Fair Representation Learning with Unreliable Labels

Yixuan Zhang[†], **Feng Zhou**[‡*], **Zhidong Li**[†], **Yang Wang**[†], **Fang Chen**[†]
[†]Data Science Institute, University of Technology Sydney, Australia
[‡]Center for Applied Statistics and School of Statistics, Renmin University of China, China
{yixuan.zhang, zhidong.li, yang.wang, fang.chen}@uts.edu.au, feng.zhou@ruc.edu.cn

Abstract

In learning with fairness, for every instance, its label can be systematically flipped to another class due to the practitioner's prejudice, namely, label bias. The existing well-studied fair representation learning methods focus on removing the dependency between the sensitive factors and the input data, but do not address how the representations retain useful information when the labels are unreliable. In fact, we find that the learned representations become random or degenerated when the instance is contaminated by label bias. To alleviate this issue, we investigate the problem of learning fair representations that are independent of the sensitive factors while retaining the task-relevant information given only access to unreliable labels. Our model disentangles the dependency between fair representations and sensitive factors in the latent space. To remove the reliance between the labels and sensitive factors, we incorporate an additional penalty based on mutual information. The learned purged fair representations can then be used in any downstream processing. We demonstrate the superiority of our method over previous works through multiple experiments on both synthetic and real-world datasets.

1 Introduction

The recent success of deploying machine learning algorithms in different high-stake application areas has increased the concerns for ethics. Due to human prejudice intervening in the labeling process, the training data collected always contains discrimination towards certain demographic groups (Lin et al., 2020; Bertrand and Mullainathan, 2004; Michelle, 2012). When decisions are made algorithmically with such unreliable labels, it affects both accuracy and fairness negatively, and further brings harm to both society and individuals (Khandani et al., 2010; Kim et al., 2015; Brennan et al., 2009). Therefore, as one critical ethical aspect, fairness-aware learning has recently experienced a surge of advances.

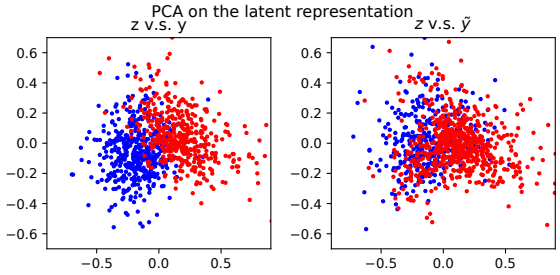


Figure 1: Principal component analysis (PCA) on the learned representations. We performed a PCA analysis between the learned representation and label for a binary classification problem on the synthetic dataset. Red points represent the positive class, while blue points represent the negative class. We compare the ideal labels (left) with unreliable labels (right): the learned representation has a strong correlation with the ideal label (obviously divided into two clusters) but a weak correlation with the unreliable label (two clusters mixed together).

Existing works on fairness extensively studied discrimination removal strategies in different training stages, i.e., pre-processing (Louizos et al., 2015a; Zemel et al., 2013; Calmon et al., 2017; Lum and Johndrow, 2016), in-processing (Bilal Zafar et al., 2015, 2016; Calders et al., 2009; Agarwal et al., 2018; Kamishima et al., 2012) and post-processing (Hardt et al., 2016). Among all these methods, fair representation learning (Louizos et al., 2015b; Creager et al., 2019; Zemel et al., 2013; Calmon et al., 2017) as a pre-processing method has gained significant attention because it is compatible with any learning algo-

Proceedings of the 26[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2023, Valencia, Spain. PMLR: Volume 206.
*Corresponding author.

rithms for downstream tasks.

In fair representation learning, one tries to learn latent representations that are informative for a particular task while removing all sensitive factors (e.g., gender or race) contained in the input data. Since the learned representations are required to preserve useful task information, an essential measure criterion, such as predictiveness, is usually applied, and this process heavily relies on the quality of labels. In the biased dataset, the label can be systematically flipped to another class due to the practitioner's prejudice, which causes the predictiveness to be improperly measured. Existing methods do not address how the representations remain useful when the given labels are unreliable. In fact, when the instance is contaminated by label bias, the learned representations by existing methods become random or degenerated (See Fig. 1 for illustration). To this end, we explore the following question in this paper: *Can we learn the representations predictable to the ideal labels (generated from fair decisions and not impacted by sensitive information) when we only have access to unreliable labels?*

To overcome the above issue, we introduce a novel model based on deep variational autoencoders (VAE) and mutual information techniques. Different from previous works, which do not care about the learning of sensitive information, our work utilizes both informative latent dimensions (fair representations) and latent nuisance dimensions (retaining sensitive information) to help us learn with unreliable labels. We require the learned fair representations to retain as much information about the data as possible while minimally informative about the sensitive factors. Furthermore, we require the learned representations to be representative and predictable of the latent labels that are not affected by the sensitive information, which we call **ideal labels**. Conclusively, we consider incorporating the information bottleneck principle (Tishby et al., 2000) into the variational autoencoders to lead the learning process toward ideal labels from two aspects. The first is that we apply the mutual information to disentangle the learned representations into sensitive and fair parts. Second, we utilize both sensitive and fair representations to recognize which instance is likely to be discriminated by the label bias and obtain the ideal labels by optimizing the difference between the mutual information. Intuitively, if the sensitive representation can predict an instance's label well, we regard this instance with a higher chance of being discriminated against and vice versa.

The major contributions of our work are summarized as follows: **(1)** As far as we know, for the problem with unreliable labels, we propose the first fair representation learning method attempting to recover the ideal labels. **(2)** We present a flexible end-to-end framework that is applicable to both group-dependent and individual label bias. **(3)** We empirically demonstrate that biased labels are adverse to both accuracy and fairness, even when the learned representations remove the bias encoded in input attributes. With unreliable labels, our experimental results on both synthetic and real data demonstrate that our framework effectively learns fair representations towards ideal labels.

2 Related Work

We formulate fair representation learning using the variational information bottleneck (VIB) principle in variational autoencoders. The information bottleneck (IB) technique is introduced by Tishby et al. (2000), and the VIB (Alemi et al., 2019) parametrizes IB via a variational lower bound. It has been used in a wide range of domains such as regularization (Alemi et al., 2019), understanding of β-VAE (Burgess et al., 2018) and compression for deep neural networks (Dai et al., 2018). Several recent works incorporate maximization of the mutual information in the variational autoencoders. For example, Dieng et al. (2019) uses skip connections to force dependency between the latent representations and observations implicitly. Kim and Mnih (2018) proposes maximizing the mutual information between learned latent representations and input data. However, the computation of mutual information is expensive. Then Zhao et al. (2019) proposed to minimize the Maximum Mean Discrepancy between the marginal of the posterior and the prior to increase the maximization of the mutual information contained in the model.

Given the general fair representation learning (FRL) framework, we can relate it to other similar approaches. Dwork et al. (2011) proposed an initial FRL framework. However, this method cannot be formulated as a generalization task due to the limitation of only working with given data. Zemel et al. (2013) then proposed an improved framework but still has limitations on representation since it uses clustering for probabilistic representation mapping. Besides, the information of sensitive attributes still risks leakage in this method. Based on Zemel et al. (2013), Louizos et al. (2015b) proposed a method to tackle these issues by using the framework of VAE (Kingma and Welling, 2014) and they injected label information to control the information leakage with observed labels. Nevertheless, all the above methods focused on the binary sensitive attribute. To further disentangle the learned representation from sensitive information, Creager et al. (2019) proposed a factorized structure in the aggregate latent labels by using the disentanglement VAE (Burgess et al., 2018; Kim and Mnih, 2019; Chen et al., 2019). Different from the work by Creager et al. (2019), for multiple sensitive attributes, we do not additionally require the corresponding dimension of s can represent each dimension of a since we do not need to modify the biased representations in the test time. We only focus on disentangling the sensitive information from z and keep z as fair as possible. Based on that, we utilize the mutual information strategy to help impute the biased

labels and learn fair representation directly from the model.

Regarding the related works of noisy label learning in fairness, most recent works use the loss correction method to derive a new fairness-aware objective under the corrupted distribution with different fairness notions. The group peer loss (Wang et al., 2021) applied the group-dependent label noise and derived fairness constraints on corrupted data with the biased label. Lamy et al. (2019) proposed a noise-tolerant fair classification method, which assumes that the sensitive attributes may contain the noises and are unreliable. Unlike the above works, our method does not require predefined fairness constraints.

3 Method

We wish to learn an invariant fair representation with unreliable labels. In learning with fairness, we observe two distinct variables: x and a, which denote non-sensitive attributes and sensitive factors, respectively. Our goal is to remove the undesired information that is related to a in the latent space. In this work, apart from learning the fair representation z, we try to learn another latent representation s, preserving all sensitive information from x that can infer a. To simplify the notations throughout the paper, we do not specify the format of sensitive attribute a, which can be either binary or multi-attribute (e.g., race v.s. race$\wedge$gender).

3.1 Fair Representation Learning with VAE

Traditionally, learning fair representations can be easily formulated as a general probabilistic model, where we can express the posterior as $p(z, s \mid x)$. Therefore, the aim to find an invariant representation z and a variant representation s can be cast as performing inference on the generative model. Following the same lines in the literature on variational inference, we assume the amortized inference distribution (encoder) is $q_\phi(z, s \mid x)$, which is used to map the observations x into the latent space. In the meanwhile, we factorize the decoder as $p_\theta(x \mid z, s)p_\theta(a \mid s)$, which enables us to return the latent representation to the input data space. We require z, s can reconstruct x back and s can reconstruct a. In this way, we can learn both useful information in the z dimension and s dimension. Then, we can easily obtain the variational lower bound of the log-likelihood $\log p(x_i, a_i)$ as follows:

$$\begin{aligned}&\sum_{i=1}^{N} \log p(x_i, a_i)\\ &\geq \sum_{i=1}^{N} \mathbb{E}_{q_\phi(z_i, s_i \mid x_i)}[\log p_\theta(x_i \mid z_i, s_i) + \log p_\theta(a_i \mid s_i)]\\ &- \mathrm{KL}[q_\phi(z_i, s_i \mid x_i) \,||\, p_\theta(z_i, s_i))] = \mathcal{F}(\phi, \theta; x_i, a_i),\end{aligned} \tag{1}$$

where $q_\phi(z_i, s_i \mid x_i) = \mathcal{N}(z_i, s_i \mid \mu_i = f_\phi(x_i), \sigma_i = e^{f_\phi(x_i)})$, $p_\theta(x_i \mid z_i, s_i) = f_\theta(z_i, s_i)$, $p_\theta(a_i \mid s_i) = f_\theta(s_i)$ and the expectation $\mathbb{E}_{q_\phi(z,s|x)}$ is empirically approximated via Monte Carlo by reparameterization trick. It is worth noting that we exclude a in the variational distribution as we require all sensitive information to be learned from x. In this way, we can extract the correlated sensitive information of a remaining in x.

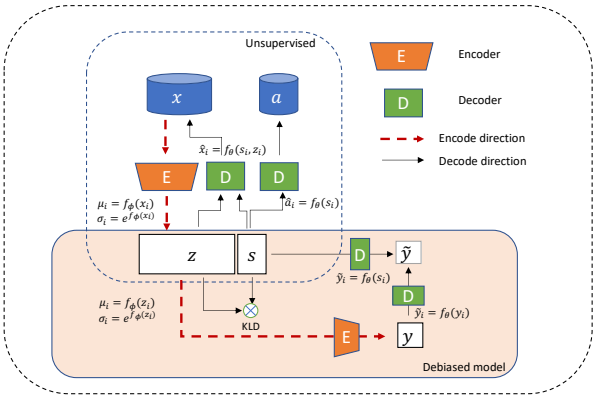


Figure 2: Framework for fair representation learning. The upper dashed rectangle is the process for the unsupervised model (i.e., existing fair representation learning framework), where we have two distinct sources: a and x. The bottom rectangle colored in orange is the learning process with unreliable labels. To characterize the bias, we assume the observed biased labels $\tilde{y}$ are conditioned on sensitive information s and ideal labels y.

Naturally, $q_\phi(z, s \mid x)$ is called the encoder, while $p_\theta(x \mid z, s)$ and $p_\theta(a \mid s)$ refers to the decoders. The model parameter (θ) and variational parameter (ϕ) are jointly optimized with the stochastic gradient variational Bayes algorithm (Kingma and Welling, 2014) according to the lower bound of the log-likelihood. In addition, this ELBO imposes a regularizer over z and s, which is the KL term in Eq. (1). It encourages the posterior $q_\phi(z, s \mid x)$ to be matched to the prior $p(z, s)$. Such a method minimizes the upper bound of the mutual information between representations z, s and input data x, i.e., $I(z, s; x)$. As a consequence, we can capture most of the salient information of x in the embedding z and s. In the optimization, the first and second terms in Eq. (1) are interpreted as reconstruction errors, while the KL term is interpreted as a regularization term.

Since we wish to remove all the sensitive information from the fair representation z, we further disentangle z_i and s_i by encouraging $I(z_i, s_i)$ to be as small as possible with the prior $q(z_i, s_i) = q(z_i)q(s_i)$. By further requiring the explicit minimization of the mutual information between the

two latent dimensions (informative and sensitive information), we obtain the objective of the following form:

$$\mathcal{L}_u(\phi, \theta; x_i, a_i) = \mathcal{F}(\phi, \theta; x_i, a_i) - \sum_{i=1}^{N} \alpha \mathrm{KL}[q(z_i, s_i) \mid\mid q(z_i)q(s_i)], \tag{2}$$

where $\alpha \geq 0$ is a regularization coefficient for the mutual information, which decides the degree of disentanglement between z and s. To compute $\mathrm{KL}[q_\phi(z, s)||q(z)q(s)]$, we use the similar approach in FactorVAE (Kim and Mnih, 2018). This term is estimated by using a discriminator that distinguishes 'real samples' from 'fake samples', and then applying the density-ratio trick for approximation (see Appendix B for more details).

3.2 Information Bottleneck View of Learning with Unreliable Labels

As mentioned in Section 1, despite the fact that we have a model encouraging independence between s and z in the latent space, we still have information about a leaked into the posterior when $\tilde{y}$ is correlated with a, which causes the degeneration of the learned representation. In our method, with the mild assumption, we model $\tilde{y}$ is systematically biased against a. This pattern can be found in many practical examples. For instance, in the recidivism prediction tasks, when participants are provided with biasing information, the predicted labels will be systematically biased against certain racial groups (Lin et al., 2020); In recruitment, the decisions are usually in favor of particular demographic groups. (Bertrand and Mullainathan, 2004). To avoid such undesirable sensitive information leakage, we instead try to maximize a penalized lower bound. In the following, we describe one way to achieve such kind of regularization by combining Eq. (2) with mutual information. A sketch of the framework is shown in Fig. 2.

We introduce y to represent the ideal labels and $\tilde{y}$ to represent the observed unreliable labels. We assume $\tilde{y}$ is conditioned on both the sensitive information s and the ideal labels y. To avoid the fair representation z being contaminated again by the sensitive information leaked in $\tilde{y}$, we inject both the label information y and biased information s. We formulate this by using information theoretic concepts. The idea is to learn s and a latent label encoding of z (i.e., y), which are highly informative for predicting $\tilde{y}$. To avoid the latent label y just simply memorizing z, we add a complexity penalty to learn a more compressive encoding of z, i.e., relevant information from z for predicting $\tilde{y}$. The penalty term is the mutual information between z and z (i.e., $I(y, z)$) we want to minimize, and this term tries to introduce some independence between z and y so that we can further remove undesired nuisance information when encoding the latent ideal label y from z.

Therefore, the IB objective is $I(y, \tilde{y}) + \beta I(s, \tilde{y}) - \gamma I(y, z)$, where $\beta \geq 0$ and $\gamma \geq 0$ are two hyperparameters controlling the regularization impacts. For $I(s, \tilde{y})$, we use the decoder parameter θ, so we can easily obtain the lower bound by dropping the constant of $H(\tilde{y})$ (it does not depend on ϕ and θ). For the IB objective $I(y, \tilde{y}) - \gamma I(y, z)$, we will construct a lower bound as follows.

According to the definition of the mutual information (Shannon, 1948), both $I(y, \tilde{y})$ and $I(y, z)$ can be expanded as:

$$I(y, \tilde{y}) = \mathbb{E}_{q_\phi(y,\tilde{y})}[\log \frac{q_\phi(\tilde{y} \mid y)}{p(\tilde{y})}],$$
$$I(y, z) = \mathbb{E}_{q_\phi(y,z)}[\log \frac{q_\phi(y \mid z)}{q_\phi(y)}]. \tag{3}$$

Since in $I(y, \tilde{y})$, the KL term involves the computation of the data distribution $p(\tilde{y})$ which causes the IB is intractable. To overcome this issues, we then introduce a decoder $p_\theta(\tilde{y} \mid y)$ to approximate $q_\phi(y, \tilde{y})$. Similarly, $I(y, z)$ requires the computation of $p(z)$, so we instead use $p_\theta(y)$ to approximate $q_\phi(y)$. Following the standard variational bounds derivation on the mutual information discussed by Barber and Agakov (2003), we obtain the lower bound of $I(y, \tilde{y})$ and the upper bound of $I(y, z)$:

$$I(y, \tilde{y}) \geq \mathbb{E}_{q_\phi(y,\tilde{y})}[\log p_\theta(\tilde{y} \mid y)] + H(\tilde{y}),$$
$$I(y, z) \leq \mathbb{E}_{q_\phi(z,y)}[\log \frac{q_\phi(y \mid z)}{p_\theta(y)}]. \tag{4}$$

Then, by dropping the constant entropy $H(\tilde{y})$ and assuming the joint distribution factorizes as $q_\phi(y, z, \tilde{y}) = q_\phi(y \mid z)p(z, \tilde{y})$, we construct the following lower bound to provide a tractable objective (see Appendix A for the detailed derivation):

$$\mathcal{F}(\theta, \phi; y_i, z_i, s_i) = \sum_{i=1}^{N} \mathbb{E}_{q_\phi(y_i|z_i)}[\log p_\theta(\tilde{y}_i \mid y_i) - \gamma \mathrm{KL}(q_\phi(y_i \mid z_i) \mid\mid p_\theta(y_i))] + \beta \sum_{i=1}^{N} \mathbb{E}_{p_\theta(\tilde{y}|s)} \log p_\theta(\tilde{y} \mid s), \tag{5}$$

where $q_\phi(y_i \mid z_i) = p_{\mathrm{Cat}}(y_i \mid \pi_i = \mathrm{softmax}(f_\phi(z_i))$, $p_\theta(\tilde{y}_i \mid y_i) = p_{\mathrm{Cat}}(\tilde{y}_i \mid \pi_i = \mathrm{softmax}(f_\theta(y_i))$ and $p_\theta(\tilde{y}_i \mid s_i) = p_{\mathrm{Cat}}(\tilde{y}_i \mid \pi_i = \mathrm{softmax}(f_\theta(s_i))$, p_{Cat} indicates the categorical distribution.

Combing Eqs. (2) and (5), the final objective function is:

$$\mathcal{F}_{Frep}(\phi, \theta; x_i, a_i, s_i, y_i, \tilde{y}_i) = \mathcal{L}_u(\theta, \phi; x_i, a_i) + \mathcal{F}(\theta, \phi; y_i, z_i, s_i). \tag{6}$$

Instead of training each layer of stochastic variables separately, we optimize the model jointly. Through this method, we utilize both unreliable label $\tilde{y}$ and the sensitive information s to guide the learning of y and obtain a better feature

extraction that remains the most useful information. Besides, optimizing the model jointly enforces the disentanglement between the unreliable label $\tilde{y}$ and the sensitive attribute a, which avoids creating the degenerated representation with respect to the ideal label y.

4 Experiment

In the following sections, we introduce our experimental setting including datasets, baselines and evaluation metrics. The implementation code is publicly available at https://github.com/co234/frl-with-unreliable-label.

4.1 Datasets and Setup

We conduct experiments on one synthetic dataset and two real-world datasets, Adult and Compas. The statistics of all datasets are shown in Table 1. We list the number of instances, the specified protected and privileged groups, as well as their corresponding number of instances. The detailed description for each dataset can be found in Appendix D. We use accuracy to measure the performance and $\Delta_{\text{DP}} = |\mathbb{E}(\hat{y} = 1 \mid a = 1) - \mathbb{E}(\hat{y} = 1 \mid a = 0)|$ to measure the fairness violation. It is worth noting that other statistical fairness notions can also be applied, and we list the results under different fairness notions in Appendix E.2. A lower Δ_{DP} indicates a minor fairness violation. We split the data into 90% train and 10% test and report the results of ten-fold experiments with random splits.

Table 1: The statistics of the synthetic, Adult and Compas datasets. We list the number of instances for the whole dataset, the number of instances in the protected and privileged groups and the fairness violation.

Dataset	# of Instances	Protected/Privileged Groups	# of Instances	Δ^*_{DP}
Synthetic	10,800	a=1/a=0	5150/5650	0.02
Adult	30,717	female/male	10,067/20,650	0.20
		female & black / rest	1943/28,774	0.19
Compas	5,554	black/white	2,874/2,680	0.15
		black & male / rest	2,848/2,706	0.14

We conduct experiments under different types of label bias: group-dependent and instance-dependent. In the group-dependent label bias setting, following Wick et al. (2019), we artificially flip the labels based on the true labels and demographic groups. For the instance-dependent setting, we flip instances inversely proportional to their distance to the decision boundary (i.e., the samples closer to the decision boundary have a higher chance of being labeled wrongly.)

4.2 Baseline Methods and Model Architectures

To evaluate the effectiveness and robustness of the proposed method, we compare our method with several VAE-based models, including the vanilla VAE (Kingma and Welling, 2014), Flexibly Fair Variational Autoendoer (FFVAE) (Creager et al., 2019) and Variational Fair Autoencoder (VFAE) (Louizos et al., 2015b). Regarding the label noisy learning methods, we compare two related noisy label learning methods: CORES^2 (Cheng et al., 2021) and Group Peer Loss (GPL) (Wang et al., 2021). For a fair comparison, we fix the autoencoder structure among the above models. The details of implementation can be found in Appendix B.

4.3 Results

Case 1: Binary sensitive attribute with group-dependent label bias. The results are shown in Table 2. The prediction performance of our method generally outperforms other baselines w.r.t. both effectiveness and robustness when we increase the ratio of label bias. Overall, when the bias ratio increases to above 20%, the accuracy of other VAE-related methods starts to drop dramatically, and the fairness violation starts to increase, which demonstrates that no matter how fair the learned z is, if we let z be predictable to the biased label $\tilde{y}$, we still obtain the biased output. This conclusion can also be obtained by comparing our method using y directly (ours+D) to our method using z (ours+LR). In the meantime, two noisy label learning methods have more steady accuracy when we increase the amount of label bias. However, since CORES^2 does not take fairness into consideration, it has an overall higher fairness violation compared to GPL. GPL deploys derived fairness constraints under corrupted distribution, so it has overall lower fairness violation compared to CORES^2. Though the performance and fairness violation of ours+LR is generally better than other baselines, it is still worse than ours+D. We obtain the same trend in all three datasets. For the synthetic dataset, we find ours+D has the overall highest accuracy with Δ_{DP} close to the Δ^*_{DP} under clean distribution. Also, we find the results of ours+LR are very close to ours+D while ours+LR has a slightly higher fairness violation and lower accuracy. Other baselines work well when the bias amount is small, but when we increase the bias impact, they are not robust to the change of bias amount. For the Adult dataset, the accuracy of ours+D is the highest among all other baselines, and at the same time, ours+D has the closest Δ_{DP} to the Δ^*_{DP}. When the label bias amount increases, the accuracy of ours+D decreases slightly, but the performance of ours+D is the most steady one compared to other baselines. For the Compas dataset, ours+D has the highest accuracy. Though GPL has the lowest Δ_{DP}, if we combine it with the accuracy, we can find the reason that Δ_{DP} is low is the weak predictiveness provided by GPL. When the corrup-

Table 2: Test accuracy (%) and fairness violation on the synthetic, Adult and Compas dataset of binary sensitive features under various corruption ratios from 10% to 40% with group-dependent label bias. We use 'ours+LR' and 'ours+D' to denote the measurement of our method using a downstream classifier on the learned representation z and the learned latent label y, respectively. We report the results in the format of mean $\pm$ standard deviation.

Method	r-10%		r-20%		r-30%		r- 40%	
	ACC	Δ_{DP}	ACC	Δ_{DP}	ACC	Δ_{DP}	ACC	Δ_{DP}
Dataset: Synthetic								
FFVAE	85.9±1.1	0.08±0.01	84.1±1.2	0.13±0.03	81.9±1.0	0.17±0.02	80.0±2.0	0.23±0.05
VFAE	81.8±2.0	0.15±0.03	82.2±1.8	0.18±0.04	81.0±2.3	0.22±0.03	79.1±1.9	0.25±0.09
VAE	85.8±1.0	0.03±0.01	85.2±0.9	0.07±0.02	82.8±1.7	0.16±0.05	79.7±1.2	0.23±0.04
CORES^2	85.5±2.3	0.01±0.01	85.6±1.1	0.07±0.01	85.5±2.1	0.06±0.00	85.3±1.1	0.08±0.01
GPL	85.5±1.1	**0.01±0.00**	85.2±1.4	0.06±0.02	84.5±1.3	0.07±0.01	83.4±1.5	0.04±0.01
Ours+LR	85.5±1.0	0.03±0.01	85.4±0.9	0.06±0.02	85.8±1.3	0.05±0.03	84.6±0.8	0.05±0.03
Ours+D	**86.3±0.9**	0.02±0.01	**86.3±1.3**	**0.04±0.00**	**86.0±1.6**	**0.04±0.01**	**85.7±1.2**	**0.03±0.01**
Dataset: Adult, a = gender								
FFVAE	82.7±2.4	0.18±0.03	82.2±1.5	0.26±0.04	79.9±1.4	0.21±0.05	76.3±1.3	0.37±0.09
VFAE	82.5±1.8	0.19±0.04	81.6±1.2	0.39±0.09	78.3±1.8	0.22±0.07	77.7±2.9	0.49±0.12
VAE	81.7±1.2	0.20±0.05	82.5±1.1	0.24±0.07	79.7±1.5	0.34±0.10	75.4±2.2	0.42±0.07
CORES^2	78.2±1.7	0.18±0.01	78.1±3.8	**0.18±0.01**	79.8±2.7	**0.17±0.01**	79.4±1.4	0.20±0.02
GPL	80.7±1.6	**0.17±0.02**	76.5±4.0	0.19±0.01	75.8±0.9	0.20±0.00	74.3±3.3	0.20±0.01
Ours+LR	82.5±1.1	0.18±0.01	**83.2±0.9**	0.20±0.02	80.1±1.9	0.19±0.03	74.8±2.8	0.33±0.06
Ours+D	**82.7±1.6**	**0.17±0.02**	83.1±0.8	0.19±0.02	**81.1±1.8**	**0.17±0.01**	**80.9±1.2**	**0.19±0.02**
Dataset: Compas, a = Race								
FFVAE	67.5±2.1	0.22±0.03	68.1±1.9	0.24±0.05	67.2±3.1	0.21±0.06	66.3±2.4	0.22±0.04
VFAE	65.9±1.8	0.18±0.05	65.4±3.7	0.24±0.06	65.2±3.5	0.19±0.05	63.6±2.2	0.24±0.04
VAE	66.8±3.4	0.25±0.04	66.9±1.4	0.27±0.02	65.4±2.1	0.26±0.04	65.7±3.7	0.23±0.04
CORES^2	66.0±2.3	0.17±0.03	66.7±2.9	0.18±0.04	66.9±2.5	0.17±0.01	66.3±3.8	0.18±0.09
GPL	63.6±2.9	**0.15±0.01**	61.2±4.3	**0.14±0.02**	64.1±3.3	**0.15±0.02**	56.8±3.2	**0.09±0.05**
Ours+LR	67.0±1.1	0.24±0.16	67.2±3.7	0.22±0.04	68.0±1.3	0.20±0.03	65.9±1.7	0.23±0.02
Ours+D	**68.5±1.2**	0.24±0.03	**68.6±1.3**	0.24±0.02	**69.1±2.6**	0.21±0.04	**66.5±1.6**	0.22±0.03

tion ratio is 40%, we can see GPL is the only method with an accuracy below 60%.

Case 2: Multiple sensitive attributes with group-dependent label bias. In this section, we conduct experiments on Adult and Compas datasets under the setting that we have 2-dimensional sensitive attributes, which are specified in Table 1. The results are shown in Table 3. For the Adult dataset, the task becomes more difficult than in the binary-sensitive attribute setting since the number of instances among the protected and privileged groups is more imbalanced than in the binary-sensitive attribute setting. We can clearly see that the accuracy and fairness of the Adult dataset for all baselines perform worse than the binary setting. Even so, ours+D still has the highest accuracy compared to the other baselines. Also, Δ_{DP} of ours+D is the lowest and closest to Δ^*_{DP}. Compared to the VAE-related methods, the two noisy learning methods perform better regarding both accuracy and fairness violation, similar to the binary-sensitive attribute setting. We notice that the accuracy of ours+LR is close to ours+D but with higher fairness violations. For the Compas dataset, the number of instances for the protected and privileged groups is balanced as in the binary sensitive attribute case. In this case, ours+D has the highest accuracy, and the performance of ours+LR is very close to ours+D. It is worth noting that, for the Compas dataset, similar to the binary-sensitive attribute setting, though GPL still has the lowest Δ_{DP}, it has the lowest accuracy at the same time, which means the Δ_{DP} is low because the predictions are not good. We do not find any obvious superiority for fairness baselines over the vanilla VAE when the corruption ratio is less than 20%.

Case 3: Instance-dependent label bias. To mimic a compelling and realistic scenario, in this section, we conduct experiments on the Adult dataset to demonstrate the results under the setting that we flip the labels randomly with the probability proportional to the reciprocal of the distance to the decision boundary under different corruption levels from 10% to 40%. We list the results of different methods in Table 4. In Table 4, we can see our method can still handle this complex setting, and the results show that both 'ours+LR' and 'ours+D' are robust to the different levels of corruption ratio from 10% to 40% w.r.t. the steady change in both accuracy and fairness violation. An interesting finding is that the results of all the baselines

Table 3: Test accuracy (%) and fairness violation on the Adult and Compas dataset of multi-sensitive features under various corruption ratios from 10% to 40% with group-dependent label bias. We use 'ours+LR' and 'ours+D' to denote the measurement of our method using a downstream classifier on the learned representation z and the learned latent label y, respectively. We report the results in the format of mean $\pm$ standard deviation.

Method	r-10%		r-20%		r-30%		r- 40%	
	ACC	Δ_{DP}	ACC	Δ_{DP}	ACC	Δ_{DP}	ACC	Δ_{DP}
Dataset: Adult, a = Race, Gender								
FFVAE	80.2±1.3	0.19±0.04	79.4±2.4	0.20±0.05	75.6±1.9	0.17±0.03	66.4±2.5	0.23±0.07
VFAE	83.1±1.2	0.23±0.05	78.1±1.4	0.38±0.10	76.7±2.5	0.39±0.08	70.9±1.1	0.46±0.14
VAE	83.0±0.9	0.17±0.05	81.1±1.2	0.22±0.09	76.1±2.9	0.31±0.13	66.9±3.1	0.34±0.11
CORES^2	80.5±1.6	0.17±0.01	79.5±2.4	**0.17±0.01**	78.4±2.6	**0.17±0.01**	77.9±1.9	**0.20±0.01**
GPL	78.7±3.3	0.17±0.02	80.4±1.7	**0.17±0.01**	79.2±3.0	0.18±0.01	76.6±2.4	0.21±0.01
Ours+LR	83.5±1.3	0.18±0.04	81.0±2.4	0.22±0.03	78.6±1.2	0.21±0.07	78.5±1.9	0.23±0.11
Ours+D	**83.7±0.8**	**0.15±0.02**	**83.4±1.4**	0.21±0.03	**80.6±2.3**	0.19±0.06	**80.0±2.1**	**0.20±0.01**
Dataset: Compas, a = Race, Gender								
FFVAE	69.6±2.4	0.26±0.07	69.26±2.5	0.25±0.06	70.3±1.4	0.25±0.09	66.7±1.5	0.22±0.07
VFAE	67.0±1.4	0.26±0.05	66.2±0.5	0.23±0.07	65.2±1.8	**0.17±0.03**	62.0±2.8	**0.16±0.02**
VAE	71.2±1.2	0.27±0.05	69.6±2.3	0.25±0.04	69.6±1.3	0.21±0.05	67.6±2.4	0.19±0.02
CORES^2	66.5±0.7	**0.20±0.01**	66.6±0.4	0.20±0.03	66.4±0.4	0.19±0.01	66.6±0.4	0.21±0.01
GPL	67.1±2.1	**0.20±0.01**	65.3±0.8	**0.20±0.01**	66.1±0.7	0.19±0.01	65.8±1.3	0.20±0.01
Ours+LR	70.4±1.1	0.27±0.03	69.4±1.3	0.25±0.05	69.4±2.3	0.21±0.05	68.9±1.4	0.19±0.01
Ours+D	**71.2±1.1**	0.26±0.02	**70.2±2.2**	0.23±0.02	**70.6±1.3**	0.21±0.03	**70.0±1.2**	0.20±0.02

Table 4: Test accuracy (%) and fairness violation on the Adult dataset of binary-sensitive features under various corruption ratios from 10% to 40% with instance-dependent label bias. We use 'ours+LR' and 'ours+D' to denote the measurement of our method using a downstream classifier on the learned representation z and the learned latent label y, respectively. We report the results in the format of mean $\pm$ standard deviation.

Method	r-10%		r-20%		r-30%		r- 40%	
	ACC	Δ_{DP}	ACC	Δ_{DP}	ACC	Δ_{DP}	ACC	Δ_{DP}
Dataset: Adult, a = Gender								
FFVAE	80.7±0.2	0.17±0.05	80.5±0.1	0.18±0.05	80.0±0.2	0.17±0.01	79.9±0.2	0.20±0.05
VFAE	79.9±0.2	**0.13±0.02**	75.2±0.2	0.23±0.06	74.2±1.1	0.22±0.06	75.1±1.3	0.23±0.07
VAE	81.7±1.0	0.16±0.01	80.8±1.0	**0.16±0.01**	80.6±1.1	0.15±0.02	79.6±1.0	0.16±0.01
CORES^2	80.1±1.1	0.18±0.01	80.9±0.4	0.17±0.01	79.6±2.9	0.18±0.01	79.1±2.9	0.18±0.01
GPL	80.2±2.0	0.18±0.01	79.3±3.9	0.18±0.02	77.2±3.2	0.18±0.01	76.4±5.7	0.17±0.02
Ours+LR	82.9±0.2	0.16±0.01	82.7±0.2	0.17±0.01	81.8±0.2	0.16±0.02	80.9±0.3	**0.12±0.03**
Ours+D	**83.5±0.1**	0.15±0.02	**83.1±0.2**	**0.16±0.01**	**82.7±0.2**	**0.14±0.01**	**81.2±0.1**	0.16±0.02

are better than the label-dependent situation, especially for the VAE-related methods. Similar to the group-dependent label bias setting, when checking the performance of two noisy label learning methods, we notice that the accuracy for GPL drops dramatically compared to the other noisy label learning method CORES^2 since GPL only considers group-dependent label bias while CORES^2 can handle with the instance-dependent bias. Regarding the several VAE-based methods, the performance has a similar trend to the group-dependent label bias. However, VFAE has the lowest accuracy and higher fairness violation compared to FFVAE and VAE due to the injection of observed labels into the learning process. This indicates that when the labels are unreliable, introducing label injection into the model might bring more risk to learning fair representations.

4.4 Ablation Studies

We use the Adult dataset to examine the effectiveness of the objective function with group-dependent label bias under a high corruption ratio (40%). The results are shown in Table 5. We compare the accuracy and fairness violation with different combinations of components in Eq. (6). Based on the results, we can see that without the mutual information regularization terms $I_\phi(y, \tilde{y})$, $I_\phi(s, \tilde{y})$ and $I_\phi(y, z)$, we have a higher fairness violation compared to the results from optimizing Eq. (6), which aligns with our expectation. Then, if we remove the regularization terms on y and

z with only $I_\phi(s, \tilde{y})$ left, we can achieve a lower fairness violation but a lower accuracy as well. On the contrary, if we remove the regularization term on s, we can obtain a higher accuracy but also a higher fairness violation.

Table 5: Ablation analysis on the objective function on the Adult dataset with high corruption ratio (40%). We test with the unsupervised model Eq. (2), the full de-biased model Eq. (1), the full de-biased model without $I_\phi(y, \tilde{y}) - \beta I_\phi(y, z)$ and the full de-biased model without $I_\phi(s, \tilde{y})$. We report test accuracy (%) and fairness violation in the format of mean $\pm$ standard deviation.

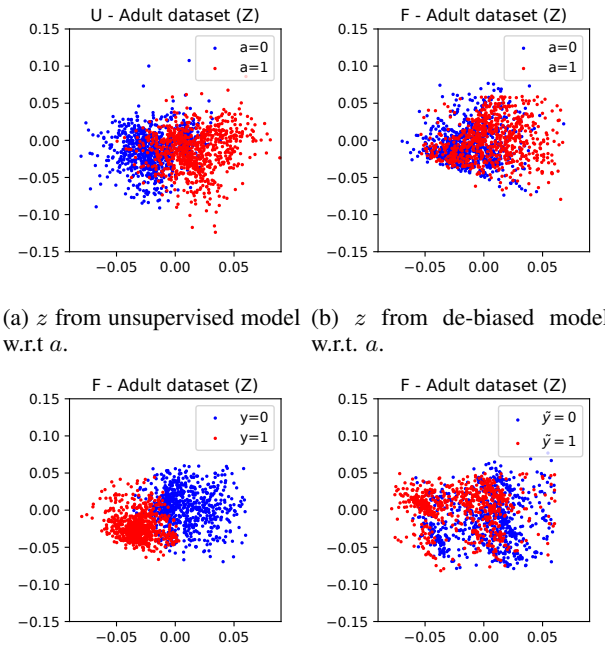
Objective Function	ACC (%)	Δ_{DP}
$\mathcal{F}_{Frep}(\theta, \phi; x_i, a_i, s_i, y_i, \tilde{y}_i)$	83.8±1.1	0.2±0.0
$\mathcal{L}_u(\theta, \phi; x_i, a_i)$	80.5±1.3	0.3±0.0
$\mathcal{L}_u(\theta, \phi; x_i, a_i) + I_\phi(y, \tilde{y}) - \beta I_\phi(y, z)$	79.2±1.5	0.3±0.1
$\mathcal{L}_u(\theta, \phi; x_i, a_i) + \beta I_\phi(s, \tilde{y})$	78.5±0.7	0.1±0.0

4.5 Analysis on Learned Representations

We also conduct the analysis of the learned representations on the Adult dataset, which is shown in Fig. 3. We set the corruption ratio as 40% and test it on both unsupervised and de-biased models. We first plot the learned representation z (after applying Kernel PCA) using the unsupervised model with objective function Eq. (2) w.r.t. two demographic groups. We can see the locations of data points in different demographic groups have a clear pattern (where the protected group locates in the upper right). From the results, we can see that when the observed labels correlate with sensitive factors, z still contains discriminative information without incorporating the latent label information. Then, in the upper right plot, we visualize z (after applying Kernel PCA) by optimizing Eq. (6) w.r.t. two demographic groups. The visualization results show that by properly injecting the label information with maximizing the mutual information in the variational encoders, the learned z can successfully reduce the discrimination (we can see the two clusters now mix together, which indicates the z cannot distinguish the demographic information). Meanwhile, we also visualize z (after Kernel PCA) w.r.t. the latent label y and biased label $\tilde{y}$ in the two bottom plots. The right bottom graph shows that though z has an ambiguous pattern to distinguish $\tilde{y}$, it still contains some randomness towards $\tilde{y}$ which we have discussed in Section 1, but z can still be divided into two clusters. While in the left bottom graph, z can be clearly divided into two clusters w.r.t. the latent label y.

4.6 Hyperparameters

We conduct experiments on different values of α, β and γ. We first test different combinations of the two hyper-



(a) z from unsupervised model w.r.t a.

(b) z from de-biased model w.r.t. a.

(c) z from de-biased model w.r.t. y.

(d) z from de-biased model w.r.t. $\tilde{y}$.

Figure 3: Representation analysis. The upper two plots depict the visualization of z v.s. a (unsupervised model), z v.s. a (optimizing Eq. (6) after applying kernel PCA. We use two colors to denote different demographic groups. The blue color denotes the protected group, and the red color denotes the privileged group. The bottom two plots visualize the learned representation z (optimizing Eq. (6)) w.r.t. the learned latent label y and observed biased label $\tilde{y}$.

Table 6: Test accuracy (%) and fairness violation on the Adult dataset with high corruption ratio (40%) under different value of β from 0.1 to 0.7 with fixed value of $\gamma = 0.5$ and $\alpha = 1$

β	0.1	0.3	0.5	0.7
ACC	78.7±0.1	79.2±0.2	80.8±0.4	80.9±0.1
DP	0.22±0.01	0.19±0.02	0.18±0.01	0.19±0.02

parameters β and γ, which control the impact of $I_\phi(y, z)$ and $I_\phi(s, \tilde{y})$ respectively. We find that the value of γ does not impact the results much, so we fix the value to 0.5 in the experiments. However, we find different β will impact the results differently. To see this, we choose $\beta \in \{0.1, 0.3, 0.5, 0.7\}$, while we fix $\alpha = 1$ and $\gamma = 0.5$. The results are listed in Table 6. We experiment with a large amount of label bias (corruption ratio of 40%), so in such a scenario, if we increase the value of β with a fixed value of γ, the model will add more attention to correct y from the discriminated samples. We also conduct experi-

ments over different values of the hyperparameter α in the compression term. However, we do not observe any patterns that the change of α will greatly affect the accuracy and fairness.

5 Conclusions

In this work, we demonstrate that when the labels are unreliable, we can still learn fair representations and latent ideal labels by introducing maximization of the mutual information in the variational autoencoders. We design our method based on two aspects: (1) learn both fair and sensitive representations in the latent space and disentangle the sensitive information from the fair dimension, and (2) construct the variational information bottleneck lower bound to discourage the latent fair labels from being similar to observed biased ones for unfair samples. We derive a novel tractable objective function for optimizing the variational lower bound. In experiments, we empirically demonstrate the superiority of our method to baselines w.r.t. effectiveness and robustness under different amounts of label bias. In addition, we show that when the observed labels are unreliable, the learned fair representations are still discriminated against the particular demographic group since the prediction in downstream tasks is still measured with unreliable labels. In our method, we rely on the prediction from latent nuisance (sensitive) information to observed biased labels to measure how unfair the sample is. This paper uses VAE as one possible solution, a possible research track in the future is to extend method in a general framework.

Acknowledgements

The authors would like to thank the anonymous reviewers for insightful comments which greatly improved the paper. This work was supported by NSFC (National Natural Science Foundation of China) Project (No. 62106121). This research was supported by Public Computing Cloud, Renmin University of China.

reference

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. M. (2018). A reductions approach to fair classification. *CoRR*, abs/1803.02453.

Alemi, A. A., Fischer, I., Dillon, J. V., and Murphy, K. (2019). Deep variational information bottleneck.

Barber, D. and Agakov, F. (2003). Information maximization in noisy channels : A variational approach. In Thrun, S., Saul, L., and Schölkopf, B., editors, *Advances in Neural Information Processing Systems*, volume 16. MIT Press.

Bertrand, M. and Mullainathan, S. (2004). Are emily and greg more employable than lakisha and jamal? a field experiment on labor market discrimination. *American Economic Review*, 94(4):991–1013.

Bilal Zafar, M., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. (2015). Fairness Constraints: Mechanisms for Fair Classification. *arXiv e-prints*, page arXiv:1507.05259.

Bilal Zafar, M., Valera, I., Gomez Rodriguez, M., and Gummadi, K. P. (2016). Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. *arXiv e-prints*.

Brennan, T., Dieterich, W., and Ehret, B. (2009). Evaluating the predictive validity of the compas risk and needs assessment system. *Criminal Justice and Behavior*, 36(1):21–40.

Burgess, C. P., Higgins, I., Pal, A., Matthey, L., Watters, N., Desjardins, G., and Lerchner, A. (2018). Understanding disentangling in β-vae.

Calders, T., Kamiran, F., and Pechenizkiy, M. (2009). Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18.

Calmon, F., Wei, D., Vinzamuri, B., Natesan Ramamurthy, K., and Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 3992–4001. Curran Associates, Inc.

Chen, R. T. Q., Li, X., Grosse, R., and Duvenaud, D. (2019). Isolating sources of disentanglement in variational autoencoders.

Cheng, H., Zhu, Z., Li, X., Gong, Y., Sun, X., and Liu, Y. (2021). Learning with instance-dependent label noise: A sample sieve approach. In *International Conference on Learning Representations*.

Creager, E., Madras, D., Jacobsen, J., Weis, M. A., Swersky, K., Pitassi, T., and Zemel, R. S. (2019). Flexibly fair representation learning by disentanglement. *CoRR*, abs/1906.02589.

Dai, B., Zhu, C., and Wipf, D. (2018). Compressing neural networks using the variational information bottleneck.

Dieng, A. B., Kim, Y., Rush, A. M., and Blei, D. M. (2019). Avoiding latent variable collapse with generative skip models. In *AISTATS*.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. (2011). Fairness Through Awareness. *arXiv e-prints*.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. *CoRR*, abs/1610.02413.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. (2012). Fairness-aware classifier with prejudice remover

regularizer. In Flach, P. A., De Bie, T., and Cristianini, N., editors, *Machine Learning and Knowledge Discovery in Databases*, pages 35–50, Berlin, Heidelberg. Springer Berlin Heidelberg.

Khandani, A. E., Kim, A. J., and Lo, A. W. (2010). Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767 – 2787.

Kim, H. and Mnih, A. (2018). Disentangling by factorising. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2649–2658. PMLR.

Kim, H. and Mnih, A. (2019). Disentangling by factorising.

Kim, S.-E., Paik, H. Y., Yoon, H., Lee, J., Kim, N., and Sung, M.-K. (2015). Sex- and gender-specific disparities in colorectal cancer risk. *World journal of gastroenterology : WJG*, 21:5167–5175.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes.

Lamy, A., Zhong, Z., Menon, A. K., and Verma, N. (2019). Noise-tolerant fair classification. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32, pages 294–306. Curran Associates, Inc.

Lin, Z., Jung, J., Goel, S., and Skeem, J. (2020). The limits of human predictions of recidivism. *Science Advances*, 6:eaaz0652.

Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R. (2015a). The Variational Fair Autoencoder. *arXiv e-prints*, page arXiv:1511.00830.

Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R. (2015b). The Variational Fair Autoencoder. *arXiv e-prints*.

Lum, K. and Johndrow, J. (2016). A statistical framework for fair predictive algorithms. *arXiv e-prints*, page arXiv:1610.08077.

Michelle, A. (2012). The new jim crow: mass incarceration in the age of colorblindness. New Press, New York.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423.

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

Wang, J., Liu, Y., and Levy, C. (2021). Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 526–536, New York, NY, USA. Association for Computing Machinery.

Wick, M., panda, s., and Tristan, J.-B. (2019). Unlocking fairness: A trade-off revisited. In Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 32, pages 8783–8792. Curran Associates, Inc.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. (2013). Learning fair representations. In Dasgupta, S. and McAllester, D., editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 325–333, Atlanta, Georgia, USA. PMLR.

Zhao, S., Song, J., and Ermon, S. (2019). Infovae: Balancing learning and inference in variational autoencoders. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

A Formulation of VIB Objective

According to Barber and Agakov (2003), we obtain the following variational bounds on the mutual information. We first derive the lower bound for $I(y, \tilde{y})$ as follows:

$$\begin{aligned} I(y,\tilde{y}) &= \int q_\phi(y,\tilde{y}) \log \frac{q_\phi(\tilde{y} \mid y) p_\theta(\tilde{y} \mid y)}{q_\phi(\tilde{y}) p_\theta(\tilde{y} \mid y)} dy d\tilde{y} \\ &= \int q_\phi(y,\tilde{y}) \log \frac{p_\theta(\tilde{y} \mid y)}{q_\phi(\tilde{y})} dy d\tilde{y} + \int q_\phi(y) \mathrm{KL}[q_\phi(\tilde{y} \mid y) \,||\, p_\theta(\tilde{y} \mid y)] dy \\ &\geq \int q_\phi(y,\tilde{y}) \log \frac{p_\theta(\tilde{y} \mid y)}{q_\phi(\tilde{y})} dy d\tilde{y} \\ &= \int q_\phi(y,\tilde{y}) \log p_\theta(\tilde{y} \mid y) dy d\tilde{y} + H(\tilde{y}). \end{aligned} \tag{7}$$

Similar to the above derivation, we then obtain the upper bound for $I(z, y)$:

$$\begin{aligned} I(z,y) &= \int q_\phi(z,y) \log \frac{q_\phi(y \mid z)}{q_\phi(y)} dy dz \\ &= \int q_\phi(z,y) \log \frac{q_\phi(y \mid z) p_\theta(y)}{q_\phi(y) p_\theta(y)} dy dz \\ &= \int q_\phi(z,y) \log \frac{q_\phi(y \mid z)}{p_\theta(y)} dy dz - \mathrm{KL}[q_\phi(y) \,||\, p_\theta(y)] dz \\ &\leq \int q_\phi(z,y) \log \frac{q_\phi(y \mid z)}{p_\theta(y)} dz dy. \end{aligned} \tag{8}$$

By dropping $H(\tilde{y})$ and using $q_\phi(y, z, \tilde{y}) = q_\phi(y \mid z) p(z, \tilde{y})$ we can get:

$$\begin{aligned} I(y,\tilde{y}) - \gamma I(z,y) &\geq \int q_\phi(y,\tilde{y}) \log p_\theta(\tilde{y} \mid y) dy d\tilde{y} + H(\tilde{y}) - \gamma \int q_\phi(z,y) \log \frac{q_\phi(y \mid z)}{p_\theta(y)} dz dy \\ &\geq \int q_\phi(y \mid z) \log p_\theta(\tilde{y} \mid y) dy - \gamma \int q_\phi(y \mid z) \log \frac{q_\phi(y \mid z)}{p_\theta(y)} dz dy \\ &= \mathbb{E}_{q_\phi(y|z)}[\log p_\theta(\tilde{y} \mid y) - \gamma \mathrm{KL}[q_\phi(y \mid z) \,||\, p_\theta(y)]]. \end{aligned} \tag{9}$$

B Implementation Details

We implement one layer encoder and decoder with the 'ReLu' activation function. For all the baseline models, we fix the encoder and decoder structures. For FFVAE, VFAE and VAE, which require downstream classifiers for the learned representations, we all use the same prediction classifier (Logistic Regression). To approximate $\mathrm{KL}[q(z, s \mid x)||p(z, s)]$, we use the same method in FactorVAE (Kim and Mnih, 2018). We train a discriminator to distinguish the fake samples drawn from $p(z)p(s) = \mathcal{N}(0, I)\mathrm{Uniform}(0, 1)$ and 'real' samples obtained from the encoder. Then, we can use

$$\mathbb{E}_{q_\phi(z,s)}[\log d(u = 1 \mid z, s) - \log d(u = 0 \mid z, s)]$$

to approximate $\mathrm{KL}[q(z \mid x)||p(z)]$, where d is the discriminator, and we use $u = 1$ to denote 'real' samples and we use $u = 0$ to denote 'fake' samples.

C Generation of Synthetic Data

We generate two multivariate Gaussian distributions for each label class. For the positive class, we have $\mu = (2, 2)$ and $\mathrm{Cov} = [[5, 1], [1, 5]]$. For the negative class, we have $\mu = (-2, -2)$ and $\mathrm{Cov} = [[10, 1], [1, 3]]$. Then we assign the sensitive attribute from a Bernoulli distribution where $p(a = 1) = \frac{p(x'|y=1)}{p(x'|y=1)+p(x'|y=0)}$ and x' is a transformed version of x which can be computed by $x' = \begin{bmatrix} cos(\phi) & -sin(\phi) \\ sin(\phi) & cos(\phi) \end{bmatrix} X$.

We totally generate 10800 synthetic samples with 2-dimensional non-sensitive attributes space and 1-d sensitive attribute space. We require the synthetic data to be as fair as possible. Therefore, the fairness violation on the synthetic data is very low, which is $\Delta_{\text{DP}} = 0.2$.

Table 7: Test accuracy (%) and fairness violation on the synthetic, Adult and Compas dataset of binary-sensitive features under various corruption ratios from 10% to 40% with instance-dependent label bias. We use 'ours+LR' and 'ours+D' to denote the measurement of our method using a downstream classifier on the learned representation z and the learned latent label y, respectively. We report the results in the format of mean $\pm$ standard deviation.

Method	r-10%		r-20%		r-30%		r- 40%	
	ACC	DEO	ACC	DEO	ACC	DEO	ACC	DEO
Dataset: Synthetic								
FFVAE	85.9±1.1	0.02±0.00	84.1±1.2	0.04±0.01	81.9±1.0	0.11±0.01	80.0±2.0	0.15±0.02
VFAE	81.8±2.0	0.03±0.00	82.2±1.8	0.12±0.01	81.0±2.3	0.24±0.05	79.1±1.9	0.28±0.04
VAE	85.8±1.0	0.02±0.00	85.2±0.9	0.05±0.01	82.8±1.7	0.11±0.02	79.7±1.2	0.22±0.05
CORES^2	85.5±2.3	0.03±0.01	85.6±1.1	0.05±0.02	85.5±2.1	0.06±0.02	85.3±1.1	0.11±0.04
GPL	85.5±1.1	0.03±0.00	85.2±1.4	0.04±0.02	84.5±1.3	0.05±0.02	83.4±1.5	0.12±0.04
Ours+LR	85.5±1.0	**0.01±0.00**	85.4±0.9	0.04±0.01	85.8±1.3	0.08±0.02	84.6±0.8	0.10±0.05
Ours+D	**86.3±0.9**	**0.01±0.00**	**86.3±1.3**	**0.02±0.00**	**86.0±1.6**	**0.04±0.01**	**85.7±1.2**	**0.06±0.00**
Dataset: Adult, a = gender								
FFVAE	82.7±2.4	0.20±0.05	82.2±1.5	0.24±0.03	79.9±1.4	0.22±0.06	76.3±1.3	0.28±0.05
VFAE	82.5±1.8	0.07±0.02	81.6±1.2	0.14±0.04	78.3±1.8	**0.07±0.02**	77.7±2.9	0.18±0.07
VAE	81.7±1.2	0.13±0.02	82.5±1.1	0.28±0.03	79.7±1.5	0.29±0.09	75.4±2.2	0.47±0.19
CORES^2	78.2±1.7	0.10±0.03	78.1±3.8	0.12±0.04	79.8±2.7	0.19±0.03	79.4±1.4	0.20±0.09
GPL	80.7±1.6	0.17±0.05	76.5±4.0	0.12±0.03	75.8±0.9	0.13±0.05	74.3±3.3	0.12±0.03
Ours+LR	82.5±1.1	0.08±0.02	**83.2±0.9**	0.10±0.03	80.1±1.9	0.13±0.05	74.8±2.8	0.18±0.04
Ours+D	**82.7±1.6**	**0.02±0.00**	83.1±0.8	**0.03±0.01**	**81.1±1.8**	0.08±0.02	**80.9±1.2**	**0.10±0.01**
Dataset: Compas, a = Race								
FFVAE	67.5±2.1	0.21±0.05	68.1±1.9	0.22±0.07	67.2±3.1	0.17±0.02	66.3±2.4	0.14±0.02
VFAE	65.9±1.8	**0.15±0.03**	65.4±3.7	**0.12±0.02**	65.2±3.5	**0.13±0.02**	63.6±2.2	**0.13±0.01**
VAE	66.8±3.4	0.21±0.07	66.9±1.4	0.23±0.11	65.4±2.1	0.21±0.08	65.7±3.7	0.26±0.09
CORES^2	66.0±2.3	0.15±0.07	66.7±2.9	0.15±0.06	66.9±2.5	0.16±0.06	66.3±3.8	0.14±0.05
GPL	63.6±2.9	0.18±0.09	61.2±4.3	0.15±0.06	64.1±3.3	0.16±0.05	56.8±3.2	0.15±0.05
Ours+LR	67.0±1.1	0.19±0.03	67.2±3.7	0.21±0.06	68.0±1.3	0.21±0.09	65.9±1.7	0.20±0.09
Ours+D	**68.5±1.2**	0.19±0.06	**68.6±1.3**	0.17±0.08	**69.1±2.6**	0.18±0.07	**66.5±1.6**	0.19±0.06

D Datasets Description

Synthetic Dataset: We use the same synthetic data generation approach mentioned in fairness constraint method (Bilal Zafar et al., 2016). We generate two multivariate Gaussian distributions for each label class. Then we randomly assign the sensitive attribute to each sample. For the synthetic data, we control the data to be fair by enforcing Δ_{DP} close to 0. For illustration, we only consider the binary sensitive attribute in synthetic data.

Adult Dataset[1]**:** The target value is whether an individual's annual income is over \$50k. The original feature dimension in this dataset is 13. After feature aggregation and encoding, the feature dimension is expanded to 35. The sensitive attributes are 'Gender' and 'Race'. In the binary sensitive attribute setting, we define 'Gender' as our interested sensitive attribute and 'Gender = Female' as the protected group. In the multi-attribute setting, we define 'Gender' and 'Race' as sensitive attributes and 'Gender = Female' combined with 'Race = Black' as the protected group.

Compas Dataset[2]**:** This data is from COMPAS, which is a tool used by judges, probation and parole officers to assess the risk of a criminal to re-offend. We focus on the prediction of 'Risk of Recidivism' (Arrest). The Compas system is found to be biased in favor of white and female defendants over a two-year follow-up period. In the binary sensitive attribute setting, we define 'Race' to be the target attribute and 'Race = Black' as the protected group. In the multi-attributes setting, we

[1] http://archive.ics.uci.edu/ml/datasets/Adult

[2] www.propublica.org/article/how-we-analyzed-the-compasrecidivism-algorithm

define both 'Race' and 'Gender' as the sensitive attributes and 'Race=Black' and 'Gender=Male' as the protected group. After feature aggregation and encoding, the feature dimension is reduced to 11.

E Additional Experiments

E.1 Evaluate Our Methods on Clean Data

We also directly evaluate our method with the same baseline models on the clean data (synthetic data). Even with the clean dataset, our method outperforms the other baselines w.r.t. accuracy and fairness. Compared with Table 2, we can see the results under different corruption ratios of our methods are very close to the results on the clean dataset.

Table 8: Test accuracy (%) and fairness violations measured on the clean synthetic dataset. We use 'ours+LR' and 'ours+D' to denote the measurement of our method using downstream classifier on the learned representation z and the learned latent label y respectively. We report the results in the format of mean $\pm$ standard deviation.

Metric	VAE	FFVAE	VFAE	CORES^2	GPL	ours+LR	ours+D
ACC	86.2 $\pm$ 1.2	86.0$\pm$1.3	83.1$\pm$0.9	86.1$\pm$1.1	86.5$\pm$2.1	86.4$\pm$0.7	86.8$\pm$1.2
DEO	0.03$\pm$0.00	0.0.03$\pm$0.01	0.05$\pm$0.01	0.01$\pm$0.00	0.02$\pm$0.00	0.01$\pm$0.00	0.01$\pm$0.00
Δ_{DP}	0.03$\pm$0.00	0.08$\pm$0.02	0.12$\pm$0.03	0.05$\pm$0.01	0.05$\pm$0.00	0.04$\pm$0.01	0.03$\pm$0.00

E.2 Other Fairness Measure

In this paper, we emphasize that our method does not need to specify the form of fairness notions in advance. We only report Δ_{DP} in Section 4, but our proposed method can be evaluated using other fairness notions. We also conduct experiments on the difference of equal opportunity (DEO) (Hardt et al., 2016), which is defined as:

$$\text{DEO} = |\mathbb{E}(\hat{y} = 1 \mid y = 1, a = 1) - \mathbb{E}(\hat{y} = 1 \mid y = 1, a = 0)|.$$

Overall, the performance is similar as measured in Δ_{DP} while both ours+D and ours+LR achieve lower fairness violations. For Compas Dataset, it is worth noting that, though FVAE has the lowest DEO, it has the lowest accuracy among all the VAE-related methods at the same time.