary to predict the sensitive attribute using the prediction from a predictor. Both the predictor and the adversary can be flexibly chosen. Since its official source code is not available, we implement the model using the same machine configurations as INFOFAIR. For fair comparison, we switch (1) the predictor to feature extractor and target predictor in our proposed framework and (2) the adversary to sensitive feature predictor in our framework. We also set the same learning rate as our framework.

Fair Classification with Fairness Constraints (FCFC) [6] measures the statistical imparity as the covariance between the sensitive attribute of a data sample and the distance of the corresponding data sample to the decision boundary of a linear classifier. We use the official implementation of FCFC provided by Zafar et al. and adopt their released parameter settings in our experiments.

GerryFair [9] ensures subgroup fairness for cost-sensitive classification through fictitious play from game-theoretic perspective. Since the relationship between α in INFOFAIR and parameters of *GerryFair* is unclear, we use the default parameters provided in the officially released source code

Generalized Demographic Parity (GDP) [24] computes the weighted total variation distance on local average prediction and global average prediction. For fair comparison, we use the official implementation, set the same backbone model for feature extraction and prediction and use the same regularization hyperparameter (0.1) as INFOFAIR.

B. Experimental Protocol and Model Architectures

The learning task we consider is fair classification with respect to categorical sensitive attribute(s). For all datasets, we take both non-sensitive features and sensitive features as input to the feature extractor. Regarding the model architecture, for *Adult Income* and *Dutch Census* datasets, the feature extractor is a 1-layer MLP with hidden dimension 32; the target predictor contains one hidden layer that calculates the log likelihood of predicting class label using the embeddings output by the feature extractor; and the sensitive feature predictor is similar to the target predictor that leverages one hidden layer to calculate the log likelihood of predicting the vectorized sensitive feature using the extracted embeddings. For *COMPAS* dataset, we set the feature extractor to be a 2-layer MLP with hidden dimension 32 in each layer, while keeping all other modules to be the same as they are for *Adult Income* and *Dutch Census* datasets.

C. Parameter Settings and Repeatability

For all datasets, we set the regularizatioin parameter $\alpha = 0.1$. The number of epochs for training is set to 100 with a patience of 5 for early stopping. Weight decay is set to 0.01. We tune the learning rate as 0.001 for *DI* and 0.0001 for *MinDiff*, *Adversarial* and our method. All learnable model parameters are optimized with Adam optimizer [43]. The starting temperature for Gumbel-Softmax is set to 1 and is divided by 2 every 50 epochs for annealing. To reduce randomness and enhance reproducibility, we run 5 different initializations with random seed from 0 to 4.

D. Machine Configurations

All three datasets are publicly available online. All models (i.e., INFOFAIR and baseline methods) are implemented with PyTorch 1.9.0 and are trained on a Linux server with 96 Intel Xeon Gold 6240R CPUs at 2.40 GHz and 4 Nvidia Tesla V100 SXM2 GPUs with 32 GB memory. We will release the source code upon publication.