Bias-Tolerant Fair Classification

Yixuan Zhang YIXUAN.ZHANG@STUDENT.UTS.EDU.AU
Feng Zhou ZHOUFENG6288@TSINGHUA.EDU.CN
Zhidong Li ZHIDONG.LI@UTS.EDU.AU
Yang Wang YANG.WANG@UTS.EDU.AU
Fang Chen FANG.CHEN@UTS.EDU.AU
Data Science Institute, University of Technology Sydney, Australia
Department of Computer Science and Technology, Tsinghua University, China

Editors: Vineeth N Balasubramanian and Ivor Tsang

Abstract

The label bias and selection bias are acknowledged as two reasons in data that will hinder the fairness of machine-learning outcomes. The label bias occurs when the labeling decision is disturbed by sensitive features, while the selection bias occurs when subjective bias exists during the data sampling. Even worse, models trained on such data can inherit or even intensify the discrimination. Most algorithmic fairness approaches perform an empirical risk minimization with predefined fairness constraints, which tends to trade-off accuracy for fairness. However, such methods would achieve the desired fairness level with the sacrifice of the benefits (receive positive outcomes) for individuals affected by the bias. Therefore, we propose a **B**ias-Tolerant **FA**ir **R**egularized **L**oss (B-FARL), which tries to regain the benefits using data affected by label bias and selection bias. B-FARL takes the biased data as input, calls a model that approximates the one trained with fair but latent data, and thus prevents discrimination without constraints required. In addition, we show the effective components by decomposing B-FARL, and we utilize the meta-learning framework for the B-FARL optimization. The experimental results on real-world datasets show that our method is empirically effective in improving fairness towards the direction of true but latent labels.

Keywords: Fairness, Loss, Label Bias, Selection Bias

1. Introduction

With the increasing adoption of autonomous decision-making systems in practice, the fairness of the outcome obtained from such systems has raised widespread concerns (Coston et al., 2019; Zafar et al., 2017a). As the decision-making systems are driven by data and models, they are vulnerable to data bias since the model can replicate the biases contained in the input data and output biased decisions (Bird et al., 2016). To address the issues, researchers proposed fairness-aware learning methods and demonstrated the potential in dealing with discrimination problems in job applicants selection (Faliagka et al., 2012), credit card approval (Khandani et al., 2010) and recidivism prediction (Brennan et al., 2009). The fairness-aware learning methods in the previous work can be categorized into (1) pre-processing methods: learn fair representations of the input data (Louizos et al., 2016; Zemel et al., 2013; Calmon et al., 2017; Lum and Johndrow, 2016); (2) in-processing

methods: incorporate fairness constraints into the objective function to achieve certain level of fairness (Zafar et al., 2017a,b; Calders et al., 2009; Agarwal et al., 2018; Kamishima et al., 2012) and (3) post-processing methods (Hardt et al., 2016): modify the learned posterior distribution of the prediction to achieve fairness. In this paper, we mainly focus on the second category, where the approaches perform an empirical risk minimization with predefined fairness constraints. These constraints, heavily dependent on predefined fairness definitions, are combined with the loss to be a fairness-aware objective function.

Model optimization based on the fairness-aware objective function creates the controversy of the trade-off between accuracy and fairness (Berk et al., 2017). The recent work of Wick et al. (2019) presented the paradox that accuracy drops due to the ignorance of *label bias* and *selection bias* when imposing fairness constraints to the model. By definition, the label bias will flip the label, e.g., from 'qualified' to 'unqualified' in recruitment data; and the selection bias will distort the ratios between the protected and unprotected group, e.g., select less positive labeled instances from the protected group. The reason that trade-off occurs is that the accuracy is still evaluated on the biased data. However, when evaluated on the bias-free data, fairness and accuracy should improve simultaneously.

In this work, inspired by the peer loss (Liu and Guo, 2020), we propose the loss function, B-FARL, that can automatically compensate both selection bias and label bias existing in input data with implicit regularizers. By minimizing the loss, the learned classifier using biased data is equivalent to the learned one using unbiased data. The peer loss is designed to handle binary label noise problems where labels are flipped randomly conditioning on the true class. It is similar to the label bias setting in our problem but has no dependence between the flip rate and sensitive features. In the design of our B-FARL, the flip rate is separately considered for distinct demographic groups (samples with different values of sensitive feature). B-FARL inherits the strength of peer loss which does not require flip rate estimation; in addition, B-FARL also does not require explicit fairness constraints or the level of fairness violation. We will show and prove that B-FARL is an appropriate loss function that guides the model to learn towards fair prediction from the biased data.

Furthermore, though peer loss does not require noise rate estimation, it requires tuning a noise rate related hyperparameter via cross validation, which is time consuming. To address this issue, we utilize the meta-learning framework. Meta-learning can learn meta-parameters (parameters to be optimized) from data directly, which is a data-driven optimization framework. Motivated by the success of hyperparameter optimization using meta-learning (Jones, 2001), we incorporate our B-FARL into the model-agnostic meta-learning (MAML) optimization framework to dynamically update the hyperparameters, which is more efficient than cross validation.

Specifically, our work makes three main contributions: **(1)** We propose the B-FARL, which enables the learning of a fair model using data containing label bias and selection bias. It is worth nothing that B-FARL does not require predefined fairness constratins but learns fairness directly from data. **(2)** We provide a theoretical analysis of the effectiveness of B-FARL by decomposing it into three indicative terms, i.e., the expected loss on the distribution of clean data, a fairness regularizer w.r.t. subgroups risk deviation, and the regularizer on the disagreement between biased and unbiased observations. **(3)** We utilize MAML framework to optimize the noise rate related hyperparameters, which is more efficient than the traditional cross validation.

2. Related Work

Fairness in machine learning Most algorithmic fairness approaches in the literature incorporate fairness constraints into the objective function (Zafar et al., 2017a,b; Calders et al., 2009; Agarwal et al., 2018; Kamishima et al., 2012) for optimization. The fairness constraints need to be predefined according to various statistical fairness criteria, such as equality opportunity (Hardt et al., 2016), equalized odds (Hardt et al., 2016) and demographic parity notion like p%-rule (Biddle, 2005). In the work of Donini et al. (2018) and Rezaei et al. (2020), they proposed to use the nonlinear measure of dependence as regularizers to approximate p%-rule or equality opportunity violations. However, the approximation could potentially hurt the performance. Besides, there are two main general drawbacks to these methods. First, the fairness criteria must be carefully chosen. Second, if the constraints can grant a fair model, testing it on the biased data will hurt the accuracy. This creates the controversy of the trade-off between accuracy and fairness. The recent work of Wick et al. (2019) analyzed the second drawback by a framework that considered label bias and selection bias. Under the bias setting, deploying fairness constraints directly to the biased data can both hurt the accuracy and fairness. To address the issue, we propose to incorporate algorithmic fairness by the label noise framework that can handle biased data learning. The most similar work is Wang et al. (2021). However, this work is fundamentally different from ours w.r.t. the problem to be solved. Their problem is how to derive fairness constraints on corrupted data in the label noise problem, while we solve the fairness problem by considering the label bias and selection bias as a special type of label noise.

Noisy label learning Most recent works of learning from noisy labels focus on modifying the loss function, which include loss correction and reweighting methods (Scott et al., 2013; Natarajan et al., 2013; Liu and Tao, 2016; Patrini et al., 2017). However, these methods require estimating the noise rate or cannot handle asymmetric noise rates. The recent work of Liu and Guo (2020) proposed a peer loss function based on the idea of peer prediction to solve label noise problems under the asymmetric noise setting. The peer loss function is defined as subtracting the loss of random sampled feature-label pair from the loss of each sample. This method does not require noise rate estimation and enables us to perform empirical risk minimization on corrupted data. The loss proposed in our work is related to the CORES^2 (COnfidence REgularized Sample Sieve) (Cheng et al., 2021) that improves the performance of peer loss by taking the expectation of the robust cross-entropy loss over the random sample pairs, encouraging a more confident prediction. This work inspires us to propose the B-FARL to solve the discrimination problem from a label bias perspective. However, this work does not in an end-to-end manner, it separates the learning process into two phases: select most clean samples in the first phase and treats the rest samples as unlabeled and retrain the model in the second phase.

3. Proposed Method

In this section, we will present our design for B-FARL. We begin with a detailed problem formulation. Next, we introduce the methodology of B-FARL followed by the analysis of B-FARL. At last, we provide the algorithm for optimizing B-FARL.

3.1. Problem Formulation

Given the triplet of random variables (X, Z, A) with sample space $\Omega = \mathcal{X} \times \{-1, 1\} \times \{0, 1\}$, X denotes the non-sensitive feature, Z denotes the clean and fair label and A is the binary sensitive feature. Let $f : X \to Z$ be a fair labeling function, which maps X to a fair and clean outcome Z. To obtain observations, we can use an observation distribution D to generate samples for the triplet. When the generative process is independent of A, we name D clean and fair distribution since the data will be fair. However, in our problem, we assume D and the generated data are latent because of discrimination. In the framework proposed by Wick et al. (2019), we can decompose the discrimination as label bias and selection bias. So, instead of observing samples from the true distribution D, we assume one can only observe samples from a corrupted distribution $\widetilde{D}$, where the labels from $\widetilde{D}$ are discriminated by sensitive feature A. We denote the discriminated label as Y and we assume Z is flipped to Y with the probability conditioning on A, i.e., $\theta_a^{\mathrm{sgn}(y)} = P(Y = y \mid Z = -y, A = a)$ in the binary classification setting. We summarize the process of labels being discriminated in Fig. 1. We also assume A is independent of X. Such a setting separates the discrimination from features and lets all the sources of discrimination be in A.

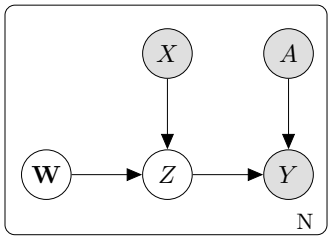

Figure 1: Generative process of bias in N observations, shaded nodes are observations, adapted from Wick et al. (2019)

The label bias is from biased decisions on the sensitive feature, e.g., gender or race. Label bias can cause the function f learned from (X, Y, A) being discriminated. On the other side, different from label bias, selection bias will affect the true ratio of two demographic groups in favor of positive outcome ($Z = 1$), and affect the data distribution D in further. We assume the selection bias occurs in the process of selecting samples from positive labeled instances among the protected group and we denote the selection bias as $\frac{r}{\sigma}$, where r is the original proportion of positive labeled instances among the protected group and $\sigma = 1$ if no selection bias occurs while $\sigma > 1$ if selection bias occurs. The selected data is denoted as $\hat{D}$ which is a subset of D. Our aim is to learn a labeling function $\hat{f}$ under the corrupted distribution $\widetilde{D}$ that can approximate the fair labeling function f and hence enable the prediction toward fairness. We propose to use noisy label learning methods to solve this problem. Some of these techniques, such as the re-weighting (Natarajan et al., 2013; Liu and Tao, 2016) or loss correction (Patrini et al., 2017) methods, require θ to be known, or they cannot handle asymmetric noise rates. To be more robust, we will eliminate such a requirement by addressing it with peer loss (Liu and Guo, 2020).

A noticeable challenge of the solution is that only label bias is convertible to the label noise, while selection bias and the combined bias cannot be directly fit into it. With the assumption that the selection bias occurs in the process of selecting positive labeled instances among the protected group, it will affect θ_0^-. Let ε_0^- denote the bias rate combining the selection bias and label bias to represent the proportion that how many data among protected group labeled as $+$ are finally observed as $-$. The relationship between ε_0^- and θ_0^- can be derived as $\theta_0^- = \frac{\sigma - r}{1 - r}\varepsilon_0^- + \frac{1 - \sigma}{1 - r}$. The full derivation can be found in the Appendix B.

3.2. B-FARL

In this section, we present our design for B-FARL based on peer loss. For each sample (x_i, y_i), the peer loss (Liu and Guo, 2020) for i is defined as

$$\ell_{peer} = \ell(f(x_i, \boldsymbol{\omega}), y_i) - \alpha \cdot \ell(f(x_{i_1}, \boldsymbol{\omega}), y_{i_2}), \tag{1}$$

where α is used as the parameter to make peer loss robust to unbalanced labels, and computed as

$$\alpha := 1 - (1 - P(Y = -1 \mid Z = +1) - P(Y = +1 \mid Z = -1)) \frac{P(Z = +1) - P(Z = -1)}{P(Y = +1) - P(Y = -1)}. \tag{2}$$

In other words, when $P(Z = +1) = P(Z = -1) = 0.5$, α is 1. In practice, α can be tuned as a hyperparameter (Liu and Guo, 2020), which means we do not require to know $P(Z = +1)$ and $P(Z = -1)$ for computing α. In Eq. (1), i_1, i_2 are independently sampled from $S/\{i\}$ ($S = \{1, 2, \ldots, N\}$) by $\frac{1}{N}$. The corresponding random variables with sensitive attribute are the triplet of $(X_{i_1}, A_{i_1}, Y_{i_2})$. If we take demographic groups into consideration, the original peer loss is re-weighted by a factor δ_a. Similar to Wang et al. (2021), it is defined as $\delta_a = \frac{1}{1-\theta_a^+-\theta_a^-}$ and hence the group-weighted peer loss for i is

$$\ell_{gp} = \delta_{a_i} \cdot \ell_{peer}. \tag{3}$$

According to Liu and Guo (2020), δ_a used to re-scale peer loss on biased data to clean data. Then we will show how B-FARL is designed by decomposing ℓ_{gp} for the protected and unprotected groups. First, we take the expectation of ℓ_{gp} w.r.t. X_{i_1} and Y_{i_2} over distribution conditioning on A as Eq. (4). There are two other reasons to take the expectation form: (1) the expectation form enables us to write the loss in terms of x_i rather than the random variable X_{i_1}, which provides convenience for computing. (2) instead of randomly sampled pairs, we use the expectation to keep the loss stable.

$$\begin{aligned}
&\frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{X_{i_1},Y_{i_2}|\widetilde{D}}[\delta_{a_i}(\ell(f(x_i,\boldsymbol{\omega}),y_i) - \alpha\cdot\ell(f(X_{i_1},\boldsymbol{\omega}),Y_{i_2}))]\\
&= \frac{1}{N}\sum_{i}^{N}\delta_{a_i}[\ell(f(x_i,\boldsymbol{\omega}),y_i) - \alpha\cdot P(A=0\mid\widetilde{D})\sum_{i'\in S_0}P(X_{i_1}=x_{i'}\mid A=0,\widetilde{D})\mathbb{E}_{Y|\widetilde{D},A=0}\ell(f(x_{i'},\boldsymbol{\omega}),Y)\\
&- \alpha\cdot P(A=1\mid\widetilde{D})\sum_{i'\in S_1}P(X_{i_1}=x_{i'}\mid A=1,\widetilde{D})\mathbb{E}_{Y|\widetilde{D},A=1}\ell(f(x_{i'},\boldsymbol{\omega}),Y)]\\
&= \frac{1}{N}\sum_{i}^{N}\delta_{a_i}[\ell(f(x_i,\boldsymbol{\omega}),y_i) - \alpha\cdot\frac{|S_0|}{N}\sum_{i'\in S_0}\frac{1}{|S_0|}\mathbb{E}_{Y|\widetilde{D},A=0}\ell(f(x_{i'},\boldsymbol{\omega}),Y)\\
&- \alpha\cdot\frac{|S_1|}{N}\sum_{i'\in S_1}\frac{1}{|S_1|}\mathbb{E}_{Y|\widetilde{D},A=1}\ell(f(x_{i'},\boldsymbol{\omega}),Y)]\\
&= \frac{1}{N}(\sum_{i\in S_0}\delta_{a_i}[\ell(f(x_i,\boldsymbol{\omega}),y_i) - \alpha\cdot\mathbb{E}_{Y|\widetilde{D},A=0}\ell(f(x_i,\boldsymbol{\omega}),Y)]\\
&+ \sum_{i\in S_1}\delta_{a_i}[\ell(f(x_i,\boldsymbol{\omega}),y_i) - \alpha\cdot\mathbb{E}_{Y|\widetilde{D},A=1}\ell(f(x_i,\boldsymbol{\omega}),Y)]),
\end{aligned} \tag{4}$$

where $S_0 = \{i|a_i = 0\}$ and $S_1 = \{i|a_i = 1\}$. Based on Eq. (4), we add intensity parameter to obtain the framework of B-FARL (L_F) as

$$L_F = \frac{1}{N}\sum_{i=1}^{N}(\ell_B(\boldsymbol{\omega}) + \boldsymbol{\beta}\boldsymbol{\ell}_A(\boldsymbol{\omega})), \tag{5}$$

with

$$\begin{aligned} \ell_B(\boldsymbol{\omega}) &= \delta_{a_i}\ell(f(x_i,\boldsymbol{\omega}),y_i), \quad \boldsymbol{\beta} = \begin{bmatrix}-\beta_0\\-\beta_1\end{bmatrix}^T, \\ \boldsymbol{\ell}_A(\boldsymbol{\omega}) &= \begin{bmatrix}\mathbb{E}_{Y|\widetilde{D},A=0}(1-a_i)\ell(f(x_i,\boldsymbol{w}),Y)\\ \mathbb{E}_{Y|\widetilde{D},A=1}a_i\ell(f(x_i,\boldsymbol{w}),Y)\end{bmatrix}, \end{aligned} \tag{6}$$

where β_0, β_1 are two hyperparameters that control the intensity of the regularizer terms ($\boldsymbol{\ell}_A$). We let δ_{a_i} and α in Eq. (4) be absorbed into β_0 and β_1. Most widely used surrogate loss functions can be used for ℓ. For example, 0-1 loss can be applied with sufficient training data (Bartlett et al., 2006) for its robustness to instance-dependent noise (Manwani and Sastry, 2013) but alternatives also can be applied such as cross entropy, logistic loss, etc. Compared to the peer loss, the two expectation regularization terms conditioning on the protected and non-protected groups can further improve the prediction performance. In section 3.3, we will show how the regularization terms help improve the performance.

3.3. Analysis of the B-FARL

In this section, we explain the effectiveness of Eq. (5) by decomposing it into components that demonstrate fairness regularization and discrimination correction. The full derivation can be found in Appendix A. B-FARL can be decomposed into the following three terms:

$$\begin{aligned} &\mathbb{E}_{\widetilde{D}}[\ell_B(\boldsymbol{\omega}) + \boldsymbol{\beta}\boldsymbol{\ell}_A(\boldsymbol{\omega})] \\ =&\underbrace{\mathbb{E}_D[\ell(f(X),Z)]}_{\text{clean model}} + \underbrace{\lambda\cdot[\mathbb{E}_{\widetilde{D}|A=0}\ell(f(X),Y) - \mathbb{E}_{\widetilde{D}|A=1}\ell(f(X),Y)]}_{\text{fairness regularization}} \\ &+\underbrace{\sum_a P(A=a)\sum_{k\in\{+1,-1\}}\sum_{l\in\{+1,-1\}}P(Z=l)\mathbb{E}_{D_{x|l,a}}(\delta_a\theta_a^{\text{sgn}(k)} - \gamma_a\cdot P(Y=k))\ell(f(x),k).}_{\text{bias regularization}} \end{aligned} \tag{7}$$

The first term is for learning with clean data. The second term shows the fairness regularization w.r.t. subgroup risks deviation which is defined in Def. 1 (Without loss of generality, we assume $\mathbb{E}_{\widetilde{D}|A=0}\ell(f(X),Y) > \mathbb{E}_{\widetilde{D}|A=1}\ell(f(X),Y)$). The last term shows the regularization effect on the biased data. Here both the regularizer effects λ in the second term and γ_a in the last term are decomposed from β_0 and β_1 in Eq. (5).

Definition 1 (Perfect fairness via subgroup risks) *We say that a predictor $f \in \mathcal{F}$ is perfectly fair w.r.t. a loss function ℓ if all subgroups attain the same average loss; i.e., in the binary sensitive attributes case (Sec. 3.2 in Williamson and Menon (2019)),*

$$\mathbb{E}_{X,Y|A=0}\ell(f(X),Y) = \mathbb{E}_{X,Y|A=1}\ell(f(X),Y). \tag{8}$$

More specifically:

- The first term is the expected loss on the distribution of clean samples.
- The second term is a fairness regularizer on the noisy distribution w.r.t. the subgroup risk measure on the noisy distribution. As explained in Williamson and Menon (2019), Def. 1 tells us under the perfect fairness, the prediction performance w.r.t. the sensitive attributes should not vary. The best case for the regularizer is perfect fairness according to Def. 1. We use the difference between average subgroup risk to measure the fairness violation and λ is the regularizer effect.
- The third term is a regularizer w.r.t. noisy loss. This loss is the penalty for the disagreement between Y and Z. The ideal situation is that $\delta_a \theta_a^{\mathrm{sgn}(k)} - \gamma_a \cdot P(Y = k)$ should be minimized, and hence the noisy term will vanish. We should point out that the selection bias is included in $\theta_1^- = \frac{\sigma - r}{1 - r}\varepsilon_1^- + \frac{1-\sigma}{1-r}$ and if $\sigma = 1$, $\theta_1^- = \varepsilon_1^-$.
- For equivalence, it is noticeable when the first term is minimized, $f(X)$ is the Bayes optimal classifier on clean data, which means the penalties of all bias do not exist. As a result, on the optimal point, all three terms are minimized so that the summation is also minimized. Therefore, classifier that can minimize the B-FARL equals classifier that can minimize the first term, which indicates the equivalence.
- The effectiveness of the first and second terms are similar to traditional loss function with fairness constraints. However, here the loss function is learned from Z while the traditional methods still use Y. Such difference endues our loss the capability to learn the correct model.

3.4. Optimization B-FARL via Model-Agnostic Meta-Learning

Meta-learning is a general framework of data-driven optimization. Most of the meta-learning methods can be viewed as a bi-level optimization which contains inner loop optimization (main optimization) and outer loop optimization (optimize the meta-parameter, e.g. hyperparmeters of inner loop). In our work, we consider the B-FARL as the main optimization goal and the re-weighting factor δ_{a_i} and regularization parameters $\boldsymbol{\beta}$ as the meta-parameters. Since δ_{a_i} for individuals among the same demographic group is the same, we can also write the first part in Eq. (5) as the following format

$$\frac{1}{N}\sum_{i=1}^{N}\ell_B(\boldsymbol{\omega}) = \frac{1}{N}[\alpha_0 \sum_{i\in\{S_0\}}\ell(f(x_i,\boldsymbol{\omega}),y_i) + \alpha_1 \sum_{i\in\{S_1\}}\ell(f(x_i,\boldsymbol{\omega}),y_i)] = \frac{1}{N}\boldsymbol{\alpha}\boldsymbol{\ell}_{D_a}, \tag{9}$$

where $\boldsymbol{\alpha} = \begin{bmatrix}\alpha_0 \\ \alpha_1\end{bmatrix}^T$ and $\boldsymbol{\ell}_{B_a} = [\sum_{i\in\{S_0\}}\ell(f(x_i,\boldsymbol{\omega}),y_i), \sum_{i\in\{S_1\}}\ell(f(x_i,\boldsymbol{\omega}),y_i)]$. Overall, the optimization can be viewed as

$$\min_{\boldsymbol{\alpha},\boldsymbol{\beta}} L_F(\boldsymbol{\omega_p}), \boldsymbol{\omega_p} = \arg\min_{\boldsymbol{\omega}} L_F(\boldsymbol{\omega}). \tag{10}$$

We split the optimization into two stages and here we define $\boldsymbol{\omega}^t$, $\boldsymbol{\beta}^t$ and $\boldsymbol{\alpha}^t$ as the corresponding variables in step t. In the meta training stage, we first initialize $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, to obtain $\boldsymbol{\omega}^1$, then fix $\boldsymbol{\omega}^1$ to obtain $\boldsymbol{\beta}^1$ and $\boldsymbol{\alpha}^1$. These two steps iteratively used to obtain $\boldsymbol{\omega}^{t+1}$, $\boldsymbol{\beta}^{t+1}$ and $\boldsymbol{\alpha}^{t+1}$. In the actual training stage, we optimize B-FARL with the updated $\boldsymbol{\beta}^{t+1}$ and $\boldsymbol{\alpha}^{t+1}$ from meta training stage. The detailed steps are summarized in Algorithm 1.

3.4.1. Meta training stage

We randomly split the training set into mini-batches with batch size n. With fixed values of $\boldsymbol{\beta}^{t+1}$ and $\boldsymbol{\alpha}^{t+1}$, we first perform the inner loop optimization and the one-step-forward weights $\boldsymbol{\omega}^{t+1}$ is updated by gradient descent with learning rate η

$$\boldsymbol{\omega}^{t+1} = \boldsymbol{\omega}^t - \eta \nabla_{\boldsymbol{\omega}^t} \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{\alpha}^t \boldsymbol{\ell}_{B_a}(\boldsymbol{\omega}^t) + \boldsymbol{\beta}^t \ell_A(\boldsymbol{\omega}^t)) \tag{11}$$

Now with updated $\boldsymbol{\omega}^{t+1}$, we then perform the outer loop optimization which updates $\boldsymbol{\beta}^{t+1}$ and $\boldsymbol{\alpha}^{t+1}$ via gradient descent with learning rate η'

$$\begin{aligned} \boldsymbol{\beta}^{t+1} &= \boldsymbol{\beta}^t - \eta' \nabla_{\boldsymbol{\beta}^t} \frac{1}{n} \sum_{i=1}^{m} (\boldsymbol{\alpha}^t \boldsymbol{\ell}_{B_a}(\boldsymbol{\omega}^{t+1}) + \boldsymbol{\beta}^t \ell_A(\boldsymbol{\omega}^{t+1})), \\ \boldsymbol{\alpha}^{t+1} &= \boldsymbol{\alpha}^t - \eta' \nabla_{\boldsymbol{\alpha}^t} \frac{1}{n} \sum_{i=1}^{m} (\boldsymbol{\alpha}^t \boldsymbol{\ell}_{B_a}(\boldsymbol{\omega}^{t+1}) + \boldsymbol{\beta}^t \ell_A(\boldsymbol{\omega}^{t+1})). \end{aligned} \tag{12}$$

3.4.2. Actual training stage

We should point out that in the meta training stage, $\boldsymbol{\omega}$ is the auxiliary as the purpose of meta training stage is to determine the optimal value for $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$. Once we have updated $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$, we train the model ($\boldsymbol{\omega}$ in the actual training stage) via gradient descent with learning rate γ

$$\boldsymbol{\omega}^{t+1} = \boldsymbol{\omega}^t - \gamma \nabla_{\boldsymbol{\omega}^t} \frac{1}{n} \sum_{i=1}^{n} (\boldsymbol{\alpha}^{t+1} \boldsymbol{\ell}_{B_a}(\boldsymbol{\omega}^t) + \boldsymbol{\beta}^{t+1} \ell_A(\boldsymbol{\omega}^t)). \tag{13}$$

Algorithm 1 Optimization for B-FARL

Initialize the hyperparameter $\boldsymbol{\beta}$ and $\boldsymbol{\alpha}$ and model weights $\boldsymbol{\omega}$
for *t=1,⋯ T* **do**
  Update the model parameter $\boldsymbol{\omega}^{t+1}$ by Eq. (11)
  Update $\boldsymbol{\beta}^{t+1}$ and $\boldsymbol{\alpha}^{t+1}$ by Eq. (12)
  Train model with $\boldsymbol{\beta}^{t+1}$ and $\boldsymbol{\alpha}^{t+1}$ by Eq. (13)
end
Obtain the prediction results

4. Experiments and Comparisons

In this section, we conduct experiments on real world data to investigate the effects of label bias and selection bias that affect accuracy and fairness and show the effectiveness of our

proposed method. Since we cannot observe the latent fair labels of the real-world data, we assume the observed data is clean and add different biases to create a biased version.

4.1. Experiment Setup

In this section, we introduce our experiment setting including the evaluation metrics and dataset descriptions.

4.1.1. Evaluation Metrics

We use two metrics: Difference of Equal Opportunity (DEO) (Hardt et al., 2016) and p%-rule (Biddle, 2005) to measure fairness violation . They are defined as

$$\text{DEO} = |P(\hat{Y} = 1 \mid A = 1, Y = 1) - P(\hat{Y} = 1 \mid A = 0, Y = 1)|,$$
$$\text{p\%} = \min\left(\frac{P(\hat{Y} = 1 \mid A = 0)}{P(\hat{Y} = 1 \mid A = 1)}, \frac{P(\hat{Y} = 1 \mid A = 1)}{P(\hat{Y} = 1 \mid A = 0)}\right).$$

A higher DEO and smaller p% indicate more fairness violation. These two indicators evaluate fairness from a different perspective. DEO considers the additional condition with the original label is positive, and p%-rule only considers the prediction results. Their combination can avoid the case that classifier pushes the results to demographic parity but neglect the true labels. In our experiment, we implement a simple Multi-Layer Perceptron (MLP) to train, and we applied binary cross-entropy loss for ℓ in Eq. (5). We use the weighted macro F1 score to measure the performance, which is the macro average weighted by the relative portion of samples within different classes. We split the data into 90% train and 10% test, and we report the results in the form of mean $\pm$ standard deviation over ten experiments with ten random splits.

4.1.2. Dataset Description

Adult Dataset[1]**:** The target value is whether an individual's annual income is over \$50k. The original feature dimension for this dataset is 13. After feature aggregation and feature encoding , the feature dimension is 35. The sensitive attribute is 'Gender', and we consider 'Gender = Female' as protected group.
German Credit Dataset[2]**:** The task of this dataset is to classify people as good or poor credit risks. The features include economical situation of each individual as well as personal information like age, gender, personal status, etc. The feature dimension is 13. In our experiment, we set 'Gender' as sensitive attribute and 'Gender = Male' as protected group.
Compas Dataset[3]**:** This data is from COMPAS, which is a tool used by judges, probation and prole officers to asses the risk of a criminal to re-offend. We focus on the predictions of 'Risk of Recidivism' (Arrest). The algorithm was found to be biased in favor of white defendants over a two-year follow-up period. We consider 'Race' to be the sensitive attribute and 'Race=Black' as protected group. After feature encoding and aggregation, the feature dimension is 11.

1. http://archive.ics.uci.edu/ml/datasets/Adult
2. https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)
3. www.propublica.org/article/how-we-analyzed-the-compasrecidivism-algorithm

dataset	number of instances	protected/unprotected groups	number of instances	Source
Adult	30,717	female/male	10,067/20,650	UCI
German Credit	900	female/male	278/622	UCI
Compas	6,492	black/white	3,325/3,167	COMPAS

Table 1: Dataset description

4.1.3. Baseline Models

From the perspective of fairness constraints, we compare to two recent fairness-aware learning methods: Rezaei et al. (2020); Donini et al. (2018); From the perspective of label bias, we compare to two related noisy label learning methods: CORES^2 (Cheng et al., 2021); Group Peer Loss (GPL) (Wang et al., 2021) as our baseline comparison. Besides, we also compare to two baseline methods: Clean and Biased, in which we train MLP on the clean data and biased data respectively.

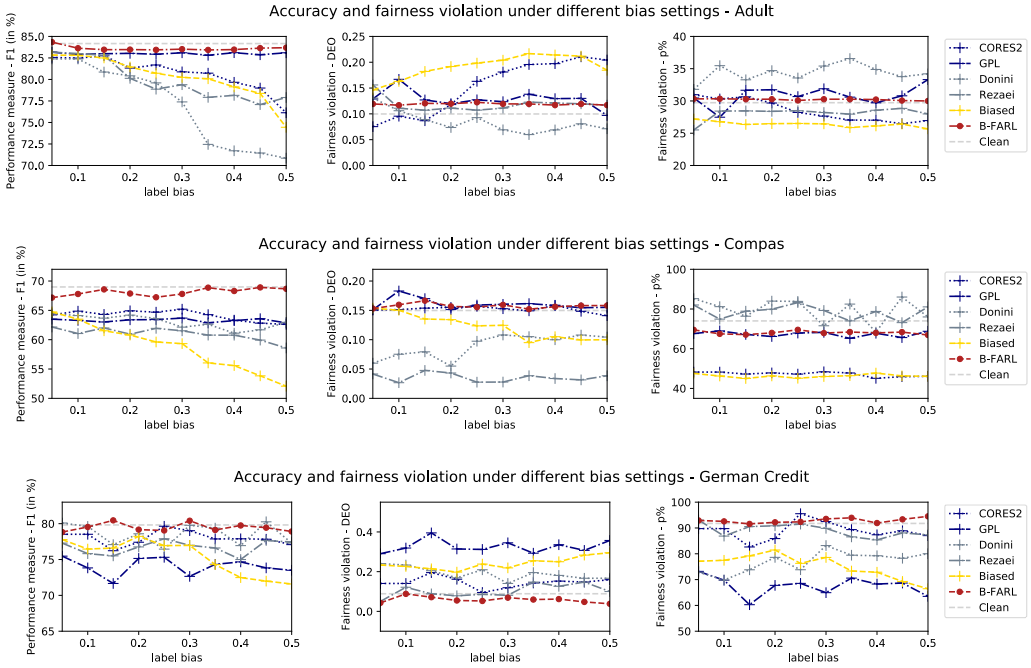

Figure 2: Accuracy and fairness violation under different label bias settings. The x-axis is the average label bias over $\{\theta_0^+, \theta_0^-, \theta_1^+, \theta_1^-\}$. We use same color to denote the methods in the same category, i.e., we use blue color to denote GPL and CORES^2, which are both noisy label learning method, and we use gray color to denote two algorithmic fairness methods.

For the efficiency, the runtime of GPL is around 20.51 minutes. B-FARL only needs 0.83 minutes. CORES^2 needs 2.32 minutes for two phases together. The incorporation of

the meta-learning framework is much more efficient, the time complexity for B-FARL is $\mathcal{O}(Td^2)$, where T is the number of iterations and d represents feature dimensions.

4.2. Comparison and Application on Real Word Data

4.2.1. Case 1: Label bias

In the first case, we test the performance of different methods under different settings of label bias with selection bias fixed. We set average label bias amount from 0.1 to 0.5 while fix the selection bias with $\sigma = 1.1$. We add bias into the train set only while keep test set clean. In the settings, we always require $\theta_0^+ > \theta_0^-$ and $\theta_1^- > \theta_1^+$.

The results are shown in Figure. 2. The prediction performance of our method generally outperforms other methods with the increase of label bias. Overall, the two algorithmic fairness methods have lower F1 scores than the two noisy label learning methods and B-FARL, though they have lower fairness violations. This demonstrates the algorithmic fairness methods will achieve a certain fairness level by "flipping" the labels of some individuals, and the low F1 indicates the flipping is in the opposite direction of the true labels. This is what we have claimed the controversy of accuracy and fairness trade-off. Also, we notice that the F1 score of two algorithmic fairness methods decreases while the fairness violation increases as the amount of label bias increases, which indicates they are not robust to the different amount of label bais; In the meantime, two noisy label learning methods, as well as B-FARL, have more steady F1 when we increase the amount of label bias. However, since CORES^2 does not take fairness into consideration, it has an overall higher fairness violation compared to GPL and B-FARL. GPL deploys derived fairness constraints under corrupted distribution, so it has overall lower fairness violation compared to CORES^2, but higher than B-FARL.

For the adult dataset, we found the results for GPL are very close to ours while GPL has a slightly higher p% value and DEO, and ours has higher accuracy and lower DEO. For the Compas dataset, the accuracy of our method is closest to the accuracy on the clean data and achieves closer p% to the benchmark for clean distribution. For the German Credit dataset, B-FARL has the highest f1, with the highest p% and lowest DEO. Overall, B-FARL is superior to the other baseline methods for optimizing towards the latent fair labels under different label bias amounts.

4.2.2. Case 2: Selection Bias

In this section, we conduct our experiments on how selection bias would affect performance and fairness violation. We fixed the label bias which we set as $\theta_0^+ = 0.25$, $\theta_0^- = 0.05$, $\theta_1^+ = 0.05$ and $\theta_1^- = 0.25$. We increase the selection bias by 2% from $\sigma = 1.01$ to $\sigma = 1.1$. Similar to the setting in Sec 4.2.1, we add selection bias to train set only.

From Fig. 3 we can see B-FARL also outperforms among all the methods with the highest F1 and low fairness violations. Unlike the experimental results of label bias, we do not observe an apparent decreasing trend as selection bias increases. However, the difference between our method and other methods are distinct. And our performance is the closest to the clean one. Also, we found GPL cannot handle selection bias very well compared to its performance under label bias. For the Adult dataset, B-FARL has the highest F1 and lowest fairness violation w.r.t. both DEO and p% measure and is close to the baseline on

clean data. The F1 score of two algorithmic fairness methods and two noisy label learning methods are close. For the Compas and German Credit dataset, B-FARL has the highest F1 score. Two algorithmic fairness methods have the highest p% value. Still, the method proposed by Donini et al. (2018) has a higher DEO violation and higher F1 than the method proposed by Rezaei et al. (2020). In contrast, the method proposed by Rezaei et al. (2020) has the lowest F1 and lowest DEO violation. This demonstrates the same phenomenon we have concluded in Sec 4.2.1. Similar to the experiment of label bias, the two noisy label learning methods have higher F1 and higher fairness violations compared to the two algorithmic fairness methods. B-FARL has the highest F1 and lowest fairness violation compared to all the methods. Overall, B-FARL is superior to the other baseline methods also under different amounts of selection bias.

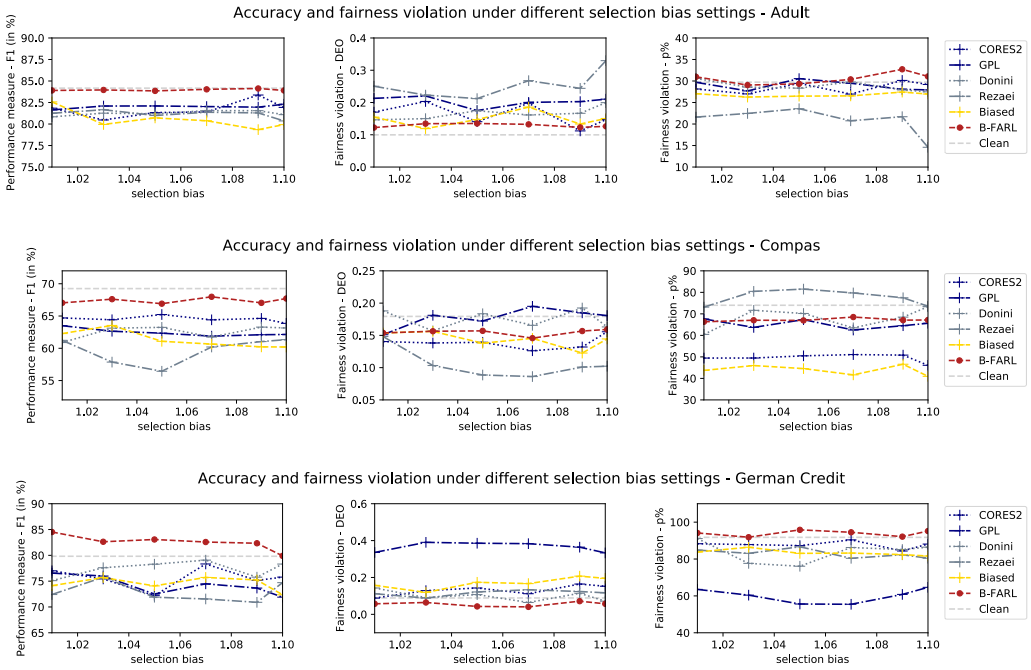

Figure 3: Accuracy and fairness violation under different selection bias settings. The x-axis is the average selection bias which is related to the proportion of positive labeled instances among the protected group. The blue color is for GPL and CORES^2, which are both noisy label learning method. The gray is for two algorithmic fairness methods.

4.3. Evaluate Our Methods on the Clean Data

We also evaluate our method on the clean data directly. We simulated ten sets of clean data according to Fig. 1. The detailed generation steps are provided in Appendix C. We found our method can achieve similar accuracy and fairness level to the baseline on the clean data. Though GPL has the highest F1 score, it also has the highest fairness violations,

this may imply GPL over-corrects the labels. In contrast, Rezaei et al. (2020) has the smallest fairness violations but with lowest F1 score, this was aligned with the results in Section 4.2.1. We found both CORES^2 and Donini et al. (2018) have accuracy and fairness drop, the former may due to the nonlinear measure of fairness constraints, which has the adverse impact of both performance and fairness, the latter may caused by the second phase of sample sieve, which introduce randomness for the semi-supervised learning.

	Clean	B-FARL	Donini	Rezaei	CORES^2	GPL
F1	98.52±1.28%	98.51±1.60%	98.22±0.85%	94.93±0.89%	94.86±2.23%	98.95±0.66%
DEO	0.62±0.61%	0.71±0.73%	0.79±0.34%	0.46±0.40%	0.87±0.47%	1.06±0.77%
p%	95.10±4.14%	95.39±3.80%	94.26±3.06%	95.88±4.32%	95.05±4.13%	94.41±4.47%

Table 2: Performance on the clean datasets

4.4. Impact of Regularization Intensity

We also examine how the regularization intensity $\boldsymbol{\beta}$ works by conducting the experiment on the 'Compas' dataset. We record the F1 score and p% value when increasingly update $\boldsymbol{\beta}$. We compute $||\boldsymbol{\beta}||$ to measure the intensity. We can see from Fig. 4, when the regularization intensity increases from around 0.2 to 0.95, the performance and p% value also increases. This demonstrates that when B-FARL is guided by appropriate regularization intensity, the accuracy and fairness improve simultaneously. However, as the intensity gets larger, we can see the p% value still increases, but the F1 score starts to decrease. This indicates that the fairness regularizer term starts to dominate as the intensity becomes larger and hence causes the results to achieve perfect fairness while neglecting the accuracy performance. However, with appropriate regularization intensity, the accuracy performance and fairness improve together.

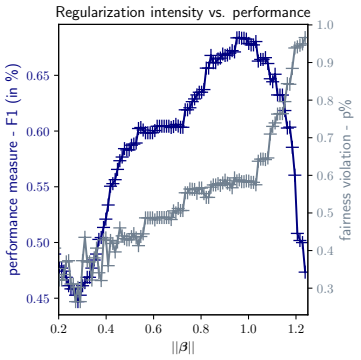

Figure 4: Regularization intensity vs. performance on Compas dataset. The x-axis is the norm of $\boldsymbol{\beta}$, the left y-axis is the F1 score (blue line) and the right y-axis is fairness measure w.r.t. p% (gray line).

5. Conclusion

In this paper, we tackle the discrimination issue from the label bias and selection bias perspective. We propose a bias-tolerant fair classification method by designing B-FARL, which is a loss having the regularization effect that can compensate both label bias and selection bias. To optimize B-FARL more efficiently, we incorporate it with the model-agnostic meta-learning framework to update the hyperparameters. Besides, We decompose the B-FARL loss into three meaningful components, including expected loss under the distribution of clean samples, fairness regularizer, and a regularizer on the disagreement between biased

and unbiased observations to demonstrate the effectiveness of B-FARL theoretically. We empirically demonstrated the superiority of our proposed framework through experiments. The possible future research directions of this work including: relax the assumption that X is independent of A for more complex data since in practice X will always contain the information from A. This can also be connected with instance-dependent label bias setting since we do not only consider the flip rate related to the true label and A, but rather include the dependency with X; combine with explainability techniques (Mary, 2019) to explore how B-FARL influence the decisions for bias correction; extend to multi-class classification tasks and enable the bias setting with multiple sensitive attributes.

References

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna M. Wallach. A reductions approach to fair classification. *CoRR*, abs/1803.02453, 2018.

Peter L Bartlett, Michael I Jordan, and Jon D McAuliffe. Convexity, classification, and risk bounds. *Journal of the American Statistical Association*, 101(473):138–156, 2006.

Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. A convex framework for fair regression, 2017.

Dan Biddle. Adverse impact and test validation: A practitioner's guide to valid and defensible employment testing. 2005.

Sarah Bird, Solon Barocas, Kate Crawford, and Hanna Wallach. Exploring or exploiting? social and ethical implications of autonomous experimentation in ai. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT-ML), New York University*, page 4, October 2016.

Tim Brennan, William Dieterich, and Beate Ehret. Evaluating the predictive validity of the compas risk and needs assessment system. *Criminal Justice and Behavior*, 36(1):21–40, 2009.

T. Calders, F. Kamiran, and M. Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pages 13–18, Dec 2009.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 3992–4001. Curran Associates, Inc., 2017.

Hao Cheng, Zhaowei Zhu, Xingyu Li, Yifei Gong, Xing Sun, and Yang Liu. Learning with instance-dependent label noise: A sample sieve approach. In *International Conference on Learning Representations*, 2021.

Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R. Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. Fair transfer learning with missing protected attributes. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, page 91–98, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450363242.

Michele Donini, Luca Oneto, Shai Ben-David, John Shawe-Taylor, and Massimiliano Pontil. Empirical risk minimization under fairness constraints. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, page 2796–2806, Red Hook, NY, USA, 2018. Curran Associates Inc.

Evanthia Faliagka, Kostas Ramantas, Athanasios Tsakalidis, and Giannis Tzimas. Application of machine learning algorithms to an online recruitment system. 01 2012.

Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *CoRR*, abs/1610.02413, 2016.

D. R. Jones. A taxonomy of global optimization methods based on response surfaces. *Journal of Global Optimization*, 21:345–383, 2001.

Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In Peter A. Flach, Tijl De Bie, and Nello Cristianini, editors, *Machine Learning and Knowledge Discovery in Databases*, pages 35–50, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg. ISBN 978-3-642-33486-3.

Amir E. Khandani, Adlar J. Kim, and Andrew W. Lo. Consumer credit-risk models via machine-learning algorithms. *Journal of Banking & Finance*, 34(11):2767 – 2787, 2010. ISSN 0378-4266.

T. Liu and D. Tao. Classification with noisy labels by importance reweighting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(3):447–461, March 2016. ISSN 0162-8828.

Yang Liu and Hongyi Guo. Peer loss functions: Learning from noisy labels without knowing noise rates. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 6226–6236. PMLR, 2020.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S. Zemel. The variational fair autoencoder. In *ICLR*, 2016.

Kristian Lum and James Johndrow. A statistical framework for fair predictive algorithms. *arXiv e-prints*, art. arXiv:1610.08077, Oct 2016.

N. Manwani and P. S. Sastry. Noise tolerance under risk minimization. *IEEE Transactions on Cybernetics*, 43(3):1146–1151, Jun 2013. ISSN 2168-2275.

Sherin Mary. *Explainable Artificial Intelligence Applications in NLP, Biomedical, and Malware Classification: A Literature Review*, pages 1269–1292. 07 2019. ISBN 978-3-030-22867-5. doi: 10.1007/978-3-030-22868-2_90.

Nagarajan Natarajan, Inderjit S Dhillon, Pradeep K Ravikumar, and Ambuj Tewari. Learning with noisy labels. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 1196–1204. Curran Associates, Inc., 2013.

Giorgio Patrini, Alessandro Rozza, Aditya Krishna Menon, Richard Nock, and Lizhen Qu. Making deep neural networks robust to label noise: A loss correction approach. *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2233–2241, 2017.

Ashkan Rezaei, Rizal Fathony, Omid Memarrast, and Brian Ziebart. Fairness for robust log loss classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34 (04):5511–5518, Apr 2020. ISSN 2159-5399.

Clayton Scott, Gilles Blanchard, and Gregory Handy. Classification with asymmetric label noise: Consistency and maximal denoising. In Shai Shalev-Shwartz and Ingo Steinwart, editors, *Proceedings of the 26th Annual Conference on Learning Theory*, volume 30 of *Proceedings of Machine Learning Research*, pages 489–511, Princeton, NJ, USA, 12–14 Jun 2013. PMLR.

Jialu Wang, Yang Liu, and Caleb Levy. Fair classification with group-dependent label noise. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 526–536, New York, NY, USA, 2021. Association for Computing Machinery. ISBN 9781450383097.

Michael Wick, swetasudha panda, and Jean-Baptiste Tristan. Unlocking fairness: A trade-off revisited. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 32, pages 8783–8792. Curran Associates, Inc., 2019.

Robert Williamson and Aditya Menon. Fairness risk measures. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 6786–6797, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

M. Zafar, I. Valera, M. Gomez-Rodriguez, and K. Gummadi. Fairness constraints: Mechanisms for fair classification. In *AISTATS*, 2017a.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1171–1180, Republic and Canton of Geneva, CHE, 2017b. International World Wide Web Conferences Steering Committee. ISBN 9781450349130.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.