Mitigating Label Bias in Machine Learning: Fairness through Confident Learning

Yixuan Zhang¹, Boyu Li², Zenan Ling³, Feng Zhou⁴*

¹China-Austria Belt and Road Joint Laboratory on AI and AM, Hangzhou Dianzi University, China
²Data Science Institute, University of Technology Sydney, Australia
³School of Electronic Information and Communications, Huazhong University of Science and Technology, China
⁴Center 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 (Khadani, Kim, and Lo 2010), recruitment (Falagka 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.
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 $(x, y)^N \in (\mathbb{R}^d, [k])^N$, where $d$ is the dimensionality of the input space and $N$ is the number of samples. 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 $\theta$ that an instance $x$ belongs to its given label $y = i$, i.e., $p(y = i | x; \theta)$.

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 $\hat{p}_n^j = P(y_n = j | x_n, \theta)$, where $j$ is the observed label and $j \in [k]$. We then use $t_j = \frac{1}{|X = j|} \sum_{x_n \in X = j} \hat{p}_n^j$ 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 $C_{y=j,z=i}$, which is given by:

$$C_{y=j,z=i} = \frac{C_{y=j,z=i}}{\sum_{i \in [k]} C_{y=j,z=i}} \times |X_y = j|.$$ (1)

Using $C_{y=j,z=i}$ 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 $C_{y=z,i}$ 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:

$$Q_{y=j,z=i} = \frac{C_{y=j,z=i}}{\sum_{j \in [k], i \in [k]} C_{y=j,z=i}}.$$ (2)

The numerator ensures that the sum of $Q_{y=j,z=i}$ over all possible values of $i$ to be equal to 1, ensuring that the overall distribution is calibrated to sum up to 1. The estimations of $C_{y=j,z=i}$ 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 approaches 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 $C_{y=j,z=i}$ 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 \( D \) represent the true underlying distribution defined by the variables \( \{X, S, Z\} \), and \( \bar{D} \) represent the observed corrupted distribution defined by \( \{X, Y\} \). 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 \( G \) 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 \( \tilde{Y} = G(X, S, Z) \). 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 \( \rho_A \), while positive instances from the disadvantaged group \( B \) are flipped with a probability of \( \rho_B \). 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 \( \theta_A \) and \( \theta_B \), 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 \( \theta_A \) and \( \theta_B \). 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 \( \theta_A \) as \( C_A \) and the one measured on \( \theta_B \) as \( C_B \). \( C_A \) and \( C_B \) are constructed by the comparison of \( t_A^j \) and \( t_B^j \) where

\[
t_A^j = \frac{1}{\sum_{x_n \in X_{y=j}} \bar{p}(y_n = j; x_n, \theta_A)}, \\
t_B^j = \frac{1}{\sum_{x_n \in X_{y=j}} \bar{p}(y_n = j; x_n, \theta_B)}.
\]

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

\[
z_n = \arg\max_{j \in [k]} \bar{p}(y_n = j; x_n, \theta),
\]

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

\[
z_A^j = \arg\max_{j \in [k]; \hat{p}(y_n = j; x_n, \theta_A) \geq t_A^j} \hat{p}(y_n = j; x_n, \theta_A), \\
z_B^j = \arg\max_{j \in [k]; \hat{p}(y_n = j; x_n, \theta_B) \geq t_B^j} \hat{p}(y_n = j; x_n, \theta_B).
\]

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 \( \theta_A \), it is likely to have a higher \( \bar{p}_n \) due to \( \theta_A \) being overly confident about the positive class, resulting in a proportionally larger \( t_A^j \), 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.
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:

\[
\begin{align*}
\tilde{t}_j^A &= \frac{1}{|X_\nu = j|} \sum_{x_n \in X_\nu = j} \psi(\hat{p}(y_n = j; x_n, \theta_A)), \\
\tilde{t}_j^B &= \frac{1}{|X_\nu = j|} \sum_{x_n \in X_\nu = j} \psi(\hat{p}(y_n = j; x_n, \theta_B)),
\end{align*}
\]  

(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 \( \psi \) 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 \( \mu \) and variance \( \nu \). 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:

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

Let \( N^A \) and \( N^B \) represent the number of instances selected with respect to \( \tilde{C}_A \) and \( \tilde{C}_B \), respectively, where \( N^A \) and \( N^B \) 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:

\[
\begin{align*}
\mu_j^A &= \tilde{t}_j^A - \frac{Q}{N^A - \nu}, \\
\mu_j^B &= \tilde{t}_j^B - \frac{Q}{N^B - \nu},
\end{align*}
\]  

(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 \( \theta_A \) and \( \theta_B \), and sharing identical architectures, are established. Unlike the original co-teaching framework proposed in the previous literature (Han et al. 2018), \( \theta_A \) and \( \theta_B \) 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 \( D^t_A \) and \( D^t_B \). We independently train \( \theta_A \) and \( \theta_B \) using \( D^t_A \) and \( D^t_B \), respectively. Following this, \( D^t \) is evaluated using \( \theta_A \) and \( \theta_B \) to obtain \( \mu_j^A \) and \( \mu_j^B \), 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 \( \tilde{D}^t_A \) and \( \tilde{D}^t_B \). The union set encompassing \( \tilde{D}^t_A \) and \( \tilde{D}^t_B \) is then removed, yielding the selected examples for model training. In experiments, we set \( N^A = N^B = N_s \). To find the optimal values for \( N_s \) and \( \nu \), hyperparameter tuning on the validation set is employed.

**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 non-
sensitive 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.

### 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 $\|E(Y = 1 | S = A) - E(\hat{Y} = 1 | S = B)\|$.  

**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 $|P(\hat{Y} = 1 | S = A, Y = 1) - P(\hat{Y} = 1 | S = B, Y = 1)|$.

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

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 $\rho_A = 0$ and $\rho_B$ is 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
overall test error when $\nu$ is fixed 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 experimental 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 $\nu$ 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 $\nu$. 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 $\nu$ 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 $\nu$ 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 $\nu$ play a role in regulating the deviation from the initial value. Larger values of $\nu$ result in greater deviations from the original value, while smaller values of $N_s$ lead to more substantial deviations from the original value.

### Hyperparameter Analysis

We explore the impact of hyperparameters $\nu$ 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 $\nu$ 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 $\nu$. 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 $\nu$ 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 $\nu$ 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 $\nu$ play a role in regulating the deviation from the initial value. Larger values of $\nu$ result in greater deviations from the original value, while smaller values of $N_s$ lead to more substantial deviations from the original value.

### 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.

<table>
<thead>
<tr>
<th>Metric</th>
<th>20%</th>
<th>40%</th>
<th>20%</th>
<th>40%</th>
</tr>
</thead>
<tbody>
<tr>
<td>CL</td>
<td>22.61±0.79</td>
<td>0.17±0.01</td>
<td>25.13±1.30</td>
<td>0.24±0.04</td>
</tr>
<tr>
<td>LongReMix</td>
<td>19.86±0.09</td>
<td>0.18±0.06</td>
<td>21.03±1.21</td>
<td>0.19±0.03</td>
</tr>
<tr>
<td>LC</td>
<td>18.89±0.37</td>
<td>0.17±0.02</td>
<td>20.10±1.11</td>
<td>0.18±0.03</td>
</tr>
<tr>
<td>GPL</td>
<td>23.50±0.30</td>
<td>0.18±0.05</td>
<td>24.25±2.36</td>
<td>0.16±0.02</td>
</tr>
<tr>
<td>Ours</td>
<td>15.75±0.62</td>
<td>0.12±0.02</td>
<td>16.83±1.48</td>
<td>0.15±0.02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metric</th>
<th>20%</th>
<th>40%</th>
<th>20%</th>
<th>40%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Law</td>
<td>30.75±0.25</td>
<td>0.15±0.02</td>
<td>31.86±1.61</td>
<td>0.17±0.02</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metric</th>
<th>20%</th>
<th>40%</th>
<th>20%</th>
<th>40%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Credit</td>
<td>9.09±0.50</td>
<td>0.05±0.01</td>
<td>9.31±0.30</td>
<td>0.09±0.02</td>
</tr>
<tr>
<td>Ours</td>
<td>19.13±0.77</td>
<td>0.01±0.01</td>
<td>19.33±0.68</td>
<td>0.02±0.01</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>20%</th>
<th>40%</th>
<th>20%</th>
<th>40%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Symmetric</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>1.51±0.26</td>
<td>0.01±0.00</td>
<td>3.60±1.59</td>
<td>0.02±0.01</td>
</tr>
<tr>
<td>T</td>
<td>0.77±0.17</td>
<td>0.00±0.00</td>
<td>0.91±0.37</td>
<td>0.01±0.00</td>
</tr>
<tr>
<td>Asymmetric</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>M</td>
<td>0.45±0.08</td>
<td>0.01±0.00</td>
<td>0.55±0.36</td>
<td>0.01±0.01</td>
</tr>
<tr>
<td>T</td>
<td>0.36±0.06</td>
<td>0.00±0.00</td>
<td>0.38±0.10</td>
<td>0.00±0.00</td>
</tr>
</tbody>
</table>

### 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-
insights from Peer loss (Liu and Guo 2020) to formulate Wang, Liu, and Levy (2021) introduced bias using group-straint under various bias models. Building on these stud-
to recover the Bayes Optimal Classifier by combining Em-
ased training data on classification and revealed the potential
on the reliability of analyses that depend on the biased out-
chova, and G’Sell (2020) present experimental findings that

Figure 2: The illustration of the impact of $\psi(x)$: Plot (a) - (c), for (b), we set $N / N_s = 0.65$, 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 $\nu$ and $N_s$ is investigated using synthetic data.

gating biased label problems in fairness, Fogliato, Choulde-
chova, 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 out-
comes. Blum and Stangl (2020) explored the impact of bi-
ased training data on classification and revealed the potential
to recover the Bayes Optimal Classifier by combining Emp-
irical Risk Minimization with the equal opportunity con-
straint under various bias models. Building on these stud-
ies, Dai (2020) introduced a comprehensive framework ad-
dressing 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 cor-
rupted 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. Con-
versely, 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 ob-
jective function during training, our focus is on data sele-
cion for explicit label bias mitigation.

**Conclusion**

This paper demonstrates that bias can be eliminated by im-
plementing confident learning and filtering the fairest in-
stances, even with access to biased labels. However, chal-
enges 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 con-
idence score thresholding compared to traditional mean es-
timation methods. Leveraging interval estimation effectively
mitigates the adverse effects of biased labels, promoting fair-
ness 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 under-
score the proposed method’s efficacy in reducing the impact
of label bias and promoting fairness.

### Table 4: Experiment Results (Asymmetric Bias Scenario)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>CL</td>
<td>19.70 ± 0.49</td>
<td>0.14 ± 0.01</td>
<td>21.48 ± 0.76</td>
<td>0.16 ± 0.01</td>
<td>33.00 ± 0.20</td>
<td>0.19 ± 0.01</td>
<td>32.08 ± 0.74</td>
<td>0.19 ± 0.01</td>
</tr>
<tr>
<td>LongReMix</td>
<td>19.86 ± 1.33</td>
<td>0.15 ± 0.06</td>
<td>18.84 ± 0.58</td>
<td>0.16 ± 0.01</td>
<td>33.35 ± 1.01</td>
<td>0.18 ± 0.07</td>
<td>42.80 ± 0.83</td>
<td>0.15 ± 0.03</td>
</tr>
<tr>
<td>LC</td>
<td>16.36 ± 0.05</td>
<td>0.13 ± 0.01</td>
<td>17.53 ± 0.12</td>
<td>0.10 ± 0.02</td>
<td>37.20 ± 0.23</td>
<td>0.11 ± 0.01</td>
<td>41.16 ± 0.20</td>
<td>0.04 ± 0.00</td>
</tr>
<tr>
<td>GPL</td>
<td>18.87 ± 0.99</td>
<td>0.07 ± 0.02</td>
<td>18.72 ± 0.64</td>
<td>0.11 ± 0.02</td>
<td>39.28 ± 1.10</td>
<td>0.10 ± 0.02</td>
<td>44.14 ± 1.11</td>
<td>0.08 ± 0.01</td>
</tr>
<tr>
<td>Ours</td>
<td>17.76 ± 0.36</td>
<td>0.07 ± 0.01</td>
<td>17.46 ± 0.66</td>
<td>0.09 ± 0.01</td>
<td>30.75 ± 0.20</td>
<td>0.14 ± 0.02</td>
<td>30.06 ± 0.17</td>
<td>0.16 ± 0.01</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Metric</th>
<th>Law 20%</th>
<th>Vio.</th>
<th>Credit 20%</th>
<th>Vio.</th>
<th>Law 40%</th>
<th>Vio.</th>
<th>Credit 40%</th>
<th>Vio.</th>
</tr>
</thead>
<tbody>
<tr>
<td>CL</td>
<td>18.12 ± 4.16</td>
<td>0.31 ± 0.05</td>
<td>22.32 ± 6.01</td>
<td>0.31 ± 0.03</td>
<td>20.43 ± 1.06</td>
<td>0.02 ± 0.02</td>
<td>20.27 ± 0.96</td>
<td>0.02 ± 0.01</td>
</tr>
<tr>
<td>LongReMix</td>
<td>9.61 ± 0.51</td>
<td>0.13 ± 0.04</td>
<td>10.15 ± 0.80</td>
<td>0.14 ± 0.08</td>
<td>21.41 ± 0.45</td>
<td>0.03 ± 0.01</td>
<td>21.46 ± 0.40</td>
<td>0.02 ± 0.01</td>
</tr>
<tr>
<td>LC</td>
<td>9.33 ± 0.07</td>
<td>0.01 ± 0.00</td>
<td>9.39 ± 0.07</td>
<td>0.01 ± 0.00</td>
<td>20.43 ± 0.09</td>
<td>0.02 ± 0.00</td>
<td>21.39 ± 0.05</td>
<td>0.02 ± 0.00</td>
</tr>
<tr>
<td>GPL</td>
<td>10.69 ± 0.50</td>
<td>0.05 ± 0.01</td>
<td>11.00 ± 0.60</td>
<td>0.06 ± 0.01</td>
<td>21.18 ± 0.56</td>
<td>0.03 ± 0.01</td>
<td>20.80 ± 0.61</td>
<td>0.02 ± 0.01</td>
</tr>
<tr>
<td>Ours</td>
<td>8.99 ± 0.53</td>
<td>0.07 ± 0.03</td>
<td>9.23 ± 0.75</td>
<td>0.08 ± 0.01</td>
<td>18.77 ± 0.21</td>
<td>0.02 ± 0.01</td>
<td>18.93 ± 0.39</td>
<td>0.01 ± 0.00</td>
</tr>
</tbody>
</table>

This 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.
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 (22JDJ110001), 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


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.


