Fair Representation Learning: An Alternative to Mutual Information

Ji Liu
Zenan Li
State Key Laboratory for Novel Software Technology,
Nanjing University, China
{mf1933059,lizn}@smail.nju.edu.cn

Yuan Yao
Feng Xu
Xiaoxing Ma
State Key Laboratory for Novel Software Technology,
Nanjing University, China
{y.yao,xf,xxm}@nju.edu.cn

Miao Xu
The University of Queensland, Australia
miao.xu@uq.edu.au

Hanghang Tong
University of Illinois at Urbana-Champaign, USA
htong@illinois.edu

ABSTRACT

Learning fair representations is an essential task to reduce bias in data-oriented decision making. It protects minority subgroups by requiring the learned representations to be independent of sensitive attributes. To achieve independence, the vast majority of the existing work primarily relaxes it to the minimization of the mutual information between sensitive attributes and learned representations. However, direct computation of mutual information is computationally intractable, and various upper bounds currently used either are still intractable or contradict the utility of the learned representations. In this paper, we introduce distance covariance as a new dependence measure into fair representation learning. By observing that sensitive attributes (e.g., gender, race, and age group) are typically categorical, the distance covariance can be converted to a tractable penalty term without contradicting the utility desideratum. Based on the tractable penalty, we propose FairDisCo, a variational method to learn fair representations. Experiments demonstrate that FairDisCo outperforms existing competitors for fair representation learning.

CCS CONCEPTS

• **Information systems** → **Data mining**; • **Computing methodologies** → *Learning latent representations*; Unsupervised learning.

KEYWORDS

fair representation learning, mutual information, distance covariance

KDD '22, August 14–18, 2022, Washington, DC, USA

ACM ISBN 978-1-4503-9385-0/22/08…$15.00
https://doi.org/10.1145/3534678.3539302

ACM Reference Format:
Ji Liu, Zenan Li, Yuan Yao, Feng Xu, Xiaoxing Ma, Miao Xu, and Hanghang Tong. 2022. Fair Representation Learning: An Alternative to Mutual Information. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '22), August 14–18, 2022, Washington, DC, USA.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3534678.3539302

1 INTRODUCTION

In many data-oriented decision making applications such as loan approval and recidivism prediction, a fundamental requirement is that the decision should be fairly made, i.e., free from *sensitive attributes* such as gender, race, or age. However, it has been found that classical machine learning and data mining systems may unintentionally output biased predictions [5, 16, 20].

In view of this, fair representation learning [31, 48] has been proposed and studied, with the goal of learning a representation free of the impact from sensitive attributes while maintaining essential expressiveness to aid the downstream decision making. Computationally, fair representation learning requires the learned representations to be *independent* from the sensitive attributes. To achieve the independence, most existing work aims to minimize the *mutual information* (MI) between the learned representations and the sensitive attributes. Due to the computational challenge of MI [2, 41], existing work proposes to minimize various upper bounds of MI. However, existing relaxations of MI may either result in intractable penalty terms that can only be solved via the unstable and sometimes counter-productive adversarial training [32], or contradict the utility desideratum [42] causing negative impact on the prediction accuracy of downstream tasks.

To address the limitations of existing work, in this paper, we propose to use *distance covariance* as a new dependence measure for fair representation learning. We focus on a fundamental group fairness notion named *demographic parity* (DP), which has attracted a lot of attention in recent years [7, 9, 12, 28, 39, 42, 48]. Essentially, DP means that different subgroups categorized according to sensitive attributes should receive positive outcome at equal rates. We show that distance covariance is a lower bound of MI, and it bears nice properties such as consistence to independence and asymptotic equivalence to MI. With the observation that sensitive attributes are typically categorical (e.g., gender, race, and age

group), we further convert the distance covariance between sensitive attributes and learned representations to a penalty term, which is tractable in optimization and does not contradict the utility terms. We further incorporate it into a variational learning framework named FairDisCo. Experimental evaluations are conducted on real datasets to demonstrate the effectiveness of the proposed approach.

In summary, our main contributions include:

- *Problem Definition.* To the best of our knowledge, we are the first to introduce distance covariance as a dependence measure for fair representation learning.
- *Analysis.* We show that distance covariance is a tighter upper bound of maximal correlation compared to MI, and it also provides a closed-form computation for the dependence between target variables.
- *Algorithm.* We propose a fair representation learning approach FairDisCo. It incorporates the distance covariance between sensitive attributes and learned representations as a penalty term into the variational optimization framework.
- *Experimental Evaluation.* We conduct experiments showing that FairDisCo can ensure near-perfect fairness while generally achieving higher utility results than the existing competitors. For example, with near-perfect fairness, the proposed approach achieves up to 14.9% accuracy improvements compared to the best competitors.

The rest of the paper is organized as follows. Section 2 analyzes the distance covariance as a dependence measure. Section 3 presents the proposed fair representation learning approach FairDisCo, and Section 4 shows the experimental results. Section 5 reviews the related work and Section 6 concludes the paper.

2 DISTANCE COVARIANCE AS DEPENDENCE MEASURE

In this section, we introduce distance covariance, and show its properties and relationships to mutual information.

2.1 Dependence Measures

For notation convenience, we start with the discussion of continuous random variables, and such discussion can be easily extended to discrete or categorical cases. For a pair of continuous random variables (X, Y) from space $\mathcal{X}\times\mathcal{Y}$, we denote their joint probability density function by $p_{(X,Y)}$, and the marginal probability density functions by p_X and p_Y, respectively.

A measure of dependence indicates in some particular manner how closely the variables X and Y are related [37, 43], and the most commonly used measure is Pearson's correlation coefficient (a.k.a. bivariate correlation):

$$\rho(X,Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathrm{Var}(X)\mathrm{Var}(Y)}}, \tag{1}$$

where $\mathrm{Cov}(X,Y)$ denotes the covariance of X and Y, and $\mathrm{Var}(X)$ and $\mathrm{Var}(Y)$ are the variances of X and Y, respectively. However, Pearson's coefficient is limited to the case when X and Y are linearly dependent, and it is highly insensitive for non-linear cases. To go beyond the linear correlation measure, researchers propose maximal correlation (a.k.a. sup correlation) [3], i.e., the supremum of the Pearson correlation over all Borel-measurable functions f and g (for which $\mathrm{Var}(f(X))$ and $\mathrm{Var}(g(Y))$ are finite and nonzero):

$$\rho_{\max}(X,Y) = \sup_{f,g} \rho\big(f(X), g(Y)\big) = \sup_{f,g} \frac{\mathrm{Cov}(f(X), g(Y))}{\sqrt{\mathrm{Var}(f(X))\mathrm{Var}(g(Y))}}. \tag{2}$$

Though the maximal correlation enjoys many nice properties [4, 34, 46], it is usually not readily computable [1, Sec. 4.5]. Hence, existing work tends to use mutual information as a measure of dependence and minimize MI to boost the independence between target variables. The MI between X and Y is defined as

$$\begin{aligned} \mathcal{I}(X,Y) &= D_{KL}(p_{(X,Y)} \| p_X \otimes p_Y) \\ &= \int_{\mathcal{Y}} \int_{\mathcal{X}} p_{(X,Y)}(x,y) \log\left(\frac{p_{(X,Y)}(x,y)}{p_X(x)p_Y(y)}\right) dx\, dy, \end{aligned} \tag{3}$$

where $D_{KL}(\cdot\|\cdot)$ is the Kullback–Leibler (KL) divergence.

We summarize the two principal properties of MI as follows, which reveal the connection to the maximal correlation (normalized maximal covariance) and illustrate the main rationale of MI as a dependence measure [14, 44].

Proposition 1. **Consistency to independence**. *MI is non-negative, i.e., $\mathcal{I}(X,Y) \geq 0$, and the equation holds if and only if X and Y are independent random variables.*

Proposition 2. **Upper bound to maximal covariance**. *For the covariance $\mathrm{Cov}(f(X), g(Y))$ of any two given functions f and g, we have*

$$\mathcal{I}(X,Y) \geq \frac{\mathrm{Cov}(f(X), g(Y))^2}{2\|f\|_\infty^2 \|g\|_\infty^2}.$$

Proposition 1 can be proved by directly using Gibbs' inequality. The complete proof of Proposition 2 can be found in Appendix A.1.

However, MI is still intractable in practice. Instead, in this paper, we propose to use the distance covariance as a dependence measure for fair representation learning. Specifically, the distance covariance refers to the (squared) Euclidean distance between $p_{(X,Y)}$ and $p_X \otimes p_Y$, i.e.,

$$\begin{aligned} \mathcal{V}^2(X,Y) &= \delta_E^2(p_{(X,Y)}, p_X \otimes p_Y) \\ &= \int_{\mathcal{Y}} \int_{\mathcal{X}} |p_{(X,Y)}(x,y) - p_X(x)p_Y(y)|^2\, dx\, dy. \end{aligned} \tag{4}$$

To normalize the distance covariance to $[0,1]$, one can divide it by the factor $\sqrt{\mathcal{V}^2(X,X)\mathcal{V}^2(Y,Y)}$, and the normalized measure is called the distance correlation [43].

2.2 Properties of Distance Covariance

Here, we start to discuss the properties of distance covariance. Existing studies have enumerated seven properties that should be satisfied by a dependence measure [3, 37], including symmetry, boundedness, monotonicity to Pearson's correlation for Gaussian variables, etc. Both distance covariance and MI satisfy the same five properties out of seven.[1]

In addition to these properties, we have the following two propositions saying that *consistency to independence* and *upper bound to maximal covariance* also hold for the distance covariance.

[1] These two measures do not satisfy properties (c) and (e) as defined in [3]. The distance correlation (i.e., normalized distance covariance) can further satisfy these two properties. However, we still use distance covariance for brevity as the two properties have little computational benefit. Due to the space limit, we do not include the proofs for these properties.

Proposition 3. **Consistency to independence**. *The distance covariance $\mathcal{V}^2(X,Y)$ is non-negative, and it achieves value zero if and only if X and Y are independent random variables.*

Proposition 4. **Upper bound to maximal covariance**. *For the covariance $\mathrm{Cov}(f(X),g(Y))$ of any two given functions f and g, we have*

$$\mathcal{V}^2(X,Y) \geq \frac{\mathrm{Cov}(f(X),g(Y))^2}{\|f^2\|_2\|g^2\|_2}.$$

Proposition 3 can be proved through Jensen's inequality [8], and the proof for Proposition 4 is in Appendix A.2.

Next, we further compare the tightness of MI and distance covariance with the following theorem, which states that distance covariance is a lower bound for MI. In other words, distance covariance is a tighter upper bound to the maximal covariance compared with MI.

Theorem 1. **Lower bound to MI**. *If $p_{(X,Y)}$ and $p_X \otimes p_Y$ is (upper) bounded by τ_1 and τ_2, respectively, then*

$$\mathcal{I}(X,Y) \geq \frac{1-\log(2)}{\max(\tau_1,\tau_2)}\mathcal{V}^2(X,Y).$$

Proof. Let p and q be any given distributions, and define $\eta(x) = (q(x)-p(x))/p(x)$. Then, the KL divergence can be rewritten as follows,

$$D_{KL}(p\|q) = \int_{\mathcal{X}} p(x)\log(p(x)/q(x))\,dx = -\int_{\mathcal{X}} p(x)\log(1+\eta(x))\,dx.$$

We define $A := \{x \mid \eta(x) > 1\} = \{x \mid q(x) > 2p(x)\}$ and $B := \{x \mid \eta(x) \leq 1\} = \{x \mid q(x) \leq 2p(x)\}$. Then, we can obtain that

(1) for $x \in A$, $(1+\eta(x)) \leq e^{a\eta(x)}$, where $a = \log(2)$;
(2) for $x \in B$, $(1+\eta(x)) \leq e^{\eta(x)-b\eta(x)^2}$, where $b = 1-\log(2)$.

Note that we have

$$\int_{\mathcal{X}} p(x)\eta(x)\,dx = \int_{\mathcal{X}} (q(x)-p(x))\,dx = 0,$$

which implies that $\int_A p(x)\eta(x)\,dx = -\int_B p(x)\eta(x)\,dx$. Putting all together, we have

$$\begin{aligned}
&D_{KL}(p\|q)\\
=&-\int_A p(x)\log(1+\eta(x))\,dx - \int_B p(x)\log(1+\eta(x))\,dx\\
\geq&-a\int_A p(x)\eta(x)\,dx - \int_B p(x)\eta(x)\,dx + b\int_B p(x)\eta(x)^2\,dx\\
=&(1-a)\int_A p(x)\eta(x)\,dx + b\int_B p(x)\eta(x)^2\,dx\\
=&(1-\log(2))\left(\int_A |q(x)-p(x)|\,dx + \int_B p(x)\left(\frac{q(x)-p(x)}{p(x)}\right)^2 dx\right).
\end{aligned}$$

Now, we switch to the mutual information, and we have

$$\begin{aligned}
\mathcal{I}(X,Y) \geq (1-\log(2))\Bigg(&\int_A |p_X(x)p_Y(y) - p_{(X,Y)}(x,y)|\,dx\,dy\\
&+\int_B \frac{|p_X(x)p_Y(y)-p_{(X,Y)}(x,y)|^2}{p_{(X,Y)}(x,y)}\,dx\,dy\Bigg),
\end{aligned}$$

where

$$A = \left\{(x,y) \mid p_X(x)p_Y(y) > 2p_{(X,Y)}(x,y)\right\},$$
$$B = \left\{(x,y) \mid p_X(x)p_Y(y) \leq 2p_{(X,Y)}(x,y)\right\}.$$

Hence, for the first summand in the parenthesis of the RHS,

$$\begin{aligned}
&\int_A |p_X(x)p_Y(y) - p_{(X,Y)}(x,y)|\,dx\,dy\\
=&\int_A \left|\frac{p_{(X,Y)}(x,y)}{p_X(x)p_Y(y)} - 1\right| p_X(x)p_Y(y)\,dx\,dy\\
\geq&\int_A \left|\frac{p_{(X,Y)}(x,y)}{p_X(x)p_Y(y)} - 1\right|^2 p_X(x)p_Y(y)\,dx\,dy\\
=&\int_A \left|\frac{p_{(X,Y)}(x,y)}{p_X(x)p_Y(y)} - 1\right|^2 \frac{(p_X(x)p_Y(y))^2}{p_X(x)p_Y(y)}\,dx\,dy\\
=&\int_A \frac{|p_X(x)p_Y(y)-p_{(X,Y)}(x,y)|^2}{p_X(x)p_Y(y)}\,dx\,dy\\
\geq&\frac{1}{\max(\tau_1,\tau_2)}\int_A |p_X(x)p_Y(y)-p_{(X,Y)}(x,y)|^2\,dx\,dy.
\end{aligned}$$

For the second summand in the parenthesis of the RHS,

$$\begin{aligned}
&\int_B \frac{|p_X(x)p_Y(y)-p_{(X,Y)}(x,y)|^2}{p_{(X,Y)}(x,y)}\,dx\,dy\\
\geq&\frac{1}{\max(\tau_1,\tau_2)}\int_B |p_X(x)p_Y(y)-p_{(X,Y)}(x,y)|^2\,dx\,dy.
\end{aligned}$$

Finally, we have

$$\begin{aligned}
\mathcal{I}(X,Y) &\geq \frac{1-\log(2)}{\max(\tau_1,\tau_2)}\int_{\mathcal{X}}\int_{\mathcal{Y}} |p_X(x)p_Y(y)-p_{(X,Y)}(x,y)|^2\,dx\,dy\\
&= \frac{1-\log(2)}{\max(\tau_1,\tau_2)}\mathcal{V}^2(X,Y),
\end{aligned}$$

which completes the proof. □

We next show an asymptotic equivalence between distance covariance and MI. Elaborately, as the distance covariance is minimized, we can consider a "nearly" independent case, i.e., $p_{(X,Y)} \approx p_X \otimes p_Y$, and define a reference distribution $r_{(X,Y)}$ as follows,

$$\begin{aligned}
p_{(X,Y)} &= r_{(X,Y)} + \epsilon\phi_{(X,Y)},\\
p_X \otimes p_Y &= r_{(X,Y)} + \epsilon\varphi_{(X,Y)},
\end{aligned} \tag{5}$$

where $\epsilon > 0$ and $\phi_{(X,Y)}, \varphi_{(X,Y)}$ are two individual perturbations. We assume they are valid additive perturbations satisfying [30]:

$$\int_{\mathcal{X}}\int_{\mathcal{Y}} \phi_{(X,Y)}(x,y)\,dx\,dy = 0,\ \int_{\mathcal{X}}\int_{\mathcal{Y}} \varphi_{(X,Y)}(x,y)\,dx\,dy = 0.$$

Theorem 2. **Asymptotic equivalence to MI**. *If the distributions $p_{(X,Y)}$ and $p_X \otimes p_Y$ can be expressed by $r_{(X,Y)}$, $\phi_{(X,Y)}$, and $\varphi_{(X,Y)}$ as in Eq. (5), MI between them can be approximated as*

$$\mathcal{I}(X,Y) \approx \frac{1}{2}\mathcal{V}^2_{1/r_{(X,Y)}}(X,Y),$$

where

$$\mathcal{V}^2_{1/r_{(X,Y)}}(X,Y) = \int_{\mathcal{X}}\int_{\mathcal{Y}} \frac{|p_{(X,Y)}(x,y)-p_X(x)p_Y(y)|^2}{r_{(X,Y)}(x,y)}\,dx\,dy.$$

This theorem can be proved by applying Taylor's theorem to the odds ratio function $f(t) = t\log(t)$ of the KL divergence [35], and details can be found in Appendix A.3. In this sense, MI can be interpreted as a weighted distance covariance when X and Y are close to independence.

3 THE PROPOSED FAIRDISCO APPROACH

In this section, we first formulate the problem under a variational framework, and then show how to compute the proposed penalty term, followed by some connection analysis to existing work.

3.1 Problem Formulation

We formulate the fair representation learning problem under the variational framework. Given a dataset $D = \{(\mathbf{x}_i, \mathbf{s}_i)\}_{i=1}^N$, where $\mathbf{x}_i \in \mathcal{X}$ and $\mathbf{s}_i \in \mathcal{S}$. The sensitive attribute $\mathbf{s}$ is usually categorical (e.g., gender, race, and age group). The goal of fair representation learning is to train an encoder to transform $(\mathbf{x}, \mathbf{s})$ to a continuous representation $\mathbf{z} \in \mathcal{Z}$, which should be expressive in terms of serving the downstream predictions while ensuring the fairness with respect to $\mathbf{s}$.

To tackle this problem, one often admits an assumption that the data point $(\mathbf{x}_i, \mathbf{s}_i)$ is generated by a random process consisting of two steps: (1) generating $\mathbf{z}_i$ and $\mathbf{s}_i$ from distributions $p(\mathbf{z})$ and $p(\mathbf{s})$, respectively; and (2) generating $\mathbf{x}_i$ from the distribution $p(\mathbf{x}|\mathbf{z}, \mathbf{s})$. This random process can be modeled as a general probabilistic graphical model [28], where the observable $\mathbf{x}$ is enforced to be generated from two individual sources $\mathbf{z}$ and $\mathbf{s}$ so that their correlation can be reduced.

Let the generative model (i.e., decoder) $p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{s})$ be parameterized by $\theta \in \Theta$, and the variational posterior (i.e., encoder) $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{s})$ be parameterized by $\phi \in \Phi$. We use a multivariate Gaussian with diagonal covariance to form the posterior $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{s})$, i.e., $q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{s}) = \mathcal{N}_\phi(\mathbf{z}; \boldsymbol{\mu}, \mathrm{diag}(\boldsymbol{\sigma}^2))$, and use a standard multivariate Gaussian $\mathcal{N}(0, \mathbf{I})$ to form the prior $p(\mathbf{z})$.

Limitations of Existing Solutions. With the above variational framework, existing work mainly incorporates an upper bound of MI as the penalty. For example, the following upper bound has been essentially used by several existing work [32, 33, 39],

$$\mathcal{I}(\mathbf{z}, \mathbf{s}) \leq \mathcal{I}(\mathbf{z}; \mathbf{x}, \mathbf{s}) \leq \mathbb{E}_{p(\mathbf{x},\mathbf{s})}[D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{s})\|p(\mathbf{z}))]. \quad (6)$$

Notice that $\mathcal{I}(\mathbf{z}; \mathbf{x}, \mathbf{s}) = \mathcal{I}(\mathbf{z}, \mathbf{s}) + \mathcal{I}(\mathbf{z}, \mathbf{x}|\mathbf{s})$. Thus, although tractable, Eq. (6) is also the upper bound of $\mathcal{I}(\mathbf{z}, \mathbf{x}|\mathbf{s})$, which represents the expressiveness of $\mathbf{z}$, and it enforces the posterior close to the prior in a similar way as β-VAE [18]. Consequently, fairness may be achieved by Eq. (6) at the expense of sacrificing expressiveness/utility. In view of this, a tighter upper bound is given by Song et al. [42]:

$$\mathcal{I}(\mathbf{z}, \mathbf{s}) \leq \mathbb{E}_{q_\phi(\mathbf{z},\mathbf{s})}[D_{KL}(p(\mathbf{s}|\mathbf{z})\|p(\mathbf{s}))]. \quad (7)$$

However, this upper bound is intractable because of posterior probability $p(\mathbf{s}|\mathbf{z})$ is unknown, and adversarial training is needed to approximate it. The similar adversarial-based upper bound is used by Creager et al. [9], which minimizes the KL divergence of two distributions using the density-ratio trick. However, it was observed that adversarial training might be unstable and sometimes counterproductive in the context of fair representation learning [32].

Our Objective Function. To tackle the limitations of MI-based penalties, we use the penalty term derived from the distance covariance to ensure the independence of $\mathbf{z}$ and $\mathbf{s}$, i.e.,

$$\mathcal{V}_\phi^2(\mathbf{z}, \mathbf{s}) = \int_{\mathcal{Z}} \int_{\mathcal{S}} |p_\phi(\mathbf{z}, \mathbf{s}) - p_\phi(\mathbf{z})p(\mathbf{s})|^2 \, d\mathbf{z} \, d\mathbf{s}. \quad (8)$$

Therefore, the optimization problem of fair representation learning can be formulated as follows,

$$\max_{\phi,\theta} \Big\{ \log p_\theta(\mathbf{x}|\mathbf{s}) - \beta \mathcal{V}_\phi^2(\mathbf{z}, \mathbf{s}) \Big\}, \quad (9)$$

where β is a balancing parameter. A larger β means more attention is paid to fairness.

For the first utility term in the above formulation, both θ and ϕ can be jointly optimized via the SGVB algorithm which maximizes a variational lower bound of the marginal likelihood [23]:

$$\begin{aligned} \log p_\theta(\mathbf{x}|\mathbf{s}) \geq\ & \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{s})}[\log p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{s})] \\ & - \mathbb{E}_{p(\mathbf{x},\mathbf{s})}[D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{s})\|p(\mathbf{z}))]. \end{aligned} \quad (10)$$

Putting together, we can obtain a tractable lower bound of Eq. (9) as follows,

$$\begin{aligned} \max_{\phi,\theta} \Big\{ & \mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x},\mathbf{s})}[\log p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{s})] - D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}, \mathbf{s})\|p(\mathbf{z})) \\ & - \beta \int_{\mathcal{Z}} \int_{\mathcal{S}} |p_\phi(\mathbf{z}, \mathbf{s}) - p_\phi(\mathbf{z})p(\mathbf{s})|^2 \, d\mathbf{z} \, d\mathbf{s} \Big\}, \end{aligned} \quad (11)$$

where minimizing the penalty (i.e., Eq. (8)) will not contradict the maximization of the utility (i.e., Eq. (10)).

3.2 Penalty Computation

We next show the computation of the penalty term in Eq. (11). We assume that $\mathbf{s}$ is a discrete/categorical variable, and we simplify the value space of $\mathbf{s}$ as $\mathcal{S} = \{1, 2..., K\}$ without loss of generality. Using the definition of Euclidean distance and the law of total probability, the penalty term can be rewritten as

$$\begin{aligned} \mathcal{V}^2(\mathbf{z}, \mathbf{s}) &= \int_{\mathcal{Z}} \sum_{k=1}^{K} |p(\mathbf{z}, \mathbf{s} = k) - p(\mathbf{z})p(\mathbf{s} = k)|^2 \, d\mathbf{z} \\ &= \sum_{k=1}^{K} p(\mathbf{s} = k)^2 \int_{\mathcal{Z}} |p(\mathbf{z}|\mathbf{s} = k) - p(\mathbf{z})|^2 \, d\mathbf{z} \\ &= \sum_{k=1}^{K} p(\mathbf{s}^k)^2 p(\mathbf{s}^{\neg k})^2 \int_{\mathcal{Z}} |p(\mathbf{z}|\mathbf{s}^k) - p(\mathbf{z}|\mathbf{s}^{\neg k})|^2 \, d\mathbf{z}, \end{aligned}$$

where the last equation is derived by the decomposition $p(\mathbf{z}) = p(\mathbf{z}, \mathbf{s} = k) + p(\mathbf{z}, \mathbf{s} \neq k)$, and we denote $p(\mathbf{s} = k)$ and $p(\mathbf{s} \neq k)$ by $p(\mathbf{s}^k)$ and $p(\mathbf{s}^{\neg k})$ for simplicity. We further compute the two terms in the integral by adding the expectation over $p(\mathbf{x}|\mathbf{s}^k)$ and $p(\mathbf{x}|\mathbf{s}^{\neg k})$, i.e.,

$$\begin{aligned} p(\mathbf{z}|\mathbf{s}^k) &= \mathbb{E}_{p(\mathbf{x}|\mathbf{s}^k)}[p(\mathbf{z}|\mathbf{x}, \mathbf{s}^k)], \\ p(\mathbf{z}|\mathbf{s}^{\neg k}) &= \mathbb{E}_{p(\mathbf{x}|\mathbf{s}^{\neg k})}[p(\mathbf{z}|\mathbf{x}, \mathbf{s}^{\neg k})]. \end{aligned}$$

Now, the distance covariance penalty only involves distributions $p(\mathbf{s})$, $p(\mathbf{x}|\mathbf{s})$, and $p(\mathbf{z}|\mathbf{x}, \mathbf{s})$. It should be noted that $\mathbf{s}$ and $\mathbf{x}$ are both observables, and thus we can estimate the distributions $p(\mathbf{s})$ based on the dataset D. For details, the approximations $q(\mathbf{s})$ is

$$q(\mathbf{s}^k) = \frac{N_k}{N}, \quad q(\mathbf{s}^{\neg k}) = \frac{N_{\neg k}}{N}$$

where N_k means the number of $\mathbf{s}$ that takes value k, and $N_{\neg k}$ means the number of $\mathbf{s}$ that do not takes value k.

Next, we adopt the variational posterior $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{s})$ as an approximation of the last term $p(\mathbf{z}|\mathbf{x},\mathbf{s})$. Using Monte Carlo estimate of the distance covariance, we have

$$\mathcal{V}^2_\phi(\mathbf{z},\mathbf{s}) \approx \sum_{k=1}^{K} \frac{N_k^2 N_{\neg k}^2}{N^4} \int_{\mathcal{Z}} |\bar{q}_\phi(\mathbf{z}|\mathbf{s}^k) - \bar{q}_\phi(\mathbf{z}|\mathbf{s}^{\neg k})|^2 \, dz, \tag{12}$$

where

$$\begin{aligned} \bar{q}_\phi(\mathbf{z}|\mathbf{s}^k) &= \frac{1}{N_k} \sum_{\substack{(\mathbf{x}_i,\mathbf{s}_i)\in\mathcal{D} \\ \mathbf{s}_i=k}} q_\phi(\mathbf{z}|\mathbf{x}_i,\mathbf{s}_i), \\ \bar{q}_\phi(\mathbf{z}|\mathbf{s}^{\neg k}) &= \frac{1}{N_{\neg k}} \sum_{\substack{(\mathbf{x}_i,\mathbf{s}_i)\in\mathcal{D} \\ \mathbf{s}_i\neq k}} q_\phi(\mathbf{z}|\mathbf{x}_i,\mathbf{s}_i). \end{aligned} \tag{13}$$

Since $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{s})$ is modeled as a Gaussian distribution, the integration term over $\mathbf{z}$ can be viewed as the squared Euclidean distance between two Gaussian mixture models $\bar{q}_\phi(\mathbf{z}|\mathbf{s}^k)$ and $\bar{q}_\phi(\mathbf{z}|\mathbf{s}^{\neg k})$ (in the following, we denote them by $\bar{q}^k_\phi$ and $\bar{q}^{\neg k}_\phi$ for brevity) [38], and its closed-form expression is [36, Sec. 8.1.8]

$$\begin{aligned} \delta^2_E(\bar{q}^k_\phi, \bar{q}^{\neg k}_\phi) = &\frac{1}{N_k^2} \sum_{\substack{i=1,\\ \mathbf{s}_i=k}}^{N} \sum_{\substack{j=1,\\ \mathbf{s}_j=k}}^{N} \mathcal{N}\big(\boldsymbol{\mu}_i; \boldsymbol{\mu}_j, \mathrm{diag}(\boldsymbol{\sigma}_i^2 + \boldsymbol{\sigma}_j^2)\big) \\ &+ \frac{1}{N_{\neg k}^2} \sum_{\substack{i=1,\\ \mathbf{s}_i\neq k}}^{N} \sum_{\substack{j=1,\\ \mathbf{s}_j\neq k}}^{N} \mathcal{N}\big(\boldsymbol{\mu}_i; \boldsymbol{\mu}_j, \mathrm{diag}(\boldsymbol{\sigma}_j^2 + \boldsymbol{\sigma}_j^2)\big) \\ &- \frac{2}{N_k N_{\neg k}} \sum_{\substack{i=1,\\ \mathbf{s}_i=k}}^{N} \sum_{\substack{j=1,\\ \mathbf{s}_j\neq k}}^{N} \mathcal{N}\big(\boldsymbol{\mu}_i; \boldsymbol{\mu}_j, \mathrm{diag}(\boldsymbol{\sigma}_i^2 + \boldsymbol{\sigma}_j^2)\big). \end{aligned} \tag{14}$$

Algorithm and Analysis. In practice, we need to update the parameters in each mini batch. For the estimation of our penalty term (i.e., Eq. (12)), it requires $O(K * d * B^2)$ time where K means the number of values of discrete variable $\mathbf{s}$, d means the dimension of latent space $\mathcal{Z}$, and B is the batch size. We can use the matrix algebra to compute the distance covariance penalty, and further accelerate its computation via parallelism. The overall algorithm is summarized in Alg. 1.

3.3 Relationship with Other Penalty Functions

In addition to MI-based penalty, VFAE [28] proposes to impose the independence constraint via maximal mean discrepancy (MMD). Specifically, it uses the distance between the data sampled from two distributions, which are actually the Gaussian mixture models in Eq. (13), as a biased empirical estimate of MMD [15]. Compared with our penalty term, VFAE conducts an additional sampling process which could hurt both accuracy and efficiency. Moreover, if we ideally compute the exact MMD through the two Gaussian mixture models rather than estimate it by the distance as done by VFAE, we have the following proposition, which shows that the closed-form expression of MMD is similar to Eq. (14) but contains an extra noise term $\frac{1}{2\gamma}I$.

Proposition 5. *Let the kernel of MMD be a radial basis function $\kappa(x,y) = \exp(-\gamma\|x-y\|^2)$, and the two Gaussian mixture models $\bar{q}_\phi(\mathbf{z}|\mathbf{s}^k)$ and $\bar{q}_\phi(\mathbf{z}|\mathbf{s}^{\neg k})$ be defined as in Eq. (13). Then, MMD between $\bar{q}^k_\phi$ and $\bar{q}^{\neg k}_\phi$ can be computed as follows.*

$$\begin{aligned} \mathrm{MMD}(\bar{q}^k_\phi, \bar{q}^{\neg k}_\phi)^2 = &\frac{1}{N_k^2} \sum_{\substack{i=1,\\ \mathbf{s}_i=k}}^{N} \sum_{\substack{j=1,\\ \mathbf{s}_j=k}}^{N} \kappa(q_\phi(\mathbf{z}|\mathbf{x}_i,\mathbf{s}_i), q_\phi(\mathbf{z}|\mathbf{x}_j,\mathbf{s}_j)) \\ &+ \frac{1}{N_{\neg k}^2} \sum_{\substack{i=1,\\ \mathbf{s}_i\neq k}}^{N} \sum_{\substack{j=1,\\ \mathbf{s}_j\neq k}}^{N} \kappa(q_\phi(\mathbf{z}|\mathbf{x}_i,\mathbf{s}_i), q_\phi(\mathbf{z}|\mathbf{x}_j,\mathbf{s}_j)) \\ &- \frac{2}{N_k N_{\neg k}} \sum_{\substack{i=1,\\ \mathbf{s}_i=k}}^{N} \sum_{\substack{j=1,\\ \mathbf{s}_j\neq k}}^{N} \kappa(q_\phi(\mathbf{z}|\mathbf{x}_i,\mathbf{s}_i), q_\phi(\mathbf{z}|\mathbf{x}_j,\mathbf{s}_j)), \end{aligned}$$

where

$$\kappa(q_\phi(\mathbf{z}|\mathbf{x}_i,\mathbf{s}_i), q_\phi(\mathbf{z}|\mathbf{x}_j,\mathbf{s}_j)) = (\frac{\pi}{\gamma})^{\frac{d}{2}} \mathcal{N}(\boldsymbol{\mu}_i; \boldsymbol{\mu}_j, \mathrm{diag}(\boldsymbol{\sigma}_1+\boldsymbol{\sigma}_2) + \frac{1}{2\gamma}\mathbf{I}).$$

Algorithm 1 The FairDisCo Algorithm

Input: dataset $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{s}_i)\}_{i=1}^N$, and penalty coefficient β;
Output: encoder $f_\phi(\mathbf{x},\mathbf{s})$ and decoder $f_\theta(\mathbf{z},\mathbf{s})$.

for each batch $\mathcal{D}_{batch} = \{(\mathbf{x}_i,\mathbf{s}_i)\}_{i=1}^B$ sampled from $\mathcal{D}$ **do**
  let $q_\phi(\mathbf{z}|\mathbf{x}_i,\mathbf{s}_i) = \mathcal{N}(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i)$ where $(\boldsymbol{\mu}_i, \boldsymbol{\sigma}_i) = f_\phi(\mathbf{x}_i,\mathbf{s}_i)$;
  calculate KL divergence loss: $L_{kl} = \sum_{i=1}^B D_{KL}(q_\phi(\mathbf{z}|\mathbf{x}_i,\mathbf{s}_i)\|p(\mathbf{z}))$;
  calculate distance covariance loss: $L_{fair} = \mathcal{V}^2_\phi(\mathbf{z},\mathbf{s})$;
  for each data point $(\mathbf{x}_i,\mathbf{s}_i) \in \mathcal{D}_{batch}$ **do**
    sample $\mathbf{z}_i \sim q_\phi(\mathbf{z}|\mathbf{x}_i,\mathbf{s}_i)$;
    if $\mathbf{x}$ is discrete **then**
      $p_\theta(\mathbf{x}|\mathbf{z}_i,\mathbf{s}_i) = \mathrm{Cat}(\boldsymbol{p}_i)$ where $\boldsymbol{p}_i = f_\theta(\mathbf{z}_i,\mathbf{s}_i)$;
    else if $\mathbf{x}$ is continuous **then**
      $p_\theta(\mathbf{x}|\mathbf{z}_i,\mathbf{s}_i) = \mathcal{N}(\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\sigma}}_i)$ where $(\hat{\boldsymbol{\mu}}_i, \hat{\boldsymbol{\sigma}}_i) = f_\theta(\mathbf{z}_i,\mathbf{s}_i)$;
    end if
  end for
  calculate reconstruction loss: $L_{re} = \sum_{i=1}^B \log(p_\theta(\mathbf{x}_i|\mathbf{z}_i,\mathbf{s}_i))$;
  put all losses together: $L = L_{kl} - L_{re} + \beta * L_{fair}$;
  update parameters ϕ and θ via the gradient descent of L;
end for

4 EXPERIMENTAL EVALUATIONS

In this section, we present the experimental results.

4.1 Experimental Setup

Datasets. We perform experiments on two large public real-world datasets that are commonly used in the fair machine learning community [28, 32, 39, 48]. The *Adult* dataset[2] contains 45,222 individuals each of which is described by some attributes (e.g., gender, education level, age, etc.). We use gender as the sensitive attribute, and the downstream task is to predict whether an individual earns more than \$50K/year. The *Heritage Health* dataset[3] contains 115,143 entries and each entry describes an patient (e.g., age, gender, physiological indexes, etc.). We use age group (i.e., whether a patient is

[2]https://archive.ics.uci.edu/ml/datasets/adult
[3]https://www.kaggle.com/c/hhp

older than 75) as the sensitive attribute, and the downstream task is to predict whether a patient will go into hospital in the next year.

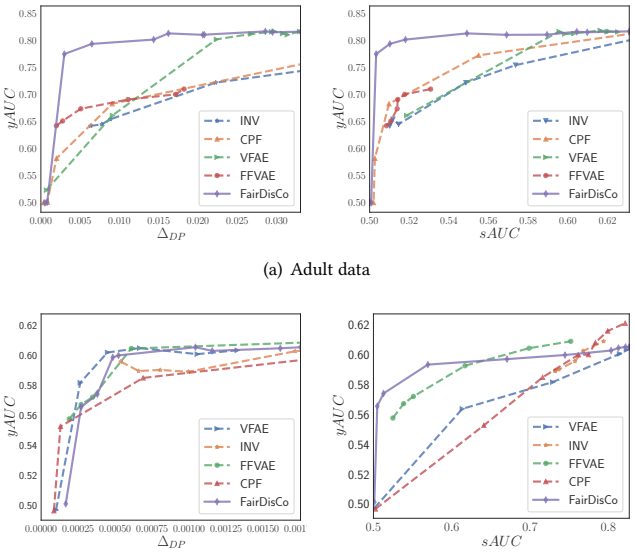
Metrics. To evaluate the effectiveness of the learned representations, we consider the following metrics. For fairness, we consider demographic parity Δ_{DP} [29] and an inference-based metric *sAUC*. Here, sAUC is the AUC result of inferring $\mathbf{s}$ from $\mathbf{z}$. For utility/expressiveness, we adopt the *yAUC* metric which is the AUC result of applying the learned representations in the downstream prediction task. The lower Δ_{DP}, higher yAUC, and closer sAUC to 0.5, the better.

Compared Methods. We compare our method to the following models: *VFAE* [28], *INV* [32], *FFVAE* [9], and *CPF* [39]. Among these competitors, *VFAE* uses an MMD penalty, and the latter three use MI-related penalties. In particular, *INV* and *CPF* relax MI to the upper bound in Eq. (6), and *FFVAE* approximates an intractable upper bound with adversarial learning.

Reproducibility. We tune the hyperparameters for each compared method. Specifically, for *VFAE*, we search the fairness penalty coefficient $\beta \in \{10^k | k = 0, 1, ..., 10\}$. For *INV*, we search its fairness penalty coefficient $\lambda \in \{0.1, 0.5, 1, 10, 50, 100, 1000\}$. For *FFVAE*, we tune its predictiveness coefficient α in $\{100, 200, 300, 500\}$ and observe little differences; thus we fix $\alpha = 300$, and tune the fairness/disentanglement coefficient $\gamma \in \{1, 5, 10, 20, 40, 60, 100\}$. We search the fairness penalty coefficient $\beta \in \{1, 1.5, 2, 2.5, 3, 5, 10, 100\}$ for *CPF*. For our own method, we set β the same as that in *VFAE*. For all datasets and baselines, we assume latent space $\mathcal{Z}$ has 8 dimensions, and use Adam optimizer with initial learning rate 0.001 and training epoch 1,000. For the Adult dataset, we use the same train/test split as existing work [28, 32, 39, 48]. For the Health dataset, we random split 80% data as the training set and use the rest as the test set. For downstream classification task and inference task (i.e., computing *sAUC*), we utilize a powerful non-linear model Random Forest as the classifier. The density $p_\theta(\mathbf{x}|\mathbf{z}, \mathbf{s})$ is modeled as a product of categorical distributions. We use one hidden layer of 64 units MLP to approximate density, ReLU as the activation function. All the experiments were carried out on a server equipped with 256GB RAM, one 16-core Intel i9-9900K CPU@3.60GHz and one NVIDIA GeForce RTX 2080Ti GPU. The datasets and the code are available at https://github.com/SoftWiser-group/FairDisCo.

4.2 Experimental Results

(A) Classification Results. We first show the fairness-utility curves of the classification results in Fig 1. For the curves, we sweep a range of hyperparameters as mentioned above for each model, run each hyperparameter 10 times, and report the mean results. We can observe from the figures that our method achieves the best tradeoffs in most cases. Although there is an inherent tradeoff between fairness and utility [51], our method can ensure near-perfect fairness ($\Delta_{DP} \approx 0, sAUC \approx 0.5$) while achieving a higher accuracy ($yAUC$) in most cases. For example, on the Adult dataset, with near-perfect fairness satisfaction (0.0064 vs. 0.0110 in Δ_{DP} and 0.5032 vs. 0.5097 in $sAUC$, respectively), FairDisCo achieves 14.9% and 13.5% accuracy improvements compared to the best competitors (*FFVAE* and *CPF*, respectively). The *FFVAE* method based on adversarial training is less effective. This is due to the difficulty and instability



(a) Adult data

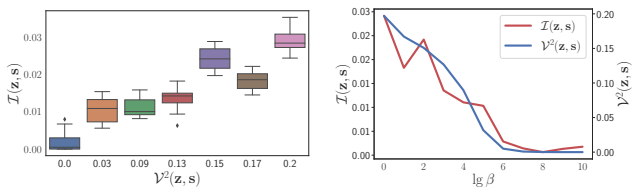
(b) Health data

Figure 1: Fairness-utility tradeoff curves. FairDisCo can ensure near-perfect fairness while generally achieving a higher accuracy than the competitors.

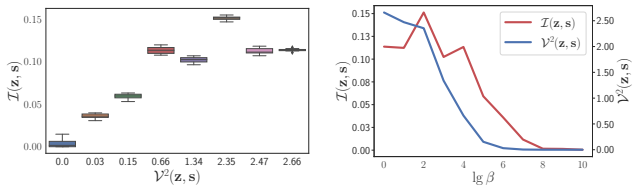
of training the discriminator. The proposed FairDisCo also outperforms the other two methods *INV* and *CPF*. The reason is that their upper bound of $\mathcal{I}(\mathbf{z}, \mathbf{s})$ is also an upper bound of $\mathcal{I}(\mathbf{z}, \mathbf{x}|\mathbf{s})$ [42]. Therefore, minimizing the upper bound may have negative impact on the utility aspect. *VFAE* is based on the MMD penalty and it is less effective than FairDisCo. This is due to the biased empirical estimate of MMD in practice.

We notice that on the Health dataset, almost every method achieves close tradeoffs between Δ_{DP} and $yAUC$, and Δ_{DP} is within a small interval $[0, 0.0015]$. We further estimate MI using the nearest-neighbor method [40]. We have $\mathcal{I}(\mathbf{s}, \mathbf{y}) = 0.00675$ for the Health data, and $\mathcal{I}(\mathbf{s}, \mathbf{y}) = 0.025431$ for the Adult data. This result indicates that the classification task on the Health dataset is inherently weakly correlated to the sensitive attribute. Therefore, DP is easier to achieve in this dataset. Still, we observe that FairDisCo is better than the competitors in tradeoffs between $sAUC$ and $yAUC$, meaning that the learned representations of FairDisCo contain less information about the sensitive attributes from the inference perspective.

(B) Mutual Information versus Distance Covariance. Next, we investigate the relationship between MI $\mathcal{I}(\mathbf{z}, \mathbf{s})$ and distance covariance $\mathcal{V}^2(\mathbf{z}, \mathbf{s})$ used in FairDisCo. The results are shown in Fig. 2, where we still use [40] to estimate $\mathcal{I}(\mathbf{z}, \mathbf{s})$ on the test set. The two figures on the left side of Fig. 2 demonstrate the correlation between between MI and distance covariance. Observe that these two measures are strongly positively correlated (with Pearson correlation coefficient 0.89). The two figures on the right side of Fig. 2 demonstrate how β controls MI and distance covariance. We can observe that these two measures have the same trend under different β's, and



(a) Adult data



(b) Health data

Figure 2: The relationship between MI and distance covariance. These two measures are strongly positively correlated, and share a similar trend converging to zero.

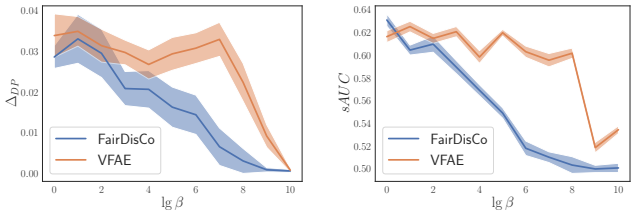
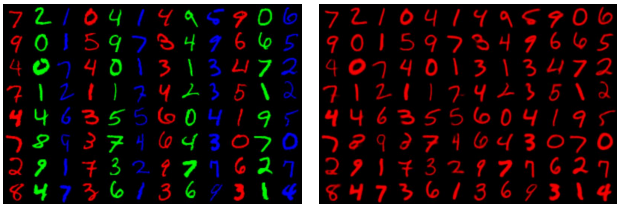


Figure 3: Fairness results under different hyperparameters.

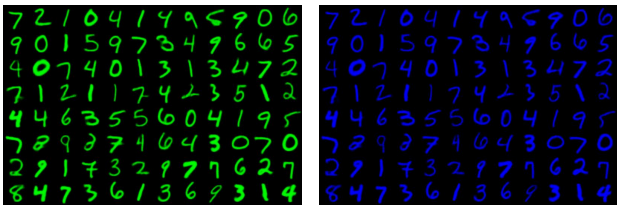
both converge to zero, which is consistent with our analysis in Section 2.2.

(C) Effect of Hyperparameters. We next investigate how the penalty coefficient β controls the fairness of FairDisCo, and show the results on the Adult dataset in Fig. 3. In the figure, we also show the results of *VFAE* for comparison since it shares the same hyperparameter setting as FairDisCo. From Fig. 3, we observe that Δ_{DP} and *sAUC* clearly decrease as β increases for both FairDisCo and *VFAE*. Additionally, FairDisCo achieves better fairness and is smoother in terms of controlling the fairness compared with *VFAE*. This is due to the fact that MMD may result in biased empirical estimation as discussed in Section 3.3.

(D) Visualization results on handwritten digit images. Finally, we provide some visualization results by applying FairDisCo on two handwritten digit datasets: Color MNIST and MNIST. Note that these two datasets do not have fairness issues and the purpose is to provide some visualization results with better interpretability. For Color MNIST dataset, we use the color of the digits as $\mathbf{s}$, and aim to encode all the non-color information into the latent representations $\mathbf{z}$. By manipulating the input $\mathbf{s}$ of the decoder in FairDisCo, we aim to generate different color digits with the same style. For MNIST dataset, we use the label of digital as $\mathbf{s}$, and we aim to generate other digits with the same style.



(a) Real digits (b) Generated red digits



(c) Generated green digits (d) Generated blue digits

Figure 4: Generated digits with different colors but the same style. Here, we set the color as the sensitive attribute. The result indicates that the color is disentangled from the learned representations.

Figure 5: Generated digits with different labels but the same style. Here, we set the digit label as the sensitive attribute. The result indicates that the digit label is disentangled from the learned representations.

The density $p_\theta(\mathbf{x}|\mathbf{z},\mathbf{s})$ is modeled as a multivariate Bernoulli distribution here. We assume latent space $\mathcal{Z}$ has 10 dimensions, and use the same encoder and decoder architecture as in [22] to approximate densities $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{s})$ and $p_\theta(\mathbf{x}|\mathbf{z},\mathbf{s})$. For the hyperparameter β, we set $lg\beta = 7$.

The results are shown in Fig. 4 and Fig. 5. For Fig. 4, we encoder a real digit by encoder $q_\phi(\mathbf{z}|\mathbf{x},\mathbf{s})$ to obtain a latent representation $\mathbf{z}$, and we manipulate $\mathbf{s} \in \{r, g, b\}$ for the decoder $p_\theta(\mathbf{x}|\mathbf{z},\mathbf{s})$ to generate a digit of the specified color with the same style. For Fig. 5, we manipulate the $\mathbf{s} \in \{0, 1, ..., 9\}$ for the decoder $p_\theta(\mathbf{x}|\mathbf{z},\mathbf{s})$ to generate a digit of the specified label with the same style. We can observe from the figures that FairDisCo successfully produces

digits with the specified colors and labels. This result indicates that the protected sensitive attribute in this experiment is indeed disentangled from the learned representations.

5 RELATED WORK

In literature, various fair representation learning methods have been proposed to ensure group fairness. For example, Zemel et al. [48] assign each data instance to certain prototypes as latent representations, and add a constraint on the prototype coefficients to ensure fairness. Louizos et al. [28] propose a variational framework for fair representation learning and add an MMD penalty to constrain the dependence between sensitive attributes and learned representations. Edwards and Storkey [11] formulate the problem as a min-max optimization problem, and train an adversary trying to predict sensitive attributes from the learned representations. Such adversarial training method is followed by several later proposals [29, 45, 49], due to the intractability of many MI-based upper bounds. Later, Song et al. [42] find that the commonly-used MI upper bound may contradict the utility of representation learning and thus achieve fairness at the expense of sacrificing utility. Therefore, they propose a tighter and intractable upper bound and solve it through adversarial training. Moyer et al. [32] identify the instability issues in adversarial training and thus present a tractable upper bound of MI without the need of adversarial training. Recently, Creager et al. [9] and Rodriguez et al. [39] adapt the progress of disentangled representation learning (e.g., FactorVAE [22] and β-VAE [18], respectively) into the fairness domain, and propose to achieve fairness via disentangling the effect of sensitive attributes from latent representations. Gitiaux and Rangwala [13] further extends β-VAE by modeling the latent representation as a binary bit stream.

In addition to demographic parity, other group fairness notions such as equalized odds [17] and accuracy parity [47] have also been studied. Essentially, these notions are defined under the *supervised* learning scenario requiring the availability of labels, while our current focus is on the *unsupervised* case. Extending our method to supervised settings is left as future work.

In addition to group fairness, individual fairness [10] and counterfactual fairness [25] have also received much recent attention. Kearns et al. [21] further study subgroup fairness, which interpolates between group fairness and individual fairness. The incompatibility and compatibility relationships between the above notions have also been studied [24, 50]. Recent work starts to consider the fairness issues in federated learning [19, 27], or when there are noisy sensitive features [6, 26].

6 CONCLUSIONS

In this paper, we advocate to use distance covariance, as a better alternative to the widely-used mutual information, to measure the dependence between sensitive attributes and learned representations. Compared with mutual information, distance covariance provides a tighter upper bound of maximal correlation. We incorporate the distance covariance as a penalty into a variational fair representation learning framework, and show that the penalty is tractable given that sensitive attributes are discrete or categorical. Experimental evaluations show that the proposed fair representation learning approach outperforms the existing competitors whose fairness penalties are based on mutual information and maximal mean discrepancy. In the future, we plan to extend the distance covariance measure to more fairness notions.

ACKNOWLEDGEMENTS

This work is supported by the National Natural Science Foundation of China (No. 62025202), the Collaborative Innovation Center of Novel Software Technology and Industrialization, and the Fundamental Research Funds for the Central Universities. Hanghang Tong is partially supported by NSF (1947135, 2134079 and 1939725). Yuan Yao is the corresponding author.

REFERENCES

[1] Narayanaswamy Balakrishnan and Chin Diew Lai. 2009. *Continuous bivariate distributions*. Springer Science & Business Media.
[2] Mohamed Ishmael Belghazi, Sai Rajeswar, Aristide Baratin, Devon Hjelm, and Aaron Courville. 2018. MINE: Mutual Information Neural Estimation. In *International Conference on Machine Learning (ICML)*.
[3] CB Bell. 1962. Mutual information and maximal correlation as measures of dependence. *The Annals of Mathematical Statistics* (1962), 587–595.
[4] Richard C Bradley. 1983. Equivalent measures of dependence. *Journal of Multivariate Analysis* 13, 1 (1983), 167–176.
[5] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science* 356, 6334 (2017), 183–186.
[6] L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. 2021. Fair classification with noisy protected attributes: A framework with provable guarantees. In *International Conference on Machine Learning (ICML)*. 1349–1361.
[7] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *International Conference on Knowledge Discovery and Data Mining (KDD)*. 797–806.
[8] Thomas M Cover. 1999. *Elements of information theory*. John Wiley & Sons.
[9] Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. 2019. Flexibly fair representation learning by disentanglement. In *International Conference on Machine Learning (ICML)*. 1436–1445.
[10] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *3rd innovations in theoretical computer science conference*. 214–226.
[11] Harrison Edwards and Amos Storkey. 2016. Censoring representations with an adversary. In *International Conference on Learning Representations (ICLR)*.
[12] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *International Conference on Knowledge Discovery and Data Mining (KDD)*. 259–268.
[13] Xavier Gitiaux and Huzefa Rangwala. 2021. Fair Representations by Compression. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.
[14] James Glimm and Arthur Jaffe. 2012. *Quantum physics: a functional integral point of view*. Springer Science & Business Media.
[15] Arthur Gretton, Karsten M Borgwardt, Malte J Rasch, Bernhard Schölkopf, and Alexander Smola. 2012. A kernel two-sample test. *The Journal of Machine Learning Research* 13, 1 (2012), 723–773.
[16] Larry Hardesty. 2018. Study finds gender and skin-type bias in commercial artificial-intelligence systems. *Retrieved April* 3 (2018), 2019.
[17] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. 3323–3331.
[18] Irina Higgins, Loic Matthey, Arka Pal, Christopher Burgess, Xavier Glorot, Matthew Botvinick, Shakir Mohamed, and Alexander Lerchner. 2017. beta-vae: Learning basic visual concepts with a constrained variational framework. In *International Conference on Learning Representations (ICLR)*.
[19] Junyuan Hong, Zhuangdi Zhu, Shuyang Yu, Zhangyang Wang, Hiroko H Dodge, and Jiayu Zhou. 2021. Federated adversarial debiasing for fair and transferable representations. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*. 617–627.
[20] Surya Mattu Julia Angwin, Jeff Larson and Lauren Kirchner. 2016. Machine Bias: There's software used across the country to predict future criminals. *ProPublica* (2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[21] Michael Kearns, Seth Neel, Aaron Roth, and Zhiwei Steven Wu. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. In *International Conference on Machine Learning (ICML)*. 2564–2572.

[22] Hyunjik Kim and Andriy Mnih. 2018. Disentangling by Factorising. In *International Conference on Machine Learning (ICML)*. 2654–2663.

[23] Diederik P. Kingma and Max Welling. 2014. Auto-Encoding Variational Bayes. In *International Conference on Learning Representations (ICLR)*.

[24] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *Innovations in Theoretical Computer Science Conference (ITCS)*, Christos H. Papadimitriou (Ed.). 43:1–43:23.

[25] MJ Kusner, J Loftus, Christopher Russell, and R Silva. 2017. Counterfactual Fairness. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*, Vol. 30.

[26] Alexandre Lamy, Ziyuan Zhong, Aditya Krishna Menon, and Nakul Verma. 2019. Noise-tolerant fair classification. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems (NeurIPS)*. 294–306.

[27] Tian Li, Shengyuan Hu, Ahmad Beirami, and Virginia Smith. 2021. Ditto: Fair and robust federated learning through personalization. In *International Conference on Machine Learning*. PMLR, 6357–6368.

[28] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S Zemel. 2016. The Variational Fair Autoencoder. In *International Conference on Learning Representations (ICLR)*.

[29] David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. 2018. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning (ICML)*. 3384–3393.

[30] Anuran Makur. 2015. *A study of local approximations in information theory*. Ph. D. Dissertation. Massachusetts Institute of Technology.

[31] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)* 54, 6 (2021), 1–35.

[32] Daniel Moyer, Shuyang Gao, Rob Brekelmans, Greg Ver Steeg, and Aram Galstyan. 2018. Invariant representations without adversarial training. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. 9102–9111.

[33] Lihao Nan and Dacheng Tao. 2020. Variational approach for privacy funnel optimization on continuous data. *J. Parallel and Distrib. Comput.* 137 (2020), 17–25.

[34] Hoang Vu Nguyen, Emmanuel Müller, Jilles Vreeken, Pavel Efros, and Klemens Böhm. 2014. Multivariate maximal correlation analysis. In *International Conference on Machine Learning*. PMLR, 775–783.

[35] Frank Nielsen and Richard Nock. 2014. On the chi square and higher-order chi distances for approximating f-divergences. *IEEE Signal Processing Letters* 21, 1 (2014), 10–13. https://doi.org/10.1109/LSP.2013.2288355

[36] Kaare Brandt Petersen, Michael Syskind Pedersen, et al. 2008. The matrix cookbook. *Technical University of Denmark* 7, 15 (2008), 510.

[37] A. Rényi. 1959. On measures of dependence. *Acta Mathematica Academiae Scientiarum Hungarica* 10, 3 (1959), 441–451. https://doi.org/10.1007/BF02024507

[38] Douglas A Reynolds. 2009. Gaussian Mixture Models. *Encyclopedia of biometrics* 741 (2009), 659–663.

[39] Borja Rodríguez-Gálvez, Ragnar Thobaben, and Mikael Skoglund. 2021. A Variational Approach to Privacy and Fairness. In *AAAI Workshop on Privacy-Preserving Artificial Intelligence*.

[40] Brian C Ross. 2014. Mutual information between discrete and continuous data sets. *PloS one* 9, 2 (2014), e87357.

[41] Jiaming Song and Stefano Ermon. 2020. Understanding the limitations of variational mutual information estimators. In *International Conference on Learning Representations (ICLR)*.

[42] Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. 2019. Learning controllable fair representations. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*. PMLR, 2164–2173.

[43] Gábor J Székely, Maria L Rizzo, and Nail K Bakirov. 2007. Measuring and testing dependence by correlation of distances. *The annals of statistics* 35, 6 (2007), 2769–2794.

[44] Michael M Wolf, Frank Verstraete, Matthew B Hastings, and J Ignacio Cirac. 2008. Area laws in quantum systems: mutual information and correlations. *Physical review letters* 100, 7 (2008), 070502.

[45] Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. 2017. Controllable invariance through adversarial feature learning. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*. 585–596.

[46] Yaming Yu. 2008. On the maximal correlation coefficient. *Statistics & Probability Letters* 78, 9 (2008), 1072–1075.

[47] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. 2017. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *International conference on world wide web (WWW)*. 1171–1180.

[48] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *International conference on machine learning (ICML)*. 325–333.

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

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

[51] Han Zhao and Geoffrey J Gordon. 2019. Inherent tradeoffs in learning fair representations. In *Annual Conference on Neural Information Processing Systems (NeurIPS)*.

A APPENDIX

A.1 Proof of Proposition 2

Proof.

$$\begin{aligned}
\mathcal{I}(X,Y) &= D_{KL}(p_{(X,Y)} \| p_X \otimes p_Y) \geq 2\delta_{TV}(p_{(X,Y)}, p_X \otimes p_Y)^2 \\
&= \frac{1}{2}\left(\int_{\mathcal{X}}\int_{\mathcal{Y}} |p_{(X,Y)}(x,y) - p_X(x)p_Y(y)|\, dx\, dy\right)^2 \\
&\geq \frac{1}{2}\left(\int_{\mathcal{X}}\int_{\mathcal{Y}} \left|\left(p_{(X,Y)}(x,y) - p_X(x)p_Y(y)\right)\frac{f(x)g(y)}{\|f\|_\infty \|g\|_\infty}\right| dx\, dy\right)^2 \\
&\geq \frac{\mathrm{Cov}(f,g)^2}{2\|f\|_\infty^2 \|g\|_\infty^2},
\end{aligned}$$

where δ_{TV} denotes the total variation distance, and the first inequality is due to Pinsker's inequality. □

A.2 Proof of Proposition 4

Proof.

$$\begin{aligned}
\mathcal{V}^2(X,Y) &= \int_{\mathcal{X}}\int_{\mathcal{Y}} |p_{(X,Y)}(x,y) - p_X(x)p_Y(y)|^2\, dx\, dy \\
&\qquad \cdot \frac{\int_{\mathcal{X}}\int_{\mathcal{Y}} |f(x)g(y)|^2\, dx\, dy}{\|fg\|_2^2} \\
&\geq \frac{1}{\|fg\|_2^2}\left(\int_{\mathcal{X}}\int_{\mathcal{Y}} \left|\left(p_{(X,Y)}(x,y) - p_X(x)p_Y(y)\right) f(x)g(y)\right| dx\, dy\right)^2 \\
&= \frac{\mathrm{Cov}(f(X), g(Y))^2}{\|fg\|_2^2},
\end{aligned}$$

where the inequality is due to Cauchy-Schwarz inequality, and $\|fg\|_2$ is defined as

$$\|fg\|_2 = \left(\int_{\mathcal{X}}\int_{\mathcal{Y}} |f(x)g(y)|^2\, dx\, dy\right)^{\frac{1}{2}}.$$

We can further use Hölder's inequality

$$\|fg\|_2^2 = \|f^2 g^2\|_1 \leq \|f^2\|_2 \|g^2\|_2,$$

and have that

$$\mathcal{V}^2(X,Y) \geq \frac{\mathrm{Cov}(f(X), g(Y))^2}{\|f^2\|_2 \|g^2\|_2},$$

which completes the proof. □

A.3 Proof of Theorem 2

Proof. Let $f(t) = t\log(t)$, using Taylor's theorem at $t = 1$ and $f(1) = 0$:

$$f(t) = f'(1)(t-1) + \frac{1}{2}f''(1)(t-1)^2 + o\left((t-1)^2\right).$$

Thus we have

$$\begin{aligned}
f\left(\frac{p_{(X,Y)}(x,y)}{p_X(x)p_Y(y)}\right) &= f'(1)\left(\frac{\epsilon(\phi(x,y) - \varphi(x,y))}{p_X(x)p_Y(y)}\right) \\
&\quad + \frac{f''(1)}{2}\left(\frac{\epsilon(\phi(x,y) - \varphi(x,y))}{p_X(x)p_Y(y)}\right)^2 + o(\epsilon^2),
\end{aligned}$$

where $\lim_{\epsilon\to 0^+} o(\epsilon^2)/\epsilon^2 = 0$. Now, we can obtain

$$\begin{aligned}
&D_{KL}(p_{(X,Y)}\|p_X\otimes p_Y)\\
&= \int_{\mathcal{X}}\int_{\mathcal{Y}} p_X(x)p_Y(y) f\left(\frac{p_{(X,Y)}}{p_X(x)p_Y(y)}\right) dx\,dy\\
&= \epsilon f'(1)\int_{\mathcal{X}}\int_{\mathcal{Y}} (\phi(x,y)-\varphi(x,y))\ dx\,dy\\
&\quad + \epsilon^2\frac{f''(1)}{2}\int_{\mathcal{X}}\int_{\mathcal{Y}}\left(\frac{(\phi(x,y)-\varphi(x,y))}{r_{(X,Y)}(x,y)+\epsilon\varphi_{(X,Y)}(x,y)}\right)^2 dx\,dy + o(\epsilon^2).
\end{aligned}$$

Since $\phi_{(X,Y)}, \varphi_{(X,Y)}$ are valid, we have

$$\begin{aligned}
&D_{KL}(p_{(X,Y)}\|p_X\otimes p_Y)\\
&= \epsilon^2\frac{f''(1)}{2}\int_{\mathcal{X}}\int_{\mathcal{Y}}\left(\frac{(\phi(x,y)-\varphi(x,y))}{r_{(X,Y)}(x,y)+\epsilon\varphi_{(X,Y)}(x,y)}\right)^2 dx\,dy + o(\epsilon^2)\\
&= \int_{\mathcal{X}}\int_{\mathcal{Y}}\frac{1}{r_{(X,Y)}(x,y)}\left(p_{(X,Y)}(x,y)-p_X(x)p_Y(y)\right)^2 dx\,dy,
\end{aligned}$$

which completes the proof. □

A.4 Proof of Proposition 5

Proof. Let $\kappa(x,y) = \exp(-\gamma\|x-y\|^2)$. Define the Hilbert space $\mathcal{H}$ as the reproducing kernel Hilbert space corresponding to κ: $\kappa(\mathbf{x},\mathbf{y}) = \langle\varphi(\mathbf{x}),\varphi(\mathbf{y})\rangle_{\mathcal{H}}$, and the mean map kernel of given distributions P and Q is

$$K(P,Q) = \mathbb{E}_{\mathbf{x}\sim P,\mathbf{y}\sim Q}\kappa(\mathbf{x},\mathbf{y}) = \langle\mathbb{E}_{\mathbf{x}\sim P}[\varphi(\mathbf{x})],\mathbb{E}_{\mathbf{y}\sim Q}[\varphi(\mathbf{y})]\rangle.$$

MMD between $\bar{q}_\phi^k$ and $\bar{q}_\phi^{\neg k}$ can then be written as

$$\begin{aligned}
\mathrm{MMD}(\bar{q}_\phi^k,\bar{q}_\phi^{\neg k})^2 &= \|\mathbb{E}_{\mathbf{z}\sim\bar{q}_\phi^k}[\varphi(\mathbf{z})] - \mathbb{E}_{\mathbf{z}\sim\bar{q}_\phi^{\neg k}}[\varphi(\mathbf{z})]\|^2\\
&= K(\bar{q}_\phi^k,\bar{q}_\phi^k) + K(\bar{q}_\phi^{\neg k},\bar{q}_\phi^{\neg k}) - 2K(\bar{q}_\phi^k,\bar{q}_\phi^{\neg k}),
\end{aligned}$$

where $\bar{q}_\phi^k$ and $\bar{q}_\phi^{\neg k}$ are two Gaussian mixture models defined in Eq. (13). Further, we have

$$K(\bar{q}_\phi^k,\bar{q}_\phi^{\neg k}) = \frac{1}{N_k N_{\neg k}}\sum_{\substack{i=1,\\ \mathbf{s}_i=k}}^{N}\sum_{\substack{j=1,\\ \mathbf{s}_j\neq k}}^{N}\kappa(q_\phi(\mathbf{z}|\mathbf{x}_i,\mathbf{s}_i), q_\phi(\mathbf{z}|\mathbf{x}_j,\mathbf{s}_j))$$

and similar results can be obtained for $K(\bar{q}_\phi^k,\bar{q}_\phi^k)$ and $K(\bar{q}_\phi^{\neg k},\bar{q}_\phi^{\neg k})$ and thus omitted for brevity, which completes the proof. □