Mitigating Label Bias in Machine Learning: Fairness through Confident Learning

Yixuan Zhang[1], **Boyu Li**[2], **Zenan Ling**[3], **Feng Zhou**[4*]

[1]China-Austria Belt and Road Joint Laboratory on AI and AM, Hangzhou Dianzi University, China
[2]Data Science Institute, University of Technology Sydney, Australia
[3]School of Electronic Information and Communications, Huazhong University of Science and Technology, China
[4]Center for Applied Statistics and School of Statistics, Renmin University of China, China
yixuan.zhang@hdu.edu.cn, boyu.li@student.uts.edu.au, lingzenan@hust.edu.cn, feng.zhou@ruc.edu.cn

Abstract

Discrimination can occur when the underlying unbiased labels are overwritten by an agent with potential bias, resulting in biased datasets that unfairly harm specific groups and cause classifiers to inherit these biases. In this paper, we demonstrate that despite only having access to the biased labels, it is possible to eliminate bias by filtering the fairest instances within the framework of confident learning. In the context of confident learning, low self-confidence usually indicates potential label errors; however, this is not always the case. Instances, particularly those from underrepresented groups, might exhibit low confidence scores for reasons other than labeling errors. To address this limitation, our approach employs truncation of the confidence score and extends the confidence interval of the probabilistic threshold. Additionally, we incorporate with co-teaching paradigm for providing a more robust and reliable selection of fair instances and effectively mitigating the adverse effects of biased labels. Through extensive experimentation and evaluation of various datasets, we demonstrate the efficacy of our approach in promoting fairness and reducing the impact of label bias in machine learning models.

Introduction

In recent decades, we have observed a shift towards an AI-driven society, where machine learning techniques have profoundly affected various aspects of our lives, including finance (Khandani, Kim, and Lo 2010), recruitment (Faliagka et al. 2012) and law (Dressel and Farid 2018). When algorithmic decisions have a significant impact on our lives, it is crucial for decision-makers or regulators to have confidence in the algorithm's performance. While deep neural networks possess the capability to learn complex patterns from input data, they encounter a fundamental challenge in the data they learn from: the labels can be influenced by sensitive information, causing the neural networks to capture and reproduce these undesirable associations. Therefore, to ensure fairness in decision-making, it becomes essential to mitigate the impact of such undesirable relationships.

Research on training fair machine learning models has then received a lot of attention. To achieve fairness, one can incorporate fairness constraints into the learning objectives (Bilal Zafar et al. 2015; Cotter, Jiang, and Sridharan 2019; Donini et al. 2018; Rezaei et al. 2020; Roh et al. 2020), or modify the model's predictions using threshold adjustments and calibration to align them with fairness constraints (Hardt, Price, and Srebro 2016; Kim, Ghorbani, and Zou 2019; Lohia et al. 2019; Petersen et al. 2021), or use data manipulation or representation learning methods before model training (Calmon et al. 2017; Choi et al. 2020; Jiang and Nachum 2020; Kamiran and Calders 2012). Nonetheless, the aforementioned methods primarily focus on modifying the machine learning model to tackle the bias problem. There has been limited effort, as indicated by Jiang and Nachum (2020), in directly addressing the biased data itself, despite the fact that often the training data itself exhibits biased features and corresponding labels.

To combat the influence of label bias, numerous approaches have been proposed. For instance, Jiang and Nachum (2020) addressed the biased data problem by formulating the label bias as a constrained optimization framework and using a re-weighting method to learn an equivalent unbiased labeling function. On the other hand, Wang, Liu, and Levy (2021) established fairness constraints by linking ground truth distribution and observed biased distribution. In this paper, we present another effective method that focuses on data selection. Intuitively, if we can select examples with labels less influenced by sensitive information, the network learned from such data will be more robust, leading to a fairer decision-making process.

The previous works aimed at mitigating label bias within the data itself using confidence scores generally encompass two phases (Northcutt, Jiang, and Chuang 2021; Cordeiro et al. 2023). In the first phase, these approaches obtain confidence scores from the trained model and subsequently eliminate erroneous instances through a probabilistic threshold. The second phase involves model retraining using the remaining clean examples from the previous step to enhance model robustness. Nevertheless, when utilizing probability thresholds to determine the true labels, such methods heavily rely on self-confidence, of which instances with low self-confidence scores are often considered to contain label errors. However, this assumption does not always hold, particularly in scenarios of imbalanced data within the protected group. In these cases, individuals often face disadvantages

*Corresponding author.

and receive negative labels, causing deep networks to assign relatively low confidence scores to such instances due to their underrepresented nature (Yang et al. 2020).

To tackle the challenges mentioned earlier, we delve into the uncertainty of the sample selection process to combat biased labels. To alleviate the uncertainty linked with low-confidence examples, we suggest expanding the confidence intervals of the probabilistic threshold using a truncation function. Unlike previous methods that use the average confidence score to determine the probabilistic threshold, our approach reduces the impact of examples with extremely high confidence scores on the threshold. This adjustment facilitates the selection of examples with relatively lower confidence scores, which are often due to their association with disadvantaged groups, resulting in underrepresentation. Additionally, to further improve robustness, we employ the co-teaching paradigm, which involves training different models from subsets of data with varying demographic information. This allows us to cross-validate the selected fair instances from different demographic groups, enhancing the overall robustness and fairness of the approach. Our contributions are four-fold:

- We introduce a data selection method that leverages confidence scores to tackle the issue of label bias. Notably, this approach is model-agnostic, thereby enabling its application with a wide range of models. This stands in contrast to many fairness-aware learning techniques aimed at addressing biased labels, which tend to be tightly integrated with specific models and the training process.
- We present an extension of confidence intervals using a robust mean estimator, aimed at minimizing uncertainty in the process of data filtering.
- We integrate the concept of co-teaching to enhance the robustness and reduce uncertainty by cross-validating selected fair instances originating from distinct demographic groups.
- We assess the effectiveness of our approach across a range of benchmark datasets. The results highlight the benefits of our method in comparison to alternative baseline techniques.

Preliminaries

In this section, we briefly review the method addressing the biased label problem based on confidence measures. Let X be the input variable, Y be the observed output variable and Z be the true labels. Consider a k-class classification problem, we denote the observed dataset as $(\mathbf{x}, y)^N \in (\mathbb{R}^d, [k])^N$, where d is the dimensionality of the input space and N is the number of examples. In the context of confident learning, an instance with low self-confidence (as stated in Definition 1) indicates a higher likelihood of being a label error (Northcutt, Jiang, and Chuang 2021). Unlike traditional confidence measures that focus on model predictions, leveraging this assumption, confident learning aims to identify examples with label errors by directly estimating the joint distribution between corrupted labels and true labels with a class-conditional noise process. To achieve this, the whole framework of confident learning combines principles of pruning noisy data, using probabilistic thresholds to estimate noise, and ranking examples to train with confidence.

Definition 1 (Self-Confidence)**.** *The predicted probability for some model θ that an instance $\mathbf{x}$ belongs to its given label $y = i$, i.e., $p(y = i; \mathbf{x}, \theta)$.*

The confident learning process typically consists of three steps: estimation, pruning, and re-training. During the estimation process, we follow four steps to estimate the joint distribution between corrupted labels and true labels. Firstly, we compute the predicted probability of the n-th sample belonging to j, denoted as $\hat{p}_n^j = P(y_n = j; \mathbf{x}_n, \theta)$, where j is the observed label and $j \in [k]$. We then use $t_j = \frac{1}{|\mathbf{X}_{y=j}|} \sum_{\mathbf{x}_n \in \mathbf{X}_{y=j}} \hat{p}_n^j$ for the n-th sample with observed label j as the probabilistic threshold for class j. Next, for each n-th sample, we determine its true label z_n is $\arg\max_j \hat{p}_n^j$ and we have $\hat{p}_n^j > t_j$. Upon obtaining z_n, we keep a count of the label information in the count matrix $C_{y,z}$. For example, if $C_{y=1,z=0} = 40$, it means there are 40 examples that were labeled as 1 but should have been labeled as 0. To partition and count label errors, we then introduce the confident joint $\bar{C}_{y=j,z=i}$, which is given by:

$$\bar{C}_{y=j,z=i} = \frac{C_{y=j,z=i}}{\sum_{i \in [k]} C_{y=j,z=i}} \times |\mathbf{X}_{y=j}|. \tag{1}$$

Using $\bar{C}_{y=j,z=i}$ instead of $C_{y=j,z=i}$ offers the advantage of mitigating the sensitivity arising from class imbalance and distribution heterogeneity. This is achieved because $\bar{C}_{y=j,z=i}$ employs per-class thresholding as a form of calibration (Northcutt, Jiang, and Chuang 2021). Then the joint distribution of y and z can be estimated using the confident joint, and the formula is expressed as:

$$Q_{y=j,z=i} = \frac{\bar{C}_{y=j,z=i}}{\sum_{j \in [k], i \in [k]} \bar{C}_{y=j,z=i}}. \tag{2}$$

The numerator ensures that the sum of $Q_{y=j,z=i}$ over all possible values of i is calibrated to match the observed marginals for all j in $[k]$. This calibration ensures that the row sums align with the observed distributions. On the other hand, the denominator calibrates the sum of $Q_{y=j,z=i}$ to be equal to 1, ensuring that the overall distribution is calibrated to sum up to 1. The estimations of $\bar{C}_{y=j,z=i}$ and $Q_{y=j,z=i}$ are primarily aimed at facilitating the subsequent pruning of corrupted data. As the dataset size grows larger, this estimation method gradually approximates the true distribution more accurately (Northcutt, Jiang, and Chuang 2021). In other words, as we gather more data, the estimates become more reliable and better reflect the underlying relationships between the observed labels and the true labels, which enhances the effectiveness of identifying and handling corrupted data points.

In the pruning step, the objective is to identify and filter out erroneous examples. This can be achieved by various pruning approaches. For instance, we can directly remove the sets of examples that constitute the count in the off-diagonals (where true labels and observed labels are different) of $\bar{C}_{y=j,z=i}$ and use the rest for training. Or we can employ $N \cdot Q_{y=j,z=i}$ to estimate the number of examples with label errors, which is denoted as $\tilde{N}$, and then select top $\tilde{N}$ examples based on the rank of predicted probability. Once the erroneous examples are filtered out, the class weights of the remaining examples can be readjusted, and the model can be retrained.

Methodology

In this section, we introduce our design for the data selection framework to eliminate label bias. We begin with a mathematical expression of the notions of label bias, and we then present the formulation of the data selection function.

Notion of Label Bias

Let $\mathcal{D}$ represent the true underlying distribution defined by the variables (X, S, Z), and $\tilde{\mathcal{D}}$ represent the observed corrupted distribution defined by (X, S, Y). Here, X refers to the non-protected attributes, S represents the sensitive attributes, and $Z = 1$ indicates the desirable outcome (typically considered the positive class). The privileged group is identified as $S = A$, while the disadvantaged group is labeled as $S = B$. In the subsequent sections, we denote the observed subset with membership $S = A$ as D_A, and the observed subset with membership $S = B$ as D_B. The entire observed dataset is denoted as D. We assume there exists a function $\mathcal{G}$ that involves flipping the labels of certain subsets of D based solely on the values of S and Z. Accordingly, the process of generating biased labels is represented as $Y = \mathcal{G}(X, S, Z)$. In line with previous work, we consider the scenario where negative examples in the privileged group might have been labeled as positive, while positive examples in the disadvantaged group could have been labeled as negative (Wick, Panda, and Tristan 2019; Dai 2020). With this context, we introduce the following definitions:

$$\rho_A = P(Y = 1 \mid S = A, Z = 0),$$
$$\rho_B = P(Y = 0 \mid S = B, Z = 1).$$

The above expressions demonstrate that negative instances from the privileged group A are subject to flipping with a probability of ρ_A, while positive instances from the disadvantaged group B are flipped with a probability of ρ_B. In Blum and Stangl (2020), the assumption is made that $\rho_A = 0$, meaning no flipping occurs for negative instances in the privileged group. On the other hand, in Fogliato, Chouldechova, and G'Sell (2020), they set $\rho_B = 0$, implying no flipping for positive instances in the disadvantaged group. In Wick, Panda, and Tristan (2019), flipping occurs for both groups, and the flip rate is symmetric, i.e., $\rho_A = \rho_B \neq 0$.

Selecting Unbiased Labels

To improve the selection of unbiased examples, inspired by Han et al. (2018), we train two models sharing the same architecture, denoted as θ_A and θ_B, on separate subsets of the dataset: D_A and D_B, respectively. Afterward, we evaluate each model on the entire dataset D and obtain the confidence measure w.r.t. both θ_A and θ_B. Following the definition of demographic parity, that predictions should be independent of sensitive attributes, we consider a model to be unbiased if its predictive performance, when trained on either D_A or D_B, is the same from the results obtained on the entire dataset D. Then, to find the instances with label bias, we use the examples that lie in the off-diagonals of the confident joint in Eq. (1). We denote the confident joint measured on θ_A as $\bar{C}_A$ and the one measured on θ_B as $\bar{C}_B$. $\bar{C}_A$ and $\bar{C}_B$ are constructed by the comparison of t_j^A and t_j^B where

$$t_j^A = \frac{1}{|\mathbf{X}_{y=j}|} \sum_{\mathbf{x}_n \in \mathbf{X}_{y=j}} \hat{p}(y_n = j; \mathbf{x}_n, \theta_A),$$
$$t_j^B = \frac{1}{|\mathbf{X}_{y=j}|} \sum_{\mathbf{x}_n \in \mathbf{X}_{y=j}} \hat{p}(y_n = j; \mathbf{x}_n, \theta_B). \quad (3)$$

Unlike conventional methods that identify biased instances assuming true labels are determined by:

$$z_n = \arg\max_{j \in [k]} \hat{p}(y_n = j; \mathbf{x}_n, \theta),$$

we adopt the assumption that the true label is determined by:

$$z_n^A = \arg\max_{j \in [k]: \hat{p}(y_n = j; \mathbf{x}_n, \theta_A) \geq t_j^A} \hat{p}(y_n = j; \mathbf{x}_n, \theta_A),$$
$$z_n^B = \arg\max_{j \in [k]: \hat{p}(y_n = j; \mathbf{x}_n, \theta_B) \geq t_j^B} \hat{p}(y_n = j; \mathbf{x}_n, \theta_B).$$

Consider the following example: if D_A is primarily composed of positively labeled instances, while the majority of instances in D_B are labeled negative, then for a given instance with observed positive label that measured on θ_A, it is likely to have a higher $\hat{p}_n$ due to θ_A being overly confident about the positive class, resulting in a proportionally larger t_j^A, and vice versa. Therefore, by setting the probabilistic threshold as outlined in Eq. (3), we can mitigate the class imbalance issue.

In addition to addressing the class imbalance problem, we account for the uncertainty associated with the probabilistic threshold, which arises from using the average of confidence scores to select instances belonging to a specific class.

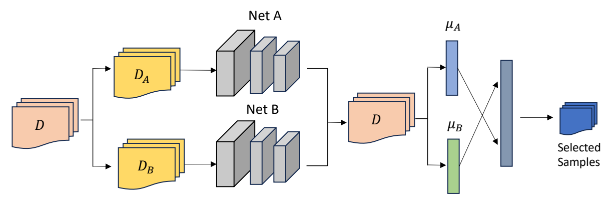

Figure 1: Our method involves training two identical models on separate datasets, D_A and D_B, and then evaluates the entire dataset D to derive a probabilistic threshold, which guides the selection of the fairest instances. Though our approach is motivated by co-teaching, the training procedure diverges from co-teaching. During training, co-teaching evaluates the selected samples to help identify the most challenging instances. In contrast, we evaluate the entire dataset to validate that the selected samples are not affected by label bias.

Such selection criteria can introduce biases, particularly for instances within the disadvantaged group. These instances may exhibit relatively lower confidence scores that are both beneath t_j^A and t_j^B, resulting in their exclusion from the selection process. However, these instances may be correctly labeled and could be valuable for generalization. To provide these instances with a chance, following the approach in Xia et al. (2022) and Chen et al. (2020), we extend a classical M-estimator (Catoni 2011) to truncate t_j^A and t_j^B as follows:

$$\begin{aligned}\bar{t}_j^A &= \frac{1}{|\mathbf{X}_{y=j}|}\sum_{\mathbf{x}_n\in\mathbf{X}_{y=j}}\psi(\hat{p}(y_n=j;\mathbf{x}_n,\theta_A)),\\ \bar{t}_j^B &= \frac{1}{|\mathbf{X}_{y=j}|}\sum_{\mathbf{x}_n\in\mathbf{X}_{y=j}}\psi(\hat{p}(y_n=j;\mathbf{x}_n,\theta_B)),\end{aligned}\tag{4}$$

where $\psi:\mathbb{R}\to\mathbb{R}$ is a non-decreasing influence function, such that the widest possible choice is $\psi(x)=\log(1+x+\frac{|x|^2}{2})$ for $x\geq 0$. The selection of ψ is motivated by the Taylor expansion of the exponential function, aiming to enhance the robustness of estimation results by mitigating the impact of extreme values. As a result, instances with extremely high confidence scores will have a limited effect on the probabilistic threshold, creating an opportunity for examples with relatively lower confidence scores that could still be correctly labeled to be included in the selection process. To achieve this, we derive the concentration inequalities for instances with low confidence scores based on Theorem 1. The derivation of Theorem 1 follows the standard process in Xia et al. (2022) and we include the proof in the Appendix for completeness.

Theorem 1. *Consider an observation set $X_N = x_1,\cdots,x_n$ with mean μ and variance ν. We utilize the non-decreasing influence function $\psi(x)=\log(1+x+\frac{x^2}{2})$. For any given $\epsilon>0$, with a probability of at least $1-2\epsilon$, we have the following inequality:*

$$\left|\frac{1}{N}\sum_{n=1}^{N}\psi(x_n)-\mu\right|\leq\frac{\nu(N+\frac{\nu\log(\epsilon^{-1})}{N^2})}{N-\nu}.$$

Let N_s^A and N_s^B represent the number of instances selected with respect to $\bar{C}_A$ and $\bar{C}_B$, respectively, where N_s^A and N_s^B are both less than or equal to the total number of instances N. Let's consider $\epsilon=\frac{1}{2N}$. Now, the threshold can be expressed as follows:

$$\begin{aligned}\mu_j^A &= \bar{t}_j^A-\frac{Q}{N_s^A-\nu},\\ \mu_j^B &= \bar{t}_j^B-\frac{Q}{N_s^B-\nu},\end{aligned}\tag{5}$$

where $Q=\nu(N+\frac{\nu\log(2N)}{N^2})$. These inequalities help us understand the behavior of the estimation when dealing with less certain or less reliable instances.

Training and Optimization

The overall training process of the proposed approach is illustrated in Fig. 1 and Algorithm 1. Initially, two networks, denoted as θ_A and θ_B, and sharing identical architectures, are established. Unlike the original co-teaching framework proposed in the previous literature (Han et al. 2018), θ_A and θ_B do not share the selected subset to update parameters; instead, they jointly evaluate the same dataset to identify biased data and enhance fairness. During each iteration, a mini-batch D^t is drawn from the dataset D. This batch is subsequently divided into two separate batches based on the grouping factor S, yielding D_A^t and D_B^t. We independently train θ_A and θ_B using D_A^t and D_B^t, respectively. Following this, D^t is evaluated using θ_A and θ_B to obtain μ_j^A and μ_j^B, calculated via the selection criteria specified in Eq. (5). Through comparison with the acquired probabilistic threshold, instances residing in the off-diagonal elements of the joint count are identified. These instances are denoted as $\tilde{D}_A^t$ and $\tilde{D}_B^t$. The union set encompassing $\tilde{D}_A^t$ and $\tilde{D}_B^t$ is then removed, yielding the selected examples for model training. In experiments, we set $N_s^A=N_s^B=N_s$. To find the optimal values for N_s and ν, hyperparameter tuning on the validation set is employed.

Algorithm 1: Training Algorithm

Input: training set D, model θ_A, θ_B and θ, epoch T, hyperparameter N_s and ν, loss function ℓ, learning rate η.
for $t=0,\cdots,T$ **do**
    Obtain mini-batch D^t from D and split it into D_A^t and D_B^t.
    $\theta_A\leftarrow\theta_A-\eta\nabla\ell(D_A^t;\theta_B)$.
    $\theta_B\leftarrow\theta_B-\eta\nabla\ell(D_B^t;\theta_B)$.
    Obtain μ_j^A and μ_j^B using Eq. (5)
    if $\hat{p}(y_n=j;\mathbf{x}_n,\theta_A)\geq\mu_j^A$ **then**
        $\hat{z}_n\leftarrow\arg\max_{j\in[k]}\hat{p}(y_n=j;\mathbf{x}_n,\theta_A)$
    else if $\hat{p}(y_n=j;\mathbf{x}_n,\theta_B)\geq\mu_j^B$ **then**
        $\hat{z}_n\leftarrow\arg\max_{j\in[k]}\hat{p}(y_n=j;\mathbf{x}_n,\theta_B)$
    end if
    Compute the count joint using Eq. (1).
    Fetch the example sets $\tilde{D}_A^t$, $\tilde{D}_B^t$ in the off-diagonal of the confident joint $\bar{C}_A$ and $\bar{C}_B$.
    $D_S^t\leftarrow D^t\backslash(\tilde{D}_A^t\cup\tilde{D}_B^t)$
    $\theta\leftarrow\theta-\eta\nabla\ell(D_S^t;\theta)$
end for

Experiments

In this section, we demonstrate the effectiveness of our methods by comparing them with several baseline models on the benchmark datasets.

Datasets

We use the synthetic data for verification and four sets of real-world data for comparison.

Synthetic We use the same setting for synthetic data generation as described in the work of Bilal Zafar et al. (2016). We generate 95,750 fair examples with 2-dimensional nonsensitive attribute space and a 1-dimensional sensitive attribute space.

Adult (Kohavi 1996) The objective of this dataset is to predict whether a person's income exceeds $50k per year. We consider two demographic groups based on gender.

ProPublica COMPAS (Brennan, Dieterich, and Ehret 2009) This dataset contains information about criminal justice. The task is to predict recidivism based on various factors. We consider two demographics based on race.

Credit Loan Data (Yeh 2016) The dataset comprises credit card default records for 30,000 applicants from April to September 2005. We consider gender as the target demographic information.

Law School Admissions (Wightman 1998) The objective of this dataset is to predict whether or not a student will pass the bar. We form two demographic groups based on gender and use the pass bar as the ground-truth label.

Dataset	N	$\|S_A\|$	$\|S_B\|$
Synthetic	95,750	72,025	23,725
Adult	46,032	31,113	14,919
COMPAS	7,214	3,518	3,696
Credit Loan Data	30,000	11,888	18,112
Law	20,798	17,491	3,307

Table 1: Dataset Description

Baselines

For all the methods, we construct a simple neural network using ReLU activation functions. Our method is evaluated against several baselines, including **confident learning (CL)** (Northcutt, Jiang, and Chuang 2021), **LongReMix** (Cordeiro et al. 2023), **label bias correction (LC)** (Jiang and Nachum 2020), and the **group peer loss (GPL)** method as described in Wang, Liu, and Levy (2021). Consistent hyperparameters are maintained across all experiments for all methods. Additional implementation details can be found in the Appendix.

Fairness Violation

We assess our performance on various datasets and methods with respect to diverse fairness metrics, which encompass (1) the demographic parity distance metric, (2) the difference of equal opportunity (DEO), and (3) the p%.

Demographic parity distance metric (Creager et al. 2019): The definition of demographic parity is that the rate of positive predictions for $S = A$ should be equivalent to that for $S = B$. This metric is formulated as $|\mathbb{E}(\hat{Y} = 1 \mid S = A) - \mathbb{E}(\hat{Y} = 1 \mid S = B)|$.

Difference of equal opportunity (DEO) (Hardt, Price, and Srebro 2016): The concept of equal opportunity states that the true positive rates for $S = A$ should be identical to those for $S = B$. The difference can be quantified as: $|P(\hat{Y} = 1 \mid S = A, Y = 1) - P(\hat{Y} = 1 \mid S = B, Y = 1)|$.

p% (Biddle 2005): This measure closely resembles the demographic parity distance metric and can be formulated as: $\min(\frac{P(\hat{Y}=1|S=A)}{P(\hat{Y}=1|S=B)}, \frac{P(\hat{Y}=1|S=A)}{P(\hat{Y}=1|S=B)})$.

Due to page limit, we present the DEO as our fairness violation metric in the table, while the outcomes pertaining to other fairness violation metrics like the demographic parity distance metric and the p% are provided in the Appendix.

Generating Biased Labels

We address two distinct categories of label bias: (1) symmetric bias, as defined by Wick, Panda, and Tristan (2019), and (2) asymmetric bias, as outlined in Blum and Stangl (2020). For the symmetric bias case, we configure $\rho_A = \rho_B$ with values chosen from the set $\{20\%, 40\%\}$. Regarding asymmetric bias, we set $\rho_A \neq \rho_B$, where ρ_A and ρ_B are both selected from $\{20\%, 40\%\}$. To ensure robust results, we perform 10 rounds of random shuffling on the training set, while retaining 10% of the biased training examples as a validation set for hyperparameter optimization.

Mean *v.s.* Truncation

We utilize the influence function $\psi(x)$ as a truncation mechanism for the probabilistic threshold, deviating from the original averaging-based approach. This modification allows for a meaningful comparison between the two metrics. Synthetic data is employed for assessment. We label the methods using Eq. (3) as "M" with hyperparameters $N_s = 0.6$ and $\nu = 10^{-2}$ fixed and we label the method using Eq. (5) as "T". Analysis of Table 3 reveals that adopting the truncated probabilistic threshold for data selection outperforms the conventional mean estimator approach, both under symmetric and asymmetric bias. Discrepancies between "M" and "T" widen with increasing bias, leading to a 3.60% error for "M", while the increase is marginal for the "T" under symmetric bias. In asymmetric bias, "T" still outperforms "M", despite higher overall accuracy. Minimal disparity is observed in fairness violation between the two methods on synthetic data. The truncated method consistently exhibits lower fairness violations, often achieving zero violations.

Comparison Results

We present the results in Table 2 and Table 4. It is evident that our approach consistently produces fair classifiers, often achieving the lowest accuracy errors and fairness violations across all methods on the four real-world datasets. Table 2 presents the results under the symmetric bias setting, where we observe that the test errors increase when the bias magnitude is 40% compared to when it's 20%. However, our method exhibits only a marginal increase in test errors and fairness violations with an elevated bias level, indicating more stable performance. It is important to highlight that while LongReMix exhibits the lowest fairness violation (measured by DEO) on the compas dataset, its test error rate is high. This outcome stems from LongRemix predominantly predicting negative outcomes for the majority of examples, resulting in a minimal DEO. However, upon reviewing the outcomes as presented in the Appendix, it becomes evident that when evaluated using alternate fairness metrics, the fairness violations are notably high. Comparing the outcomes in Table 4, it is apparent that all the

	Adult				Compas			
	20%		40%		20%		40%	
Metric	Err.(%)(↓)	Vio.(↓)	Err.(%)(↓)	Vio.(↓)	Err.(%)(↓)	Vio.(↓)	Err.(%)(↓)	Vio.(↓)
CL	$22.61_{\pm 0.79}$	$0.17_{\pm 0.01}$	$25.13_{\pm 1.30}$	$0.24_{\pm 0.04}$	$32.69_{\pm 1.64}$	$0.20_{\pm 0.02}$	$49.75_{\pm 7.29}$	$0.19_{\pm 0.02}$
LongReMix	$19.66_{\pm 3.09}$	$0.18_{\pm 0.06}$	$21.03_{\pm 1.21}$	$0.19_{\pm 0.03}$	$40.89_{\pm 2.39}$	$\mathbf{0.09}_{\pm 0.02}$	$43.21_{\pm 1.61}$	$\mathbf{0.09}_{\pm 0.02}$
LC	$18.89_{\pm 0.37}$	$0.17_{\pm 0.02}$	$20.10_{\pm 0.11}$	$0.18_{\pm 0.03}$	$34.21_{\pm 0.42}$	$0.15_{\pm 0.03}$	$35.19_{\pm 0.23}$	$0.16_{\pm 0.01}$
GPL	$23.50_{\pm 2.40}$	$0.18_{\pm 0.05}$	$24.25_{\pm 2.36}$	$0.16_{\pm 0.02}$	$44.88_{\pm 3.18}$	$0.18_{\pm 0.03}$	$46.81_{\pm 3.01}$	$0.17_{\pm 0.02}$
Ours	$\mathbf{15.75}_{\pm 0.62}$	$\mathbf{0.12}_{\pm 0.02}$	$\mathbf{16.83}_{\pm 1.48}$	$\mathbf{0.15}_{\pm 0.02}$	$\mathbf{30.75}_{\pm 0.25}$	$0.15_{\pm 0.02}$	$\mathbf{31.86}_{\pm 1.61}$	$0.17_{\pm 0.02}$
	Law				Credit			
	20%		40%		20%		40%	
Metric	Err.(↓)	Vio.(↓)	Err.(↓)	Vio.(↓)	Err.(↓)	Vio.(↓)	Err.(↓)	Vio.(↓)
CL	$15.49_{\pm 3.56}$	$0.28_{\pm 0.05}$	$16.01_{\pm 2.15}$	$0.26_{\pm 0.03}$	$21.10_{\pm 1.02}$	$0.04_{\pm 0.02}$	$20.50_{\pm 0.03}$	$0.03_{\pm 0.01}$
LongReMix	$10.25_{\pm 1.40}$	$0.13_{\pm 0.09}$	$9.55_{\pm 0.45}$	$0.11_{\pm 0.02}$	$20.10_{\pm 0.99}$	$0.03_{\pm 0.02}$	$21.97_{\pm 1.99}$	$0.04_{\pm 0.01}$
LC	$9.54_{\pm 0.07}$	$\mathbf{0.05}_{\pm 0.01}$	$9.67_{\pm 0.05}$	$0.06_{\pm 0.01}$	$19.58_{\pm 0.27}$	$0.02_{\pm 0.01}$	$20.49_{\pm 0.11}$	$\mathbf{0.02}_{\pm 0.01}$
GPL	$10.44_{\pm 0.56}$	$0.06_{\pm 0.01}$	$10.26_{\pm 0.46}$	$\mathbf{0.05}_{\pm 0.01}$	$21.70_{\pm 1.06}$	$0.02_{\pm 0.01}$	$22.82_{\pm 1.03}$	$0.03_{\pm 0.02}$
Ours	$\mathbf{9.09}_{\pm 0.50}$	$\mathbf{0.05}_{\pm 0.01}$	$\mathbf{9.31}_{\pm 0.30}$	$0.09_{\pm 0.02}$	$\mathbf{19.13}_{\pm 0.77}$	$\mathbf{0.01}_{\pm 0.01}$	$\mathbf{19.33}_{\pm 0.68}$	$\mathbf{0.02}_{\pm 0.01}$

Table 2: Experiment Results (Symmetric Bias Scenario): Each row pertains to a specific method. The table illustrates the test errors (%) and fairness violations of Confident Learning (CL), LongReMix, Label Correction methods (LC), Group Peer Loss (GPL), and our proposed approach.

methods have better performance than the results displayed in Table 2. Despite this, our method maintains superior or comparable predictive accuracy when contrasted with other approaches, and it consistently exhibits the lowest levels of fairness violations. An interesting observation is that the results of confident learning itself do not consistently yield fair classifiers compared to label bias correction methods for fairness. This is due to its reliance on class-dependent noise, leading to higher test errors and fairness violations. The reason for this lies in the fact that confident learning does not incorporate demographic information, unlike the label bias correction methods, resulting in less effective performance.

Hyperparameter Analysis

We explore the impact of hyperparameters ν in the range of $\{10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}$ and N_s in the range of $\{0.5, 0.6, 0.7, 0.8, 0.9\}$ to examine their impact using synthetic data. The experiment's focus is on assessing the influence of these two hyperparameters. In Fig. 2, we conduct an experiment by keeping $N_s = 0.75$ fixed and varying the value of ν from 10^{-4} to 10^{-1}. Fig. 2 demonstrates that the overall test error when $\rho_A = \rho_B = 0.4$ is higher compared to the error rate when $\rho_A = \rho_B = 0.2$ as we change N_s and ν. We do not report the fairness violation since they are overall very small (close to 0) and consistently exhibit no apparent variation when we change the value of ν and N_s. Regarding N_s, we hold $\nu = 10^{-2}$ constant and change the value from 0.5 to 0.9. The graphical representation of how N_s and ν influence the probabilistic threshold is presented in Fig. 2. The plot reveals that the influence function smooths higher confidence scores. Additionally, both N_s and ν play a role in regulating the deviation from the initial value. Larger values of ν result in greater deviations from the original value, while smaller values of N_s lead to more substantial deviations from the original value.

	Symmetric			
	20%		40%	
	Err.(%)	Vio.	Err.(%)	Vio.
M	$1.51_{\pm 0.26}$	$0.01_{\pm 0.01}$	$3.60_{\pm 1.59}$	$0.02_{\pm 0.01}$
T	$\mathbf{0.77}_{\pm 0.17}$	$\mathbf{0.00}_{\pm 0.00}$	$\mathbf{0.91}_{\pm 0.37}$	$\mathbf{0.01}_{\pm 0.00}$
	Asymmetric			
	20%		40%	
	Err.(%)	Vio.	Err.(%)	Vio.
M	$0.45_{\pm 0.08}$	$0.01_{\pm 0.01}$	$0.55_{\pm 0.36}$	$0.01_{\pm 0.01}$
T	$\mathbf{0.36}_{\pm 0.06}$	$\mathbf{0.00}_{\pm 0.00}$	$\mathbf{0.38}_{\pm 0.10}$	$\mathbf{0.00}_{\pm 0.00}$

Table 3: Comparisons between the mean estimator and the truncation method under symmetric and asymmetric bias.

Related Work

In this study, we introduced a data selection framework to mitigate label bias and achieve fairness. Existing data selection techniques (Loshchilov and Hutter 2017; Kawaguchi and Lu 2019; Jiang et al. 2019; Coleman et al. 2019; Loshchilov and Hutter 2015; Killamsetty et al. 2020; Mindermann et al. 2022) successfully address biased labels based on loss, but they are not well-suited for fairness. Our approach is loss-agnostic and specifically designed for fairness. Though FairBatch (Roh et al. 2021) introduces adaptive data selection to achieve fairness, the entire framework neglects label bias. Regarding the research on miti-

	Adult				Compas			
	20%		40%		20%		40%	
Metric	Err.(%)(↓)	Vio.(↓)	Err.(%)(↓)	Vio.(↓)	Err.(%)(↓)	Vio.(↓)	Err.(%)(↓)	Vio.(↓)
CL	$19.70_{\pm0.49}$	$0.14_{\pm0.01}$	$21.48_{\pm0.76}$	$0.16_{\pm0.01}$	$33.00_{\pm0.20}$	$0.19_{\pm0.01}$	$32.08_{\pm0.74}$	$0.19_{\pm0.01}$
LongReMix	$19.86_{\pm1.33}$	$0.15_{\pm0.06}$	$18.84_{\pm0.98}$	$0.16_{\pm0.01}$	$33.35_{\pm1.01}$	$0.18_{\pm0.07}$	$42.80_{\pm0.83}$	$0.15_{\pm0.03}$
LC	$\mathbf{16.36}_{\pm0.05}$	$0.13_{\pm0.01}$	$17.53_{\pm0.12}$	$0.10_{\pm0.02}$	$37.20_{\pm0.23}$	$0.11_{\pm0.01}$	$41.16_{\pm0.20}$	$\mathbf{0.04}_{\pm0.00}$
GPL	$18.87_{\pm0.99}$	$\mathbf{0.07}_{\pm0.02}$	$18.72_{\pm0.64}$	$0.11_{\pm0.02}$	$39.28_{\pm1.10}$	$\mathbf{0.10}_{\pm0.02}$	$41.44_{\pm1.11}$	$0.08_{\pm0.01}$
Ours	$17.76_{\pm0.36}$	$\mathbf{0.07}_{\pm0.01}$	$\mathbf{17.46}_{\pm0.66}$	$\mathbf{0.09}_{\pm0.01}$	$\mathbf{30.75}_{\pm0.20}$	$0.14_{\pm0.02}$	$\mathbf{30.06}_{\pm0.17}$	$0.16_{\pm0.01}$
	Law				Credit			
	20%		40%		20%		40%	
Metric	Err.(%)(↓)	Vio.(↓)	Err.(%)(↓)	Vio.(↓)	Err.(%)(↓)	Vio.(↓)	Err.(%)(↓)	Vio.(↓)
CL	$18.12_{\pm4.16}$	$0.31_{\pm0.05}$	$22.32_{\pm6.01}$	$0.31_{\pm0.03}$	$20.43_{\pm1.06}$	$0.02_{\pm0.02}$	$20.27_{\pm0.96}$	$0.02_{\pm0.01}$
LongReMix	$9.61_{\pm0.51}$	$0.13_{\pm0.04}$	$10.15_{\pm0.80}$	$0.14_{\pm0.08}$	$21.41_{\pm0.45}$	$0.03_{\pm0.01}$	$21.46_{\pm0.40}$	$0.02_{\pm0.01}$
LC	$9.33_{\pm0.07}$	$\mathbf{0.01}_{\pm0.00}$	$9.39_{\pm0.07}$	$\mathbf{0.01}_{\pm0.00}$	$20.43_{\pm0.09}$	$0.02_{\pm0.00}$	$21.39_{\pm0.05}$	$0.02_{\pm0.00}$
GPL	$10.69_{\pm0.50}$	$0.05_{\pm0.01}$	$11.00_{\pm0.60}$	$0.06_{\pm0.01}$	$21.18_{\pm0.56}$	$0.03_{\pm0.01}$	$20.80_{\pm0.61}$	$0.02_{\pm0.01}$
Ours	$\mathbf{8.99}_{\pm0.53}$	$0.07_{\pm0.03}$	$\mathbf{9.23}_{\pm0.75}$	$0.08_{\pm0.01}$	$\mathbf{18.77}_{\pm0.21}$	$\mathbf{0.02}_{\pm0.01}$	$\mathbf{18.93}_{\pm0.39}$	$\mathbf{0.01}_{\pm0.00}$

Table 4: Experiment Results (Asymmetric Bias Scenario): Each row pertains to a specific method. The table illustrates the test errors (%) and fairness violations of Confident Learning (CL), LongReMix, Label Correction methods (LC), Group Peer Loss (GPL), and our proposed approach.

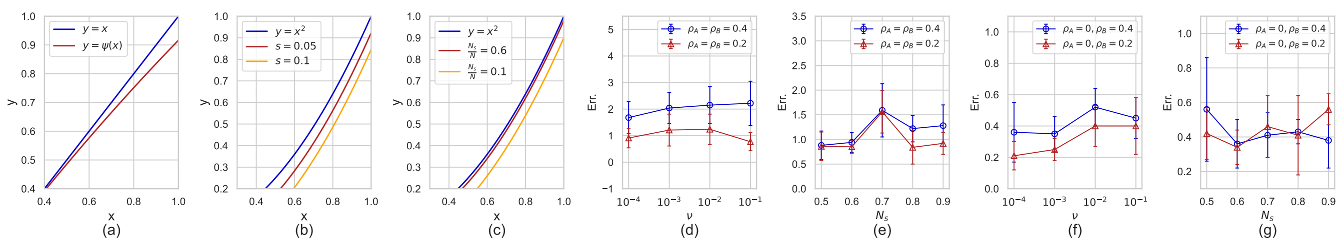

Figure 2: The illustration of the impact of $\psi(x)$: Plot (a) - (c), for (b), we set $\frac{N_s}{\nu} = 0.65$, for (c), we set $\nu = 0.5$; Influence of Hyperparameters on Probabilistic Threshold: Plot (d) - (g). Plot (d) and (e) are under the symmetric bias setting, while Plot (f) and (g) are under the asymmetric bias setting. The impact of hyperparameters ν and N_s is investigated using synthetic data.

gating biased label problems in fairness, Fogliato, Chouldechova, and G'Sell (2020) present experimental findings that even minor biases in observed labels can cast uncertainty on the reliability of analyses that depend on the biased outcomes. Blum and Stangl (2020) explored the impact of biased training data on classification and revealed the potential to recover the Bayes Optimal Classifier by combining Empirical Risk Minimization with the equal opportunity constraint under various bias models. Building on these studies, Dai (2020) introduced a comprehensive framework addressing label bias and data distribution shift for fairness. Wang, Liu, and Levy (2021) introduced bias using group-dependent label noise and proposed a surrogate loss function to handle various fairness constraints in the presence of corrupted data with biased labels. Notably, they incorporated insights from Peer loss (Liu and Guo 2020) to formulate the group peer loss, addressing group-dependent label noise. Conversely, Jiang and Nachum (2020) approached label bias through a constrained optimization framework, proposing a re-weighting method to correct instances affected by label bias. In contrast to previous approaches modifying the objective function during training, our focus is on data selection for explicit label bias mitigation.

Conclusion

This paper demonstrates that bias can be eliminated by implementing confident learning and filtering the fairest instances, even with access to biased labels. However, challenges arise with instances of low self-confidence, especially for underrepresented groups, which may still contain biased labels and remain unselected. To address this, the proposed approach utilizes the lower bound of confidence intervals, significantly enhancing the robustness and reliability of confidence score thresholding compared to traditional mean estimation methods. Leveraging interval estimation effectively mitigates the adverse effects of biased labels, promoting fairness in machine learning models. Furthermore, incorporating co-teaching enhances fair instance selection, contributing to the overall effectiveness of the framework. Extensive experimentation and evaluation on various datasets underscore the proposed method's efficacy in reducing the impact of label bias and promoting fairness.

Acknowledgments

We thank the anonymous reviewers for their constructive feedback and suggestions on how to strengthen the paper. This work was supported by NSFC Project (No. 62106121), the MOE Project of Key Research Institute of Humanities and Social Sciences (22JJD110001), the fund for building world-class universities (disciplines) of Renmin University of China (No. KYGJC2023012), China-Austria Belt and Road Joint Laboratory on Artificial Intelligence and Advanced Manufacturing of Hangzhou Dianzi University, and the Public Computing Cloud, Renmin University of China.

References

Biddle, D. 2005. Adverse Impact and Test Validation: A Practitioner's Guide to Valid and Defensible Employment Testing.

Bilal Zafar, M.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2015. Fairness Constraints: Mechanisms for Fair Classification. *arXiv e-prints*, arXiv:1507.05259.

Bilal Zafar, M.; Valera, I.; Gomez Rodriguez, M.; and Gummadi, K. P. 2016. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. *arXiv e-prints*, arXiv:1610.08452.

Blum, A.; and Stangl, K. 2020. Recovering from Biased Data: Can Fairness Constraints Improve Accuracy? *Symposium on Foundations of Responsible Computing (FORC)*, 1.

Brennan, T.; Dieterich, W.; and Ehret, B. 2009. Evaluating the Predictive Validity of the Compas Risk and Needs Assessment System. *Criminal Justice and Behavior*, 36(1): 21–40.

Calmon, F.; Wei, D.; Vinzamuri, B.; Natesan Ramamurthy, K.; and Varshney, K. R. 2017. Optimized Pre-Processing for Discrimination Prevention. In Guyon, I.; Luxburg, U. V.; Bengio, S.; Wallach, H.; Fergus, R.; Vishwanathan, S.; and Garnett, R., eds., *Advances in Neural Information Processing Systems 30*, 3992–4001. Curran Associates, Inc.

Catoni, O. 2011. Challenging the empirical mean and empirical variance: a deviation study. arXiv:1009.2048.

Chen, P.; Jin, X.; Li, X.; and Xu, L. 2020. A generalized Catoni's M-estimator under finite α-th moment assumption with $\alpha \in (1, 2)$. arXiv:2010.05008.

Choi, K.; Grover, A.; Singh, T.; Shu, R.; and Ermon, S. 2020. Fair Generative Modeling via Weak Supervision. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 1887–1898. PMLR.

Coleman, C.; Yeh, C.; Mussmann, S.; Mirzasoleiman, B.; Bailis, P.; Liang, P.; Leskovec, J.; and Zaharia, M. 2019. Selection Via Proxy: Efficient Data Selection For Deep Learning. *CoRR*, abs/1906.11829.

Cordeiro, F. R.; Sachdeva, R.; Belagiannis, V.; Reid, I.; and Carneiro, G. 2023. LongReMix: Robust learning with high confidence samples in a noisy label environment. *Pattern Recognition*, 133: 109013.

Cotter, A.; Jiang, H.; and Sridharan, K. 2019. Two-Player Games for Efficient Non-Convex Constrained Optimization. In Garivier, A.; and Kale, S., eds., *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, volume 98 of *Proceedings of Machine Learning Research*, 300–332. PMLR.

Creager, E.; Madras, D.; Jacobsen, J.; Weis, M. A.; Swersky, K.; Pitassi, T.; and Zemel, R. S. 2019. Flexibly Fair Representation Learning by Disentanglement. *CoRR*, abs/1906.02589.

Dai, J. 2020. Label Bias, Label Shift: Fair Machine Learning with Unreliable Labels.

Donini, M.; Oneto, L.; Ben-David, S.; Shawe-Taylor, J.; and Pontil, M. 2018. Empirical risk minimization under fairness constraints. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, NIPS'18, 2796–2806. Red Hook, NY, USA: Curran Associates Inc.

Dressel, J.; and Farid, H. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, 4(1): eaao5580.

Faliagka, E.; Ramantas, K.; Tsakalidis, A.; and Tzimas, G. 2012. Application of Machine Learning Algorithms to an online Recruitment System.

Fogliato, R.; Chouldechova, A.; and G'Sell, M. 2020. Fairness Evaluation in Presence of Biased Noisy Labels. In Chiappa, S.; and Calandra, R., eds., *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, 2325–2336. PMLR.

Han, B.; Yao, Q.; Yu, X.; Niu, G.; Xu, M.; Hu, W.; Tsang, I.; and Sugiyama, M. 2018. Co-teaching: Robust training of deep neural networks with extremely noisy labels. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of Opportunity in Supervised Learning. *CoRR*, abs/1610.02413.

Jiang, A. H.; Wong, D. L.; Zhou, G.; Andersen, D. G.; Dean, J.; Ganger, G. R.; Joshi, G.; Kaminsky, M.; Kozuch, M.; Lipton, Z. C.; and Pillai, P. 2019. Accelerating Deep Learning by Focusing on the Biggest Losers. *CoRR*, abs/1910.00762.

Jiang, H.; and Nachum, O. 2020. Identifying and Correcting Label Bias in Machine Learning. In Chiappa, S.; and Calandra, R., eds., *Proceedings of the Twenty Third International Conference on Artificial Intelligence and Statistics*, volume 108 of *Proceedings of Machine Learning Research*, 702–712. PMLR.

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

Kawaguchi, K.; and Lu, H. 2019. Ordered SGD: A New Stochastic Optimization Framework for Empirical Risk Minimization. In *International Conference on Artificial Intelligence and Statistics*.

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

Killamsetty, K.; Sivasubramanian, D.; Ramakrishnan, G.; of Texas at Dallas, R. I. U.; of Technology Bombay Institution One, I. I.; and Two, I. 2020. GLISTER: Generalization based Data Subset Selection for Efficient and Robust Learning. In *AAAI Conference on Artificial Intelligence*.

Kim, M. P.; Ghorbani, A.; and Zou, J. 2019. Multiaccuracy: Black-Box Post-Processing for Fairness in Classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '19, 247–254. New York, NY, USA: Association for Computing Machinery. ISBN 9781450363242.

Kohavi, R. 1996. Scaling up the Accuracy of Naive-Bayes Classifiers: A Decision-Tree Hybrid. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, 202–207. AAAI Press.

Liu, Y.; and Guo, H. 2020. 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*, 6226–6236. PMLR.

Lohia, P.; Natesan Ramamurthy, K.; Bhide, M.; Saha, D.; Kush, R.; and Puri, R. 2019. Bias Mitigation Post-processing for Individual and Group Fairness. 2847–2851.

Loshchilov, I.; and Hutter, F. 2015. Online Batch Selection for Faster Training of Neural Networks. *CoRR*, abs/1511.06343.

Loshchilov, I.; and Hutter, F. 2017. Decoupled Weight Decay Regularization. In *International Conference on Learning Representations*.

Mindermann, S.; Brauner, J. M.; Razzak, M. T.; Sharma, M.; Kirsch, A.; Xu, W.; Höltgen, B.; Gomez, A. N.; Morisot, A.; Farquhar, S.; and Gal, Y. 2022. Prioritized Training on Points that are Learnable, Worth Learning, and not yet Learnt. In Chaudhuri, K.; Jegelka, S.; Song, L.; Szepesvari, C.; Niu, G.; and Sabato, S., eds., *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 15630–15649. PMLR.

Northcutt, C.; Jiang, L.; and Chuang, I. 2021. Confident Learning: Estimating Uncertainty in Dataset Labels. *J. Artif. Int. Res.*, 70: 1373–1411.

Petersen, F.; Mukherjee, D.; Sun, Y.; and Yurochkin, M. 2021. Post-processing for Individual Fairness. In Beygelzimer, A.; Dauphin, Y.; Liang, P.; and Vaughan, J. W., eds., *Advances in Neural Information Processing Systems*.

Rezaei, A.; Fathony, R.; Memarrast, O.; and Ziebart, B. 2020. Fairness for Robust Log Loss Classification. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04): 5511–5518.

Roh, Y.; Lee, K.; Whang, S.; and Suh, C. 2020. FR-Train: A Mutual Information-Based Approach to Fair and Robust Training. In III, H. D.; and Singh, A., eds., *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, 8147–8157. PMLR.

Roh, Y.; Lee, K.; Whang, S. E.; and Suh, C. 2021. FairBatch: Batch Selection for Model Fairness. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Wang, J.; Liu, Y.; and Levy, C. 2021. Fair Classification with Group-Dependent Label Noise. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, 526–536. New York, NY, USA: Association for Computing Machinery. ISBN 9781450383097.

Wick, M.; Panda, S.; and Tristan, J.-B. 2019. *Unlocking Fairness: A Trade-off Revisited*. Red Hook, NY, USA: Curran Associates Inc.

Wightman, L. F. 1998. *LSAC national longitudinal bar passage study : executive summary, summary of the report*. LSAC research report series. Newtown, Pa: Law School Admission Council.

Xia, X.; Liu, T.; Han, B.; Gong, M.; Yu, J.; Niu, G.; and Sugiyama, M. 2022. Sample Selection with Uncertainty of Losses for Learning with Noisy Labels. In *International Conference on Learning Representations*.

Yang, K.; Qinami, K.; Fei-Fei, L.; Deng, J.; and Russakovsky, O. 2020. Towards Fairer Datasets: Filtering and Balancing the Distribution of the People Subtree in the ImageNet Hierarchy. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, 547–558. New York, NY, USA: Association for Computing Machinery. ISBN 9781450369367.

Yeh, I.-C. 2016. default of credit card clients. UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C55S3H.