Fairness with Adaptive Weights

Junyi Chai [1] **Xiaoqian Wang** [1]

Abstract

Fairness is now an important issue in machine learning. There are arising concerns that automated decision-making systems reflect real-world biases. Although a wide range of fairness-related methods have been proposed in recent years, the *under-representation* problem has been less studied. Due to the uneven distribution of samples from different populations, machine learning models tend to be biased against minority groups when trained by minimizing the average empirical risk across all samples. In this paper, we propose a novel adaptive reweighing method to address representation bias. The goal of our method is to achieve group-level balance among different demographic groups by learning adaptive weights for each sample. Our approach emphasizes more on error-prone samples in prediction and enhances adequate representation of minority groups for fairness. We derive a closed-form solution for adaptive weight assignment and propose an efficient algorithm with theoretical convergence guarantees. We theoretically analyze the fairness of our model and empirically verify that our method strikes a balance between fairness and accuracy. In experiments, our method achieves comparable or better performance than state-of-the-art methods in both classification and regression tasks. Furthermore, our method exhibits robustness to label noise on various benchmark datasets.

1. Introduction

As machine learning techniques are widely applied in many fields, the fairness of machine learning has become an important issue. Automated decision-making algorithms, if not properly constrained, may make decisions that are discriminated to certain population groups. These groups are characterized by sensitive attributes, including age, gender, race, etc. There have been many cases showing that fairness is not a trivial problem in machine learning, and that fairness could not be achieved by simply removing the sensitive attributes from data. Research on COMPAS dataset (Chouldechova, 2017) shows that a well-calibrated classification algorithm tends to classify black defendants as of higher risk while classify white defendants as of lower risk. This discrimination results in much higher false positive rate for black defendants. Offers of same-day delivery coverage by Amazon are severely biased with respect to race, even if the algorithm does not take into consideration the race of customers (Ingold & Soper, 2016). For machine learning algorithms used in various fields of society, only having high classification accuracy is not enough. It is crucial to make sure that the algorithm is not biased against specific populations and does not reveal real-world discrimination, as many real-world datasets are unevenly distributed or biased.

Current approaches for fairness are mainly based on two different assumptions. Works including (Krasanakis et al., 2018) and (Jiang & Nachum, 2020) assume that the discrimination arises from the biased labeling process, which could result in wrongly labelled samples, and that there exists an underlying unbiased label mapping which maps the original label to the unbiased label based on different fairness criteria. The goal is to approximate the mapping by label-based methods, like adjusting the labels of sensitive groups (Jiang & Nachum, 2020) or adjusting the loss function values in sensitive groups (Krasanakis et al., 2018). Another widely applied assumption in fairness approaches is that the discrimination arises from the sensitive information that is correlated to sensitive attributes, and by removing the sensitive information from input data we are able to build classifiers that meet the specified fairness constraints. Methods including (Creager et al., 2019) are proposed based on this assumption. However, few methods on fairness have explicitly addressed the issue of representation bias, which arises due to insufficient amount of data in certain groups or subgroups. This kind of bias is also referred to as *under-representation*. One intuitive idea regarding this problem is: if all the groups and subgroups are adequately represented with enough data, the bias would be greatly reduced, if not eliminated. For representation bias, the sensitive attributes are not necessarily correlated with input features,

[1]Elmore Family School of Electrical and Computer Engineering, Purdue University. Correspondence to: Xiaoqian Wang <joywang@purdue.edu>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

while most works on fairness assume that the sensitive attributes are correlated with input features. Thus, conventional methods on fairness do not always work regarding underrepresentation. One related area regarding this issue is imbalanced classification, where the goal is to improve the classification accuracy of minor groups. However, methods on imbalanced classification are more concerned about overall classification accuracy, and fairness is not considered as an important metric in imbalanced classification.

We draw inspiration from cost-sensitive learning on imbalanced classification, where the goal is to improve classification accuracy of minor groups by assigning different weights to major groups and minor groups. These methods, however, have one drawback that all samples within the same group are assigned with equal weight. During the training stage, samples that are very far from the decision boundary are very unlikely to be wrongly classified, while samples that lie close to the decision boundary are more likely to be misclassified. One simple solution to this issue is to assign weights proportional to classification error. However, such attempt does no guarantee fairness, and training under such assignment could be very unstable and fail to converge.

In this paper, we propose an adaptive weight assignment strategy to improve fairness. Compared to the two reweighing methods mentioned above, our method is a sample-based method that addresses representation bias by constraining the model to be more careful with minority groups, which differs from previous methods by estimating the underlying labels or by removing the correlation between sensitive attributes and modified input features. Our method learns adaptive weights for samples within the same group based on the probability of the samples being misclassified. During the training stage, only partial samples (those that are more likely to be misclassified) within one group are assigned with positive weights, while other samples (those that are easier to be correctly classified) are assigned with zero weights. By controlling the amounts of samples assigned with positive weights, our method controls the trade-off between accuracy and fairness, and fairness criteria is guaranteed by constraining the sum of weights in different groups. The algorithm is trained and tested on the original datasets with features and labels unaffected, but the desired fairness criteria are still achieved. This improves the *interpretability* of learning algorithms since the natural meaning of features is preserved.

We summarize our contribution as follows:

1. We formulate a novel sample-based reweighing method with a closed-form solution on weight assignment for mitigating representation bias.

2. By balancing between demographic groups without specifying fairness metrics during training, our reweighing method improves fairness and robustness under different metrics with theoretical guarantee.

3. Our model improves fairness in both classification and regression tasks and is robust to label noise.

2. Related Work

Most works on fairness focus on binary classification under binary sensitive attributes. Generally, these methods can be divided into three categories: preprocessing, in-processing and post-processing. Methods on preprocessing are proposed to modify the input data to eliminate potentially biased information and the classifier is subsequentially trained and applied on the modified data. Approaches to preprocess data include sample selection (Roh et al., 2021), fair representation learning (Tan et al., 2020; Madras et al., 2018), disentanglement of sensitive information and preprocessed data (Creager et al., 2019), fair data generation (Xu et al., 2019; Jang et al., 2021b), data mapping (Calmon et al., 2017), etc. In-processing methods achieves fairness by imposing extra constraints during training to obtain a fairer classifier. Methods in this category add fairness constraints to the objective in terms of model regularization (Aghaei et al., 2019; Berk et al., 2017), adversarial layers (Adel et al., 2019), uncorrelation constraints between decision boundary and sensitive attribute (Zafar et al., 2017), and mutual information (Roh et al., 2020), etc. Post-processing focuses on how to modify output in different demographic groups to achieve fairness (Hardt et al., 2016; Jang et al., 2021a), which adjusts decision thresholds of different groups according to specific fairness criteria.

Compared with fair classification, fair regression has received less attention. Agarwal et al. (2019) propose two criteria for fair regression and propose a reduction-based fair regression model. Chzhen et al. (2020) derive a close form expression for the optimal fair predictor based on optimal transport. Steinberg et al. (2020) propose an efficient way to estimate mutual information between the prediction and sensitive information, and uses such information as regularization for fair regression.

Imbalanced classification methods are generally divided into two parts: resampling and cost-sensitive learning. Conventionally, methods on resampling try to achieve balance between groups by either oversampling the minority or undersampling the majority. Recent works seek to preform resampling in both major groups and minor groups. Bao et al. (2019) perform classification using clustering centers in learned latent space, which is equivalent to undersampling all demographic groups. Cost-sensitive learning targets at solving the problem by assigning different weights to different samples, instead of directly sampling the dataset, such that samples within minor groups or samples of greater

importance are more carefully treated compared with unit weights. One intuitive idea is to assign higher weights to minor groups during training such that the cost of misclassification of minor groups are higher than that of major groups. Zhang et al. (2018) propose to improve the performance of deep belief network for imbalanced classification by assigning weights optimized by evolutionary algorithm. Huang et al. (2019) propose to achieve representation balance between groups by constraining the embedding to maintain inter-cluster margins both within and between classes. Other methods, including (Wang et al., 2017), propose to use transfer learning to transfer representation of major groups to minor groups for an imbalanced dataset.

There are other works that address the problem of biased distribution. Methods including (Choi et al., 2019) and (Kim et al., 2019) are proposed based on minimizing the mutual information of bias label and embedding through adversarial training. Alvi et al. (2018) propose to add extra confusion loss in the training objective to make sure that the learned feature embeddings are invariant to bias information.

Our work is different from previous works on fairness in that we achieve fairness from a perspective of group balance, instead of data rectification, which we fulfill by adaptively assigning weights to ensure a fair and sufficient representation of samples within each group. Our method does not impose particular fairness constraints in the objective function. The weakness of imposing specific constraints is that the model could only perform well in terms of the specific fairness notion imposed, and linear relaxations of fairness constraints could be too relaxed (Lohaus et al., 2020) in terms of bias mitigation. Compared with methods on cost-sensitive learning, our method takes into consideration both the protection of minor groups and the difference of samples within each group. Kamiran & Calders (2012) propose to assign weight per sample with the goal of achieving independence between sensitive attribute and label.

3. Method

3.1. Fairness Notions

Consider a binary classifier f_θ with parameter θ which maps input features x to binary label $\hat{y} \in \{0, 1\}$. Denote as $y \in \{0, 1\}$ the original label of feature x, and $s \in \{0, 1\}$ the sensitive attribute. The learned mapping of classifier can be formulated as: $\hat{y} = f_\theta(x, s)$.

Disparate treatment (Zafar et al., 2017) exists when the classifier makes different predictions on samples from different demographic groups given that the input features are identical. To eliminate disparate treatment, the classifier should achieve calibration across groups: $p(\hat{y}|x, s) = p(\hat{y}|x)$.

Disparate impact (Kamiran & Calders, 2012) measures the statistical parity, i.e. the difference in positive outcome rate between different groups. Disparate impact between different groups is eliminated when the prediction $\hat{y}$ is independent of s: $p(\hat{y}|s = 0) = p(\hat{y}|s = 1)$. However, by simply eliminating disparate impact we are not guaranteed a fair classifier. Under the condition that the distribution of training samples is uneven, by restricting this fairness objective the classifier could make good decisions on major group, while making poor (or even random) decisions on minor group. Besides, achieving zero disparate impact could be against an optimal classifier when statistical features of different demographic groups vary.

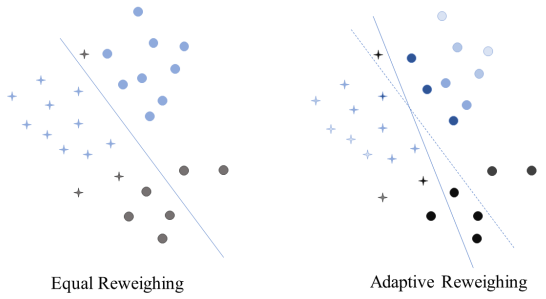


Figure 1. Demonstration of our adaptive reweighing method. Points of different shapes indicate samples of different labels, different colors refer to different groups and darker color indicates higher weight. In equal reweighing, all points within one group are assigned with the same weight, while our method takes into account both how likely samples are to be misclassified and balance among different groups. In this way, samples in each group are better represented, and fairness is achieved.

Disparate mistreatment (Hardt et al., 2016) arises when the misclassification rates, measured by false positives and false negatives of different groups are different. Compared with disparate impact, disparate mistreatment relies on the original labels, and is thus a post-hoc criterion. Works including (Chouldechova, 2017) demonstrate that unless the classifier achieves accuracy of 100%, it is impossible to achieve all disparate mistreatment criteria at once. Thus, disparate FPR (false positive rate) and FNR (false negative rate) are commonly adopted to measure disparate mistreatment:

$$p(\hat{y} \neq y|y = 1, s) = p(\hat{y} \neq y|y = 1),$$
$$p(\hat{y} \neq y|y = 0, s) = p(\hat{y} \neq y|y = 0).$$

Fairness notions for regression problem shares similar properties as those of classification problems. Consider a training set $\{(x_i, y_i, s_i), 1 \leq i \leq N\}$ and a regression model which maps input feature x to $\hat{y} \in [0, 1]$, a common way to measure the performance of regression model is by mean squared loss (MSE):

$$l_{mse} = \frac{1}{N} \sum_{i=1}^{N} (\hat{y}_i - y_i)^2.$$

One way to measure disparity of regression model is bounded group loss, which requires that the predictor's average remains under acceptable level for each group. As proposed by Agarwal et al. (2019), the MSE of each demographic group should be upper-bounded by a certain value: $\frac{1}{N_s}\sum_{s_i=s}(\hat{y}_i - y_i)^2 \leq \epsilon_s, \forall s$.

Another way to measure disparity is statistical parity (Agarwal et al., 2019). A predictor satisfies statistical parity if the output is independent of the protected attribute: $p[f(x) \geq a|s] = p[f(x) \geq a]$.

3.2. Problem Formulation

For a given dataset $\{x_1, x_2, ... x_n\}$ of n samples, the vanilla training objective without reweighing can be formulated as:

$$\min_{\theta} \frac{1}{n}\sum_{i=1}^{n} L(y_i, f_\theta(x_i)). \tag{1}$$

One problem of this unweighted training is that the classifier tends to focus more on major group. To resolve representation bias in learning algorithm, one simple way is to assign weight to samples in each group as $w_i = \frac{c}{n'}$, where n' denotes the number of samples in the group that the i-th sample belongs to, and c is a constant. This method, however, does not take into account difference between samples within groups, and does not always improve fairness considering different metrics. By assigning weight to samples of different groups, we want to make sure that the classifier pays more attention to samples that are wrongly or likely to be wrongly classified. Besides, we want to make sure that the weighted samples are balanced between groups. This can be achieved by constraining the sum of weights within groups. Thus, the reweighing problem can be formulated as:

$$\max_{w} \sum_{s}\sum_{i\in g_s} w_i L(y_i, \hat{y}_i) \;\; s.t. \;\; \sum_{i\in g_s} w_i = c, w \geq 0, \tag{2}$$

where g_s collects the indices of samples belonging to the demographic group s.

The optimization problem in (2) can be partitioned by groups as follows for each demographic group:

$$\max_{w} \sum_{i=1}^{n'} w_i L(y_i, \hat{y}_i) \;\; s.t. \;\; w^T 1 = c, w \geq 0.$$

Here we use n' instead of n to denote samples within one group because we want to perform reweighing within each specific group. It is very easy to tell that the solution to this problem is to assign all the weight to samples of the largest loss in each group and all other weights zero. However, for group fairness we are concerned about the partial of samples that are likely to be wrongly classified, not the only one sample that are most likely to be misclassified. Besides, such one-hot encoding would lead to very unstable training, and the algorithm could possibly fail to converge. To address this problem, we introduce one more regularization term, and the overall optimization problem regarding w for each group is:

$$\begin{aligned} &\max_{w} \sum_{i=1}^{n'} w_i L(y_i, \hat{y}_i) - \alpha \left\|w\right\|_2^2 \\ &s.t. \;\; w^T 1 = c, w \geq 0\,. \end{aligned} \tag{3}$$

And by adjusting the value of α we can adjust the amount of samples that receive non-zero weights. As α approaches infinity, the first term gradually becomes trivial, and the solution changes from that of (2) to equal weight $\frac{c}{n'}$.

3.3. Solution of Adaptive Weights

We derive the closed-form solution of adaptive weights w in Problem (3) as follows:

Theorem 3.1. *Consider a classifier with parameter θ such that $\hat{y}_i = f_\theta(x_i)$. Without loss of generality, assume that losses $l_i = L(y_i, \hat{y}_i), i = 1, 2, \ldots, n'$ in a given sensitive group are sorted in descending order such that $l_i \geq l_j, \forall i > j$. Then the optimal solution $w^* \in \mathbb{R}^{n'}$ of Problem (3) is:*

$$w_i^* = \max(\frac{l_i - \lambda}{2\alpha}, 0), \;\; i = 1, 2, ..., n',$$

where $\lambda = \frac{\sum_{j=1}^{k'} l_j - 2\alpha c}{k'}$, and k' is determined by $\sum_{j=1}^{k'} l_j - k' l_{k'+1} > 2\alpha c > \sum_{j=1}^{k'} l_j - k' l_{k'}$. When $\sum_{j=1}^{k'} l_j - 2\alpha c \leq 0, \forall k'$, we can still follow the update rule except that we now have $k' = n'$ and $\lambda \leq 0$.

Proof Sketch: The closed-form solution of Problem (3) can be obtained using the method of Lagrange multipliers. We discuss the optimal Lagrange multiplier λ via Karush–Kuhn–Tucker (KKT) conditions. Detailed proof of Theorem 3.1 is in the Appendix.

Insights: From Theorem 3.1 we can see that both α and c have an impact on the optimal w^*. When α and c are relatively large, all samples within the group receive non-zero weights, and part of the samples receive loss-based adaptive weights, while the other samples receive equal weights. Under this condition our method acts like a combination of adaptive weighing and equal weighing.

We also notice that our training objective within each subgroup is similar to that of (Hashimoto et al., 2018). In this way, our method can be seen as a combination of group-level balance and robustness within each group, i.e., our method not only achieves fairness, but also introduces distributional

robustness within each subgroup to the predictor. We validate the effectiveness of robustness within each group in experimental section.

Intuitively, our method achieves two goals: group balance and error-prone reweighing. During training, the constraint c for each subgroup is chosen as the number of samples in the major subgroup. In this way, our method can also be seen as adjusting the distribution of reweighed training samples in minor groups to ensure fairness. Connection between our reweighed minimization problem and fairness metric is stated as below:

Theorem 3.2. *Consider a classifier f_θ with parameter θ such that $\hat{y}_i = f_\theta(x_i)$. Given the adaptive weight w^* by optimizing Problem (3), under the L_1-norm loss or the cross-entropy loss for $L(y_i, \hat{y}_i)$, the following fairness metrics*

- *Disparate mistreatment:*

$$\sum_s \quad (|p(\hat{y} \neq y | y = 1, s) - p(\hat{y} \neq y | y = 1)| + |p(\hat{y} \neq y | y = 0, s) - p(\hat{y} \neq y | y = 0)|)$$

- *Equal opportunity:*

$$\sum_s (|p(\hat{y} \neq y | y = 1, s) - p(\hat{y} \neq y | y = 1)|)$$

are upper bounded by our weighted loss up to a multiplicative constant: $\sum_{i=1}^{n} w_i^* L(y_i, \hat{y}_i)$.

Proof Sketch: The proof of Theorem 3.2 is obtained by using the formulation of w^* in (3). Since w^* is the closed-form solution for the maximization problem in (3), the weighted loss upper bounds the equal loss (where samples in each demographic share the same weight). Detailed proof of Theorem 3.2 can be founded in Appendix.

Insights: Theorem 3.2 indicates that disparate mistreatment and equal opportunity of the classifier f_θ is bounded by optimizing the adaptively weighted loss function, i.e., our method optimizes simultaneously w.r.t. both classification accuracy and fairness.

3.4. Training Algorithm and Convergence

Our training algorithm follows the idea of (Lu et al., 2020). We directly update new weights per sample rather than partially editing existing ones. In r-th iteration, the training algorithm performs the following optimization:

$$w = \arg \max_{w^T 1 = c, w \geq 0} \sum_{i=1}^{N} w_i l_i - (c_1 + \alpha) \|w\|^2,$$

$$\theta = \arg \min_{\theta} \sum_{i=1}^{N} w_i l_i + c_2 \left\|\theta - \theta^r\right\|^2,$$

where θ represents the parameters of classifier f_θ and $l_i = L(y_i, f_\theta(x_i))$.

To adaptively adjust the sample weights based on the performance of classifier, we first pre-train the unconstrained classifier with equal weights to find hard samples and obtain training loss. We then update the weight according to the closed-form solution as stated in the previous subsection and train the classifier with such weight. This process is iterated until the weight w converges. The detailed training is shown in Algorithm 1. Compared with the baseline classifier, our Algorithm 1 has only one extra step – updating w in Step 3. This step takes $O(nlogn)$ time due to the sorting of l_i in Theorem 3.1, where n is the number of training samples.

As proved by Lu et al. (2020), if the training objective satisfies strong convexity and Lipschitz continuity w.r.t. the minimization problem, and satisfies strongly concavity w.r.t. the maximization problem, then with a proper α, the training objective is guaranteed to converge:

$$\mathbf{F}^{r+1} - \mathbf{F}^r < -c_1 \left\|w^{r+1} - w^r\right\|^2 - c_2 \left\|\theta^{r+1} - \theta^r\right\|^2.$$

where $\mathbf{F}^r$ is the objective value of (3) at the r-th iteration, c_1 and c_2 are constants.

Algorithm 1 Adaptive Reweighing Algorithm

Pre-train a classifier f with parameter θ^{prev} by optimizing (1). Set $w_i \leftarrow 1, w_i^{\text{prev}} \leftarrow \infty, \forall i$. Set c_1 to be sufficiently small, and $c_2 = 1$.

while $\sum_i (w_i - w_i^{\text{prev}})^2 > \delta^2$ **do**

1. Set $w_i^{\text{prev}} \leftarrow w_i, \forall i$;
2. Set $\theta \leftarrow \theta^{\text{prev}}$;
3. Update w by solving (3) in each group:

$$\max_w \sum_{i=1}^{n'} w_i L(y_i, \hat{y}_i) - (c_1 + \alpha) \|w\|^2,$$
$$s.t.\ w^T 1 = c, w \geq 0;$$

4. Train the classifier f_θ by minimizing the weighted training loss with penalty:

$$\arg \min_\theta \sum_{i=1}^{n} w_i L(y_i, f_\theta(x_i)) + c_2 \|\theta - \theta^{\text{prev}}\|^2.$$

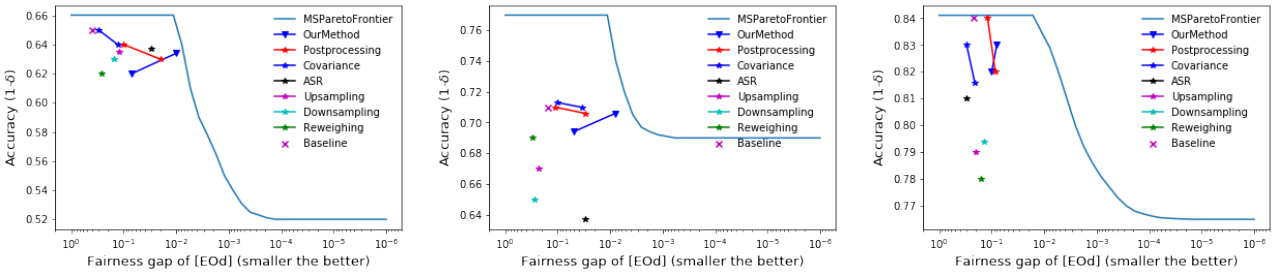
end while

Return f_θ, w

4. Experiments

4.1. Experimental Setup

We evaluate our model on three benchmark classification datasets: Adult dataset (Dua & Graff, 2017), the UCI German credit risk dateset (Dua & Graff, 2017), the ProPublica



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

Figure 2. Pareto frontier on COMPAS, Adult, and German credit datasets.

Table 1. Experimental results on Adult dataset.

Method	Baseline	Reweighing	Undersampling	Oversampling	ASR	Postprocessing	Covariance	Ours
Accuracy	84.14±0.34	78.36±0.82	79.48±0.31	79.34±0.42	81.52±0.26	82.21±0.44	81.67±0.42	**82.44±0.33**
Disparate Impact	18.40±0.29	23.22±0.26	18.49±0.43	19.65±0.36	**0.31±0.07**	13.02±1.02	19.11±2.67	17.14±0.47
Disparate TPR	14.60±0.95	**1.59±4.70**	8.90±1.12	6.49±1.52	29.81±3.46	4.11±0.76	9.20±1.41	1.74±1.01
Disparate TNR	7.87±0.63	13.76±3.41	4.18±0.84	13.76±0.68	**2.84±0.85**	4.02±1.21	12.24±2.03	7.17±0.62

Table 2. Experimental results on German dataset.

Method	Baseline	Reweighing	Undersampling	Oversampling	ASR	Postprocessing	Covariance	Ours
Accuracy	71.62±2.02	68.24±2.96	65.52±3.25	67.32±2.45	70.14±1.51	70.61±0.22	**71.23±0.36**	70.60±1.37
Disparate Impact	14.23±5.56	7.06±5.94	11.17±6.01	9.16±4.23	5.13±0.95	4.25±0.47	4.72±0.13	**3.46±1.34**
Disparate TPR	8.34±7.46	21.02±3.47	13.81±11.79	12.14±8.31	1.13±0.52	2.51±0.12	1.23±0.20	**0.64±0.43**
Disparate TNR	6.35±4.82	9.96±4.23	14.59±4.84	11.37±4.17	1.27±0.61	1.03±0.07	1.15±0.31	**0.14±0.12**

Table 3. Experimental results on COMPAS dataset.

Method	Baseline	Reweighing	Undersampling	Oversampling	ASR	Postprocessing	Covariance	Ours
Accuracy	65.23±1.39	62.24±2.47	63.34±2.41	63.50±2.42	63.75±1.27	63.42±1.14	**64.11±1.46**	63.41±1.35
Disparate Impact	22.29±4.76	9.13±3.16	8.45±2.68	8.55±2.83	2.31±0.25	2.33±0.10	7.36±1.03	**1.82±0.11**
Disparate TPR	21.14±7.14	6.46±2.14	9.32±3.86	7.02±3.44	1.07±0.33	1.06±0.16	3.38±0.71	**1.02±0.09**
Disparate TNR	17.41±3.72	19.11±3.22	5.77±1.73	5.25±1.40	1.14±0.21	1.20±0.21	10.28±2.33	**0.24±0.17**

Table 4. Experimental results of nonlinear classifier on Adult dataset.

Method	Baseline	Reweighing	Undersampling	Oversampling	ASR	Postprocessing	Covariance	Ours
Accuracy	83.14±0.54	77.63±0.74	77.76±0.81	78.73±0.47	80.67±0.53	81.47±0.63	80.21±0.47	**81.78±0.17**
Disparate Impact	17.62±0.63	23.61±0.31	18.21±0.25	18.23±0.26	**0.31±0.04**	13.17±1.67	18.37±2.67	15.52±0.16
Disparate TPR	15.37±1.33	**1.76±5.16**	8.24±1.12	7.73±2.13	27.21±3.37	4.27±1.06	9.14±1.34	1.67±0.92
Disparate TNR	7.47±1.16	13.34±4.21	4.17±1.26	13.26±2.41	**2.31±0.95**	4.43±1.65	11.34±1.84	6.87±0.62

Table 5. Experimental results of nonlinear classifier on German dataset.

Method	Baseline	Reweighing	Undersampling	Oversampling	ASR	Postprocessing	Covariance	Ours
Accuracy	71.62±1.76	68.19±2.75	66.24±2.67	67.74±2.46	70.13±1.51	70.63±0.67	**71.06±0.52**	70.62±1.14
Disparate Impact	15.26±7.27	8.14±5.13	11.47±6.67	9.86±3.67	6.17±1.51	4.21±0.87	4.43±0.44	**3.16±1.11**
Disparate TPR	8.45±6.67	19.94±4.46	13.71±12.34	12.67±7.74	1.53±0.81	3.01±0.83	1.16±0.21	**0.63±0.58**
Disparate TNR	6.32±4.17	9.41±4.47	15.52±5.67	11.84±5.14	1.77±0.41	1.21±0.17	1.15±0.36	**0.32±0.15**

Table 6. Experimental results of nonlinear classifier on COMPAS dataset.

Method	Baseline	Reweighing	Undersampling	Oversampling	ASR	Postprocessing	Covariance	Ours
Accuracy	64.17±1.13	61.18±1.78	62.76±2.26	62.35±2.13	63.17±1.21	63.14±1.16	**63.64±1.31**	63.23±1.64
Disparate Impact	21.37±5.24	10.17±2.27	8.83±2.69	8.67±3.12	2.41±0.31	3.24±0.11	7.43±1.22	**2.23±0.87**
Disparate TPR	22.21±8.17	6.85±2.13	8.86±3.11	7.44±2.57	1.82±0.46	1.31±0.17	3.13±0.76	**1.16±0.08**
Disparate TNR	17.64±3.46	18.85±4.41	5.41±1.68	6.13±1.25	1.71±0.43	1.24±0.23	11.47±2.63	**0.69±0.34**

COMPAS dataset (Larson et al., 2016), and two regression datasets: Law School (Wightman, 1998), Communities & Crime (CRIME) dataset. Details of the datasets are in the Appendix. For pre-processing of the above datasets, We perform one-hot coding on each non-numerical feature and normalize each numerical feature by subtracting the mean value and scaling to unit variance.

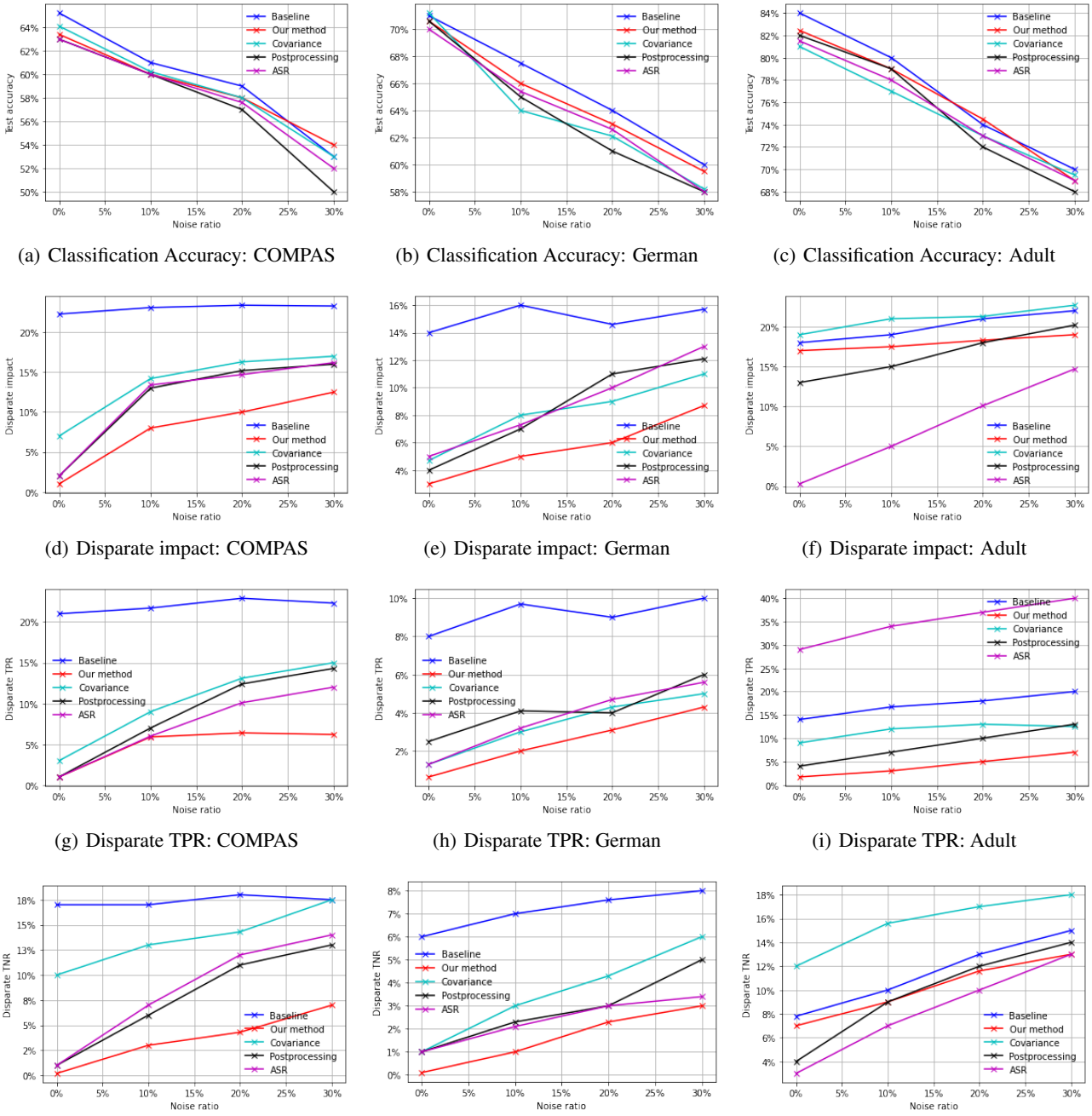
For classification, we build all classifiers based on logistic regression. We use accuracy as the evaluation metric. For fairness metrics we adopt disparate impact, disparate TPR and disparate TNR. We compare our method with six related methods: *Baseline*: logistic regression without fairness constraints; *Reweighing*: logistic regression with assigning balancing weights to different population groups; *Undersampling*: balancing between different demographic groups by randomly selecting samples and form the new training dataset such that samples in each group are of the same number; *Oversampling*: balancing between different groups by duplicating samples in minor groups such that samples in each group are of the same number; *Adaptive sensitive reweighing (ASR)*: logistic regression with adaptive sensitive reweighing (Krasanakis et al., 2018). *Covariance*: logistic regression with linear covariance constraints to mitigate disparate impact and mistreatment (Zafar et al., 2017). *Postprocessing*: logistic regression with postprocessing (Hardt et al., 2016).

For regression task, we build all models based on Regularized Least Squares (RLS). We use mean squared error (MSE) as the evaluation metric and statistical parity (SP) as fairness metric. Similar to the idea of fair classification, we define subgroups of each training set based on the sensitive information and the cutoff threshold $\mathbb{1}[y \geq 0.5]$. For instance, we treat all samples of $s = s_0$ and $y \geq 0.5$ as one subgroup. We compare our method with four related methods: *Baseline*: RLS without fairness constraints. *Reweighing*: RLS with assigning balancing weights to different subgroups. *Oversampling*: RLS with duplicating samples in minor groups such that each group contains same number of samples. *Undersampling*: RLS with major group randomly sampled such that each group contains same number of samples. *Fair Wasserstein barycenters (FWB)* : RLS with fair Wasserstein barycenter (Chzhen et al., 2020).

Table 7. Experimental results on Law school dataset.

Method	MSE	SP
Baseline	0.114±0.003	15.20±4.34%
Oversampling	0.152±0.004	9.62±3.17%
Undersampling	0.163±0.002	8.57±4.52%
FWB (Chzhen et al., 2020)	0.141±0.004	**2.13±0.13**%
Our method	**0.135±0.004**	2.16±0.19%

We repeat experiments on each dataset five times and before each repetition we randomly spilt data into 80% training data and 20% test data. All the methods evaluated are trained and tested on the same data partitions each time. Values of hyperparameter α in our method are set by performing cross-validation on training data in the value range of 1 to 20. The hyperparameters for the comparing methods are tuned as suggested by the authors (Krasanakis et al., 2018; Hardt et al., 2016; Zafar et al., 2017).



(a) Classification Accuracy: COMPAS (b) Classification Accuracy: German (c) Classification Accuracy: Adult

(d) Disparate impact: COMPAS (e) Disparate impact: German (f) Disparate impact: Adult

(g) Disparate TPR: COMPAS (h) Disparate TPR: German (i) Disparate TPR: Adult

(j) Disparate TNR: COMPAS (k) Disparate TNR: German (l) Disparate TNR: Adult

Figure 3. Change of accuracy and fairness under different noise ratio on COMPAS, Adult and German credit datasets.

4.2. Fairness in Classification

For classification dataset, we first evaluate the trade-off between fairness and accuracy using Pareto frontier. Kim et al. (2020) propose to model the classifier performance in terms of both accuracy and fairness using an 8-dimensional fairness vector z. In this way, the fairness constraints can be modelled as a linear system with specific fairness matrices. As suggested by the authors, by solving the optimization problem of z with model-specific constraints and fairness regularization terms, we can obtain the Pareto frontier with the (ϵ, δ) solutions. In Figure 2, the blue curve is the model-specific Pareto frontier, which represents the optimal trade-off that could be achieved for a post-processed logistic regression classifier. Smaller distance to the Pareto

Table 8. Experimental results on CRIME dataset.

Method	MSE	SP
Baseline	0.037±0.003	50.63±6.75%
Oversampling	0.052±0.004	21.37±7.73%
Undersampling	0.047±0.006	19.42±6.63%
FWB (Chzhen et al., 2020)	**0.042±0.004**	12.10±1.19%
Our method	0.043±0.004	**11.47±1.63%**

frontier means better fairness-accuracy trade-off, and solutions lie right on the curve represent an ideal classifier. As can be seen from the figures, our method lies closer to the frontier than other methods, which demonstrates that our method achieves better fairness-accuracy trade-off.

We further show more fairness and accuracy results in Tables 1 to 3. Among all the compared methods, our method achieves the best fairness metrics on German credit dataset and COMPAS dataset and the best accuracy on Adult dataset excluding baseline. Our method achieves low disparate TPR and TNR on both German credit dataset and COMPAS dataset, and low disparate TPR on Adult dataset. It can be seen clearly from Table 1 that ASR achieves low disparate impact and disparate TNR, but at the cost of much higher disparate TPR than baseline. Covariance achieves relatively good performance on German credit dataset, but the disparate TNR on Adult dataset and COMPAS dataset are relatively high. Our method achieves low disparate TPR without increasing disparate TNR and disparate impact, and works better in removing imbalance between positive subgroups. Methods including reweighing and resampling achieve good performance in terms of one certain fairness metric, but their performance in terms of other fairness metrics is not satisfactory, and the classification accuracies are relatively low. Besides, the standard deviation of metrics of these methods are relatively high, which shows the instability of simply performing reweighing or resampling on training data without imposing further constraints. By contrast, our method achieves fairness with more stable performance.

We further show results on non-linear classifiers in Table 4 to 6. Specifically, we build all classifiers based on multilayer perceptron (MLP). Our method still achieves best or comparable performance on all three datasets, which also validates the effectiveness of our method.

It is noteworthy that our method achieves relatively bad performance on Adult dataset in disparate impact. However as discussed before, disparate impact alone is not an adequate metric to evaluate fairness as it merely computes the difference in positive outcome rates, without considering the base rate. And our method achieves comparable disparate TPR and TNR than the best results with less standard deviation. More results on sensitivity w.r.t. hyperparameters are shown in the Appendix.

4.3. Classification with Noisy Label

Results on fair classification under average noisy label are shown in Figure 3. Specifically, we apply half the noise corruption on major group and half the corruption on minor group. Due to distribution disparity in different groups, this would result in a group-dependent noise. As proved by Wang et al. (2021), imposing fairness constraints under group-dependent noise would induce a deviation on fairness under clean data. However, our method naturally bypasses such problem, since our method achieves fairness through group balance and loss-based reweighing, without directly imposing fairness constraints. Besides, our loss-based reweighing achieves distributionally robust within each subgroup. Compared with other methods, our method achieves better robustness in terms of both fairness and accuracy under different noise ratios, while performance of other methods become very unstable as the noise ratio increases.

4.4. Regression Results

We further validate our method for fair regression. As the results shown in Table 7 and 8, amongst all compared methods, our method achieves the lowest MSE on Law school dataset excluding baseline and the lowest SP on Communities & crime dataset. Compared with Oversampling and Undersampling, our method better achieves fairness and MSE, which validates the effectiveness of our loss-based reweighing. Besides, the results show that our method is competitive with state-of-the-art methods regarding both MSE and SP.

5. Conclusion

Representation bias is an important yet less studied problem in fairness. In this paper we discuss how reweighing helps in mitigating representation bias. We propose a sample-level adaptive reweighing method to flexibly solve this problem. Instead of considering specific fairness criteria, our method achieves fairness by balancing different groups. The objective function for weight assignment consists of two parts: the first part underscoring wrongly-classified samples in different groups, and the second part addressing the trade-off between fairness and equal weights. We derive closed-form solution for the weight assignment, and experiments on five benchmark datasets show that our method can achieve fairness with relatively small sacrifice in accuracy or MSE.

Acknowledgements

This work was partially supported by the EMBRIO Institute, contract #2120200, a National Science Foundation (NSF) Biology Integration Institute, Purdue's Elmore ECE Emerging Frontiers Center, and NSF IIS #1955890.

References

Adel, T., Valera, I., Ghahramani, Z., and Weller, A. One-network adversarial fairness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 2412–2420, 2019.

Agarwal, A., Dudík, M., and Wu, Z. S. Fair regression: Quantitative definitions and reduction-based algorithms. In *International Conference on Machine Learning*, pp. 120–129. PMLR, 2019.

Aghaei, S., Azizi, M. J., and Vayanos, P. Learning optimal and fair decision trees for non-discriminative decision-making. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 1418–1426, 2019.

Alvi, M., Zisserman, A., and Nellaker, C. Turning a blind eye: Explicit removal of biases and variation from deep neural network embeddings. *arXiv preprint arXiv:1809.02169*, 2018.

Bao, F., Deng, Y., Kong, Y., Ren, Z., Suo, J., and Dai, Q. Learning deep landmarks for imbalanced classification. *IEEE Transactions on Neural Networks and Learning Systems*, 31(8):2691–2704, 2019.

Berk, R., Heidari, H., Jabbari, S., Joseph, M., Kearns, M., Morgenstern, J., Neel, S., and Roth, A. A convex framework for fair regression. *arXiv preprint arXiv:1706.02409*, 2017.

Calders, T., Karim, A., Kamiran, F., Ali, W., and Zhang, X. Controlling attribute effect in linear regression. In *2013 IEEE 13th international conference on data mining*, pp. 71–80. IEEE, 2013.

Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., and Varshney, K. R. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, pp. 3992–4001, 2017.

Choi, J., Gao, C., Messou, J. C., and Huang, J.-B. Why can't i dance in the mall? learning to mitigate scene bias in action recognition. In *Advances in Neural Information Processing Systems*, pp. 853–865, 2019.

Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017.

Chzhen, E., Denis, C., Hebiri, M., Oneto, L., and Pontil, M. Fair regression with wasserstein barycenters. *arXiv preprint arXiv:2006.07286*, 2020.

Creager, E., Madras, D., Jacobsen, J.-H., Weis, M., Swersky, K., Pitassi, T., and Zemel, R. Flexibly fair representation learning by disentanglement. In *International Conference on Machine Learning*, pp. 1436–1445, 2019.

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

Feng, L., Shu, S., Lin, Z., Lv, F., Li, L., and An, B. Can cross entropy loss be robust to label noise? In *International Joint Conference on Artificial Intelligence*, pp. 2206–2212, 2020.

Hardt, M., Price, E., and Srebro, N. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pp. 3315–3323, 2016.

Hashimoto, T., Srivastava, M., Namkoong, H., and Liang, P. Fairness without demographics in repeated loss minimization. In *International Conference on Machine Learning*, pp. 1929–1938. PMLR, 2018.

Huang, C., Li, Y., Chen, C. L., and Tang, X. Deep imbalanced learning for face recognition and attribute prediction. *IEEE transactions on pattern analysis and machine intelligence*, 42(11):2781–2794, 2019.

Ingold, D. and Soper, S. Amazon doesn't consider the race of its customers. should it? *Bloomberg*, 2016.

Jang, T., Shi, P., and Wang, X. Group-aware threshold adaptation for fair classification. *arXiv preprint arXiv:2111.04271*, 2021a.

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

Jiang, H. and Nachum, O. Identifying and correcting label bias in machine learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 702–712. PMLR, 2020.

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

Kim, B., Kim, H., Kim, K., Kim, S., and Kim, J. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9012–9020, 2019.

Kim, J. S., Chen, J., and Talwalkar, A. Model-agnostic characterization of fairness trade-offs. *arXiv preprint arXiv:2004.03424*, 2020.

Krasanakis, E., Spyromitros-Xioufis, E., Papadopoulos, S., and Kompatsiaris, Y. Adaptive sensitive reweighting to mitigate bias in fairness-aware classification. In *Proceedings of the 2018 World Wide Web Conference*, pp. 853–862, 2018.

Larson, J., Mattu, S., Kirchner, L., and Angwin, J. Compas analysis. *GitHub, available at: https://github.com/propublica/compas-analysis*, 2016.

Lohaus, M., Perrot, M., and Von Luxburg, U. Too relaxed to be fair. In *International Conference on Machine Learning*, pp. 6360–6369. PMLR, 2020.

Lu, S., Tsaknakis, I., Hong, M., and Chen, Y. Hybrid block successive approximation for one-sided non-convex min-max problems: algorithms and applications. *IEEE Transactions on Signal Processing*, 68:3676–3691, 2020.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. Learning adversarially fair and transferable representations. In *International Conference on Machine Learning*, pp. 3384–3393, 2018.

Roh, Y., Lee, K., Whang, S., and Suh, C. Fr-train: A mutual information-based approach to fair and robust training. In *International Conference on Machine Learning*, pp. 8147–8157. PMLR, 2020.

Roh, Y., Lee, K., Whang, S., and Suh, C. Sample selection for fair and robust training. *Advances in Neural Information Processing Systems*, 34:815–827, 2021.

Steinberg, D., Reid, A., O'Callaghan, S., Lattimore, F., McCalman, L., and Caetano, T. Fast fair regression via efficient approximations of mutual information. *arXiv preprint arXiv:2002.06200*, 2020.

Tan, Z., Yeom, S., Fredrikson, M., and Talwalkar, A. Learning fair representations for kernel models. In *International Conference on Artificial Intelligence and Statistics*, pp. 155–166. PMLR, 2020.

Wang, J., Liu, Y., and Levy, C. Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pp. 526–536, 2021.

Wang, Y.-X., Ramanan, D., and Hebert, M. Learning to model the tail. In *Advances in Neural Information Processing Systems*, pp. 7029–7039, 2017.

Wightman, L. F. Lsac national longitudinal bar passage study. lsac research report series. 1998.

Xu, D., Wu, Y., Yuan, S., Zhang, L., and Wu, X. Achieving causal fairness through generative adversarial networks. In *International Joint Conference on Artificial Intelligence*, pp. 1452–1458, 2019.

Zafar, M. B., Valera, I., Rogriguez, M. G., and Gummadi, K. P. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, pp. 962–970. PMLR, 2017.

Zhang, C., Tan, K. C., Li, H., and Hong, G. S. A cost-sensitive deep belief network for imbalanced classification. *IEEE transactions on neural networks and learning systems*, 30(1):109–122, 2018.

A. Proof of Theorem 1

Proof. Problem (4) can be written as

$$\min_{w} \mathbf{L}(\boldsymbol{w}) = \sum_{k=1}^{n_c} \mathbf{L}(\boldsymbol{w}_k)$$

$$\sum_{k=1}^{n_c} \mathbf{L}(\boldsymbol{w}_k) = \sum_{k=1}^{n_c} \Big[-\sum_{i=1}^{n_k}\sum_{j=1}^{n} w_{kij} l_{kij} + \alpha \Big(\sum_{i=1}^{n_k}\sum_{j=1}^{n} |w_{kij}| \Big)^2 \Big], s.t. \sum_{i=1}^{n_k}\sum_{j=1}^{n} w_{kij} = c, w_k \geq 0. \tag{4}$$

We can handle the n_c groups in Problem (4) separately. The optimization problem *w.r.t.* the k-th group can be formulated as follows:

$$\min_{\mathbf{u}} -\mathbf{l}^T\mathbf{u} + \mathbf{u}^T\mathbf{u}, \;\; s.t. \;\; \mathbf{u} \geq \mathbf{0}, \mathbf{u}^T\mathbf{1} = c, \tag{5}$$

where $\mathbf{u} = [w_{k11}, w_{k12}, \ldots, w_{kn_k1}, w_{k21}, \ldots, w_{kn_kn}]$. The Lagrangian function of Problem (5) is

$$\min_{\mathbf{u}} -\mathbf{l}^T\mathbf{u} + \mathbf{u}^T\mathbf{u} - \eta^T\mathbf{u} + \lambda(\mathbf{u}^T\mathbf{1} - c), \tag{6}$$

where $\eta \geq \mathbf{0}$ and $\lambda \geq \mathbf{0}$ are Lagrangian multipliers. Take derivative of Problem (6) *w.r.t.* $\mathbf{u}$ and set it to zero, we get

$$\eta - \lambda\mathbf{1} + \mathbf{1} = 2m\mathbf{1},$$

where $m = \mathbf{1}^T\mathbf{u}$. From the KKT condition we can derive $\eta^T\mathbf{u} = 0$. Consequently, we can derive

$$\begin{cases} w_q = 0 & \Longrightarrow \eta_q > 0 \;\; \Longrightarrow l_q - \lambda < 0\,, \\ w_q > 0 & \Longrightarrow \eta_q = 0 \;\; \Longrightarrow l_q - \lambda \geq 0\,, \end{cases} \tag{7}$$

where $q \in \{1, ..., n \times n_k\}$. Without loss of generality, suppose l is a sorted vector such that $l_1 > l_2 > \cdots > l_{n_kn}$, and suppose there is a $k' \in \{1, ..., n \times n_k\}$ that satisfies $l_{k'} \geq \lambda > l_{k'+1}$ where then according to Eq. (7) we can derive

$$\begin{cases} u_q = 0 & if \quad i > k', \\ u_q = \frac{l_i - \lambda}{2\alpha} & if \quad 1 \geq i \geq k'. \end{cases}$$

And notice that $\mathbf{u}^T\mathbf{1} = c$. we can derive the value of λ as follows:

$$\sum_{j=1}^{k'} \frac{l_i - \lambda}{2\alpha} = c \quad \Longrightarrow \quad \lambda = \frac{\sum_{j=1}^{k'} l_j - 2\alpha c}{k'}.$$

Combining (7), we know the value of k' satisfies:

$$\sum_{j=1}^{k'} l_j - k' l_{k'+1} > 2\alpha c > \sum_{j=1}^{k'} l_j - k' l_{k'}.$$

Here the optimal solution lies within feasible region when $\sum_{j=1}^{k'} l_i - 2\alpha c \geq 0$ holds true. When $2\alpha c$ is very large, all samples within the group receive non-zero weights. □

B. Proof of Theorem 2

Proof. Our proposed optimization problem can be divided into two parts:

$$\mathbf{w}^* = \max_{\mathbf{w}} \mathbf{w}^T \mathbf{l} - \alpha \left\| \mathbf{w} \right\|_2^2, \tag{8}$$

$$\mathbf{l}^* = \arg\min_{\mathbf{l}} \mathbf{w}^T \mathbf{l} - \alpha \left\| \mathbf{w} \right\|_2^2, \tag{9}$$

During k-th iteration, both $\mathbf{w}$ and $\mathbf{l}$ is updated to the optimum. Consider the baseline reweighing method, where samples within each demographic group are assigned with equal weight $\frac{c}{N}$. Denote such weight as $\mathbf{u}$, in current iteration we have

$$\mathbf{w}^{*T} \mathbf{l} - \alpha \left\| \mathbf{w}^* \right\|_2^2 \geq \mathbf{u}^T \mathbf{l} - \alpha \left\| \mathbf{u} \right\|_2^2, \tag{10}$$

Since $\mathbf{u}$ assigns equal weight per sample, it is same as calculating average loss within each group, and $\mathbf{u}^T \mathbf{l}$ can be further written as

$$\mathbf{u}^T \mathbf{l} = c(|\bar{\mathbf{l}^*_{a=0}}| + |\bar{\mathbf{l}^*_{a=1}}|), \tag{11}$$

Where $\bar{\mathbf{l}^*_{a=i}}$ represents average loss in each subgroup and c represents the sum of weight in each subgroup. In our experiment, the classifier is chosen as logistic regression, and $\mathbf{l}$ represents cross-entropy loss. As proved by (Feng et al., 2020), the cross-entropy loss $\mathbf{l}$ is lower-bounded by its corresponding mean absolute error $\mathbf{l}_{MAE}$, thus we have

$$|\bar{\mathbf{l}^*_{a=0}}| + |\bar{\mathbf{l}^*_{a=1}}| \geq |\bar{\mathbf{l}^*_{MAE,a=0}}| + |\bar{\mathbf{l}^*_{MAE,a=1}}|, \tag{12}$$

Since $l_{MAE} = |h_{pred} - y_{true}|$, denote as l_y the corresponding $0-1$ classification loss, on correctly-classified samples the following inequality always hold:

$$l_{MAE} \geq l_y, \tag{13}$$

And on wrongly classified samples:

$$l_{MAE} + 0.5 \geq l_y, \tag{14}$$

Thus (12) can be further lower-bounded by:

$$|\bar{\mathbf{l}^*_{MAE,a=0}}| + |\bar{\mathbf{l}^*_{MAE,a=1}}| \geq |\bar{\mathbf{l}^*_{y,a=0}}| + |\bar{\mathbf{l}^*_{y,a=1}}| - c' \geq |\bar{\mathbf{l}^*_{y,a=0}} - \bar{\mathbf{l}^*_{y,a=1}}| - c' \tag{15}$$

Where $c' = 0.5 \sum_{n=1}^{N} \mathbf{1}(y_{pred} \neq y)$. Denote as $N_{i,a=j}$ the number of samples with label i in j-th group, we can express the average $0-1$ loss within each subgroup as:

$$\bar{\mathbf{l}^*_{0,a=j}} = \frac{1}{N_{0,a=j}} \mathbf{1}(y_{pred} \neq 0) \tag{16}$$

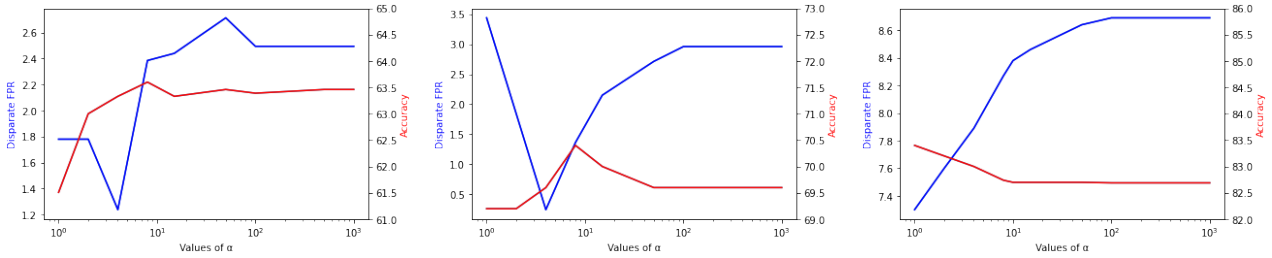
$$\bar{\mathbf{l}^*_{1,a=j}} = \frac{1}{N_{1,a=j}} \mathbf{1}(y_{pred} \neq 1) \tag{17}$$

Which corresponds exactly to the FPR and FNR of j-th group. Combining (10), (12), (15), (16), and (17), we have

$$\mathbf{w}^{*T} \mathbf{l} \geq c(|\text{FPR}_{a=0} - \text{FPR}_{a=1}| + |\text{FNR}_{a=0} - \text{FNR}_{a=1}|) - c' \tag{18}$$

$$\geq c(|\text{FPR}_{a=0} - \text{FPR}_{a=1}| + |\text{FNR}_{a=0} - \text{FNR}_{a=1}|) - 0.5N. \tag{19}$$

Which shows that our reweighed minimization problem corresponding exactly to minimizing the upper bound of equalized odds during each iteration. □



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

Figure 4. Change of disparate FPR and accuracy with varying α on three datasets. Lower disparate FPR shows better fairness. Higher accuracy shows better classification.

C. Experiments Supplement

C.1. Dataset Details

We evaluate our model on five benchmark datasets. The first three are for classification tasks, and the last two are for regression:

- **Adult**: The Adult dataset (Dua & Graff, 2017) contains 65,123 samples with 11 attributes and one binary label indicating whether an individual's annual income exceeds $50K$. The sensitive attribute in this dataset is *sex*.
- **German**: The UCI German credit risk dateset (Dua & Graff, 2017) contains 1,000 samples with 20 attributes and one binary label indicating whether a client is classified highly risky. The sensitive attribute in this dataset is *sex*.
- **COMPAS**: The ProPublica COMPAS dataset (Larson et al., 2016) contains 7,215 samples with 11 attributes. The goal is to predict whether the defendant re-offend within two years. Following the protocol in earlier fairness methods (Zafar et al., 2017), we only select white and black individuals in COMPAS dataset, which contains 6,150 samples in total. The sensitive attribute in this dataset is *race*.
- **Law School**: The Law School Admissions Councils National Longitudinal Bar Passage Study (Wightman, 1998) contains 20,649 samples with 38 attributes. The goal is to predict a student's GPA. Following the protocol in (Chzhen et al., 2020), we normalize GPA to [0,1]. The sensitive attribute in this dataset is *race*.
- **Communities & Crime (CRIME)**: The Communities and Crime dataset contains 1,994 samples with 128 attributes. The goal is to predict the number of violent crimes per 10^5 population. Following (Calders et al., 2013), we normalize the number of crimes to [0,1] and treat black population and non-black population as the sensitive groups.

C.2. Sensitivity to Hyperparameters

Here we discuss the trade-off between performance and fairness on three datasets. Figure 4 show the change of accuracy and disparate TNR (same as disparate FPR) with increasing α. As discussed in Section 3, higher α means more samples are assigned with non-zero weights and the difference of weights within groups become smaller. On COMPAS dataset and German credit dataset, as α increases, the disparate FPR first decreases, then gradually increases and finally stabilizes. On COMPAS dataset, as α increases, the classification accuracy also increases from 61.5% to 63.5%, which shows that more training samples with non-zero weights benefit the performance of classifier. However, the effect of α on accuracy on German credit dataset and Adult dataset are relatively small, and the change of accuracy is only about 0.7%.

It is worth noticing that although there is no consistent trade-off between fairness and accuracy as α changes, on all the three datasets, as α increases, the fairness discrepancy also increases, which shows α has direct control over fairness, i.e., focusing on hard samples helps the classifier to achieve fairness. However, focusing on hard samples does not always harm classification accuracy. Our assumption about accuracy is that if easy samples lie far apart from classification hyperplane, i.e., such samples are unlikely to be wrongly classified, then focusing on hard samples will also help improve accuracy. However, if hard samples are relatively close to easy samples, then strictly forcing the classifier to focus on hard samples will have a negative affect on accuracy.