Achieving Fairness through Separability: A Unified Framework for Fair Representation Learning

Taeuk Jang
Purdue University

Hongchang Gao
Temple University

Pengyi Shi
Purdue University

Xiaoqian Wang
Purdue University

Abstract

Fairness is a growing concern in machine learning as state-of-the-art models may amplify social prejudice by making biased predictions against specific demographics such as race and gender. Such discrimination raises issues in various fields such as employment, criminal justice, and trust score evaluation. To address the concerns, we propose learning fair representation through a straightforward yet effective approach to project intrinsic information while filtering sensitive information for downstream tasks. Our model consists of two goals: one is to ensure that the latent data from different demographic groups is non-separable (i.e., make the latent data distribution independent of the sensitive feature to improve fairness); the other is to maximize the separability of latent data from different classes (i.e., maintain the discriminative power of data for the sake of the downstream tasks like classification). Our method adopts a non-zero-sum adversarial game to minimize the distance between data from different demographic groups while maximizing the margin between data from different classes. Moreover, the proposed objective function can be easily generalized to multiple sensitive attributes and multi-class scenarios as it upper bounds popular fairness metrics in these cases. We provide theoretical analysis of the fairness of our model and validate w.r.t. both fairness and predictive performance on benchmark datasets.

1 Introduction

Machine learning has made tremendous progress in autonomous driving, natural language processing, and many other fields. Although recent models excel in various applications and even outperform humans in some fields, we should be cautious when leveraging the power as models can make biased predictions across different demographic groups.

The biased prediction in machine learning is a rising concern due to the potential societal and legal impacts. Recent research found that various types of discrimination pervade artificial intelligence. For example, the future risk assessment software COMPAS has been found to be biased to different population groups (Angwin et al., 2016; Dressel and Farid, 2018). In particular, African Americans who do not commit future crimes are more likely to be mistaken as high-risk than Caucasians with similar profiles, i.e., higher false positive rate in the unprivileged group. Moreover, machine learning based job hiring platforms assign higher qualification scores for unqualified men than qualified women (Lahoti et al., 2019).

One major reason behind discrimination in model prediction is data bias (Barocas and Selbst, 2016). Since data is collected based on human decisions, data bias occurs in various forms (Mehrabi et al., 2019). Models trained on such data can replicate or even amplify the bias existing in the data and make biased predictions. However, addressing the biased prediction is not a trivial problem. One straightforward solution, *fairness through blindness*, simply removes *sensitive features* (e.g., race or gender) from the data. However, due to the feature redundancy, it remains *sensitive-relevant features* that are indicative of the sensitive features (Pedreshi et al., 2008). For example, *race* is excluded from the data, the model can learn from the sensitive-relevant features (e.g., ZIP code) and make the biased prediction.

Recent works aim to improve fairness by optimizing dual objectives: utility (e.g., accuracy) and fairness (e.g., equal opportunity (Hardt et al., 2016)). These methods often focus on specific fairness metrics and target certain downstream tasks such as classification (Rezaei et al., 2020), clustering (Huang et al., 2019)), and image generation (Sarhan et al., 2020). Particularly, LAFTR (Madras et al., 2018) employs respective loss functions for demographic parity and equalized odds to learn transferably fair repre-

Proceedings of the 27[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2024, Valencia, Spain. PMLR: Volume 238.

sentation for classification in an adversarial learning approach. In information theory literature, some works proposed to obfuscate sensitive information from the variational posterior distribution. For instance, Creager et al. (2019) aimed at minimizing total correlation (Chen et al., 2018), while Song et al. (2019); Cui et al. (2023) proposed to minimize mutual information between latent representation and sensitive information. However, these methods usually require approximations in the objective functions since the underlying distributions are intractable.

In contrast, this paper introduces fair representation learning by regularizing the data distribution from a new perspective. The proposed method learns a latent distribution where data from different *sensitive groups* — characterized by sensitive features — are *non-separable*, while data from different *classes* — characterized by the target label — are maximally *separable*. We achieve the non-separability on latent distribution w.r.t. sensitive features by minimizing the upper bound of the separability and increasing the separability w.r.t. target label by maximizing the marginal distance of decision boundaries among different classes. Specifically, we learn such representation by enforcing the latent distribution w.r.t. sensitive features to be non-separable even with a very powerful classifier, and meanwhile, the latent distribution w.r.t. target label to be easily classified with a simple linear classifier. The classifier proposed in our model is used to regularize the latent distribution but is not tied up with any specific fairness metric or downstream task. The proposed method is simple but effective and also can be generalized to multi-class classifications with multiple sensitive attributes. In addition, our *unified objective* is capable of improving various fairness metrics, supported by rigorous theoretical analysis. We provide theoretical proof that popular fairness metrics, e.g., demographic parity, equal opportunity, and equalized odds, are upper bounded by the proposed objective function. Further, we present experimental results on three datasets to validate our model in utility, fairness, and quality of the latent data distribution. We summarize our contribution as follows:

- We propose a novel fair representation learning method by enforcing the non-separability w.r.t. sensitive features while retaining the maximal discriminative power w.r.t. target label.
- Our approach offers a unified objective that improves various fairness metrics and enhances robustness.
- The proposed method is straightforward yet powerful and applicable to multi-label and multi-sensitive attribute scenarios.
- We provide comprehensive theoretical analysis, establishing the upper bound of the major fairness metrics including demographic parity, equal opportunity, and equalized odds.

2 Related Work

To mitigate the challenge of *prediction outcome discrimination* and *prediction quality disparity* (Du et al., 2019), a variety of approaches have been proposed. For instance, pre-processing methods explore feature transformation to remove the dependence on sensitive features (Louizos et al., 2015). Other methods conduct re-sampling (Jiang and Nachum, 2019) and re-weighting (Chai and Wang, 2022a) based on the sensitive characteristics to mitigate bias from data. This approach is further expanded to address the cases when sensitive features are unavailable (Liu et al., 2021; Lahoti et al., 2020; Jang and Wang, 2023; Chai and Wang, 2022b). However, it has been found that merely transforming or perturbing data can harm performance (Corbett-Davies et al., 2017). To address this, Jang et al. (2021) augment the dataset with synthetic counterparts to perfectly balance the distribution.

Another approach to alleviating discrimination involves constraining the posteriors to satisfy specific fairness metrics. Hardt et al. (2016) introduced equalized odds, which balances the true positive and true negative rate from different demographic groups. Additionally, they learn a threshold to yield a fairer outcome from an unfair (black-box) model in a post-processing manner. Subsequent works (Kim et al., 2020; Jang et al., 2022) extend to various fairness notions and explore contextualized optimal trade-off, i.e., Pareto frontiers.

To learn fair representation in the latent space, zero-sum adversarial learning is often employed (Beutel et al., 2017). This typically involves two players engaging in the same loss function, where one tries to maximize and the other adversarially minimizes the fairness violations. For instance, Madras et al. (2018) propose to adopt different group fairness metrics as the adversarial objectives and analyze the balance between utility and fairness. However, Roy and Boddeti (2019) argue that likelihood-based adversary may result in sub-optimal performance in sensitive information leakage. To address this, they introduce a non-zero-sum game framework for fair representation learning, investigating its equilibrium and convergence of optimization. In addition, Jovanović et al. (2023) study fairness certificate, which provides a fairness guarantee of arbitrary classifiers trained on top of learned representation. Normalizing flow is also explored for learning fair representation (Balunovic et al., 2022). (Cotter et al., 2019) studied general non-zero-sum optimization in non-convex or non-differentiable settings. However, few work has studied comprehensive theoretical analysis regarding the relationship between fairness metrics and non-zero-sum objectives.

Variational inference (Kingma and Welling, 2013) is also employed to learn representation oblivious to sensitive information. Specifically, Creager et al. (2019) suggest minimizing total correlation (Chen et al., 2018) to ensure that

the learned representation is independent of sensitive information. Similarly, Song et al. (2019) proposed to minimize mutual information between latent representation and sensitive information, while Liu et al. (2022) adopt distance covariance. However, these methods usually require approximations in the objective functions since the underlying distributions are intractable.

Since the aforementioned methods adopted adversarial learning, which has a potential risk of unstable optimization process and empirically poor trade-off between fairness and utility Sarhan et al. (2020), alternative approaches have been proposed to learn fair representation. For instance, disentanglement learning method (Sarhan et al., 2020) redirects the sensitive information to latent subspace by orthogonal prior regularization. Contrastive learning (Oh et al., 2022) offers another non-adversarial approach that enforces non-sensitive latents to be closer while sensitive latents to be apart. However, these methods often have complicated training procedures, e.g., orthogonal prior sampling, and contrastive sampling, compared to adversarial learning. Additionally, these methods do not provide a clear theoretical understanding, which can make it difficult to interpret the mechanism.

Our work also falls into fair representation learning, however, our method is distinguished from the existing works in the following points. First, in contrast to the methods, which learn a representation to adversarially maximize the prediction loss of sensitive features (Madras et al., 2018; Adel et al., 2019), our approach aims at learning distribution non-separable in different sensitive groups by maximizing the entropy for fair representation. The maximization of prediction loss on sensitive features may not ensure the independence between the latent representation and the sensitive feature in a general case (e.g., sensitive feature with multiple values), while our method remains effective as we directly optimize the feature extractor over a uniform distribution w.r.t. sensitive feature. Second, the proposed representation ensures maximizing the mean and minimizing the variance of margin among multiple classes, which is adaptive to multi-class classification and is robust to outliers in the classification tasks. Lastly, our work is differentiated from other non-adversarial learning methods (Oh et al., 2022; Sarhan et al., 2020), which primarily rely on empirical evidence. Conversely, our approach to learning the optimal fair margin of the representation is more efficient and provides a theoretical guarantee on various fairness metrics with multiple sensitive attributes involved.

3 FSNS: Fair Representation - Sensitive Non-Separable & Label Separable

3.1 Problem Definition

As discussed in the previous section, the biased prediction in traditional machine learning models originates from bias in the data. In order to address the discrimination in prediction, it is important to alleviate data bias, i.e., mitigating the disparity in the distribution of data from different sensitive groups. However, due to the existence of sensitive-relevant features, it is hard to directly filter the sensitive information in the original data space.

Here we propose to learn fair representation in the latent data space while maintaining the non-sensitive information to fulfill downstream tasks. Take the downstream task of classification as an example, with the input data $\mathbf{x}$, the sensitive feature $\mathbf{a}$, and the target label $\mathbf{y}$. Our goal is to learn a data representation $\mathbf{z}$ that is both *fair* (i.e., independent to the sensitive feature) and *discriminative* (i.e., maintaining the maximal discriminative power w.r.t. $\mathbf{y}$).

To achieve this goal, we propose our model **FSNS** (**F**air representation that **S**eparable on target label and **N**on-**S**eparable on sensitive features) with the following two strategies: 1) we minimize the distance between distributions from different sensitive groups such that a powerful classifier cannot predict which sensitive group does the data belong to; 2) we maximize the margin between the distributions from different classes such that a simple classifier can easily classify data w.r.t. target label in the latent space. This allows the model to project the data to a fair space where demographic discrimination is minimized, and the predictive power is maximally maintained.

3.2 Illustration of the FSNS Model

As illustrated in Fig. 1, the FSNS model consists of three modules: a feature extractor (H_θ) to learn the fair representation, a simple classifier (W), such as SVM, to predict the target label, and a sensitive feature predictor (C_ϕ). The goal is to learn fair latent representation such that the distribution w.r.t. the sensitive features is non-separable (data from different sensitive groups cannot be discriminated even with a powerful predictor C_ϕ), while the distribution w.r.t. the target label is maximally separable (data from different classes can be easily discriminated with a simple classifier W).

3.2.1 Notations

Given a data sample $\mathbf{x} \in \mathbb{R}^d$, we use one-hot encoding to represent the sensitive feature $\mathbf{a} \in \{0,1\}^k$ and target label $\mathbf{y} \in \{0,1\}^c$, where k is the number of possible values of the sensitive feature and c is the number of classes. Take

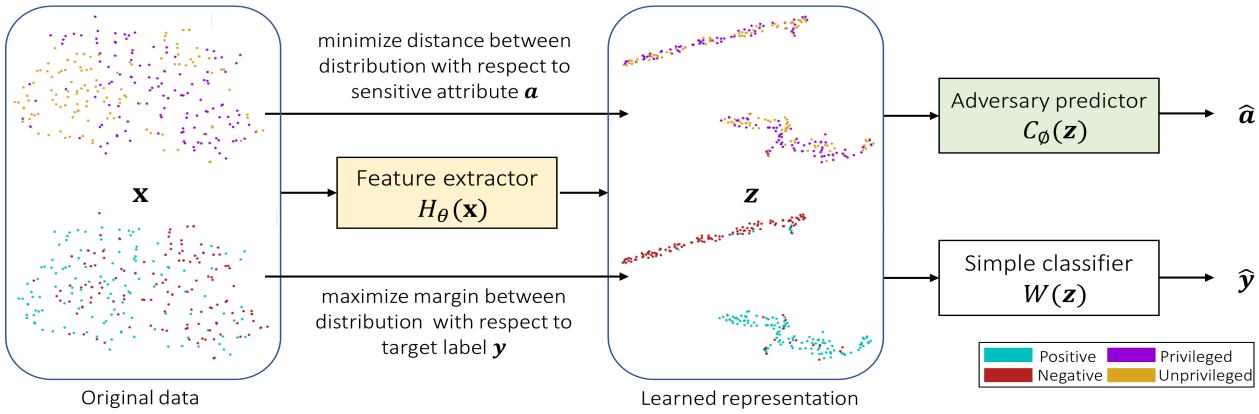


Figure 1: Illustration of FSNS model. In the original feature space, distributions from different sensitive groups are easily distinguishable, i.e., privileged (purple) v.s. unprivileged (gold), since **a** is correlated with **x**, thereby causing bias in data. In contrast, the distributions from different classes, i.e., positive (cyan) v.s. negative (brown), are not well separable. To learn fair representation while maintaining the predictive performance, the feature extractor H_θ projects the original space into latent space that is non-separable w.r.t. the sensitive features while maximally separable w.r.t. the target label.

race as an example for the sensitive feature,[1] then the goal of fair prediction is to eliminate bias among the *race* groups ($k = 4$): Asian, White, Black, and Hispanic.

For a given data sample $\mathbf{x}$, the feature extractor provides representation $H_\theta(\mathbf{x})$ which are used as input for the sensitive feature predictor C_ϕ and the classifier W. We denote the predicted sensitive feature as $C_\phi(H_\theta(\mathbf{x})) = \hat{\mathbf{a}} = [\hat{a}_1,\ \hat{a}_2,\ \ldots,\ \hat{a}_k]^\top \in [0,1]^k$, and the predicted label as $W(H_\theta(\mathbf{x})) = \hat{\mathbf{y}} = [\hat{y}_1,\ \hat{y}_2,\ \ldots,\ \hat{y}_c]^\top \in [0,1]^c$. The predicted outcomes can be interpreted as a probability distribution on all the possible values, e.g., $\hat{y}_j$ is the predicted probability of outcome being in the j-th class. Hence, we also have $\sum_{j=1}^{k} \hat{a}_j = 1$ and $\sum_{j=1}^{c} \hat{y}_j = 1$.

We denote the feature distribution for data samples $\mathbf{x}$ as $\mathcal{X}$, and the distribution for sensitive feature $\mathbf{a}$ as $\mathcal{A}$. We define the loss function that measures the difference between the sensitive feature $\mathbf{a}$ and the predicted outcome as

$$\mathcal{L}\big(\mathbf{a}, C_\phi(H_\theta(\mathbf{x}))\big) = \sum_{j=1}^{k} \mathcal{L}(a_j, \hat{a}_j)\,. \tag{1}$$

Examples of $\mathcal{L}$ includes cross-entropy loss, ℓ_1-norm loss, etc. For example, if $k = 3$ for the sensitive feature, $\mathbf{a}$ takes value $[1, 0, 0]^\top$, and considers the ℓ_1-norm loss, or say, the mean absolute error (MAE), then Eqn. (1) equals

$$\mathcal{L}\big(\mathbf{a}, C_\phi(H_\theta(\mathbf{x}))\big) = |1 - \hat{a}_1| + |\hat{a}_2| + |\hat{a}_3|$$

with $\hat{a}_j$ being the j-th coordinate from the predicted outcome vector $C_\phi(H_\theta(\mathbf{x}))$.

[1] It is notable that our model can be easily adapted to multiple sensitive features by enumerating the combination of possible values in multiple features.

3.2.2 Loss function and training

For the two predictors, we train C_ϕ to minimize the following sensitive loss $\mathcal{L}_a$ such that C_ϕ is a complex predictor, e.g., MLP, to predict the sensitive feature $\mathbf{a}$:

$$\mathcal{L}_a = \mathbb{E}_{(\mathbf{x},\mathbf{a})\sim\mathcal{X}\times\mathcal{A}}\left[\mathcal{L}\big(\mathbf{a}, C_\phi(H_\theta(\mathbf{x}))\big)\right]\,, \tag{2}$$

where $\mathcal{L}$ is defined in Eqn. (1). On the contrary, we build W as a simple linear support vector machine (SVM) to predict the target label $\mathbf{y}$.

For the feature extractor $H_\theta(\mathbf{x})$, we train it to achieve the following two goals.

The first goal is to learn fair representation such that the latent distribution from different sensitive groups is non-separable. To accomplish this, we propose to minimize the following fair loss $\mathcal{L}_{fair}$, inspired by Roy and Boddeti (2019):

$$\mathcal{L}_{fair} = \frac{1}{|\mathcal{A}|}\sum_{\mathbf{a}\in\mathcal{A}} \mathbb{E}_{\mathbf{x}\sim\mathcal{X}_\mathbf{a}}\left[\mathcal{L}\big(\bar{\mathbf{a}}, C_\phi(H_\theta(\mathbf{x}))\big)\right]\,, \tag{3}$$

where $\bar{\mathbf{a}} = [1/k,\ \ldots,\ 1/k]^\top$ denotes the non-informative, uniform prediction that assigns equal distribution over each of the k classes in the sensitive feature. We denote $\mathcal{X}_\mathbf{a}$ to represent the feature distribution with sensitive feature $\mathbf{a} \in \mathcal{A}$. By updating H_θ to minimize $\mathcal{L}_{fair}$, H_θ learns to generate fair representation as it maximizes the entropy (uncertainty) in inferring sensitive information such that even a powerful predictor C_ϕ cannot differentiate among the sensitive groups. Moreover, such non-zero-sum adversarial approach offers superior equilibrium and convergence properties than traditional zero-sum adversarial game (Roy and Boddeti, 2019). In Theorem 2 below, we

provide rigorous proof that minimizing this loss function leads to primary fairness metrics, including demographic parity, equal opportunity, and equalized odds.

The second goal of $H_\theta(\mathbf{x})$ is to learn a representation with discriminative power for classification such that the latent distribution from different classes is maximally separable. It turns out that simply maximizing the minimum margin of SVM has poor generalization (Zhang and Zhou, 2019), and instead, the margin distribution should be considered (Gao and Zhou, 2013). Following (Zhang and Zhou, 2017), we optimize H_θ to maximize the average margin and minimize the margin variance, which characterizes the margin distribution in SVM to maintain the maximal discriminative power w.r.t. the target label with stable performance.

Given a sample data $\mathbf{x} \in \mathbb{R}^d$, denote t to be the true class for the target label, i.e., $\mathbf{y} \in \{0,1\}^c$ takes the value with the t-th component being 1 and all others being 0. With the linear classifier $W = [\mathbf{w}_1,\ \mathbf{w}_2,\ \ldots,\ \mathbf{w}_c] \in \mathbb{R}^{d\times c}$, denote the mean margin of classifier as

$$M = \frac{1}{n}\sum_{i=1}^{n}[\mathbf{w}_t^\top H_\theta(\mathbf{x}_i) - \max_{l\neq t}\mathbf{w}_l^\top H_\theta(\mathbf{x}_i)]$$

and the variances of the margin of a sample $\mathbf{x}_i$ as:

$$\xi_i = \max\big(M - \big(\mathbf{w}_t^\top H_\theta(\mathbf{x}_i) - \max_{l\neq t}\mathbf{w}_l^\top H_\theta(\mathbf{x}_i)\big) - \gamma, 0\big),$$

$$\epsilon_i = \max\big(\big(\mathbf{w}_t^\top H_\theta(\mathbf{x}_i) - \max_{l\neq t}\mathbf{w}_l^\top H_\theta(\mathbf{x}_i)\big) - M - \gamma, 0\big),$$

where ξ_i and ϵ_i are two types of deviations (less or greater than margin mean), and $\gamma \geq 0$ is the soft margin.

We optimize H_θ to minimize the max-margin loss function $\mathcal{L}_{mm}$ defined as:

$$\mathcal{L}_{mm} = \Omega(\mathbf{w}) + \mathop{\mathbb{E}}_{(\mathbf{x},\mathbf{y})\sim\mathcal{X}\times\mathcal{Y}}\left[-M + \lambda\frac{\xi^2+\mu\epsilon^2}{(1-\gamma)^2}\right], \quad (4)$$

where $\Omega(\cdot)$ is ℓ_2-norm regularizer, λ and μ are trading-off parameters introduced in (Zhang and Zhou, 2017). Note that we can scale the weight $\mathbf{w}$ in SVM to fix margin mean M to be 1, in turn, we have $(1-\gamma)^2$ instead of $(M-\gamma)^2$ in Eqn. (4). Therefore, our goal is to optimize three modules to minimize the following objectives:

$$\begin{aligned}
&\arg\min_{H_\theta, W} \quad \mathcal{L}_{mm}(\mathbf{y}, W(H_\theta(\mathbf{x}))) + \lambda_{fair}\mathcal{L}_{fair}\big(\bar{\mathbf{a}}, C_\phi(H_\theta(\mathbf{x}))\big),\\
&\arg\min_{C_\phi} \qquad\qquad\qquad \mathcal{L}_a\big(\mathbf{a}, C_\phi(H_\theta(\mathbf{x}))\big).
\end{aligned}$$

Note that FSNS has min-max property but all modules are optimized to minimize their objectives, which yields *non-zero-sum* game. We summarize the optimization steps of our algorithm in Algorithm 1, where $\tilde{\mathcal{L}}$ indicates the empirical loss. As a reminder, SVM is only used to learn fair representations in pre-processing. We have the flexibility of choosing any classifier for downstream tasks after learning the representation.

Algorithm 1 Optimization Algorithm of FSNS Model

Input Dataset $\{(\mathbf{x}_i, \mathbf{a}_i, \mathbf{y}_i)\}_{i=1}^n$, number of mini-batches n_b, learning rate α_θ, α_w and α_ϕ. SVM hyper-parameters γ, λ, μ.
Output Fair representation learning model H_θ.
Initialize W, C_ϕ randomly.
Initialize H_θ with the latent representation by pre-training our model to just classify w.r.t. $\mathbf{y}$.
while not converge **do**
  for t = 1, 2, ..., n_b **do**
    1. Update C_ϕ using gradient descent w.r.t. $\tilde{\mathcal{L}}_a$.
    2. Update H_θ using gradient descent w.r.t. both $\tilde{\mathcal{L}}_{fair}$ and $\tilde{\mathcal{L}}_{mm}$.
    3. Update W using gradient descent w.r.t. $\tilde{\mathcal{L}}_{mm}$.
  end for
end while

3.3 Theoretical Property

In this section, we study the theoretical property of FSNS regarding minimizing $\mathcal{L}_{fair}$, i.e., maximizing the entropy on sensitive attribute prediction.

Theorem 1. *Consider the optimal classifier $C_\phi^* : \mathbb{R}^h \to [0,1]^k$ for the sensitive feature prediction. Denote $\mathcal{L}_{fair}(C_\phi^*) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{X}}\left[\mathcal{L}\big(\bar{\boldsymbol{a}}, C_\phi^*(H_\theta(\boldsymbol{x}))\big)\right]$, i.e., the fair loss using C_ϕ^* and the representation learned by H_θ in FSNS. Under the ℓ_1-norm loss or the cross-entropy loss for $\mathcal{L}$, the following fairness metrics for multi-class classification with multiple sensitive groups are bounded by $\mathcal{L}_{fair}$ for some classifier $W : \mathbb{R}^h \to [0,1]^c$ for the target label prediction:*

- *the demographic disparity:*

$$\begin{aligned}
\mathcal{L}_{fair}(C_\phi^*) &\geq \Delta_{DP}\\
&= \max_{a,b\in\mathcal{A}}\Big\|\mathbb{E}\big[W(H_\theta(\boldsymbol{x}))|A=a\big]\\
&\qquad - \mathbb{E}\big[W(H_\theta(\boldsymbol{x}))|A=b\big]\Big\|_1,
\end{aligned}$$

- *the equalized odds:*

$$\begin{aligned}
&\mathcal{L}_{fair}(C_\phi^*)\\
&\quad\geq \Delta_{EO}\\
&\quad= C\max_{a,b\in\mathcal{A}}\sum_{j=1}^{c}\Big(\Big\|\mathbb{E}\big[W(H_\theta(\boldsymbol{x}))|A=a, Y=j\big]\\
&\qquad - \mathbb{E}\big[W(H_\theta(\boldsymbol{x}))|A=b, Y=j\big]\Big\|_1\Big),
\end{aligned}$$

where $C = \min_{j\in\{1,\cdots,c\}}\big(P(y\sim\mathcal{Y}_j)\big)$.

Proof of Theorem 2 can be found in the supplementary materials. Since encoder H_θ minimizes w.r.t. $\mathcal{L}_{fair}$, Theorem

2 indicates that the fairness metrics are bounded by $\mathcal{L}_{fair}$ of the optimal sensitive attribute predictor C^*_ϕ for any classifier takes learned representation $H_\theta(\mathbf{x})$ as input. Moreover, when $\mathcal{L}_{fair}(C^*_\phi) \approx 0$, all these fairness metrics are also close to 0.

In the supplementary materials, we explain the role of using C^*_ϕ from a min-max perspective. At a high level, when we use C^*_ϕ in the fairness loss to train H_θ, we maximize a lower bound for $\mathcal{L}_{fair}$, where this lower bound relates to the separability of the classifier, i.e., the ability that this classifier can tell apart the sensitive feature predicted from data drawn from different groups. Thus, when minimizing $\mathcal{L}_{fair}$ to be close to 0, we make sure that the maximal separability (produced when using C^*_ϕ) will be controlled to be close to 0. The min-max interaction enables us to minimize the distance between distributions from different sensitive groups such that even the most powerful classifier cannot predict which sensitive group the data belongs to. This intuitively motivates the theoretical analysis of the proposed learning fair representation. When comparing our method with the related min-max methods (Adel et al., 2019; Madras et al., 2018), our FSNS model improves various fairness metrics in a unified objective and preserves higher accuracy in classification (validated in results in Section 4) by our max-margin objective in Eqn. (4).

4 Experiments

We conduct experiments to evaluate how our model FSNS affects the utility (accuracy in classification tasks) and fairness by comparing with the state-of-the-art models.

4.1 Experimental Setup

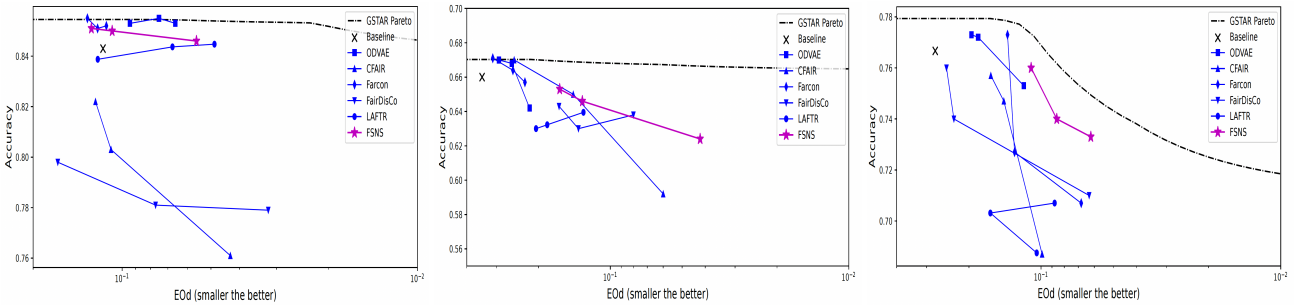
We compare our model with the following recent fair representation learning methods. LAFTR (Madras et al., 2018) is a fair representation learning model that adopts fairness metrics as the zero-sum adversarial objectives. ODVAE (Sarhan et al., 2020) learns fair disentanglement in the latent space by regularizing the posterior with orthogonal priors. CFair (Zhao et al., 2020) proposed to minimize the balanced error rate (BER) (Feldman et al., 2015) along with the conditional alignment of latent representation. Farcon (Oh et al., 2022) adopted a contrastive learning approach to learn fair representation by randomly swapping latent codes in the observation set. FairDisCo (Liu et al., 2022) is a variational method that employs distance covariance for latent independence without adversarial learning. Baseline is a network with the same structure as our feature extractor H_θ and classifier W. The difference between our model and the baseline is that we include the fairness module C_ϕ in the objective to validate the necessity of the fairness module of FSNS.

Our FSNS model consists of the feature extractor (H_θ) to learn fair representation, a classifier (W) to predict the target label, and a sensitive feature predictor (C_ϕ). Feature extractor and sensitive feature predictor are 4-layer fully connected neural networks with leaky ReLU as the activation function. All components in FSNS model is updated via ADAM optimizer (Kingma and Ba, 2014). For hyperparameter tuning, we used grid search on $\lambda_{fair} \in [0, 10]$ for balancing the fairness and accuracy of FSNS. We use Pytorch and Scikit-learn toolbox to implement our code and run the algorithm on a machine with four Quadro RTX 6000 GPUs and Intel I9-9960X.

We evaluate the performance of the methods on several fairness benchmark datasets. Specifically, we evaluate binary classification on the Adult Census Income Data (Kohavi, 1996), COMPAS, German credit data considering both a binary sensitive attribute and multiple sensitive attributes, i.e., intersectional bias (Ghosh et al., 2021), scenarios. We evaluate multi-class classification on ACSIncome, ACSTravelTime dataset (Ding et al., 2021). Since the original label is continuous value (dollars for ACSIncome and commuting time in minutes for ACSTravelTime), we categorize the label into three classes, i.e., $c = 3$, using equidistant quantiles (Denis et al., 2021). Detailed description of the dataset is as below:

- **Adult**: data from the UCI repository Kohavi (1996): The data contains 48,842 instances described by 14 features (workclass, age, education, sex, race, etc.) and the goal is to predict whether the income exceeds 50K USD per year. The feature *sex* is used as the sensitive feature.
- **Compas**[2]: The dataset includes 6,167 samples described by 401 features with the outcome showing if each person gets rearrested within two years. The feature *sex* is used as the sensitive feature in this dataset.
- **German** credit data from the UCI repository Dua and Graff (2019): The dataset contains 1,000 samples described by 20 features and the goal is to predict the credit risks. The feature *sex* is used as the sensitive feature.
- **ACSIncome** Ding et al. (2021): data from American Community Survey (ACS) Public Use Microdata Sample (PUMS). It is to predict an individual's income in dollars. It contains various information including COW (class of worker), educational level, etc of 1,664,500 samples from all states in the US.
- **ACSTravelTime** Ding et al. (2021): data from ACS PUMS. It is to predict an individual's commute time in minutes. It contains features including educational level, marital status, occupation, etc of 1,466,648 samples from all states in the US.

[2]https://github.com/propublica/compas-analysis

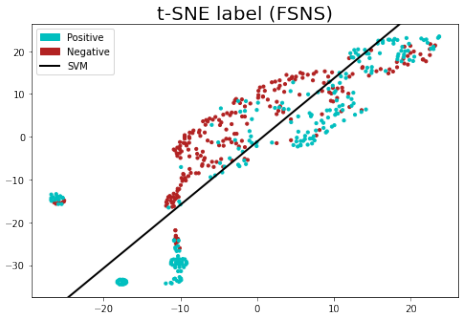


(a) Adult Dataset (b) COMPAS Dataset (c) German Dataset

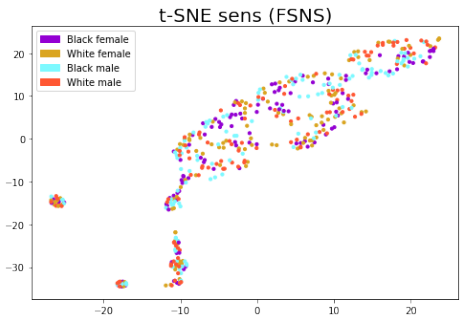
Figure 2: Comparison of accuracy-fairness trade-off on three datasets. Overall, FSNS achieves the best trade-off, i.e., best accuracy at the same level of fairness violations.

All features in each dataset are normalized to the range of $[0, 1]$. For each dataset, we randomly split it into training (70%), validation (15%), and test (15%) sets, and report the average results on the test set in 5 repetitions.

To measure fairness, we adopt famous *equalized odds* (EOd, the absolute difference of false positive rate and true positive rate between different sensitive groups) (Hardt et al., 2016).



(a) Label



(b) Sensitive attributes

Figure 3: t-SNE visualization of learned representation of FSNS on Adult dataset with intersectional sensitive attributes (gender and race).

4.2 Evaluation on Utility and Fairness

Fig. 2 presents the comparison of the utility and fairness trade-off by varying weights on the fairness constraints by each method. GSTAR *Pareto frontier* (Jang et al., 2022) provides the best achievable accuracy-fairness trade-offs in a model-specific manner. Note that the x-axis of the plot is inverted, meaning that the upper-right side corresponds to a better trade-off between fairness and accuracy.

Among the comparing methods, FSNS consistently exhibits a superior or competitive trade-off between accuracy and fairness compared with baselines. Notably, FSNS often yields results that closely align with the top-right corner of the Pareto frontier, which is desired. Specifically, FSNS achieves the highest or comparable accuracy at the same fairness violation level. Moreover, in COMPAS and german datasets, FSNS achieves the least fairness violation at a similar accuracy level. The results validate that FSNS effectively balances accuracy and fairness, which achieves either the best or comparable fairness with a marginal sacrifice of accuracy even with its simple structure.

4.3 Qualitative Analysis of the Learned Representation

We visualize the learned representation using t-SNE visualization (Maaten and Hinton, 2008) for Adult dataset in Fig 4. The ideal fair representation should preserve the predictiveness of the target label while remaining agnostic to sensitive information. Compared to other fair representation learning methods, FSNS (last column) shows better separation w.r.t. the target label (first row), while it is hard to distinguish the distribution between different sensitive groups in FSNS (second row). Whereas, the sensitive information can be easily distinguished in other distributions (LAFTR and ODVAE). Notably, FSNS delivers promising results with the simple structure without additional techniques such as adversarial learning (LAFTR) or variational inference (ODVAE). More results on other datasets and methods are included in the Supplementary.

4.4 Multiple Sensitive Attribute Scenarios

To validate the effectiveness of FSNS on multiple sensitive attributes, we here consider the combination of the two binary sensitive attributes. For adult dataset, we consider $\{gender, race\}$ as sensitive attributes, $\{gender, race\}$ for

Table 1: The result on the multiple sensitive attribute scenario on benchmark datasets. FSNS consistently achieves better or comparable fairness compared to baselines while preserving accuracy.

(a) Adult dataset. Gender and race as sensitive attributes.

		Baseline		FSNS		ODVAE		CFAIR		Farcon		FairDisCo	
Gender	Race	TPR ↑	FPR ↓	TPR ↑	FPR ↓	TPR ↑	FPR ↓	TPR ↑	FPR ↓	TPR ↑	FPR ↓	TPR ↑	FPR ↓
Female	Black	0.444	0.049	0.389	0.0	0.5	0.027	0.5	0.063	0.428	0.063	0.895	0.552
Female	White	0.622	0.126	0.417	0.028	0.566	0.038	0.613	0.065	0.519	0.078	0.922	0.559
Male	Black	0.484	0.071	0.385	0.015	0.452	0.03	0.703	0.066	0.511	0.063	0.934	0.569
Male	White	0.580	0.135	0.462	0.047	0.535	0.088	0.660	0.126	0.548	0.128	0.942	0.674
ACC ↑		0.810		<u>0.836</u>		<u>0.836</u>		**0.842**		0.812		0.798	
EOD ↓		0.255		**0.119**		<u>0.141</u>		0.223		0.185		0.170	

(b) COMPAS dataset. Gender and race as sensitive attributes.

		Baseline		FSNS		ODVAE		CFAIR		Farcon		FairDisCo	
Race	Gender	TPR ↑	FPR ↓	TPR ↑	FPR ↓	TPR ↑	FPR ↓	TPR ↑	FPR ↓	TPR ↑	FPR ↓	TPR ↑	FPR ↓
Black	Female	0.592	0.261	0.374	0.079	0.648	0.298	0.315	0.03	0.687	0.308	0.661	0.270
Black	Male	0.413	0.123	0.25	0.055	0.457	0.151	0.5	0.274	0.446	0.205	0.476	0.269
White	Female	0.422	0.149	0.222	0.015	0.467	0.164	0.556	0.209	0.422	0.119	0.439	0.212
White	Male	0.261	0.058	0.131	0.0	0.435	0.076	0.528	0.135	0.522	0.079	0.516	0.262
ACC ↑		**0.689**		0.666		<u>0.684</u>		0.680		0.682		0.626	
EOD ↓		0.535		<u>0.328</u>		0.435		0.429		0.454		**0.280**	

(c) German dataset. Gender and age as sensitive attributes.

		Baseline		FSNS		ODVAE		CFAIR		Farcon		FairDisCo	
Gender	Age	TPR ↑	FPR ↓	TPR ↑	FPR ↓	TPR ↑	FPR ↓	TPR ↑	FPR ↓	TPR ↑	FPR ↓	TPR ↑	FPR ↓
Female	Young	0.75	0.0	0.5	0.0	0.427	0.333	0.429	0.666	0.571	0.333	0.125	0.0
Female	Old	0.0	0.154	0.333	0.231	0.333	0.375	0.333	0.5	0.333	0.375	0.364	0.222
Male	Young	0.5	0.058	0.5	0.118	0.4	0.074	1.0	0.259	0.6	0.148	0.263	0.083
Male	Old	0.458	0.069	0.583	0.205	0.353	0.1	0.648	0.317	0.383	0.083	0.158	0.061
ACC ↑		**0.787**		<u>0.733</u>		0.707		0.680		0.720		0.725	
EOD ↓		0.904		<u>0.398</u>		**0.368**		0.978		0.494		0.461	

Table 2: Multi-class classification results. FSNS shows the best fairness improvement while preserving accuracy.

		Baseline	FSNS	ODVAE	CFAIR	Farcon	FairDisCo
ACSIncome	ACC ↑	**0.692**	0.683	0.618	0.678	0.690	0.619
	DP ↓	0.113	<u>0.056</u>	0.156	0.070	0.105	**0.025**
	EOD ↓	0.246	**0.104**	0.186	<u>0.126</u>	0.232	0.130
ACSTravelTime	ACC ↑	**0.483**	0.465	0.482	0.448	0.469	0.416
	DP ↓	0.062	0.025	0.068	<u>0.021</u>	**0.015**	0.025
	EOD ↓	0.159	**0.027**	0.107	<u>0.057</u>	0.061	0.076

COMPAS, and $\{gender, age\}$ for german dataset.

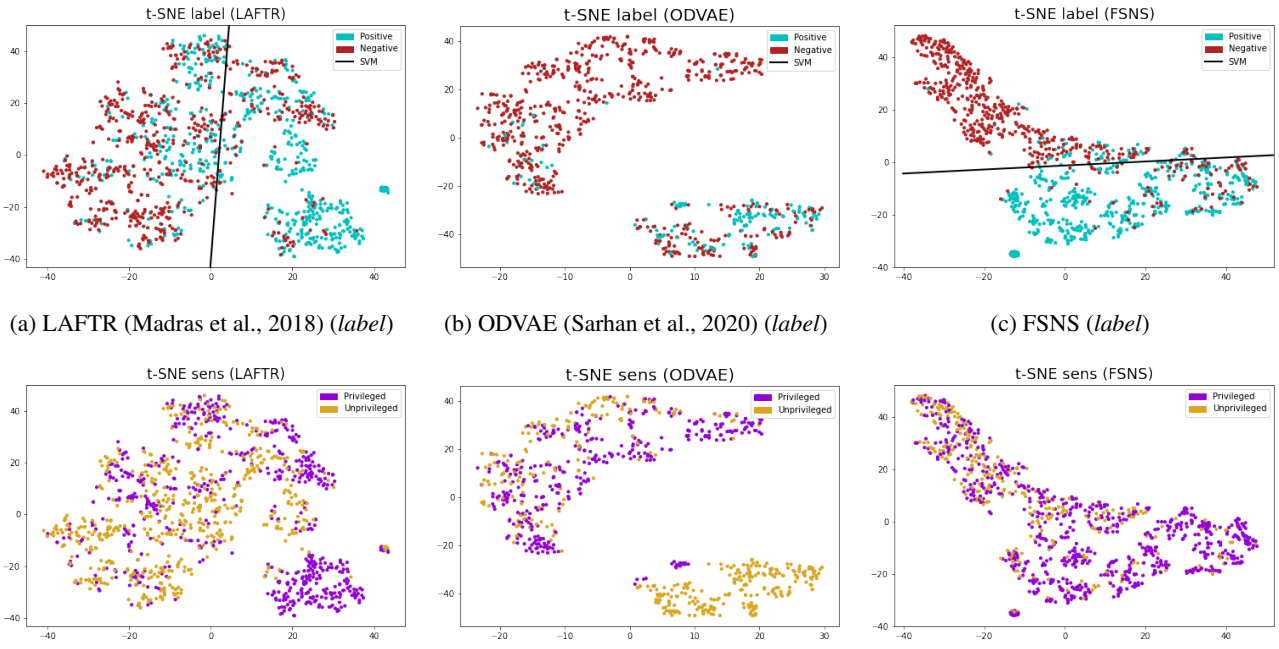
To measure fairness considering intersectionality (multiple sensitive attributes) Kearns et al. (2017), we formulate EOD by following the maximum difference of performance measures between two groups Wang et al. (2022); Ghosh et al. (2021) as:

$$\text{EOD} = \max_{a,b \in \mathcal{A}} |TPR_a - TPR_b| + |FPR_a - FPR_b|$$

In Table 1, we summarize the results on intersectional fairness on the three datasets. Note that LAFTR (Madras et al., 2018) is omitted since it is specifically designed to address fairness violations for a single binary sensitive attribute. The results demonstrate that FSNS effectively generalizes to a multiple sensitive attribute scenario. Specifically, FSNS significantly improved EOD violation and even improved accuracy compared to the baseline. Fig 3 presents a t-SNE visualization of the learned representation by FSNS. The distribution qualitatively reveals that the proposed approach, which aims at learning separability on target label and non-separability on sensitive attributes, is effective in achieving intersectional fairness scenarios.

4.5 Multi-class Classification Scenarios

To evaluate multi-class classification, we conduct experiments on two fairness benchmark datasets: ACSIncome, ACSTravelTime (Ding et al., 2021). ACSIncome aims to predict the annual income of a person, with race as the binary sensitive attribute (white vs black). The ACSTravelTime dataset aims to predict the commuting time of the



(a) LAFTR (Madras et al., 2018) (*label*) (b) ODVAE (Sarhan et al., 2020) (*label*) (c) FSNS (*label*)

(d) LAFTR (Madras et al., 2018) (*sens*) (e) ODVAE (Sarhan et al., 2020) (*sens*) (f) FSNS (*sens*)

Figure 4: t-SNE visualization of the learned representation from different methods. The first row shows the distribution w.r.t. the target label in classification, with a black line showing a linear SVM trained on the corresponding data representation. The second row depicts the same distribution with the sensitive feature-based coloring scheme. Ideal representation is separated by target label while less separable by sensitive feature.

working population over the age of 16, with race as the binary sensitive attribute (white vs black). For both datasets, the original label is a continuous value (dollars for ACSIncome and commuting time in minutes for ACSTravelTime). We divide the label into three classes using equidistant quantiles (Denis et al., 2021).

In Table 2, we summarize the results of multi-class classification tasks. In general, FSNS achieves the best fairness violations while maintaining comparable accuracy. This validates our claim that the proposed framework is theoretically and empirically effective for mitigating bias in multi-class and multi-group fairness problems in classification tasks.

5 Conclusion

In this paper, we introduce FSNS, a novel fair representation learning model to mitigate bias in the original data and maintain predictive power for downstream tasks. Unlike previous methods, our approach does not rely on techniques that can complicate the training process, including zero-sum adversarial game, variational inference, or contrastive learning. Instead, we learn fair representation by maximizing the entropy of sensitive information to improve fairness. To that end, we minimize the upper bound of the separability of latent data from different sensitive groups while maximizing the margin of data from different classes. The proposed method is simple in structure and comprises a unified objective, which empirically yields minimum performance sacrifice alongside a theoretical guarantee of upper bound on various fairness metrics. We present both theoretical analysis and empirical evidence to validate our model in multi-class classification with multiple sensitive attributes. The extensive results suggest that FSNS achieves a better trade-off between utility and fairness and results robust to data poisoning.

As most real-world fairness problems are related to sensitive social problems including rights and privacy, there exist some limitations to our FSNS model. FSNS requires sensitive attribute information to learn fair representation. However, this information may not be available due to certain laws and regulations or security reasons. Besides, there can be intentional or unintentional mistakes in the collected sensitive information. Our model would be difficult to apply to such data uncertainty. Also, the balance w.r.t. both the protected attribute and the target label can affect the outcome. For some datasets, we found that upsampling by duplicating some samples to balance the dataset improves the result in fairness. Therefore, potential future research topics to improve fairness include: 1) how to balance or generate synthetic data that has better quality than simply duplicating, 2) how to learn fair representation robustly with omitted or mislabeled data.

Acknowledgements

This work was partially supported by Purdue's Elmore ECE Emerging Frontiers Center, and NSF IIS 1955890, IIS 2146091.

References

Adel, T., Valera, I., Ghahramani, Z., and Weller, A. (2019). One-network adversarial fairness. In *AAAI*, volume 33, pages 2412–2420.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias: there's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*.

Balunovic, M., Ruoss, A., and Vechev, M. (2022). Fair normalizing flows. In *International Conference on Learning Representations*.

Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, 104:671–732.

Beutel, A., Chen, J., Zhao, Z., and Chi, E. H. (2017). Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*.

Chai, J. and Wang, X. (2022a). Fairness with adaptive weights. In *ICML*, pages 2853–2866. PMLR.

Chai, J. and Wang, X. (2022b). Self-supervised fair representation learning without demographics. *NeurIPS*, 35:27100–27113.

Chen, R. T., Li, X., Grosse, R. B., and Duvenaud, D. K. (2018). Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *SIGKDD*, pages 797–806.

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

Creager, E., Madras, D., Jacobsen, J.-H., Weis, M., Swersky, K., Pitassi, T., and Zemel, R. (2019). Flexibly fair representation learning by disentanglement. In *International conference on machine learning*, pages 1436–1445. PMLR.

Cui, Y., Chen, M., Zheng, K., Chen, L., and Zhou, X. (2023). Controllable universal fair representation learning. In *Proceedings of the ACM Web Conference 2023*, pages 949–959.

Denis, C., Elie, R., Hebiri, M., and Hu, F. (2021). Fairness guarantee in multi-class classification. *arXiv preprint arXiv:2109.13642*.

Ding, F., Hardt, M., Miller, J., and Schmidt, L. (2021). Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34.

Dressel, J. and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Sci. Adv.*, 4(1):eaao5580.

Du, M., Yang, F., Zou, N., and Hu, X. (2019). Fairness in deep learning: A computational perspective. *arXiv preprint arXiv:1908.08843*.

Dua, D. and Graff, C. (2019). UCI machine learning repository.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *KDD*, pages 259–268.

Feng, L., Shu, S., Lin, Z., Lv, F., Li, L., and An, B. (2020). Can cross entropy loss be robust to label noise? In Bessiere, C., editor, *IJCAI*, pages 2206–2212. International Joint Conferences on Artificial Intelligence Organization.

Gao, W. and Zhou, Z.-H. (2013). On the doubt about margin explanation of boosting. *AIJ*, 203:1–18.

Ghosh, A., Genuit, L., and Reagan, M. (2021). Characterizing intersectional group fairness with worst-case comparisons. In *Artificial Intelligence Diversity, Belonging, Equity, and Inclusion*, pages 22–34. PMLR.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In *NIPS*, pages 3315–3323.

Huang, L., Jiang, S., and Vishnoi, N. (2019). Coresets for clustering with fairness constraints. In *NeurIPS*, pages 7587–7598.

Jang, T., Shi, P., and Wang, X. (2022). Group-aware threshold adaptation for fair classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6988–6995.

Jang, T. and Wang, X. (2023). Difficulty-based sampling for debiased contrastive representation learning. In *CVPR*, pages 24039–24048.

Jang, T., Zheng, F., and Wang, X. (2021). Constructing a fair classifier with generated fair data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7908–7916.

Jiang, H. and Nachum, O. (2019). Identifying and correcting label bias in machine learning. *arXiv preprint arXiv:1901.04966*.

Jovanović, N., Balunovic, M., Dimitrov, D. I., and Vechev, M. (2023). Fare: Provably fair representation learning with practical certificates.

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. (2017). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. arxiv 2017. *arXiv preprint arXiv:1711.05144*.

Kim, J. S., Chen, J., and Talwalkar, A. (2020). Fact: A diagnostic for group fairness trade-offs. In *ICML*, pages 5264–5274. PMLR.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

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

Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., and Chi, E. (2020). Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33:728–740.

Lahoti, P., Gummadi, K. P., and Weikum, G. (2019). ifair: Learning individually fair data representations for algorithmic decision making. In *ICDE*, pages 1334–1345. IEEE.

Liu, E. Z., Haghgoo, B., Chen, A. S., Raghunathan, A., Koh, P. W., Sagawa, S., Liang, P., and Finn, C. (2021). Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR.

Liu, J., Li, Z., Yao, Y., Xu, F., Ma, X., Xu, M., and Tong, H. (2022). Fair representation learning: An alternative to mutual information. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1088–1097.

Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R. (2015). The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *J. Mach. Learn. Res.*, 9(Nov):2579–2605.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. (2018). Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

Oh, C., Won, H., So, J., Kim, T., Kim, Y., Choi, H., and Song, K. (2022). Learning fair representation via distributional contrastive disentanglement. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1295–1305.

Pedreshi, D., Ruggieri, S., and Turini, F. (2008). Discrimination-aware data mining. In *SIGKDD*, pages 560–568.

Rezaei, A., Fathony, R., Memarrast, O., and Ziebart, B. (2020). Fairness for robust log loss classification. In *AAAI*, volume 34, pages 5511–5518.

Roh, Y., Lee, K., Whang, S., and Suh, C. (2020). Fr-train: A mutual information-based approach to fair and robust training. In *ICML*, pages 8147–8157. PMLR.

Roy, P. C. and Boddeti, V. N. (2019). Mitigating information leakage in image representations: A maximum entropy approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2586–2594.

Sarhan, M. H., Navab, N., Eslami, A., and Albarqouni, S. (2020). Fairness by learning orthogonal disentangled representations. In *European Conference on Computer Vision*, pages 746–761. Springer.

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

Wang, A., Ramaswamy, V. V., and Russakovsky, O. (2022). Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 336–349.

Zhang, T. and Zhou, Z.-H. (2017). Multi-class optimal margin distribution machine. In *ICML*, pages 4063–4071. JMLR. org.

Zhang, T. and Zhou, Z.-H. (2019). Optimal margin distribution machine. *IEEE Transactions on Knowledge and Data Engineering*, 32(6):1143–1156.

Zhao, H., Coston, A., Adel, T., and Gordon, G. J. (2020). Conditional learning of fair representations. In *ICLR*.

Checklist

The checklist follows the references. For each question, choose your answer from the three possible options: Yes, No, Not Applicable. You are encouraged to include a justification to your answer, either by referencing the appropriate section of your paper or providing a brief inline description (1-2 sentences). Please do not modify the questions. Note that the Checklist section does not count towards the page limit. Not including the checklist in the first submission won't result in desk rejection, although in such case we will ask you to upload it during the author response period and include it in camera ready (if accepted).

In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   (b) Complete proofs of all theoretical results. [Yes]
   (c) Clear explanations of any assumptions. [Yes]

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [Yes]

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets. [Yes]
   (b) The license information of the assets, if applicable. [Not Applicable]
   (c) New assets either in the supplemental material or as a URL, if applicable. [Yes]
   (d) Information about consent from data providers/curators. [Not Applicable]
   (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable]

5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Supplementary Material for "Achieving Fairness through Separability: A Unified Framework for Fair Representation Learning"

6 Proof of Theorem 1

Here, we prove Theorem 1 in the multi-class classification setting with multiple sensitive groups, i.e., $a \in \{1, \cdots, k\}$. In the multiclass setting, the output $C_\phi : \mathbb{R}^h \to [0,1]^k$. Denote the feature distributions as $\mathcal{X}_a = P(\cdot|A=a)$.

6.1 Lemmas

We first prove the following two lemmas, one for L_1-norm loss (MAE) and the other for cross-entropy loss.

Lemma 1. *Under the L_1-norm loss, for any given sensitive feature predictor C_ϕ, we have*

$$\mathcal{L}_{fair} \geq \max_{a,b \in \mathcal{A}} \mathbb{E}_{\boldsymbol{x}_a \sim \mathcal{X}_a, \boldsymbol{x}_b \sim \mathcal{X}_b} \Big[\big\| C_\phi(H_\theta(\boldsymbol{x}_a)) - C_\phi(H_\theta(\boldsymbol{x}_b)) \big\|_1 \Big]. \tag{5}$$

Proof. For a given sensitive feature predictor C_ϕ, the fair loss function is formulated as:

$$\mathcal{L}_{fair} \quad = \quad \sum_{a \in \mathcal{A}} \mathbb{E}_{\mathbf{x} \sim \mathcal{X}_a} \Big[\mathcal{L}\big(\bar{\mathbf{a}}, C_\phi(H_\theta(\mathbf{x}))\big) \Big], \tag{6}$$

where $\bar{\mathbf{a}} = \frac{1}{k} \cdot \in [0,1]^k$ puts equal distribution on each of the predicted outcomes. Note that $\mathcal{L}_{fair}$ as defined in Eqn. (6) is the sum of multiple expectations from different sensitive groups. Specifically, $\mathbb{E}_{\mathbf{x} \sim \mathcal{X}_a}$ means the expectation conditional on event $\mathbf{x} \sim \mathcal{X}_a$, i.e., when features are generated from $\mathcal{X}_a$ with sensitive feature a. Summing over the different sensitive groups as in $\mathcal{L}_{fair}$ is to isolate the bias of the fairness loss toward any particular group a, and *equally weight* each group to prevent from weighing by the statistics i.e., $P(\mathbf{x} \sim \mathcal{X}_a)$.

When using the L_1-norm loss (MAE), we can simplify $\mathcal{L}$ within the expectation in Eqn. (6) above as $\mathcal{L}_{\text{MAE}} = \|C_\phi(H_\theta(\mathbf{x})) - \bar{\mathbf{a}}\|_1$. Then, we have

$$\begin{aligned}
\mathcal{L}_{fair} = & \sum_{a \in \mathcal{A}} \mathbb{E}_{\mathbf{x}_a \sim \mathcal{X}_a} \left[\left\| C_\phi(H_\theta(\mathbf{x}_a)) - \bar{\mathbf{a}} \right\|_1 \right] \\
\geq & \max_{a,b \in \mathcal{A}} \left(\mathbb{E}_{\mathbf{x}_a \sim \mathcal{X}_a} \left[\left\| C_\phi(H_\theta(\mathbf{x}_a)) - \bar{\mathbf{a}} \right\|_1 \right] + \mathbb{E}_{\mathbf{x}_b \sim \mathcal{X}_b} \left[\left\| C_\phi(H_\theta(\mathbf{x}_b)) - \bar{\mathbf{a}} \right\|_1 \right] \right) \\
\geq & \max_{a,b \in \mathcal{A}} \mathbb{E}_{\mathbf{x}_a \sim \mathcal{X}_a, \mathbf{x}_b \sim \mathcal{X}_b} \left[\left\| C_\phi(H_\theta(\mathbf{x}_a)) - \bar{\mathbf{a}} - C_\phi(H_\theta(\mathbf{x}_b)) + \bar{\mathbf{a}} \right\|_1 \right] \\
= & \max_{a,b \in \mathcal{A}} \mathbb{E}_{\mathbf{x}_a \sim \mathcal{X}_a, \mathbf{x}_b \sim \mathcal{X}_b} \left[\left\| C_\phi(H_\theta(\mathbf{x}_a)) - C_\phi(H_\theta(\mathbf{x}_b)) \right\|_1 \right].
\end{aligned}$$

□

Here, the first inequality comes from the fact that L1 norm is non-negative – thus, only selecting two sensitive groups would be smaller than summing over all $a \in \mathcal{A}$. The equality is achieved when there are only two sensitive groups. The second inequality follows the triangular inequality $||X|| + ||Y|| \geq ||X - Y||$.

Lemma 2. *Under the categorical cross-entropy (CE) distance, for any given sensitive feature predictor $C_\phi : \mathbb{R}^h \to [0,1]^k$, we have*

$$\mathcal{L}_{fair} \geq \frac{1}{k} \max_{a,b\in\mathcal{A}} \mathbb{E}_{\boldsymbol{x}_a\sim\mathcal{X}_a,\boldsymbol{x}_b\sim\mathcal{X}_b}\Big[\mathcal{L}_{CE}\big(e_a, C_\phi(H_\theta(\boldsymbol{x}_a))\big) + \mathcal{L}_{CE}\big(e_b, C_\phi(H_\theta(\boldsymbol{x}_b))\big)\Big], \tag{7}$$

where $e_a \in \{0,1\}^k$ is one-hot encoding of the sensitive attribute a.

Here, the CE distance $\mathcal{L}_{CE}$ is defined as

$$\mathcal{L}_{CE}\big(e_a, C_\phi(H_\theta(\mathbf{x}_a))\big) = -\sum_{i=1}^{k}[i=a]\log(C_\phi(H_\theta(\mathbf{x}_b))_i), \tag{8}$$

where $C_\phi(\cdot)_i$ denotes i-th index of the output and is indicator function.

Proof. For a given sensitive feature predictor C_ϕ, the fair loss function follows (6). For the cross-entropy loss with multiple sensitive groups $a \in \{1,\cdots,k\}$, we have

$$\begin{aligned}
\mathcal{L}_{fair} &= \sum_{a\in\mathcal{A}} \mathbb{E}_{\mathbf{x}\sim\mathcal{X}_a}\Big[\mathcal{L}_{CE}\big(\bar{\mathbf{a}}, C_\phi(H_\theta(\mathbf{x}))\big)\Big] \\
&= \frac{1}{k}\sum_{a\in\mathcal{A}} \mathbb{E}_{\mathbf{x}_a\sim\mathcal{X}_a}\Big[-\sum_{i=1}^{k}\log C_\phi(H_\theta(\mathbf{x}_a))_i\Big] \\
&\geq \frac{1}{k}\max_{a,b\in\mathcal{A}} \mathbb{E}_{\mathbf{x}_a\sim\mathcal{X}_a,\mathbf{x}_b\sim\mathcal{X}_b}\Big[-\sum_{i=1}^{k}\Big(\log C_\phi(H_\theta(\mathbf{x}_a))_i + \log C_\phi(H_\theta(\mathbf{x}_b))_i\Big)\Big] \\
&\geq \frac{1}{k}\max_{a,b\in\mathcal{A}} \mathbb{E}_{\mathbf{x}_a\sim\mathcal{X}_a,\mathbf{x}_b\sim\mathcal{X}_b}\Big[-\sum_{i=1}^{k}\Big([i=a]\log C_\phi(H_\theta(\mathbf{x}_a))_i + [i=b]\log(C_\phi(H_\theta(\mathbf{x}_b))_i)\Big)\Big] \\
&= \frac{1}{k}\max_{a,b\in\mathcal{A}} \mathbb{E}_{\mathbf{x}_a\sim\mathcal{X}_a,\mathbf{x}_b\sim\mathcal{X}_b}\big[\mathcal{L}_{CE}\big(e_a, C_\phi(H_\theta(\mathbf{x}_a))\big) + \mathcal{L}_{CE}\big(e_b, C_\phi(H_\theta(\mathbf{x}_b))\big)\big]\ .
\end{aligned} \tag{9}$$

□

Lemma 3. *Consider a given classifier $C_\phi : \mathbb{R}^h \to [0,1]^k$ for the sensitive feature prediction. Denote $\mathcal{L}_{fair} = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{X}}\big[\mathcal{L}\big(\bar{\boldsymbol{a}}, C_\phi(H_\theta(\boldsymbol{x}))\big)\big]$, i.e., the fair loss using C_ϕ and the representation learned by H_θ in FSNS. Under the ℓ_1-norm loss or the cross-entropy loss for $\mathcal{L}$, the following fairness metrics for multi-class with multiple sensitive groups, $|\mathcal{A}| \geq 2$, are bounded by $\mathcal{L}_{fair}$:*

- *Condition on the group:*

$$\mathcal{L}_{fair} \geq \max_{a,b\in\mathcal{A}} \Big\|\mathbb{E}\big[C_\phi(H_\theta(\boldsymbol{x}))|A=a\big] - \mathbb{E}\big[C_\phi(H_\theta(\boldsymbol{x}))|A=b\big]\Big\|_1,$$

- *Condition on the group and label:*

$$\mathcal{L}_{fair} \geq\ C\max_{a,b\in\mathcal{A}}\sum_{j=1}^{c}\Big(\Big\|\mathbb{E}\big[C_\phi(H_\theta(\boldsymbol{x}))|A=a,Y=j\big] - \mathbb{E}\big[C_\phi(H_\theta(\boldsymbol{x}))|A=b,Y=j\big]\Big\|_1\Big),$$

We prove Lemma 3 for the L_1-norm loss first and then for the CE loss.

Proof for Lemma 3: L_1-norm loss. First, for the demographic parity, the result follows straightforwardly from Lemma 1. For a classifier $C_\phi : \mathbb{R}^h \to [0,1]^k$ as the sensitive feature predictor, we have

$$\begin{aligned}
\mathcal{L}_{fair} &\geq \max_{a,b\in\mathcal{A}} \mathbb{E}_{\mathbf{x}_a\sim\mathcal{X}_a,\mathbf{x}_b\sim\mathcal{X}_b}\Big[\big\|C_\phi(H_\theta(\mathbf{x}_a)) - C_\phi(H_\theta(\mathbf{x}_b))\big\|_1\Big] \\
&\geq \max_{a,b\in\mathcal{A}} \Big\|\mathbb{E}_{\mathbf{x}_a\sim\mathcal{X}_a}\big[C_\phi(H_\theta(\mathbf{x}_a))\big] - \mathbb{E}_{\mathbf{x}_b\sim\mathcal{X}_b}\big[C_\phi(H_\theta(\mathbf{x}_b))\big]\Big\|_1 \\
&= \max_{a,b\in\mathcal{A}} \Big\|\mathbb{E}\big[C_\phi(H_\theta(\mathbf{x}))|A=a\big] - \mathbb{E}\big[C_\phi(H_\theta(\mathbf{x}))|A=b\big]\Big\|_1,
\end{aligned}$$

where the second inequality follows Jensen's inequality and the feature distributions are independent, and the last row is the demographic disparity among k groups. Since H_θ in FSNS model minimizes w.r.t. $\mathcal{L}_{fair}$, the demographic disparity of f is bounded by this minimized $\mathcal{L}^*_{fair}$. In other words, when $\mathcal{L}_{fair}$ is close to 0, the demographic disparity is also close to 0.

Next, for the equalized odds metric, we denote $\mathcal{Y}_j$ as the conditional distribution for $y = j$ for $j \in \{1, \cdots, c\}$ in c-class classification task. Following Lemma 1 and decomposing by the target label provides:

$$
\begin{aligned}
\mathcal{L}_{fair} &\geq \max_{a,b\in\mathcal{A}} \mathbb{E}_{\mathbf{x}_a\sim\mathcal{X}_a,\mathbf{x}_b\sim\mathcal{X}_b}\Big[\big\|C_\phi(H_\theta(\mathbf{x}_a)) - C_\phi(H_\theta(\mathbf{x}_b))\big\|_1\Big] \\
&= \max_{a,b\in\mathcal{A}} \sum_{j=1}^{c} P(y\sim\mathcal{Y}_j)\cdot \mathbb{E}_{y\sim\mathcal{Y}_j,\mathbf{x}_a\sim\mathcal{X}_a,\mathbf{x}_b\sim\mathcal{X}_b}\Big[\big\|C_\phi(H_\theta(\mathbf{x}_a)) - C_\phi(H_\theta(\mathbf{x}_b))\big\|_1\Big] \\
&\geq \max_{a,b\in\mathcal{A}} \sum_{j=1}^{c} P(y\sim\mathcal{Y}_j)\cdot \Big\|\mathbb{E}_{y\sim\mathcal{Y}_j,\mathbf{x}_a\sim\mathcal{X}_a}\big[C_\phi(H_\theta(\mathbf{x}_a))\big] - \mathbb{E}_{y\sim\mathcal{Y}_j,\mathbf{x}_b\sim\mathcal{X}_b}\big[C_\phi(H_\theta(\mathbf{x}_b))\big]\Big\|_1 \\
&= \max_{a,b\in\mathcal{A}} \sum_{j=1}^{c} P(y\sim\mathcal{Y}_j)\cdot \Big\|\mathbb{E}\big[C_\phi(H_\theta(\mathbf{x}))|A=a,Y=j\big] - \mathbb{E}\big[C_\phi(H_\theta(\mathbf{x}))|A=b,Y=j\big]\Big\|_1 \\
&\geq C \max_{a,b\in\mathcal{A}} \sum_{j=1}^{c} \Big(\Big\|\mathbb{E}\big[C_\phi(H_\theta(\mathbf{x}))|A=a,Y=j\big] - \mathbb{E}\big[C_\phi(H_\theta(\mathbf{x}))|A=b,Y=j\big]\Big\|_1\Big),
\end{aligned}
\tag{10}
$$

where $C = \min_{j\in\{1,\cdots,c\}}\big(P(y\sim\mathcal{Y}_j)\big)$. The second inequality again follows Jensen's inequality by interchanging the absolute value and (conditional) expectation. □

For the binary classification with binary sensitive group scenario, i.e., $a, y \in \{0, 1\}$ and $C_\phi(\cdot) \in [0, 1]$, we can easily show the standard equal opportunity metric from Eqn. (10) as

$$
\mathcal{L}_{fair} \geq k\cdot\Big|\mathbb{E}\big[C_\phi(H_\theta(\mathbf{x}))|A=0,Y=1\big] - \mathbb{E}\big[C_\phi(H_\theta(\mathbf{x}))|A=1,Y=1\big]\Big|,
$$

where $0 \leq k = P(y\sim\mathcal{Y}_1) \leq 1$.

Proof for Lemma 3: CE loss. The proof follows a similar procedure as for the L_1-norm loss. It is important to note the relationship between the CE loss and the L_1-norm loss. When the target label is one-hot encoded, i.e., categorical, the cross-entropy (CE) is proven to be lower bounded by mean absolute error (MAE) in Theorem 1 of Feng et al. (2020) as

$$
\mathcal{L}_{CE}\big(e_y, f(\mathbf{x})\big) \geq \frac{1}{2}\mathcal{L}_{MAE}\big(e_y, f(\mathbf{x})\big)
$$

for any predicted outcomes $f(\mathbf{x})$ and one-hot encoded target $e_y \in \{0, 1\}^c$, where $e_{yj} = 1$ if $j = y$, otherwise 0. The proof of this result applies the Taylor expansion to the $\log$ function; see details in Feng et al. Feng et al. (2020) for derivations.

Hence, applying Lemma 2 and the inequality property of CE loss Feng et al. (2020), we can get the bound of CE loss for a sensitive feature classifier $C_\phi : \mathbb{R}^h \to [0, 1]^k$ as

$$
\begin{aligned}
\mathcal{L}_{fair} &\geq \frac{1}{k}\max_{a,b\in\mathcal{A}} \mathbb{E}_{\mathbf{x}_a\sim\mathcal{X}_a,\mathbf{x}_b\sim\mathcal{X}_b}\Big[\mathcal{L}_{CE}\big(e_a, C_\phi(H_\theta(\mathbf{x}_a))\big) + \mathcal{L}_{CE}\big(e_b, C_\phi(H_\theta(\mathbf{x}_b))\big)\Big]. \\
&\geq \frac{1}{2k}\max_{a,b\in\mathcal{A}} \mathbb{E}_{\mathbf{x}_a\sim\mathcal{X}_a,\mathbf{x}_b\sim\mathcal{X}_b}\Big[\big\|e_a - C_\phi(H_\theta(\mathbf{x}_a))\big\|_1 + \big\|e_b - C_\phi(H_\theta(\mathbf{x}_b))\big\|_1\Big] \\
&\geq \frac{1}{2k}\max_{a,b\in\mathcal{A}} \mathbb{E}_{\mathbf{x}_a\sim\mathcal{X}_a,\mathbf{x}_b\sim\mathcal{X}_b}\Big[\big\|C_\phi(H_\theta(\mathbf{x}_a)) - C_\phi(H_\theta(\mathbf{x}_b))\big\|_1\Big] \\
&\geq \frac{1}{2k}\max_{a,b\in\mathcal{A}} \Big\|\mathbb{E}_{\mathbf{x}_a\sim\mathcal{X}_a}\big[C_\phi(H_\theta(\mathbf{x}_a))\big] - \mathbb{E}_{\mathbf{x}_b\sim\mathcal{X}_b}\big[C_\phi(H_\theta(\mathbf{x}_b))\big]\Big\|_1 \\
&= \frac{1}{2k}\max_{a,b\in\mathcal{A}} \Big\|\mathbb{E}\big[C_\phi(H_\theta(\mathbf{x}))|A=a\big] - \mathbb{E}\big[C_\phi(H_\theta(\mathbf{x}))|A=b\big]\Big\|_1.
\end{aligned}
$$

The third inequality can be shown by direct comparison. To see this, the L_1-norm in the second row (within the conditional expectation) for the k-dimensional space can be written as:

$$\begin{aligned}\left\|e_a - C_\phi(H_\theta(\mathbf{x}_a))\right\|_1 + \left\|e_b - C_\phi(H_\theta(\mathbf{x}_b))\right\|_1 &= \left|1 - C_\phi(H_\theta(\mathbf{x}_a))_a\right| + \left|1 - C_\phi(H_\theta(\mathbf{x}_b))_b\right| \\ &+ \sum_{i\in\mathcal{A}\setminus\{a,b\}}^{k} \Big(C_\phi(H_\theta(\mathbf{x}_a))_i + C_\phi(H_\theta(\mathbf{x}_b))_i\Big) + C_\phi(H_\theta(\mathbf{x}_a))_b + C_\phi(H_\theta(\mathbf{x}_b))_a.\end{aligned}$$

Similarly, the third row can be written as

$$\begin{aligned}\left\|C_\phi(H_\theta(\mathbf{x}_a)) - C_\phi(H_\theta(\mathbf{x}_b))\right\|_1 &= \left|C_\phi(H_\theta(\mathbf{x}_a))_a - C_\phi(H_\theta(\mathbf{x}_b))_a\right| \\ &+ \left|C_\phi(H_\theta(\mathbf{x}_a))_b - C_\phi(H_\theta(\mathbf{x}_b))_b\right| + \sum_{i\in\mathcal{A}\setminus\{a,b\}}^{k} \left|C_\phi(H_\theta(\mathbf{x}_a))_i - C_\phi(H_\theta(\mathbf{x}_b))_i\right|\end{aligned}$$

Note that C_ϕ outputs probability vectors, i.e., $\sum_i C_\phi(H_\theta(\mathbf{x}))_i = 1$ and $C_\phi(H_\theta(\mathbf{x}))_i \in [0,1]$ for each i. Then directly comparing each term of the decomposition, we can easily achieve the inequality as:

$$\left\|e_a - C_\phi(H_\theta(\mathbf{x}_a))\right\|_1 + \left\|e_b - C_\phi(H_\theta(\mathbf{x}_b))\right\|_1 \geq \left\|C_\phi(H_\theta(\mathbf{x}_a)) - C_\phi(H_\theta(\mathbf{x}_b))\right\|_1.$$

Thus, under the CE loss, when $\mathcal{L}_{fair}$ is close to 0, the demographic disparity is also close to 0. The proof for equalized odds and equal opportunity follows from the same procedure as in Eqn. (10) for the L_1-norm loss. □

6.2 Insight from the Min-Max Perspective

The results in Lemmas 1 and 2 indicate that for any given classifier C_ϕ, optimizing H_θ w.r.t. minimizing $\mathcal{L}_{fair}$ leads to minimizing the upper bound of the separability of C_ϕ (in terms of L_1-norm and CE loss, respectively). In the algorithm, we choose C_ϕ^* in $\mathcal{L}_{fair}$ when minimizing this $\mathcal{L}_{fair}$ loss. We now explain the role in using C_ϕ^* for $\mathcal{L}_{fair}$. We use the L_1-norm loss to illustrate; the insights are the same when applying the CE loss.

First, recall that C_ϕ^* is obtained as

$$C_\phi^* = \quad \arg\min_{C_\phi} \mathbb{E}_{(\mathbf{x},a)\sim\mathcal{X}\times\mathcal{A}}\big[\mathcal{L}\big(a, C_\phi(H_\theta(\mathbf{x}))\big)\big],$$

where $a \in \{1, \cdots, k\}$ is the sensitive feature given in the data. Under the L_1 norm loss, the expectation within the $\arg\min$ can be further rewritten as

$$\begin{aligned}\mathbb{E}_{(\mathbf{x},a)\sim\mathcal{X}\times\mathcal{A}}\left[\mathcal{L}\big(a, C_\phi(H_\theta(\mathbf{x}))\big)\right] &= \sum_{a\in\mathcal{A}} P(\mathbf{x}\sim\mathcal{X}_a)\, \mathbb{E}_{\mathbf{x}_a\sim\mathcal{X}_a}\Big[\|e_a - C_\phi(H_\theta(\mathbf{x}_a))\|_1\Big] \\ &\geq \rho \sum_{a\in\mathcal{A}} \mathbb{E}_{\mathbf{x}_a\sim\mathcal{X}_a}\Big[\|e_a - C_\phi(H_\theta(\mathbf{x}_a))\|_1\Big] \\ &\geq \rho\, \mathbb{E}_{\mathbf{x}_a\sim\mathcal{X}_a,\mathbf{x}_b\sim\mathcal{X}_b}\Big[\|e_a - C_\phi(H_\theta(\mathbf{x}_a))\|_1 + \|e_b - C_\phi(H_\theta(\mathbf{x}_b))\|_1\Big] \quad \text{for } a,b\in\mathcal{A} \\ &\geq \rho\, \mathbb{E}_{\mathbf{x}_a\sim\mathcal{X}_a,\mathbf{x}_b\sim\mathcal{X}_b}\Big[2 - \|C_\phi(H_\theta(\mathbf{x}_a)) - C_\phi(H_\theta(\mathbf{x}_b))\|_1\Big] \\ &\geq \rho\Big(2 - \mathbb{E}_{\mathbf{x}_a\sim\mathcal{X}_a,\mathbf{x}_b\sim\mathcal{X}_b}\Big[\|C_\phi(H_\theta(\mathbf{x}_a)) - C_\phi(H_\theta(\mathbf{x}_b))\|_1\Big]\Big)\end{aligned} \tag{11}$$

where $\rho = \min_{a\in\mathcal{A}}\big(P(\mathbf{x}\sim\mathcal{X}_a)\big)$ and e_a is one-hot vector with $e_{aj} = 1$ only if $j = a$, otherwise 0. The last row is the separability w.r.t. C_ϕ. Since C_ϕ^* minimizes $\mathcal{L}$, we have

$$\begin{aligned}\mathbb{E}_{(\mathbf{x},a)\sim\mathcal{X}\times\mathcal{A}}\Big[\mathcal{L}\big(a, C_\phi(H_\theta(\mathbf{x}))\big)\Big] &\geq \mathbb{E}_{(\mathbf{x},a)\sim\mathcal{X}\times\mathcal{A}}\Big[\mathcal{L}\left(a, C_\phi^*(H_\theta(\mathbf{x}))\right)\Big] \\ &\geq \rho\Big(2 - \mathbb{E}_{\mathbf{x}_a\sim\mathcal{X}_a,\mathbf{x}_b\sim\mathcal{X}_b}\Big[\left\|C_\phi^*(H_\theta(\mathbf{x}_a)) - C_\phi^*(H_\theta(\mathbf{x}_b))\right\|_1\Big]\Big) \\ &\geq \rho\Big(2 - \max_{a,b\in\mathcal{A}} \mathbb{E}_{\mathbf{x}_a\sim\mathcal{X}_a,\mathbf{x}_b\sim\mathcal{X}_b}\Big[\left\|C_\phi^*(H_\theta(\mathbf{x}_a)) - C_\phi^*(H_\theta(\mathbf{x}_b))\right\|_1\Big]\Big),\end{aligned}$$

In other words, for any C_ϕ

$$2 - \frac{1}{\rho}\,\mathbb{E}_{(\mathbf{x},a)\sim\mathcal{X}\times\mathcal{A}}\Big[\mathcal{L}\big(a, C_\phi(H_\theta(\mathbf{x}))\big)\Big] \leq \max_{a,b\in\mathcal{A}} \mathbb{E}_{\mathbf{x}_a\sim\mathcal{X}_a,\mathbf{x}_b\sim\mathcal{X}_b}\Big[\big\|C^*_\phi(H_\theta(\mathbf{x}_a)) - C^*_\phi(H_\theta(\mathbf{x}_b))\big\|_1\Big].$$

Thus, plugging C^*_ϕ into Eqn. (5), we get

$$\begin{aligned}\mathcal{L}_{fair} \geq\ & \max_{a,b\in\mathcal{A}} \mathbb{E}_{\mathbf{x}_a\sim\mathcal{X}_a,\mathbf{x}_b\sim\mathcal{X}_b}\Big[\big\|C^*_\phi(H_\theta(\mathbf{x}_a)) - C^*_\phi(H_\theta(\mathbf{x}_b))\big\|_1\Big] \\ \geq\ & 2 - \frac{1}{\rho}\,\mathbb{E}_{(\mathbf{x},a)\sim\mathcal{X}\times\mathcal{A}}\Big[\mathcal{L}\big(a, C_\phi(H_\theta(\mathbf{x}))\big)\Big], \qquad \forall C_\phi.\end{aligned}$$

In other words, when we minimize the fairness loss such that it is close to 0, we make sure that the maximal separability (produced when using C^*_ϕ) will be controlled to be close to 0. This min-max interaction enables us to find an efficient predictor yet maintaining fairness.

6.3 Proof for Theorem 1

The interpretation of the proposed framework from a min-max perspective in Section 6.2 provides a compelling motivation for developing theoretical analyses of fair representation learning.

Theorem 2. *Consider the optimal classifier $C^*_\phi : \mathbb{R}^h \to [0,1]^k$ for the sensitive feature prediction. Denote $\mathcal{L}_{fair}(C^*_\phi) = \mathbb{E}_{\boldsymbol{x}\sim\mathcal{X}}\Big[\mathcal{L}\big(\bar{\boldsymbol{a}}, C^*_\phi(H_\theta(\boldsymbol{x}))\big)\Big]$, i.e., the fair loss using C^*_ϕ and the representation learned by H_θ in FSNS. Under the ℓ_1-norm loss or the cross-entropy loss for $\mathcal{L}$, the following fairness metrics for multi-class classification with multiple sensitive groups are bounded by $\mathcal{L}_{fair}$ for some classifier $W : \mathbb{R}^h \to [0,1]^c$ for the target label prediction:*

- *the demographic disparity:*

$$\mathcal{L}_{fair}(C^*_\phi) \geq \Delta_{DP} = \max_{a,b\in\mathcal{A}} \Big\|\mathbb{E}\big[W(H_\theta(\boldsymbol{x}))|A=a\big] - \mathbb{E}\big[W(H_\theta(\boldsymbol{x}))|A=b\big]\Big\|_1,$$

- *the equalized odds:*

$$\mathcal{L}_{fair}(C^*_\phi) \geq \Delta_{EO} = C \max_{a,b\in\mathcal{A}} \sum_{j=1}^{c}\left(\Big\|\mathbb{E}\big[W(H_\theta(\boldsymbol{x}))|A=a, Y=j\big] - \mathbb{E}\big[W(H_\theta(\boldsymbol{x}))|A=b, Y=j\big]\Big\|_1\right),$$

where $C = \min_{j\in\{1,\cdots,c\}}\big(P(y\sim\mathcal{Y}_j)\big)$.

Proof. Recall that in Lemma 3, we showed that the following lower bound holds for $\mathcal{L}_{fair}$ for any classifier C_ϕ:

- Condition on the group:

$$\mathcal{L}_{fair}(C_\phi) \geq \max_{a,b\in\mathcal{A}} \Big\|\mathbb{E}\big[C_\phi(H_\theta(\mathbf{x}))|A=a\big] - \mathbb{E}\big[C_\phi(H_\theta(\mathbf{x}))|A=b\big]\Big\|_1,$$

- Condition on the group and label:

$$\mathcal{L}_{fair}(C_\phi) \geq\ C \max_{a,b\in\mathcal{A}} \sum_{j=1}^{c}\left(\Big\|\mathbb{E}\big[C_\phi(H_\theta(\mathbf{x}))|A=a, Y=j\big] - \mathbb{E}\big[C_\phi(H_\theta(\mathbf{x}))|A=b, Y=j\big]\Big\|_1\right),$$

which also holds for C^*_ϕ. In addition, the optimal classifier C^*_ϕ, which minimizes the prediction loss of sensitive attributes, maximizes $\mathbb{E}_{\mathbf{x}_a\sim\mathcal{X}_a,\mathbf{x}_b\sim\mathcal{X}_b}\Big[\|C_\phi(H_\theta(\mathbf{x}_a)) - C_\phi(H_\theta(\mathbf{x}_b))\|_1\Big]$ as provided in Eqn. (11). Thus, the RHS of the first inequality above can be lower bounded by

$$\mathbb{E}_{\mathbf{x}_a\sim\mathcal{X}_a,\mathbf{x}_b\sim\mathcal{X}_b}\Big[\|C^*_\phi(H_\theta(\mathbf{x}_a)) - C^*_\phi(H_\theta(\mathbf{x}_b))\|_1\Big] \geq \mathbb{E}_{\mathbf{x}_a\sim\mathcal{X}_a,\mathbf{x}_b\sim\mathcal{X}_b}\Big[\|C_\phi(H_\theta(\mathbf{x}_a)) - C_\phi(H_\theta(\mathbf{x}_b))\|_1\Big],$$

which holds for any classifier C_ϕ that takes the learned representation from H_θ as the input.

Inspired by the adversarial setting in the theoretical analysis of fair representation by Madras et al.Madras et al. (2018), we substitute the classifier in the RHS with W, i.e., the target label predictor, which gives us

$$\mathbb{E}_{\mathbf{x}_a\sim\mathcal{X}_a,\mathbf{x}_b\sim\mathcal{X}_b}\Big[\|C^*_\phi(H_\theta(\mathbf{x}_a)) - C^*_\phi(H_\theta(\mathbf{x}_b))\|_1\Big] \geq \mathbb{E}_{\mathbf{x}_a\sim\mathcal{X}_a,\mathbf{x}_b\sim\mathcal{X}_b}\Big[\|W(H_\theta(\mathbf{x}_a)) - W(H_\theta(\mathbf{x}_b))\|_1\Big]. \tag{12}$$

This substitution is intuitively supported considering two boundary cases: 1) when Y is independent of A ($Y \perp A$); 2) when Y is identical to A ($Y = A$). If the target label is independent of A, RHS would be 0. In contrast, if Y depends solely on A, then the equality holds. In either case, there exists some classifier W such that (12) holds. Additionally, it holds in general as the design of FSNS ensures that the predictor C^*_ϕ is more powerful (4-layered MLP) compared to the target predictor W (linear SVM) (C^*_ϕ corresponds to the optimal adversary in Madras et al. (2018) and can be proved to upper bound the difference in the RHS for any given learner W in the binary classification setting, since it dominates the naive adversary of choosing the opposite of W). In other words, the sensitive feature predictions made by C^*_ϕ are more confident and separable than the target prediction made by W on any given sensitive groups. Then, we can rewrite RHS, which is defined as

$$\Delta_{DP} = \mathbb{E}_{\mathbf{x}_a\sim\mathcal{X}_a,\mathbf{x}_b\sim\mathcal{X}_b}\Big[\|W(H_\theta(\mathbf{x}_a)) - W(H_\theta(\mathbf{x}_b))\|_1\Big].$$

Therefore, fair loss on optimal classifier C^*_ϕ upper bounds demographic parity violation as:

$$\mathcal{L}_{fair}(C^*_\phi) \geq \mathbb{E}_{\mathbf{x}_a\sim\mathcal{X}_a,\mathbf{x}_b\sim\mathcal{X}_b}\Big[\|C^*_\phi(H_\theta(\mathbf{x}_a)) - C^*_\phi(H_\theta(\mathbf{x}_b))\|_1\Big] \geq \Delta_{DP}. \tag{13}$$

Thus, minimizing the proposed fair loss function $\mathcal{L}_{fair}$ with respect to the encoder H_θ results in learning a fair representation by minimizing the upper bound of fairness violation. Intuitively, this can be interpreted as if separability w.r.t. sensitive attribute is minimized, we can reduce the disparity of prediction for any downstream tasks. Similar results can be easily extended to equalized odds (Δ_{EO}). It is worth noting that with the unified fair loss function $\mathcal{L}_{fair}$, we can achieve the upper bound for both demographic parity and equalized odds, while LAFTR Madras et al. (2018) requires specific loss terms for each fairness goal. □

7 Additional Experimental Results

The source code of FSNS can be found in the Repository [3].

7.1 Evaluation of the Learned Representation

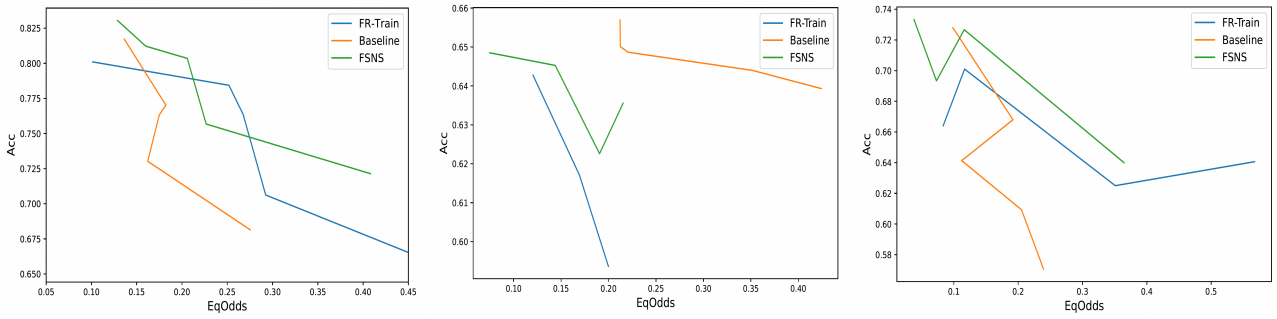
Table 3: The prediction accuracy on the target label (Label) in columns 1,2 and sensitive feature (Sens) in columns 3,4 using the original data or the latent representation from FSNS. Higher accuracy on "Label" prediction is desired for the prediction task. *Lower accuracy* on "Sens" prediction indicates *fairer* representation that contains less sensitive information.

Dataset	Label (Orig)	Label (FSNS)	Sens (Orig)	Sens (FSNS)
Adult	0.749	0.861	0.869	0.741
COMPAS	0.536	0.872	0.704	0.657
German	0.651	1.000	0.754	0.679

We further quantitatively evaluate the quality of the learned representation from FSNS. To verify our FSNS model generates a fair representation that we cannot infer the sensitive feature from it, we trained an auxiliary classifier to predict sensitive features from the distribution, which has four layers, and each layer has 64 units with ReLU activation.

Table 3 summarizes the results of classification accuracy of target label and sensitive attribute based on different distributions. As per columns 3 and 4 in Table 3, we achieve lower predictive accuracy from our representation than that from the original data. Higher accuracy w.r.t. sensitive feature indicates that the network can easily infer the sensitive feature from the distribution, which is *undesirable*. Therefore, this demonstrates that the original data contains plenty of information

[3] https://github.com/Taeuk-Jang/FSNS



(a) Adult Dataset (b) COMPAS Dataset (c) German Dataset

Figure 5: Illustration of the trend in accuracy and equalized odds as the ratio of training data poisoning increases.

about sensitive features, i.e., sensitive-relevant feature, and directly training a network can cause discrimination in the predictive performance. In contrast, FSNS significantly boosts target label accuracy (*desired*) while preventing sensitive information leakage. This also consent with the qualitative analysis of the representation in Fig. 3 in the main paper.

In Figure 6, 7, and 8, we further illustrate t-SNE visualization Maaten and Hinton (2008) of all comparing methods on three benchmark datasets. Row 1 and 2 in each figure illustrate the learned representation colored by different labels, and row 3 and 4 depict the same representation colored by different sensitive attributes. Similar to the result in the main paper, FSNS achieves the best overall separability w.r.t. target label, while achieving inseparability on the sensitive attribute.

7.2 Robustness of Learned Representation

Generalization and robustness are also critical problems in classification tasks. A recent study Roh et al. (2020) showed that training a model aimed only at robustness or fairness could suffer from severe performance degradation when it is trained on noisy training data (e.g., target label poisoning). So they propose FR-Train to holistically aims at both goals. We consider the margin distribution in representation space instead of general performance losses (e.g., cross-entropy, MSE). Thus FSNS is optimized to minimize the variance of the margin, and this leverages better generalization and robustness to the unseen data (i.e., test set).

By following the setup of Roh et al. (2020), we poison $[10\%, 20\%, 30\%, 40\%, 50\%]$ of the training data in the privileged group ($\mathbf{a} = 1$) by flipping the labels to maximize the accuracy degradation. We then evaluate the results on the clean testing set. In Fig. 5, we compare FSNS with a recent fair representation learning method incorporating robustness to data poisoning, FR-Train Roh et al. (2020), as well as the baseline on the three datasets. Although all methods experienced some loss in both accuracy and fairness as the degree of poisoning increased, we observed that the FSNS consistently outperformed the other methods at the same level of equalized odds. For example in adult dataset, note that even at the severe poisoning (40%), we maintain relatively high accuracy (73%) while others are lower than 70%. Moreover, in german dataset, the accuracy vastly decreases in baseline and fairness violation is amplified in FR-Train. Thus, the results suggest that FSNS is the most robust to severe label noise and maintains the highest accuracy. This validates the effectiveness of FSNS in learning a fair representation with a unified objective that takes into account both margin distribution and fairness considerations, resulting in improved robustness and fairness.

7.3 Time Complexity Analysis

We conducted comparison of the training time on the Adult dataset with the same hardware as shown in the Table 4. The result depicts the efficiency of FSNS, attributed from its unified loss and simple structure. Although slightly slower than CFAIR, FSNS consistently outperforms across various experiments as reported in the main paper. Such efficiency adds practical value and competitive edge of FSNS.

	FSNS	ODVAE	CFAIR	Farcon	FairDisCo	LAFTR (EOd)
Time	20min 26sec	33min 43sec	**15min 6sec**	54min 58sec	25min 17sec	25min 52sec

Table 4: Computational complexity of the comparing methods in Adult dataset.

References

Adel, T., Valera, I., Ghahramani, Z., and Weller, A. (2019). One-network adversarial fairness. In *AAAI*, volume 33, pages 2412–2420.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. (2016). Machine bias: there's software used across the country to predict future criminals. and it's biased against blacks. *ProPublica*.

Balunovic, M., Ruoss, A., and Vechev, M. (2022). Fair normalizing flows. In *International Conference on Learning Representations*.

Barocas, S. and Selbst, A. D. (2016). Big data's disparate impact. *Calif. L. Rev.*, 104:671–732.

Beutel, A., Chen, J., Zhao, Z., and Chi, E. H. (2017). Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*.

Chai, J. and Wang, X. (2022a). Fairness with adaptive weights. In *ICML*, pages 2853–2866. PMLR.

Chai, J. and Wang, X. (2022b). Self-supervised fair representation learning without demographics. *NeurIPS*, 35:27100–27113.

Chen, R. T., Li, X., Grosse, R. B., and Duvenaud, D. K. (2018). Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., and Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *SIGKDD*, pages 797–806.

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

Creager, E., Madras, D., Jacobsen, J.-H., Weis, M., Swersky, K., Pitassi, T., and Zemel, R. (2019). Flexibly fair representation learning by disentanglement. In *International conference on machine learning*, pages 1436–1445. PMLR.

Cui, Y., Chen, M., Zheng, K., Chen, L., and Zhou, X. (2023). Controllable universal fair representation learning. In *Proceedings of the ACM Web Conference 2023*, pages 949–959.

Denis, C., Elie, R., Hebiri, M., and Hu, F. (2021). Fairness guarantee in multi-class classification. *arXiv preprint arXiv:2109.13642*.

Ding, F., Hardt, M., Miller, J., and Schmidt, L. (2021). Retiring adult: New datasets for fair machine learning. *Advances in Neural Information Processing Systems*, 34.

Dressel, J. and Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Sci. Adv.*, 4(1):eaao5580.

Du, M., Yang, F., Zou, N., and Hu, X. (2019). Fairness in deep learning: A computational perspective. *arXiv preprint arXiv:1908.08843*.

Dua, D. and Graff, C. (2019). UCI machine learning repository.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *KDD*, pages 259–268.

Feng, L., Shu, S., Lin, Z., Lv, F., Li, L., and An, B. (2020). Can cross entropy loss be robust to label noise? In Bessiere, C., editor, *IJCAI*, pages 2206–2212. International Joint Conferences on Artificial Intelligence Organization.

Gao, W. and Zhou, Z.-H. (2013). On the doubt about margin explanation of boosting. *AIJ*, 203:1–18.

Ghosh, A., Genuit, L., and Reagan, M. (2021). Characterizing intersectional group fairness with worst-case comparisons. In *Artificial Intelligence Diversity, Belonging, Equity, and Inclusion*, pages 22–34. PMLR.

Hardt, M., Price, E., and Srebro, N. (2016). Equality of opportunity in supervised learning. In *NIPS*, pages 3315–3323.

Huang, L., Jiang, S., and Vishnoi, N. (2019). Coresets for clustering with fairness constraints. In *NeurIPS*, pages 7587–7598.

Jang, T., Shi, P., and Wang, X. (2022). Group-aware threshold adaptation for fair classification. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 6988–6995.

Jang, T. and Wang, X. (2023). Difficulty-based sampling for debiased contrastive representation learning. In *CVPR*, pages 24039–24048.

Jang, T., Zheng, F., and Wang, X. (2021). Constructing a fair classifier with generated fair data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 7908–7916.

Jiang, H. and Nachum, O. (2019). Identifying and correcting label bias in machine learning. *arXiv preprint arXiv:1901.04966*.

Jovanović, N., Balunovic, M., Dimitrov, D. I., and Vechev, M. (2023). Fare: Provably fair representation learning with practical certificates.

Kearns, M., Neel, S., Roth, A., and Wu, Z. S. (2017). Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. arxiv 2017. *arXiv preprint arXiv:1711.05144*.

Kim, J. S., Chen, J., and Talwalkar, A. (2020). Fact: A diagnostic for group fairness trade-offs. In *ICML*, pages 5264–5274. PMLR.

Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kingma, D. P. and Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

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

Lahoti, P., Beutel, A., Chen, J., Lee, K., Prost, F., Thain, N., Wang, X., and Chi, E. (2020). Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33:728–740.

Lahoti, P., Gummadi, K. P., and Weikum, G. (2019). ifair: Learning individually fair data representations for algorithmic decision making. In *ICDE*, pages 1334–1345. IEEE.

Liu, E. Z., Haghgoo, B., Chen, A. S., Raghunathan, A., Koh, P. W., Sagawa, S., Liang, P., and Finn, C. (2021). Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pages 6781–6792. PMLR.

Liu, J., Li, Z., Yao, Y., Xu, F., Ma, X., Xu, M., and Tong, H. (2022). Fair representation learning: An alternative to mutual information. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1088–1097.

Louizos, C., Swersky, K., Li, Y., Welling, M., and Zemel, R. (2015). The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*.

Maaten, L. v. d. and Hinton, G. (2008). Visualizing data using t-sne. *J. Mach. Learn. Res.*, 9(Nov):2579–2605.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. (2018). Learning adversarially fair and transferable representations. *arXiv preprint arXiv:1802.06309*.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., and Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

Oh, C., Won, H., So, J., Kim, T., Kim, Y., Choi, H., and Song, K. (2022). Learning fair representation via distributional contrastive disentanglement. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1295–1305.

Pedreshi, D., Ruggieri, S., and Turini, F. (2008). Discrimination-aware data mining. In *SIGKDD*, pages 560–568.

Rezaei, A., Fathony, R., Memarrast, O., and Ziebart, B. (2020). Fairness for robust log loss classification. In *AAAI*, volume 34, pages 5511–5518.

Roh, Y., Lee, K., Whang, S., and Suh, C. (2020). Fr-train: A mutual information-based approach to fair and robust training. In *ICML*, pages 8147–8157. PMLR.

Roy, P. C. and Boddeti, V. N. (2019). Mitigating information leakage in image representations: A maximum entropy approach. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2586–2594.

Sarhan, M. H., Navab, N., Eslami, A., and Albarqouni, S. (2020). Fairness by learning orthogonal disentangled representations. In *European Conference on Computer Vision*, pages 746–761. Springer.

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

Wang, A., Ramaswamy, V. V., and Russakovsky, O. (2022). Towards intersectionality in machine learning: Including more identities, handling underrepresentation, and performing evaluation. In *2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 336–349.

Zhang, T. and Zhou, Z.-H. (2017). Multi-class optimal margin distribution machine. In *ICML*, pages 4063–4071. JMLR. org.

Zhang, T. and Zhou, Z.-H. (2019). Optimal margin distribution machine. *IEEE Transactions on Knowledge and Data Engineering*, 32(6):1143–1156.

Zhao, H., Coston, A., Adel, T., and Gordon, G. J. (2020). Conditional learning of fair representations. In *ICLR*.

(a) Original Data (*label*) (b) LAFTR (*label*) (c) ODVAE (*label*)

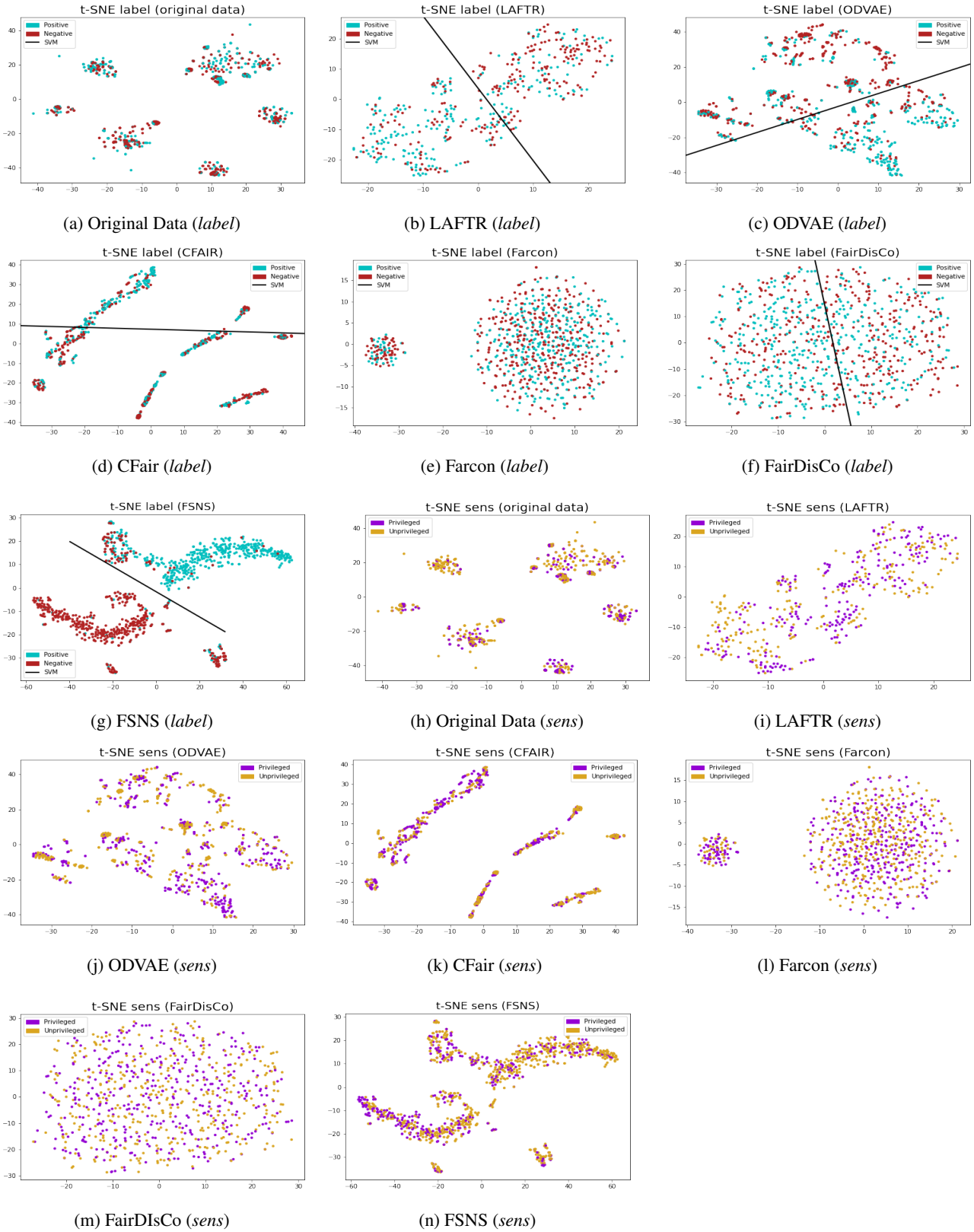
(d) CFair (*label*) (e) Farcon (*label*) (f) FairDisCo (*label*)

(g) FSNS (*label*) (h) Original Data (*sens*) (i) LAFTR (*sens*)

(j) ODVAE (*sens*) (k) CFair (*sens*) (l) Farcon (*sens*)

(m) FairDisCo (*sens*) (n) FSNS (*sens*)

Figure 6: t-SNE visualization of the learned representation from different methods in Adult dataset. Row 1 and 2 show the distribution w.r.t. the target label in classification, with a black line showing a linear SVM trained on the corresponding data representation. Row 3 and 4 depict the same distribution with the sensitive feature-based coloring scheme. Ideal representation should be separable by target label while less separable by sensitive feature.



(a) Original Data (*label*) (b) LAFTR (*label*) (c) ODVAE (*label*)

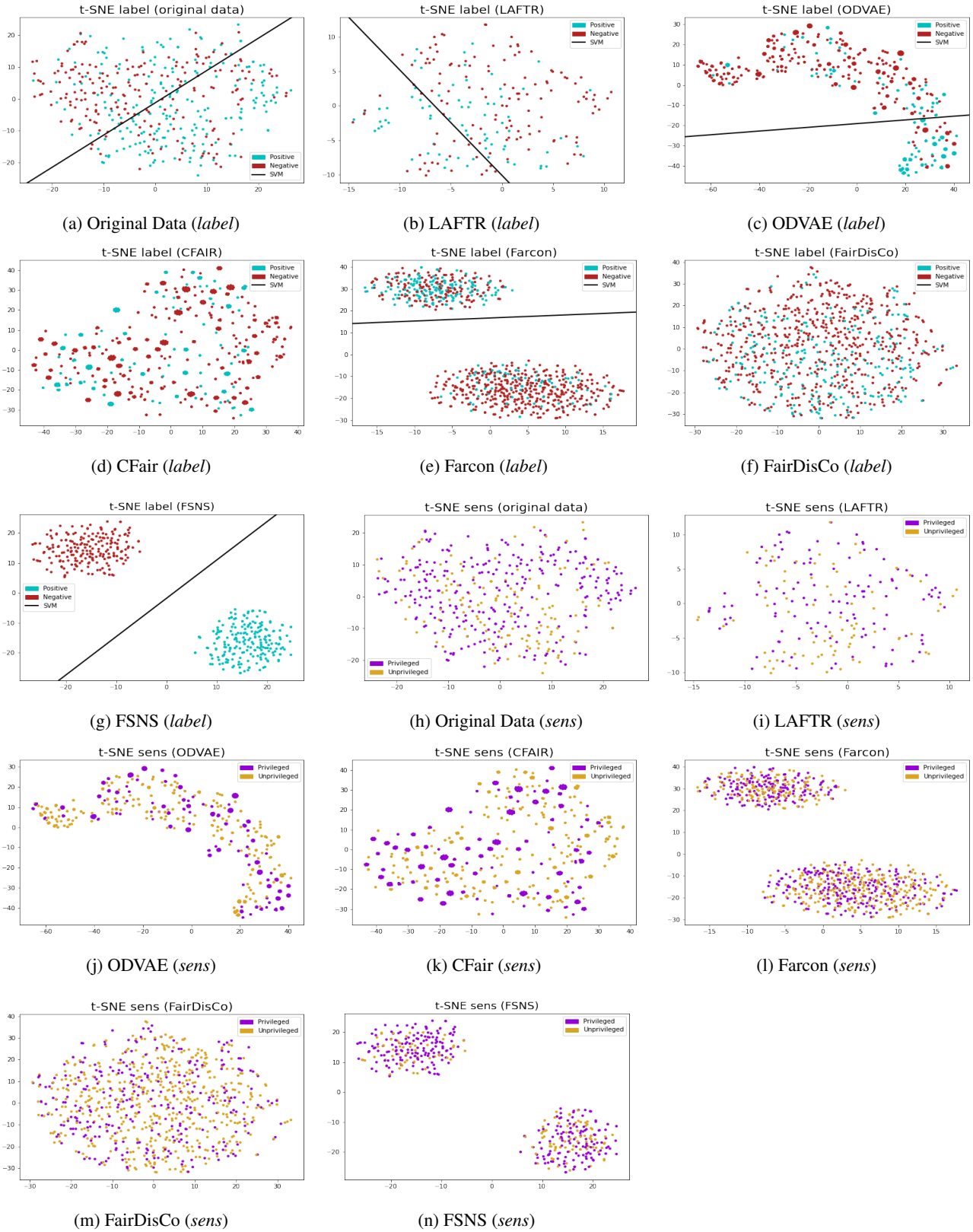
(d) CFair (*label*) (e) Farcon (*label*) (f) FairDisCo (*label*)

(g) FSNS (*label*) (h) Original Data (*sens*) (i) LAFTR (*sens*)

(j) ODVAE (*sens*) (k) CFair (*sens*) (l) Farcon (*sens*)

(m) FairDIsCo (*sens*) (n) FSNS (*sens*)

Figure 7: t-SNE visualization of the learned representation from different methods in COMPAS dataset. Row 1 and 2 show the distribution w.r.t. the target label in classification, with a black line showing a linear SVM trained on the corresponding data representation. Row 3 and 4 depict the same distribution with the sensitive feature-based coloring scheme. Ideal representation should be separable by target label while less separable by sensitive feature.



(a) Original Data (*label*) (b) LAFTR (*label*) (c) ODVAE (*label*)

(d) CFair (*label*) (e) Farcon (*label*) (f) FairDisCo (*label*)

(g) FSNS (*label*) (h) Original Data (*sens*) (i) LAFTR (*sens*)

(j) ODVAE (*sens*) (k) CFair (*sens*) (l) Farcon (*sens*)

(m) FairDisCo (*sens*) (n) FSNS (*sens*)

Figure 8: *t*-SNE visualization of the learned representation from different methods in german dataset. Row 1 and 2 show the distribution w.r.t. the target label in classification, with a black line showing a linear SVM trained on the corresponding data representation. Row 3 and 4 depict the same distribution with the sensitive feature-based coloring scheme. Ideal representation should be separable by target label while less separable by sensitive feature.