Fairness Without Harm: An Influence-Guided Active Sampling Approach

Jinlong Pang
UC Santa Cruz
jpang14@ucsc.edu

Jialu Wang
UC Santa Cruz
faldict@ucsc.edu

Zhaowei Zhu
Docta.ai
zzw@docta.ai

Yuanshun Yao
Meta GenAI
kevinyao@meta.com

Chen Qian
UC Santa Cruz
cqian12@ucsc.edu

Yang Liu*
UC Santa Cruz
yangliu@ucsc.edu

Abstract

The pursuit of fairness in machine learning (ML), ensuring that the models do not exhibit biases toward protected demographic groups, typically results in a compromise scenario. This compromise can be explained by a Pareto frontier where given certain resources (e.g., data), reducing the fairness violations often comes at the cost of lowering the model accuracy. In this work, we aim to train models that mitigate group fairness disparity without causing harm to model accuracy. Intuitively, acquiring more data is a natural and promising approach to achieve this goal by reaching a better Pareto frontier of the fairness-accuracy tradeoff. The current data acquisition methods, such as fair active learning approaches, typically require annotating sensitive attributes. However, these sensitive attribute annotations should be protected due to privacy and safety concerns. In this paper, we propose a tractable active data sampling algorithm that does not rely on training group annotations, instead only requiring group annotations on a small validation set. Specifically, the algorithm first scores each new example by its influence on fairness and accuracy evaluated on the validation dataset, and then selects a certain number of examples for training. We theoretically analyze how acquiring more data can improve fairness without causing harm, and validate the possibility of our sampling approach in the context of risk disparity. We also provide the upper bound of generalization error and risk disparity as well as the corresponding connections. Extensive experiments on real-world data demonstrate the effectiveness of our proposed algorithm. Our code is available at github.com/UCSC-REAL/FairnessWithoutHarm.

1 Introduction

Machine Learning (ML) has dramatically impacted numerous optimization and decision-making processes across various domains, such as credit scoring [62] and demand forecasting [14]. Algorithmic fairness embraces the principle, often enforced by law and regulations, that the decision-maker should not exhibit biases toward protected group membership [84], identified by characteristics such as race, gender, or disability. However, the pursuit of fairness unavoidably results in a compromise scenario where reducing the fairness violations usually leads to a degradation in accuracy, which has been observed and verified by numerous literature [52, 25, 81, 69, 83]. Theoretically, the phenomenon can be understood through a Pareto frontier on the tradeoff between group fairness and accuracy [72, 51, 10, 74]. That is, as illustrated in Figure 1, given certain resources such as training data, when

*Corresponding Author: Yang Liu

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

a model has reached a point on the Pareto frontier, without more data resources, it is impossible that one can improve fairness without worsening off model accuracy.

One major source of unfairness and a major cause of the fairness-accuracy tradeoff is biased training data. If an unbiased and "fairer" dataset is available, we will be hopeful that unfairness can be alleviated without compromising accuracy. Furthermore, such a "fairer" dataset would allow for obtaining a fair and accurate model through the standard empirical risk minimization (ERM) with cross-entropy (CE) loss. The above observation points to a promising way to improve fairness via actively acquiring more informative data, aiming to shift towards a better Pareto frontier of the fairness-accuracy trade-off [1, 46]. However, existing approaches that seek more data, such as fair active learning [4], typically require annotating sensitive attributes for training data. In practice, these sensitive attribute information such as race and gender, should be protected due to privacy regulations [35, 5, 66]. In the normal active learning scenario, collecting more data with sensitive attributes heightens privacy and safety risks due to the increased probability of leaking sensitive information.

Therefore, we ask the following question: *When not disclosing more annotations of training sensitive attributes, how can we acquire more data to improve model fairness without sacrificing accuracy?*

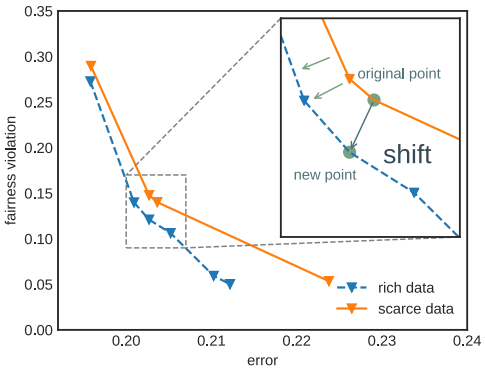


Figure 1: We compare the Pareto frontiers between the model trained with scarce data and that trained with rich data. Acquiring more data is capable of shifting the Pareto frontier toward lower disparity and lower error rates. In consequence, we can reach a new trade-off point that offers improved fairness and accuracy simultaneously, surpassing the original trade-off point.

In this paper, we propose a tractable active data sampling algorithm in a *training sensitive attributes-free* way, which solely requires sensitive attributes on a *small validation set*. Particularly, the algorithm evaluates each example's influence on fairness and accuracy using the validation dataset for ranking and then selects a certain number of examples to supplement the training set for training. We name our solution Fair Influential Sampling (FIS). The core challenge is approximating the corresponding influences of each new example without accessing its sensitive attributes. Technically, we evaluate the importance (influences) of each new example by comparing its gradient to that derived from the entire validation set. This comparison helps quantify the hypothesized change of group fairness disparity metric when adding this example to the training set. As a result, the requirement of training sensitive attributes can be relaxed, as gradient derivation serves as a role of fairness constraints to measure the group fairness disparity. The main contributions of our work are summarized as follows.

- We develop a tractable active data sampling algorithm (Algorithm 1) that does not rely on training sensitive attributes. The algorithm scores each new example based on the combined influences of prediction and fairness and then opts for a certain number of examples for training. [Section 4]
- We theoretically analyze how acquiring more data can improve fairness without harm from a distribution shift perspective view, and validate the possibility of our sampling approach in the context of risk disparity. We also provide the upper bound of generalization error and risk disparity as well as the corresponding connections (Theorem 5.1 and Theorem 5.2). [Section 5]
- Empirical experiments on real-world datasets (CelebA, Adult, and COMPAS) substantiate our claims, indicating the effectiveness and potential of our proposed algorithm in achieving fairness for ML classifiers. [Section 6]

2 Related work

Fairness-accuracy tradeoff There are numerous works that have been successful at mitigating fairness disparities [24, 32, 2, 78, 69]. However, these works typically rely on protected sensitive attributes of training examples to measure the fairness disparities across groups. Moreover, a fairness-

accuracy tradeoff has been shown, meaning that enforcing fair constraints heavily degrades the model performance [52, 25, 81]. Notably, Chen et al. [20] characterized the change of the fairness violation when the data distribution is shifted. Except for training sensitive attributes, this paper does not work in the classical regime of the fairness-accuracy tradeoff. By properly collecting new data, we can improve both accuracy and fairness, which cannot be achieved by working on a static training dataset that naturally incurs such a tradeoff. Besides, compared to prior works [52, 56], our method does not require additional assumptions about the classifier and the characteristics of the training/testing datasets (e.g., distribution shifts). Relevant work [42] utilizes the influence function to reweight the data examples but requires re-training. Our method focuses on soliciting additional samples from an external dataset while [42] reweights the existing and fixed training dataset.

Active learning The core idea of active learning is to rank unlabeled instances by developing specific measures, including uncertainty [41, 45], representativeness [22], inconsistency [70], variance [34], and error [59]. A related line of work [48, 71, 31, 76] concentrates on ranking unlabeled instances based on the influence function. Compared to these studies with a focus on prediction performance, our work poses a distinct challenge taking into account fairness violations. Our approach is more closely with the *fair active learning* approach [4]. However, this framework still relies on training sensitive attributes and then unavoidably encounters the tradeoff between fairness and accuracy.

Fair classifiers without demographics There are various studies to achieve fairness without demographics. For example, Zhao et al. [82] explores the correlations between sensitive attributes and non-sensitive attributes to learn fair and accurate classifiers. Yan et al. [77] investigates the class imbalance problem with a KNN-based pre-processing method. Chai et al. [16] utilizes the soft labels from an overfitting teacher model to train a student model to avoid using demographics. A line of research establishes theoretical connections between features and attributes to avoid using demographic information, employing methods like causal graphs [63], correlation shifts [58], and demographic shifts [30]. In contrast, our approach refrains from making assumptions. Another line of work utilizes distributionally robust optimization (DRO) to reduce fairness disparity without relying on training sensitive attributes [33, 38, 47, 40, 67, 64]. Although these works evaluate the worst-case group performance in the context of fairness, their approaches differ as they do not strive to equalize the loss across groups. Besides, in these studies, accuracy and worst-case accuracy are used as fairness metrics to showcase the efficacy of the proposed algorithms. However, these fairness metrics are restrictive and inconsistent with common definitions such as demographic parity (DP).

Fair classification The fairness-aware learning algorithms, in general, can be categorized into pre-processing, in-processing, and post-processing methods. Pre-processing methods typically reweigh or distort the data examples to mitigate the identified biases [7, 9, 65, 60, 15, 17, 80, 18, 68]. More relevant to this work is the *importance reweighting*, which assigns weights to training examples [37, 36, 23, 21, 57, 44]. Our algorithm bears similarity to a specific case of importance reweighting, particularly the 0-1 reweighting applied to newly added data. Other parallel studies utilize importance weighting to learn a complex generative model in a weakly supervised setting [23, 21], or to mitigate representation bias in training datasets [44]. Post-processing methods typically enforce fairness on a learned model through calibration [26, 27, 32], but these work might not achieve the best fairness-accuracy tradeoff [75, 55]. In contrast, these post-processing works still require sensitive attributes during the inference phase. Recent work [19] develops a *bias score* classifier that operates independently of sensitive attributes; however, it is constrained to binary classifications.

3 Preliminaries

Problem setup We consider a standard K-class classification task whose training (test) data distribution is $\mathcal{P}$ ($\mathcal{Q}$). Let $P := \{z_n\}_{n=1}^{|P|}$ represent the *training dataset* following distribution $\mathcal{P}$, where $|P|$ denotes the corresponding sample size. Each example, denoted as $z_n := (x_n, y_n)$, comprises two random variables: the *feature vector* x and the *label* y. The model trained on P is evaluated by the *test dataset* $Q := \{z^\circ\}_{n=1}^{|Q|}$, where $(\cdot)^\circ$ denotes that the data follows distribution $\mathcal{Q}$, each example $z_n^\circ := (x^\circ, y^\circ, s^\circ)$, and the sensitive group s_n° often refers to characteristics such as race, gender, etc. To align the fairness requirements on the test set Q with the model trained on P, a

popular way is to exploit the sensitive attributes s [24, 72] or their proxies [84] in P and use them to formulate Lagrangians during training. However, extending these approaches to the active learning setting would require disclosing more sensitive attributes during sampling and training [4], which *contradicts* our goal. To avoid disclosing more sensitive attributes, we align the fairness requirements by a small hold-out *validation dataset* $Q_v := \{z_n^\circ\}_{n=1}^{|Q_v|}$ that is independent and identically distributed (IID) with the test set Q. We defer more technical details to Section 4.

Following the general active learning setting [4], we would acquire new examples from a large *unlabeled dataset* $U := \{x'_n\}_{n=1}^{|U|}$ within a limited labeling budget $B(\ll |U|)$ [48, 71, 31]. Denote the solicited example by $z'_n := (x'_n, y'_n)$, where y'_n is the ground-truth label. Note that the protected sensitive attributes from datasets P and U remain undisclosed during sampling and training. In this paper, we aim to incrementally update a model that was initially trained on P using standard ERM, by incorporating newly solicited data z'_n, such that the model can improve fairness without worsening model accuracy. Thus, the core challenge is efficiently determining new examples that induce a significantly better Pareto frontier. In the proceeding section 4, we shall delve into how to acquire new data from unlabeled set.

Fairness definition Note that this work focuses solely on active sampling to build a fairer dataset, which is then used to train the model through standard ERM with Cross-Entropy (CE) loss. Without relying on additional assumptions about the model or training/testing dataset, an intuitive and natural approach is to analyze the expected risk. Therefore, we introduce the concept of *risk disparity* as an intermediate-term for theoretical analysis of fairness.

Definition 3.1 (Risk disparity [33, 79, 3])**.** *Define Q_k as the sub-distribution of Q corresponding to group k. Given the optimized model parameters $\mathbf{w}^P$ trained on set P, risk disparity is defined as: $\mathcal{R}_{Q_k}(\mathbf{w}^P) - \mathcal{R}_{\mathcal{Q}}(\mathbf{w}^P)$, where $\mathcal{R}_{\mathcal{Q}}(\mathbf{w}) := \mathbb{E}_{z\sim\mathcal{Q}}[\ell(\mathbf{w}, z)]$ denotes expected risk induced on target distribution $\mathcal{Q}$.*

Definition 3.1 naturally quantifies the discrepancy in a trained model's performance between a specific group set Q_k and the entire test set Q. That is, a model can be deemed fair if it exhibits consistent performance for a group set Q_k as compared to the test dataset Q. In settings such as face or speech recognition, this fairness definition implies the necessity for all demographic groups to receive the same quality service [13]. For completeness, we also include two well-known definitions of fairness:

Definition 3.2 (Demographic Parity (`DP`))**.** *A classifier f adheres to demographic parity concerning the sensitive attribute s if: $\mathbb{E}[f(\mathbf{w}, x)] = \mathbb{E}[f(\mathbf{w}, x)|s]$.*

Definition 3.3 (Equalized Odds (`EOd` [32])**.** *A classifier f meets the equalized odds with respect to the sensitive attribute s if: $\mathbb{E}[f(\mathbf{w}, x)|y] = \mathbb{E}[f(\mathbf{w}, x)|y, s]$.*

Even though there may be a general incompatibility between risk disparity and popular group fairness metrics like DP and EOd, under the criteria of the proposed fairness notion, these definitions could be encouraged [61, 33]. More details and proof can be found in the Appendix B.

Proposition 3.1. *(Informal) Under appropriate conditions, the risk disparity can serve as a lower bound for fairness disparities based on common fairness definitions, such as DP and EOd.*

Remark 3.1 (Connections to other fairness definitions)**.** *Definition 3.1 targets group-level risk fairness, which has similar granularity to other fairness notions such as accuracy parity [78], device-level parity [43], small accuracy loss for groups [79, 10, 50, 33], and bounded group loss [3].*

4 Improving fairness without harm via data influential sampling

In this section, we first introduce how to measure the importance (influence) of each example on accuracy and fairness without using the corresponding sensitive attributes, respectively. Then, we propose an influence-guided sampling algorithm that actively acquires new data based on the influences for further training.

4.1 Finding influential examples

To avoid using training sensitive attributes, our primary idea is to find newly acquired data that assists in creating a "fairer" dataset, which allows for training a fair and accurate model via standard ERM.

Initially, we explore whether newly acquired data enhances fairness by examining the training process, where the model is typically updated using gradient descent. The change of model parameters by performing one step gradient descent on newly acquired data z' is

$$\mathbf{w}_{t+1} = \mathbf{w}_t - \eta \cdot \partial_{\mathbf{w}_t} \ell(\mathbf{w}_t, z') \tag{1}$$

where η refers to the learning rate and $\ell(\cdot)$ is the training loss function. It should be noted that before we solicit the true labels of samples z', we first use proxy labels. In the following subsection 4.2, we will present a strategy for proxy labels. Training on z' affects the model's prediction on validation data z_n° regarding both accuracy and fairness. If the updated model $\mathbf{w}_{t+1}$ outperforms the previous one $\mathbf{w}_t$ evaluated on the validation dataset in terms of fairness and accuracy, this acquired data z' helps to reduce the fairness disparity without worsening accuracy.

To separately measure the accuracy and fairness performance of the updated model on the validation set, we introduce two types of loss functions: **fairness loss** $\underline{\phi(\mathbf{w}, z_n^\circ)}$ and **accuracy loss** $\underline{\ell(\mathbf{w}, z_n^\circ)}$, where validation data $z_n^\circ = (x_n^\circ, y_n^\circ, s_n^\circ)$. Note that these loss functions are developed for sampling, not for training. Besides, training loss function $\ell(\cdot)$ can be reused as the accuracy loss function due to the same update target. One can identify training loss $\ell(\cdot, z')$ and accuracy loss $\ell(\cdot, z_n^\circ)$ based on the input data used. Without loss of generality, we assume that $\ell(\cdot)$ and $\phi(\cdot)$ are differentiable w.r.t. $\mathbf{w}$. Here, we do not restrict the generality of the fairness loss function; it can be any smoothed version of fairness metrics such as DP or EOd. Following this, we develop the influence of the accuracy and fairness components for finding the samples, respectively.

Influence of accuracy component When model parameters are updated from $\mathbf{w}_t$ to $\mathbf{w}_{t+1}$ by adding a new example z', the influence of model's accuracy on one validation example z_n° is:

$$\mathsf{Infl}_{\mathsf{acc}}(z', z_n^\circ; \mathbf{w}_t, \mathbf{w}_{t+1}) := \ell(\mathbf{w}_{t+1}, z_n^\circ) - \ell(\mathbf{w}_t, z_n^\circ).$$

For ease of notation, we use $\mathsf{Infl}_{\mathsf{acc}}(z', z_n^\circ)$ to represent $\mathsf{Infl}_{\mathsf{acc}}(z', z_n^\circ; \mathbf{w}_t, \mathbf{w}_{t+1})$. By applying first-order Taylor expansion, we obtain the following closed-form statement:

Lemma 4.1. *The accuracy influence of new example z' on the validation dataset Q_v is:*

$$\mathit{Infl}_{\mathit{acc}}(z') := \sum\nolimits_{n \in |Q_v|} \mathit{Infl}_{\mathit{acc}}(z', z_n^\circ) \approx -\eta \sum\nolimits_{n \in |Q_v|} \langle \partial_{\mathbf{w}_t} \ell(\mathbf{w}_t, z'), \partial_{\mathbf{w}_t} \ell(\mathbf{w}_t, z_n^\circ) \rangle \tag{2}$$

Intuitively, the more negative $\mathsf{Infl}_{\mathsf{acc}}(z')$ is, the more positive the model accuracy (performance) that example z' can provide.

Influence of fairness component When model parameters are updated from $\mathbf{w}_t$ to $\mathbf{w}_{t+1}$ by adding a new example z', the influence of model's fainess on one validation example z_n° is:

$$\mathsf{Infl}_{\mathsf{fair}}(z', z_n^\circ; \mathbf{w}_t, \mathbf{w}_{t+1}) := \phi(\mathbf{w}_{t+1}, z_n^\circ) - \phi(\mathbf{w}_t, z_n^\circ). \tag{3}$$

For simplicity, we write $\mathsf{Infl}_{\mathsf{fair}}(z', z_n^\circ; \mathbf{w}_t, \mathbf{w}_{t+1})$ as $\mathsf{Infl}_{\mathsf{fair}}(z', z_n^\circ)$. Then, similarly, we have:

Lemma 4.2. *The fairness influence of new example z' on the validation dataset Q_v is:*

$$\mathit{Infl}_{\mathit{fair}}(z') := \sum\nolimits_{n \in |Q_v|} \mathit{Infl}_{\mathit{fair}}(z', z_n^\circ) \approx -\eta \sum\nolimits_{n \in |Q_v|} \langle \partial_{\mathbf{w}_t} \ell(\mathbf{w}_t, z'), \partial_{\mathbf{w}_t} \phi(\mathbf{w}_t, z_n^\circ) \rangle \tag{4}$$

Similar to the accuracy component, the greater the negativity of $\mathsf{Infl}_{\mathsf{fair}}(z')$ is, the greater the positive impact that the example z' has on fairness.

Intuitions These two components evaluate the accuracy and fairness impact of each example by comparing the gradient originating from a single data sample with the gradient derived from the entire validation set, respectively. This comparison helps quantify the potential advantage of including this specific example in training. For instance, if the gradient obtained from one example has a similar direction to the gradient from the validation set, it indicates that incorporating this example contributes to enhancing the model's fairness or accuracy.

Algorithm 1 Fair influential sampling (FIS)

Input: training set P, unlabeled set U, validation set Q_v, new acquired set $S_t = \{\}, \forall t \in [T]$ rounds, number of new selected examples in each round r, tolerance ϵ.
Warmup: Train a classifier f solely on P by minimizing the empirical risk R_ℓ. Obtain model parameters $\mathbf{w}_1$ and validation accuracy (on Q_v) VAL_0.
for t **in** $\{1, 2, \cdots, T\}$ **do**
  Guess proxy label $\hat{y}'$ for new examples $\hat{z}'$ using Eq. (5).
  Compute the influence of accuracy and fairness component using Eq. (2) and Eq. (4):
    $$S_{\text{original}} = \{\mathsf{Infl}_{\mathsf{fair}}(\hat{z}') \mid \mathsf{Infl}_{\mathsf{acc}}(\hat{z}') \leq 0, \mathsf{Infl}_{\mathsf{fair}}(\hat{z}') \leq 0, \hat{z}' \in U\}$$
  while $|S_t| < r$ **do**
    Find top-$(r - |S_t|)$ annotated examples $\hat{z}'_n$ based on the lowest fairness influence and then inquire about true labels y':
      $$\{z'_n\} \xleftarrow{\text{inquiring}} \{\hat{z}'_n\} \leftarrow \mathsf{Top}\text{-}(r - |S_t|)(S_{\text{original}})$$
    $S_t \leftarrow S_t \cup \{z'_n \mid \mathsf{Infl}_{\mathsf{acc}}(z'_n) \leq 0, \mathsf{Infl}_{\mathsf{fair}}(z'_n) \leq 0\}$
    $U \leftarrow U \cap S_t; \quad S_{\text{original}} \leftarrow S_{\text{original}} \cap S_t$
  end while
  Continue to train the model f on the set S_t via standard ERM. Obtain the updated model parameters $\mathbf{w}_{t+1}$. If the model's validation accuracy (on Q_v) VAL_t does not meet the desired threshold VAL_0, reject the updated model.
end for
Output: models $\{\mathbf{w}_t \mid \text{VAL}_t > \text{VAL}_0 - \epsilon\}$

Training sensitive attributes are not disclosed One can easily check that neither the influence of accuracy nor fairness components require the sensitive attributes of any example z', as the example z' only appears in the first-order gradient of the accuracy loss $\partial_{\mathbf{w}_t} \ell(\mathbf{w}_t, z')$. In the fairness component, calculating the $\partial_{\mathbf{w}_t} \phi(\mathbf{w}_t, z_n^\circ)$ only relies on validation example z_n°'s sensitive attributes. Here, we also validate how accurate the first-order estimation of the influence is in comparison to the real influence [39], and find that the estimated influences for most of the examples are very close to their actual influence values. We refer the readers to Appendix C.1 for more details.

Even without disclosing training sensitive attributes, the correlations between non-demographic features and demographic information may still lead to privacy leakage issues [82]. To address this potential privacy concern, we provide further discussions and theoretical analysis using differential privacy in Appendix F.

4.2 Algorithm: fair influential sampling (FIS)

Following Lemma 4.1 and Lemma 4.2, we can efficiently select those examples with the most negative fairness influence and negative accuracy influence. This sampling method aids in reducing fairness disparities without worsening model accuracy.

Labeling Before presenting our sampling algorithm, it is necessary to address the problem of not accessing the true labels of new solicited examples. Lacking the label information for new examples poses a challenge in determining the corresponding influence on accuracy and fairness, a fact that is substantiated by Lemma 4.1 and Lemma 4.2. Intuitively, one can always recruit human annotators to get the ground-truth labels for those unlabeled examples. However, it is impractical due to the limited labeling budgets. To tackle this problem, another common approach is utilizing a model that has been effectively trained on dataset P to produce proxy labels, which approximate the calculation of influences for examples from a substantial unlabeled dataset U. It's important to note that these proxy labels are exclusively used during the sampling phase. To maintain good model performance, we still need to inquire about the true labels of the selected data examples for subsequent training. Here, we propose to annotate the proxy labels with the model trained on the labeled set P. In particular, we introduce a strategy that employs lowest-influence labels for annotating label $\hat{y}'$ given x':

$$\hat{y}' = \underset{k \in \{1, \cdots, K\}}{\arg\min} \; |\mathsf{Infl}_{\mathsf{acc}}(x', k)|, \tag{5}$$

Here, we denote $\hat{z}' := (x', \hat{y}')$ for the proxy labels.

Proposed algorithm The full procedure is outlined in Algorithm 1. Note that the tolerance ϵ is applied to monitor the performance drop in validation accuracy. In Line 2, we initiate the process by training a classifier f solely on dataset P, that is, performing a warm start. Subsequently, T-round sampling iterations are applied to acquire more examples to dataset P. Following the iterative fashion, FIS guesses labels using Eq. (5) in Line 4. Then, we calculate the scores for proxy examples based on the accuracy and fairness influence using Eq. (2) and Eq. (4) respectively. In Lines 6-12, we would opt for r samples based on the influence scores, and inquire about the true labels of these examples. However, due to the gap between proxy labels $\hat{y}'$ and true label y, the accuracy and fairness influences of the top-r samples based on inquired true labels z' may not necessarily satisfy the same conditions ($\mathsf{Infl}_{\mathsf{acc}}(\hat{z}') \leq 0, \mathsf{Infl}_{\mathsf{fair}}(\hat{z}') \leq 0$). Therefore, we use a while loop to iteratively select the top-$(r - |S_t|)$ examples for labeling until we obtain r samples whose fairness influences based on true labels meet the conditions. Subsequently, in Line 11, we update sets U and S_{original} to prevent duplicate sampling. In Line 13, we would continue training using new examples with true inquired labels from set S_t. We save the model parameters at each round as checkpoints $\mathbf{w}_t$. To avoid potential accuracy drops incurred by excessively large random perturbations, we exclusively choose and offer models for output whose validation accuracy exceeds the initial validation accuracy VAL_0. Although we propose a specific strategy for guessing labels, our algorithm is flexible and compatible with other labeling methods. A comparative analysis of computational costs is detailed in Appendix C.2.

5 How more data improves fairness without harm?

In general, acquiring new data to supplement the original training dataset would potentially raise the distribution shift problem, affecting both accuracy and fairness. In this section, from a distribution shift perspective view, we first present a generalization error bound (accuracy side, Theorem 5.1) and risk disparity bound (fairness side, Theorem 5.2). The theoretical results jointly provide a high-level key insight that controlling the negative impact of distribution shift on generalization error, which refers to the model accuracy, could allow for improving fairness without harm. This theoretical insight validates the possibility of our sampling approach.

Without loss of generality, we discretize the whole distribution space and suppose that the train/test distributions are both drawn from a series of component distributions $\{\pi_1, \cdots, \pi_I\}$ [28]. Then, the empirical risk $\mathcal{R}_P(\mathbf{w})$ calculated over a training set P can be reformulated by splitting samples based on the component distributions:

$$\mathcal{R}_P(\mathbf{w}) := \mathbb{E}_{z \in P}[\ell(\mathbf{w}, z)] = \sum\nolimits_{i=1}^{I} p^{(P)}(\pi = i) \cdot \mathbb{E}_{z \sim \pi_i}[\ell(\mathbf{w}, z)].$$

where $p^{(P)}(\pi = i)$ represents the frequencies of examples in P drawn from component distribution π_i. Then, we can define the measure of probability distance between two sets or distributions as $\text{dist}(\mathcal{A}, \mathcal{B}) := \sum_{i=1}^{I} |p^{(\mathcal{A})}(\pi = i) - p^{(\mathcal{B})}(\pi = i)|$. To reflect the implicit unfairness in the models, we introduce two basic assumptions in convergence analysis [43].

Assumption 5.1 (L-Lipschitz Continuous)**.** *There exists a constant $L > 0$, for any $\mathbf{v}, \mathbf{w} \in \mathbb{R}^d$, $\mathcal{R}_P(\mathbf{v}) \leq \mathcal{R}_P(\mathbf{w}) + \langle \nabla \mathcal{R}_P(\mathbf{w}), \mathbf{v} - \mathbf{w} \rangle + \frac{L}{2} \|\mathbf{v} - \mathbf{w}\|_2^2$.*

Assumption 5.2 (Bounded Gradient on Random Sample)**.** *The stochastic gradients on any sample z are uniformly bounded,* i.e.*, $\mathbb{E}[||\nabla \mathcal{R}_P(\mathbf{w}_t, z)||^2] \leq G^2$, and training epoch $t \in [1, \cdots, T]$.*

Analogous to Assumption 5.2, we further make a mild assumption to bound the loss over the component distributions π_i according to the corresponding model, that is, $\mathbb{E}_{z \sim \pi_i}[\ell(\mathbf{w}^P, z)] \leq G_P, \forall i \in I$, where G_P is a bounding constant. For completeness, we first analyze the upper bound of generalization error, specifically from the standpoint of distribution shifts. Omitted proof can be found in Appendix D.

Theorem 5.1 (Generalization error bound)**.** *Let dist$(\mathcal{P}, \mathcal{Q})$, G_P be defined therein. With probability at least $1 - \delta$ with $\delta \in (0, 1)$, the generalization error bound of the model trained on dataset P is*

$$\mathcal{R}_{\mathcal{Q}}(\mathbf{w}^P) \leq \underbrace{G_P \cdot \mathit{dist}(\mathcal{P}, \mathcal{Q})}_{\textit{distribution shift}} + \sqrt{\frac{\log(4/\delta)}{2|P|}} + \mathcal{R}_P(\mathbf{w}^P).$$

Note that the generalization error bound is predominantly impacted by the shift in distribution, especially when we consider an overfitting model, i.e., the empirical risk $\mathcal{R}_P(\mathbf{w}^P) \to 0$.

Theorem 5.2 (Upper bound of risk disparity). *Suppose $\mathcal{R}_{\mathcal{Q}}(\cdot)$ follows Assumption 5.1. Let dist$(\mathcal{P}, \mathcal{Q})$, G_P, dist$(\mathcal{P}_k, \mathcal{Q}_k)$ and dist(P_k, P) be defined therein. The initial learning rate η_0 satisfies $\eta_0^2 < \frac{1}{\sqrt{2}TL}$, where T is the number of training epochs. Given model $\mathbf{w}^P$ and $\mathbf{w}^k$ trained exclusively on group k's data P_k, with probability at least $1-\delta$ with $\delta \in (0,1)$, then the upper bound of risk disparity is*

$$\mathcal{R}_{\mathcal{Q}_k}(\mathbf{w}^P) - \mathcal{R}_{\mathcal{Q}}(\mathbf{w}^P) \leq \underbrace{G_k \cdot dist(\mathcal{P}_k, \mathcal{Q}_k) + G_P \cdot dist(\mathcal{P}, \mathcal{Q})}_{\text{distribution shift}} + \underbrace{4L^2G^2 \cdot dist(P_k, P)^2}_{\text{group gap}} + \Upsilon \quad (6)$$

where $\Upsilon = \sqrt{\frac{\log(4/\delta)}{2|P|}} + \sqrt{\frac{\log(4/\delta)}{2|P_k|}} + \varpi + \varpi_k$. Note that $\mathbb{E}_{z\sim\pi_i}[\ell(\mathbf{w}^k, z)] \leq G_k$, $\varpi = \mathcal{R}_P(\mathbf{w}^P) - \mathcal{R}^*_{\mathcal{Q}}(\mathbf{w}^{\mathcal{Q}})$ and $\varpi_k = \mathcal{R}_{P_k}(\mathbf{w}^k) - \mathcal{R}^*_{\mathcal{Q}_k}(\mathbf{w}^{\mathcal{Q}_k})$. ϖ and ϖ_k can be regarded as constants because $\mathcal{R}_P(\mathbf{w}^P)$ and $\mathcal{R}_{P_k}(\mathbf{w}^k)$ correspond to the empirical risks, $\mathcal{R}^*_{\mathcal{Q}}(\mathbf{w}^{\mathcal{Q}})$ and $\mathcal{R}^*_{\mathcal{Q}_k}(\mathbf{w}^{\mathcal{Q}_k})$ represent the ideal minimal empirical risk of model $\mathbf{w}^{\mathcal{Q}}$ trained on distribution $\mathcal{Q}$ and $\mathcal{Q}_k$, respectively.

Interpretations of Theorem 5.2 Eq. (6) illustrates several aspects that induce unfairness. (1) *Group biased data*. For group-level fairness, the more balanced the data is, the smaller the risk disparity would be; (2) *distribution shift*. For source/target distribution, the closer the distributions are, the smaller the performance gap would be; (3) *Data size*. For training data size, a larger data size (potentially eliminating data bias across groups) would lead to a smaller performance gap.

Main observation Theorem 5.1 underscores how the generalization error is impacted by distribution shifts. Theorem 5.2 implies that the risk disparity is essentially influenced by the distribution shift and the inherent group gap term. In practice, approaches that mitigate the group gap, such as imposing fairness regularizers, acquiring new data, or reweight the training data samples [37], will inevitably incur additional distribution shifts between the training and test data. The incurred distribution shift further leads to a performance drop due to the generalization error in Theorem 5.1. Nonetheless, one theoretical insight is that if one can control the negative impacts of potential distribution shifts through generalization error while implementing fairness-enhancing strategies, it is possible to achieve the goal of improving fairness without causing harm. This high-level insight supports the effectiveness of our proposed sampling approach, in which we acquire new data to reduce the group gap through fairness components while preventing the potential adverse impacts of distribution shifts using the accuracy influence component.

6 Empirical results

In this section, we empirically demonstrate the disparate impact across groups and present the effectiveness of the proposed Fair Influential Sampling method to mitigate the disparity.

6.1 Experimental setup

We evaluate the performance of our algorithm on three real-world datasets across three different modalities: CelebA [49], UCI Adult [8] and Compas [6]. We implement the fairness loss $\phi(\cdot)$ based on three common group fairness metrics: difference of demographic parity (`DP`), difference of equality of opportunity (`EOp`), and difference of equal odds (`EOd`). We compare our method with five baselines: 1) Base (ERM): directly train the model on the training dataset P; 2) Random: train the model on dataset P and randomly sampled data from U with inquired true labels; 3) BALD [12]: active sampling according to the mutual information; 4) ISAL [48]: selects unlabeled examples based on the calculated influence in an active learning setting. We apply model predictions as pseudo-labels; 5) Just Train Twice (JTT) [47]: reweighting those misclassified examples for re-training. Here, we examine a weight of 20 for misclassified examples, marked as JTT-20. Recall that we present the average result of the classifier $\mathbf{w}_t$ outputs from Algorithm 1. The general term "*fairness violation*" is utilized to quantify the absolute differences based on fairness metrics, such as DP and EOd. More details on datasets and hyper-parameters are provided in Appendix E.1.

6.2 Main results

Note that all the experimental results presented subsequently are from three independent trials, each conducted with distinct random seeds. We present the primary results as tuples in the form

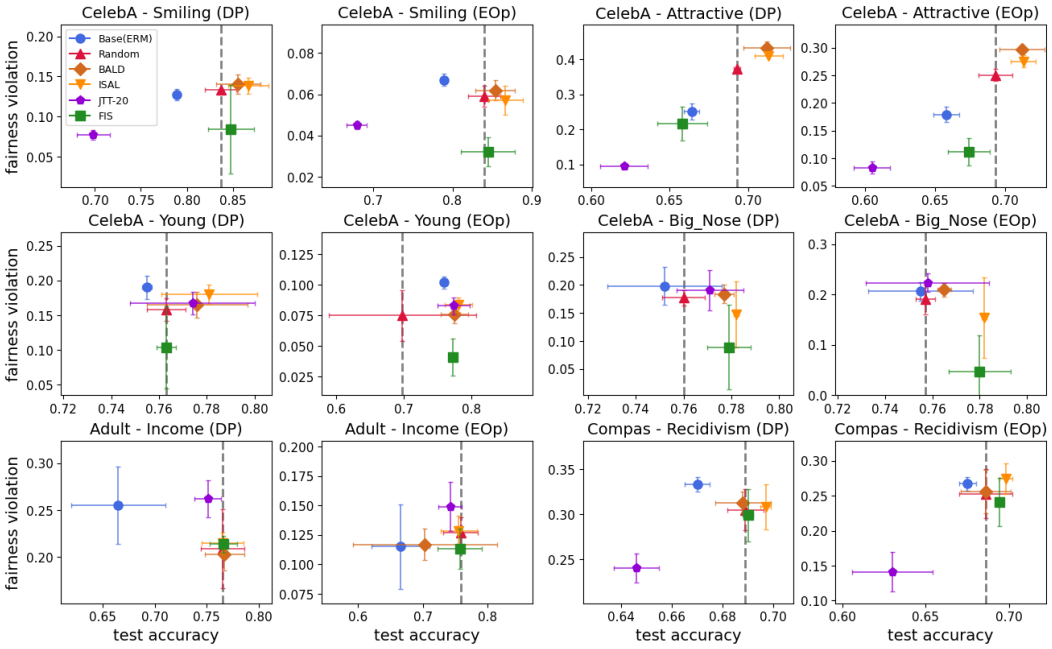


Figure 2: Main results on CelebA, Adult and Compas datasets. The Y axis shows `fairness_violation`; X axis denotes `test_accuracy`. **CelebA:** Four binary targets: `Smiling`, `Attractive`, `Young`, and `Big_Nose`; Sensitive attribute: `gender`. **Adult**: Binary target: `Income`; Sensitive attribute: `Age`. **Compas**: Binary target: `Recidivism`; Sensitive attribute: `Race`. We select two fairness metrics `DP` and `Eop` to measure fairness violations for each setting. The vertical dotted line at the random baseline accuracy helps easily identify which results achieve fairness without sacrificing performance (accuracy).

(`test_accuracy`, `fairness_violation`) to facilitate comparison of the fairness-accuracy tradeoff. Due to space limits, we provide a full version of the experimental results (tables) in Appendix E.2.

Results on image datasets Initially, we train a vision transformer using a patch size of $(8, 8)$ on the CelebA face attribute dataset [49]. We select four binary classification targets, including `Smiling`, `Attractive`, `Young`, and `Big Nose`. The sensitive attribute is `gender`. 2% of the labeled data is allocated for training, while the remaining 98% is reserved for sampling purposes. Then, the test dataset is split into two independent portions: a new test set and a validation set, with 10% of the test data randomly designated as the hold-out validation set. For ease of computation, only the last two layers of the model are used to calculate the influence of accuracy and fairness components. For Figure 2, one main observation is that FIS outperforms baselines with a significant margin on three fairness metrics while maintaining the same accuracy level. This improvement, as indicated in Theorem 5.2, can be attributed to FIS assigning priority to new examples based on the fairness influence, then avoiding accuracy reduction via their accuracy influence.

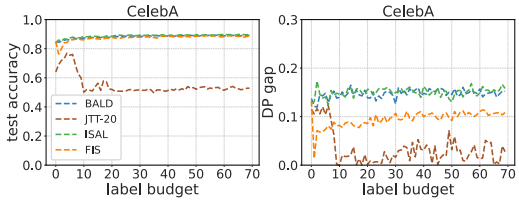
Results on tabular datasets Next, we work with multi-layer perceptron (MLP) with two layers trained on the Adult [8] and Compas dataset [6], respectively. We select `age` as the sensitive attribute for the Adult dataset and `race` for the Compas dataset. For two datasets, we resample the data to balance the class and group membership [17]. The whole dataset is split into training and test sets at a 4:1 ratio. Then, we randomly re-select 20% of the training set for initial training and the remaining 80% for sampling. Also, 20% examples of the test set are selected to form a validation set. The MLP model is a two-layer ReLU network with a hidden size of 64. We utilize the whole model parameters to compute the influence of accuracy and fairness for examples. Figure 2 summarizes the main results of the Adult and Compas datasets. On the Adult dataset, we observe that our sampling method achieves the lowest violation for equality of opportunity and has a comparable performance for the DP metric. Besides, our algorithm achieves a much better accuracy-fairness trade-off than

other baselines on the Compas dataset. JTT-20 achieves a lower fairness violation with the price of a significant accuracy drop compared to other baselines.

6.3 Ablation study

What is the impact of label budgets? Here, we examine how the varying label budgets r affect the trade-off between accuracy and fairness. For ease of comparison, we adhere to a consistent label budget per round to illustrate their respective impacts. As shown in Figure 3, our method consistently preserves a lower fairness violation than the BALD and ISAL baselines with a similar test accuracy. While we observe that the JTT-20 algorithm can achieve a near-zero fairness violation under a limited budget on the CelebA dataset, we argue that the model accuracy is rather uninformative (about 50%). More empirical results can be found in Appendix E.3.

How does the validation set size affect the performance? We further explore the impact of adjusting the validation set size on our algorithm's performance. We present the test accuracy and fairness violations across different validation set sizes on the CelebA dataset. Note that the default validation set size is set to 1% of the whole dataset size. In particular, the minimum scale of the validation set size is set to $1/5\times$ (nearly 400 CelebA images). The results in Figure 3 indicate that our algorithm still retains the test accuracy and fairness violation when we vary the validation set size. Additional results conducted on Adult and Compas datasets are provided in Appendix E.4.



	CelebA - Smiling		
	(Test_acc↑, DP↓)	(Test_acc↑, EOp↓)	(Test_acc↑, EOd↓)
$1\times$	(0.848, 0.084)	(0.876, 0.031)	(0.864, 0.030)
$1/2\times$	(0.872, 0.105)	(0.891, 0.042)	(0.880, 0.028)
$1/5\times$	(0.872, 0.117)	(0.863, 0.057)	(0.886, 0.028)

Figure 3: **Left:** The impact of label budgets on the test accuracy & DP gap in the **CelebA** dataset. **Right:** The impact of the validation set size on (`test_accuracy`, `fairness_violation`) results.

7 Conclusions and limitations

In this work, we are interested in facilitating ML models that mitigate group fairness disparity without harming model accuracy. To achieve this, different from current active sampling methods, we propose a tractable fair influential sampling method FIS, which avoids the need for training group annotations during the sampling or training phase, thereby preventing the potential exposure of sensitive information. In particular, this algorithm acquires data samples from a large dataset for training based on the influence of fairness and accuracy evaluated on the auxiliary validation dataset. Empirical experiments on real-world data validate the efficacy of our proposed method.

Nonetheless, we recognize that our method has limitations. Although the proposed sampling algorithm does not require sensitive attribute information from the massive data, it relies on a clean and informative validation set that contains the sensitive attributes of data examples. We consider this as a reasonable requirement in practice, given the relatively modest size of the validation set. Besides, one potential concern is that the sampling strategy may become inefficient when the collected validation set is noisy. However, this practical issue can be heavily alleviated by using loss correction methods [54, 84] or noise-tolerant fairness loss functions to rectify the error terms [53, 73, 29]. In future work, we aim to address this limitation by developing more robust sampling strategies that can perform effectively even with noisy validation sets.

Acknowledgment This work is partially supported by the National Science Foundation (NSF) under grants IIS-2143895, IIS-2040800, IIS-2416896, and CCF-2023495. Additionally, Pang and Qian were partially supported by NSF Grants 2322919, 2420632, 2426031, and 2114113.

Broader Impact

This paper presents work whose goal is to advance the field of fairness in machine learning. There are many potential societal consequences of our work. While the proposed algorithm does intend to infer sensitive attributes of data examples that may be protected by privacy regulations, it does not necessitate direct access to such sensitive information. On the other hand, our work can serve as an effective approach leading to mitigating the disparity with a limited annotation budget. We have thoroughly examined the potential ethical implications of our work and, based on our assessment, do not identify any issues that we deem necessary to emphasize here specifically.

References

[1] Pareto effiency and production possibility frontier. https://en.wikipedia.org/wiki/Production%E2%80%93possibility_frontier.

[2] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning (ICML '18)*, 2018.

[3] Alekh Agarwal, Miroslav Dudík, and Zhiwei Steven Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*, pages 120–129. PMLR, 2019.

[4] Hadis Anahideh, Abolfazl Asudeh, and Saravanan Thirumuruganathan. Fair active learning. *Expert Systems with Applications*, 199:116981, 2022.

[5] McKane Andrus, Elena Spitzer, Jeffrey Brown, and Alice Xiang. What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 249–260, 2021.

[6] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. 2016.

[7] Abolfazl Asudeh, Zhongjun Jin, and HV Jagadish. Assessing and remedying coverage for a given dataset. In *2019 IEEE 35th International Conference on Data Engineering (ICDE)*, pages 554–565. IEEE, 2019.

[8] Arthur Asuncion and David Newman. Uci machine learning repository, 2007.

[9] Fabio Azzalini, Chiara Criscuolo, and Letizia Tanca. Fair-db: Functional dependencies to discover data bias. In *EDBT/ICDT Workshops*, 2021.

[10] Ananth Balashankar, Alyssa Lees, Chris Welty, and Lakshminarayanan Subramanian. What is fair? exploring pareto-efficiency for fairness constrained classifiers. *arXiv preprint arXiv:1910.14120*, 2019.

[11] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning: Limitations and opportunities*. MIT Press, 2023.

[12] Frédéric Branchaud-Charron, Parmida Atighehchian, Pau Rodríguez, Grace Abuhamad, and Alexandre Lacoste. Can active learning preemptively mitigate fairness issues? *arXiv preprint arXiv:2104.06879*, 2021.

[13] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[14] Réal André Carbonneau, Kevin Laframboise, and Rustam M. Vahidov. Application of machine learning techniques for supply chain demand forecasting. *Eur. J. Oper. Res.*, 184:1140–1154, 2008.

[15] L Elisa Celis, Vijay Keswani, and Nisheeth Vishnoi. Data preprocessing to mitigate bias: A maximum entropy based approach. In *International conference on machine learning*, pages 1349–1359. PMLR, 2020.

[16] Junyi Chai, Taeuk Jang, and Xiaoqian Wang. Fairness without demographics through knowledge distillation. *Advances in Neural Information Processing Systems*, 35:19152–19164, 2022.

[17] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16:321–357, 2002.

[18] Irene Chen, Fredrik D Johansson, and David Sontag. Why is my classifier discriminatory? *Advances in neural information processing systems*, 31, 2018.

[19] Wenlong Chen, Yegor Klochkov, and Yang Liu. Post-hoc bias scoring is optimal for fair classification. *arXiv preprint arXiv:2310.05725*, 2023.

[20] Yatong Chen, Reilly Raab, Jialu Wang, and Yang Liu. Fairness transferability subject to bounded distribution shift. *Advances in Neural Information Processing Systems*, 35:11266–11278, 2022.

[21] Kristy Choi, Aditya Grover, Trisha Singh, Rui Shu, and Stefano Ermon. Fair generative modeling via weak supervision. In *International Conference on Machine Learning*, pages 1887–1898. PMLR, 2020.

[22] Sanjoy Dasgupta and Daniel Hsu. Hierarchical sampling for active learning. In *Proceedings of the 25th international conference on Machine learning*, pages 208–215, 2008.

[23] Maurice Diesendruck, Ethan R Elenberg, Rajat Sen, Guy W Cole, Sanjay Shakkottai, and Sinead A Williamson. Importance weighted generative networks. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2019, Würzburg, Germany, September 16–20, 2019, Proceedings, Part II*, pages 249–265. Springer, 2020.

[24] Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. *Advances in neural information processing systems*, 31, 2018.

[25] 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*, pages 2803–2813. PMLR, 2020.

[26] Michael Feldman. *Computational fairness: Preventing machine-learned discrimination*. PhD thesis, 2015.

[27] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pages 259–268, 2015.

[28] Vitaly Feldman. Does learning require memorization? a short tale about a long tail. In *Proceedings of the 52nd Annual ACM SIGACT Symposium on Theory of Computing*, pages 954–959, 2020.

[29] Lei Feng, Senlin Shu, Zhuoyi Lin, Fengmao Lv, Li Li, and Bo An. Can cross entropy loss be robust to label noise? In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 2206–2212, 2021.

[30] Stephen Giguere, Blossom Metevier, Yuriy Brun, Bruno Castro da Silva, Philip S Thomas, and Scott Niekum. Fairness guarantees under demographic shift. In *Proceedings of the 10th International Conference on Learning Representations (ICLR)*, 2022.

[31] Zayd Hammoudeh and Daniel Lowd. Training data influence analysis and estimation: A survey. *arXiv preprint arXiv:2212.04612*, 2022.

[32] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

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

[34] Steven CH Hoi, Rong Jin, and Michael R Lyu. Large-scale text categorization by batch mode active learning. In *Proceedings of the 15th international conference on World Wide Web*, pages 633–642, 2006.

[35] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–16, 2019.

[36] Heinrich Jiang and Ofir Nachum. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*, pages 702–712. PMLR, 2020.

[37] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and information systems*, 33(1):1–33, 2012.

[38] Polina Kirichenko, Pavel Izmailov, and Andrew Gordon Wilson. Last layer re-training is sufficient for robustness to spurious correlations. *arXiv preprint arXiv:2204.02937*, 2022.

[39] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.

[40] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. *Advances in neural information processing systems*, 33:728–740, 2020.

[41] David D Lewis and Jason Catlett. Heterogeneous uncertainty sampling for supervised learning. In *Machine learning proceedings 1994*, pages 148–156. Elsevier, 1994.

[42] Peizhao Li and Hongfu Liu. Achieving fairness at no utility cost via data reweighing with influence. In *International Conference on Machine Learning*, pages 12917–12930. PMLR, 2022.

[43] Tian Li, Maziar Sanjabi, Ahmad Beirami, and Virginia Smith. Fair resource allocation in federated learning. In *International Conference on Learning Representations*, 2019.

[44] 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*, pages 9572–9581, 2019.

[45] Michael Lindenbaum, Shaul Markovitch, and Dmitry Rusakov. Selective sampling for nearest neighbor classifiers. *Machine learning*, 54:125–152, 2004.

[46] Richard G Lipsey. An introduction to positive economics-4. 1975.

[47] 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*, pages 6781–6792. PMLR, 2021.

[48] Zhuoming Liu, Hao Ding, Huaping Zhong, Weijia Li, Jifeng Dai, and Conghui He. Influence selection for active learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 9274–9283, 2021.

[49] 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*, pages 3730–3738, 2015.

[50] Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Fairness with minimal harm: A pareto-optimal approach for healthcare. *arXiv preprint arXiv:1911.06935*, 2019.

[51] Natalia Martinez, Martin Bertran, and Guillermo Sapiro. Minimax pareto fairness: A multi objective perspective. In *International conference on machine learning*, pages 6755–6764. PMLR, 2020.

[52] Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In *Conference on Fairness, accountability and transparency*, pages 107–118. PMLR, 2018.

[53] Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In *Advances in neural information processing systems*, pages 1196–1204, 2013.

[54] Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1944–1952, 2017.

[55] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems*, pages 5680–5689, 2017.

[56] Flavien Prost, Hai Qian, Qiuwen Chen, Ed H Chi, Jilin Chen, and Alex Beutel. Toward a better trade-off between performance and fairness with kernel-based distribution matching. *arXiv preprint arXiv:1910.11779*, 2019.

[57] Maan Qraitem, Kate Saenko, and Bryan A Plummer. Bias mimicking: A simple sampling approach for bias mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20311–20320, 2023.

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

[59] Nicholas Roy and Andrew McCallum. Toward optimal active learning through monte carlo estimation of error reduction. *ICML, Williamstown*, 2:441–448, 2001.

[60] Shubham Sharma, Yunfeng Zhang, Jesús M Ríos Aliaga, Djallel Bouneffouf, Vinod Muthusamy, and Kush R Varshney. Data augmentation for discrimination prevention and bias disambiguation. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 358–364, 2020.

[61] Changjian Shui, Gezheng Xu, Qi Chen, Jiaqi Li, Charles X Ling, Tal Arbel, Boyu Wang, and Christian Gagné. On learning fairness and accuracy on multiple subgroups. *Advances in Neural Information Processing Systems*, 35:34121–34135, 2022.

[62] Naeem Siddiqi. *Credit risk scorecards: developing and implementing intelligent credit scoring*, volume 3. John Wiley & Sons, 2012.

[63] Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. Fairness violations and mitigation under covariate shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 3–13, 2021.

[64] Nimit Sohoni, Jared Dunnmon, Geoffrey Angus, Albert Gu, and Christopher Ré. No subclass left behind: Fine-grained robustness in coarse-grained classification problems. *Advances in Neural Information Processing Systems*, 33:19339–19352, 2020.

[65] Ki Hyun Tae and Steven Euijong Whang. Slice tuner: A selective data acquisition framework for accurate and fair machine learning models. In *Proceedings of the 2021 International Conference on Management of Data*, pages 1771–1783, 2021.

[66] Michael Veale and Reuben Binns. Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data. *Big Data & Society*, 4(2):2053951717743530, 2017.

[67] Akshaj Kumar Veldanda, Ivan Brugere, Sanghamitra Dutta, Alan Mishler, and Siddharth Garg. Hyper-parameter tuning for fair classification without sensitive attribute access. *arXiv preprint arXiv:2302.01385*, 2023.

[68] Jialu Wang, Yang Liu, and Caleb Levy. Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 526–536, New York, NY, USA, 2021. Association for Computing Machinery.

[69] Jialu Wang, Xin Eric Wang, and Yang Liu. Understanding instance-level impact of fairness constraints. In *International Conference on Machine Learning*, pages 23114–23130. PMLR, 2022.

[70] Ran Wang, Sam Kwong, and Degang Chen. Inconsistency-based active learning for support vector machines. *Pattern Recognition*, 45(10):3751–3767, 2012.

[71] Tianyang Wang, Xingjian Li, Pengkun Yang, Guosheng Hu, Xiangrui Zeng, Siyu Huang, Cheng-Zhong Xu, and Min Xu. Boosting active learning via improving test performance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 8566–8574, 2022.

[72] Yuyan Wang, Xuezhi Wang, Alex Beutel, Flavien Prost, Jilin Chen, and Ed H Chi. Understanding and improving fairness-accuracy trade-offs in multi-task learning. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 1748–1757, 2021.

[73] Hongxin Wei, Lei Feng, Xiangyu Chen, and Bo An. Combating noisy labels by agreement: A joint training method with co-regularization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13726–13735, 2020.

[74] Susan Wei and Marc Niethammer. The fairness-accuracy pareto front. *Statistical Analysis and Data Mining: The ASA Data Science Journal*, 15(3):287–302, 2022.

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

[76] Yichen Xie, Han Lu, Junchi Yan, Xiaokang Yang, Masayoshi Tomizuka, and Wei Zhan. Active finetuning: Exploiting annotation budget in the pretraining-finetuning paradigm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23715–23724, 2023.

[77] Shen Yan, Hsien-te Kao, and Emilio Ferrara. Fair class balancing: Enhancing model fairness without observing sensitive attributes. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1715–1724, 2020.

[78] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.

[79] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P Gummadi. Fairness constraints: A flexible approach for fair classification. *The Journal of Machine Learning Research*, 20(1):2737–2778, 2019.

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

[81] Han Zhao and Geoffrey J Gordon. Inherent tradeoffs in learning fair representations. *The Journal of Machine Learning Research*, 23(1):2527–2552, 2022.

[82] Tianxiang Zhao, Enyan Dai, Kai Shu, and Suhang Wang. Towards fair classifiers without sensitive attributes: Exploring biases in related features. In *Proceedings of the Fifteenth ACM International Conference on Web Search and Data Mining*, pages 1433–1442, 2022.

[83] Zhaowei Zhu, Tianyi Luo, and Yang Liu. The rich get richer: Disparate impact of semi-supervised learning. *arXiv preprint arXiv:2110.06282*, 2021.

[84] Zhaowei Zhu, Yuanshun Yao, Jiankai Sun, Hang Li, and Yang Liu. Weak proxies are sufficient and preferable for fairness with missing sensitive attributes. In *International Conference on Machine Learning*, pages 43258–43288. PMLR, 2023.

Appendix

The Appendix is organized as follows.

- Section A provides more details of the related work.
- Section B explores the relationship between our proposed fairness notion risk disparity and common fairness metrics, such as DP and EOd. In particular, we provide the full proof for Proposition 3.1.
- Section C provides a detailed analysis of the FIS algorithm including 1) evaluating first-order influence estimations against real influence, 2) a comparative analysis of computational costs, and 3) the exploration of the labeling strategies.
- Section D presents the full proofs for the Lemmas and Theorems shown in Section 4 and Section 5.
- Section E presents detailed descriptions of all datasets, corresponding parameter settings, and full version of the experimental results. In particular, to demonstrate FIS's advantage at the same levels of information, we introduce a new baseline called `Random+Val`.
- Section F discusses the privacy concern potentially caused by the correlations between non-demographic and demographic features.

A More details of related work

Active learning The core idea of active learning is to rank unlabeled instances by developing specific significant measures, including uncertainty [41, 45], representativeness [22], inconsistency [70], variance [34], and error [59]. Each of these measures has its criterion to determine the importance of instances for enhancing classifier performance. For example, uncertainty considers the most important unlabeled instance to be the nearest one to the current classification boundary. A related line of work [48, 71, 31, 76] concentrates on ranking unlabeled instances based on the influence function. Compared to these studies with a focus on prediction accuracy, our work poses a distinct challenge taking into account fairness violations. We note that adopting a particular sampling strategy can lead to distribution shifts between the training and testing data. What's worse, even though fairness is satisfied within the training dataset, the model may still exhibit unfair treatments on the test dataset due to the distribution shift. Therefore, it becomes imperative for the sampling approach to also account for its potential impacts on fairness.

Pareto optimality In the field of fairness in machine learning, Pareto optimality indicates the theoretical frontier of fairness accuracy tradeoff, meaning that fairness can not be improved without worsening model accuracy. Existing methods primarily focus on seeking the Pareto frontier of fairness accuracy tradeoff for the neural network classifier, instead of reaching a better one. For example, Balashankar et al. [10] first analyzes the Pareto optimality for classifiers within the context of fairness constraints. Wang et al. [72] explores the multi-dimensional Pareto frontiers of the fairness-accuracy tradeoff in the multi-task setting. Another works [51, 50] target to obtain a Pareto efficient classifier to reduce worst-case group risks by formulating group fairness as a multiple-objective optimization problem, where each group risk is an objective function.

Fair classification The fairness-aware learning algorithms, in general, can be categorized into pre-processing, in-processing, and post-processing methods. Pre-processing methods typically reweigh or distort the data examples to mitigate the identified biases [7, 9, 65, 60, 15, 17, 80, 18]. More relevant to us is the *importance reweighting*, which assigns different weights to different training examples to ensure fairness across groups [37, 36, 23, 21, 57, 44]. Our proposed algorithm bears similarity to a specific case of importance reweighting, particularly the 0-1 reweighting applied to newly added data. The main advantage of our work, however, lies in its ability to operate without needing access to the sensitive attributes of either the new or training data. Other parallel studies utilize importance weighting to learn a complex fair generative model in a weakly supervised setting [23, 21], or to mitigate representation bias in training datasets [44]. Post-processing methods typically enforce fairness on a learned model through calibration [26, 27, 32]. Although this approach is likely to decrease the disparity of the classifier, by decoupling the training from the fairness enforcement, this procedure may not lead to the best trade-off between fairness and accuracy [75, 55]. In contrast, our work can achieve a better trade-off between fairness and accuracy, because we reduce the fairness disparity by mitigating the adverse effects of distribution shifts on generalization error. Additionally, these post-processing techniques necessitate access to the sensitive attribute during the inference phase, which is often not available in many real-world scenarios.

B Understanding risk disparity through common fairness metrics

Here, we provide a brief proof to illustrate Proposition 3.1, that is, the relationship between risk parity and common fairness metrics, such as DP and EOd. For completeness, here we provide a more detailed version of Proposition 3.1, as well as its proof.

Proposition 3.1. Consider a binary classification scenario involving two demographic groups $S \in \{k, k'\}$. When two groups are balanced, i.e., $\frac{\mathbb{P}(Y=y|S=k)}{\mathbb{P}(Y=y|S=k')} = 1$, where $\hat{Y}$ denotes the predicted label, risk disparity can serve as a lower bound for fairness disparities based on EOd. Similarly, if the group sufficiency ratio can be calibrated by 1, i.e., $\mathbb{P}(Y = y|S = k, \hat{Y} = y) = \mathbb{P}(Y = y|S = k', \hat{Y} = y) = R$ where R is calibration score, risk disparity can be formulated as a lower bound for DP-based fairness disparity.

Proof. Consider a scenario involving two demographic groups $S \in \{k, k'\}$, alongside a 0-1 loss function.

Fairness metric EOd For EOd, the risk disparity can be reformulated as

$$
\begin{aligned}
&|\mathcal{R}_{Q_k}(\mathbf{w}^P) - \mathcal{R}_{Q_{k'}}(\mathbf{w}^P)| \\
=&\left|\frac{1}{|Q_k|}\sum\nolimits_{(x_n,y_n)\in Q_k}\mathbb{I}\left(\hat{y}_n \neq y_n\right) - \frac{1}{|Q_{k'}|}\sum\nolimits_{(x_n,y_n)\in Q_{k'}}\mathbb{I}\left(\hat{y}_n \neq y_n\right)\right| \\
=&\left|\mathbb{P}(\hat{Y} \neq Y \mid S = k) - \mathbb{P}(\hat{Y} \neq Y \mid S = k')\right| \\
=&\left|\sum_y \Big(\mathbb{P}\left(Y = y \mid S = k'\right)\mathbb{P}(\hat{Y} = y \mid S = k', Y = y) - \mathbb{P}(Y = y \mid S = k)\mathbb{P}(\hat{Y} = y \mid S = k, Y = y)\Big)\right| \\
=&\left|\sum_y \mathbb{P}(Y = y \mid S = k') \cdot \Big(\mathbb{P}(\hat{Y} = y \mid S = k', Y = y) - \omega_{\text{EOd}}(y)\mathbb{P}(\hat{Y} = y \mid S = k, Y = y)\Big)\right| \\
=&\Big|\mathbb{P}(Y = 1|S = k') \cdot \Big(\mathbb{P}(\hat{Y} = 1 \mid S = k', Y = 1) - \omega_{\text{EOd}}(y = 1) \cdot \mathbb{P}(\hat{Y} = 1 \mid S = k, Y = 1)\Big) \qquad \text{\# binary case} \\
&\quad + \mathbb{P}(Y = 0|S = k') \cdot \Big(\mathbb{P}(\hat{Y} = 0 \mid S = k', Y = 0) - \omega_{\text{EOd}}(y = 0) \cdot \mathbb{P}(\hat{Y} = 0 \mid S = k, Y = 0)\Big)\Big| \\
=&\Big|\mathbb{P}(Y = 1|S = k') \cdot \Big(\mathbb{P}(\hat{Y} = 1 \mid S = k', Y = 1) - \omega_{\text{EOd}}(y = 1) \cdot \mathbb{P}(\hat{Y} = 1 \mid S = k, Y = 1)\Big) \\
&\quad + \mathbb{P}(Y = 0|S = k') \cdot \Big(\omega_{\text{EOd}}(y = 0) \cdot \mathbb{P}(\hat{Y} = 1 \mid S = k, Y = 0) - \mathbb{P}(\hat{Y} = 1 \mid S = k', Y = 0)\Big) \\
&\quad + \mathbb{P}(Y = 0|S = k') \cdot (1 - \omega_{\text{EOd}}(y = 0))\Big| \\
=&\Big|\mathbb{P}(Y = 1|S = k') \cdot \Big(\mathbb{P}(\hat{Y} = 1 \mid S = k', Y = 1) - \mathbb{P}(\hat{Y} = 1 \mid S = k, Y = 1)\Big) \\
&\quad - \mathbb{P}(Y = 0|S = k') \cdot \Big(\mathbb{P}(\hat{Y} = 1 \mid S = k', Y = 0) - \mathbb{P}(\hat{Y} = 1 \mid S = k, Y = 0)\Big)\Big| \qquad \text{\# balanced groups} \\
\leq&\mathbb{P}(Y = 1|S = k') \cdot \Big|\mathbb{P}(\hat{Y} = 1 \mid S = k', Y = 1) - \mathbb{P}(\hat{Y} = 1 \mid S = k, Y = 1)\Big| \\
&\quad + \mathbb{P}(Y = 0|S = k') \cdot \Big|\mathbb{P}(\hat{Y} = 1 \mid S = k', Y = 0) - \mathbb{P}(\hat{Y} = 1 \mid S = k, Y = 0)\Big| \\
\leq&\underbrace{\sum_{y\in\{0,1\}}\Big|\mathbb{P}(\hat{Y} = 1 \mid S = k', Y = y) - \mathbb{P}(\hat{Y} = 1 \mid S = k, Y = y)\Big|}_{\text{EOd-based fairness disparity}}
\end{aligned}
$$

where we define $\omega_{\text{EOd}}(y) := \frac{\mathbb{P}(Y=y|S=k)}{\mathbb{P}(Y=y|S=k')}$, serving as the bias weight. Here, we make a mild assumption that the two demographic groups are balanced, i.e., $\omega_{\text{EOd}}(y) = 1$. Then, we can see that the last item measures the fairness disparity based on EOd for binary classification problems. Thus, we can claim that reducing risk disparity can promote the fairness metric EOd.

Fairness metric DP Similarly, for DP, we formulate the risk disparity as

$$
\begin{aligned}
&|\mathcal{R}_{Q_k}(\mathbf{w}^P) - \mathcal{R}_{Q_{k'}}(\mathbf{w}^P)| \\
=&\left|\frac{1}{|Q_k|}\sum\nolimits_{(x_n,y_n)\in Q_k}\mathbb{I}\left(\hat{y}_n \neq y_n\right) - \frac{1}{|Q_{k'}|}\sum\nolimits_{(x_n,y_n)\in Q_{k'}}\mathbb{I}\left(\hat{y}_n \neq y_n\right)\right| \\
=&\left|\mathbb{P}(\hat{Y} \neq Y \mid S=k) - \mathbb{P}(\hat{Y} \neq Y \mid S=k')\right| \\
=&\left|\sum_y \Big(\mathbb{P}(\hat{Y}=y \mid S=k')\mathbb{P}(Y=y \mid S=k', \hat{Y}=y) - \mathbb{P}(\hat{Y}=y \mid S=k)\mathbb{P}(Y=y \mid S=k, \hat{Y}=y)\Big)\right| \\
=&\left|\sum_y \mathbb{P}(Y=y|S=k', \hat{Y}=y)\cdot\Big(\mathbb{P}(\hat{Y}=y \mid S=k') - \omega_{\text{DP}}(y)\mathbb{P}(\hat{Y}=y \mid S=k)\Big)\right| \\
=&\Big|\mathbb{P}(Y=1|S=k', \hat{Y}=1)\cdot\Big(\mathbb{P}(\hat{Y}=1 \mid S=k') - \omega_{\text{DP}}(y=1)\cdot\mathbb{P}(\hat{Y}=1 \mid S=k)\Big) \qquad \text{\# binary case} \\
&\quad + \mathbb{P}(Y=0|S=k', \hat{Y}=0)\cdot\Big(\mathbb{P}(\hat{Y}=0 \mid S=k') - \omega_{\text{DP}}(y=0)\cdot\mathbb{P}(\hat{Y}=0 \mid S=k)\Big)\Big| \\
=&\Big|\mathbb{P}(Y=1|S=k', \hat{Y}=1)\cdot\Big(\mathbb{P}(\hat{Y}=1 \mid S=k') - \omega_{\text{DP}}(y=1)\cdot\mathbb{P}(\hat{Y}=1 \mid S=k)\Big) \\
&\quad + \mathbb{P}(Y=0|S=k', \hat{Y}=0)\cdot\Big(\omega_{\text{DP}}(y=0)\cdot\mathbb{P}(\hat{Y}=1 \mid S=k) - \mathbb{P}(\hat{Y}=1 \mid S=k')\Big) \\
&\quad + \mathbb{P}(Y=0|S=k', \hat{Y}=0)\cdot(1-\omega_{\text{DP}}(y=0))\Big| \qquad \text{\# Calibrating group sufficiency} \\
=&\Big|\mathbb{P}(Y=1|S=k', \hat{Y}=1)\cdot\Big(\mathbb{P}(\hat{Y}=1 \mid S=k') - \mathbb{P}(\hat{Y}=1 \mid S=k)\Big) \\
&\quad + \mathbb{P}(Y=0|S=k', \hat{Y}=0)\cdot\Big(\mathbb{P}(\hat{Y}=1 \mid S=k) - \mathbb{P}(\hat{Y}=1 \mid S=k')\Big)\Big| \\
=&\Big|\mathbb{P}(Y=1|S=k', \hat{Y}=1) - \mathbb{P}(Y=0|S=k', \hat{Y}=0)\Big|\cdot\Big|\mathbb{P}(\hat{Y}=1 \mid S=k') - \mathbb{P}(\hat{Y}=1 \mid S=k)\Big| \\
\leq&\underbrace{\Big|\mathbb{P}(\hat{Y}=1 \mid S=k') - \mathbb{P}(\hat{Y}=1 \mid S=k)\Big|}_{\text{DP-based fairness disparity}}
\end{aligned}
$$

where we define $\omega_{\text{DP}}(y) := \frac{\mathbb{P}(Y=y|S=k,\hat{Y}=y)}{\mathbb{P}(Y=y|S=k',\hat{Y}=y)}$. In fact, $\omega_{\text{DP}}(y)$ measures the group sufficiency ratio [61]. Note that the group sufficiency is closely related to the idea of calibration [11]. Thus, we make a mild assumption that the group sufficiency ratio $\omega_{\text{DP}}(y)$ can be calibrated by 1, i.e., $\mathbb{P}(Y=y|S=k,\hat{Y}=y) = \mathbb{P}(Y=y|S=k',\hat{Y}=y) = R$, where R is a calibration score. When two demographic groups are balanced, i.e., $\omega_{\text{DP}}(y)=1$, we can see that the last inequality indicates the DP-based fairness disparity for binary classification problems. Therefore, we can also claim that the fairness metric DP can also be encouraged when reducing the risk disparity.

□

C Detailed analysis of the fair influential sampling algorithm

C.1 Evaluating first-order influence estimations against real influence

Recall that the influence score derived in this paper primarily utilizes a first-order approach. Here, we will demonstrate how accurate the first-order estimation of the influence is in comparison to the real influence.

C.2 Comparative analysis of computational costs

Recall that the proposed algorithm FIS needs to pre-calculate the accuracy loss and fairness loss for evaluating the performance of a certain example. However, the extra computation cost is comparable to the cost of traditional model training. Note that the main extra computation cost in FIS (Algorithm

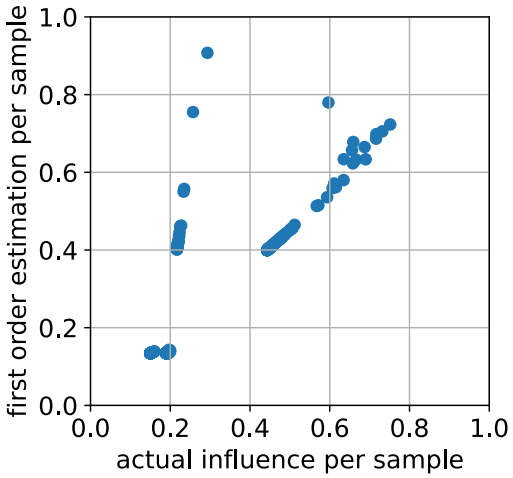


Figure 4: We validate how accurate the first-order estimation of the influence is in comparison to the real influence. The x-axis represents the actual influence per sample, and the y-axis represents the estimated influence. We observe that while some of the examples are away from the diagonal line (which indicates the estimation is inaccurate), the estimated influences for most of the data samples are very close to their actual influence values.

1) mainly results from model gradients. Let p denote the number of model parameters, then the cost for computing the gradients is $O(p)$ per sample. Specifically, in each round that involves sampling, we need to calculate three parts of gradients: the gradients of $|U|$ unlabeled instances, the average gradient of $|Q_v|$ validation instances w.r.t. accuracy loss, and the gradient of $|Q_v|$ validation instances w.r.t. fairness loss. Note that in general, $|Q_v| \ll |U|$. In practical implementation, to speed up the calculation of gradients over $|U|$ instances, we randomly sample 0.2-0.5% of the unlabeled dataset in each sampling batch. Additionally, we can increase the number of newly selected examples for each round (r) to save computation costs. In our experiments, we usually have 10-20 sampling rounds. For instance, running one experiment for the CelebA dataset on a single GPU roughly requires about 4 hours.

C.3 Exploration of the labeling strategies

Note that we provide a strategy that employs lowest-influence labels for annotating labels. This section will explore an alternative strategy that employs model predictions for the purpose of labeling. For completeness, we outline the two proposed labeling strategies as follows.

Strategy I **Use low-influence labels.** That is, $\hat{y} = \arg\min_{k\in[K]} |\mathsf{Infl}_{\mathsf{acc}}(x', k)|$, which corresponds to using the most uncertain point.

Strategy II **Rely on model prediction.** That is, $\hat{y} = \arg\max_{k\in[K]} f(x; \mathbf{w})[k]$, where $f(x; \mathbf{w})[y]$ indicates the model's prediction probability on label y.

Remark C.1. *Suppose that the model is trained with cross-entropy loss. The labels obtained through* ***Strategy II*** *are sufficient to minimize the influence of the prediction component, i.e.,* $\mathsf{Infl}_{acc}(x', k)$. *That said, the* ***Strategy II*** *will produce similar labels as* ***Strategy I****.*

Proof. Based on the definition of the influence of the prediction component, as delineated in Eq. (2), it becomes evident that the most uncertain points are obtained when the proxy labels closely align with the true labels. Consequently, the model predictions used in Strategy II also approximate the true labels to minimize the cross-entropy loss. Thus, in a certain sense, Strategy I and Strategy II can be considered equivalent. □

D Omitted proofs

In this section, we present complete proofs for the lemmas and theorems in Section 4 and 5, respectively.

D.1 Proof of Lemma 4.1

Lemma 4.1. *The influence of predictions on the validation dataset Q_v can be denoted by*

$$\mathit{Infl}_{\mathit{acc}}(z') := \sum_{n\in|Q_v|} \mathit{Infl}_{\mathit{acc}}(z', z_n^\circ) \approx -\eta \sum\nolimits_{n\in|Q_v|} \langle \partial_{\mathbf{w}_t}\ell(\mathbf{w}_t, z'), \partial_{\mathbf{w}_t}\ell(\mathbf{w}_t, z_n^\circ)\rangle$$

Proof. Taking the first-order Taylor expansion, we will have

$$\ell(\mathbf{w}_{t+1}, z_n^\circ) \approx \ell(\mathbf{w}_t, z_n^\circ) + \left\langle \frac{\partial \ell(\mathbf{w}, z_n^\circ)}{\partial f(\mathbf{w}, x_n^\circ)}\bigg|_{\mathbf{w}=\mathbf{w}_t}, f(\mathbf{w}_{t+1}, x_n^\circ) - f(\mathbf{w}_t, x_n^\circ)\right\rangle.$$

where we take this expansion with respect to $f(\mathbf{w}, x_n^\circ)$. Similarly, we have

$$\begin{aligned} f(\mathbf{w}_{t+1}, x_n^\circ) - f(\mathbf{w}_t, x_n^\circ) &\approx \left\langle \frac{\partial f(\mathbf{w}, x_n^\circ)}{\partial \mathbf{w}}, \mathbf{w}_{t+1} - \mathbf{w}_t\right\rangle\bigg|_{\mathbf{w}=\mathbf{w}_t} \\ &= -\eta \left\langle \frac{\partial f(\mathbf{w}, x_n^\circ)}{\partial \mathbf{w}}, \frac{\partial \ell(\mathbf{w}, z')}{\partial \mathbf{w}}\right\rangle\bigg|_{\mathbf{w}=\mathbf{w}_t}. \end{aligned}$$

where the last equality holds due to Eq. (1). Therefore,

$$\begin{aligned} \ell(\mathbf{w}_{t+1}, z_n^\circ) - \ell(\mathbf{w}_t, z_n^\circ) &\approx -\eta \left\langle \frac{\partial \ell(\mathbf{w}, z_n^\circ)}{\partial f(\mathbf{w}, x_n^\circ)}, \left\langle \frac{\partial f(\mathbf{w}, x_n^\circ)}{\partial \mathbf{w}}, \frac{\partial \ell(\mathbf{w}, z')}{\partial \mathbf{w}}\right\rangle\right\rangle\bigg|_{\mathbf{w}=\mathbf{w}_t} \\ &= -\eta \left\langle \frac{\partial f(\mathbf{w}, x_n^\circ)}{\partial \mathbf{w}}, \frac{\partial \ell(\mathbf{w}, z')}{\partial \mathbf{w}}\right\rangle\bigg|_{\mathbf{w}=\mathbf{w}_t}. \end{aligned}$$

Then the accuracy influence on the validation dataset V can be denoted by

$$\mathsf{Infl}_{\mathsf{acc}}(z') := \sum_{n\in|Q_v|} \mathsf{Infl}_{\mathsf{acc}}(z', z_n^\circ) \approx -\eta \left\langle \partial_{\mathbf{w}_t}\ell(\mathbf{w}_t, z'), \sum_{n\in|Q_v|} \partial_{\mathbf{w}_t}\ell(\mathbf{w}_t, z_n^\circ)\right\rangle$$

□

D.2 Proof of Lemma 4.2

Lemma 4.2. *The influence of fairness on the validation dataset Q_v can be denoted by*

$$\mathit{Infl}_{\mathit{fair}}(z') := \sum_{n\in|Q_v|} \mathit{Infl}_{\mathit{fair}}(z', z_n^\circ) \approx -\eta \sum_{n\in|Q_v|} \langle \partial_{\mathbf{w}_t}\ell(\mathbf{w}_t, z'), \partial_{\mathbf{w}_t}\phi(\mathbf{w}_t, z_n^\circ)\rangle$$

Proof. By first-order approximation, we have

$$\phi(\mathbf{w}_{t+1}, z_n^\circ) \approx \phi(\mathbf{w}_t, z_n^\circ) + \left\langle \frac{\partial \phi(\mathbf{w}_t, z_n^\circ)}{\partial f(\mathbf{w}, x_n^\circ)}\bigg|_{\mathbf{w}=\mathbf{w}_t}, f(\mathbf{w}_{t+1}, x_n^\circ) - f(\mathbf{w}_t, x_n^\circ)\right\rangle.$$

Recall by first-order approximation, we have

$$f(\mathbf{w}_{t+1}, x_n^\circ) - f(\mathbf{w}_t, x_n^\circ) \approx -\eta \left\langle \frac{\partial f(\mathbf{w}, x_n^\circ)}{\partial \mathbf{w}}, \frac{\partial \ell(\mathbf{w}, z')}{\partial \mathbf{w}}\right\rangle\bigg|_{\mathbf{w}=\mathbf{w}_t}.$$

Therefore,

$$\phi(\mathbf{w}_{t+1}, z_n^\circ) - \phi(\mathbf{w}_t, z_n^\circ) \approx -\eta \left\langle \frac{\partial \ell(\mathbf{w}, z')}{\partial \mathbf{w}}, \frac{\partial \phi(\mathbf{w}_t, z_n^\circ)}{\partial \mathbf{w}}\right\rangle\bigg|_{\mathbf{w}=\mathbf{w}_t}$$

Note the loss function in the above equation should be ℓ since the model is updated with ℓ-loss. Therefore,

$$\mathsf{Infl}_{\mathsf{fair}}(z') = \sum_{n\in|Q_v|} \mathsf{Infl}_{\mathsf{fair}}(z', z_n^\circ) \approx -\eta \sum_{n\in|Q_v|} \left\langle \frac{\partial \ell(\mathbf{w}, z')}{\partial \mathbf{w}}, \frac{\partial \phi(\mathbf{w}_t, z_n^\circ)}{\partial \mathbf{w}} \right\rangle \Bigg|_{\mathbf{w}=\mathbf{w}_t}.$$

□

D.3 Proof of Theorem 5.1

Theorem 5.1. *(Generalization error bound). Let dist$(\mathcal{P}, \mathcal{Q})$, G_P be defined therein. With probability at least $1-\delta$ with $\delta \in (0,1)$, the generalization error bound of the model trained on dataset P is*

$$\mathcal{R}_{\mathcal{Q}}(\mathbf{w}^P) \leq \underbrace{G_P \cdot dist(\mathcal{P}, \mathcal{Q})}_{\text{distribution shift}} + \sqrt{\frac{\log(4/\delta)}{2|P|}} + \mathcal{R}_P(\mathbf{w}^P). \tag{7}$$

Note that the generalization error bound is predominantly influenced by the shift in distribution when we think of an overfitting model, i.e., the empirical risk $\mathcal{R}_P(\mathbf{w}^P) \to 0$. The detailed proof is presented as follows.

Proof. The generalization error bound is

$$\begin{aligned}\mathcal{R}_{\mathcal{Q}}(\mathbf{w}^P) &= \underbrace{\left(\mathcal{R}_{\mathcal{Q}}(\mathbf{w}^P) - \mathcal{R}_{\mathcal{P}}(\mathbf{w}^P)\right)}_{\text{distribution shift}} + \underbrace{\left(\mathcal{R}_{\mathcal{P}}(\mathbf{w}^P) - \mathcal{R}_P(\mathbf{w}^P)\right)}_{\text{Hoeffding's inequality}} + \underbrace{\mathcal{R}_P(\mathbf{w}^P))}_{\text{empirical risk}} \\ &\leq G_P \cdot \text{dist}(\mathcal{P}, \mathcal{Q}) + \sqrt{\frac{\log(4/\delta)}{2|P|}} + \mathcal{R}_P(\mathbf{w}^P)\end{aligned}$$

For the first term (distribution shift), we have

$$\begin{aligned}\mathcal{R}_{\mathcal{Q}}(\mathbf{w}^P) - \mathcal{R}_{\mathcal{P}}(\mathbf{w}^P) &= \mathbb{E}_{z\sim\mathcal{Q}}[\ell(\mathbf{w}^P, z)] - \mathbb{E}_{z\sim\mathcal{P}}[\ell(\mathbf{w}^P, z)] \\ &= \sum_{i=1}^{I} p^{(\mathcal{Q})}(\pi = i)\mathbb{E}_{z\sim\pi_i}[\ell(\mathbf{w}^P, z)] - \sum_{i=1}^{I} p^{(\mathcal{P})}(\pi = i)\mathbb{E}_{z\sim\pi_i}[\ell(\mathbf{w}^P, z)] \\ &\leq \sum_{i=1}^{I} |p^{(\mathcal{P})}(\pi = i) - p^{(\mathcal{Q})}(\pi = i)|\mathbb{E}_{z\sim\pi_i}[\ell(\mathbf{w}^P, z)] \\ &\leq G_P \cdot \text{dist}(\mathcal{P}, \mathcal{Q}).\end{aligned}$$

where we define $\text{dist}(\mathcal{P}, \mathcal{Q}) = \sum_{i=1}^{I} |p^{(\mathcal{P})}(\pi = i) - p^{(\mathcal{Q})}(\pi = i)|$ and $\mathbb{E}_{z\sim\pi_i}[\ell(\mathbf{w}^P, z)] \leq G_P, \forall i \in I$ because of Assumption 5.2. To avoid misunderstanding, we use a subscript P of the constant G to clarify the corresponding model $\mathbf{w}^P$. Then, for the second term (Hoeffding inequality), with probability at least $1-\delta$, we have $|\mathcal{R}_{\mathcal{P}}(\mathbf{w}^P) - \mathcal{R}_P(\mathbf{w}^P)| \leq \sqrt{\frac{\log(4/\delta)}{2|P|}}$. □

D.4 Proof of Theorem 5.2

Theorem 5.2. *(Upper bound of fairness disparity). Suppose $\mathcal{R}_{\mathcal{Q}}(\cdot)$ follows Assumption 5.1. Let dist$(\mathcal{P}, \mathcal{Q})$, G_P, dist$(\mathcal{P}_k, \mathcal{Q}_k)$ and dist(P_k, P) be defined therein. The initial learning rate $\eta_0^2 < \frac{1}{\sqrt{2}TL}$, where T denotes the number of training epochs. Given model $\mathbf{w}^P$ and $\mathbf{w}^k$ trained exclusively on group k's data P_k, with probability at least $1-\delta$ with $\delta \in (0,1)$, then the upper bound of the fairness disparity is*

$$\mathcal{R}_{\mathcal{Q}_k}(\mathbf{w}^P) - \mathcal{R}_{\mathcal{Q}}(\mathbf{w}^P) \leq \underbrace{G_P \cdot dist(\mathcal{P}, \mathcal{Q})}_{\text{distribution shift}} + \underbrace{4L^2G^2 \cdot dist(P_k, P)^2}_{\text{group gap}} + G_k \cdot dist(\mathcal{P}_k, \mathcal{Q}_k) + \Upsilon.$$

where $\Upsilon = \sqrt{\frac{\log(4/\delta)}{2|P|}} + \sqrt{\frac{\log(4/\delta)}{2|P_k|}} + \varpi + \varpi_k$. *Note that* $\mathbb{E}_{z\sim\pi_i}[\ell(\mathbf{w}^k, z)] \leq G_k$, $\varpi = \mathcal{R}_P(\mathbf{w}^P) - \mathcal{R}^*_{\mathcal{Q}}(\mathbf{w}^{\mathcal{Q}})$ *and* $\varpi_k = \mathcal{R}_{P_k}(\mathbf{w}^k) - \mathcal{R}^*_{\mathcal{Q}_k}(\mathbf{w}^{\mathcal{Q}_k})$. *Specifically,* ϖ *and* ϖ_k *can be regarded as constants because* $\mathcal{R}_P(\mathbf{w}^P)$ *and* $\mathcal{R}_{P_k}(\mathbf{w}^k)$ *correspond to the empirical risks,* $\mathcal{R}^*_{\mathcal{Q}}(\mathbf{w}^{\mathcal{Q}})$ *and* $\mathcal{R}^*_{\mathcal{Q}_k}(\mathbf{w}^{\mathcal{Q}_k})$ *represent the ideal minimal empirical risk of model* $\mathbf{w}^{\mathcal{Q}}$ *trained on distribution* $\mathcal{Q}$ *and* $\mathcal{Q}_k$*, respectively. Moreover, these quantities* ϖ *and* ϖ_k *are not taken into account during the training phase, but rather in relation to the final model.*

Proof. First of all, we have

$$\begin{aligned}
&\mathcal{R}_{\mathcal{Q}_k}(\mathbf{w}^P) - \mathcal{R}_{\mathcal{Q}}(\mathbf{w}^P) \\
=&(\mathcal{R}_{\mathcal{Q}}(\mathbf{w}^{P_k}) - \mathcal{R}_{\mathcal{Q}}(\mathbf{w}^P)) + (\mathcal{R}_{\mathcal{Q}_k}(\mathbf{w}^P) - \mathcal{R}_{\mathcal{Q}}(\mathbf{w}^{P_k})) \\
=&(\mathcal{R}_{\mathcal{Q}}(\mathbf{w}^{P_k}) - \mathcal{R}_{\mathcal{Q}}(\mathbf{w}^P)) + (\mathcal{R}_{\mathcal{Q}_k}(\mathbf{w}^P) - \mathcal{R}_{\mathcal{Q}_k}(\mathbf{w}^{P_k})) + (\mathcal{R}_{\mathcal{Q}_k}(\mathbf{w}^{P_k}) - \mathcal{R}_{\mathcal{Q}}(\mathbf{w}^{P_k})) \\
\leq&(\mathcal{R}_{\mathcal{Q}}(\mathbf{w}^{P_k}) - \mathcal{R}_{\mathcal{Q}}(\mathbf{w}^P)) + (\mathcal{R}_{\mathcal{Q}_k}(\mathbf{w}^P) - \mathcal{R}_{\mathcal{Q}_k}(\mathbf{w}^{P_k}))
\end{aligned}$$

where $\mathbf{w}^{P_k}$ represents the model trained exclusively on group k's data. For simplicity, when there is no confusion, we use $\mathbf{w}^k$ to substitute $\mathbf{w}^{P_k}$. The inequality $\mathcal{R}_{\mathcal{Q}_k}(\mathbf{w}^k) - \mathcal{R}_{\mathcal{Q}}(\mathbf{w}^k) \geq 0$ holds because the model tailored for a single group k can not generalize well to the entirety of the test set Q.

Then, for the first term, we have

$$\begin{aligned}
\mathcal{R}_{\mathcal{Q}}(\mathbf{w}^k) - \mathcal{R}_{\mathcal{Q}}(\mathbf{w}^P) &\overset{(a)}{\leq} \langle \nabla\mathcal{R}_{\mathcal{Q}}(\mathbf{w}^P), \mathbf{w}^k - \mathbf{w}^P\rangle + \frac{L}{2}\|\mathbf{w}^k - \mathbf{w}^P\|^2 \\
&\overset{(b)}{\leq} L\|\mathbf{w}^k - \mathbf{w}^P\|^2 + \frac{1}{2L}\|\nabla\mathcal{R}_{\mathcal{Q}}(\mathbf{w}^P)\|^2 \\
&\overset{(c)}{\leq} \underbrace{L\|\mathbf{w}^k - \mathbf{w}^P\|^2}_{\text{group gap}} + \underbrace{(\mathcal{R}_{\mathcal{Q}}(\mathbf{w}^P) - \mathcal{R}^*_{\mathcal{Q}}(\mathbf{w}^{\mathcal{Q}}))}_{\text{train-test model gap}}
\end{aligned}$$

where inequality (a) holds because of the L-smoothness of expected loss $\mathcal{R}_{\mathcal{Q}}(\cdot)$, i.e., Assumption 5.1. Specifically, inequality (b) holds because, by Cauchy-Schwarz inequality and AM-GM inequality, we have

$$\langle \nabla\mathcal{R}_{\mathcal{Q}}(\mathbf{w}^P), \mathbf{w}^k - \mathbf{w}^P\rangle \leq \frac{L}{2}\|\mathbf{w}^k - \mathbf{w}^P\|^2 + \frac{1}{2L}\|\nabla\mathcal{R}_{\mathcal{Q}}(\mathbf{w}^P)\|^2.$$

Then, inequality (c) holds due to the L-smoothness of $\mathcal{R}_{\mathcal{Q}}(\cdot)$ (Assumption 5.1), we can get a variant of Polak-Łojasiewicz inequality, which follows

$$\|\nabla\mathcal{R}_{\mathcal{Q}}(\mathbf{w}^P)\|^2 \leq 2L(\mathcal{R}_{\mathcal{Q}}(\mathbf{w}^P) - \mathcal{R}^*_{\mathcal{Q}}(\mathbf{w}^{\mathcal{Q}})).$$

where $(\mathcal{R}^*_{\mathcal{Q}}(\mathbf{w}^{\mathcal{Q}}))$ denotes the ideal minimal empirical risk of model $\mathbf{w}^{\mathcal{Q}}$ trained on distribution $\mathcal{Q}$. Following a similar idea, for the second term, we also have

$$\mathcal{R}_{\mathcal{Q}_k}(\mathbf{w}^P) - \mathcal{R}_{\mathcal{Q}_k}(\mathbf{w}^k) \leq L\|\mathbf{w}^P - \mathbf{w}^k\|^2 + (\mathcal{R}_{\mathcal{Q}_k}(\mathbf{w}^k) - \mathcal{R}^*_{\mathcal{Q}_k}(\mathbf{w}^{\mathcal{Q}_k}))$$

Combined with two terms, we have

$$\mathcal{R}_{\mathcal{Q}_k}(\mathbf{w}^P) - \mathcal{R}_{\mathcal{Q}}(\mathbf{w}^P) \leq \underbrace{(\mathcal{R}_{\mathcal{Q}}(\mathbf{w}^P) - \mathcal{R}^*_{\mathcal{Q}}(\mathbf{w}^{\mathcal{Q}}))}_{\text{train-test model gap}} + \underbrace{2L\|\mathbf{w}^k - \mathbf{w}^P\|^2}_{\text{group model gap}} + (\mathcal{R}_{\mathcal{Q}_k}(\mathbf{w}^k) - \mathcal{R}^*_{\mathcal{Q}_k}(\mathbf{w}^{\mathcal{Q}_k}))$$

Lastly, integrating with Lemmas D.1, D.2 and D.3, we can finish the proof. □

Lemma D.1. *(Train-test model gap) With probability at least* $1-\delta$*, given the model* $\mathbf{w}^P$ *trained on train set* P*, we have*

$$\mathcal{R}_{\mathcal{Q}}(\mathbf{w}^P) - \mathcal{R}^*_{\mathcal{Q}}(\mathbf{w}^{\mathcal{Q}}) \leq G_P \cdot dist(\mathcal{P}, \mathcal{Q}) + \sqrt{\frac{\log(4/\delta)}{2|P|}} + \varpi.$$

where $dist(\mathcal{P}, \mathcal{Q}) = \sum_{i=1}^{I} |p^{(\mathcal{P})}(\pi = i) - p^{(\mathcal{Q})}(\pi = i)|$ *and* $\mathbb{E}_{z\sim\pi_i}[\ell(\mathbf{w}^P, z)] \leq G_P, \forall i \in I$*, and a constant* $\varpi := \mathcal{R}_P(\mathbf{w}^P) - \mathcal{R}^*_{\mathcal{Q}}(\mathbf{w}^{\mathcal{Q}})$.

Proof. First of all, we have,

$$\begin{aligned}\mathcal{R}_{\mathcal{Q}}(\mathbf{w}^P) - \mathcal{R}^*_{\mathcal{Q}}(\mathbf{w}^{\mathcal{Q}}) &= \Big(\mathcal{R}_{\mathcal{Q}}(\mathbf{w}^P) - \mathcal{R}_{\mathcal{P}}(\mathbf{w}^P)\Big) + \mathcal{R}_{\mathcal{P}}(\mathbf{w}^P) - \mathcal{R}^*_{\mathcal{Q}}(\mathbf{w}^{\mathcal{Q}}) \\ &\le G \cdot \text{dist}(\mathcal{P}, \mathcal{Q}) + \mathcal{R}_{\mathcal{P}}(\mathbf{w}^P) - \mathcal{R}^*_{\mathcal{Q}}(\mathbf{w}^{\mathcal{Q}}) \\ &\le \underbrace{G \cdot \text{dist}(\mathcal{P}, \mathcal{Q})}_{\text{distribution shift}} + \underbrace{\Big(\mathcal{R}_{\mathcal{P}}(\mathbf{w}^P) - \mathcal{R}_P(\mathbf{w}^P)\Big)}_{\text{Hoeffding's inequality}} + \underbrace{\Big(\mathcal{R}_P(\mathbf{w}^P) - \mathcal{R}^*_{\mathcal{Q}}(\mathbf{w}^{\mathcal{Q}})\Big)}_{\text{overfitting \& ideal case}}\end{aligned}$$

For the first term (distribution shift), we have

$$\begin{aligned}\mathcal{R}_{\mathcal{Q}}(\mathbf{w}^P) - \mathcal{R}_{\mathcal{P}}(\mathbf{w}^P) &= \mathbb{E}_{z\sim\mathcal{Q}}[\ell(\mathbf{w}^P, z)] - \mathbb{E}_{z\sim\mathcal{P}}[\ell(\mathbf{w}^P, z)] \\ &= \sum_{i=1}^{I} p^{(\mathcal{Q})}(\pi = i)\mathbb{E}_{z\sim\pi_i}[\ell(\mathbf{w}^P, z)] - \sum_{i=1}^{I} p^{(\mathcal{P})}(\pi = i)\mathbb{E}_{z\sim\pi_i}[\ell(\mathbf{w}^P, z)] \\ &\le \sum_{i=1}^{I} |p^{(\mathcal{P})}(\pi = i) - p^{(\mathcal{Q})}(\pi = i)|\mathbb{E}_{z\sim\pi_i}[\ell(\mathbf{w}^P, z)] \\ &\le G_P \cdot \text{dist}(\mathcal{P}, \mathcal{Q}).\end{aligned}$$

where we define $\text{dist}(\mathcal{P}, \mathcal{Q}) = \sum_{i=1}^{I} |p^{(\mathcal{P})}(\pi = i) - p^{(\mathcal{Q})}(\pi = i)|$ and $\mathbb{E}_{z\sim\pi_i}[\ell(\mathbf{w}^P, z)] \le G_P, \forall i \in I$ because of Assumption 5.2. For the second term, with probability at least $1-\delta$, we have $|\mathcal{R}_{\mathcal{P}}(\mathbf{w}^P) - \mathcal{R}_P(\mathbf{w}^P)| \le \sqrt{\frac{\log(4/\delta)}{2|P|}}$. Note that the third term $\mathcal{R}_P(\mathbf{w}^P) - \mathcal{R}^*_{\mathcal{Q}}(\mathbf{w}^{\mathcal{Q}})$ can be regarded as a constant ϖ.because $\mathcal{R}_P(\mathbf{w}^P)$ is the empirical risk and $\mathcal{R}^*_{\mathcal{Q}}(\mathbf{w}^{\mathcal{Q}})$ is the ideal minimal empirical risk of model $\mathbf{w}^{\mathcal{Q}}$ trained on distribution $\mathcal{Q}$.

Therefore, with probability at least $1-\delta$, given model $\mathbf{w}^P$,

$$\mathcal{R}_{\mathcal{Q}}(\mathbf{w}^P) - \mathcal{R}^*_{\mathcal{Q}}(\mathbf{w}^{\mathcal{Q}}) \le G_P \cdot \text{dist}(\mathcal{P}, \mathcal{Q}) + \sqrt{\frac{\log(4/\delta)}{2|P|}} + \varpi.$$

□

Lemma D.2. *(Group model gap) Suppose Assumptions 5.1 and 5.2 hold for empirical risk $\mathcal{R}_P(\cdot)$. The initial learning rate $\eta_0^2 < \frac{1}{\sqrt{2}TL}$, where T denotes the number of training epochs. Then, we have*

$$\|\mathbf{w}^k - \mathbf{w}^P\|^2 \le 2LG^2\bigg(\sum_{i=1}^{I}\Big|p^{(k)}(\pi = i) - p^{(P)}(\pi = i)\Big|\bigg)^2.$$

Proof. According to the above definition, we similarly define the following empirical risk $\mathcal{R}_{P_k}(\mathbf{w})$ over group k's data P_k by splitting samples according to their marginal distributions, shown as follows.

$$\mathcal{R}_{P_k}(\mathbf{w}) := \sum_{i=1}^{I} p^{(k)}(\pi = i)\mathbb{E}_{z\sim\pi_i}[\ell(\mathbf{w}, z)].$$

Let η_t indicate the learning rate of epoch t. Then, for each epoch t, group k's optimizer performs SGD as follows:

$$\mathbf{w}_t^k = \mathbf{w}_{t-1}^k - \eta_t \sum_{i=1}^{I} p^{(k)}(\pi = i)\nabla_{\mathbf{w}}\mathbb{E}_{z\sim\pi_i}[\ell(\mathbf{w}_{t-1}^k, z)].$$

For any epoch $t+1$, we have

$$
\begin{aligned}
&\|\mathbf{w}_{t+1}^k - \mathbf{w}_{t+1}^P\|^2 \\
&= \|\mathbf{w}_t^k - \eta_t \sum_{i=1}^{I} p^{(k)}(\pi=i)\nabla_{\mathbf{w}}\mathbb{E}_{z\sim\pi_i}[\ell(\mathbf{w}_t^k, z)] - \mathbf{w}_t^P + \eta_t \sum_{i=1}^{I} p^{(P)}(\pi=i)\nabla_{\mathbf{w}}\mathbb{E}_{z\sim\pi_i}[\ell(\mathbf{w}_t^P, z)]\|^2 \\
&\leq \|\mathbf{w}_t^k - \mathbf{w}_t^P\|^2 + \eta_t^2 \|\sum_{i=1}^{I} p^{(k)}(\pi=i)\nabla_{\mathbf{w}}\mathbb{E}_{z\sim\pi_i}[\ell(\mathbf{w}_t^k, z)] - \sum_{i=1}^{I} p^{(P)}(\pi=i)\nabla_{\mathbf{w}}\mathbb{E}_{z\sim\pi_i}[\ell(\mathbf{w}_t^P, z)]\|^2 \\
&\leq \|\mathbf{w}_t^k - \mathbf{w}_t^P\|^2 + 2\eta_t^2 \|\sum_{i=1}^{I} p^{(P)}(\pi=i) L_{\pi_i} \Big[\nabla_{\mathbf{w}}\mathbb{E}_{z\sim\pi_i}[\ell(\mathbf{w}_t^k, z)] - \nabla_{\mathbf{w}}\mathbb{E}_{z\sim\pi_i}[\ell(\mathbf{w}_t^P, z)]\Big]\|^2 \\
&\quad + 2\eta_t^2 \|\sum_{i=1}^{I} \Big(p^{(k)}(\pi=i) - p^{(P)}(\pi=i)\Big)\nabla_{\mathbf{w}}\mathbb{E}_{z\sim\pi_i}[\ell(\mathbf{w}_t^P, z)]\|^2 \\
&\leq \|\mathbf{w}_t^k - \mathbf{w}_t^P\|^2 + 2\eta_t^2 \Big(\sum_{i=1}^{I} p^{(k)}(\pi=i) L_{\pi_i}\Big)^2 \|\mathbf{w}_t^k - \mathbf{w}_t^P\|^2 \\
&\quad + 2L\eta_t^2 g_{max}^2(\mathbf{w}_t^Q)\Big(\sum_{i=1}^{I} |p^{(k)}(\pi=i) - p^{(P)}(\pi=i)|\Big)^2 \\
&\leq \Big(1 + 2\eta_t^2 \Big(\sum_{i=1}^{I} p^{(k)}(\pi=i) L_{\pi_i}\Big)^2\Big)\|\mathbf{w}_t^k - \mathbf{w}_t^P\|^2 \\
&\quad + 2L\eta_t^2 g_{max}^2(\mathbf{w}_t^Q)\Big(\sum_{i=1}^{I} |p^{(k)}(\pi=i) - p^{(P)}(\pi=i)|\Big)^2 \\
&\leq (1 + 2\eta_t^2 L^2)\|\mathbf{w}_t^k - \mathbf{w}_t^P\|^2 + 2L\eta_t^2 G^2 \Big(\sum_{i=1}^{I} |p^{(k)}(\pi=i) - p^{(P)}(\pi=i)|\Big)^2.
\end{aligned}
$$

where the third inequality holds since we assume that $\nabla_{\mathbf{w}}\mathbb{E}_{z\sim\pi_i}[\ell(\mathbf{w}, z)]$ is L_{π_i}-Lipschitz continuous, *i.e.*, $||\nabla_{\mathbf{w}}\mathbb{E}_{z\sim\pi_i}[\ell(\mathbf{w}_t^k, z)] - \nabla_{\mathbf{w}}\mathbb{E}_{z\sim\pi_i}[\ell(\mathbf{w}_t^P, z)]|| \leq L_{\pi_i}||\mathbf{w}_t^k - \mathbf{w}_t^P||$, and denote $g_{max}(\mathbf{w}_t^P) = \max_{i=1}^{I} ||\nabla_{\mathbf{w}}\mathbb{E}_{z\sim\pi_i}[\ell(\mathbf{w}_t^P, z)]||$. The last inequality holds because the above-mentioned assumption that $L = L_{\pi_i} = L_{\pi}, \forall i \in I$, *i.e.*, Lipschitz-continuity will not be affected by the samples' classes. Then, $g_{max}(\mathbf{w}_t^P) \leq G$ because of Assumption 5.2.

For T training epochs, we have

$$
\begin{aligned}
&\|\mathbf{w}_T^k - \mathbf{w}_T^P\|^2 \\
&\leq (1 + 2\eta_T^2 L^2)\|\mathbf{w}_{T-1}^k - \mathbf{w}_{T-1}^P\|^2 + 2L\eta_T^2 G^2 \Big(\sum_{i=1}^{I} |p^{(k)}(\pi=i) - p^{(P)}(\pi=i)|\Big)^2 \\
&\leq \prod_{t=0}^{T}(1 + 2\eta_t^2 L^2)^t \|\mathbf{w}_0^k - \mathbf{w}_0^P\|^2 + 2LG^2 \sum_{t=0}^{T}(\eta_t^2(1 + 2\eta_t^2 L^2))^{T-t}\Big(\sum_{i=1}^{I} |p^{(k)}(\pi=i) - p^{(P)}(\pi=i)|\Big)^2 \\
&\leq 2LG^2 \sum_{t=0}^{T}(\eta_t^2(1 + 2\eta_t^2 L^2))^{T-t}\Big(\sum_{i=1}^{I} |p^{(k)}(\pi=i) - p^{(P)}(\pi=i)|\Big)^2.
\end{aligned}
$$

where the last inequality holds because the initial models are the same, i.e., $\mathbf{w}_0 = \mathbf{w}_0^k = \mathbf{w}_0^P, \forall k$. When the condition $\eta_t^2 < \eta_0^2 < \frac{1}{\sqrt{2}TL}$ satisfies, $2LG^2 \sum_{t=0}^{T}(\eta_t^2(1 + 2\eta_t^2 L^2))^{T-t}$ can be simplifited as $2LG^2$, which is independent of the learning algorithm. This condition is easy to be satisfied since the learning rate η_t is a small value (< 0.0001) and usually set to be decay with the training epoch (i.e., $\eta_{t+1} \leq \eta_t$). □

Lemma D.3. *With probability at least* $1-\delta$*, given the model* $\mathbf{w}^k$ *trained on group* k*'s dataset* P_k*, we have*

$$\mathcal{R}_{\mathcal{Q}_k}(\mathbf{w}^k) - \mathcal{R}^*_{\mathcal{Q}_k}(\mathbf{w}^{\mathcal{Q}_k}) \leq G_k \cdot \mathit{dist}(\mathcal{P}_k, \mathcal{Q}_k) + \sqrt{\frac{\log(4/\delta)}{2|P_k|}} + \varpi_k.$$

where $\mathit{dist}(\mathcal{P}_k, \mathcal{Q}_k) = \sum_{i=1}^{I} |p^{(\mathcal{P}_k)}(\pi = i) - p^{(\mathcal{Q}_k)}(\pi = i)|$ *and* $\mathbb{E}_{z\sim\pi_i}[\ell(\mathbf{w}^k, z)] \leq G_k, \forall i \in I$*, and* $\varpi_k := \mathcal{R}_{P_k}(\mathbf{w}^k) - \mathcal{R}^*_{\mathcal{Q}_k}(\mathbf{w}^{\mathcal{Q}_k})$.

Proof. Building upon the proof idea presented in Lemma D.1, for completeness, we provide a full proof here. Firstly, we have,

$$\begin{aligned}
&\mathcal{R}_{\mathcal{Q}_k}(\mathbf{w}^k) - \mathcal{R}^*_{\mathcal{Q}_k}(\mathbf{w}^{\mathcal{Q}_k}) \\
&= \underbrace{(\mathcal{R}_{\mathcal{Q}_k}(\mathbf{w}^k) - \mathcal{R}_{\mathcal{P}_k}(\mathbf{w}^k))}_{\text{distribution shift}} + \underbrace{(\mathcal{R}_{\mathcal{P}_k}(\mathbf{w}^k) - \mathcal{R}_{P_k}(\mathbf{w}^k))}_{\text{Hoeffding's inequality}} + \underbrace{(\mathcal{R}_{P_k}(\mathbf{w}^k) - \mathcal{R}^*_{\mathcal{Q}_k}(\mathbf{w}^{\mathcal{Q}_k}))}_{\text{overfitting \& ideal case}}
\end{aligned}$$

For the first term, we have

$$\begin{aligned}
&\mathcal{R}_{\mathcal{Q}_k}(\mathbf{w}^k) - \mathcal{R}_{\mathcal{P}_k}(\mathbf{w}^k) \\
&= \sum_{i=1}^{I} p^{(\mathcal{Q}_k)}(\pi = i)\mathbb{E}_{z\sim\pi_i}[\ell(\mathbf{w}^k, z)] - \sum_{i=1}^{I} p^{(\mathcal{P}_k)}(\pi = i)\mathbb{E}_{z\sim\pi_i}[\ell(\mathbf{w}^k, z)] \\
&\leq \sum_{i=1}^{I} |p^{(\mathcal{P}_k)}(\pi = i) - p^{(\mathcal{Q}_k)}(\pi = i)|\mathbb{E}_{z\sim\pi_i}[\ell(\mathbf{w}^k, z)] \\
&\leq G_k \cdot \mathrm{dist}(\mathcal{P}_k, \mathcal{Q}_k).
\end{aligned}$$

where $\mathrm{dist}(\mathcal{P}_k, \mathcal{Q}_k) := \sum_{i=1}^{I} |p^{(\mathcal{P}_k)}(\pi = i) - p^{(\mathcal{Q}_k)}(\pi = i)|$ and $\mathbb{E}_{z\sim\pi_i}[\ell(\mathbf{w}^k, z)] \leq G_k, \forall i \in I$ due to Assumption 5.2. Recall that the constant G_k clarifies the bound of loss on the corresponding model $\mathbf{w}^k$. For the second term, with probability at least $1-\delta$, we have $|\mathcal{R}_{\mathcal{P}_k}(\mathbf{w}^k) - \mathcal{R}_{P_k}(\mathbf{w}^k)| \leq \sqrt{\frac{\log(4/\delta)}{2|P_k|}}$. For the third term, we define $\varpi_k := \mathcal{R}_{P_k}(\mathbf{w}^k) - \mathcal{R}^*_{\mathcal{Q}_k}(\mathbf{w}^{\mathcal{Q}_k})$, which can be regarded as a constant. This is because $\mathcal{R}_{P_k}(\mathbf{w}^k)$ represents empirical risk and $\mathcal{R}^*_{\mathcal{Q}_k}(\mathbf{w}^{\mathcal{Q}_k})$ is the ideal minimal empirical risk of model $\mathbf{w}^{\mathcal{Q}_k}$ trained on sub-distribution $\mathcal{Q}_k$.

Therefore, with probability at least $1-\delta$, given model $\mathbf{w}^k$,

$$\mathcal{R}_{\mathcal{Q}_k}(\mathbf{w}^k) - \mathcal{R}^*_{\mathcal{Q}_k}(\mathbf{w}^{\mathcal{Q}_k}) \leq G_k \cdot \mathrm{dist}(\mathcal{P}_k, \mathcal{Q}_k) + \sqrt{\frac{\log(4/\delta)}{2|P_k|}} + \varpi_k.$$

□

E More experimental results

E.1 Datasets and parameter settings

We empirically evaluate FIS on the CelebA dataset, an image dataset commonly used in the fairness literature [49]. We also evaluate FIS on two tabular datasets: UCI Adult [8] and Compas dataset [6].

E.1.1 CelebA dataset

Dataset details CelebA [49] is an image dataset with 202,599 celebrity face images annotated with 40 attributes, including gender, hair colour, age, smiling, etc. The sensitive attribute is gender: $S = \{\text{Men, Women}\}$. We select four binary classification targets, including smiling, attractive, young, and big nose. For example, the task is to predict whether a person in an image is young ($Y = 1$) or non-young ($Y = 0$), among other attribute predictions.

Hyper-parameter details In all our experiments using the CelebA dataset, we train a vision transformer with patch size $(8, 8)$ using SGD optimizer and a batch size of 128. The epochs are split into two phases: warm-up epochs (5 epochs) and training epochs (10 epochs). The default label budget per round, which represents the number of solicited data samples, is set to 256. Additionally, the default values for learning rate, momentum, and weight decay are 0.01, 0.9, and 0.0005, respectively. We initially allocate 2% of the training set for training purposes and the remaining 98% for sampling. Then, we randomly select 10% of the test data for validation. For JTT, we explore 10% data for training purposes with weights $\lambda = 20$ for retraining misclassified examples.

E.1.2 UCI Adult dataset

Dataset details. The Adult dataset [8] predicts whether an individual's annual income falls below or exceeds $50K$, denoted as $Y = 0$ and $Y = 1$, respectively. This prediction is based on a variety of continuous and categorical attributes, including education level, age, gender, occupation, etc. The default sensitive attribute in this dataset is gender $S = \{\text{Men, Women}\}$ [80]. In particular, we also group this dataset using age $S = \{\text{Teenager, Non-teenager}\}$. To achieve a balanced age distribution in the dataset, individuals with an age of less than 30 are grouped as "Teenagers". The dataset contains a total of 45,000 instances. The dataset exhibits an imbalance: there are twice as many men as women, and only 15% of those with high incomes are women.

Hyper-parameter details. In the experiments using the Adult dataset, we train a two-layer ReLU network with a hidden size of 64. The epochs are split into two phases: warm-up epochs (100 epochs) and training epochs (60 epochs). The default label budget per round, which represents the number of solicited data samples, is set to 1024. Additionally, the default values for learning rate, momentum, and weight decay are 0.00001, 0.9, and 0.0005, respectively. We resample the datasets to balance the class and group membership [17]. The dataset is randomly split into a train and a test set in a ratio of 80 to 20. Then, we randomly re-select 20% of the training set for initial training and the remaining 80% for sampling. Also, 20% examples of the test set are selected to form a validation set. We utilize the whole model to compute the prediction influence and fairness for examples. Then, we randomly select 10% of the test data for validation. For JTT, we explore 30% data for training purposes with weights $\lambda = 20$ for retraining misclassified examples.

E.1.3 Compas dataset

Dataset details. Compas dataset, also known as the Correctional Offender Management Profiling for Alternative Sanctions dataset, is a collection of data related to criminal defendants. It contains information on approximately 6,172 individuals who were assessed for risk of re-offending. The primary task associated with this dataset is predicting whether a defendant will re-offend ($Y = 1$) or not ($Y = 0$) within a certain time frame after their release. The sensitive attribute is often considered to be race, specifically whether the individual is classified as African American or not.

Hyper-parameter details. In the experiments using the Compas dataset, we train a multi-layer neural network with one hidden layer consisting of 64 neurons. The epochs are split into two phases: warm-up epochs (20 epochs) and training epochs (50 epochs). The default label budget per round, which represents the number of solicited data samples, is set to 128. Furthermore, the default values for learning rate, momentum, and weight decay are 0.01, 0.9, and 0.0005, respectively. We resample the datasets to balance the class and group membership [17]. The dataset is initially split into training and test sets at an 80-20 ratio. Then, we further split 20% of the training set for initial training, reserving the remaining 80% for sampling. Additionally, 20% of the test set is selected to create a validation set. We use the entire model to calculate prediction influence and evaluate fairness for the dataset examples.

E.2 Full version of experimental results

E.3 Exploring the impact of label budgets

In our study, we examine how varying label budgets r influence the balance between accuracy and fairness. We present the results of test accuracy and fairness disparity across different label budgets on the CelebA, Compas, and Adult datasets. In these experiments, we use the demographics parity

Table 2: We report the (`test_accuracy`, `fairness_violation`) for evaluating the performance on the **CelebA dataset** with two binary classification targets `Young` and `Big Nose`. We select `gender` as the sensitive attribute.

$\epsilon = 0.05$	**CelebA - Young**		
	(Test_acc↑, DP↓)	(Test_acc↑, EOp↓)	(Test_acc↑, EOd↓)
Base(ERM)	(0.755 ± 0.002, 0.190 ± 0.017)	(0.759 ± 0.005, 0.102 ± 0.005)	(0.755 ± 0.002, 0.182 ± 0.018)
Random	(0.763 ± 0.008, 0.158 ± 0.016)	(0.698 ± 0.109, 0.075 ± 0.021)	(0.766 ± 0.011, 0.156 ± 0.017)
BALD	(0.776 ± 0.021, 0.165 ± 0.019)	(0.775 ± 0.020, 0.076 ± 0.007)	(0.779 ± 0.005, 0.162 ± 0.021)
ISAL	(0.781 ± 0.020, 0.180 ± 0.014)	(0.781 ± 0.020, 0.084 ± 0.006)	(0.780 ± 0.021, 0.173 ± 0.012)
JTT-20	(0.774 ± 0.026, 0.167 ± 0.016)	(0.774 ± 0.024, 0.083 ± 0.007)	(0.772 ± 0.023, 0.171 ± 0.025)
FIS	(0.763 ± 0.004, 0.104 ± 0.059)	(0.773 ± 0.003, 0.041 ± 0.015)	(0.763 ± 0.005, 0.118 ± 0.074)

$\epsilon = 0.05$	**CelebA - Big Nose**		
	(Test_acc↑, DP↓)	(Test_acc↑, EOp↓)	(Test_acc↑, EOd↓)
Base(ERM)	(0.752 ± 0.024, 0.198 ± 0.034)	(0.755 ± 0.022, 0.206 ± 0.018)	(0.755 ± 0.022, 0.183 ± 0.029)
Random	(0.760 ± 0.009, 0.177 ± 0.014)	(0.757 ± 0.004, 0.190 ± 0.029)	(0.759 ± 0.006, 0.167 ± 0.029)
BALD	(0.777 ± 0.004, 0.184 ± 0.016)	(0.765 ± 0.003, 0.209 ± 0.014)	(0.770 ± 0.004, 0.170 ± 0.015)
ISAL	(0.782 ± 0.001, 0.148 ± 0.059)	(0.782 ± 0.001, 0.154 ± 0.080)	(0.779 ± 0.006, 0.145 ± 0.065)
JTT-20	(0.771 ± 0.014, 0.191 ± 0.036)	(0.758 ± 0.026, 0.223 ± 0.018)	(0.764 ± 0.019, 0.193 ± 0.016)
FIS	(0.779 ± 0.009, 0.089 ± 0.076)	(0.780 ± 0.013, 0.046 ± 0.072)	(0.772 ± 0.015, 0.062± 0.081)

Table 3: The performance results of (`test_accuracy`, `fairness_violation`) on the **Adult dataset**. The sensitive attribute is `age`.

$\epsilon = 0.05$	**Income (age)**		
	(Test_acc↑, DP↓)	(Test_acc↑, EOp↓)	(Test_acc↑, EOd↓)
Base(ERM)	(0.665 ± 0.045, 0.255 ± 0.041)	(0.665 ± 0.045, 0.115 ± 0.036)	(0.665 ± 0.045, 0.158 ± 0.030)
Random	(0.765 ± 0.021, 0.209 ± 0.042)	(0.758 ± 0.027, 0.127 ± 0.013)	(0.764 ± 0.018, 0.133 ± 0.027)
BALD	(0.767 ± 0.019, 0.203 ± 0.017)	(0.703 ± 0.111, 0.117 ± 0.013)	(0.763 ± 0.022, 0.128 ± 0.014)
ISAL	(0.765 ± 0.020, 0.215 ± 0.011)	(0.755 ± 0.028, 0.128 ± 0.013)	(0.761 ± 0.024, 0.138 ± 0.009)
JTT-20	(0.751 ± 0.013, 0.262 ± 0.020)	(0.742 ± 0.018, 0.149 ± 0.021)	(0.745 ± 0.014, 0.171 ± 0.012)
FIS	(0.766 ± 0.013, 0.214 ± 0.009)	(0.757 ± 0.034, 0.113 ± 0.017)	(0.763 ± 0.011, 0.143 ± 0.023)

Table 4: The performance results of (`test_accuracy`, `fairness_violation`) on the **Compas dataset**. The selected sensitive attribute is `race`.

$\epsilon = 0.05$	**Recidivism**		
	(Test_acc↑, DP↓)	(Test_acc↑, EOp↓)	(Test_acc↑, EOd↓)
Base(ERM)	(0.675 ± 0.005, 0.333 ± 0.008)	(0.675 ± 0.005, 0.267 ± 0.010)	(0.675 ± 0.005, 0.284 ± 0.010)
Random	(0.689 ± 0.007, 0.305 ± 0.023)	(0.686 ± 0.016, 0.253 ± 0.035)	(0.688 ± 0.006, 0.256 ± 0.023)
BALD	(0.688 ± 0.011, 0.313 ± 0.012)	(0.686 ± 0.015, 0.256 ± 0.031)	(0.688 ± 0.011, 0.263 ± 0.011)
ISAL	(0.697 ± 0.002, 0.308 ± 0.025)	(0.698 ± 0.004, 0.274 ± 0.022)	(0.697 ± 0.001, 0.260 ± 0.026)
JTT-20	(0.646 ± 0.009, 0.240 ± 0.016)	(0.630 ± 0.024, 0.141 ± 0.028)	(0.646 ± 0.009, 0.200 ± 0.007)
FIS	(0.690 ± 0.002, 0.299 ± 0.029)	(0.694 ± 0.002, 0.241 ± 0.035)	(0.698 ± 0.005, 0.252 ± 0.030)

(DP) as our fairness metric. For convenience, we maintain a fixed label budget per round, using rounds of label budget allocation to demonstrate its impact. The designated label budgets per round for the CelebA, Compas, and Adult are 256, 128, and 512, respectively. In the following figures, the x-axis is both the number of label budget rounds. The y-axis for the left and right sub-figures are test accuracy and DP gap, respectively. As observed in Figures 5-7, compared to the three baselines (BALD, JTT-20, and ISAL), our approach substantially reduces the DP gap without sacrificing test accuracy.

Specifically, on the Adult dataset, both accuracy and fairness violation converge to similar numerical values when the budget is lower than 20, suggesting a potential overfitting of the model to insufficient training examples. With a larger budget, our algorithms outperform other baseline methods, achieving higher accuracy and lower demographic disparity.

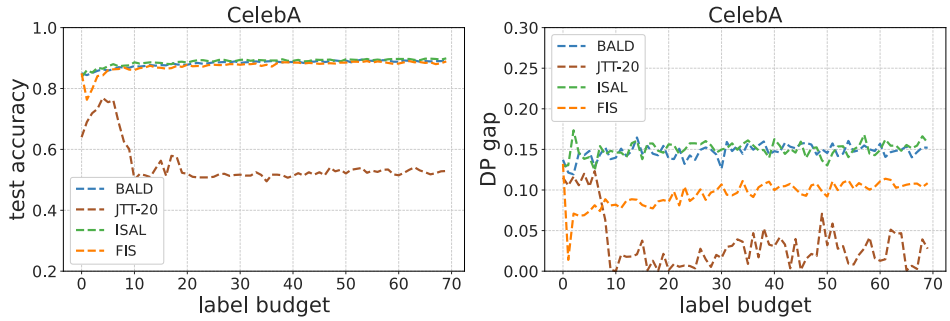


Figure 5: The impact of label budgets on the test accuracy & DP gap in the CelebA dataset. The binary classification targets is `Smiling`.

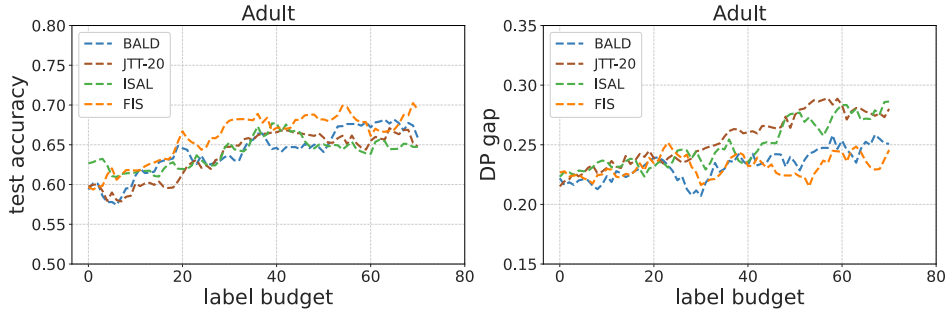


Figure 6: The impact of label budgets on the test accuracy & DP gap in the Adult dataset. The sensitive attribute is `sex`.

E.4 The role of validation dataset size

In this subsection, we explore the impact of adjusting the validation set size on our algorithm's performance. We present the test accuracy and fairness disparity across different validation set sizes on the CelebA, Compas, and Adult datasets. Note that the default validation set size for image and tabular datasets is set to 1% and 4% of the whole dataset size, respectively. That is, the default validation set sizes are 1996 (CelebA), 1800 (Adult), and 247 (Compas) instances, respectively. In particular, given the smaller size of the default validation set, the minimum scale of the validation set size is set to $1/5\times$ (nearly 400 CelebA images). Tables 5 and 6 present the performance results on the CelebA, UCI Adult, and Compas datasets, respectively.

Table 5: The performance results of (`test_accuracy`, `fairness_violation`) on the **CelebA dataset** when the validation set size is reduced to $1/2\times$ and $1/5\times$. Our algorithm retains the test accuracy and fairness violation when we vary the validation set size.

$\epsilon = 0.05$	**CelebA - Smiling**			**CelebA - Attractive**		
	(Test_acc↑, DP↓)	(Test_acc↑, EOp↓)	(Test_acc↑, EOd↓)	(Test_acc↑, DP↓)	(Test_acc↑, EOp↓)	(Test_acc↑, EOd↓)
$1\times$	(0.848, 0.084)	(0.876, 0.031)	(0.864, 0.030)	(0.680, 0.285)	(0.695, 0.148)	(0.692, 0.148)
$1/2\times$	(0.872, 0.105)	(0.891, 0.042)	(0.880, 0.028)	(0.648, 0.249)	(0.688, 0.188)	(0.678, 0.147)
$1/5\times$	(0.872, 0.117)	(0.863, 0.057)	(0.886, 0.028)	(0.604, 0.171)	(0.707, 0.209)	(0.645, 0.145)
$\epsilon = 0.05$	**CelebA - Young**			**CelebA - Big_Nose**		
	(Test_acc↑, DP↓)	(Test_acc↑, EOp↓)	(Test_acc↑, EOd↓)	(Test_acc↑, DP↓)	(Test_acc↑, EOp↓)	(Test_acc↑, EOd↓)
$1\times$	(0.766, 0.139)	(0.775, 0.043)	(0.769, 0.168)	(0.771, 0.156)	(0.765, 0.129)	(0.758, 0.155)
$1/2\times$	(0.735, 0.093)	(0.762, 0.067)	(0.769, 0.055)	(0.771, 0.054)	(0.761, 0.162)	(0.748, 0.096)
$1/5\times$	(0.743, 0.107)	(0.780, 0.097)	(0.757, 0.166)	(0.772, 0.095)	(0.750, 0.300)	(0.760, 0.156)

E.5 Benchmarking model performance with validation set enhancements

Note that we resort to an additional validation set for developing FIS. To demonstrate FIS's advantage at the same levels of information, we introduce a new baseline called `Random+Val`. This method

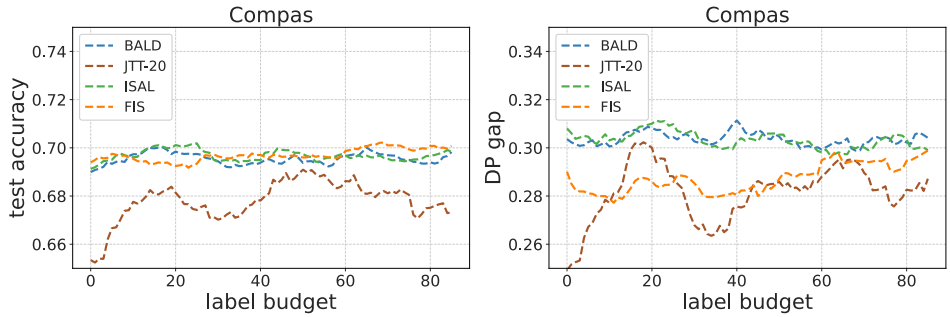


Figure 7: The impact of label budgets on the test accuracy & DP gap in the Compas dataset. The sensitive attribute is `race`.

Table 6: We examine the performance results of (`test_accuracy`, `fairness_violation`) on the tabular datasets (**Left**: Adult; **Right**: Compas) when the validation set size is reduced to $1/2\times$, $1/4\times$, and $1/20\times$. We observe that our algorithm still retains the test accuracy and fairness violation when we vary the validation set size.

$\epsilon = 0.05$	**Adult - Income(Age)**			**Compas - Recidivism**		
	(Test_acc↑, DP↓)	(Test_acc↑, EOp↓)	(Test_acc↑, EOd↓)	(Test_acc↑, DP↓)	(Test_acc↑, EOp↓)	(Test_acc↑, EOd↓)
$1\times$	(0.757, 0.198)	(0.718, 0.124)	(0.750, 0.125)	(0.690, 0.313)	(0.696, 0.249)	(0.702, 0.257)
$1/2\times$	(0.717, 0.259)	(0.634, 0.123)	(0.736, 0.143)	(0.683, 0.270)	(0.680, 0.247)	(0.693, 0.244)
$1/4\times$	(0.749, 0.196)	(0.750, 0.121)	(0.747, 0.137)	(0.677, 0.283)	(0.682, 0.276)	(0.680, 0.244)
$1/20\times$	(0.721, 0.205)	(0.706, 0.148)	(0.706, 0.179)	(0.689, 0.289)	(0.668, 0.236)	(0.683, 0.252)

involves continuing to train the model with a randomly sampled validation set. Specifically, we start with the `Random`'s last saved checkpoint and train it further using the validation set. In particular, we would incorporate a fairness regularizer with dynamic weight to train the model using validation data, considering its sensitive attributes to reduce fairness disparity. Due to its small size, we limit training to 10 epochs to avoid overfitting. The performance results of `Random`, `Random+Val`, and FIS are given in Table 7.

Table 7: Comparative analysis of (`test_accuracy`, `fairness_violation`) in the CelebA, Adult and Compas datasets. The table illustrates that even at the same information level (using the validation set to train), FIS can obtain better performances. Similarly, we highlight all the fairer but without sacrificing accuracy results in boldface compared to `Random`.

$\epsilon = 0.05$	**CelebA - Smiling**			**CelebA - Attractive**		
	(Test_acc↑, DP↓)	(Test_acc↑, EOp↓)	(Test_acc↑, EOd↓)	(Test_acc↑, DP↓)	(Test_acc↑, EOp↓)	(Test_acc↑, EOd↓)
Random	(0.853, 0.132)	(0.863, 0.053)	(0.861, 0.031)	(0.696, 0.367)	(0.708, 0.253)	(0.696, 0.243)
Random + Val	(0.801, 0,115)	(0.872, 0.139)	(0.879, 0.153)	(0.638, 0.199)	(0.699, 0.366)	(0.699, 0.355)
FIS	**(0.877, 0.122)**	**(0.886, 0.040)**	**(0.882, 0.023)**	**(0.680, 0.285)**	**(0.695, 0.148)**	**(0.692, 0.148)**
$\epsilon = 0.05$	**Adult - Income(Age)**			**Compas - Recidivism**		
	(Test_acc↑, DP↓)	(Test_acc↑, EOp↓)	(Test_acc↑, EOd↓)	(Test_acc↑, DP↓)	(Test_acc↑, EOp↓)	(Test_acc↑, EOd↓)
Random	(0.745, 0.236)	(0.729, 0.136)	(0.748, 0.151)	(0.696, 0.316)	(0.698, 0.258)	(0.694, 0.269)
Random + Val	**(0.762, 0.161)**	(0.743, 0.266)	**(0.788, 0.131)**	(0.604, 0.136)	(0.623, 0.147)	(0.628, 0.153)
FIS	**(0.751, 0.205)**	**(0.725, 0.130)**	**(0.750, 0.125)**	**(0.688, 0.316)**	**(0.695, 0.250)**	**(0.693, 0.254)**

F Potential privacy leakage from non-demographic and demographic feature correlations

Note that some non-demographic features may correlate with sensitive attributes (demographic information), which could lead to potential privacy leakage issues [82]. However, it is crucial to clarify that our method specifically limits its query to the data. We only query the true labels for a selected subset of unlabeled examples. While correlations between non-demographic variables and sensitive attributes may exist in the underlying data, our method itself does not introduce any additional privacy leakage beyond what is inherent in the original dataset. On the other hand, our

work's primary focus is not on addressing privacy concerns related to demographic information. Rather, our main objective is to reduce fairness disparities while maintaining a favorable utility trade-off. We achieve this without explicitly incorporating any additional sensitive information into the process and without directly engaging with the complex privacy implications of demographic data usage.

To address this potential privacy concern, we can analyze it via differential privacy. Consider a function $\phi(\cdot)$ that maps non-demographic information (feature $\boldsymbol{x}$ and label y) to a sensitive attribute s. Due to insufficient information, $\phi(\cdot)$ is not deterministic, making it unable to precisely estimate s. If a deterministic mapping function existed, it would unavoidably lead to privacy leakage. Therefore, we assume that given the same features and label, the function $\phi(\cdot)$ output might vary. For example, in the Adult dataset, $\boldsymbol{x}$ includes age and credit history, y is the income, and s is the gender. In practice, both men and women have a probability of earning more than 50K (i.e., $y = 1$) given the same age, credit history, etc (same feature $\boldsymbol{x}$). Suppose $P(\phi(\boldsymbol{x}, y) = s \mid S = s, \boldsymbol{x}, y) \leq 1 - \epsilon_0$ and $P(\phi(\boldsymbol{x}, y) = s \mid S = s', \boldsymbol{x}, y) \geq \epsilon_1$ for all $\boldsymbol{x}, y, s, s'$ (where $s \neq s'$). Here, S represents the sensitive attribute variable, which is unknown to $\phi(\cdot)$. Then, we have

$$\frac{\mathbb{P}(\phi(\boldsymbol{x}, y) = \tilde{s} | S = s, \boldsymbol{x}, y)}{\mathbb{P}(\phi(\boldsymbol{x}, y) = \tilde{s} | S = s', \boldsymbol{x}, y)} \leq \frac{\max \mathbb{P}(\phi(\boldsymbol{x}, y) = \tilde{s} \mid S = s, \boldsymbol{x}, y)}{\min \mathbb{P}(\phi(\boldsymbol{x}, y) = \tilde{s} \mid S = s', \boldsymbol{x}, y)} \leq \frac{1 - \epsilon_0}{\epsilon_1} = e^{\varepsilon}.$$

where $\varepsilon = \ln\left(\frac{1-\epsilon_0}{\epsilon_1}\right)$. In practice, if the mapping function is too strong, i.e. $\ln\left(\frac{1-\epsilon_0}{\epsilon_1}\right)$ is too large, we can add additional noise to reduce their informativeness and therefore better protect privacy.