Mitigating Algorithmic Bias with Limited Annotations

Anonymous authors
Paper under double-blind review

Abstract

Existing work on fairness modeling commonly assumes that sensitive attributes for all instances are fully available, which may not be true in many real-world applications due to the high cost of acquiring sensitive information. When sensitive attributes are not disclosed or available, it is needed to manually annotate a small part of the training data to mitigate bias. However, the skewed distribution across different sensitive groups preserves the skewness of the original dataset in the annotated subset, which leads to non-optimal bias mitigation. To tackle this challenge, we propose Active Penalization Of Discrimination (APOD), an interactive framework to guide the limited annotations towards maximally eliminating the effect of algorithmic bias. The proposed APOD integrates discrimination penalization with active instance selection to efficiently utilize the limited annotation budget, and it is theoretically proved to be capable of bounding the algorithmic bias. According to the evaluation on five benchmark datasets, APOD outperforms the state-of-the-arts baseline methods under the limited annotation budget, and shows comparable performance to fully annotated bias mitigation, which demonstrates that APOD could benefit real-world applications when sensitive information is limited. The source code of the proposed method is available at: https://anonymous.4open.science/r/APOD-fairness-4C02.

1 Introduction

Although deep neural networks (DNNs) have been demonstrated with great performance in many real-world applications, it shows discrimination towards certain groups or individuals (Caton & Haas, 2020; Tolmeijer et al., 2020; Rajkomar et al., 2018; Bobadilla et al., 2020), especially in high-stake applications, e.g., loan approvals (Steel & Angwin, 2010), policing (Goel et al., 2016), targeted advertisement (Sweeney, 2013), college admissions (Zimdars, 2010), or criminal risk assessments (Angwin et al., 2016). Social bias widely exists in many real-world data (Mehrabi et al., 2021; Chen et al., 2018; Li & Vasconcelos, 2019; Chuang & Mroueh, 2021). For example, the Adult dataset (Dua & Graff, 2017a) contains significantly more low-income female instances than males. Recent studies revealed that training a DNN model on biased data may inherit and even amplify the social bias and lead to unfair predictions in downstream tasks (Dwork et al., 2012; Creager et al., 2019; Sun et al., 2019; Kusner et al., 2017; Dai & Wang, 2021).

The problem of bias mitigation is challenging due to the skewed data distribution (Hashimoto et al., 2018; Azzalini & Valle, 1996; Azzalini, 2005) across different demographic groups. For example, in the Adult dataset, instances of female with high income are significantly less than the ones with low income (Dua & Graff, 2017a). Also, in the German credit dataset, the majority of people younger than 35 show a bad credit history (Dua & Graff, 2017b). The effect of the skewed distribution on model fairness is illustrated in a binary classification task (e.g. positive class denoted as gray + and •, negative class as red + and •) with two sensitive groups (e.g. group 0 denoted as + and +, group 1 as • and •) shown in Figure 1. In Figure 1 (a), the positive instances (+) are significantly less than negative instances (+) in group 0, which leads to a classification boundary deviating from the fair one. Existing work on fairness modeling can be categorized into two groups with or without sensitive attributes (Du et al., 2020; Kleinberg et al., 2018). The first group relied on full exposure of sensitive attributes in training data, such as Fair Mixup (Chuang & Mroueh, 2021), Rationale regularization (Du et al., 2019; Zafar et al., 2017), Adversarial learning (Arduini et al., 2020) and Group DRO (Sagawa et al., 2019). However, the sensitive information may not be disclosed

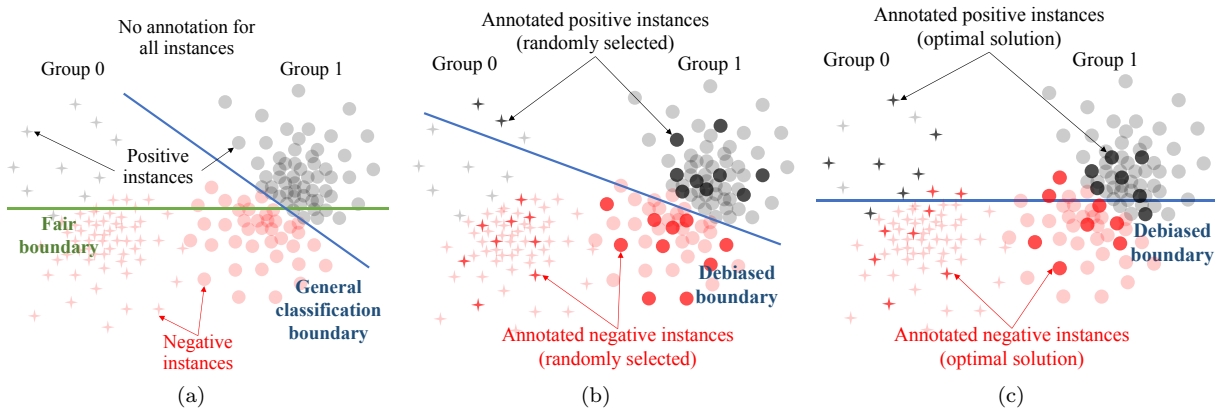


Figure 1: (a) The general classification boundary without bias mitigation deviates from the fair boundary due to the skewed distribution across four underlying subgroups (i.e. $+$, $+$, $\bullet$ and $\bullet$). (b) The annotation budget is set as 30. The randomly annotated data subset follows the same skewed distribution across the subgroups. The classification model is still unfair on the entire dataset. (c) With the same annotation budget, the optimal solution should select a more representative subset, which mitigates algorithmic bias on the entire dataset.

or available in some real world scenarios (Zhao et al., 2021; Dai & Wang, 2021; Chen et al., 2019; Kallus et al., 2021), and the cost of annotating sensitive attributes by human experts is high (Anahideh et al., 2020), which leads to the limited applications of this group of work to the real-world scenarios.

The second group of work formulates the fairness without dependency on sensitive information, such as ARL (Lahoti et al., 2020), ReBias (Bahng et al., 2020), LfF (Nam et al., 2020) and JTT (Liu et al., 2021). However, those works rely on heuristic clustering of training instances to form potential demographic groups for the bias mitigation, which may deteriorate the fairness performance to some extent (Wang et al., 2020). To tackle the issue, some work involves the human expert providing a partially annotated dataset for the bias mitigation (Anahideh et al., 2020). However, only a small portion of the dataset is annotated due to the limitation of human labor efforts. An intuitive solution is to randomly select a small portion of instances for annotation and target semi-supervised bias mitigation (Jung et al., 2021; Awasthi et al., 2021). However, as shown in Figure 1 (b), the randomly selected instances will follow the same skewed distribution across sensitive groups, which still preserves the bias information in the classifier. In such a manner, it is highly likely to achieve a non-optimal solution, which is fair only on the annotated dataset but not the entire dataset. Therefore, it is needed to have a unified framework, which integrates the selection of a representative subset for annotation with model training towards the global fairness, as shown in Figure 1 (c).

In this work, we propose Active Penalization Of Discrimination (APOD), a novel interactive framework which integrates the penalization of discrimination with active instance selection, for bias mitigation in the real-world scenarios where sensitive information is limited. Specifically, APOD iterates between the model debiasing and active instance selection to gradually approach the global fairness. For debiasing the model, APOD enables bias penalization in an end-to-end manner via adopting a fairness regularizer. In the active instance selection, an annotated data subset is constructed via recursive selections of representative data instances from the subgroup where the model shows the worst performance, such that it can maximally expose the existing bias of the model for subsequent debiasing. Finally, we provide theoretical and experimental analysis to demonstrate the effectiveness of APOD. The contributions of this work are summarized as follows:

- We propose an interactive framework APOD to integrate the bias mitigation with efficient active instance selection when the annotation of sensitive attributes is very limited.
- We propose the relaxed reformulation of the fairness objective, and theoretically prove that APOD could improve model fairness via bounding the relaxed fairness metric.
- The effectiveness of APOD is thoroughly demonstrated by the experiments on five benchmark datasets, which shows APOD is competitive with state-of-the-art methods using fully disclosed sensitive attributes.

2 Preliminaries

In this section, we first introduce the notations used in this work, and give the problem definition of bias mitigation in the active scenario. Then, we introduce the fairness metrics.

2.1 Notation and Problem Definition

Without loss of generality, we follow the existing work (Chuang & Mroueh, 2021; Zhang et al., 2018; Lahoti et al., 2020) to consider a classification task in this work. Specifically, we aim to learn a DNN classifier f with the input feature $\boldsymbol{x} \in \mathcal{X}$, label $y \in \mathcal{Y} = \{0, 1\}$ and sensitive attribute $a \in \mathcal{A} = \{0, 1\}$, where $\mathcal{X}$ and $\mathcal{Y}$ denote the feature and label space, respectively. The instances with sensitive attribute $A = 0$ and $A = 1$ belong to the unprivileged and privileged groups, respectively. Let $\mathcal{D} = \{(\boldsymbol{x}_i, y_i) \mid 1 \leq i \leq N\}$ denote the entire dataset, which consists of the annotated set $\mathcal{S} = \{(\boldsymbol{x}_i, y_i, a_i)\}$ and unannotated set $\mathcal{U} = \{(\boldsymbol{x}_i, y_i)\}$, i.e., the value of the sensitive attribute is known for instances in $\mathcal{S}$, but it is unknown for instances in $\mathcal{U}$. The proposed interactive bias mitigation is illustrated in Figure 2 (a). Specifically, in each iteration, an instance $(\boldsymbol{x}^*, y^*)$ is selected from unannotated dataset $\mathcal{U}$ for human experts; the experts essentially do the job of mapping $\mathcal{X} \times \mathcal{Y} \to \mathcal{X} \times \mathcal{Y} \times \mathcal{A}$, by providing the annotation of sensitive attribute a^* for the selected instance $(\boldsymbol{x}^*, y^*)$. After that, the classifier is updated and debiased using the partially annotated dataset including the newly annotated instance $(\boldsymbol{x}^*, y^*, a^*)$, where the new classifier will then be involved for the instance selection in the next iteration. This loop terminates if the human-annotation budget is reached.

Such an active scenario to debias f is time-consuming for deep neural networks, due to the fact that f is retrained in each iteration. To improve the efficiency of learning, the classifier f is split into body $f_b : \mathcal{X} \to \mathbb{R}^M$ and head $f_h : \mathbb{R}^M \to \mathbb{R}^{|\mathcal{Y}|}$, where the body f_b denotes the first several layers, and the head f_h denotes the remaining layers of the classifier such that $\hat{y}_i = \arg\max\{f_h(f_b(\boldsymbol{x}_i|\theta_b)|\theta_h)\}$. The body f_b learns the instance embedding $\boldsymbol{h}_i = f_b(\boldsymbol{x}_i|\theta_h)$, where $\boldsymbol{h}_i \in \mathbb{R}^M$ denotes the embedding of $\boldsymbol{x}_i$, and M denotes the dimension of embedding space. The head f_h contributes to fair classification via having $\hat{y}_i = \arg\max\{f_h(\boldsymbol{h}_i|\theta_h)\}$, where $f_h(\boldsymbol{h}_i|\theta_h) \in \mathbb{R}^{|\mathcal{Y}|}$ and $\hat{y}_i \in \mathcal{Y}$. Instead of updating the entire classifier, the classifier body f_b is pretrained and fixed during the bias mitigation, where f_b is pretrained to minimize the cross-entropy loss without annotations of sensitive attributes. In such a manner, the mitigation of unfairness relies on debiasing the classifier head f_h. This strategy with a fixed classifier body during the bias mitigation has been proved to be effective enough in existing works (Du et al., 2021; Slack et al., 2020).

2.2 Fairness Evaluation Metrics

In this work, we follow the existing work (Mehrabi et al., 2021; Gajane & Pechenizkiy, 2017; Chuang & Mroueh, 2021; Du et al., 2021) to consider two metrics to evaluate fairness: Equality of Opportunity (Hardt et al., 2016; Verma & Rubin, 2018) and Equalized Odds (R. et al., 2020; Verma & Rubin, 2018). These two metrics are measured based on the true positive rate $\text{TPR}_{A=a} = \mathbb{P}(\hat{Y} = 1|A = a, Y = 1)$ and the false positive rate $\text{FPR}_{A=a} = \mathbb{P}(\hat{Y} = 1|A = a, Y = 0)$ for $a \in \mathcal{A}$.

Equality of Opportunity requires the unprivileged group ($A = 0$) and privileged groups ($A = 1$) have equal probability of an instance from the positive class being assigned to positive outcome, which is defined as $\mathbb{P}(\hat{Y} = 1|A = 0, Y = 1) = \mathbb{P}(\hat{Y} = 1|A = 1, Y = 1)$. In this work, we apply EOP given as follows to evaluate Equality of Opportunity,

$$\text{EOP} = \frac{\text{TPR}_{A=0}}{\text{TPR}_{A=1}} = \frac{\mathbb{P}(\hat{Y} = 1 \mid A = 0, Y = 1)}{\mathbb{P}(\hat{Y} = 1 \mid A = 1, Y = 1)}. \tag{1}$$

Equalized Odds expects favorable outcomes to be independent of the sensitive attribute, given the ground-truth prediction, which can be formulated as $\mathbb{P}(\hat{Y} = 1|A = 0, Y = y) = \mathbb{P}(\hat{Y} = 1|A = 1, Y = y)$ for $y \in \mathcal{Y}$. To evaluate Equalized Odds, ΔEO combines the difference of TPR and FPR across two sensitive groups as

$$\Delta\text{EO} = \Delta\text{TPR} + \Delta\text{FPR}, \tag{2}$$

where $\Delta\text{TPR} = \text{TPR}_{A=0} - \text{TPR}_{A=1}$ and $\Delta\text{FPR} = \text{FPR}_{A=0} - \text{FPR}_{A=1}$. Under the above definitions, EOP close to 1 and ΔEO close to 0 indicate fair classification results.

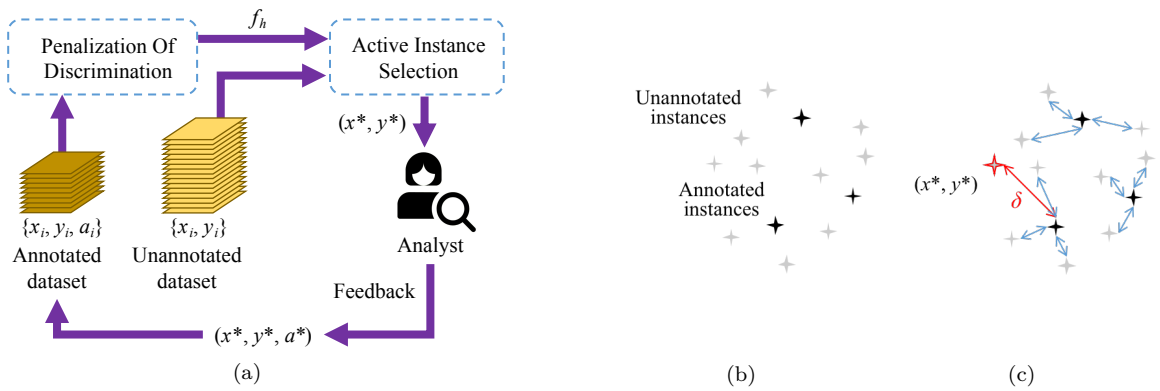


Figure 2: (a) The APOD pipeline alternates between POD and AIS. (b) Individual selection: The annotated and unannotated instances from subgroup $\mathcal{U}_{\tilde{a}}^{\tilde{c}}$, where $\tilde{a} = 0$ and $\tilde{c} = 1$. (c) Each unannotated instance is connected to an annotated instance determined by $\min_{\boldsymbol{x}_j \in \mathcal{S}} ||\boldsymbol{h}_i - \boldsymbol{h}_j||_2$ (marked as blue arrows). The red δ denotes the largest distance pair which selects the best candidate for annotation.

3 Active Penalization Of Discrimination

In this section, we introduce the Active Penalization Of Discrimination (APOD) framework to mitigate algorithmic bias under a limited annotation budget. As shown in Figure 2 (a), APOD integrates Penalization Of Discrimination (POD) and Active Instance Selection (AIS) in a unified and iterative framework. Specifically, in each iteration, POD focuses on debiasing the classifier head f_h on the partially annotated dataset $\{(\boldsymbol{x}_i, y_i, a_i) \in \mathcal{S}\}$ and $\{(\boldsymbol{x}_i, y_i) \in \mathcal{U}\}$, while AIS selects the optimal instance $(\boldsymbol{x}^*, y^*)$ from the unannotated dataset $\mathcal{U}$ that can further promote bias mitigation. Sensitive attributes of the selected instances will be annotated by human experts: $(\boldsymbol{x}^*, y^*) \rightarrow (\boldsymbol{x}^*, y^*, a^*)$. After that, these instances will be moved from the unannotated dataset $\mathcal{U} \leftarrow \mathcal{U} \setminus \{(\boldsymbol{x}^*, y^*)\}$ to the annotated dataset $\mathcal{S} \leftarrow \mathcal{S} \cup \{(\boldsymbol{x}^*, y^*, a^*)\}$ for debiasing the classifier in the next iteration. The POD and AIS are introduced as follows.

3.1 Penalization Of Discrimination (POD)

POD learns a fair classifier head f_h via bias penalization on both annotated instances $\{(\boldsymbol{x}_i, y_i, a_i) \in \mathcal{S}\}$ and unannotated instances $\{(\boldsymbol{x}_i, y_i) \in \mathcal{U}\}$. To be concrete, POD considers a regularization term, consisting of the true and false positive rate difference[1], to balance the model performance on different subgroups. In this way, given $\boldsymbol{h}_i = f_b(\boldsymbol{x}_i|\theta_b)$, f_h is updated to minimize the hybrid loss function given by

$$L = \sum_{i=1}^{N} l(\boldsymbol{h}_i, y_i; \theta_h) + \lambda(\Delta\text{TPR}^2 + \Delta\text{FPR}^2), \tag{3}$$

where $l(\boldsymbol{h}_i, y_i; \theta_h)$ denotes the cross-entropy loss, and the term $\Delta\text{TPR}^2 + \Delta\text{FPR}^2$ penalizes the bias in f_h to improve fairness, controlled by the hyper-parameter λ.

However, Equation (3) is not feasible to debias f_h in an end-to-end manner, since neither TPR nor FPR is differentiable with respect to the parameters θ_h. It is thus necessary to reformulate ΔTPR and ΔFPR, which involves the parameterization of true and false positive rate with respect to θ_h, respectively. For notation convenience and without the loss of generality, we unify the formulation of true and false positive rates by

$$p_a(y, c) = \mathbb{P}(\hat{Y} = c \mid Y = y, A = a), \tag{4}$$

[1]The combination of TPR and FPR is representative enough accross different fairness metrics. POD is flexible to use other metrics as the regularizer for the bias mitigation.

where we take $y = 1, c = 1$ to have $p_a(1,1) = \text{TPR}_{A=a}$ and $y = 0, c = 1$ to have $p_a(0,1) = \text{FPR}_{A=a}$. To parameterize $p_a(y,c)$ with respect to θ_h, we reformulate it as follows

$$p_a(y,c) = \frac{\sum_{(\boldsymbol{x}_i,y_i,a_i)\in\mathcal{S}_a^y} \mathbf{1}_{\hat{y}_i=c}}{|\mathcal{S}_a^y|} = \frac{\sum\limits_{(\boldsymbol{x}_i,y_i,a_i)\in\mathcal{S}_a^y} \text{sgn}(f_h^c(\boldsymbol{h}_i) - f_h^{1-c}(\boldsymbol{h}_i))}{|\mathcal{S}_a^y|} \tag{5}$$

$$\approx \frac{\sum\limits_{(\boldsymbol{x}_i,y_i,a_i)\in\mathcal{S}_a^y} \lambda(f_h^c(\boldsymbol{h}_i) - f_h^{1-c}(\boldsymbol{h}_i))}{|\mathcal{S}_a^y|} \triangleq \lambda\tilde{p}_a(y,c), \tag{6}$$

where $\text{sgn}(x) = 0$ for $x < 0$ and $\text{sgn}(x) = 1$ for $x \geq 0$. Here we relax $\text{sgn}(x)$ with a linear function[2] λx in the approximation of Equation (5) to make $p_a(y,c)$ differentiable with respect to θ_h; $\mathcal{S}_a^y = \{(\boldsymbol{x}_i, y_i, a_i) \in \mathcal{S} \mid a_i = a, y_i = y\}$ for $a \in \mathcal{A}, y \in \mathcal{Y}$; and $f_h^i(\boldsymbol{h})$ denotes element i of $f_h(\boldsymbol{h})$ for $i \in \mathcal{Y}$. Based on the relaxed regularization term, f_h is updated to minimize the loss function given by

$$L = \frac{1}{N}\sum_{i=1}^{N} l(\boldsymbol{h}_i, y_i; \theta_h) + \lambda \sum_{y\in\mathcal{Y}} \left[\tilde{p}_0(y,1) - \tilde{p}_1(y,1)\right]^2, \tag{7}$$

where the estimation of cross-entropy $\frac{1}{N}\sum_{i=1}^{N} l(\boldsymbol{h}_i, y_i; \theta_h)$ includes both annotated and unannotated instances; the regularization term $[\tilde{p}_0(y,1) - \tilde{p}_1(y,1)]^2$ is calculated using the annotated instances; and the hyper-parameter λ controls the importance of regularization.

3.2 Active Instance Selection (AIS)

In each iteration, AIS selects instances from the unannotated dataset $\mathcal{U}$ to annotate the sensitive attribute values. The newly annotated instances are merged with the dataset for debiasing the classifier head in subsequent iterations. The AIS process consists of two steps: (1) **Group selection** is to select the subgroup $\mathcal{U}_{\tilde{a}}^{\tilde{c}} = \{(\boldsymbol{x}_i, y_i) \in \mathcal{U} \mid a_i = \tilde{a}, y_i = \tilde{c}\}$ on which the model has the worst performance; (2) **Individual selection** is to select the optimal instance from $\mathcal{U}_{\tilde{a}}^{\tilde{c}}$, which can mostly expose the existing bias of the model for promoting the bias mitigation in the next iteration.

3.2.1 Group Selection

The group selection is motivated by the observation that adding more instances to the subgroup having the worst classification accuracy can improve the fairness by increasing its contribution to the average loss (Hashimoto et al., 2018; Lahoti et al., 2020). Specifically, for group selection, the unannotated dataset $\mathcal{U}$ is splitted into $\{\mathcal{U}_a^c\}_{a\in\mathcal{A},c\in\mathcal{Y}}$, where $\mathcal{U}_a^c = \{(\boldsymbol{x}_i, y_i) \in \mathcal{U} | a_i = a, y_i = c\}$ denotes a subgroup of unannotated instances. We estimate the classification accuracy $p_a(c,c) = \mathbb{P}(\hat{Y} = c | A = a, Y = c)$ to evaluate f on each subgroup $\mathcal{U}_a^c$ for $a \in \mathcal{A}$ and $c \in \mathcal{Y}$, respectively, following Equation (4). In this way, the subgroup $\mathcal{U}_{\tilde{a}}^{\tilde{c}} = \{(\boldsymbol{x}_i, y_i) \in \mathcal{U} | a_i = \tilde{a}, y_i = \tilde{c}\}$ which suffers from the worst accuracy is selected by

$$\tilde{a}, \tilde{c} = \arg\min_{a\in\mathcal{A},c\in\mathcal{Y}} p_a^*(c,c), \tag{8}$$

where $p_a^*(c,c) = p_a(c,c) - (p_0(c,c) + p_1(c,c))/2$ denotes the centralized classification accuracy after considering the performance divergence of the classifier on different classes. For example, in Figure 1 (b), we select the subgroup with the worst accuracy $\mathcal{U}_0^1$ which corresponds to the positive instances from group 0, due to the fact that $p_0^*(1,1) < p_0^*(0,0), p_1^*(0,0), p_1^*(1,1)$.

Note that $p_a^*(c,c)$ cannot be estimated without the annotations of sensitive attribute. We thus learn another classifier head $f_a : \mathbb{R}^M \to \mathbb{R}^{|\mathcal{A}|}$ to predict the sensitive attribute $\hat{a} = \arg\max f_a(\boldsymbol{h}_i | \theta_a)$ for the unannotated instances $\boldsymbol{x}_i \in \mathcal{U}$, where f_a is updated on the annotated dataset $\mathcal{S}$ by minimizing the cross-entropy loss

$$\theta_a^* = \frac{1}{|\mathcal{S}|} \sum_{(\boldsymbol{x}_i,y_i,a_i)\in\mathcal{S}} l(\boldsymbol{h}_i, a_i; \theta_a). \tag{9}$$

[2]It also has other choices for the relaxation, e.g. sigmoid and tanh functions. The linear function is chosen for simplicity.

3.2.2 Individual Selection

Individual selection aims to proactively select the most representative instances for annotation, which can maximally promote bias mitigation. Since the classifier f has the worst accuracy on subgroup $\mathcal{U}_{\tilde{a}}^{\tilde{c}}$, reducing the classification error on $\mathcal{U}_{\tilde{a}}^{\tilde{c}}$ would improve fairness, where $\tilde{a}$ and $\tilde{c}$ are chosen through group selection in Equation (8). The strategy of individual selection is to expand the annotated dataset to reduce δ-cover of subgroup $\mathcal{U}_{\tilde{a}}^{\tilde{c}}$ (Sener & Savarese, 2018). Specifically, the annotated dataset $\mathcal{S}$ enables δ-cover of the entire dataset $\mathcal{D}$ if $\forall \boldsymbol{x}_i \in \mathcal{D}$, $\exists \boldsymbol{x}_j \in \mathcal{S}$ such that $||\boldsymbol{x}_i - \boldsymbol{x}_j||_2 \leq \delta$, where δ denotes the coverage radius given by

$$\delta = \max_{\boldsymbol{x}_i \in \mathcal{D}} \min_{\boldsymbol{x}_j \in \mathcal{S}} ||\boldsymbol{x}_i - \boldsymbol{x}_j||_2. \tag{10}$$

Furthermore, it is observed that the generalization error of a model approaches the training error[3] if the coverage radius δ is small (Sener & Savarese, 2018). Following such scheme, we select the instance in subgroup $\mathcal{U}_{\tilde{a}}^{\tilde{c}}$, which could decrease δ-coverage to reduce the classification error on $\mathcal{U}_{\tilde{a}}^{\tilde{c}}$. To be concrete, the distance between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$ is measured by $||\boldsymbol{h}_i - \boldsymbol{h}_j||_2$, where $\boldsymbol{h}_i = f_b(\boldsymbol{x}_i|\theta_b)$ and $\boldsymbol{h}_j = f_b(\boldsymbol{x}_j|\theta_b)$ are the embeddings of $\boldsymbol{x}_i$ and $\boldsymbol{x}_i$, respectively. We have the instance $(\boldsymbol{x}^*, y^*)$ selected for annotation following the max-min rule

$$(\boldsymbol{x}^*, y^*) = \operatorname*{arg\,max}_{(\boldsymbol{x}_i, y_i) \in \mathcal{U}_{\tilde{a}}^{\tilde{c}}} \min_{(\boldsymbol{x}_j, y_j) \in \mathcal{S}} ||\boldsymbol{h}_i - \boldsymbol{h}_j||_2. \tag{11}$$

The individual selection strategy is illustrated in Figures 2 (b) and (c), where δ reduction guides the individual selection. The candidate instances in $\mathcal{U}_{\tilde{a}}^{\tilde{c}}$ and annotated instances are shown in Figure 2 (b). The distances between each candidate instance and annotated instances are measured in embedding space $||\boldsymbol{h}_i - \boldsymbol{h}_j||_2$, where the minimal one is marked as a blue arrow. The red instance marked by (x^*, y^*) in Figure 2 (c) indicates the best instance candidate to be annotated.

Algorithm 1: Active Penalization Of Discrimination.

1 **Input**: initial annotated dataset $\mathcal{S}$.
2 **Output**: classifier body f_b and head f_h.
3 $\theta_b^*, \theta_h^* = \arg\min \sum_{i=1}^{N} l(\boldsymbol{x}_i, y_i; \theta_b, \theta_h)$.
4 **while** *within budget limit* **do**
5 # Penalization Of Discrimination.
6 $\theta_h^* = \text{POD}(\mathcal{S},\ f_b,\ f_h)$
7 # Active Instance Selection.
8 $(\boldsymbol{x}^*, y^*) = \text{AIS}(f_b, f_h)$
9 $\mathcal{S} = \mathcal{S} \cup \{(\boldsymbol{x}^*, y^*, a^*)\}$ and $\mathcal{U} = \mathcal{U} \setminus \{(\boldsymbol{x}^*, y^*)\}$

Algorithm 2: Penalization Of Discrimination (POD)

1 **Input**: annotated dataset $\mathcal{S}$, classifier body f_b and classifier head f_h.
2 **Output**: fair classifier head f_h^*.
3 **while** *not converged* **do**
4 For $a \in \mathcal{A}$ and $y \in \mathcal{Y}$, estimate $\tilde{p}_a(y, 1)$ given by Equation (6).
5 Update the classifier head f_h to minimize the loss function in Equation (7).
6 **return** f_h^*

3.3 The APOD Algorithm

The details of APOD are summarized in Algorithm 1. Initially, APOD learns the biased f_b and f_h, and randomly samples a small set of annotated instances $\mathcal{S}$. In each iteration, APOD first learns f_a to predict the sensitive attribute of unannotated instances; then debiases f_h via POD (line 5); after this, APOD selects the optimal instance $(\boldsymbol{x}^*, y^*)$ for annotation via AIS (line 6) and merges the selected instance with the annotated dataset (line 7); POD and AIS are given in Algorithms 2 and 3, respectively; the iteration stops once the number of annotated instance reaches the budget.

3.4 Theoretical Analysis

We theoretically investigate the proposed APOD to guarantee that bias mitigation is globally achieved, as shown in Theorem 1. We then demonstrate the effectiveness of AIS (both group selection and individual selection) in Remark 1. The proof of Theorem 1 is given in Appendix A.

[3]The training error is less than generalization error in most cases.

Algorithm 3: Active Instance Selection (AIS).

Input: classifier body f_b and classifier head f_h.
Output: the selected instance $(\boldsymbol{x}^*, y^*)$.
Update f_a to minimize $\frac{1}{|\mathcal{S}|}\sum_{(\boldsymbol{x}_i,y_i,a_i)\in\mathcal{S}} l(\boldsymbol{h}_i, a_i; \theta_a)$.
Estimate the sensitive attribute $\hat{a}_i = \arg\max f_a(\boldsymbol{h}_i \mid \theta_a)$ for $\boldsymbol{x}_i \in \mathcal{U}$.
For $a \in \mathcal{A}$ and $c \in \mathcal{Y}$, estimate the classification accuracy $p_a(c,c) = \mathbb{P}(\hat{Y} = c|\hat{A} = a, Y = c)$ on subgroup $\mathcal{U}_a^c$.
For $a\in\mathcal{A}$ and $c\in\mathcal{Y}$, centralize $p_a(c,c)$ into $p_a^*(c,c)$ by $p_a^*(c,c) = p_a(c,c) - \frac{p_0(c,c)+p_1(c,c)}{2}$.
Execute the group selection by

$$\tilde{a}, \tilde{c} = \arg\min_{a\in\mathcal{A}, c\in\mathcal{Y}} p_a^*(c,c).$$

Execute the individual selection by

$$(\boldsymbol{x}^*, y^*) = \arg\max_{(\boldsymbol{x}_i,y_i)\in\mathcal{U}_{\tilde{a}}^{\tilde{c}}} \min_{(\boldsymbol{x}_j,y_j,a_j)\in\mathcal{S}} ||\boldsymbol{h}_i - \boldsymbol{h}_j||_2.$$

Theorem 1. *Assume the loss value on the training set satisfies* $\frac{1}{|\mathcal{S}|}\sum_{(\boldsymbol{x}_i,y_i,a_i)\in\mathcal{S}} l(\boldsymbol{h}_i, y_i; \theta_h) \leq \epsilon$[4]*, and* $l(\boldsymbol{h}, y; \theta_h)$ *and* f_h *satisfy* K_l*- and* K_h*-Lipschitz continuity*[5]*, respectively. The generalization loss difference between unprivileged group and privileged group has the following upper bound with probability* $1-\gamma$*,*

$$\left|\int_{\mathcal{X}_0}\int_{\mathcal{Y}} p(\boldsymbol{x},y) l(\boldsymbol{h}, y; \theta_h) \mathrm{d}\boldsymbol{x}\mathrm{d}y - \int_{\mathcal{X}_1}\int_{\mathcal{Y}} p(\boldsymbol{x},y) l(\boldsymbol{h}, y; \theta_h) \mathrm{d}\boldsymbol{x}\mathrm{d}y\right| \leq \epsilon + \min\left\{\sqrt{-L^2 \log\gamma (2N_{\tilde{a}})^{-1}}, (K_l + K_h L)\delta_{\tilde{a}}\right\}, \tag{12}$$

where $\tilde{a} = \arg\max_{a\in\mathcal{A}} \int_{\mathcal{X}_a}\int_{\mathcal{Y}} p(\boldsymbol{x},y) l(\boldsymbol{h}, y; \theta_h)\mathrm{d}\boldsymbol{x}\mathrm{d}y$*;* $\mathcal{X}_a = \{\boldsymbol{x}_i \in \mathcal{D} | a_i = a\}$*;* $\delta_{\tilde{a}} = \max_{\boldsymbol{x}_i\in\mathcal{X}_{\tilde{a}}} \min_{(\boldsymbol{x}_j,y_j,a_j)\in\mathcal{S}} ||\boldsymbol{h}_i - \boldsymbol{h}_j||_2$*;* $N_{\tilde{a}} = |\{(\boldsymbol{x}_i, y_i, a_i) | a_i = \tilde{a}, (\boldsymbol{x}_i, y_i, a_i) \in \mathcal{S}\}|$*;* $L = \max_{(\boldsymbol{x}_i,y_i)\in\mathcal{U}} l(\boldsymbol{h}_i, y_i; \theta_h)$*; and* $\boldsymbol{h}_i = f_b(\boldsymbol{x}_i|\theta_b)$*.*

In Theorem 1, the global fairness is formalized via considering the generalization error difference between the unprivileged and privileged group as the relaxed fairness metric, and APOD contributes to the global fairness via explicitly tightening the upper bound of the relaxed fairness metric. We demonstrate the details that AIS can iteratively tighten the bound in Remark 1.

Remark 1. *In each iteration of APOD, the group selection reduces the value of* $\sqrt{-L^2 \log\gamma (2N_{\tilde{a}})^{-1}}$ *by merging a new instance* $(\boldsymbol{x}_i, y_i, a_i)|_{a_i=\tilde{a}}$ *to the annotated dataset* $\mathcal{S}$ *to increase the value of* $N_{\tilde{a}} = |\{(\boldsymbol{x}_i, y_i, a_i) \in \mathcal{S} | a_i = \tilde{a}\}|$*. Here, we adopt an approximation given by Equation (13) due to the negative relationship between the accuracy and the generalization loss,*

$$\tilde{a} = \arg\min_{a\in\mathcal{A}} p_a^*(c,c) \approx \arg\max_{a\in\mathcal{A}} \int_{\mathcal{X}_a}\int_{\mathcal{Y}_c} p(\boldsymbol{x},y) l(\boldsymbol{h}, y; \theta_h)\mathrm{d}\boldsymbol{x}\mathrm{d}y, \tag{13}$$

where $\mathcal{Y}_c = \{y = c \mid y \in \mathcal{Y}\}$ *for* $c \in \mathcal{Y}$*. Meanwhile, the individual selection reduces the value of* $\delta_{\tilde{a}}$ *by selecting an instance following Equation (11). With the combination of group selection and individual selection, APOD contributes to the decline of* $\min\{\sqrt{-L^2 \log\gamma (2N_{\tilde{a}})^{-1}}, (K_l + K_h L)\delta_{\tilde{a}}\}$*, which leads to tightening the upper bound of the fairness metric in Equation (12).*

Remark 1 reveals that both group selection and individual selection of the two-step AIS are effective in tightening the upper bound of relaxed fairness metric. Compared to AIS, we consider two compositional instance selection methods: one with group selection alone, where we randomly select an instance $(\boldsymbol{x}^*, y^*)$ from the subgroup $\mathcal{U}_{\tilde{a}}^{\tilde{c}}$ satisfying $\tilde{a}, \tilde{c} = \arg\min_{a\in\mathcal{A}, c\in\mathcal{Y}} p_a^*(c,c)$; and another with individual selection alone, where an instance is selected via $(\boldsymbol{x}^*, y^*) = \arg\max_{(\boldsymbol{x}_i,y_i)\in\mathcal{U}} \min_{(\boldsymbol{x}_j,y_j,a_j)\in\mathcal{S}} ||\boldsymbol{h}_i - \boldsymbol{h}_j||_2$ without the selection of subgroup. According to Remark 1, the compositional methods merely enable to reduce one of the terms $(2N_{\tilde{a}})^{-1}$ or $\delta_{\tilde{a}}$ in Equation (12), which are less effective than the two-step AIS as an unit.

[4] ϵ can be very small if the classifier head f_h has been well-trained on the annotated dataset $\mathcal{S}$.

[5] $l(\boldsymbol{h}, y; \theta_h)$ and f_h satisfy $|l(\boldsymbol{h}_i, y; \theta_h) - l(\boldsymbol{h}_j, y; \theta_h)| \leq K_l ||\boldsymbol{h}_i - \boldsymbol{h}_j||_2$ and $|p(y|\boldsymbol{x}_i) - p(y|\boldsymbol{x}_j)| \leq K_h ||\boldsymbol{h}_i - \boldsymbol{h}_j||_2$, respectively, where the likelihood function $p(y \mid \boldsymbol{x}_i) = \mathrm{softmax}(f_h(\boldsymbol{h}_i|\theta_h))$.

4 Experiment

In this section, we conduct experiments to evaluate APOD, aiming to answer the following research questions: **RQ1**: In terms of comparison with state-of-the-art baseline methods, does APOD achieve more effective mitigation of unfairness under the same annotation budget? **RQ2**: Does APOD select more informative annotations for bias mitigation than baseline methods? **RQ3**: How does the ratio of annotated instances affect the mitigation performance of APOD? **RQ4**: Do both group selection and individual selection in the AIS contribute to bias mitigation?

4.1 Experimental Setup

Datasets. The experiments are conducted on the MEPS, Loan default, German credit, Adult and CelebA datasets. The details about the datasets including the size and spliting of the datasets, the predicted and sensitive attributes, and the annotation budget are given in Appendix B.

Implementation Details. The experiment on each dataset follows the pipeline of **pre-training**, **debiasing**, and **head-selection**. Each step is shown as follows.

Pre-training: We pre-train $f_h(f_b(\bullet \mid \theta_b) \mid \theta_h)$ on the whole training set without sensitive annotations for 50 epochs; and pre-train $f_a(f_b(\bullet \mid \theta_b) \mid \theta_a)$ for 10 epochs using two randomly selected annotated instances from each group. The pre-trained f_b, f_h and f_a provide initial solutions of the classifier for the bias mitigation.

Debiasing: We adopt APOD to debias the classifier head $f_h(\bullet \mid \theta_h)$ for several iterations. Specifically, the number of iterations equals the available annotation number, where APOD selects one instance for annotation, debiases $f_h(\bullet \mid \theta_h)$ and retrains $f_a(\bullet \mid \theta_a)$ for 10 epochs in each iteration, and back up the snapshot parameter θ_h and θ_a in the last epoch of each iteration. In the Pre-training and Debiasing stages, the parameters θ_b, θ_h and θ_a are updated using the Adam optimizer with a learning rate of 10^{-3}, mini-batch size 256 and a dropout probability of 0.5. The DNN architectures and detailed hyper-parameter settings on different datasets are given in Appendix C.

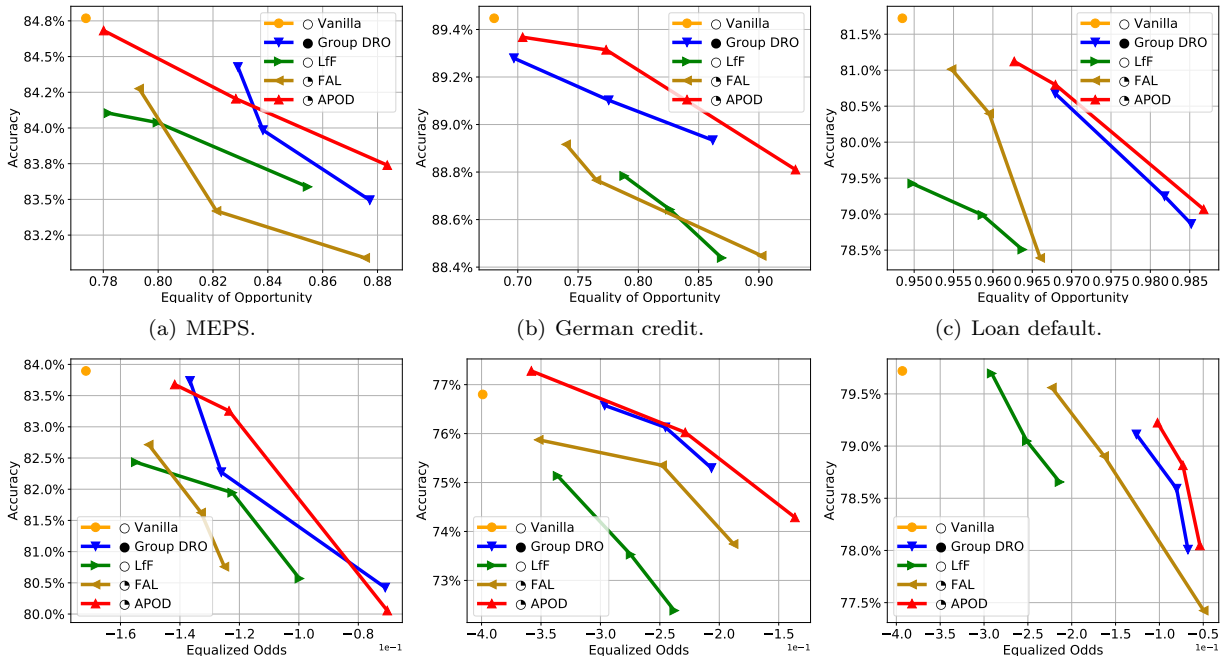
Head-selection: We use the trained $f_a(f_b(\bullet \mid \theta_b) \mid \theta_a)$ to generate the proxy sensitive annotations for the validation dataset so that the fairness metrics can be estimated on the validation dataset. The optimal debiased classifier head f_h is selected to maximize the summation of accuracy and fairness score on the validation dataset. We merge the selected f_h with the pre-trained f_b and test the classifier $f_h(f_b(\bullet \mid \theta_b) \mid \theta_h)$ on the test dataset. This pipeline is executed five times to reduce the effect of randomness, and the average testing performance and the standard deviation are reported in the remaining sections.

4.2 Bias Mitigation Performance Analysis (RQ1)

In this section, APOD is compared with state-of-the-art baseline methods of bias mitigation.

Baseline Methods. Vanilla: The classifier is trained to minimize the cross-entropy loss without bias mitigation. **Group DRO** (Sagawa et al., 2019): Group DRO utilizes all sensitive information to minimize the classification loss on the unprivileged group to reduce the performance gap between different sensitive groups. **Learning from Failure (LfF)** (Nam et al., 2020): As a debiasing method that relies on proxy sensitive annotations, LfF adopts generalized cross-entropy loss to learn a proxy annotation generator, and proposes a re-weighted cross entropy loss to train the debiased model. **Fair Active Learning (FAL)** (Anahideh et al., 2020): The instance selection in FAL is to maintain a subset of annotated instances for model training, which is not guided by gradient-based model debiasing. More details are given in the Appendix D.

To have a fair comparison, we unify the splitting of datasets for all methods, and set the same annotation budget for APOD and FAL. The mitigation performance is indicated by the fairness-accuracy curves (Chuang & Mroueh, 2021), where the hyperparameter λ of APOD varies in the range of $(0, 2]$, and the hyperparameter setting of baseline methods can be referred to Appendix C. We give the fairness-accuracy curves of each method on the five benchmark datasets in Figures 3 (a)-(f), respectively, where ●, ○ and ◔ indicate the bias mitigation relies on entire-, zero- or partial- annotation of the training dataset, respectively. Finally, we follow existing work (Bechavod & Ligett, 2017) to evaluate mitigation performance using the fairness metric



(a) MEPS. (b) German credit. (c) Loan default.

(d) Adult. (e) CelebA-wavy hair. (f) CelebA-young.

Figure 3: Accuracy-fairness curve; Algorithm: Vanilla training, Group DRO, LfF, FAL and APOD; Dataset: (a) MEPS, (b) German credit, (c) Loan default, (d) Adult, (e) CelebA-wavy hair, (f) CelebA-young.

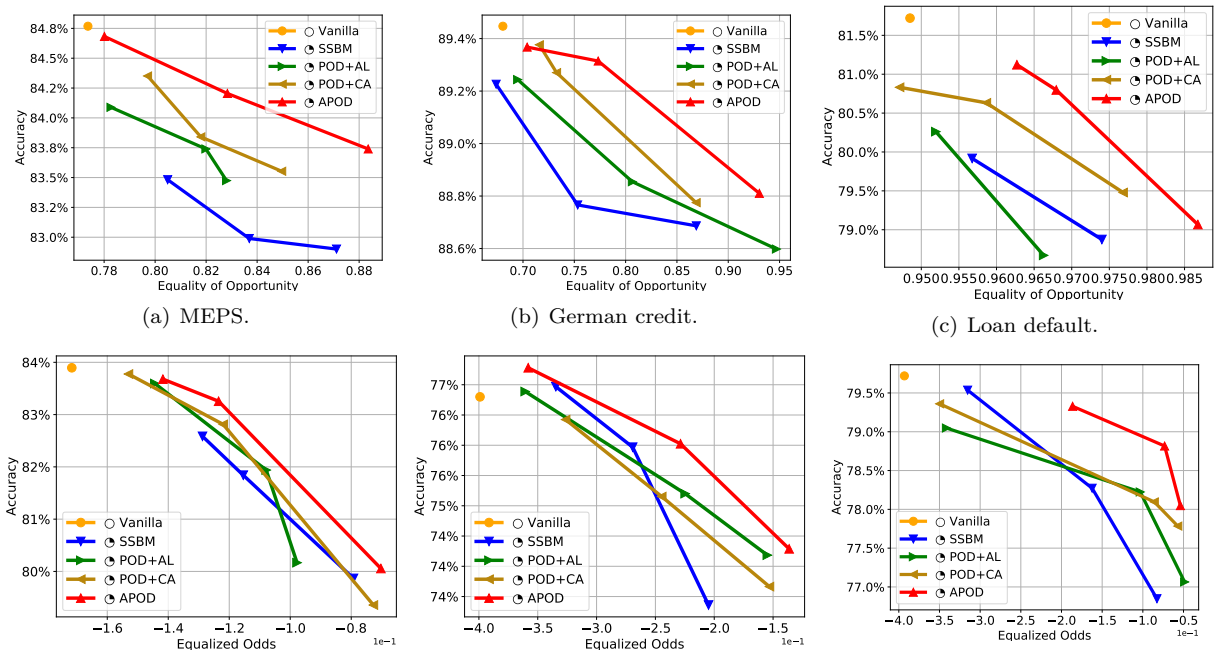
EOP on the MEPS, German credit and Loan default datasets, and using the fairness metric ΔEO on the remaining datasets (Du et al., 2021). In general, we have the following observations:

- APOD outperforms FAL on the five datasets under the same annotation budget in terms of the mitigation performance at the same level of accuracy. This demonstrates the superiority of APOD applied to the scenarios with limited sensitive information.
- APOD needs very few (less than 3% of the dataset) sensitive annotations, and shows comparable mitigation performance to Group DRO (Group DRO requires a fully annotated dataset). This indicates the capacity of APOD for bias mitigation under a limitation of sensitive annotations.
- APOD outperforms LfF which relies on the proxy annotation of sensitive attributes. It indicates that the limited human-annotated sensitive information in our framework is more beneficial than proxy annotations on the entire dataset to bias mitigation.

4.3 Annotation Effectiveness Analysis (RQ2)

In this section, APOD is compared with a semi-supervised method and two state-of-the-art active learning methods to demonstrate that AIS contributes to more informative sensitive annotations for the bias mitigation. Specifically, the **Semi-supervised bias mitigation** (SSBM) works based on a partially annoatated dataset without instance selection during bias mitigation. The active learning-based methods adopt existing selection strategies to replace the AIS in APOD including **POD+Active learning with uncertainty sampling** (POD+AL) (Ren et al., 2020; Coleman et al., 2019), and **POD+Active learning with Core-set Approach** (POD+CA) (Sener & Savarese, 2018).

Baselines Methods. **Vanilla**: The classifier is trained to minimize the cross-entropy loss without bias mitigation. **SSBM**: The semi-supervised bias mitigation initially samples a data subset for annotations via random selection, then adopts POD to debias the classifier on the partially annoatated dataset. **POD+AL**: The AIS in APOD is replaced by active learning with uncertainty sampling, where an instance is selected



(a) MEPS. (b) German credit. (c) Loan default.

(d) Adult. (e) CelebA-wavy hair. (f) CelebA-young.

Figure 4: Accuracy-fairness curve; Algorithm: Vanilla training, SSBM, POD + AL, POD + CA and APOD; Dataset: (a) MEPS, (b) German credit, (c) Loan default, (d) Adult, (e) CelebA-wavy hair, (f) CelebA-young.

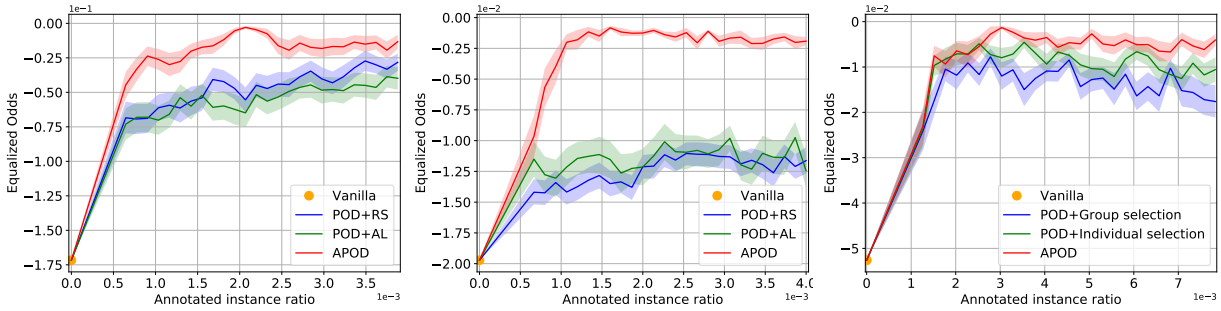
to maximize the Shannon entropy of model prediction. **POD+CA**: AIS is replaced by active learning with core-set approach, where an instance is selected to maximize the coverage of the entire unannotated dataset. More details on the baseline methods are given in the Appendix D.

To unify the experiment condition, all methods have the same annotation budget and have λ in the range of $(0, 2]$. The fairness-accuracy curves on the five datasets are given in Figures 4 (a)-(f), respectively. Based on the mitigation performance, we have the following observations:

- Compared to the semi-supervised method and the active learning-based methods, APOD achieves better mitigation performance at the same level of accuracy, indicating the proposed AIS selects more informative annotations than those methods for bias mitigation.
- Different from POD+AL and POD+CA which sample the annotated instances from the whole dataset in each iteration, APOD interactively selects more representative instances from different subgroups in different iterations, i.e. $\mathcal{U}_a^y$ for $a \in \mathcal{A}$ and $y \in \mathcal{Y}$, which contributes to more effective bias mitigation.
- SSBM shows almost the worst mitigation performance among all of the methods, because the initially randomly selected subset preserves the skewness of the original dataset, leading to non-optimal bias mitigation, which is consistent with our discussion in Section 1.

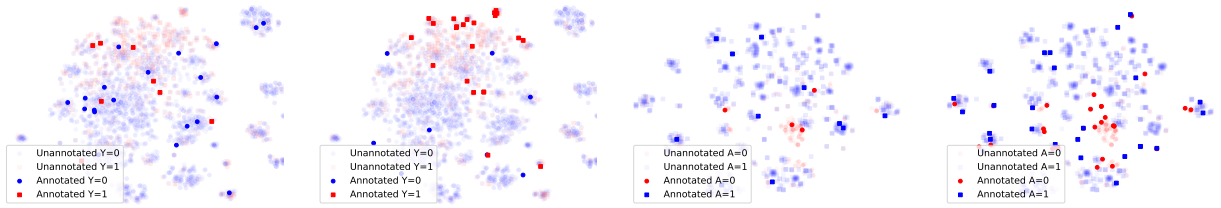
4.4 Annotation Ratio Analysis (RQ3)

We now evaluate the effect of the the annotation ratio (that is, the ratio of the annotated instances to the training instances) on bias mitigation. Specifically, we tune λ in the range of $(0, 2]$, and find that $\lambda = 0.5$ and 0.1 can provide a good accuracy-fairness trade-off on the Adult and Loan default datasets, respectively. In addition to the existing baseline methods, we also consider replacing AIS in APOD into random selection (**POD+RS**) for comparison. Since one instance is selected for annotation in each iteration of APOD, the Equalized Odds of the snapshot model in each iteration is estimated and plotted versus the



(a) Adult. (b) Loan default. (c) MEPS. Ablation result.

Figure 5: Effect of the annotation ratio to APOD, POD+RS and POD+AL on (a) Adult and (b) Loan default dataset. (c) Mitigation performance of APOD, POD+Group selection and POD+Individual selection.



(a) Random selection. (b) APOD. (c) Random selection. (d) APOD.

Figure 6: Comparison of APOD and Random selection in terms of the annotated instances from different groups. (a) Annotated instances by Random selection. (b) Annotated instances by APOD. (c) Annotated positive instances (Y=1) by Random selection. (d) Annotated positive instances (Y=1) by APOD.

annotation ratio on the Adult and Loan default datasets in Figures 5 (a) and (b), respectively. We also give the error bar to show the standard deviation of each method. Overall, we have the following observations:

- All methods achieve better mitigation performance as the annotation ratio increases due to the distribution of the annotated set becoming consistent with the entire dataset.

- APOD shows better mitigation performance than POD+AL and POD+RS at the same level of annotation ratios. This indicates the selection of annotated instances by AIS significantly leads to a reduction of bias. In contrast, the bias mitigation of POD+RS merely derives from the increasing annotations.

- APOD shows significantly higher improvement in bias mitigation than the baseline methods even when the annotation ratio is small, and enables the mitigation to converge to a higher level at smaller annotation ratios (i.e., earlier) than baseline methods.

4.5 Ablation Study (RQ4)

To demonstrate the effectiveness of group selection and individual selection, APOD is compared with two compositional methods: POD+Group selection and POD+Individual selection. The three methods are tested with the same hyperparameter setting on the MEPS dataset. The value of the fairness metric is given in Figure 5 (c). It is observed that both POD+Group selection and POD+Individual selection show considerable degradation in mitigation performance compared to APOD. It empirically validates Remark 1 that both group selection and individual selection in AIS contribute to tightening the upper bound of the relaxed fairness metrics, thus contributing to bias mitigation.

4.6 Visualization of Annotated Instances

We visualize the tSNE embeddings of the annotated instances to trace the active instance selection of APOD. Specifically, Figures 6 (a)-(d) illustrate the the tSNE embeddings of the annotated instances selected by APOD and random selection on the MEPS and Adult datasets, respectively. We use different colors to indicate different groups, where positive instances (Y=1) are less than negative ones (Y=0), and the unprivileged group (A=0) is smaller than the privileged group (A=1). According to Figures 6 (a)-(d), we have the following observations:

- The annotated instances of Random selection in Figures 6 (a) and (c) are consistent with Figure 1 (b), which follows the skewed distribution of original dataset, and leads to non-optimal mitigation of bias.
- The annotated instances of APOD in Figures 6 (b) and (d) are consistent with the optimal annotating in Figure 1 (c), where the annotated subset shows less skewness compared to the original distribution.
- APOD selects more annotated instances from the unprivileged group $\{(x_i, y_i, a_i)|y_i = 1, a_i = 0\}$ than random selection. In such a manner, APOD improves the contribution of unprivileged group to the average loss, which contributes to the bias mitigation.

5 Conclusion

In this paper, we propose APOD, an iterative framework for active bias mitigation under the limitation of sensitive annotations. In each iteration, APOD guides the active instance selection to discover the optimal instance for annotation, and maximally promotes bias mitigation based on the partially annotated dataset through penalization of discrimination. Theoretical analysis indicates that APOD contributes to effective bias mitigation via bounding the relaxed fairness metrics. Experiment results further demonstrate the effectiveness of APOD on five benchmark datasets, where it outperforms baseline methods under the same annotation budget and has a desirable outcome of bias mitigation even when most of the sensitive annotations are unavailable. This also indicates its benefit to real-world applications, especially when the disclosed or available sensitive information is very limited.

Broader Impact Statement

Machine learning systems require a mitigation of algorithmic bias to have fair decisions in the high-stake scenarios, e.g. loan approvals, college admissions, and criminal risk assessments, etc. However, the sensitive annotation may be related with user private information in the high-stake scenarios. It may increase the risk of user privacy being leaked to use a fully annotated dataset for bias mitigation. This work proposes a solution to debias a machine learning algorithm using less 3% sensitive annotations, solving the fairness problem in the high-stake scenarios while protecting the sensitive information of most users. It remains a challenge to take advantage of various modalities of expert knowledge to improve bias mitigation.

References

Hadis Anahideh, Abolfazl Asudeh, and Saravanan T. Fair active learning. *arXiv preprint arXiv:2001.01796*, 2020.

J Angwin, J Larson, S Mattu, and L Kirchner. There's software used across the country to predict future criminals. *ProPublica*, 2016.

Mario Arduini, Lorenzo Noci, Federico Pirovano, Ce Zhang, Yash Raj Shrestha, and Bibek Paudel. Adversarial learning for debiasing knowledge graph embeddings. *arXiv preprint arXiv:2006.16309*, 2020.

Pranjal Awasthi, Alex Beutel, Matthäus Kleindessner, Jamie Morgenstern, and Xuezhi Wang. Evaluating fairness of machine learning models under uncertain and incomplete information. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 206–214, 2021.

Adelchi Azzalini. The skew-normal distribution and related multivariate families. *Scandinavian Journal of Statistics*, 32(2):159–188, 2005.

Adelchi Azzalini and A Dalla Valle. The multivariate skew-normal distribution. *Biometrika*, 83(4):715–726, 1996.

Hyojin Bahng, Sanghyuk Chun, Sangdoo Yun, Jaegul Choo, and Seong Joon Oh. Learning de-biased representations with biased representations. In *International Conference on Machine Learning*, pp. 528–539. PMLR, 2020.

Yahav Bechavod and Katrina Ligett. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044*, 2017.

Rachel KE Bellamy, Kuntal Dey, Michael Hind, Samuel C Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, et al. Ai fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias. *arXiv preprint arXiv:1810.01943*, 2018.

Jesús Bobadilla, Raúl Lara-Cabrera, Ángel González-Prieto, and Fernando Ortega. Deepfair: deep learning for improving fairness in recommender systems. *arXiv preprint arXiv:2006.05255*, 2020.

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

Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? *arXiv preprint arXiv:1805.12002*, 2018.

Jiahao Chen, Nathan Kallus, Xiaojie Mao, Geoffry Svacha, and Madeleine Udell. Fairness under unawareness: Assessing disparity when protected class is unobserved. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 339–348, 2019.

Ching-Yao Chuang and Youssef Mroueh. Fair mixup: Fairness via interpolation. *arXiv preprint arXiv:2103.06503*, 2021.

Steven B Cohen. The medical expenditure panel survey: an overview. *Effective Clinical Practice*, 5(3), 2002.

Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, J. Leskovec, and M. Zaharia. Selection via proxy: Efficient data selection for deep learning. *arXiv preprint arXiv:1906.11829*, 2019.

Elliot Creager, David Madras, Jörn-Henrik Jacobsen, Marissa Weis, Kevin Swersky, Toniann Pitassi, and Richard Zemel. Flexibly fair representation learning by disentanglement. In *International Conference on Machine Learning*, pp. 1436–1445. PMLR, 2019.

Enyan Dai and S. Wang. Say no to the discrimination: Learning fair graph neural networks with limited sensitive attribute information. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*, pp. 680–688, 2021.

Mengnan Du, Ninghao Liu, Fan Yang, and Xia Hu. Learning credible deep neural networks with rationale regularization. In *2019 IEEE International Conference on Data Mining (ICDM)*, pp. 150–159. IEEE, 2019.

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. (2007), 2017a. URL `http://archive.ics.uci.edu/ml`.

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

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.

Pratik Gajane and Mykola Pechenizkiy. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184*, 2017.

Sharad Goel, Justin M Rao, Ravi Shroff, et al. Precinct or prejudice? understanding racial disparities in new york city's stop-and-frisk policy. *Annals of Applied Statistics*, 10(1):365–394, 2016.

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

Tatsunori Hashimoto, Megha Srivastava, Hongseok Namkoong, and Percy Liang. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pp. 1929–1938. PMLR, 2018.

Sangwon Jung, Sanghyuk Chun, and Taesup Moon. Learning fair classifiers with partially annotated group labels. *arXiv preprint arXiv:2111.14581*, 2021.

Nathan Kallus, Xiaojie Mao, and Angela Zhou. Assessing algorithmic fairness with unobserved protected class using data combination. *Management Science*, 2021.

Jon Kleinberg, Jens Ludwig, Sendhil Mullainathan, and Ashesh Rambachan. Algorithmic fairness. In *Aea papers and proceedings*, volume 108, pp. 22–27, 2018.

Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *arXiv preprint arXiv:1703.06856*, 2017.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H Chi. Fairness without demographics through adversarially reweighted learning. *arXiv preprint arXiv:2006.13114*, 2020.

Yi Li and Nuno Vasconcelos. Repair: Removing representation bias by dataset resampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9572–9581, 2019.

Evan Z Liu, Behzad Haghgoo, Annie S Chen, Aditi Raghunathan, Pang Wei Koh, Shiori Sagawa, Percy Liang, and Chelsea Finn. Just train twice: Improving group robustness without training group information. In *International Conference on Machine Learning*, pp. 6781–6792. PMLR, 2021.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*, pp. 3730–3738, 2015.

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.

Junhyun Nam, Hyuntak Cha, Sungsoo Ahn, Jaeho Lee, and Jinwoo Shin. Learning from failure: Training debiased classifier from biased classifier. *arXiv preprint arXiv:2007.02561*, 2020.

Yaniv R., Stephen B., and Emmanuel J C. Achieving equalized odds by resampling sensitive attributes. *arXiv preprint arXiv:2006.04292*, 2020.

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.

Pengzhen Ren, Yun Xiao, Xiaojun Chang, Po-Yao Huang, Zhihui Li, Xiaojiang Chen, and Xin Wang. A survey of deep active learning, 2020.

Shiori Sagawa, Pang Wei Koh, T. B Hashimoto, and P. Liang. Distributionally robust neural networks for group shifts: On the importance of regularization for worst-case generalization. *arXiv preprint arXiv:1911.08731*, 2019.

Ozan Sener and Silvio Savarese. Active learning for convolutional neural networks: A core-set approach, 2018.

Dylan Slack, Sorelle A Friedler, and Emile Givental. Fairness warnings and fair-maml: learning fairly with minimal data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 200–209, 2020.

Emily Steel and Julia Angwin. On the web's cutting edge, anonymity in name only. *The Wall Street Journal*, 4, 2010.

Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. *arXiv preprint arXiv:1906.08976*, 2019.

Latanya Sweeney. Discrimination in online ad delivery. *Communications of the ACM*, 56(5):44–54, 2013.

Suzanne Tolmeijer, Markus Kneer, Cristina Sarasua, Markus Christen, and Abraham Bernstein. Implementations in machine ethics: a survey. *ACM Computing Surveys (CSUR)*, 53(6):1–38, 2020.

Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pp. 1–7. IEEE, 2018.

Serena Wang, Wenshuo Guo, Harikrishna Narasimhan, Andrew Cotter, Maya Gupta, and Michael I Jordan. Robust optimization for fairness with noisy protected groups. *arXiv preprint arXiv:2002.09343*, 2020.

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.

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

Tianxiang Zhao, Enyan Dai, Kai Shu, and Suhang Wang. You can still achieve fairness without sensitive attributes: Exploring biases in non-sensitive features. *arXiv preprint arXiv:2104.14537*, 2021.

Anna Zimdars. Fairness and undergraduate admission: a qualitative exploration of admissions choices at the university of oxford. *Oxford Review of Education*, 36(3):307–323, 2010.

Appendix

A Proof of Theorem 1

In this section, we first propose Corollary 1, then adopt the corollary to prove Theorem 1.

Corollary 1. *For $p(y), f(y), g(y) > 0$, we have*

$$\int_{\mathcal{Y}} p(y)f(y)\mathrm{d}y \leq \int_{\mathcal{Y}} q(y)f(y)\mathrm{d}y + \int_{\mathcal{Y}} |p(y)-q(y)|f(y)\mathrm{d}y$$
$$\int_{\mathcal{Y}} p(y)f(y)\mathrm{d}y \leq \int_{\mathcal{Y}} p(y)g(y)\mathrm{d}y + \int_{\mathcal{Y}} p(y)|f(y)-g(y)|\mathrm{d}y$$

Proof.

$$\begin{aligned}\int_{\mathcal{Y}} p(y)f(y)\mathrm{d}y &= \int_{\mathcal{Y}} q(y)f(y)\mathrm{d}y + \int_{\mathcal{Y}} [p(y)-q(y)]f(y)\mathrm{d}y \\ &\leq \int_{\mathcal{Y}} q(y)f(y)\mathrm{d}y + \int_{\mathcal{Y}} |p(y)-q(y)|f(y)\mathrm{d}y \\ \int_{\mathcal{Y}} p(y)f(y)\mathrm{d}y &= \int_{\mathcal{Y}} p(y)g(y)\mathrm{d}y + \int_{\mathcal{Y}} p(y)[f(y)-g(y)]\mathrm{d}y \\ &\leq \int_{\mathcal{Y}} p(y)g(y)\mathrm{d}y + \int_{\mathcal{Y}} p(y)|f(y)-g(y)|\mathrm{d}y\end{aligned}$$

□

After proving Corollary 1, we return to prove the theorem.

Theorem 1. *Assume the loss value on the training set satisfies $\frac{1}{|\mathcal{S}|}\sum_{(\boldsymbol{x}_i,y_i,a_i)\in\mathcal{S}} l(\boldsymbol{h}_i, y_i; \theta_h) \leq \epsilon$[6], and $l(\boldsymbol{h}, y; \theta_h)$ and f_h satisfy K_l- and K_h-Lipschitz continuity[7], respectively. The generalization loss difference between unprivileged group and privileged group has the following upper bound with probability $1-\gamma$,*

$$\left|\int_{\mathcal{X}_0}\int_{\mathcal{Y}} p(\boldsymbol{x},y)l(\boldsymbol{h},y;\theta_h)\mathrm{d}\boldsymbol{x}\mathrm{d}y - \int_{\mathcal{X}_1}\int_{\mathcal{Y}} p(\boldsymbol{x},y)l(\boldsymbol{h},y;\theta_h)\mathrm{d}\boldsymbol{x}\mathrm{d}y\right| \leq \epsilon + \min\left\{\sqrt{-L^2\log\gamma(2N_{\tilde{a}})^{-1}}, (K_l+K_hL)\delta_{\tilde{a}}\right\}, \tag{14}$$

where $\tilde{a} = \arg\max_{a\in\mathcal{A}} \int_{\mathcal{X}_a}\int_{\mathcal{Y}} p(\boldsymbol{x},y)l(\boldsymbol{h},y;\theta_h)\mathrm{d}\boldsymbol{x}\mathrm{d}y$; $\mathcal{X}_a = \{\boldsymbol{x}_i \in \mathcal{D} | a_i = a\}$; $\delta_{\tilde{a}} = \max_{\boldsymbol{x}_i\in\mathcal{X}_{\tilde{a}}}\min_{(\boldsymbol{x}_j,y_j,a_j)\in\mathcal{S}}||\boldsymbol{h}_i - \boldsymbol{h}_j||_2$; $N_{\tilde{a}} = |\{(\boldsymbol{x}_i,y_i,a_i)|a_i=\tilde{a}, (\boldsymbol{x}_i,y_i,a_i)\in\mathcal{S}\}|$; $L = \max_{(\boldsymbol{x}_i,y_i)\in\mathcal{U}} l(\boldsymbol{h}_i,y_i;\theta_h)$; and $\boldsymbol{h}_i = f_b(\boldsymbol{x}_i|\theta_b)$.

Proof. According to the upper bound of generalization error, the generalization error for group $\boldsymbol{x} \in \mathcal{X}_a$ for $\forall a \in \mathcal{A}$ is bounded with probability $1-\gamma$,

$$g_a = \int_{\mathcal{X}_a}\int_{\mathcal{Y}} p(\boldsymbol{x},y)l(\boldsymbol{h},y;\theta_h)\mathrm{d}\boldsymbol{x}\mathrm{d}y \leq \epsilon + \sqrt{-L^2\log\gamma(2N_a)^{-1}},$$

where $L = \max_{(\boldsymbol{x}_i,y_i)\in\mathcal{U}} l(\boldsymbol{h}_i,y_i;\theta_h)$. Moreover, we consider the upper bound of absolute gap $|g_0 - g_1| \leq \max_{a\in\mathcal{A}} g_a$. The generalization error difference between the two groups is bounded with probability $1-\gamma$ as follow,

$$\left|\int_{\mathcal{X}_0}\int_{\mathcal{Y}} p(\boldsymbol{x},y)l(\boldsymbol{h},y;\theta_h)\mathrm{d}\boldsymbol{x}\mathrm{d}y - \int_{\mathcal{X}_1}\int_{\mathcal{Y}} p(\boldsymbol{x},y)l(\boldsymbol{h},y;\theta_h)\mathrm{d}\boldsymbol{x}\mathrm{d}y\right| \leq \epsilon + \sqrt{-L^2\log\gamma(2N_{\tilde{a}})^{-1}}, \tag{15}$$

where $\tilde{a} = \arg\max_{a\in\mathcal{A}} \int_{\mathcal{X}_a}\int_{\mathcal{Y}} p(\boldsymbol{x},y)l(\boldsymbol{h},y;\theta_h)\mathrm{d}\boldsymbol{x}\mathrm{d}y$.

[6] ϵ is small if the classifier head f_h has been well-trained on the annotated dataset $\mathcal{S}$.

[7] $l(\boldsymbol{h},y;\theta_h)$ and f_h satisfy $|l(\boldsymbol{h}_i,y;\theta_h) - l(\boldsymbol{h}_j,y;\theta_h)| \leq K_l||\boldsymbol{h}_i - \boldsymbol{h}_j||_2$ and $|p(y|\boldsymbol{x}_i) - p(y|\boldsymbol{x}_j)| \leq K_h||\boldsymbol{h}_i - \boldsymbol{h}_j||_2$, respectively, where the likelihood function $p(y \mid \boldsymbol{x}_i) = \mathrm{softmax}(f_h(\boldsymbol{h}_i|\theta_h))$.

To prove the second bound of the generalization error difference, let $\mathcal{N}(\boldsymbol{x}_i)$ denote the nearest neighbour of $\boldsymbol{x}_i \in \mathcal{X}$ which belongs to the annotated dataset, i.e. $\mathcal{N}(\boldsymbol{x}_i) = \arg\min_{(\boldsymbol{x}_j, y_j, a_j)\in\mathcal{S}} ||\boldsymbol{h}_j - \boldsymbol{h}_i||_2$; let $\boldsymbol{h}_i^{\mathcal{N}}$ denote the embedding of $\mathcal{N}(\boldsymbol{x}_i)$; and let $d_{\boldsymbol{x}_i}$ denote the distance between $\boldsymbol{x}_i$ and $\mathcal{N}(\boldsymbol{x}_i)$ in the embedding space, i.e. $d_{\boldsymbol{x}_i} = \min_{(\boldsymbol{x}_j, y_j, a_j)\in\mathcal{S}} ||\boldsymbol{h}_i - \boldsymbol{h}_j||_2$. Accoding to Corollary 1, with $p(y) = p(y|\boldsymbol{x}_i)$, $q(y) = p(y|\mathcal{N}(\boldsymbol{x}_i))$ and $f(y) = l(\boldsymbol{h}_i, y; \theta_h)$, the generalization error can be bounded by

$$\int_{\mathcal{Y}} p(y|\boldsymbol{x}_i) l(\boldsymbol{h}_i, y; \theta_h) \mathrm{d}y \leq \int_{\mathcal{Y}} p(y|\mathcal{N}(\boldsymbol{x}_i)) l(\boldsymbol{h}_i, y; \theta_h) \mathrm{d}y + \int_{\mathcal{Y}} \big|p(y|\boldsymbol{x}_i) - p(y|\mathcal{N}(\boldsymbol{x}_i))\big| l(\boldsymbol{h}_i, y; \theta_h) \mathrm{d}y. \tag{16}$$

Note that the classifier head f_h satisfies K_h-Lipschitz continuity $|p(y|\boldsymbol{x}_i) - p(y|\mathcal{N}(\boldsymbol{x}_i))| \leq K_h ||\boldsymbol{h}_i - \boldsymbol{h}_i^{\mathcal{N}}||_2 = K_h d_{\boldsymbol{x}_i}$ and $l(\boldsymbol{h}, y; \theta_h) \leq L$, the second term in the right-side of Equation (16) is bounded by

$$\int_{\mathcal{Y}} \big|p(y|\boldsymbol{x}_i) - p(y|\mathcal{N}(\boldsymbol{x}_i))\big| l(\boldsymbol{h}_i, y; \theta_h) \mathrm{d}y \leq K_h L d_{\boldsymbol{x}_i}. \tag{17}$$

Furthermore, taking $p(y) = p(y|\mathcal{N}(\boldsymbol{x}_i))$, $f(y) = l(\boldsymbol{h}_i, y; \theta_h)$ and $g(y) = l(\boldsymbol{h}_i^{\mathcal{N}}, y; \theta_h)$ into Corollary 1, we have the first term in the right-side of Equation (16) can be bounded by

$$\int_{\mathcal{Y}} p(y|\mathcal{N}(\boldsymbol{x}_i)) l(\boldsymbol{h}, y; \theta_h) \mathrm{d}y \leq \int_{\mathcal{Y}} p(y|\mathcal{N}(\boldsymbol{x}_i)) l(\boldsymbol{h}_i^{\mathcal{N}}, y; \theta_h) \mathrm{d}y + \int_{\mathcal{Y}} p(y|\mathcal{N}(\boldsymbol{x}_i)) |l(\boldsymbol{h}, y; \theta_h) - l(\boldsymbol{h}_i^{\mathcal{N}}, y; \theta_h)| \mathrm{d}y \leq \epsilon + K_l d_{\boldsymbol{x}_i}, \tag{18}$$

where we have $\int_{\mathcal{Y}} p(y|\mathcal{N}(\boldsymbol{x}_i)) l(\boldsymbol{h}_i^{\mathcal{N}}, y; \theta_h) \mathrm{d}y \leq \epsilon$ due to the upper bound of training error; and we have

$$\int_{\mathcal{Y}} p(y|\mathcal{N}(\boldsymbol{x}_i)) |l(\boldsymbol{h}_i, y; \theta_h) - l(\boldsymbol{h}_i^{\mathcal{N}}, y; \theta_h)| \mathrm{d}y \leq K_l d_{\boldsymbol{x}_i}, \tag{19}$$

due to the K_l-Lipschitz continuity of the loss function.

Taking Equations (17) and (18) into Equation (16), the generalization error on group $\boldsymbol{x} \in \mathcal{X}_a$ can be bounded by

$$\int_{\mathcal{X}_a} \int_{\mathcal{Y}} p(\boldsymbol{x}, y) l(\boldsymbol{h}_i, y; \theta_h) \mathrm{d}y \mathrm{d}\boldsymbol{x} \leq \epsilon + (K_l + K_h L)\delta_a, \tag{20}$$

where $\delta_a = \max_{\boldsymbol{x}_i \in \mathcal{X}_a} d_{\boldsymbol{x}_i} = \max_{\boldsymbol{x}_i \in \mathcal{X}_a} \min_{\boldsymbol{x}_j \in \mathcal{S}} ||\boldsymbol{h}_i - \boldsymbol{h}_j||_2$ denotes the max-min distance between the unannotated and annotated instances in the embedding space. Note that $a \in \mathcal{A}$, we take $\tilde{a} = \arg\max_{a\in\mathcal{A}} \int_{\mathcal{X}_a} \int_{\mathcal{Y}} p(\boldsymbol{x}, y) l(\boldsymbol{h}, y; \theta_h) \mathrm{d}y \mathrm{d}\boldsymbol{x}$. The generalization error difference between the two groups can be bounded by

$$\left| \int_{\mathcal{X}_0} \int_{\mathcal{Y}} p(\boldsymbol{x}, y) l(\boldsymbol{h}, y; \theta_h) \mathrm{d}\boldsymbol{x} \mathrm{d}y - \int_{\mathcal{X}_1} \int_{\mathcal{Y}} p(\boldsymbol{x}, y) l(\boldsymbol{h}, y; \theta_h) \mathrm{d}\boldsymbol{x} \mathrm{d}y \right| \leq \epsilon + (K_l + K_h L)\delta_{\tilde{a}}. \tag{21}$$

Combine Equation (21) with (15), we have the generalization error gap between group $x \in \mathcal{X}_0$ and group $x \in \mathcal{X}_1$ bounded as follow with probability $1 - \gamma$,

$$\begin{aligned} &\left| \int_{\mathcal{X}_0} \int_{\mathcal{Y}} p(\boldsymbol{x}, y) l(\boldsymbol{h}, y; \theta_h) \mathrm{d}\boldsymbol{x} \mathrm{d}y - \int_{\mathcal{X}_1} \int_{\mathcal{Y}} p(\boldsymbol{x}, y) l(\boldsymbol{h}, y; \theta_h) \mathrm{d}\boldsymbol{x} \mathrm{d}y \right| \\ &\qquad \leq \epsilon + \min\left\{ \sqrt{-L^2 \log \gamma (2N_{\tilde{a}})^{-1}}, (K_l + K_h L)\delta_{\tilde{a}} \right\}. \end{aligned}$$

□

B Details about the Datasets

The experiments are conducted on the MEPS[8], Loan default[9], German credit[10], Adult[11] and CelebA[12] datasets to demonstrate the proposed framework is effective to mitigate the socially influential bias such as

[8] https://github.com/Trusted-AI/AIF360/tree/master/aif360/data/raw/meps
[9] https://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients
[10] https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)
[11] http://archive.ics.uci.edu/ml/datasets/Adult
[12] http://mmlab.ie.cuhk.edu.hk/projects/CelebA.html

Table 1: Details about the datasets.

	Adult	MEPS	Loan default	German credit	CelebA
Domain	Social	Medical	Financial	Financial	Social
Data formate	Tabular	Tabular	Tabular	Tabular	Image
Predicted attribute	Salary	Utilization	Defaulting	Credit	Wavy hair, Young
Sensitive attribute	Gender	Race	Age	Age	Gender
Number of instance	30162	15830	30000	4521	5000
Number of attribute	13	41	8	16	160×160
Train, Validate, Test splitting	0.25, 0.25, 0.5	0.25, 0.25, 0.5	0.25, 0.25, 0.5	0.25, 0.25, 0.5	0.25, 0.25, 0.5
Annotation budget	4‰	8‰	4‰	2%	3%

Table 2: Detailed hyper-parameter setting.

	Adult	MEPS	Loan default	German Credit	CelebA-hair	CelebA-young
Classifier body f_b	Perceptron	Perceptron	Perceptron	Perceptron	ResNet-18	ResNet-18
Classifier head f_h	2-layer MLP	3-layer MLP	2-layer MLP	3-layer MLP	3-layer MLP	3-layer MLP
Classifier head f_a	Perceptron	Perceptron	Perceptron	Perceptron	Perceptron	Perceptron
Embedding dim M	64	32	64	32	256	256
Hidden-layer dim	32	32	32	32	64	128

the gender, race or age bias. The statistics of the datasets is given in Table 1. The details about the datasets are described as follows.

- **MEPS**: The task on this dataset is to predict whether a person would have a *high* or *low* utilization based on other features (*region, marriage, etc.*). The *Race* of each person is the sensitive attribute, where the two sensitive groups are *white* and *non-white* (Cohen, 2002). The vanilla model shows discrimination towards the non-white group. The annotation budget is 8‰ [13].

- **Loan default**: The task is to predict whether a person would default the payment of loan based on personal information (*Bill amount, education, etc.*), where the sensitive attribute is *age*, and the two sensitive groups are people *above 35* and those *below 35* (Bellamy et al., 2018). The vanilla trained model shows discrimination towards the younger group. The annotation budget is 4‰.

- **German credit**: The goal of this dataset is to predict whether a person has *good* or *bad* credit risks based on other features (*balance, job, education, etc.*). *Age* is the sensitive attribute, where the two sensitive groups are people *older than 35* and those *not older than 35* (Dua & Graff, 2017b). The vanilla trained model shows discrimination towards the younger group. The annotation budget is 2%.

- **Adult**: The task for this dataset is to predict whether a person has *high* (more than 50K/yr) or *low* (less than 50K/yr) income based on other features (*education, occupation, working hours, etc.*). *Gender* is considered as the sensitive attribute for this dataset (Dua & Graff, 2017a). Thus, we have two sensitive groups *male* and *female*. The vanilla trained classification model shows discrimination towards the female group. The annotation budget is 4‰.

- **CelebA**: This is a large-scale image dataset of human faces (Liu et al., 2015). We consider two tasks for this dataset: i) identifying whether a person has wavy hair; ii) identifying whether a person is young. *Gender* is the sensitive feature, where the two sensitive groups are *male* and *female*. The vanilla trained model shows discrimination towards male in task i) and female in tasks ii), respectively. The annotation budget is 3%.

C Detailed Hyper-parameter Setting

The detailed hyper-parameter setting is given in Table 2.

[13]The cost of annotating 8‰ of training instances is affordable.

D Details about the Baseline Methods

We introduce details on the baseline methods in this section.

- **Group DRO**: Group DRO maintains a distribution $\boldsymbol{q} = [q_0, q_1]$ over the sensitive groups $a \in \mathcal{A}$, and updates the classifier $f(\bullet \mid \theta_f)$ via the min-max optimization given by

$$\theta_f = \arg\min_{\theta} \max_{\boldsymbol{q}} \sum_{a \in \mathcal{A}} \frac{q_a}{N_a} \sum_{(\boldsymbol{x}_i, y_i) \sim \mathcal{D}_a} l(\boldsymbol{x}_i, y_i; \theta), \tag{22}$$

  where $\mathcal{D}_a = \{\boldsymbol{x}_i, y_i \mid a_i = a\}$ depends on fully-annotated training set to generate the sensitive groups.

- **FAL**: Original FAL depends on the annotation of sensitive attribute to have active instance selection. Hence, we consider an improved version of the original framework to adapt to the problem in this work. Specifically, our improved FAL updates the classifier to minimize the cross-entropy loss on the annotated dataset. The annotated instances are selected by

$$(\boldsymbol{x}^*, y^*) = \arg\max_{(x,y) \in \mathcal{U}} \alpha \mathrm{ACC}(f_t) + (1-\alpha)[\mathcal{F}(f_t) - \mathcal{F}(f_{t-1})], \tag{23}$$

  where f_t denotes the classifier learned on the annotated dataset $\mathcal{S}$; $\mathcal{F}(f_t)$ denotes the fairness score of classifier f_t, which is the value of Equalized Odds in our experiment; α controls the trade-off between accuracy and fairness; and we have α in the range of $[0.5, 1]$ in our experiments.

- **LfF**: LfF adopts generalized cross entropy loss to learn the biased model f_B to provide proxy annotation, and simultaneously learn the debiased model f_D towards minimizing the cross entropy re-weighted by the proxy annotation. f_B and f_D are updated by

$$\begin{aligned} \theta_B^* &= \min_{\theta_B} \sum_{i=1}^{N} \frac{1 - p(\boldsymbol{x}_i; \theta_B)^q}{q}, \\ \theta_D^* &= \arg\min_{\theta_D} \sum_{i=1}^{N} \frac{l(\boldsymbol{x}_i, \hat{y}; \theta_B) l(\boldsymbol{x}_i, \hat{y}; \theta_D)}{l(\boldsymbol{x}_i, \hat{y}; \theta_B) + l(\boldsymbol{x}_i, \hat{y}; \theta_D)}, \end{aligned} \tag{24}$$

  where we control the hyper-parameter q in the range of $[2.5, 3]$ in our experiments.

- **SSBM**: This method initially randomly select a subset for annoatation, then adopts POD for the bias mitigation.

- **POD+RS**: Different from SSBM, this method randomly selects an annotated instance and adopts POD for bias mitigation in each iteration. The random instance selection and POD executes iteratively. This method is designed for studying the effect of annotation ratio to the mitigation performance.

- **POD+AL**: This method adopts POD for bias mitigation. Different from APOD, the annotated instances are selected by uncertainty sampling. Specifically, we calculate the Shannon entropy of the model prediction for each instance in the unannotated dataset. For $\boldsymbol{x}_i \in \mathcal{U}$, we have the entropy given by

$$\mathcal{H}(\boldsymbol{x}_i) = -p_{\hat{y}_i=1} \log_2 p_{\hat{y}_i=1} - p_{\hat{y}_i=0} \log_2 p_{\hat{y}_i=0}. \tag{25}$$

  where $[p_{\hat{y}_i=1}, p_{\hat{y}_i=0}] = \mathrm{softmax}[f(\boldsymbol{h}_i|\theta_h)]$; and $f(\boldsymbol{h}_i|\theta_h) \in \mathcal{Y}$. The instance for annotation is selected by

$$(\boldsymbol{x}^*, y^*) = \arg\max_{(\boldsymbol{x}_i, y_i) \in \mathcal{U}} \mathcal{H}(\boldsymbol{x}_i). \tag{26}$$

- **POD+CA**: This method adopts POD for bias mitigation. Different from APOD, POD+CA selects the instance for annotation following the max-min rule given by

$$(\boldsymbol{x}^*, y^*) = \arg\max_{\boldsymbol{x}_i \in \mathcal{U}} \min_{\boldsymbol{x}_j \in \mathcal{S}} ||\boldsymbol{h}_i - \boldsymbol{h}_j||_2, \tag{27}$$

  where $\mathcal{S}$ and $\mathcal{U}$ denote the annotated and unannotated datasets, respectively.