You Only Debias Once: Towards Flexible Accuracy-Fairness Trade-offs at the Inference Time

Anonymous authors
Paper under double-blind review

Abstract

Deep neural networks are prone to various bias issues, jeopardizing their applications for high-stake decision-making. Existing fairness methods typically offer a fixed accuracy-fairness trade-off at the inference time, since the weight of the well-trained model is a fixed point (fairness-optimum) in the weight space [1]. Nevertheless, more flexible accuracy-fairness trade-offs at the inference time are practically desired since: 1) stakes of the same downstream task can vary for different individuals, and 2) different regions have diverse laws or regularization for fairness. If using the previous fairness methods, we have to train multiple models, each offering a specific level of accuracy-fairness trade-off. This is often computationally expensive, time-consuming, and difficult to deploy, making it less practical for real-world applications. To address this problem, we propose *You Only Debias Once* (YODO) to achieve in-situ flexible accuracy-fairness trade-offs at the inference time, using *a single model* that trained only once [2]. Instead of pursuing one individual fixed point (fairness-optimum) in the weight space, we aim to find a "line" in the weight space that connects the accuracy-optimum and fairness-optimum points using a single model. Points (models) on this line implement varying levels of accuracy-fairness trade-offs. At the inference time, by manually selecting the specific position of the learned "line", our proposed method can achieve arbitrary accuracy-fairness trade-offs for different end-users and scenarios. Experimental results on tabular and image datasets show that YODO achieves flexible trade-offs between model accuracy and fairness, at ultra-low overheads. Our codes are anonymously available at https://anonymous.4open.science/r/yodo-BB81.

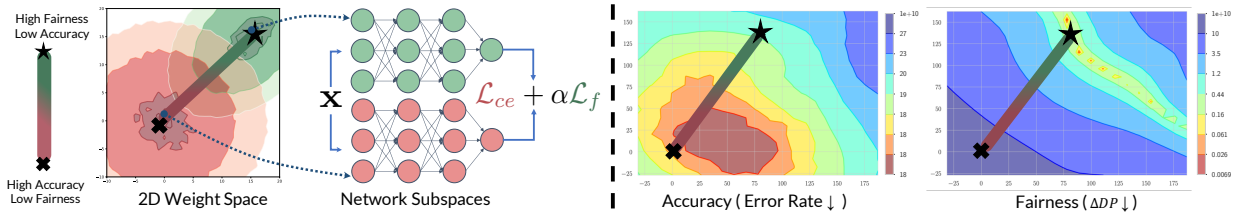
1 Introduction

Deep neural networks (DNNs) are prone to be biased with respect to sensitive attributes (Mehrabi et al., 2021; Du et al., 2020; Dwork et al., 2012; Binns, 2018; Caton & Haas, 2020; Barocas et al., 2017), which raises concerns about the application of deep learning model on high-stake decision-making, such as credit scoring (Petrasic et al., 2017), criminal justice (Berk et al., 2021), job market (Hu & Chen, 2018), healthcare (Rajkomar et al., 2018; Ahmad et al., 2020; Bjarnadóttir & Anderson, 2020; Grote & Keeling, 2022) and education Bøyum (2014); Brunori et al. (2012); Kizilcec & Lee (2022). Decisions made by these biased algorithms could have a long-lasting even life-long impact on people's lives and may affect underprivileged groups negatively. This concern about biased models has aroused wide interest from both academic and industrial researchers in analyzing and achieving fairness for DNNs.

Many studies have shown that achieving fairness in deep learning models involves a trade-off with model accuracy (Bertsimas et al., 2011; Menon & Williamson, 2018; Zhao & Gordon, 2019; Bakker et al., 2019). This has led to the development of various methods aimed at addressing fairness in deep learning models (Edwards & Storkey, 2015; Louppe et al., 2017; Chuang & Mroueh, 2020), but typically these methods have been

[1] The *weight space* of a model refers to the space of all possible values that the model's trainable parameters can take. If a neural network has m trainable parameters, then the dimension of the weight space is m. Each point, a m-dimensional vector, in the weight space corresponds to a specific model. For visualization purposes, the weight space is often reduced to a 2D space, as shown in Figure 1.

[2] This is also what we mean by *one-time training* in this paper.



(a) The Proposed Model

(b) The Loss Landscape of YODO

Figure 1: **(a)**:The overview of our proposed method. 2D Weight Space indicated the landscape of model accuracy and fairness. ✖ indicates the accuracy-optimum weight with high accuracy but low fairness, and ★ indicates the fairness-optimum weight with low accuracy but high fairness. Network Subspaces shows the different subspaces correspond with different objectives (i.e., accuracy $\mathcal{L}_{ce}$ and fairness $\mathcal{L}_f$). **(b)**: The loss landscape of the model accuracy (error rate) and fairness (ΔDP) in the same weight space of our proposed method. The weight space is reduced to two dimensions (Garipov et al., 2018). The different points indicate different objectives, ✖ indicates the accuracy–optimum endpoint in the weight space, while ★ indicates the fairness-optimum endpoint in the weight space. The dataset is `ACS-I` with gender as sensitive attribute.

designed to achieve a fixed level of accuracy-fairness trade-off. In real-world applications, however, the appropriate trade-off between accuracy and fairness may vary depending on the context and the needs of different stakeholders/regions. Thus, it is important to have flexible trade-offs at the inference time due to:

1) Downstream tasks with different stakes can have varying fairness requirements, depending on the individuals involved. According to a survey by Srivastava et al. (2019), people prioritize accuracy over fairness when the stakes are high in certain domains. For example, in healthcare (e.g., cancer prediction), accuracy should be favored over fairness, as prioritizing fairness can lead to "fatal" consequences, such as missing cancer diagnoses at higher rates (Chen et al., 2018; Srivastava et al., 2019). Since the stakes of downstream tasks differ among individuals, the expected level of fairness may also vary significantly. Thus, in high-stakes domains, the trade-off between accuracy and fairness should be flexible and controllable at the inference time.

2) Different regions have diverse laws or regulations for fairness. The use of decision-making systems is regulated by local laws and policies. However, countries may exhibit differences in the importance of fairness in various applications. For example, the labor market and employees may expect fairness to be much stronger in Germany than in North America (Gerlach et al., 2008). Therefore, developers of machine learning models should consider the varying fairness requirements when applying their algorithms in different regions.

Unfortunately, while urgently needed, flexible trade-offs between model accuracy and fairness at inference time remain underexplored. Thus, we aim to answer the following question in this paper:

Can we achieve flexible accuracy-fairness trade-offs at the inference time using a single model that trained only once?

It is an open but challenging problem to achieve flexible accuracy-fairness trade-offs. One solution is to train multiple models for different trade-offs, which limits its practical use due to the significant training time and memory overhead. Another alternative solution is the post-processing method for fairness, which may lead to suboptimal model accuracy and require sensitive attributes at the inference time (Woodworth et al., 2017).

To address the above question, we propose *You Only Debias Once* (YODO) to achieve flexible fairness-accuracy trade-offs via learning an *objective-diverse* neural network subspace that contains accuracy-optimum and fairness-optimum points in the weight space. As illustrated in Figure 1(a), we design an objective-diverse neural network subspace, which contains two endpoints (the red network and the green network in the weight space.) The two endpoints[3] are encouraged to accuracy-optimum (✖ in Figure 1) and fairness-optimum (★ in Figure 1) solutions, respectively. During the training time, we encourage the two endpoints to converge to accuracy-optimum and fairness-optimum points in the weight space, and also encourage the "line" between the two endpoints to achieve transitional solutions. Specifically, we combine the two endpoints with random weight and then optimize them with a corresponding combination of accuracy and fairness objectives. At the

[3]One endpoint represents a specific model parameter.

inference time, we enable the model to achieve arbitrary accuracy-fairness trade-offs with the trained model by manually selecting a trade-off coefficient to determine the model weight (namely, select a position of the line). Our **contributions** are highlighted as follows:

- We propose a new target, achieving in-situ flexible accuracy-fairness trade-offs at the inference time while only trained once, which adaptively meets the requirements of flexible fairness in real-world applications.
- We achieve the above in-situ flexible accuracy-fairness trade-offs by introducing an **objective-diverse** neural network subspace to achieve accuracy-fairness trade-offs. The subspace has two different endpoints in weight space, which are optimized for accuracy-optimum and fairness-optimum at different endpoints of subspaces and the "line" between the two endpoints to achieve transitional solutions. Thus it can achieve flexible trade-offs by customizing the endpoints at the inference time, with ultra-low overheads.
- Experimental results validate the effectiveness and efficiency of the proposed method on both tabular and image data. The result of experiments shows that our proposal achieves comparable performance with only trained once with various neural network architectures when compared to trained models for single use or fairness on a single attribute only. The visualization experiment provides the insight that our method gradually learns fair representation, which guarantees fair prediction.

2 Preliminaries

In this section, we introduce the notation used throughout this paper and provide an overview of the preliminaries of our work, including algorithmic fairness and neural network subspace learning.

Notations. We use $\{\mathbf{x}, y, s\}$ to denote a data instance, where $\mathbf{x} \in \mathbb{R}^d$, $y \in \{0, 1\}$, $s \in \{0, 1\}$ are feature, label, sensitive attribute, respectively. $\hat{y} \in [0, 1]$ denotes the predicted value by machine learning model $\hat{y} = f(\mathbf{x}; \theta) : \mathbb{R}^d \to [0, 1]$ with trainable parameter $\theta \in \mathbb{R}^m$ in the m-dimensional weight space, which is represented as a m-dimensional flatten vector. $\mathcal{D}$ denotes the data distribution of $(\mathbf{x}, y)$ and $\mathcal{D}_0/\mathcal{D}_1$ denotes the distribution of the data with sensitive attribute 0/1. In this work, we consider the fair binary classification ($y \in \{0, 1\}$) with binary sensitive attributes ($s \in \{0, 1\}$).

2.1 Algorithmic Fairness

For simplicity and without loss of generality, we focus on the group fairness definition of demographic parity (DP) (Dwork et al., 2012). DP is a fairness metric that aims to ensure that the proportion of positive outcomes is equal between different demographic groups in the population. To measures the DP between two demographic groups, ΔDP is defined as the absolute difference in the positive prediction rates of the machine learning model between the two demographic groups. This definition is proposed to ensure that the model treats different groups equally. Demographic parity (DP) is defined as follows:

Definition 2.1 (Demographic Parity)**.** *Demographic Parity requires the predicted values $\hat{y}$ to be independent of the sensitive attribute s, that is, $P(\hat{y}|s = 0) = P(\hat{y}|s = 1)$.*

One relaxed metric to measure demographic parity has also been proposed by Edwards & Storkey (2015) and widely used by Agarwal et al. (2018); Wei et al. (2019;?); Taskesen et al. (2020); Madras et al. (2018); Chuang & Mroueh (2020). The relaxed metric ΔDP of demographic parity is defined as follows:

$$\Delta\mathrm{DP}(f) = \left|\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_0} f(\mathbf{x}) - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_1} f(\mathbf{x})\right|, \tag{1}$$

One practical and effective way to achieve demographic parity is to formulate a penalized optimization with ΔDP as gap regularization. The overall objective for a fixed level of accuracy-fairness trade-offs is as follows:

$$\mathcal{L} = \mathcal{L}_{ce}(f(\mathbf{x}; \theta), y) + A \cdot \mathcal{L}_f(f(\mathbf{x}; \theta), y) = \mathcal{L}_{ce} + A \cdot \mathcal{L}_f, \tag{2}$$

where $\mathcal{L}_{ce}$ is the objective function (e.g., cross-entropy) of the downstream task, and A is a fixed hyperparameter to balance the model accuracy and fairness. In the following, we set A to 1 for simplicity.

In addition to DP, we also consider the fairness metric of Equality of Opportunity (EO) in our experiments, as described in Section 5.6. By examining both DP and EO, we can gain a more comprehensive understanding of the performance of our approach and the trade-offs between fairness and accuracy.

2.2 Neural Network Subspaces

The optimization of the neural network is to find a minimum (often a local minimum) in a high-dimensional weight space. Wortsman et al. (2021) and Benton et al. (2021) proposed a method to learn a functionally diverse neural network subspace, which is parameterized by two sets of network weights, ω_1 and ω_2. At training time, a network is sampled from the line defined by this pair of weights, $\theta = (1-\alpha)\omega_1 + \alpha\omega_2$ for $\alpha \in [0, 1]$, and the two sets of weights ω_1 and ω_2 are optimized by one back-propagation simultaneously, making the learning process more efficient. By parameterizing the neural network subspace using two sets of weights, this method allows for greater control over the optimization of the neural network. This is because the user can specify the values of α to control the degree of mixing between ω_1 and ω_2. Different from this method that the learning objective of both ω_1 and ω_2 are model accuracy for the downstream task, our proposed method lets the ω_1 and ω_2 learn different target, accuracy and fairness, respectively. The method also allows for the exploration of the trade-offs between accuracy and fairness, as the two sets of weights can be optimized to achieve different levels of accuracy and fairness.

3 Methodology

In this section, we introduce our method YODO, which learns an objective-diverse neural network subspace targeting flexible trade-offs between accuracy and fairness at the inference time while training the model once.

3.1 You Only Debias Once

Our goal is to find an objective-diverse subspace (the "line") in weight space comprised of accuracy-optimum (✖ in Figure 1) and fairness-optimum (★ in Figure 1) neural networks, as illustrated in Figure 1. Specifically, we first parameterize two sets of trainable weights $\omega_1 \in \mathbb{R}^n$ and $\omega_2 \in \mathbb{R}^n$ for one neural network, and then optimize weights ω_1, ω_2 to achieve that $f(\mathbf{x}; \omega_1)$ is high for model accuracy and $f(\mathbf{x}; \omega_2)$ is high for model fairness. Thus $f(\mathbf{x}; (1-\alpha)\omega_1 + \alpha\omega_2))$ achieves α-controlled accuracy-fairness trade-offs. In other words, we aim to learn a "line" between ω_1 and ω_2 in the weight space to achieve flexible accuracy-fairness trade-offs at the inference time.

In the training process, we aim to optimize ω_1, ω_2 with objective functions targeting accuracy and fairness objectives, respectively. Thus two endpoints ω_1 and ω_2 correspond with objective function for downstream task $\mathcal{L}_{ce}$ and the objective function for fairness. $\mathcal{L}_{ce} + \mathcal{L}_f$ [4]. Thus the learned ω_1, ω_2 will be accuracy-optimum (✖ in Figure 1) and fairness-optimum (★ in Figure 1) points in the weight space. Based on the linear combination of two endpoints, we try to minimize the loss as

$$\mathcal{L} = (1-\alpha) \underbrace{\mathcal{L}_{ce}}_{\omega_1(✖)} + \alpha \underbrace{(\mathcal{L}_{ce} + \mathcal{L}_f)}_{\omega_2(★)} = \mathcal{L}_{ce} + \alpha\mathcal{L}_f, \tag{3}$$

where $\theta = (1-\alpha)\omega_1 + \alpha\omega_2$ for $\alpha \in [0, 1]$. $\mathcal{L}_{ce}$ is instantiated as the cross-entropy loss for downstream tasks (i.e., binary classification task), and $\mathcal{L}_f$ is instantiated as demographic parity difference $\Delta\text{DP}[f(x; \theta)]$.

Since we seek to optimize the different levels of fairness constraints, for each $\alpha \in [0, 1]$, we propose to minimize $\mathbb{E}_{(\mathbf{x},y)\sim\text{D}}[\mathcal{L}_{ce}(f(\mathbf{x}; \theta), y) + \alpha\mathcal{L}_f(f(\mathbf{x}; \theta), y)]$, thus our training objective is to minimize

$$\begin{aligned} \mathbb{E}_{\alpha\sim\text{U}}\big[\mathbb{E}_{(\mathbf{x},y)\sim\text{D}}[\mathcal{L}_{ce}(f(\mathbf{x}; \theta), y) + \alpha\mathcal{L}_f(f(\mathbf{x}; \theta), y)]\big], \\ s.t. \quad \theta = (1-\alpha)\omega_1 + \alpha\omega_2, \end{aligned} \tag{4}$$

where U denotes the uniform distribution between 0 and 1 and the trade-off hyperparameter α follows uniform distribution, i.e., $\alpha \sim \text{U}$.

Why Does YODO Achieve Flexible Accuracy-fairness Trade-offs in Terms of Objective Function? For each α in U, the objective $\mathbb{E}_{(\mathbf{x},y)\sim\text{D}}[\mathcal{L}_{ce}(f(\mathbf{x}; \theta), y) + \alpha\mathcal{L}_f(f(\mathbf{x}; \theta), y)]$ will be optimized to the minima, leading to different accuracy-fairness trade-offs with different α. In other words, our model can be regarded as

[4] For simplicity, here we set the hyperparameters A in Equation (2) to 1.

training infinite models with different fairness constraints in the training phase. Such a mechanism guarantees that YODO could achieve flexible trade-offs with one-pass training at the inference time. In the following, we analyze our model from the model optimization perspective.

3.2 Optimization Leading to Objective-Diverse Subspace

In this section, we discuss the model optimization for YODO. In each training batch, we randomly sample $\alpha \sim \mathrm{U}_{[0,1]}$, and then we use $\theta = (1-\alpha)\omega_1 + \alpha\omega_2$ as the model parameters. We calculate the gradients for ω_1 and ω_2 with respect to objective function (Equation (4)) as follows:

$$\frac{\partial \mathcal{L}}{\partial \omega_i} = \frac{\partial \mathcal{L}}{\partial \omega_i} = \frac{\partial \mathcal{L}}{\partial \theta}\frac{\partial \theta}{\partial \omega_i}. \tag{5}$$

Consider that $\theta = (1-\alpha)\omega_1 + \alpha\omega_2$, the gradients for the endpoints ω_1 and ω_2 are

$$\begin{aligned} \frac{\partial(\mathcal{L}_{ce} + \alpha\mathcal{L}_f)}{\partial \omega_1} &= \alpha\frac{\partial(\mathcal{L}_{ce} + \alpha\mathcal{L}_f)}{\partial \theta}, \\ \frac{\partial(\mathcal{L}_{ce} + \alpha\mathcal{L}_f)}{\partial \omega_2} &= (1-\alpha)\frac{\partial(\mathcal{L}_{ce} + \alpha\mathcal{L}_f)}{\partial \theta}, \end{aligned} \tag{6}$$

From Equations (5) and (6), we can see that gradients for ω_1 and ω_2 are related to the $\mathcal{L}_{ce}$ and $\mathcal{L}_f$ with different values of the scale coefficient (i.e., α and $1-\alpha$). The optimization of the two endpoints ω_1 and ω_2 only depends on the gradient $\frac{\partial(\mathcal{L}_{ce}+\alpha\mathcal{L}_f)}{\partial\theta}$, which are the most time-consuming operations in the back-propagation during the model training. This indicates that our method does not require any extra cost in the back-propagation phase since we only compute $\frac{\partial(\mathcal{L}_{ce}+\alpha\mathcal{L}_f)}{\partial\theta}$ once.

Why Does YODO Achieve Flexible Accuracy-fairness Trade-offs in Terms of Gradient? Equation (6) indicates the optimization of the weights ω_1 and ω_2, as well as the objective function, are both controlled by the hyperparameter α. In the following analysis, we will examine the optimization of these weights with different values of α.

- When $\alpha = 0$, the gradients of ω_2 in Equation (6) is calculated by the loss $\frac{\partial(\mathcal{L}_{ce}+\alpha\mathcal{L}_f)}{\partial\omega_2} = (1-0)\frac{\partial(\mathcal{L}_{ce}+0*\mathcal{L}_f)}{\partial\theta} = \frac{\partial(\mathcal{L}_{ce})}{\partial\theta}$ and the gradients of ω_1 is 0, indicating weight ω_2 will be only optimized by the accuracy objective.
- When $\alpha = 1$, the gradients in Equation (6) is calculated by the loss $\frac{\partial(\mathcal{L}_{ce}+\alpha\mathcal{L}_f)}{\partial\omega_1} = 1 * \frac{\partial(\mathcal{L}_{ce}+1*\mathcal{L}_f)}{\partial\theta} = \frac{\partial(\mathcal{L}_{ce}+\mathcal{L}_f)}{\partial\theta}$ and the gradients of ω_1 is 0, indicating weight ω_1 will be only optimized by the fairness objective.
- When $0 < \alpha < 1$, the weights ω_1 and ω_2 will be only optimized by the linear combination of accuracy objective and fairness objective.

From the analysis of gradient, we can conclude that *the optimization tends to encourage the two endpoints ω_1 and ω_2 to accuracy-optimum and fairness-optimum solutions in the weight space.* As illustrated in Figure 1(b), we plot the landscape of the error rate (accuracy) and the ΔDP (fairness). The landscapes are depicted with a trained YODO. ✖ and ★ indicate the two endpoints, which are encouraged to learn accuracy and fairness objectives, respectively. The landscape shows that our method learns an objective-diverse neural network subspace and optimizes the endpoints to accuracy-optimum and fairness-optimum solutions.

Algorithm 1 YODO

Require: Training set $\mathcal{S}$, balance hyperparameters β
1: Initialize each ω_i independently.
2: **for** batch in $\mathcal{S}$ **do**
3: Randomly sample an α
4: Calculate interpolated weight $\omega =\leftarrow (1-\alpha)\omega_1 + \alpha\omega_2$
5: Calculate loss $\mathcal{L} = \mathcal{L}ce + \alpha\mathcal{L}f + \beta\mathcal{L}_{reg}$
6: Back-propagate $\mathcal{L}$ to update ω_1 and ω_2 using Equation (6).
7: **end for**

Model Optimization. To promote the diversity of the neural network subspaces being learned, we follow the approach proposed in (Wortsman et al., 2021) and aim to learn two distinct endpoints, i.e., the

accuracy-optimum and the fairness-optimum endpoint in the weight space. Specifically, we constrain the cosine similarity of the pair of endpoints (ω_1 and ω_2) to be zero. To achieve this, we augment the original learning objective (Equation (4)) with a cosine similarity regularization term $\mathcal{L}_{reg}$, which is defined as $\mathcal{L}_{reg} = \frac{\langle \omega_1, \omega_2 \rangle^2}{|\omega_1|_2^2 |\omega_2|_2^2}$. which encourages the two endpoints to be as dissimilar as possible, effectively diversifying the learned subspaces. Thus the final objective function will be $\mathcal{L} = \mathcal{L}ce + \alpha \mathcal{L}f + \beta \mathcal{L}_{reg}$. The algorithm of YODO is presented in Algorithm 1.

4 Discussion

In this section, we provide further analysis of our proposed method, including the model complexity, and model deployment. Additionally, we discuss the relationship between our work and related research.

Model Complexity. we analyze the time and space complexity of YODO. Compared to the memory utilization of the models with fixed accuracy-fairness trade-off, our method utilizes twofold memory usage. However, it can achieve in-situ flexible trade-offs at inference time. In practice, if the downstream task needs N kinds of different accuracy-fairness trade-offs, we need N different trained models with different fixed fairness constraints, which also need N times training. In this sense, YODO only requires storing two sets of neural network weights and one-time training. For time complexity, the extra computational cost stems from the linear combination of the two sets of neural network weights at the inference time, which are time-saving and negligible. At the training time, the computational cost is from gradient computation of ω_1/ω_2 (Equation (6)).

Model Deployment. We discuss the advantages of the proposed method in the model deployment phase. YODO can be recalibrated for accuracy-fairness configuration before/after deployment for a fixed/flexible accuracy-fairness trade-offs. For fixed accuracy-fairness trade-offs, we can set predefined α for YODO without retraining the model. For flexible accuracy-fairness trade-offs, we can define an API for users to input their α to meet their own requirements. Another advantage is that our method can output the varying predictive values for individuals (experiment in Figure 7), which can provide a reference for the decision makers to examine the realization degree of fairness.

Difference and Relation to Previous Works. Wortsman et al. (2021); Benton et al. (2021) proposed to learn the neural network subspaces, which aims to find a flat subspace for high accuracy. The goal of these two papers is to learn a *functionally diverse* neural network subspace in which the model accuracy is all high at each endpoint. Our method creatively proposes to achieve flexible fairness at the inference time via learning a *objective-diverse* neural network subspace containing the optimum solutions for multiple objectives (e.g., accuracy and fairness). Nunez et al. (2023) proposed a similar method to achieve accuracy-efficiency trade-offs, which is similar to our proposed method. However, this paper proposed to train a degenerate subspace with a single point in weight space. Instead, we target to train an objective-diverse neural network subspace by training the linear combined weights with corresponding linear combinations of objectives (i.e., accuracy and fairness).

5 Experiments

We empirically evaluate the performance of our proposed method in this section. First, we evaluate the effectiveness of the proposed method on both tabular and image datasets in Section 5.1. Then, we investigate the in-situ flexible accuracy-fairness trade-offs of YODO in Section 5.2. Moreover, we investigate the visualization of hidden representation and the distribution of the prediction values in Sections 5.3 and 5.4, which gains insights into why our method can achieve flexible accuracy-fairness trade-offs. We also provide a case study on `CelebA` dataset in Section 5.5. The major **Obs**ervations of the experiments are highlighted in **boldface**.

Experimental Setting. In our experiments, we evaluate the performance of our proposed method on multiple datasets, including both tabular and image data. For tabular dataset, we adopt `UCI Adult` (Dua & Graff, 2017), `KDD Census` (Dua & Graff, 2017), `ACS-I` (ncome) (Ding et al., 2021), `ACS-E` (mployment) (Ding

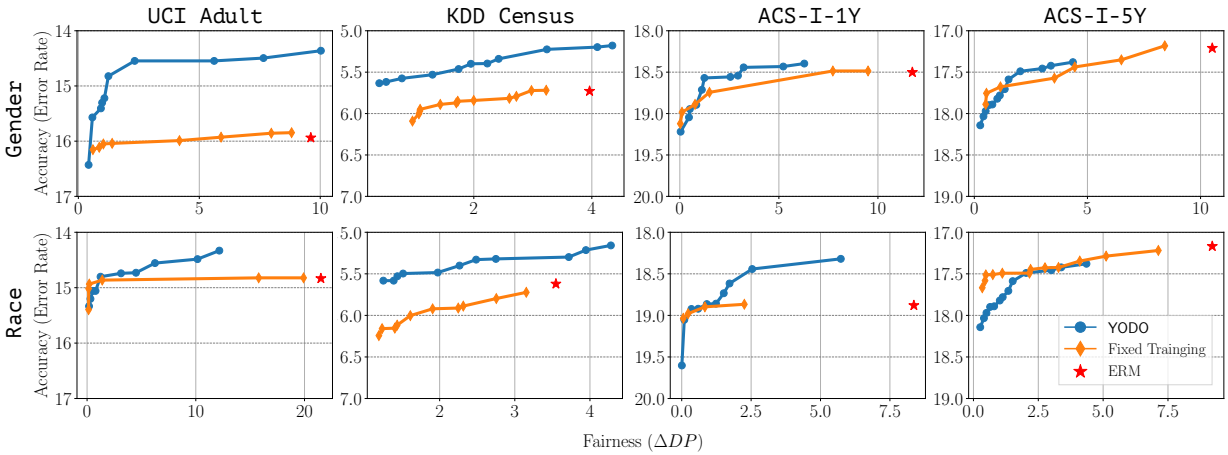


Figure 2: The Pareto frontier of the model accuracy and fairness. The first row is the fairness performance with respect to gender sensitive attribute, while the second row is race sensitive attribute. The model performance metric is Error Rate (lower is better), and the fairness metric is ΔDP (lower is better, Equation (1)).

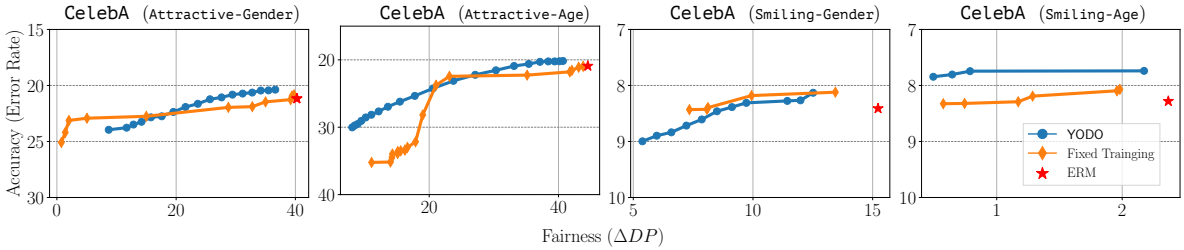


Figure 3: The Pareto frontier of the model performance and fairness on the `CelebA` dataset. The downstream task is to predict whether a person is Attractive (Smiling) or not. The sensitive attribute we considered is gender and age. The x-axis represents the difference in demographic parity (ΔDP) between the sensitive groups, while the y-axis represents the error rate of the model. Our proposed one-time training model achieves a comparable trade-off between performance and fairness, as indicated by the Pareto frontier when compared to a set of fixed-trained models.

et al., 2021), which are widely used benchmark datasets in the fairness and machine learning community. For image data analysis, we use the `CelebA` dataset, which is a large-scale face attributes dataset with more than 200K celebrity images. This dataset is commonly used in the field of computer vision and has been used in various studies related to fairness in machine learning. The additional experimental setting is presented in Appendix A.

Baselines. We included two baseline methods, namely **Fixed Training** and **ERM**, in our experiments to validate the effectiveness of our proposed method. Fixed Training, as an important baseline, trains multiple models for different fixed accuracy-fairness trade-offs. Training the multiple fixed models requires expensive time and memory costs. For each point in Figures 2 and 10, we trained one fixed model using the objective function shown in Equation (2) with different values of A, which is set to $[0, 1]$ with an interval of 0.05. For our experimental setting, we have to train 21 individual models from scratch for each trade-off hyperparameter. Another baseline is Empirical Risk Minimization (ERM), which is to minimize the empirical risk of downstream task (marked as $\star$ in the result).

5.1 Will YODO Achieve Flexible Trade-offs only with Training Once?

In this section, we validate the effectiveness of our proposed `YODO` on real-world datasets, including tabular data and image data. We present the results in Figures 2, 3 and 10 and use Pareto frontier (Emmerich & Deutz, 2018) to evaluate our proposed method and the baseline. Pareto frontier is widely used to evaluate the accuracy-fairness trade-offs by Kim et al. (2020); Liu & Vicente (2020); Wei & Niethammer (2020) and

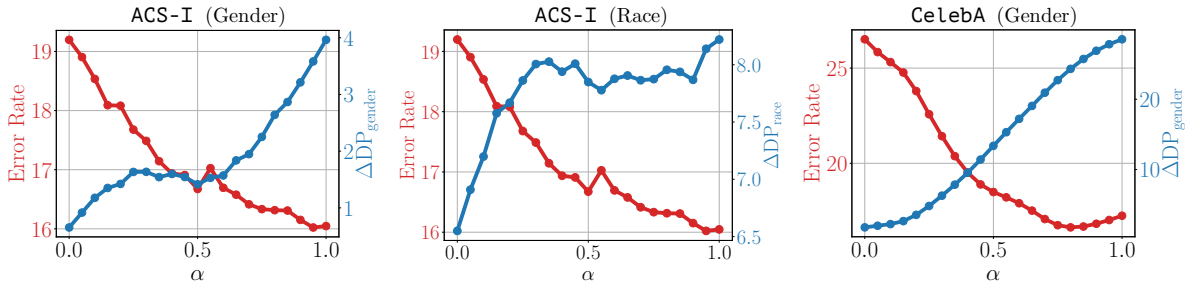


Figure 4: The accuracy-fairness trade-offs at the inference time with respect to α for three different datasets: `ACS-I` dataset with gender as the sensitive attribute (**Left**), `ACS-I` dataset with race as the sensitive attribute (**Middle**), `CelebA` dataset with gender as the sensitive attribute (**Right**). We observed that the fine-grained accuracy-fairness trade-offs could be achieved by selecting different values of α, providing more nuanced accuracy-fairness trade-offs. Note that the results are obtained at inference time with a single trained model.

characterizes a model's achievable accuracy for given fairness conditions. Pareto frontier characterizes a model's achievable accuracy for given fairness conditions, as a measurement to understand the trade-offs between model accuracy and fairness. The Pareto frontier indicates the set of models that achieve the best trade-off between performance and fairness. The detail of the experiments can be found at Appendix B.3. The results on the tabular dataset are presented in Figures 2 and 10. And the results on the image dataset are presented in Figure 3. From these figures, we make the following major observations:

Obs.1: Even though YODO only needs to be trained once, it performs similarly to baseline (Fixed Training), or even better in some cases. After conducting experiments on real-world datasets, we compared the Pareto frontier of YODO with that of the **Fixed Training** baseline. We found that Pareto frontier of YODO coincides with that of **Fixed Training** in most figures, indicating it can achieve comparable or even better performance than baseline. The result makes YODO readily usable for real-world applications requiring flexible fairness. It is worth highlighting that our proposed method only needs one-time training to obtain the Pareto frontier at inference time, making it computationally efficient.

Obs.2: YODO achieves better fairness performance than baseline (Fixed Training) in some cases. From the results of `UCI Adult`, `KDD Census`, and `ACS-I` dataset, we observed that our method achieves better fairness performance in some cases. Specifically, the Pareto frontier represented in blue in the figures can reach a lower ΔDP value than the baseline methods. This observation suggests that our method is highly effective in achieving flexible fairness and can be readily used in real-world applications where fairness is crucial for the downstream task.

Obs.3: On some datasets (e.g., `UCI Adult`, `KDD Census`, `CelebA`), YODO even outperforms the ERM baseline. Upon comparing our results on the `UCI Adult` and `KDD Census` datasets, we observed that the Pareto frontier of our method covers the point of ERM. This finding suggests that our approach outperforms ERM in both model accuracy and fairness performance. It also demonstrates that our proposed objective-diverse neural network subspace can achieve flexible accuracy-fairness trade-offs while potentially improving model accuracy. This advantage is likely attributed to the powerful capacity of the proposed objective-diverse subspace.

Obs.4: YODO achieves a more smooth Pareto frontier than baseline (Fixed Training). On most datasets, especially on `CelebA` image data, our proposed method demonstrates a smoother Pareto frontier compared to the baselines. For example, on `CelebA` Attractive-Age and Attractive-Gender in Figure 3, the Pareto frontier is notably smoother than baselines. A smooth Pareto frontier implies that our proposal is able to achieve more fine-grained accuracy-fairness trade-offs, which can provide a more dependable prediction for high-stakes decision-making scenarios.

5.2 Can the Trade-Offs Be Flexible by Controlling Parameter α?

To explore the effect of hyperparameter α on the accuracy-fairness trade-offs, we conducted experiments on the `ACS-I` and `CelebA` datasets with different values of α. α is a hyperparameter that controls the

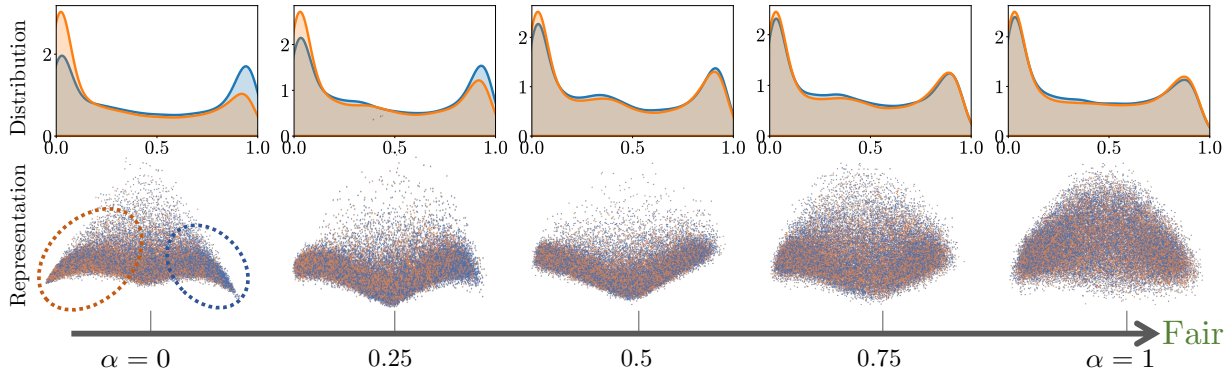


Figure 5: The changing distribution of the prediction values and the representations as α increases. Blue indicates male and orange indicates female. **Top**: The distribution is estimated with kernel density estimation (Terrell & Scott, 1992). The distributions are more polarized between males and females with $\alpha = 0$, but become more similar with $\alpha = 1$. As α increases, the distributions for male and female groups become more similar, indicating achieving demographic parity. **Bottom**: The visualization of the representation with the values of α are from 0 to 1. The figure at the far left ($\alpha = 0$) shows that the representations for males (⬭) and females (⬭) are distinctly separate, indicating that the representations contain more sensitive information. The figure on the far right ($\alpha = 1$) shows the representations of different groups are mixed together, indicating that the representation contains little sensitive information.

trade-off between model accuracy and fairness in machine learning models. A higher value of α emphasizes more on fairness, while a lower value prioritizes accuracy. By adjusting the value of α, we can make the accuracy-fairness trade-off flexible and controllable. We evaluate the model accuracy and fairness performance with the different values of α at the inference time. The results are presented in Figure 4 and we observed:

Obs.5: The accuracy-fairness trade-offs can be controlled by α at the inference time. Our experiments on both tabular and `CelebA` datasets showed that as we increase the value of hyperparameter α, the error rate (lower is better) gradually increases while ΔDP gradually decreases. The solution at $\alpha = 0$ prioritizes accuracy-optimum (lowest error rate), while the solution at $\alpha = 1$ prioritizes fairness-optimum (lowest ΔDP). These results demonstrate that our proposed method can achieve flexible accuracy-fairness trade-offs, with distinct linear correlations between the trade-off and the value of α. Based on these findings, our approach can provide more reliable and controllable predictions by customizing α at inference time.

5.3 How Does the Distribution of the Predictive Values Vary with Changing α ?

In this section, we examine the distribution of predictive values to investigate the varying trade-offs of our method. We specifically plot the distribution of predictive values for different groups (male and female) to verify the flexible accuracy-fairness trade-offs provided by our approach. The experiments are conducted on `ACS-I` dataset and the results are presented in Figure 5. We have the following observation:

Obs.6: The distribution of predictive values for different groups becomes increasingly similar as the value of α increases, indicating that our model becomes more fair as the values of α increase. Additionally, we found that the distributions of the predictive values of different groups follow the same distribution, showing the predictive values are independent of sensitive attributes. The varying distribution of predictive values provides valuable insights into why our model can achieve flexible accuracy-fairness trade-offs from a distributional perspective..

5.4 How Does the Representations Vary with Changing α ?

In this section, we visualize the hidden representation of different groups to provide further analysis and insight into the proposed method, `YODO`. The visualization of representations is widely used to inspect the fairness issue of DNNs from the manifold perspective (Du et al., 2021). With the different combination with different αs, we visualize the hidden representation in Figure 5 and we observed that:

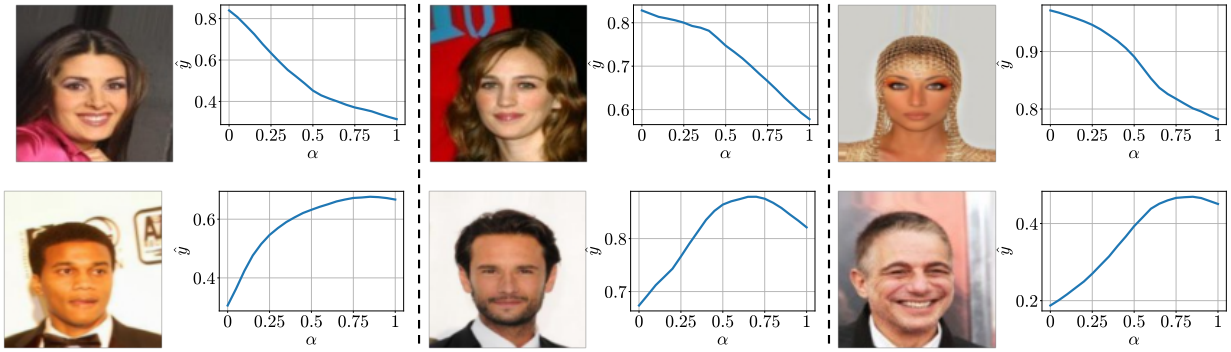


Figure 7: Case study on `CelebA` dataset. The sensitive attribute is gender. The downstream task is to predict whether a person is attractive or not. The results show that the YODO can provide the instance-level prediction change for practitioners to examine the fairness performance of our proposed method. By providing instance-level prediction changes, YODO enabled practitioners to better understand the model's fairness performance and make informed decisions.

Obs.7: The disparity of the representation of different groups becomes smaller and smaller with the increasing α. The figure on the far left in Figure 5 shows that the representation of male and female groups are distinctly separate, indicating that the representations contain more sensitive information. The figure on the far right shows the representations of different groups mixed together, indicating that the representations contain little sensitive information. The sensitive information embedded in the representation indicates that the two endpoint biases correspond to model accuracy and fairness, respectively. This phenomenon provides insight into YODO, that it can learn a fair representation to guarantee a fair prediction.

5.5 How Do the Prediction Value Vary in Instance Level? A Case Study on Image Data

In this section, we experiment on the `CelebA` dataset to investigate how the predicted values changes and the varying prediction with the various α in the instance level. The sensitive attribute we consider is Gender. The downstream task is to predict whether a person is attractive or not. The female group has more positive samples than the Male group, leading the biased prediction. We first present the mean of the prediction with respect to varying α for each group (i.e., male and female) in Figure 6. We also present some cases to investigate the effect of the change of the values of α in Figure 7. Based on our analysis of Figures 6 and 7, we make the following observations:

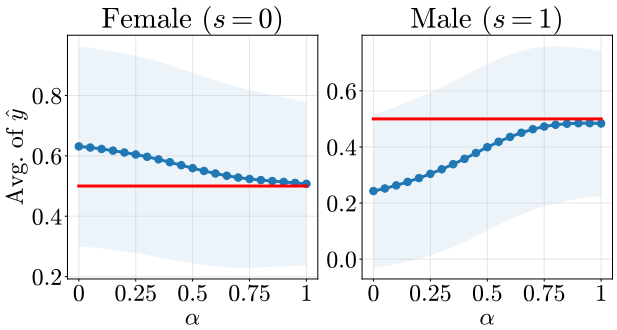


Figure 6: The mean of the prediction values for females and males with respect to varying α. The red line indicates the predictive value $\hat{y} = 0.5$. The blue line is the mean of predictive values of each group (i.e., male and female). The results illustrate that the mean prediction values for males and females become more similar as α increases., indicating a fairer prediction.

Obs.8: The mean of the predictive values of different groups (e.g., female, male) approaches 0.5 with the increasing α. Analysis of Figure 7 shows that, overall, YODO tends to decrease the predictive values for the Female group and increase the predictive values for the Male group, resulting in lower ΔDP values and indicating fairer predictions overall. Our observations demonstrate that our proposed method can mitigate the unfairness caused by the dataset's gender imbalance and ensure that individuals from different groups are treated equitably.

Obs.9: YODO can provide instance-level explanation for group fairness. In `CelebA` dataset, the Female group has more positive (attractive) samples than the other one, leading to a higher predictive value for the male group. Figure 7 shows that YODO tends to lower the predictive values of the Female group while higher the predictive values of the Male group. Such a tendency would result in fairer predictive

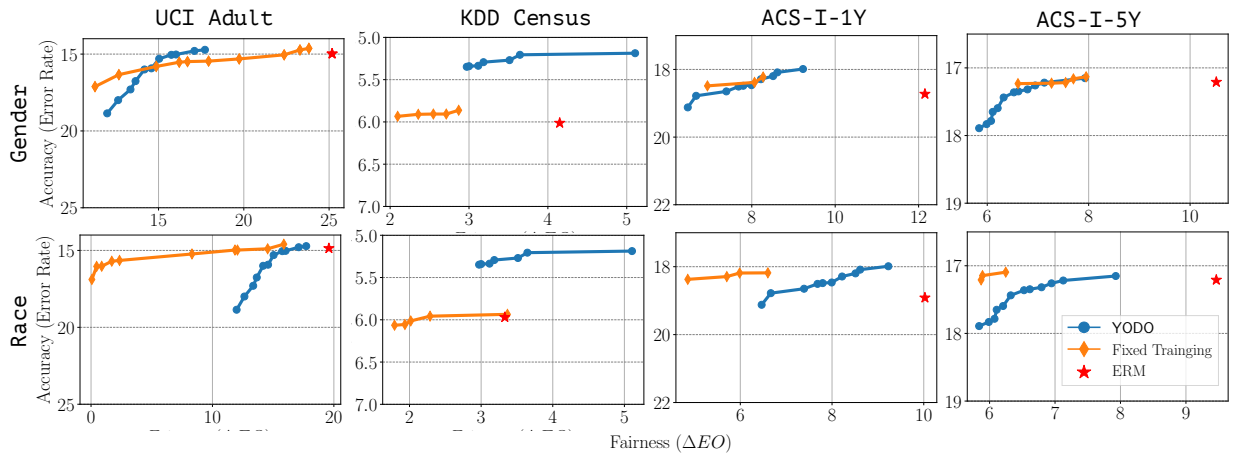


Figure 8: The Pareto frontier of the model accuracy and fairness. The first row is the fairness performance with respect to gender sensitive attribute, while the second row is race sensitive attribute. The model performance metric is Error Rate (lower is better), and the fairness metric is ΔEO (lower is better, Equation (1).

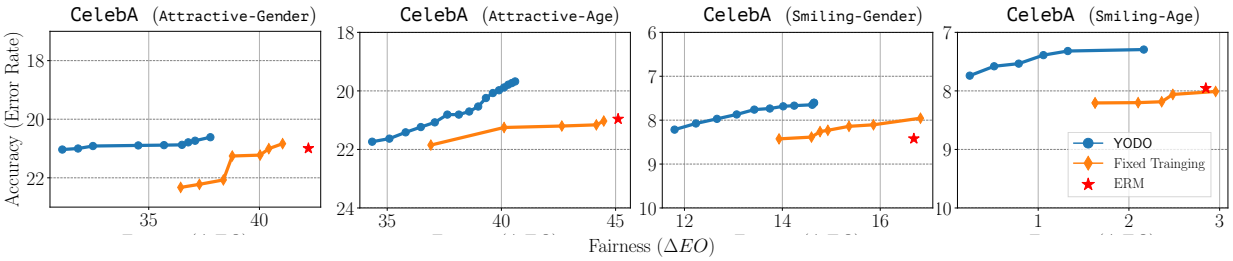


Figure 9: The Pareto frontier of the model performance and fairness on `CelebA` dataset. The downstream task is to predict whether a person is Attractive (Smiling) or not. The sensitive attribute is gender and age.

results. This observation lays a solid foundation for the real-world application of our proposed method. Our proposed method YODO offers an instance-level explanation for individuals. This allows for a more transparent and interpretable machine learning model, which can help build trust with users. For example, Figure 7 demonstrates how YODO can lower the predictive values for the Female group and increase the predictive values for the Male group, resulting in a fairer outcome for individuals belonging to each group. Our method can provide individualized explanations and trustworthy model for end-users.

5.6 How Does YODO Perform on Anther Group Fairness, Equality of Opportunity?

In this section, we experiment on the fairness metric Equality of Opportunity (EO) (Hardt et al., 2016). EO measures whether a classifier provides equal opportunities to individuals from different groups. By satisfying this criterion, we can ensure that protected groups are not disproportionately negatively impacted by the model's predictions.

Definition 5.1 (Equality of opportunity)**.** *A classifier satisfies this definition if both protected and unprotected groups have an equal probability of a subject in a positive class having a negative predictive value.*

For example, this implies that the probability of an applicant with an actual good credit score being incorrectly assigned a bad predicted credit score should be the same for both male and female applicants. The formal definition of EO is as the following:

$$\Delta\text{EO}(f) = \left|\mathbb{E}_{\mathbf{x}\sim\mathcal{D}_0,y=1} f(\mathbf{x}) - \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_1,y=1} f(\mathbf{x})\right|, \tag{7}$$

We conduct the experiment using ΔEO as a fairness constraint and metric and present the results on tabular and image data. The results are presented in Figures 8, 9 and 11. We have the following observations:

Obs.10: YODO performs similarly to baseline (Fixed Training) in terms of Equality of Opportunity, or even better in some cases. This observation indicates it can achieve comparable or even better performance than the baseline. And this experiment also makes our method readily usable for real-world applications, which require flexible fairness. The result shows the effectiveness of our proposal on other fairness metrics, demonstrating its overall effectiveness in promoting fairness.

6 Related Works

In this section, we discuss three areas of research that are related to our work, namely group fairness, accuracy-fairness trade-offs, and neural network subspace learning.

Fairness in Machine Learning Recently, algorithmic fairness (Pogodin et al., 2023; Cruz et al., 2023; Guo et al., 2023; Li et al., 2023; Pham et al., 2023; Jung et al., 2023; Han et al., 2023; Chung et al., 2023; Roh et al., 2021; Guldogan et al., 2023; Roh et al., 2023; Schrouff et al., 2022a;b; Vogel et al., 2020; Coston et al., 2020; Maheshwari et al., 2022) is required legally or morally in machine learning systems and various fairness definitions in machine learning systems have been proposed to meet the requirement of fairness expectation. The fairness can typically be classified into *individual fairness* or *group fairness*, which can be achieved via pre/in/post-processing. In this paper, we focus on the in-processing group fairness, which measures the statistical parity between subgroups defined by the sensitive attributes, such as gender or race (Zemel et al., 2013; Louizos et al., 2015; Hardt et al., 2016; Chuang & Mroueh, 2020; Zafar et al., 2017; Madras et al., 2018; Joseph et al., 2016). Nevertheless, these constraints are trained and satisfied during training, the model may expect different accuracy-fairness trade-offs at the inference time. In contrast, Our proposed method, YODO, aims to address this issue by enabling flexible accuracy-fairness trade-offs at the inference time.

Accuracy-Fairness Trade-offs Many existing works investigate the trade-offs between the model accuracy and fairness(Cooper et al., 2021; Kim et al., 2020; Bertsimas et al., 2011; Menon & Williamson, 2018; Liu & Vicente, 2020; Kleinberg, 2018; Haas, 2019; Rothblum & Yona, 2021; Wei & Niethammer, 2020; Barlas et al., 2021; Dutta et al., 2020; Maity et al., 2020; Chouldechova & Roth, 2018; Zhang et al., 2022; Vogel et al., 2021). Maity et al. (2020) discusses the existence of the accuracy-fairness trade-offs. (Kim et al., 2020; Dressel & Farid, 2018; Zliobaite, 2015; Dutta et al., 2020; Blum & Stangl, 2019; Wick et al., 2019) shows that fairness and model accuracy conflict with each another, and achieved fairness often comes with a necessary cost in loss of model accuracy. Cooper et al. (2021); Kim et al. (2020) re-examines the trade-offs and concludes that unexamined assumptions may result in emergent unfairness. These studies emphasize the need to develop models that balance accuracy and fairness while also considering the unique characteristics of the data and the potential for emergent biases.

Neural Network Subspaces Learning The idea of learning a neural network subspace is concurrently proposed by Wortsman et al. (2021) and Benton et al. (2021). Multiple sets of network weights are treated as the corners of a simplex, and an optimization procedure updates these corners to find a region in weight space in which points inside the simplex correspond to accurate networks. Garipov et al. (2018) learning a connection between two independently trained neural networks, considering curves and piecewise linear functions with fixed endpoints. Wortsman et al. (2021) and Benton et al. (2021) concurrently proposed to learn a functionally diverse subspace. Our proposed method YODO differs from these methods by specifying each subspace to the different learning objectives and allowing flexible fairness levels at the inference time.

7 Conclusion

In this paper, we proposed YODO, a novel method that achieves accuracy-fairness trade-offs at inference time to meet the diverse requirements of fairness in real-world applications. Our approach is the first to achieve flexible trade-offs between model accuracy and fairness through the use of an objective-diverse neural network subspace. Our extensive experiments demonstrate the effectiveness and practical value of the proposed approach, and we offer a detailed analysis of the underlying mechanisms by examining the distributions of predictive values and hidden representations. By enabling in-situ flexibility, our approach can provide more nuanced control over the trade-offs between accuracy and fairness, thereby advancing the state-of-the-art in the field of algorithmic fairness.

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pp. 60–69, 2018.

Muhammad Aurangzeb Ahmad, Arpit Patel, Carly Eckert, Vikas Kumar, and Ankur Teredesai. Fairness in machine learning for healthcare. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3529–3530, 2020.

Michiel A Bakker, Alejandro Noriega-Campero, Duy Patrick Tu, Prasanna Sattigeri, Kush R Varshney, and AS Pentland. On fairness in budget-constrained decision making. In *KDD Workshop of Explainable Artificial Intelligence*, 2019.

Pinar Barlas, Kyriakos Kyriakou, Olivia Guest, Styliani Kleanthous, and Jahna Otterbacher. To" see" is to stereotype: Image tagging algorithms, gender recognition, and the accuracy-fairness trade-off. *Proceedings of the ACM on Human-Computer Interaction*, 4(CSCW3):1–31, 2021.

Solon Barocas, Moritz Hardt, and Arvind Narayanan. Fairness in machine learning. *Nips tutorial*, 1:2017, 2017.

Gregory Benton, Wesley Maddox, Sanae Lotfi, and Andrew Gordon Gordon Wilson. Loss surface simplexes for mode connecting volumes and fast ensembling. In *International Conference on Machine Learning*, pp. 769–779. PMLR, 2021.

Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, 50(1):3–44, 2021.

Dimitris Bertsimas, Vivek F Farias, and Nikolaos Trichakis. The price of fairness. *Operations research*, 59(1): 17–31, 2011.

Reuben Binns. Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency*, pp. 149–159. PMLR, 2018.

Margrét Vilborg Bjarnadóttir and David Anderson. Machine learning in healthcare: Fairness, issues, and challenges. In *Pushing the Boundaries: Frontiers in Impactful OR/OM Research*, pp. 64–83. INFORMS, 2020.

Avrim Blum and Kevin Stangl. Recovering from biased data: Can fairness constraints improve accuracy? *arXiv preprint arXiv:1912.01094*, 2019.

Steinar Bøyum. Fairness in education–a normative analysis of oecd policy documents. *Journal of Education Policy*, 29(6):856–870, 2014.

Paolo Brunori, Vito Peragine, and Laura Serlenga. Fairness in education: The italian university before and after the reform. *Economics of Education Review*, 31(5):764–777, 2012.

Simon Caton and Christian Haas. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*, 2020.

Irene Y Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 3543–3554, 2018.

Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *CoRR*, abs/1810.08810, 2018. URL http://arxiv.org/abs/1810.08810.

Ching-Yao Chuang and Youssef Mroueh. Fair mixup: Fairness via interpolation. In *International Conference on Learning Representations*, 2020.

Hyung Won Chung, Xavier Garcia, Adam Roberts, Yi Tay, Orhan Firat, Sharan Narang, and Noah Constant. Unimax: Fairer and more effective language sampling for large-scale multilingual pretraining. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=kXwdL1cWOAi.

A Feder Cooper, Ellen Abrams, and NA NA. Emergent unfairness in algorithmic fairness-accuracy trade-off research. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 46–54, 2021.

Amanda Coston, Alan Mishler, Edward H. Kennedy, and Alexandra Chouldechova. Counterfactual risk assessments, evaluation, and fairness. In *FAT**, pp. 582–593, 2020. URL https://doi.org/10.1145/3351095.3372851.

André Cruz, Catarina G Belém, João Bravo, Pedro Saleiro, and Pedro Bizarro. FairGBM: Gradient boosting with fairness constraints. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=x-mXzBgCX3a.

Frances Ding, Moritz Hardt, John Miller, and Ludwig Schmidt. Retiring adult: New datasets for fair machine learning. *NeurIPS*, 2021.

Julia Dressel and Hany Farid. The accuracy, fairness, and limits of predicting recidivism. *Science advances*, 4 (1):eaao5580, 2018.

Mengnan Du, Fan Yang, Na Zou, and Xia Hu. Fairness in deep learning: A computational perspective. *IEEE Intelligent Systems*, 2020.

Mengnan Du, Subhabrata Mukherjee, Guanchu Wang, Ruixiang Tang, Ahmed Awadallah, and Xia Hu. Fairness via representation neutralization. *Advances in Neural Information Processing Systems*, 34, 2021.

Dheeru Dua and Casey Graff. UCI machine learning repository, 2017. URL http://archive.ics.uci.edu/ml.

Sanghamitra Dutta, Dennis Wei, Hazar Yueksel, Pin-Yu Chen, Sijia Liu, and Kush Varshney. Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing. In *International Conference on Machine Learning*, pp. 2803–2813. PMLR, 2020.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226, 2012.

Harrison Edwards and Amos Storkey. Censoring representations with an adversary. *arXiv preprint arXiv:1511.05897*, 2015.

Michael TM Emmerich and André H Deutz. A tutorial on multiobjective optimization: fundamentals and evolutionary methods. *Natural computing*, 17(3):585–609, 2018.

Timur Garipov, Pavel Izmailov, Dmitrii Podoprikhin, Dmitry Vetrov, and Andrew Gordon Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pp. 8803–8812, 2018.

Knut Gerlach, David Levine, Gesine Stephan, and Olaf Struck. Fairness and the employment contract: North american regions versus germany. *Cambridge Journal of Economics*, 32(3):421–439, 2008.

Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pp. 249–256. JMLR Workshop and Conference Proceedings, 2010.

Thomas Grote and Geoff Keeling. Enabling fairness in healthcare through machine learning. *Ethics and Information Technology*, 24(3):39, 2022.

Ozgur Guldogan, Yuchen Zeng, Jy yong Sohn, Ramtin Pedarsani, and Kangwook Lee. Equal improvability: A new fairness notion considering the long-term impact. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=dhYUMMy0_Eg.

Dongliang Guo, Zhixuan Chu, and Sheng Li. Fair attribute completion on graph with missing attributes. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=9vcXCMp9VEp.

Christian Haas. The price of fairness-a framework to explore trade-offs in algorithmic fairness. In *40th International Conference on Information Systems, ICIS 2019*. Association for Information Systems, 2019.

Xudong Han, Timothy Baldwin, and Trevor Cohn. Everybody needs good neighbours: An unsupervised locality-based method for bias mitigation. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=pOnhudsvzR.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29:3315–3323, 2016.

Lily Hu and Yiling Chen. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference*, pp. 1389–1398, 2018.

Matthew Joseph, Michael Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in learning: classic and contextual bandits. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 325–333, 2016.

Sangwon Jung, Taeeon Park, Sanghyuk Chun, and Taesup Moon. Exact group fairness regularization via classwise robust optimization. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Q-WfHzmiG9m.

Joon Sik Kim, Jiahao Chen, and Ameet Talwalkar. Fact: A diagnostic for group fairness trade-offs. In *International Conference on Machine Learning*, pp. 5264–5274. PMLR, 2020.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

René F Kizilcec and Hansol Lee. Algorithmic fairness in education. In *The ethics of artificial intelligence in education*, pp. 174–202. Routledge, 2022.

Jon Kleinberg. Inherent trade-offs in algorithmic fairness. In *Abstracts of the 2018 ACM International Conference on Measurement and Modeling of Computer Systems*, pp. 40–40, 2018.

Puheng Li, James Zou, and Linjun Zhang. FaiREE: fair classification with finite-sample and distribution-free guarantee. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=shzu8d6_YAR.

Suyun Liu and Luis Nunes Vicente. Accuracy and fairness trade-offs in machine learning: A stochastic multi-objective approach. *arXiv preprint arXiv:2008.01132*, 2020.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Large-scale celebfaces attributes (celeba) dataset. 2018.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

Gilles Louppe, Michael Kagan, and Kyle Cranmer. Learning to pivot with adversarial networks. *NeurIPS*, 30, 2017.

David Madras, Elliot Creager, Toniann Pitassi, and Richard Zemel. Learning adversarially fair and transferable representations. *International Conference on Machine Learning*, 2018.

Gaurav Maheshwari, Pascal Denis, Mikaela Keller, and Aurélien Bellet. Fair nlp models with differentially private text encoders. *CoRR*, abs/2205.06135, 2022. URL https://doi.org/10.48550/arXiv.2205.06135.

Subha Maity, Debarghya Mukherjee, Mikhail Yurochkin, and Yuekai Sun. There is no trade-off: enforcing fairness can improve accuracy. *arXiv preprint arXiv:2011.03173*, 2020.

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

Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, pp. 107–118. PMLR, 2018.

Elvis Nunez, Maxwell Horton, Anish Prabhu, Anurag Ranjan, Ali Farhadi, and Mohammad Rastegari. Lcs: Learning compressible subspaces for efficient, adaptive, real-time network compression at inference time. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3818–3827, 2023.

Kevin Petrasic, Benjamin Saul, James Greig, Matthew Bornfreund, and Katherine Lamberth. Algorithms and bias: What lenders need to know. *White & Case*, 2017.

Thai-Hoang Pham, Xueru Zhang, and Ping Zhang. Fairness and accuracy under domain generalization. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=jBEXnEMdNOL.

Roman Pogodin, Namrata Deka, Yazhe Li, Danica J. Sutherland, Victor Veitch, and Arthur Gretton. Efficient conditionally invariant representation learning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=dJruFeSRym1.

Alvin Rajkomar, Michaela Hardt, Michael D Howell, Greg Corrado, and Marshall H Chin. Ensuring fairness in machine learning to advance health equity. *Annals of internal medicine*, 169(12):866–872, 2018.

Yuji Roh, Kangwook Lee, Steven Whang, and Changho Suh. Sample selection for fair and robust training. *Advances in Neural Information Processing Systems*, 34:815–827, 2021.

Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. Improving fair training under correlation shifts. *arXiv preprint arXiv:2302.02323*, 2023.

Guy N Rothblum and Gal Yona. Consider the alternatives: Navigating fairness-accuracy tradeoffs via disqualification. *arXiv preprint arXiv:2110.00813*, 2021.

Jessica Schrouff, Natalie Harris, Oluwasanmi Koyejo, Ibrahim Alabdulmohsin, Eva Schnider, Krista Opsahl-Ong, Alex Brown, Subhrajit Roy, Diana Mincu, Christina Chen, Awa Dieng, Yuan Liu, Vivek Natarajan, Alan Karthikesalingam, Katherine A. Heller, Silvia Chiappa, and Alexander D'Amour. Maintaining fairness across distribution shift: do we have viable solutions for real-world applications? *CoRR*, abs/2202.01034, 2022a. URL https://arxiv.org/abs/2202.01034.

Jessica Schrouff, Natalie Harris, Oluwasanmi O Koyejo, Ibrahim Alabdulmohsin, Eva Schnider, Krista Opsahl-Ong, Alexander Brown, Subhrajit Roy, Diana Mincu, Chrsitina Chen, Awa Dieng, Yuan Liu, Vivek Natarajan, Alan Karthikesalingam, Katherine A Heller, Silvia Chiappa, and Alexander D'Amour. Diagnosing failures of fairness transfer across distribution shift in real-world medical settings. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022b. URL https://openreview.net/forum?id=K-A4tDJ6HHf.

Megha Srivastava, Hoda Heidari, and Andreas Krause. Mathematical notions vs. human perception of fairness: A descriptive approach to fairness for machine learning. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2459–2468, 2019.

Bahar Taskesen, Viet Anh Nguyen, Daniel Kuhn, and Jose Blanchet. A distributionally robust approach to fair classification. *arXiv preprint arXiv:2007.09530*, 2020.

George R Terrell and David W Scott. Variable kernel density estimation. *The Annals of Statistics*, pp. 1236–1265, 1992.

Robin Vogel, Aurélien Bellet, and Stéphan Clémençon. Learning fair scoring functions: Fairness definitions, algorithms and generalization bounds for bipartite ranking. *CoRR*, abs/2002.08159, 2020. URL https://arxiv.org/abs/2002.08159.

Robin Vogel, Aurélien Bellet, and Stéphan Clémençon. Learning fair scoring functions: Bipartite ranking under roc-based fairness constraints. In *AISTATS*, pp. 784–792, 2021. URL http://proceedings.mlr.press/v130/vogel21a.html.

Dennis Wei, Karthikeyan Natesan Ramamurthy, and Flavio du Pin Calmon. Optimized score transformation for fair classification. *arXiv preprint arXiv:1906.00066*, 2019.

Susan Wei and Marc Niethammer. The fairness-accuracy pareto front. *arXiv preprint arXiv:2008.10797*, 2020.

Michael Wick, Jean-Baptiste Tristan, et al. Unlocking fairness: a trade-off revisited. *Advances in neural information processing systems*, 32, 2019.

Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *Conference on Learning Theory*, pp. 1920–1953, 2017.

Mitchell Wortsman, Maxwell Horton, Carlos Guestrin, Ali Farhadi, and Mohammad Rastegari. Learning neural network subspaces. *arXiv preprint arXiv:2102.10472*, 2021.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pp. 962–970. PMLR, 2017.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pp. 325–333. PMLR, 2013.

Guanhua Zhang, Yihua Zhang, Yang Zhang, Wenqi Fan, Qing Li, Sijia Liu, and Shiyu Chang. Fairness reprogramming. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=Nay_rOB-dZv.

Han Zhao and Geoff Gordon. Inherent tradeoffs in learning fair representations. *Advances in neural information processing systems*, 32:15675–15685, 2019.

Indre Zliobaite. On the relation between accuracy and fairness in binary classification. *arXiv preprint arXiv:1505.05723*, 2015.

In this appendix, we provide the two additional experiments in Appendix A, including the experiment on `ACS-E` dataset and the distribution of the predictive values on `CelebA` dataset. We also provide a more detailed experimental setting in Appendix B.

A Additional Experiments

A.1 Experiments on `ACS-E` Dataset

In this appendix, we present the results of complementary experiments conducted on the `ACS-E` dataset with gender as the sensitive attribute. The results for two-group fairness metrics, ΔDP and ΔEO, are presented in Figures 10 and 11, respectively. These results are complementary to the ones presented in Figures 2 and 8.

The experimental results on the `ACS-E` dataset with gender as the sensitive attribute show that our proposed method achieves comparable accuracy-fairness trade-offs to the baselines for both group fairness metrics ΔDP and ΔEO. This finding is consistent with the results of the experiments on other datasets with different sensitive attributes presented in Figures 2 and 8.

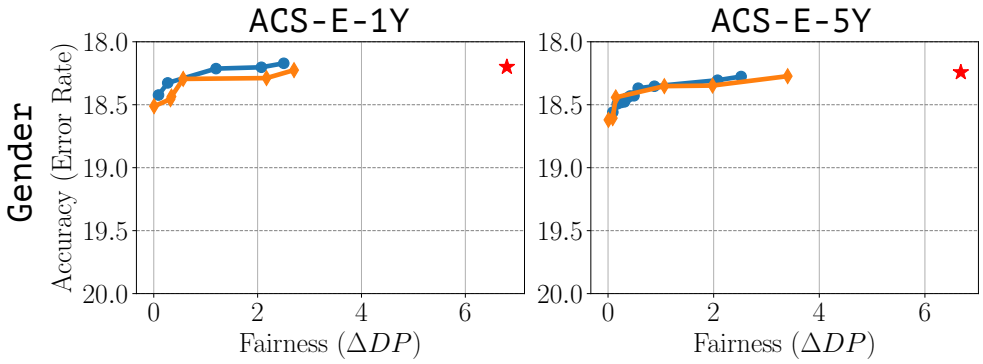


Figure 10: The Pareto frontier of the model performance and fairness on `ACS-E` dataset spanning 1 year or 5 years on ΔDP metric. The sensitive attribute is gender.

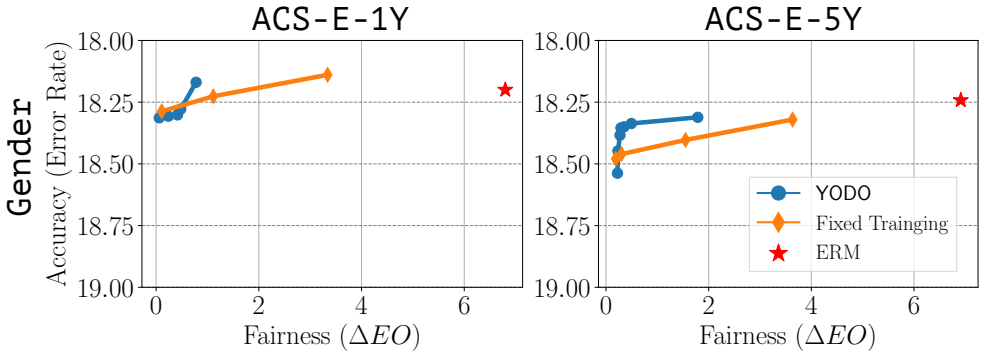


Figure 11: The Pareto frontier of the model performance and fairness on `ACS-E` data spanning 1 year or 5 years on ΔEO metric. The sensitive attribute is gender.

A.2 The Distribution of Predictive Values on Image Dataset

In this appendix, we provide the distribution of the predictive values on the image dataset, `CelebA`. The distribution of the predictive values of different groups are more and more similar with the increase of the values of α, indicating that our model is more and more fair with the increase of the values of α. This result shows that our proposed method encourages the distribution to follow the same distribution, and further guarantees fair prediction.

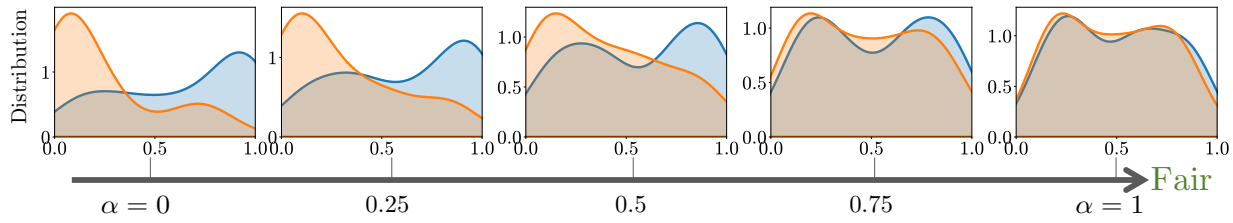


Figure 12: The distribution of the prediction values with different α on `CelebA` dataset. The distribution of the predictive values of different groups (i.e., Male, Female) becomes more and more similar with the increasing α. The distributions are more polarised between Male and Female with $\lambda = 0$ while the distribution is nearly the same with $\lambda = 1$.

Table 1: The summary of the datasets used in our experiment.

Dataset	Data Type	Task	Sensitive Attributes	#Instances
`UCI Adult`	Tabular	Income	Gender, Race	48,842
`KDD Census`	Tabular	Income	Gender, Race	299,285
`ACS-Income` (1 year)	Tabular	Income	Gender, Race	265,171
`ACS-Income` (5 years)	Tabular	Income	Gender, Race	1,315,945
`ACS-Emploement` (1 year)	Tabular	Employment	Gender	136,965
`ACS-Emploement` (5 years)	Tabular	Employment	Gender	665,137
`CelebA`	Image	Attractive	Gender, Age	202,599

B Experiment Details

In this section, we provide a detailed description of our experiment setting, including the description of the datasets, neural network architectures, and experiment settings used in our experiments.

B.1 Dataset

In the appendix, we provide more details about the datasets used in the experiments. These include tabular datasets such as `UCI Adult`, `KDD Census`, `ACS-I`, `ACS-E`, as well as the image dataset `CelebA`. We provide the statistics of the datasets in Table 1. In the following, we provide a comprehensive description of each dataset used in our experiments:

- **`UCI Adult`**[5] (Dua & Graff, 2017): This dataset is extracted from the 1994 Census database. The downstream task is to predict whether the personal income is over 50K a year. The sensitive attributes in this dataset are gender and race.
- **`KDD Census`**[6] (Dua & Graff, 2017): This data set contains 299285 census data extracted from the 1994 and 1995 by the U.S. Census Bureau. The data contains 41 demographic and employment related variables. The instance weight indicates the number of people in the population that each record represents due to stratified sampling. The sensitive attributes are gender and race.
- **`ACS-I`**(ncome)[7] (Ding et al., 2021): The task is to predict whether the income of an individual is above $\$50,000$. The source data was filtered to only include individuals above the age of 16, who reported usual working hours of at least 1 hour per week in the past year, and an income of at least $\$100$. The threshold of the income is $\$50,000$. We use two data from this dataset which spans 1 year or 5 years. The sensitive attributes we consider for this dataset are gender and race.

[5] https://archive.ics.uci.edu/ml/datasets/adult
[6] https://archive.ics.uci.edu/ml/datasets/Census-Income+(KDD)
[7] https://github.com/zykls/folktables

- **ACS-E**(mployment)[8] (Ding et al., 2021): The task is to predict whether an individual is employed and the individuals are between the ages of 16 and 90. We use two data from this dataset which spans 1 year or 5 years. The sensitive attribute we consider for this dataset is gender.
- **CelebA**[9] (Liu et al., 2018): The CelebFaces Attributes (CelebA) dataset a large-scale face attributes dataset consisting of more than 200K celebrity images and each image has 40 face attributes. The downstream task is to predict whether the person is attractive (smiling) or not, formulated as binary classification tasks. We consider Male (gender) and Young (age) as sensitive attributes.

B.2 Neural Network Architectures.

The experiments were conducted using a Multilayer Perceptron (MLP) neural network architecture for tabular data and a ResNet-18 architecture for image data.

- **Tabular**. Tabular data is structured data with a fixed number of input features. For our experiments, we used a two-layer Multilayer Perceptron (MLP) with 256 hidden neurons. The MLP architecture is commonly used for tabular data.
- **Image**. We used a ResNet-18 architecture for image data in our experiments. The used ResNet-18 is has been pre-trained on the ImageNet dataset. And we fine-tuned it for our task. The ResNet-18 architecture is widely used for image classification tasks, as it can handle the high dimensionality of image data and learn hierarchical representations of the input features.

B.3 Experiment setting

In our experiment, we train the neural network with Adam(Kingma & Ba, 2014). Since the results include two metrics (accuracy and fairness), thus we trained the neural network with fixed training epochs. For UCI Adult, KDD Census dataset, we train 2 epochs, and we trained 8 epochs for ACS-I and ACS-E datasets. For the CelebA dataset, we trained for 30 epochs. We use the Xavier (Glorot & Bengio, 2010) to initialize the neural networks. For baseline **Fixed Training**, we set A in $[0, 1]$ with interval of 0.05, resulting in 20 models.

[8] https://github.com/zykls/folktables
[9] https://mmlab.ie.cuhk.edu.hk/projects/CelebA.html