FAIR MACHINE LEARNING UNDER LIMITED DEMOGRAPHICALLY LABELED DATA

Mustafa Safa Ozdayi
The University of Texas at Dallas
mustafa.ozdayi@utdallas.edu

Murat Kantarcioglu
The University of Texas at Dallas
muratk@utdallas.edu

Rishabh Iyer
The University of Texas at Dallas
rishabh.iyer@utdallas.edu

ABSTRACT

Research has shown that, machine learning models might inherit and propagate undesired social biases encoded in the data. To address this problem, fair training algorithms are developed. However, most algorithms assume we know demographic/sensitive data features such as gender and race. This assumption falls short in scenarios where collecting demographic information is not feasible due to privacy concerns, and data protection policies. A recent line of work develops fair training methods that can function without any demographic feature on the data, that are collectively referred as Rawlsian methods. Yet, we show in experiments that, Rawlsian methods tend to exhibit relatively high bias. Given this, we look at the middle ground between the previous approaches, and consider a setting where we know the demographic attributes for only a small subset of our data. In such a setting, we design fair training algorithms which exhibit both good utility, and low bias. In particular, we show that our techniques can train models to significantly outperform Rawlsian approaches even when 0.1% of demographic attributes are available in the training data. Furthermore, our main algorithm can accommodate multiple training objectives easily. We expand our main algorithm to achieve robustness to label noise in addition to fairness in the limited demographics setting to highlight that property as well.

1 INTRODUCTION

Researchers have illustrated that machine learning (ML) models might exhibit discriminatory behavior. Buolamwini & Gebru (2018) show that many of the commercial face recognition systems tend have higher rates of error on people with darker skin color. Jeff Larson & Angwin (2016) have analyzed COMPAS, a software used by some of the U.S courts to asses defendants' likelihood to reoffend. The authors unveiled that, the software overestimates the likelihood to reoffend for black defendants, and underestimates for white defendants. Finally, it was reported that Amazon had to abandon the use of its ML based recruitment tool because it was disproportionately penalizing the women candidates (Dastin).

To address such issues, many fair training algorithms are developed such as Pleiss et al. (2017); Donini et al. (2018); Celis et al. (2019). At its core, these algorithms try to maximize the utility of the trained models while trying to keep the bias exhibited by them minimal with respect to some bias metric(s). In the recent literature, parity-based metrics, such as equality of odds difference and average of odds difference (Hardt et al. (2016b)), are among the most popular. This is perhaps because these metrics somewhat correspond to the fairness notions put forward by

the government agencies, such as the 80% rule of U.S. Equal Employment Opportunity Commission.[1]

However, the existing parity-based fair training approaches require the demographic/sensitive attributes, such as gender and race, to be available on the data. This renders them unusable in scenarios where collecting demographic attributes from the individuals is not possible due to privacy concerns and certain data regulations, such as European Union's GDPR (EUG). As has been noted in a survey that is conducted with industry practitioners in Holstein et al. (2019), this is a significant problem that needs to be addressed before we can see the adaptation of fair training algorithms by the industry.

To alleviate the aforementioned problem, Hashimoto et al. (2018) and Lahoti et al. (2020) developed fair training algorithms that do not rely on demographic attributes at all. However, as we experimentally show in this work, these *Rawlsian*[2] algorithms tend to exhibit high bias when parity-based metrics are concerned.

In short, existing parity-based fair training approaches require demographic attributes on data, and this limits their usability in real-world scenarios. In contrast, Rawlsian approaches do not require any demographic attribute, but they perform relatively poor as far as parity-based bias metrics are concerned. Given these observations, we consider the middle ground between the previous approaches, and consider a setting where we know the demographic attributes for only a small subset of our data (henceforth, we refer this setting as limited demographics setting). *In summary, our goal is to develop fair training algorithms which perform good with respect to parity-based bias metrics in the limited demographics setting.* As suggested in Holstein et al. (2019), limited demographics setting is realistic for many scenarios. This is because it is likely that the trained models will be validated before they are released to the wild, and validation requires data with demographic information as long as parity-based metrics are concerned[3].

Concretely, we show and evaluate two ways to do fair training in the limited demographics setting. We start with considering a strawman solution that imputes unknown demographic attributes. In other words, we first train a predictor model for demographic attributes on the demographically labeled portion of the data, and then predict and fill the demographic attributes for the rest of the data. Then, any existing parity-based fair training algorithm can be applied. By using this approach, we illustrate the performance of some of the existing parity-based algorithms in the limited demographics setting. Our results suggest that, even if we have only a few dozen demographically labeled data at hand, *the strawman solution tends to outperform the state-of-the-art Rawlsian algorithm* for parity-based bias metrics.

After evaluating the strawman solution, we improve upon it by developing a novel fair training algorithm that is particularly suited to limited demographics setting. By experiments, we show that our algorithm degrades more gracefully than the strawman solution, and exhibits lower bias as the size of demographically labeled data gets smaller. Furthermore, flexibility of our formulation allows us to expand our algorithm to accommodate other training objectives in addition to fairness. For example, we know that real-world data, and sometimes even carefully curated benchmarks such as ImageNet (Deng et al. (2009)) and Cifar10 (Krizhevsky et al. (2009)), might contain label noise as reported by Northcutt et al. (2021). Given this, we expand our algorithm to achieve robustness to label noise in addition to the fairness in the limited demographics setting.
We summarize our contributions below.

- We adapt some of the existing parity-based fair training algorithms to limited demographics setting with a strawman solution, and compare the strawman solution with state-of-the-art

[1]This rule states that, companies should be hiring protected groups at a rate that is at least 80% of that of white men (80R).

[2]These algorithms are referred as Rawlsian because they are based on some min-max optimization. This formulation somewhat corresponds to Rawl's principle of fairness which states (among other things) "Social and economic inequalities are to be arranged so that they are to the **greatest benefit of the least advantaged members** of society, consistent with the just savings principle Rawls (1999)"

[3]For example, in general, creditors may not request or collect information about an applicant's race, color, religion, national origin, or sex. **Exceptions to this rule generally involve situations in which the information is necessary to test for compliance with fair lending rules.** [CFBP Consumer Law and Regulations, 12 CFR §1002.5]

Rawlsian method of Lahoti et al. (2020). We show that, the strawman solution tends to exhibit lower bias even when there is only a few dozen demographically labeled data at hand.

- We develop a novel fair training algorithm, based on bilevel optimization, named *BiFair*, that is particularly suited to limited demographics setting. Our experiments show that BiFair tends to exhibit lower bias compared to the strawman solution as the size of the demographically labeled data gets smaller.
- Our formulation is flexible, and can be easily expanded to contain multiple training objectives. To illustrate the flexibility of our formalization, we extend BiFair to be robust against noise in the labels. This gives us a fair training algorithm that is robust to noisy labels in the limited demographics setting.

We organize the rest of our paper as follows: in Section 2, we provide the necessary background to the reader, and discuss the related work. In Section 3, we present the BiFair algorithm. In Section 4, we present our experiments. In Section 5, we discuss some aspects of our work as well as provide avenues for further research. Finally in Section 6, we recap our work, and conclude the paper.

2 Background and Related Work

2.1 Fairness in ML

Fairness is a multifaceted concept that has different definitions based on context it is considered. Our main focus in this work is supervised classification, and in this domain, parity-based fairness definitions, such as statistical parity, equalized odds and equality of opportunity of Hardt et al. (2016a), are the most prominent in the recent literature. For example, AIF360 of Bellamy et al. (2018), a popular toolkit for fairness research, benchmarks its results by using parity-based bias metrics.

Algorithms that train fair models are typically grouped under three categories: pre-processing, in-processing, and post-processing. Pre-processing methods are applied to the data prior to the training. The goal is to transform the training data in a way such that, when a model is later trained on the transformed data, it exhibits good fairness performance (Kamiran & Calders (2012); Feldman et al. (2015); Zemel et al. (2013); Calmon et al. (2017)). The in-processing methods are applied at the training time, for example by adding regularization terms or encoding hard constraints on the training objective (Kamishima et al. (2012); Bechavod & Ligett (2017); Zhang et al. (2018); Agarwal et al. (2018); Celis et al. (2019); Donini et al. (2018)). Finally, post-processing methods are applied to an already trained model. They try to limit the bias of the model by adjusting the model's outputs directly, such as by negating its output on certain inputs (Kamiran et al. (2012); Pleiss et al. (2017); Hardt et al. (2016b)). The particular algorithm we develop, BiFair, falls into the category of in-processing methods.

Our work differs from the existing fairness literature primarily by the setting which we consider. As mentioned before, most of the previous work assume the existence of demographic attributes on all of the data. This assumption might not be realistic due to privacy concerns and data regulations. On the other hand, some works develop fair training algorithms that can function without any demographic attribute (Hashimoto et al. (2018); Lahoti et al. (2020)), but we show these algorithms fall short in performance when parity-based bias metrics are concerned. In contrast, we develop approaches that perform well for parity-based metrics with limited demographic data.

It is also worth noting the work of Roh et al. (2020a), which develops an algorithm that can train both fair, and robust models. Like this work, we also consider robustness and fairness together, but we do so as an extension of our main contribution. Yet again, our work differs from this work by the limited demographics setting we consider as well.

2.2 Bilevel Optimization

A bilevel optimization is type a nested optimization, where optimality of an *outer* problem is subject to the optimality of an *inner* problem. A general formulation of the bilevel optimization is given below,

$$\min_{x,y} f(x, y^\star) \text{ subject to } y^\star \in \arg\min_{y} g(x, y).$$

Table 1: Various parity-based bias metrics that are used to quantify the bias exhibited by models, and corresponding loss functions that can be plugged into BiFair (see Equation 1). Lower values for bias metrics indicate fairer models. As can be seen, many parity-based bias metrics can simply be expressed as utility loss differences across groups, and yield differentiable loss functions.

Metric	Definition	Fairness Loss
Statistical Parity Difference (SPD)	$\lvert \text{PPR}^{p} - \text{PPR}^{up} \rvert$	$\lvert L^{p}_{u\lvert 1} - L^{up}_{u\lvert 1} \rvert$
Equality of Opportunity Difference (EOD)	$\lvert \text{TPR}^{p} - \text{TPR}^{up} \rvert$	$\lvert L^{p,fav}_{u} - L^{up,fav}_{u} \rvert$
Average Odds Difference (AOD)	$0.5 \cdot (\lvert \text{FPR}^{p} - \text{FPR}^{up} \rvert + \lvert \text{TPR}^{p} - \text{TPR}^{up} \rvert)$	$\lvert L^{p,unfav}_{u} - L^{up,unfav}_{u} \rvert + \lvert L^{p,fav}_{u} - L^{up,unfav}_{u} \rvert$

PPR, FPR, and TPR denote the positive predictive rate, false positive rate, and true positive rate, respectively. L_u denotes the utility loss function we use to train the model, such as cross-entropy loss for logistic regression and neural networks. The superscripts *p*, and *up* denote the privileged, and unprivileged groups. Similarly, the superscripts *fav* and *unfav* denote the favorable, and the unfavorable label. For example, $L^{p,fav}_{u}$ denotes the average utility loss computed over privileged and favorable samples. Note that, in contrast to EOD and AOD, SPD does not take ground-truth labels of inputs into account. Therefore, we have to compute the utility loss by setting the target as 1 for SPD, denoted as $L_{u|1}$, regardless of the actual label of the data instance.

Here, $f : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is referred as the outer problem, and $g : \mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ is referred as the inner problem. It is important to note the dependence of the outer problem to the inner problem. Due to this, one cannot solve inner and outer problems simultaneously, but rather, inner problem has to be solved first before one can treat the outer problem.

Recent years have seen the successful application of bilevel optimization to various areas in ML, such as meta-learning (Finn et al. (2017); Nichol et al. (2018); Rajeswaran et al. (2019)), or scalable high-dimensional hyperparameter optimization (Lorraine et al. (2020); Franceschi et al. (2018)). From the recent works that use bilevel optimization, the most similar works to ours are due to Ren et al. (2018); Jenni & Favaro (2018); Roh et al. (2020b). In all of these work, bilevel optimization is used to learn a set of weights on the training dataset as in our work. In Ren et al. (2018), the learned weights ensure good training under datasets with severe label imbalance, and/or noisy labels. In Jenni & Favaro (2018), weights are learned in a way to ensure good generalization of the trained model. Finally, Roh et al. (2020b) uses bilevel optimization to develop a fair training algorithm as in our work, but their algorithm requires that all the training data to have demographic information, where we consider a setting with limited demographic data in this work as mentioned before.

3 Fair Training with Limited Demographic Data

In this section, we tackle the problem of training fair models when we know the demographic attributes for only a subset of the data. To this end, we first provide a strawman solution that adapts existing fair training algorithms that require demographic information on all the dataset. It turns out that, even this approach works fairly well against Rawlsian approaches as far as parity-based metrics are concerned as we experimentally show later. Then, we develop a fair training algorithm that is particularly suited to limited demographics setting by leveraging the bilevel optimization framework. Our experiments in Section 4 shows this approach degrades more gracefully than the strawman approach as the size of the demographically labeled data gets smaller.

3.1 Problem Setting and Notation

We consider the problem of binary supervised classification for a dataset $\{X, y, s\}_{i=1}^{n}$ with a model M_θ parameterized by $\theta \in R^d$. Here, $X^{(i)} \in \mathbb{R}^m$ is the set of features, $y^{(i)} \in \{0, 1\}$ is the label, and $s^{(i)} \in \{0, 1\}$ is the demographic/sensitive feature (e.g., race) of the ith sample. A sample is referred as *favorable* if its label is 1, and *unfavorable* otherwise. Similarly, the samples with a sensitive attribute of 1 are referred as *privileged*, and *unprivileged* otherwise. By definition, privileged samples appear more frequently with the favorable label in the overall population, e.g., it could be that 2/3 of favorable labels belong to the privileged group, where as this figure is 1/3 for unprivileged samples. Crucially, we assume that we know s_i values only a for a subset of our dataset. Our goal is to learn a model can predict the labels of new samples with high accuracy, while exhibiting low bias value with respect to a parity-based bias metric(s), such as those defined in Table 1.

3.2 A STRAWMAN SOLUTION

Let D_{tr} denote our training dataset, and let D_{dem} be the subset of D_{tr} whose demographic attributes are known. Then, we can adapt any existing parity-based fair training algorithm to limited demographics setting. First, (i) train a predictor model for demographic features using D_{dem}, then, (ii) fill the demographic features for the remaining of training data by having the model predict demographic attributes from other features, and finally, (iii) run the fair training algorithm on D_{tr} as usual.

Although this data imputation method is pretty straightforward, there are few points to discuss. First, we can see that the performance of this approach is essentially determined by how accurately we can predict demographic features from the other features. So when using this approach, we are implicitly assuming that there exists a correlation between the sensitive feature (s_i), and other features (X_i). In our experiments, we have observed this assumption generally holds. Regardless, this correlation can change from dataset to dataset, and might affect the stability of the strawman approach. Therefore, it is worth to note that the next algorithm we develop makes no such assumption.

3.3 BIFAIR

We now describe a fair training algorithm that is particularly suited to limited demographics setting, named ***BiFair***, by leveraging the bilevel optimization framework. Briefly, we introduce a set of weights w on the training dataset, such that training on the weighted dataset yields both good utility, and low bias for the model. We learn the values of w concurrent to the model training by solving a bilevel optimization problem. Concretely, let L_u be a loss function that we use to train the model. For example, L_u could be the hinge loss if our model is a SVM, or it could be cross-entropy if it is a neural network. Further, let L_f be differentiable fairness loss that is associated with a bias metric, such that, by minimizing L_f, we can reduce the bias of the model. Then, we can formulate our learning objective as follows,

$$\begin{aligned} w^*, \theta^* \in \arg\min_{w,\theta} & L_f(M_{\theta^\star}, D_{dem}), \\ \text{subject to } \theta^\star \in \arg\min_{\theta} & \sum_i w^{(i)} \cdot L_u(M_\theta, D_{tr}^{(i)}). \end{aligned} \tag{1}$$

As is seen, in the inner optimization, we minimize L_u on the weighted training dataset. In the outer optimization, the weights are adjusted to minimize the fairness loss on the portion of the dataset where we have access to the demographic attributes (hence, we can compute any of the fairness losses given in Table 1 on that portion).

Many models of practical interest yields no closed-form solution to the inner problem, but rather, are optimized by iterative methods, e.g., by gradient descent. So, finding an optimal solution to the inner problem formulated in Equation 1 is usually a costly process. One workaround of this, is to relax the inner problem by finding an approximate solution to it as presented in Domke (2012). Briefly, we can approximate the solution to the inner problem by taking a few steps of gradient descent, and then compute the gradient for the outer problem at the approximated solution. We can then update the training data weights using the gradient of the outer problem, and repeat this until the outer level problem converges. This gives us an algorithm that is straightforward to implement with ML frameworks that provide automatic differentiation such as PyTorch (Paszke et al. (2019)) and TensorFlow (Abadi et al. (2015)). The pseudo-code of our algorithm is presented in Algorithm 1. Briefly, the lines 3-9 correspond to approximating the inner problem, and in line 14 and 15, we compute the gradient for the outer problem, and update the dataset weights, respectively.

3.4 EXTENDING BIFAIR FOR ROBUSTNESS

As mentioned before in Section 1, the bilevel optimization is flexible in the sense that, we can target multiple objectives at the same time by minimal change in the formulation. To highlight this, we consider fairness and robustness together. In what follows, we assume D_{tr} might contain noisy labels, and D_{dem} has clean labels. With such an assumption, we can add robustness to BiFair by simply adding the utility loss to the outer-level problem. That is, the new formulation becomes,

$$w^*, \theta^* \in \arg\min_{w,\theta} L_f(M_{\theta^\star}, D_{dem}) + \lambda \cdot L_u(M_{\theta^\star}, D_{dem}),$$
$$\text{subject to } \theta^\star \in \arg\min_{\theta} \sum_i w^{(i)} \cdot L_u(M_\theta, D_{tr}^{(i)}). \tag{2}$$

where $\lambda \geq 0$ is a scalar hyperparameter introduced to control the trade-off between utility and fairness. As can be seen, in the new formulation, the dataset weights are updated by taking the utility loss computed over the clean-labeled D_{dem} into account, and this gives us robustness. To accommodate for this change in Algorithm 1, we only need to add $L_{u_{dem}}$ to line 13, and consequently compute the gradient over $L_f + \lambda \cdot L_{u_{dem}}$ in line 14.

Algorithm 1: BiFair with automatic differentiation for supervised classification. For each outer iteration, we find an approximate solution to the inner problem (lines 3-9). Then, we compute the fairness loss L_f on the demographically labeled portion of the data (lines 10-13). Finally, we update the weights of the training datasets (lines 14-15). Note that, we use demographic features only on line 13. Also, computing $\nabla_w L_f$ requires us to maintain the computation graph of the inner-loop. This computation graph is only freed at line 15, after we compute $\nabla_w L_f$.

Input : Training dataset D_{tr} with demographically labeled portion denoted as $D_{dem} \subseteq D_{tr}$, and corresponding batch sizes B_{tr}, and B_{dem}, number of outer iterations T_{out}, and number of inner iterations *per* outer iteration T_{in}

Output : Trained model M_θ parameterized by θ

1 Initialize θ and w randomly
2 **for** $t_{out} \leftarrow 1$ **to** T_{out} **do**
3 **for** $t_{in} \leftarrow 1$ **to** T_{in} **do**
4 $X_{tr}, y_{tr}, \leftarrow$ SampleMiniBatch(D_{tr}, B_{tr})
5 $\hat{y}_{tr} \leftarrow$ ForwardPass(M_θ, X_{tr})
6 $L_u \leftarrow \frac{1}{B_{tr}} \sum_i w^{(i)} \cdot L_u(\hat{y}_{tr}^{(i)}, y_{tr}^{(i)})$
7 $\nabla_\theta L_u \leftarrow$ BackProp(L_u, θ)
8 $\theta \leftarrow$ OptimizerUpdate$(\theta, \nabla_\theta L_u)$ `// SGD, Adam etc.`
9 **end**
10 $X_{dem}, y_{dem}, s_{dem} \leftarrow$ SampleMiniBatch(D_{dem}, B_{dem}) `// Demographic features are retrieved here`
11 $\hat{y}_{dem} \leftarrow$ ForwardPass(M_θ, X_{dem})
12 $L_{u_{dem}} \leftarrow \frac{1}{B_{dem}} \sum_i L_u(\hat{y}_{dem}^{(i)}, y_{dem}^{(i)})$
13 $L_f \leftarrow$ ComputeFairnessLoss$(L_{u_{dem}}, y_{dem}, s_{dem})$ `// See Table 1 for what this loss can be`
14 $\nabla_w L_f \leftarrow$ BackProp(L_f, w)
15 $w \leftarrow$ OptimizerUpdate$(w, \nabla_w L_f)$
16 **end**

4 Experiments

In this section, we evaluate the performance of the strawman solution, and BiFair via experiments and compare them against several baselines as well. Our implementation is in PyTorch (Paszke et al. (2019)) with Higher library (Grefenstette et al. (2019)), and our code and scripts to replicate our results are publicly available at `https://github.com/TinfoilHat0/BiFair`.

4.1 Experimental Setting

Evaluation Metrics: In our experiments, we record four metrics of interest. Three of them are bias metrics, AOD, EOD, and SPD, that are listed and defined in Table 1. In addition to these, we use the balanced accuracy as in Bellamy et al. (2018) to quantify the utility of the models. We choose balanced accuracy over accuracy because most real-world datasets used in fairness research have label-imbalance, including the ones we use, and this makes accuracy a poor metric of choice. Balanced accuracy (BAcc) is equal to the accuracy when there is no label-imbalance, and is defined as follows,

$$\text{BAcc} = \frac{\text{TNR} + \text{TPR}}{2},$$

where TNR and TPR stand for true negative rate, and true positive rate, respectively. For bias metrics, we note that lower values indicate better performance, i.e., the model is fairer when bias values are lower.

Baselines: We consider four baselines in our evaluations: *unconstrained training*, two parity-based fair training algorithms sampled from the AIF360 toolkit Bellamy et al. (2018), *Kamiran reweighing* of Kamiran & Calders (2012) and *Prejudice Remover* of Kamishima et al. (2012), and one Rawlsian fair training algorithm, *Adversarially Reweighted Learning* (ARL) of Lahoti et al. (2020). To the best of our knowledge, ARL is the state-of-the-art Rawlsian method. When choosing other fair training algorithms, we have taken benchmarks presented in AIF 360 toolkit into account, and have chosen the best performing algorithms in its respective category. When the results between two algorithms were too close to call for a clear winner, we have chosen the algorithm that we deem as simpler to implement. We provide a brief description of each baseline below.

Unconstrained Training simply refers to the training of model without any fairness constraint, i.e., we are doing empirical loss minimization as usual.
Kamiran Reweighing of Kamiran & Calders (2012) is a pre-processing technique that aims to ensure statistical independence between the label and the sensitive attribute by assigning weights to data points. Particularly, the technique first computes the expected probability, pr_{exp}, for each combination of sensitive attribute and label under the assumption that the sensitive attribute and the label are independent. Then, they measure the observed probability, pr_{obs}, for each combination of sensitive attribute and label in the training dataset. Finally, each data point is assigned the weight pr_{exp}/pr_{obs} based on the value of their sensitive attribute and their label.
Prejudice Remover of Kamishima et al. (2012) is an in-processing technique that tries to ensure statistical independence between the model's prediction, and the sensitive attribute. To do so, the empirical mutual information between the model's prediction and the sensitive attribute is added as a regularization term in the training objective.
Adversarially Reweighted Learning (ARL) of Lahoti et al. (2020) is an in-processing technique that models the fair training as a min-max game between a *learner* model, and an *adversary* model. The goal of the learner model is to minimize the empirical loss over a weighted dataset where the adversary tries to assign higher weights to samples in which learner model performs worse as indicated by its loss value. Crucially, since the reweighing is done only according to individual loss values of samples, this algorithm does not require any sensitive attribute to be known.

Datasets: As for the datasets, we use the Adult and Bank datasets from the UCI repository (Dua & Graff (2017)). In the Adult dataset, the goal is to predict whether a person's income is greater than of \$50k USD a year, the sensitive feature is gender, and men are the privileged. For the Bank dataset, the goal is to predict whether a client will make a deposit subscription or not, the sensitive attribute is age, and old people (older than 25 years) are privileged. We provide some statistics for the datasets we use, and briefly discuss their implications for fairness in Figure 1.

Setup: For each dataset, we ensure a train/validation/test split of $60\%/20\%/20\%$ where we assume model can access the sensitive attributes for only a subset of the training dataset. The sensitive attributes are treated as meta-features, and are not fed to the input layer of models to ensure uniformity across baselines. In all of our experiments, we train a logistic regression model using the Adam optimizer of Kingma & Ba (2014), and in the case of BiFair, we also use Adam to update the training dataset weights. Each model is trained until the validation loss stagnates for 5 epochs, i.e., we use early-stopping to decide when to stop training. We report the final measurements done on the test dataset where each measurement is averaged over 10 runs. Measurements are plotted as barcharts for the sake of presentation, and exact values are provided in Appendix B.

4.2 Experimental Results

Strawman vs. ARL: We first illustrate the performance of our strawman approach against unconstrained training and ARL. To do so, we adapt Kamiran Reweighing, and Prejudice Remover to limited demographics setting by imputing unknown demographic attributes as described in Section 3.2. The results are presented and discussed in detail in Figure 2. In general, we observe that the strawman solution exhibits considerably less bias than ARL even when the size of the demographically labeled data is as small as 1%-0.1% of the training data.

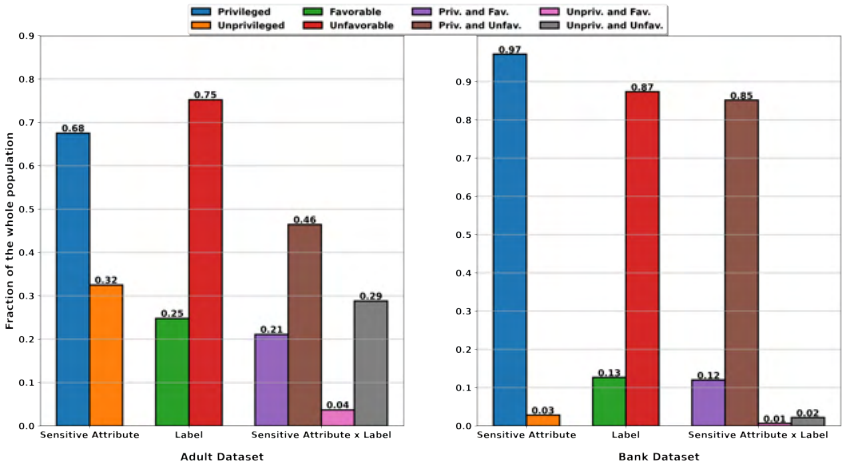


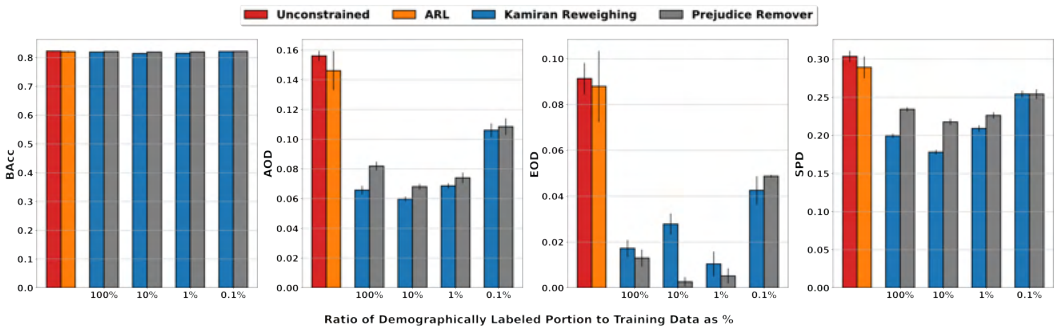
Figure 1: Several statistics regarding fairness for the datasets that we use. First, note that both datasets suffer from severe label imbalance, which justifies our choice of balanced accuracy over accuracy to measure the utility. Further, samples with privileged sensitive attribute value (men for Adult, and old people for Bank) appear more frequently with the favorable label overall. This is not surprising as privileged samples also appear much more frequently than unprivileged samples. However, it is interesting to note that, conditioned on the sensitive attribute, unprivileged samples are more likely to be favorable for the Bank dataset. Concretely, we have $Pr[\text{Label=Favorable} \mid \text{Sensitive Attribute=Privileged}] = 0.12$ and $Pr[\text{Label=Favorable} \mid \text{Sensitive Attribute=Unprivileged}] = 0.23$ in the Bank dataset. We suspect that, due to this, the bias exhibited by the models over the Bank dataset was generally much lower. Consequently, the effect of fair training approaches was more visible in the Adult dataset.

BiFair vs. Strawman: Now that we have seen strawman approach can outperform ARL with little demographically labeled data, we compare it with BiFair. For BiFair, we use the AOD loss (see Table 1) as its fairness loss. This is because AOD encompasses EOD, and unlike SPD, it takes ground-truth labels of data points into account. So, it does not necessary lead to a drop in the utility of the model. We plot and discuss the results in Figure 3. In brief, the plots suggest that the performance of BiFair degrades more gracefully as the size of the demographically labeled portion gets smaller. Consequently, BiFair tends to outperform the strawman approach especially when the size of the demographically labeled portion is small. To highlight this better, we conducted additional experiments for cases where the demographically labeled portion is less than 1% of the training data, and provide them in Appendix (Figure 5).

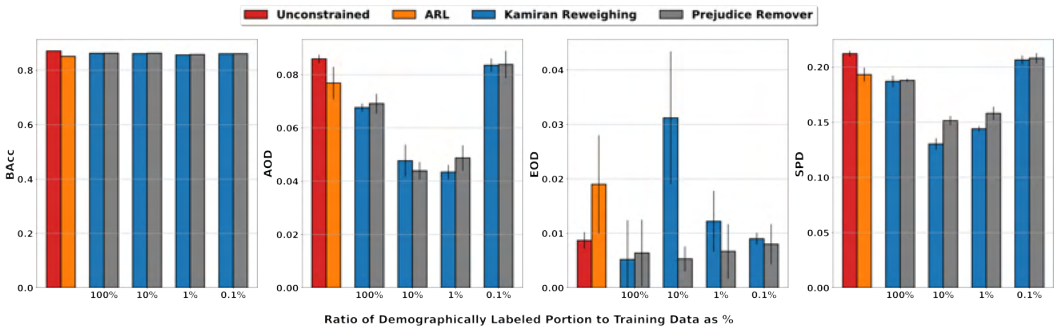
Fairness under Noisy Labels: As we discussed in Section 3.4, we can trivially extend BiFair, and make it robust to label noise assuming our demographically labeled portion has clean labels (see Equation 2). To highlight this property of BiFair, we use the following setting: we first ensure 0.1% of the training has known demographics, and clean class-labels. For the rest of the data, we flip the class labels with 1/2 probability. The results are presented and discussed in Figure 4. As can be observed, every other algorithm than BiFair fails to provide any utility, i.e., their accuracies are about 50%. Meanwhile, BiFair achieves a much better accuracy, while still reducing the bias considerably.

5 Discussion

We now discuss some aspects of our results that we find interesting, and suggest some avenues for future work. First, when evaluating our strawman approach, we have observed that the bias values of models can sometimes get lower as the size of demographically labeled portion shrinks. Given our prediction accuracy for demographic attributes strictly decreases, as the size of demographically labeled portion decreases, this suggests to us that, some amount of noise on the demographic attributes might act as a regularization effect. Consequently, a carefully-controlled noise amount on demographic attributes might perhaps be used to train fairer models. Therefore, we believe looking at the relationship between noisy demographic attributes, and bias can be an interesting avenue of research.

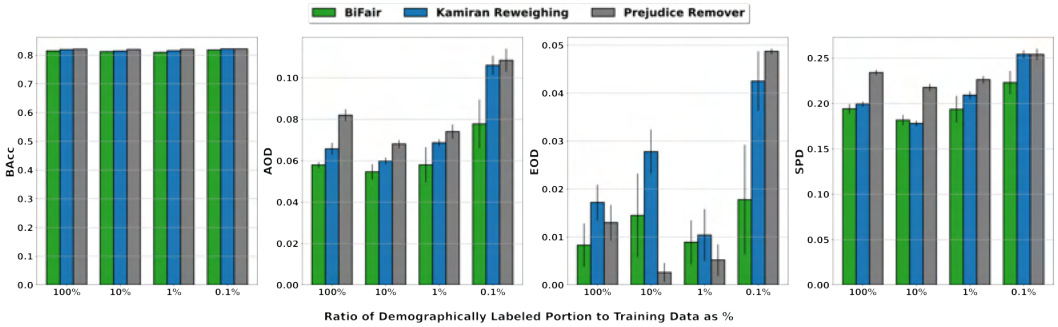


(a) Adult dataset

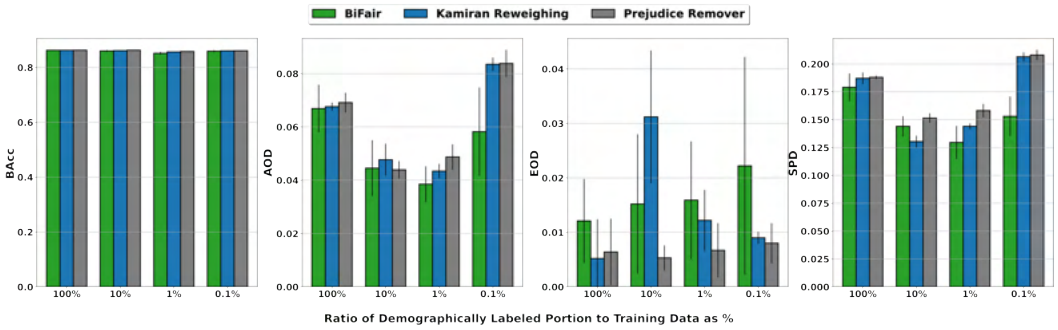


(b) Bank dataset

Figure 2: The performance of strawman approach against unconstrained training and ARL. First of all, we observe that all methods perform similarly in terms of accuracy.s for the bias metrics, we see all the methods tend to exhibit lower bias compared to the unconstrained training as expected. However, overall results indicate that ARL is the least effective fair training approach of all. For example, for Adult dataset, we see that the strawman approach outperforms ARL by a visible margin for all bias metrics even with only 0.1% demographically labeled data ($\approx$ 20 samples). As for Bank dataset, ARL tends to perform slightly better than the strawman approach for AOD and SPD metrics for 0.1% case, and worse for every other case. Finally, it is worth looking at how the strawman solution scales with the size of demographically labeled data. Across all metrics, 0.1% case has higher bias than 100% case. With too little demographically labeled data, we cannot predict demographic attributes too accurately as expected. However, we observe that, as we move from 100% to 0.1%, bias values gets lower in some cases. We think that this happens because some amount of noise in the demographic attributes acts as a regularizer, and results in better generalization performance for bias metrics. This is especially visible in Bank dataset if we compare 100% case and 10%-1% case for AOD.

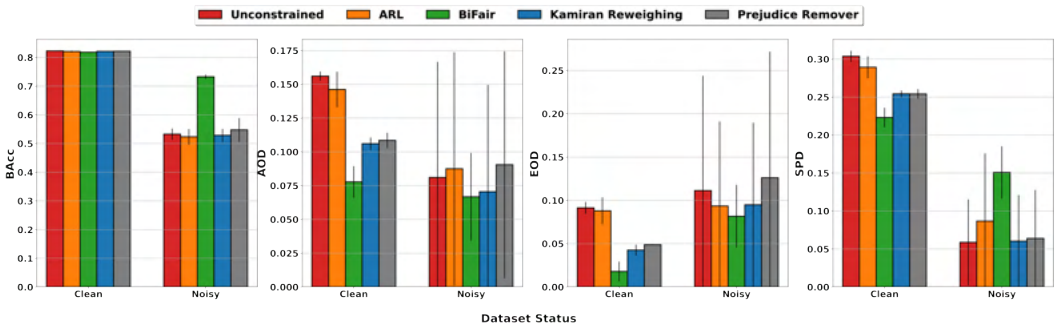


(a) Adult dataset

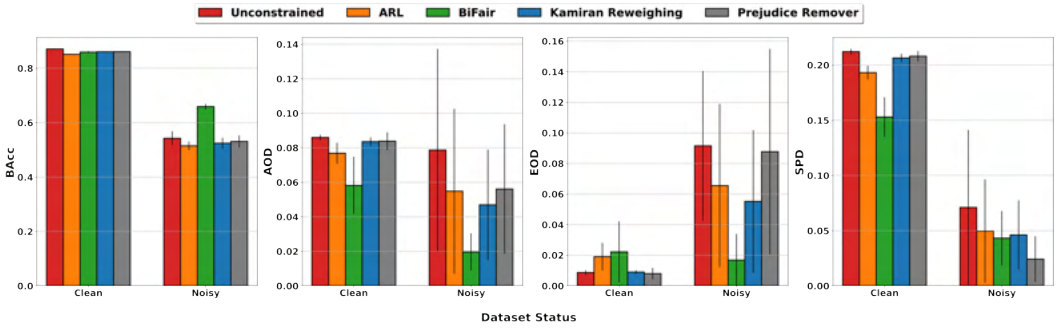


(b) Bank dataset

Figure 3: The performance of BiFair against the strawman approach. Again, we see all the approaches tend to perform similar with respect to the accuracy. However, we see that, BiFair tends to degrade more gracefully compared to the strawman approach, and tends to exhibit better fairness as the demographically labeled data gets smaller. This is especially visible for the 0.1% case. For Adult dataset, it outperforms the strawman approach across all the bias metrics, and for Bank dataset, it is only worse for EOD. We suspect this happens because the Bank datasets exhibits high AOD and low EOD by default (see results of Unconstrained in Figure 2 for Bank). So, perhaps there is a tug-of-war between AOD and EOD metrics for this dataset, and BiFair increases EOD slightly as it brings down AOD.



(a) Adult dataset



(b) Bank dataset

Figure 4: Comparison of algorithms under clean, and noisy label settings. As is seen, every other algorithm other than BiFair is susceptible to label noise. In the noisy setting, the accuracy of every algorithm other than BiFair is merely above 50%. Briefly put, they fail to provide any utility. On the other hand, BiFair maintains much better accuracy with low bias values.

As for BiFair, it is worth to briefly discuss the performance overhead, First, note that, we have to maintain the computation graph of the inner optimization (lines 3-9 in Algorithm 1) until gradient of the outer optimization is computed (line 16 in Algorithm 1). Consequently, the memory usage of our algorithm scales linearly with T_{in}. Although we have observed that we get good performance even with small values of T_{in} in our experimental evaluation[4], the high memory requirements could pose a problem for certain datasets, and large models (e.g., ResNet-101 of He et al. (2016)). One straightforward way to reduce the memory usage is due to the truncated backpropagation method presented in Shaban et al. (2019). With truncated backpropagation, we maintain the computation graph only for the last few iterations regardless of the value of T_{in}. Consequently, this makes memory requirements independent of T_{in}. For example, we can take 100 inner iterations, yet maintain the computation graph only for the last 5 steps. This is likely to give a better performance than only taking 5 inner iterations (see experimental analysis of Shaban et al. (2019)). With respect to computation cost, we see that our algorithm does a forward-backward pass for each inner iteration (line 5 and 7), and then another forward-backward pass at the outer level (line 11 and 15), per iteration. Consequently, this induces a computation overhead factor of T_{in} + 1 over regular training, which does a single forward-backward pass per iteration. Approaches such as implicit gradients of Rajeswaran et al. (2019) might be used to reduce the extra computation cost.

Finally, it is worth to note that, although we focused on supervised classification in this work, one can trivially adapt our main formulation given in Equation 1 to other tasks, such as regression. So, it might be interesting to quantify the performance of our approaches for other tasks under the limited demographics setting.

6 Conclusion

We briefly recap our main results before concluding the paper. First, rather surprisingly, we have demonstrated that even a straightforward strawman solution can adapt existing fair training algorithms to limited demographics setting with rather good performance. This implies that industry practitioners who have limited access to demographic attributes can adapt their existing pipelines and algorithms to this setting rather easily. Second, we have developed a novel algorithm, named BiFair, that is particularly suited to limited demographics setting. Through experiments, we showed that BiFair scales more gracefully as the size of the demographically labeled portion gets smaller, and overall, it tends to exhibit lower bias than the strawman solution. Further, we have expanded BiFair to make it robust to noisy labels, and illustrated that it can provide both good fairness, and good utility under heavy label noise in the limited demographics setting. In general, we emphasize our main formulation presented in Equation 1 is quite flexible, and it accommodate for multiple training objectives with little modification as we have shown for fairness and robustness. In summary, we have developed and evaluated approaches to train fair models in a setting where we know demographic/sensitive attributes for only a subset of the data at hand. To the best of our knowledge, this is the first work to consider such setting in the context of fairness research. Overall, we believe this is a well-motivated setting, and corresponds to what most industry practitioners face in real world. So, we hope that our work will motivate researchers to consider and analyze this particular setting further.

[4]Concretely, all of our results are obtained with $T_{in} = 2$. We did not observe a noticeable improvement in performance for higher values of T_{in}.

REFERENCES

Understanding the 80% rule. URL https://www.jfmeltonlaw.com/articles/understanding-the-80-rule/.

General data protection regulation. URL https://en.wikipedia.org/wiki/General_Data_Protection_Regulation.

Martin Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Ian Goodfellow, Andrew Harp, Geoffrey Irving, Michael Isard, Yangqing Jia, Rafal Jozefowicz, Lukasz Kaiser, Manjunath Kudlur, Josh Levenberg, Dandelion Mané, Rajat Monga, Sherry Moore, Derek Murray, Chris Olah, Mike Schuster, Jonathon Shlens, Benoit Steiner, Ilya Sutskever, Kunal Talwar, Paul Tucker, Vincent Vanhoucke, Vijay Vasudevan, Fernanda Viégas, Oriol Vinyals, Pete Warden, Martin Wattenberg, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL https://www.tensorflow.org/. Software available from tensorflow.org.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 60–69. PMLR, 10–15 Jul 2018. URL http://proceedings.mlr.press/v80/agarwal18a.html.

Yahav Bechavod and Katrina Ligett. Learning fair classifiers: A regularization-inspired approach. *arXiv preprint arXiv:1707.00044*, pp. 1733–1782, 2017.

Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John Richards, Diptikalyan Saha, Prasanna Sattigeri, Moninder Singh, Kush R. Varshney, and Yunfeng Zhang. AI Fairness 360: An extensible toolkit for detecting, understanding, and mitigating unwanted algorithmic bias, October 2018. URL https://arxiv.org/abs/1810.01943.

Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In Sorelle A. Friedler and Christo Wilson (eds.), *Proceedings of the 1st Conference on Fairness, Accountability and Transparency*, volume 81 of *Proceedings of Machine Learning Research*, pp. 77–91, New York, NY, USA, 23–24 Feb 2018. PMLR. URL http://proceedings.mlr.press/v81/buolamwini18a.html.

Flavio P Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 3995–4004, 2017.

L Elisa Celis, Lingxiao Huang, Vijay Keswani, and Nisheeth K Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 319–328, 2019.

Jeffrey Dastin. Amazon scraps secret ai recruiting tool that showed bias against women. Online; accessed May 21, 2021.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Justin Domke. Generic methods for optimization-based modeling. In Neil D. Lawrence and Mark Girolami (eds.), *Proceedings of the Fifteenth International Conference on Artificial Intelligence and Statistics*, volume 22 of *Proceedings of Machine Learning Research*, pp. 318–326, La Palma, Canary Islands, 21–23 Apr 2012. PMLR. URL http://proceedings.mlr.press/v22/domke12.html.

Michele Donini, Luca Oneto, Shai Ben-David, John Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. *arXiv preprint arXiv:1802.08626*, 2018.

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

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*, KDD '15, pp. 259–268, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450336642. doi: 10.1145/2783258.2783311. URL https://doi.org/10.1145/2783258.2783311.

Chelsea Finn, Pieter Abbeel, and Sergey Levine. Model-agnostic meta-learning for fast adaptation of deep networks. In *International Conference on Machine Learning*, pp. 1126–1135. PMLR, 2017.

Luca Franceschi, Paolo Frasconi, Saverio Salzo, Riccardo Grazzi, and Massimiliano Pontil. Bilevel programming for hyperparameter optimization and meta-learning. In *International Conference on Machine Learning*, pp. 1568–1577. PMLR, 2018.

Edward Grefenstette, Brandon Amos, Denis Yarats, Phu Mon Htut, Artem Molchanov, Franziska Meier, Douwe Kiela, Kyunghyun Cho, and Soumith Chintala. Generalized inner loop meta-learning. *arXiv preprint arXiv:1910.01727*, 2019.

Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016a. URL https://proceedings.neurips.cc/paper/2016/file/9d2682367c3935defcb1f9e247a97c0d-Paper.pdf.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, pp. 3323–3331, Red Hook, NY, USA, 2016b. Curran Associates Inc. ISBN 9781510838819.

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

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

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*, pp. 1–16, 2019.

Lauren Kirchner Jeff Larson, Surya Mattu and Julia Angwin. How we analyzed the compas recidivism algorithm, 2016. URL https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. Online; accessed May 14, 2021.

Simon Jenni and Paolo Favaro. Deep bilevel learning. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 618–633, 2018.

Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

Faisal Kamiran, Asim Karim, and Xiangliang Zhang. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*, pp. 924–929. IEEE, 2012.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 35–50. Springer, 2012.

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

Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research). 2009. URL http://www.cs.toronto.edu/~kriz/cifar.html.

Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed Chi. Fairness without demographics through adversarially reweighted learning. In *34th Conference on Neural Information Processing Systems*. Curran Associates, Inc., 2020.

Jonathan Lorraine, Paul Vicol, and David Duvenaud. Optimizing millions of hyperparameters by implicit differentiation. In *International Conference on Artificial Intelligence and Statistics*, pp. 1540–1552. PMLR, 2020.

Alex Nichol, Joshua Achiam, and John Schulman. On first-order meta-learning algorithms. *arXiv preprint arXiv:1803.02999*, 2018.

Curtis G. Northcutt, Anish Athalye, and Jonas Mueller. Pervasive label errors in test sets destabilize machine learning benchmarks, 2021.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems 32*, pp. 8024–8035. Curran Associates, Inc., 2019. URL http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *arXiv preprint arXiv:1709.02012*, 2017.

Aravind Rajeswaran, Chelsea Finn, Sham Kakade, and Sergey Levine. Meta-learning with implicit gradients. *arXiv preprint arXiv:1909.04630*, 2019.

John Rawls. *A theory of justice: Revised edition*. Harvard university press, 1999.

Mengye Ren, Wenyuan Zeng, Bin Yang, and Raquel Urtasun. Learning to reweight examples for robust deep learning. In *International Conference on Machine Learning*, pp. 4334–4343. PMLR, 2018.

Yuji Roh, Kangwook Lee, Steven Whang, and Changho Suh. Fr-train: A mutual information-based approach to fair and robust training. In *International Conference on Machine Learning*, pp. 8147–8157. PMLR, 2020a.

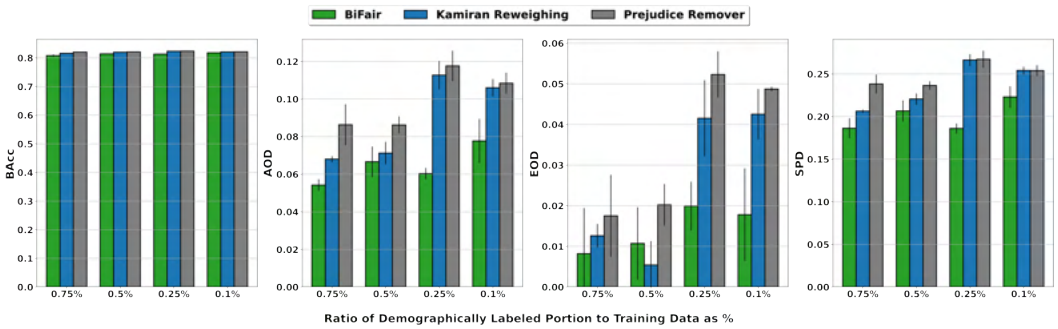
Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. Fairbatch: Batch selection for model fairness. *arXiv preprint arXiv:2012.01696*, 2020b.

Amirreza Shaban, Ching-An Cheng, Nathan Hatch, and Byron Boots. Truncated back-propagation for bilevel optimization. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 1723–1732. PMLR, 2019.

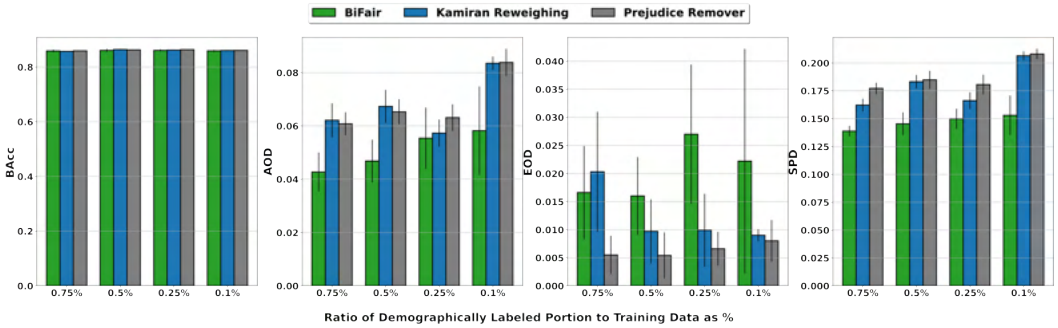
Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pp. 325–333. PMLR, 2013.

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

A ADDITIONAL EXPERIMENTS FOR BIFAIR



(a) Adult dataset



(b) Bank dataset

Figure 5: The performance of BiFair against the strawman approach when the demographically labeled portion is less than 1% of the training data. As is seen, BiFair is better across all metrics for 0.25% and 0.1% case for the Adult dataset. For the Bank dataset, it is better for AOD and SPD, and worse for EOD (which we have speculated why in Figure 3).

B BARCHART VALUES

Table 2: Values that are used to generate Figure 2, Figure 3, clean dataset part of Figure 4, and Figure 5.

Algorithm	Demographic Portion (%)	BAcc	AOD	EOD	SPD
Unconstrained	–	0.822 ± 0.0004	0.156 ± 0.0034	0.091 ± 0.0068	0.303 ± 0.0073
ARL	–	0.820 ± 0.0029	0.146 ± 0.0132	0.088 ± 0.0155	0.289 ± 0.0144
Kamiran Reweighing	100%	0.819 ± 0.0006	0.066 ± 0.0029	0.017 ± 0.0037	0.199 ± 0.0028
Kamiran Reweighing	10%	0.814 ± 0.0007	0.060 ± 0.0016	0.028 ± 0.0046	0.178 ± 0.0030
Kamiran Reweighing	1%	0.815 ± 0.0009	0.069 ± 0.0017	0.010 ± 0.0054	0.209 ± 0.0043
Kamiran Reweighing	0.75%	0.816 ± 0.0007	0.068 ± 0.0016	0.013 ± 0.0029	0.206 ± 0.0020
Kamiran Reweighing	0.5%	0.820 ± 0.0003	0.071 ± 0.0060	0.005 ± 0.0059	0.221 ± 0.0065
Kamiran Reweighing	0.25%	0.823 ± 0.0010	0.113 ± 0.0075	0.042 ± 0.0094	0.266 ± 0.0069
Kamiran Reweighing	0.1%	0.821 ± 0.0006	0.106 ± 0.0046	0.043 ± 0.0062	0.254 ± 0.0044
Prejudice Remover	100%	0.820 ± 0.0006	0.082 ± 0.0029	0.013 ± 0.0037	0.234 ± 0.0028
Prejudice Remover	10%	0.819 ± 0.0010	0.068 ± 0.0020	0.003 ± 0.0020	0.218 ± 0.0041
Prejudice Remover	1%	0.820 ± 0.0005	0.074 ± 0.0035	0.005 ± 0.0033	0.226 ± 0.0042
Prejudice Remover	0.75%	0.820 ± 0.0007	0.086 ± 0.0109	0.018 ± 0.0101	0.238 ± 0.0111
Prejudice Remover	0.5%	0.821 ± 0.0008	0.086 ± 0.0045	0.020 ± 0.0051	0.236 ± 0.0049
Prejudice Remover	0.25%	0.823 ± 0.0017	0.118 ± 0.0081	0.052 ± 0.0057	0.267 ± 0.0100
Prejudice Remover	0.1%	0.821 ± 0.0008	0.108 ± 0.0056	0.049 ± 0.0006	0.254 ± 0.0063
BiFair	100%	0.814 ± 0.0015	0.058 ± 0.0014	0.008 ± 0.0045	0.194 ± 0.0051
BiFair	10%	0.812 ± 0.0020	0.055 ± 0.0037	0.014 ± 0.0087	0.182 ± 0.0057
BiFair	1%	0.809 ± 0.0021	0.058 ± 0.0085	0.009 ± 0.0046	0.194 ± 0.0148
BiFair	0.75%	0.808 ± 0.0040	0.054 ± 0.0031	0.008 ± 0.0112	0.186 ± 0.0118
BiFair	0.5%	0.814 ± 0.0014	0.067 ± 0.0081	0.011 ± 0.0089	0.207 ± 0.0123
BiFair	0.25%	0.813 ± 0.0014	0.060 ± 0.0030	0.020 ± 0.0060	0.186 ± 0.0059
BiFair	0.1%	0.818 ± 0.0016	0.078 ± 0.0117	0.018 ± 0.0114	0.223 ± 0.0128

(a) Adult dataset

Algorithm	Demographic Portion (%)	BAcc	AOD	EOD	SPD
Unconstrained	–	0.871 ± 0.0005	0.086 ± 0.0016	0.009 ± 0.0015	0.212 ± 0.0026
ARL	–	0.851 ± 0.0018	0.077 ± 0.0061	0.019 ± 0.0090	0.193 ± 0.0062
Kamiran Reweighing	100%	0.862 ± 0.0010	0.068 ± 0.0015	0.005 ± 0.0072	0.187 ± 0.0052
Kamiran Reweighing	10%	0.861 ± 0.0010	0.048 ± 0.0060	0.031 ± 0.0122	0.130 ± 0.0053
Kamiran Reweighing	1%	0.856 ± 0.0015	0.043 ± 0.0027	0.012 ± 0.0056	0.144 ± 0.0026
Kamiran Reweighing	0.75%	0.857 ± 0.0012	0.062 ± 0.0064	0.020 ± 0.0107	0.162 ± 0.0057
Kamiran Reweighing	0.5%	0.864 ± 0.0019	0.067 ± 0.0062	0.010 ± 0.0057	0.183 ± 0.0061
Kamiran Reweighing	0.25%	0.862 ± 0.0013	0.057 ± 0.0051	0.010 ± 0.0065	0.166 ± 0.0075
Kamiran Reweighing	0.1%	0.860 ± 0.0019	0.083 ± 0.0025	0.009 ± 0.0011	0.206 ± 0.0039
Prejudice Remover	100%	0.863 ± 0.0010	0.069 ± 0.0037	0.006 ± 0.0061	0.188 ± 0.0016
Prejudice Remover	10%	0.863 ± 0.0011	0.044 ± 0.0033	0.005 ± 0.0023	0.151 ± 0.0043
Prejudice Remover	1%	0.858 ± 0.0010	0.049 ± 0.0048	0.007 ± 0.0050	0.158 ± 0.0060
Prejudice Remover	0.75%	0.859 ± 0.0009	0.061 ± 0.0043	0.006 ± 0.0034	0.177 ± 0.0051
Prejudice Remover	0.5%	0.863 ± 0.0009	0.065 ± 0.0047	0.005 ± 0.0041	0.185 ± 0.0081
Prejudice Remover	0.25%	0.863 ± 0.0012	0.063 ± 0.0050	0.007 ± 0.0030	0.181 ± 0.0088
Prejudice Remover	0.1%	0.861 ± 0.0016	0.084 ± 0.0051	0.008 ± 0.0037	0.208 ± 0.0046
BiFair	100%	0.862 ± 0.0016	0.067 ± 0.0089	0.012 ± 0.0077	0.179 ± 0.0125
BiFair	10%	0.860 ± 0.0031	0.045 ± 0.0105	0.015 ± 0.0128	0.144 ± 0.0091
BiFair	1%	0.852 ± 0.0051	0.038 ± 0.0067	0.016 ± 0.0108	0.130 ± 0.0149
BiFair	0.75%	0.859 ± 0.0046	0.043 ± 0.0073	0.017 ± 0.0083	0.139 ± 0.0048
BiFair	0.5%	0.861 ± 0.0057	0.047 ± 0.0080	0.016 ± 0.0069	0.145 ± 0.0102
BiFair	0.25%	0.861 ± 0.0039	0.055 ± 0.0115	0.027 ± 0.0124	0.150 ± 0.0091
BiFair	0.1%	0.859 ± 0.0037	0.058 ± 0.0166	0.022 ± 0.0200	0.153 ± 0.0178

(b) Bank dataset

Table 3: Values that are used to generate noisy dataset part of Figure 4. Note that, for this setting, every algorithm has access to a clean-labeled subset of the training data whose size is of 0.1%, and whose demographic attributes are known.

Algorithm	BAcc	AOD	EOD	SPD
Unconstrained	0.532 ± 0.0197	0.081 ± 0.0856	0.111 ± 0.1325	0.059 ± 0.0565
ARL	0.523 ± 0.0270	0.087 ± 0.0863	0.093 ± 0.0976	0.087 ± 0.0892
Kamiran Reweighing	0.528 ± 0.0225	0.070 ± 0.0790	0.095 ± 0.0948	0.060 ± 0.0608
Prejudice Remover	0.547 ± 0.0417	0.090 ± 0.0840	0.126 ± 0.1457	0.064 ± 0.0638
BiFair	0.733 ± 0.0069	0.067 ± 0.0325	0.082 ± 0.0364	0.151 ± 0.0345

(a) Adult dataset

Algorithm	BAcc	AOD	EOD	SPD
Unconstrained	0.542 ± 0.0258	0.079 ± 0.0586	0.091 ± 0.0490	0.071 ± 0.0703
ARL	0.515 ± 0.0156	0.055 ± 0.0478	0.066 ± 0.0534	0.050 ± 0.0468
Kamiran Reweighing	0.524 ± 0.0201	0.047 ± 0.0320	0.055 ± 0.0466	0.046 ± 0.0312
Prejudice Remover	0.531 ± 0.0221	0.056 ± 0.0375	0.088 ± 0.0672	0.024 ± 0.0207
BiFair	0.659 ± 0.0097	0.020 ± 0.0108	0.017 ± 0.0172	0.043 ± 0.0246

(b) Bank dataset