arXiv:2412.11072v1 [cs.LG] 15 Dec 2024

Navigating Towards Fairness with Data Selection

Yixuan Zhang[1], Zhidong Li[2], Yang Wang[2], Fang Chen[2], Xuhui Fan[3], Feng Zhou[4,5]*
[1]School of Statistics and Data Science, Southeast University, China
[2]Data Science Institute, University of Technology Sydney, Australia
[3]School of Computing, Macquarie University, Australia
[4]Center for Applied Statistics and School of Statistics, Renmin University of China, China
[5]Beijing Advanced Innovation Center for Future Blockchain and Privacy Computing, China
zh1xuan@hotmail.com, {zhidong.li, yang.wang, fang.chen}@uts.edu.au, xuhui.fan@mq.edu.au, feng.zhou@ruc.edu.cn

Abstract

Machine learning algorithms often struggle to eliminate inherent data biases, particularly those arising from unreliable labels, which poses a significant challenge in ensuring fairness. Existing fairness techniques that address label bias typically involve modifying models and intervening in the training process, but these lack flexibility for large-scale datasets. To address this limitation, we introduce a data selection method designed to efficiently and flexibly mitigate label bias, tailored to more practical needs. Our approach utilizes a zero-shot predictor as a proxy model that simulates training on a clean holdout set. This strategy, supported by peer predictions, ensures the fairness of the proxy model and eliminates the need for an additional holdout set, which is a common requirement in previous methods. Without altering the classifier's architecture, our modality-agnostic method effectively selects appropriate training data and has proven efficient and effective in handling label bias and improving fairness across diverse datasets in experimental evaluations.

Introduction

Fairness is a critical and essential problem in real-world applications. In recent years, it has attracted great attention, especially in high-stake domains such as finance (Khandani, Kim, and Lo 2010; Yeh and Lien 2009; Mukerjee et al. 2002), law (Brennan, Dieterich, and Ehret 2009; Lin et al. 2020), recruiting (Faliagka et al. 2012; Bertrand and Mullainathan 2004), school admissions (Moore 1998) and medicine (Kim et al. 2015). Although many fairness-aware learning methods have been proposed recently, they often assume that the data collected for training is representative of the true data distribution. However, these methods are still validated on so-called "clean data," which neglects the impact of *label bias*.

Recent studies (Wang, Liu, and Levy 2021; Dai 2020; Zhang et al. 2023; Konstantinov and Lampert 2022) have increasingly focused on the adverse impacts of label bias and have proposed methodologies to improve fairness in this setting. These efforts involve adjusting models and intervening in the training process by considering the amount of bias present in labels during fair learning. The objective of these methods is to develop a fair labeling function that reflects the true underlying distribution, thereby improving the robustness of the fairness techniques. However, when handling large-scale datasets, these methods often lack flexibility due to the necessity of training with the complete dataset, which can slow down model convergence and substantially increase training costs.

To meet the needs of complex machine learning systems in practical applications, employing data selection methods to filter out a subset of useful data for training proves to be an effective solution. Apart from efficiency improvement, data selection techniques have been very effective in mitigating the impact of noisy data. These methods typically prioritize training with either difficult (Loshchilov and Hutter 2015; Katharopoulos and Fleuret 2018; Jiang et al. 2019) or easy samples (Bengio et al. 2009) based on the training loss. Nevertheless, such a singular filtering strategy limits the ability to handle the diversity of real-world situations since the difficulty of samples often arises from incorrect annotations, inherent ambiguity, or atypical patterns (Mindermann et al. 2022; Deng, Cui, and Zhu 2023). To overcome this limitation, Mindermann et al. (2022) introduces a new data selection criterion, the reducible holdout loss selection (RHO-LOSS), based on the impact on generalization loss, which further prevents the selection of redundant or noisy samples.

Inspired by this new selection criterion, this paper aims to extend this method to the field of fairness. We revisit the derivation of RHO-LOSS, aligning the predicted posterior distribution of selected samples with the fair data distribution. Instead of relying on a holdout set and training an auxiliary validation model, we establish a more accurate approximation by incorporating zero-shot predictors from pre-trained models. This approach eliminates the need for an additional holdout set. To further prevent discrimination leakage in pre-trained models, we implement the peer prediction mechanism during training. This ensures fairness when using the zero-shot predictor to evaluate generalization loss.

We conduct comprehensive empirical evaluations on several benchmark datasets. The experiments on the image classification tasks demonstrate the effectiveness of our proposed data selection principle, which can adaptively select fair instances that are less impacted by label bias. Although not explicitly emphasized in this paper, we also offer a solution to reduce the impact of selection bias by resampling the

*Corresponding author.

Copyright © 2025, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

selected data. It is worth noting that the proposed method is *modality-agnostic*, as it only provides a principle of data selection and can be compatible with any log-likelihood or cross-entropy-based classifiers. Meanwhile, it achieves faster convergence as we only select "good" data points for model training. Our method is robust and achieves superior performance with respect to accuracy and fairness metrics under different bias settings. Our contributions are three-fold:

- We propose a data selection principle that enables the learning of a fair labeling function by selecting fair and balanced instances in the training set to fix both label and selection bias. Notably, unlike most noisy label learning methods, our method does not require noise rate estimation.
- Our method is general, modality-agnostic, and compatible with any classifier or neural network based on log-likelihood or cross-entropy-based.
- Our method converges faster and achieves better test accuracy than alternative baselines.

Preliminaries

In this section, we briefly introduce the related background knowledge of online batch selection and present the concept of the data selection principle based on the data impact on the generalization loss, as proposed in Mindermann et al. (2022).

Consider a dataset $\mathcal{D} = \{(x_i, y_i, s_i)\}_{i=1}^n$, where x represents the non-sensitive features, $y \in \{0, 1\}$ is the binary label, and $s \in \{0, 1\}$ denotes the sensitive variables. Given a model parameterized by θ, $f_\theta : \mathcal{X} \to \mathbb{R}^K$ (K is the number of class), in online batch learning, a batch B_t of size N_B is drawn from the training dataset $\mathcal{D}$ at each training step t. The objective of online batch selection is to pick samples from B_t based on specific ranking criteria and use them to construct a smaller batch b_t with size N_b for model updates. The ranking function represents the criteria of which data should be selected, and previous methods usually pick the "hard" data points based on the ranking of training loss from highest to lowest. Then, the common gradient descent is performed to minimize the loss using b_t, which is denoted as $\sum_{i=1}^{N_b} L(y_i, f(x_i))$. By iteratively doing this, we can select data points that minimize the loss of training set. However, this kind of selection method will tend to pick redundant data points or outliers since it focuses on the highest training loss and lack the flexibility due to the simplistic selection criteria.

In contrast, Mindermann et al. (2022) established selection criteria based on the data impact on the model's generalization loss, thus effectively addressing the limitations of previous methods. To revisit this approach, we follow the framework of online batch selection. Let $\mathcal{D}_t$ represent the observed data prior to training step $t + 1$. Given a sample (x, y) drawn from batch B_{t+1}, we assume, for simplicity, that only one data point is selected at a time. If this data point is chosen, the updated predictive distribution in a Bayesian view will be $p(y' \mid x', \mathcal{D}_t \cup (x, y))$. This distribution should ideally align with the true data-generating distribution $\dot{p}(x', y')$, and to achieve this goal, one nature way is to minimize the KL divergence between the predictive distribution and the data-generating distribution, $\mathbb{E}_{\dot{p}(x')} KL[\dot{p}(y' \mid x') || p(y' \mid x', \mathcal{D}_t \cup (x, y)]$, which can be equivalently expressed as:

$$-\mathbb{E}_{\dot{p}(x',y')}[\log p(y' \mid x', \mathcal{D}_t \cup (x, y)] + \text{const.}, \quad (1)$$

where the const. denotes the negative entropy of the data generating distribution, and is agnostic to the optimization. In order to evaluate the impact of the selected data point on the generalization loss, we utilize the holdout samples $\mathcal{D}^* = \{(x_i^*, y_i^*)\}_{i=1}^m$ from the data-generating distribution $\dot{p}(x', y')$. By leveraging the Monte Carlo to approximate expectations with empirical averages, the aim now becomes:

$$\arg\max_{(x,y)\in B_{t+1}} \frac{1}{m} \sum_{i=1}^{m} [\log p(y_i^* \mid x_i^*, \mathcal{D}_t \cup (x, y))], \quad (2)$$

and this optimization is equivalent to select a data point $(x, y) \in B_{t+1}$ that mostly maximizes $\log p(\mathcal{D}^* \mid \mathcal{D}_t \cup (x, y))$, which corresponds to the generalization loss. By applying Bayes' rule, we obtain:

$$p(\mathcal{D}^* \mid \mathcal{D}_t \cup (x, y)) = \frac{p(y \mid x, \mathcal{D}^*, \mathcal{D}_t)}{p(y \mid x, \mathcal{D}_t)} p(\mathcal{D}^* \mid \mathcal{D}_t). \quad (3)$$

The term $p(\mathcal{D}^* \mid \mathcal{D}_t, x)$ becomes $p(\mathcal{D}^* \mid \mathcal{D}_t)$ is due to a single data point x does not influence the model's update performance. Approximating $p(y \mid x, \mathcal{D}^*, \mathcal{D}_t)$ with $p(y \mid x, \mathcal{D}^*)$ and dropping the term $p(\mathcal{D}^* \mid \mathcal{D}_t)$ (independent of (x, y)), the final approximated tractable selection function of RHO-LOSS is given by:

$$\begin{aligned} &\arg\max_{(x,y)\in B_{t+1}} \log p(y \mid x, \mathcal{D}^*) - \log p(y \mid x, \mathcal{D}_t) \\ &= \arg\max_{(x,y)\in B_{t+1}} L[y \mid x, \mathcal{D}_t] - L[y \mid x, \mathcal{D}^*]. \end{aligned} \quad (4)$$

The first term, $L[y \mid x, \mathcal{D}_t]$, represents the training loss using model trained on the training set $\mathcal{D}_t$, while $L[y \mid x, \mathcal{D}^*]$ is the irreducible holdout loss using model trained on the holdout set $\mathcal{D}^*$. Thus, the aim of selecting a data point $(x, y) \in B_{t+1}$ that mostly maximize the log-likelihood on the holdout set (Eq. (2)) can be approximated by Eq. (4).

Method

Building on the concept of RHO-LOSS, we now turn our attention to how it connects with fairness and how label bias impacts data selection.

Refining Selection Principle for Fairness

We begin by revisiting the derivation of RHO-LOSS and incorporating fairness principles (demographic parity[1]) to refine its formulation. Consider a holdout dataset $\mathcal{D}^* = \{x_i^*, y_i^*, s_i^*\}_{i=1}^m$, where all samples are generated from a fair distribution $p(x, y, s) = p(y \mid x)p(x)p(s)$, meaning labels are not influenced by sensitive attributes. While fairness

[1] We use demographic parity as an illustration here, other fairness metrics are introduced in the experiment.

has various interpretations (Barocas, Hardt, and Narayanan 2018), we follow a common assumption in fairness literature and it is worth noting that, such a fair distribution is an idealized condition and is not observed in practice. Our objective now is to select a sample (x, y) from batch B_{t+1}, belonging to the demographic group defined by s, ensuring that the updated predictive distribution closely aligns with the fair data distribution. Similar to the derivation in RHO-LOSS, by expanding the KL expression and applying Monte Carlo estimation using additional holdout samples, the optimization problem becomes:

$$\underset{(x,y)\in B_{t+1}}{\arg\max} \sum_{s} p(s) \left[\log p(\mathcal{D}^* \mid \mathcal{D}_t \cup (x, y), s)\right]. \tag{5}$$

We estimate $p(s)$ using $\frac{C_s}{m}$, where C_s represents the number of samples belonging to demographic group s under fair distribution. For demographic group s, using Bayes' rule, the updated predictive distribution is as follows:

$$p(\mathcal{D}^* \mid \mathcal{D}_t \cup (x, y), s) = \frac{p(y \mid x, \mathcal{D}^*, \mathcal{D}_t, s)}{p(y \mid x, \mathcal{D}_t, s)} p(\mathcal{D}^* \mid \mathcal{D}_t). \tag{6}$$

By plugging Eq. (6) into Eq. (5) and omitting the term that is irrelevant to the selected sample, we reformulate the objective as:

$$\underset{(x,y)\in B_{t+1}}{\arg\max} \sum_{s} \frac{C_s}{m} [\log p(y \mid x, \mathcal{D}^*, \mathcal{D}_t, s) - \log p(y \mid x, \mathcal{D}_t, s)]. \tag{7}$$

The second term in Eq. (7) is straightforward to compute, but the first term is challenging to estimate because it involves both the training and holdout data. To address this, RHO-LOSS approximates the term with $\log p(y \mid x, \mathcal{D}^*)$, while Deng, Cui, and Zhu (2023) approximates it by finding a lower bound. We adopt the lower bound approximation, which is expressed as follows:

$$\log p(y \mid x, \mathcal{D}^*, \mathcal{D}_t, s) \geq \mathbb{E}_{p(\theta \mid \mathcal{D}^*)} \log p(y \mid x, \theta, s), \tag{8}$$

where θ is the model parameter. For completeness, we include the full derivation in the Appendix. Substituting Eq. (8) into Eq. (7), the resulting objective is as follows:

$$\underset{(x,y)\in B_{t+1}}{\arg\max} \sum_{s} \frac{C_s}{m} \left[\alpha \mathbb{E}_{p(\theta \mid \mathcal{D}^*)} \log p(y \mid x, \theta, s) - \log p(y \mid x, \mathcal{D}_t, s)\right], \tag{9}$$

where $\alpha \in [0, 1]$ is a scaling factor that determines the trade-off between the original inequality and the lower bound.

Although we obtain this lower bound to approximate the first term in Eq. (7), it cannot be directly calculated. To further make it tractable, we build on the framework outlined by Deng, Cui, and Zhu (2023) to approximate this lower bound using zero-shot predictors as a proxy for the validation model in Mindermann et al. (2022), thereby eliminating the need to gather additional holdout data. Zero-shot predictors have demonstrated promising transfer performance across a wide range of downstream tasks due to training on extensive datasets. Consequently, we use the following approximation:

$$\mathbb{E}_{p(\theta \mid \mathcal{D}^*)} \log p(y \mid x, \theta, s) \approx \log p(y \mid \tilde{f}(x), s), \tag{10}$$

where $\tilde{f}$ represents a zero-shot predictor derived from a pre-trained model used as a validation model. The approximation is considered reasonable because the expectation is taken over the posterior of θ, and it is assumed that the dataset used to train the zero-shot predictor is sufficiently large, resulting in a very narrow posterior distribution. Consequently, we can directly extract the posterior mean, represented by $\tilde{f}$, and substitute it into $\log p(y \mid x, \theta; s)$ as an effective approximation of $\mathbb{E}_{p(\theta \mid \mathcal{D}^*)} \log p(y \mid x, \theta; s)$. Then, substitute Eq. (10) into Eq. (9), the selection function with fairness considerations can be reformulated as:

$$\underset{(x,y)\in B_{t+1}}{\arg\max} \sum_{s} \frac{C_s}{m} \left[L[y \mid x, \mathcal{D}_t, s] - \alpha L[y, \tilde{f}(x), s]\right]. \tag{11}$$

The final objective is structured similarly to RHO-LOSS: the first term denotes the training loss, while the second term approximates the holdout loss via a zero-shot predictor.

Fair Data Selection with Peer Prediction Mechanism

In this section, we present our revised selection approach to enhance fairness. Previous methods rely on a clean holdout set for training validation models, which is not always feasible in practice. Instead, we use zero-shot predictors as a proxy model, eliminating the need for a holdout set (see Eq. (11)). Despite the pre-trained model's robust ability to effectively extract and utilize fundamental patterns from large datasets, it may still inherit and propagate label bias, failing to adequately reflect the data-generating distribution of the current task, especially in terms of fairness. This misalignment can lead to persistent discrimination in the proxy model's approximations, mainly because most pre-trained models neglect to emphasize fairness during training, resulting in inherent biases. Even models developed with fairness considerations can introduce biases in downstream tasks due to distribution shifts (Jiang et al. 2023; Schrouff et al. 2022; Chowdhury and Chaturvedi 2023). To address these issues and tackle inherent label bias, we incorporate a peer prediction mechanism into our zero-shot predictor.

Peer Prediction Mechanism The correlated agreement type (Dasgupta and Ghosh 2013; Shnayder et al. 2016) of peer prediction mechanism involves two agents: one that provides noisy labels and another that mimics the Bayes optimal classifier. Liu and Guo (2020) crafted a scoring function that encourages truthful reporting by ensuring that the optimal classifier maximizes its score with accurate predictions. By minimizing the negative scoring function as a loss (referred to as *peer loss*), the resulting classifier closely approximates the Bayes optimal classifier, effectively addressing challenges posed by noisy labels. The loss function is formulated as: $L[y_i \mid x_i] - \gamma L[y_{i_2} \mid x_{i_1}]$, where x_{i_1} and y_{i_2} are independently sampled from the training set, excluding the (x_i, y_i) pair, and $\gamma \in [0, 1]$ is a parameter that makes the loss robust to imbalanced labels.

Final Fair Data Selection Objective Drawing on the principles of the peer prediction mechanism, we implement a similar strategy to eliminate label bias and hence ensure fairness. We define $\mathcal{D}_s$ and $\mathcal{D}_{s'}$ as the subsets containing sensitive information for $S = s$ and $S = s'$, respectively. In a fair setting, predictions should be independent of s. To achieve this, we modify the original peer loss by sampling x_{i_1} from $\mathcal{D}_s$ and y_{i_2} from $\mathcal{D}_{s'}$, and the corresponding random variables are the pair of (X_{i_1}, Y_{i_2}). The intuition behind this design is that we assume the bias exists, and we construct such cross-group pairings to create a "biased" version. This setup allows us to estimate the bias rate conditioned on different demographic groups using the peer loss function. Since this loss function involves pairing two randomly selected instances, it inherently contains some randomness. To stabilize the results, we calculate the expectation across these randomly selected instances, and the equivalent expectation version is (details in the Appendix):

$$L[y_i, \tilde{f}(x_i), s] - \gamma \mathbb{E}_{Y|\mathcal{D}_{s'}}[L[Y, \tilde{f}(x_i), s]]. \tag{12}$$

Combining Eq. (12) into Eq. (11) and simplifying Eq. (11), with the weights $\frac{C_s}{m}$ summing to 1, we obtain the final selection function:

$$\begin{aligned} \arg\max_{(x,y)\in B_{t+1}} \; & L[y \mid x, \mathcal{D}_t, s] + (1-\alpha) L[y, \tilde{f}(x), s] \\ & - \gamma \mathbb{E}_{Y|\mathcal{D}_{s'}}[L[Y, \tilde{f}(x), s]]. \end{aligned} \tag{13}$$

Resampling to Deal with Selection Bias In addition to addressing label bias, our implementation also tackles selection bias, which contributes to label imbalance. Assuming selection bias impacts the statistical independence between demographic groups S and target labels Z (Kamiran and Calders 2012), we resample based on the discrepancies between the actual counts of individuals in demographic group $s = i$ with clean label $z = j$ ($C_{s,z}$) and their expected counts ($\mathbb{E}[C_{s,z}]$). We estimate $\mathbb{E}[C_{s,z}]$ using $p(s)p(z)N$, where $p(s)$ and $p(z)$ are the empirically measured probabilities. To correct imbalances, we upsample subgroups where $C_{s,z} < \mathbb{E}[C_{s,z}]$ and downsample those where $C_{s,z} > \mathbb{E}[C_{s,z}]$. Although the clean label Z is unobserved, we assume that the labels of selected instances are fair and treat them accordingly. This resampling strategy is applied to each selected batch to ensure data balance.

Implementation The training process is outlined in Algorithm 1. We employ a zero-shot predictor with a derived expectation version of the adapted peer loss as a surrogate model to simulate the estimated loss on the holdout set. For each instance in batch B_{t+1}, we evaluate Eq. (13) and select the top-N_b instances to form a smaller batch b_{t+1}. To address selection bias, we further pick samples based on the severity of selection bias to ensure balance among different demographic groups. We then update the parameters θ of the target model using this selected data.

Why the fair selection principle pick instances less influenced by label bias? Our data selection method integrates a zero-shot predictor with a peer prediction mechanism to address label bias, using the derived loss as a validation measure against a fair distribution. By evaluating each data point's impact on this validation loss, we avoid instances likely discriminated by sensitive information and exclude irrelevant outliers. This approach also conceptually relates to the Query By Committee (QBC) algorithm (Seung, Opper, and Sompolinsky 1992; Freund et al. 1997), which selects examples based on classifier disagreement (here, between the current model and $\tilde{f}$). Following Cheng et al. (2021); Zhang et al. (2021) (detailed in the Appendix), we demonstrate that our method effectively promotes fairness and mitigates label bias through a breakdown of Eq. (12):

Algorithm 1: Fair data selection to address label bias issue

```
Input: training set $\mathcal{D}$, $N_b$, $N_B$, $T$, $\alpha$, $\gamma$, zero-shot predictor $\tilde{f}$ and a target model $f_\theta$.
Initialize $\theta_0$.
for $t$ in $0, \cdots, T$ do
    Randomly select $N_B$ instances to construct $B_{t+1}$;
    For each sample $(x_i, y_i, s_i)$ in $B_{t+1}$, estimate and compute the objective in Eq. (13);
    Select top-$N_b$ samples to construct $b_{t+1}$;
    Drop instances from $b_{t+1}$ if $C_{s,z} > \mathbb{E}[C_{s,z}]$, otherwise, bootstrap;
    Perform gradient descent and update $\theta$ with resampled data.
end for
```

$$\begin{aligned} & \mathbb{E}_{\tilde{D}} \left[L[Y, \tilde{f}(X), S] - \gamma \mathbb{E}_{Y|\mathcal{D}_{s'}} L[Y, \tilde{f}(x), s] \right] \\ & = \underbrace{\mathbb{E}_D \Delta_s L[Y, \tilde{f}(X), S]}_{\text{fair model}} \\ & \underbrace{+ \sum_{j\in[K]} \sum_{i\in[K]} \sum_{s\in[S]} P(S=s) P(Z=i) [\Delta_{D_u} L[j, \tilde{f}(X), s]]}_{\text{Penalty on noisy loss}} \\ & \underbrace{-\gamma \sum_{j\in[K]} p_j^s \left[\mathbb{E}_{D_{X|S=1}} L[j, \tilde{f}(X)] - \mathbb{E}_{D_{X|S=0}} L[j, \tilde{f}(X)] \right],}_{\text{Penalty for the disagreement between demographic groups}} \end{aligned} \tag{14}$$

where $\tilde{D}$ denotes the observed distribution, which contains label bias, while D represents the underlying clean fair distribution. $\Delta_s = 1 - P(Y = 1 \mid Z = 0, S = s) - P(Y = 0 \mid Z = 1, S = s)$, $p_j^s = P(Y = j, S = 0) - P(Y = j, S = 1)$, $T_{ij}^s = P(Y = j \mid Z = i, S = s)$, $\Delta_{D_u} = \mathbb{E}_{D|Z=i,S=s}(U_{ij}^s - \gamma P(Y = j \mid S = s))$ and $U_{ij}^s = T_{ij}^s$ if $i \neq j$, and $U_{ij}^s = T_{jj}^s - \mathbb{E}_{D|Z=j,S=s} T_{jj}^s$ if $i = j$. The equation's first term captures the model's clean loss on the fair distribution. The second term adds penalties for noisy losses, adjusting labels to account for observed demographic disparities. The third term imposes penalties for demographic discrepancies, ensuring fair performance across all groups.

Experiment

In the subsequent sections, we first describe our experimental setup, covering datasets, baselines, and evaluation met-

Method/Dataset	LFW+a(0.2)		LFW+a(0.4)		CelebA(0.2)		CelebA(0.4)	
	ACC	Δ_{DP}	ACC	Δ_{DP}	ACC	Δ_{DP}	ACC	Δ_{DP}
CLIP	90.8	0.11	83.2	0.21	62.2	0.60	61.6	0.61
Grad Norm	77.5$_{\pm0.7}$	0.02$_{\pm0.01}$	75.6$_{\pm0.8}$	0.07$_{\pm0.01}$	76.4$_{\pm2.8}$	0.28$_{\pm0.07}$	74.0$_{\pm1.9}$	0.26$_{\pm0.02}$
Grad Norm IS	82.6$_{\pm1.4}$	0.43$_{\pm0.13}$	75.4$_{\pm0.2}$	0.78$_{\pm0.01}$	82.6$_{\pm2.1}$	0.40$_{\pm0.14}$	75.5$_{\pm3.4}$	0.42$_{\pm0.07}$
Uniform	89.0$_{\pm0.9}$	0.03$_{\pm0.01}$	79.8$_{\pm5.1}$	0.07$_{\pm0.04}$	82.0$_{\pm1.3}$	0.29$_{\pm0.08}$	81.0$_{\pm3.1}$	0.31$_{\pm0.09}$
RHO-LOSS	89.5$_{\pm0.5}$	0.06$_{\pm0.02}$	83.9$_{\pm1.8}$	0.08$_{\pm0.20}$	80.5$_{\pm0.9}$	0.22$_{\pm0.07}$	79.7$_{\pm1.7}$	0.28$_{\pm0.04}$
Ours-s	90.0$_{\pm1.3}$	0.02$_{\pm0.01}$	86.2$_{\pm0.7}$	0.08$_{\pm0.05}$	85.5$_{\pm0.6}$	**0.21**$_{\pm0.02}$	84.5$_{\pm0.6}$	0.21$_{\pm0.01}$
Ours	**90.9**$_{\pm0.6}$	**0.01**$_{\pm0.00}$	**88.7**$_{\pm0.7}$	**0.04**$_{\pm0.01}$	**86.5**$_{\pm0.6}$	**0.21**$_{\pm0.02}$	**85.2**$_{\pm1.7}$	**0.20**$_{\pm0.01}$

Table 1: Test accuracy (%) and fairness violation (Δ_{DP}) on CelebA and LFW+a with various symmetrical label bias amount of 20% and 40%. We report results in the format of mean ± standard deviation.

Method/Dataset	LFW+a(0.2)		LFW+a(0.4)		CelebA(0.2)		CelebA(0.4)	
	p% ratio	Δ_{DEO}	p% ratio	Δ_{DEO}	p% ratio	Δ_{DEO}	p% ratio	Δ_{DEO}
CLIP	88.9	0.03	77.4	0.07	38.3	0.32	37.3	0.33
Grad Norm	92.7$_{\pm1.5}$	0.03$_{\pm0.01}$	91.1$_{\pm0.6}$	-	77.6$_{\pm7.3}$	0.39$_{\pm0.02}$	76.0$_{\pm8.9}$	0.55$_{\pm0.23}$
Grad Norm IS	90.4$_{\pm2.6}$	**0.01**$_{\pm0.01}$	90.2$_{\pm0.5}$	-	31.9$_{\pm0.0}$	-	32.5$_{\pm0.0}$	-
Uniform	97.9$_{\pm1.9}$	**0.01**$_{\pm0.01}$	90.1$_{\pm5.2}$	**0.14**$_{\pm0.09}$	76.3$_{\pm3.9}$	0.52$_{\pm0.33}$	68.6$_{\pm5.9}$	0.77$_{\pm0.10}$
RHO-LOSS	92.3$_{\pm2.8}$	0.03$_{\pm0.01}$	90.4$_{\pm2.4}$	-	80.8$_{\pm5.9}$	0.44$_{\pm0.14}$	**81.0**$_{\pm0.1}$	0.42$_{\pm0.01}$
Ours	**98.3**$_{\pm0.4}$	**0.01**$_{\pm0.00}$	**94.6**$_{\pm3.7}$	**0.14**$_{\pm0.08}$	**84.1**$_{\pm5.3}$	**0.38**$_{\pm0.15}$	77.2$_{\pm1.1}$	**0.35**$_{\pm0.02}$

Table 2: Other fairness measure with p%-rule and Δ_{DEO} on CelebA and LFW+a with various symmetrical label bias amount of 20% and 40%. '-' denotes the invalid measure of DEO due to low accuracy.

rics. Next, we compare our methods against existing state-of-the-art data selection techniques across various image classification tasks (CelebFaces Attributes (**CelebA**) (Liu et al. 2015) and modified Labeled Faces in the Wild Home (**LFW+a**) (Wolf, Hassner, and Taigman 2011)), considering different amounts of label bias. We examine our selection criteria through detailed ablation studies.

Benchmark Datasets. We evaluate the performance of our proposed method using two image datasets: CelebA and LFW+a. The CelebA dataset is utilized to discern the label HeavyMakeup, considering gender ("Female") as the sensitive variable where biases have been noted towards female. In the LFW+a dataset, we augment each image with additional attributes like gender and race (same in CelebA), aiming to classify the identity's gender. The sensitive variable here is "WavyHair", where literature has shown a strong correlation regarding males. Each dataset is divided into training, validation, and test sets.

Baselines. To evaluate our method's effectiveness and robustness, we compare our method with several selection methods on the image tasks. These include uniform sampling (**Uniform**), gradient norm selection (**Grad Norm**, which selects data points with high gradient norms) (Katharopoulos and Fleuret 2018), and gradient norm with importance sampling (**Grad Norm IS**) (Katharopoulos and Fleuret 2018). Additionally, we compare it with **RHO-LOSS** (Mindermann et al. 2022). We implement two variants of our method: one includes resampling of b_{t+1} (**Ours**), and the other does not (**Ours-s**).

Evaluation Metrics. We use accuracy to evaluate prediction performance and measure fairness violation with $\Delta_{\text{DP}} = |\mathbb{E}(\hat{Y} = 1 \mid S = 1) - \mathbb{E}(\hat{Y} = 1 \mid S = 0)|$. A lower Δ_{DP} indicates less fairness violation. We also conduct experiments on the difference of equal opportunity (DEO) (Hardt, Price, and Srebro 2016), which is defined as $\Delta_{\text{DEO}} = |\mathbb{E}(\hat{Y} = 1 \mid Y = 1, S = 1) - \mathbb{E}(\hat{Y} = 1 \mid Y = 1, S = 0)|$, and the p%-rule, p% $= \min(\frac{P(\hat{Y}=1|S=0)}{P(\hat{Y}=1|S=1)}, \frac{P(\hat{Y}=1|S=1)}{P(\hat{Y}=1|S=0)})$. A lower Δ_{DEO} suggests less fairness violation, while a lower p%-rule indicates higher fairness violation.

Setup. In our experiments addressing label bias, we introduce symmetrical label biases of 20% and 40%. We use the AdamW optimizer (learning rate 0.001, weight decay 0.01), we set a batch size of $N_b = 32$ and a batch ratio $\frac{N_b}{N_B} = 0.1$, consistent with the RHO-LOSS setup. For the LFW+a dataset, we employ ResNet-18 (He et al. 2016), and for CelebA, we use ResNet-50 across all methods, along with a zero-shot predictor based on CLIP-RN50. We vary α and γ within the set $\{0.1, 0.3, 0.5, 0.7, 0.9\}$. Results are averaged over three random trials. All experiments are performed with GPUs (NVIDIA GeForce RTX 3090 with 86GB memory).

Comparison Results

Results are displayed in Table 1 for 20% and 40% bias amount. In the meantime, we report the fairness measure using p%-rule and Δ_{DEO} in Table 2. Our proposed method consistently demonstrates the highest accuracy and minimal fairness violation as bias increases. The outcomes illustrate

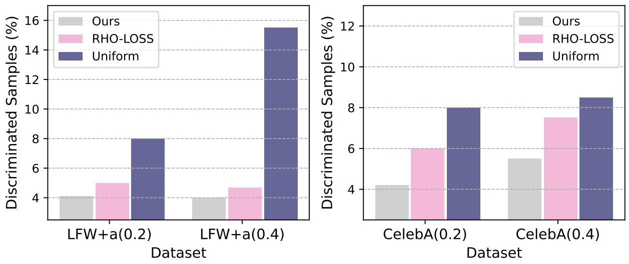

Figure 1: Proportion of selected instances discriminated by label bias using the proposed method (Ours), RHO-LOSS, and Uniform Sampling. The left plot corresponds to the LFW+a dataset, and the right plot corresponds to the CelebA dataset. Overall, we can observe that the proposed method has the lowest rate of discriminated sample selection.

the importance of addressing label bias to prevent heightened bias in the output. While other baselines work well with low bias, they are not robust when the bias amount increases. Interestingly, we find that gradient norm selection has the worst performance, even worse than uniform sampling, especially for large label bias amount. This phenomenon shows that selecting data by high variance will tend to pick "dirty" points that are affected by the sensitive information.

Analysis of Properties of Selected Data

In this section, we analyze our proposed method by evaluating the proportion of fair instances to the selected data. Fig. 1 reveal a pattern: as the amount of bias increases, the proportion of selected instances that are fair starts to decline due to the increased difficulty in distinguishing fair instances from unfair ones. However, our proposed selection method still maintains the highest ratio of selected fair instances, and significantly exceeds that of uniform sampling. This observation reinforces our method's superiority over other data selection methods in different label bias settings.

Ablation Studies

In this section, we conduct ablation studies on zero-shot predictor, model architecture, and important hyperparameters.

Zero-shot Predictor. In our experiments, we initially used CLIP-RN50 as the proxy model and have now extended testing to include ViT-B/16 (Dosovitskiy et al. 2021) and the validation model from RHO-LOSS. Results in Table 3 show consistent performance across different zero-shot backbones. For ViT-B/16 and RHO-LOSS models, baseline accuracies are 55% and 65% under 20% label bias, and 52% and 64% under 40% label bias, respectively. ViT-B/16 accuracies are slightly lower than RN50, but our data selection method with peer prediction maintains a comparable performance on CelebA at a 0.4 bias rate. RHO-LOSS model performance is slightly higher than RN50, but results are similar. This confirms the robustness of our method, which performs effectively regardless of the zero-shot predictor used.

Backbone. We test our proposed method with different Backbone architecture on CelebA dataset, including the variant of ResNet (ResNet-18 and ResNet-50) and the DenseNet-121 (Huang et al. 2017). The results are displayed in Table 3. From the results, we can see enhance model complexity generally improves both predictive performance and fairness, though these improvements are not significant. Overall, the impact of the model's structure on its performance is not obvious, which implies our proposed method is robust to the backbone structure.

Component		CelebA(0.2)		CelebA(0.4)	
		ACC	Δ_{DP}	ACC	Δ_{DP}
Zero-shot Predictor	CLIP-RN50	$85.5_{\pm 0.6}$	$0.21_{\pm 0.02}$	$85.2_{\pm 1.7}$	$0.23_{\pm 0.06}$
	ViT-B/16	$86.4_{\pm 0.2}$	$0.20_{\pm 0.03}$	$84.5_{\pm 1.2}$	$0.19_{\pm 0.02}$
	R-V	$85.6_{\pm 0.4}$	$0.23_{\pm 0.05}$	$85.5_{\pm 0.9}$	$0.23_{\pm 0.04}$
Backbone Model	ResNet-18	$84.8_{\pm 1.4}$	$0.22_{\pm 0.02}$	$84.1_{\pm 0.2}$	$0.22_{\pm 0.01}$
	ResNet-50	$86.5_{\pm 0.6}$	$0.21_{\pm 0.02}$	$85.2_{\pm 1.7}$	$0.20_{\pm 0.01}$
	DenseNet-121	$85.3_{\pm 0.2}$	$0.21_{\pm 0.01}$	$84.3_{\pm 0.4}$	$0.19_{\pm 0.02}$

Table 3: Test accuracy (%) and fairness violations (Δ_{DP}) for a variant zero-shot predictor (R-V denotes the validation model used in RHO-LOSS) and the backbone of our method.

Hyperparameters. We then analyze the effects of three hyperparameters: γ, α, and the selection ratio. In the first plot in Fig. 2, we plot the difference of test accuracy and fairness violation with γ varying, we set $\alpha = 0.1$. We can see that the accuracy slightly improves as γ increases and the fairness violation appears downward trend as γ increases. This align with the effect of γ that controls the fairness level (fairness and accuracy should improve simultaneously when the data is unbiased (Wick, panda, and Tristan 2019)). In the second plot in Fig. 2, we set $\gamma = 0.3$ and test different values of α, we can see a similar trend for accuracy, but for the fairness violation, remains fairly stable. For the selection ratio, by default, is 10% in the experiment. In the third plot in Fig. 2, we plot the change of accuracy and fairness violation w.r.t. the selection ratio. We can see the accuracy increases when the selection ratio increases from 0.1 to 0.3 and then begins to drop. This demonstrates that increasing the selection ratio also increases the likelihood of picking data points that are not fair enough. However, due to the effect of the second term in Eq. (14) acting as the fair regularizer, the fairness violation shows minimal variation.

Convergence and Accuracy

We conduct an experiment about convergence speed by testing the epochs required to reach target test accuracy on the two image datasets. Interestingly, due to our proposed data selection procedure, it has a faster convergence speed. In Table 4, we can see that the proposed method converges much faster than uniform sampling and RHO-LOSS. With fewer epochs, our proposed method can achieve the target accuracy point. In the meantime, as shown in Table 1, the proposed method has the highest test accuracy compared to the other two methods. These results also align with Fig. 1, and can be explained as the proposed method is able to pick the fairest instances and therefore improve both accuracy and fairness at the same time.

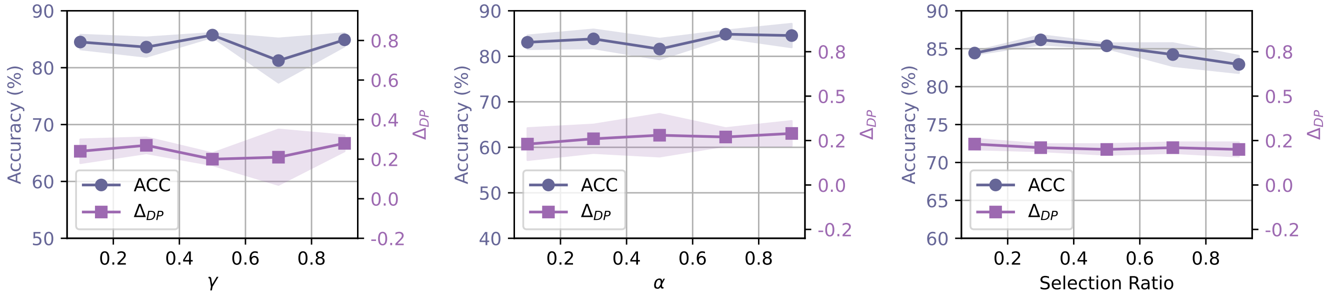

Figure 2: Ablation studies on critical hyperparameters, including γ, α, and selection ratio, on the CelebA dataset with a 40% label bias amount. We use blue to denote accuracy (left axis) and purple to denote fairness violation (right axis).

Dataset	ACC	Uniform	RHO-LOSS	Ours
LFW+a(0.2)	70%	41	32	**21**
	80%	73	65	**55**
LFW+a(0.4)	70%	77	69	**23**
	80%	-	87	**78**
CelebA(0.2)	70%	65	33	**25**
	80%	117	57	**42**
CelebA(0.4)	70%	94	46	**29**
	80%	125	-	**85**

Table 4: Epochs required to reach target test accuracy.

Related Work

Fair Learning with Label Bias Fairness remains a critical and essential concern in real-world applications. A primary source of unfairness is label bias, which is typically modeled as fair (clean) labels being systematically flipped for individuals from certain demographic groups (Wick, panda, and Tristan 2019), i.e., if S is a sensitive attribute and Z a fair label, $P(Y = i \mid Z = j, S)$ occurs with $i \neq j$ and $i, j \in 0, 1$. A growing number of studies are exploring fair learning in settings with noisy labels to address this issue. For example, Wang, Liu, and Levy (2021) applied group-dependent label noise and derived fairness constraints on corrupted data. Jiang and Nachum (2020) propose a reweighting method to correct instances affected by label bias. Building on these, Dai (2020) presents a framework to understand the combined effects of label bias and data distribution shifts in the context of fairness from a fundamental perspective. These works share a common framework that considers the noise rate of labels to improve the robustness of fair learning methods. However, these approaches involve modifying models and intervening in the training process, which limits their flexibility for complex systems and large-scale datasets. To overcome this limitation, we adopt a data selection framework that enhances both flexibility and efficiency, making it suitable for practical applications.

Data Selection Methods As data sizes increase, using all available data for training becomes inefficient. Therefore, data selection methods have been developed to selectively train on only the most useful data, thereby enhancing efficiency and reducing computational costs. Previous approaches like curriculum learning (Bengio et al. 2009) typically choose data points from easy to hard. This method can lead to redundancy, as once such data points are learned, they should not be learned again. Other methods select data based on high training loss or high prediction uncertainty (Loshchilov and Hutter 2017; Kawaguchi and Lu 2019; Jiang et al. 2019; Coleman et al. 2019; Loshchilov and Hutter 2015). A common issue with these approaches is their tendency to choose outliers or noisy points when focusing on high loss or uncertainty. To address this issue, new selection methods that assess the data's impact on generalization loss, derived from a holdout set (Killamsetty et al. 2020; Mindermann et al. 2022), have been developed. Addressing the problem of less principled approximations in these methods and circumventing the need for a holdout set, Deng, Cui, and Zhu (2023) implemented a Bayesian approach, enhancing the validity of approximations and eliminating the reliance on a holdout set through the use of a zero-shot predictor. Inspired by this work, we also utilize the peer prediction mechanism to ensure the fairness of the zero-shot predictor.

Conclusions

This paper addresses label bias in fairness for large-scale datasets by proposing a fair data selection strategy that aligns the RHO-LOSS criterion with a fair distribution and uses a zero-shot predictor to eliminate the need for a clean holdout set. The approach enhances model fairness and accuracy while ensuring the selected data better represents a fair distribution through peer prediction mechanism, suitable for training robust models across various real-world applications. The proposed selection method is constructed from three aspects: (1) deriving a tractable selection function to pick data less affected by label bias, (2) eliminating the need for an additional holdout set previously required for validation, and (3) incorporating a peer prediction mechanism to ensure the fairness of the validation model. Experiments demonstrate the method's effectiveness in mitigating label bias, achieving faster convergence, and higher test accuracy compared to alternatives. While this work primarily focuses on fairness within the defined scope, future research will explore more complex settings, including out-of-distribution (OOD) scenarios and the role of confounders in fairness.

Acknowledgments

We sincerely thank the reviewers for their careful reading of the manuscript and their valuable suggestions. This work was supported by the NSFC Project (No. 62106121), the MOE Project of Key Research Institute of Humanities and Social Sciences (22JJD110001), the fundamental research funds for the central universities, and the research funds of Renmin University of China (24XNKJ13).

References

Barocas, S.; Hardt, M.; and Narayanan, A. 2018. Fairness and Machine Learning Limitations and Opportunities.

Bengio, Y.; Louradour, J.; Collobert, R.; and Weston, J. 2009. Curriculum Learning. In *Proceedings of the 26th Annual International Conference on Machine Learning*, ICML '09, 41–48. New York, NY, USA: Association for Computing Machinery. ISBN 9781605585161.

Bertrand, M.; and Mullainathan, S. 2004. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, 94(4): 991–1013.

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.

Cheng, H.; Zhu, Z.; Li, X.; Gong, Y.; Sun, X.; and Liu, Y. 2021. Learning with Instance-Dependent Label Noise: A Sample Sieve Approach. In *International Conference on Learning Representations*.

Chowdhury, S. B. R.; and Chaturvedi, S. 2023. Sustaining fairness via incremental learning. In *Proceedings of the Thirty-Seventh AAAI Conference on Artificial Intelligence and Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence and Thirteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'23/IAAI'23/EAAI'23. AAAI Press. ISBN 978-1-57735-880-0.

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.

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

Dasgupta, A.; and Ghosh, A. 2013. Crowdsourced judgement elicitation with endogenous proficiency. In *Proceedings of the 22nd International Conference on World Wide Web*, WWW '13, 319–330. New York, NY, USA: Association for Computing Machinery. ISBN 9781450320351.

Deng, Z.; Cui, P.; and Zhu, J. 2023. Towards Accelerated Model Training via Bayesian Data Selection. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; Uszkoreit, J.; and Houlsby, N. 2021. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale. In *International Conference on Learning Representations*.

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

Freund, Y.; Seung, H. S.; Shamir, E.; and Tishby, N. 1997. Selective Sampling Using the Query by Committee Algorithm. *Mach. Learn.*, 28(2–3): 133–168.

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

He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep Residual Learning for Image Recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, 770–778. IEEE Computer Society.

Huang, G.; Liu, Z.; van der Maaten, L.; and Weinberger, K. Q. 2017. Densely Connected Convolutional Networks. In *CVPR*, 2261–2269. IEEE Computer Society. ISBN 978-1-5386-0457-1.

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.

Jiang, Z.; Han, X.; Jin, H.; Wang, G.; Chen, R.; Zou, N.; and Hu, X. 2023. Chasing Fairness Under Distribution Shift: A Model Weight Perturbation Approach. In *Thirty-seventh Conference on Neural Information Processing Systems*.

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

Katharopoulos, A.; and Fleuret, F. 2018. Not All Samples Are Created Equal: Deep Learning with Importance Sampling. In Dy, J.; and Krause, A., eds., *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 2525–2534. PMLR.

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, S.-E.; Paik, H. Y.; Yoon, H.; Lee, J.; Kim, N.; and Sung, M.-K. 2015. Sex- and gender-specific disparities in colorectal cancer risk. *World journal of gastroenterology : WJG*, 21: 5167–5175.

Konstantinov, N.; and Lampert, C. H. 2022. Fairness-Aware PAC Learning from Corrupted Data. *Journal of Machine Learning Research*, 23(160): 1–60.

Lin, Z. J.; Jung, J.; Goel, S.; and Skeem, J. 2020. The limits of human predictions of recidivism. *Science Advances*, 6(7): eaaz0652.

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.

Liu, Z.; Luo, P.; Wang, X.; and Tang, X. 2015. Deep Learning Face Attributes in the Wild. In *2015 IEEE International Conference on Computer Vision (ICCV)*, 3730–3738.

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.

Moore, J. S. 1998. An Expert System Approach to Graduate School Admission Decisions and Academic Performance Prediction. *Omega*, 26(5): 659–670.

Mukerjee, A.; Biswas, R.; Kalyanmoy, Y.; Amrit, D.; and Mathur, P. 2002. Multi-objective Evolutionary Algorithms for the Risk-return Trade-off in Bank Loan Management. *International Transactions in Operational Research*, 9.

Schrouff, J.; Harris, N.; Koyejo, O. O.; Alabdulmohsin, I.; Schnider, E.; Opsahl-Ong, K.; Brown, A.; Roy, S.; Mincu, D.; Chen, C.; Dieng, A.; Liu, Y.; Natarajan, V.; Karthikesalingam, A.; Heller, K. A.; Chiappa, S.; and D'Amour, A. 2022. Diagnosing failures of fairness transfer across distribution shift in real-world medical settings. In Oh, A. H.; Agarwal, A.; Belgrave, D.; and Cho, K., eds., *Advances in Neural Information Processing Systems*.

Seung, H. S.; Opper, M.; and Sompolinsky, H. 1992. Query by committee. In *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, COLT '92, 287–294. New York, NY, USA: Association for Computing Machinery. ISBN 089791497X.

Shnayder, V.; Agarwal, A.; Frongillo, R.; and Parkes, D. C. 2016. Informed Truthfulness in Multi-Task Peer Prediction. In *Proceedings of the 2016 ACM Conference on Economics and Computation*, EC '16, 179–196. New York, NY, USA: Association for Computing Machinery. ISBN 9781450339360.

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. In Wallach, H.; Larochelle, H.; Beygelzimer, A.; d'Alché-Buc, F.; Fox, E.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 32, 8783–8792. Curran Associates, Inc.

Wolf, L.; Hassner, T.; and Taigman, Y. 2011. Effective Unconstrained Face Recognition by Combining Multiple Descriptors and Learned Background Statistics. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10): 1978–1990.

Yeh, I.; and Lien, C.-H. 2009. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients. *Expert Systems with Applications*, 36: 2473–2480.

Zhang, Y.; Zhou, F.; Li, Z.; Wang, Y.; and Chen, F. 2021. Bias-tolerant Fair Classification. In Balasubramanian, V. N.; and Tsang, I., eds., *Proceedings of The 13th Asian Conference on Machine Learning*, volume 157 of *Proceedings of Machine Learning Research*, 840–855. PMLR.

Zhang, Y.; Zhou, F.; Li, Z.; Wang, Y.; and Chen, F. 2023. Fair Representation Learning with Unreliable Labels. In Ruiz, F.; Dy, J.; and van de Meent, J.-W., eds., *Proceedings of The 26th International Conference on Artificial Intelligence and Statistics*, volume 206 of *Proceedings of Machine Learning Research*, 4655–4667. PMLR.

Reproducibility Checklist

1. This paper:

- Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes)
- Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes)
- Provides well marked pedagogical references for less-familiare readers to gain background necessary to replicate the paper (yes)

2. Does this paper make theoretical contributions? (yes)

2.1 If yes, please complete the list below.

- All assumptions and restrictions are stated clearly and formally. (yes)
- All novel claims are stated formally (e.g., in theorem statements). (yes)
- Proofs of all novel claims are included. (yes)
- Proof sketches or intuitions are given for complex and/or novel results. (yes)
- Appropriate citations to theoretical tools used are given. (yes)
- All theoretical claims are demonstrated empirically to hold. (yes)
- All experimental code used to eliminate or disprove claims is included. (NA)

3. Does this paper rely on one or more datasets? (yes)

3.1 If yes, please complete the list below.

- A motivation is given for why the experiments are conducted on the selected datasets (NA)
- All novel datasets introduced in this paper are included in a data appendix. (NA)
- All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. (NA)
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations. (yes)
- All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available. (yes)
- All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing. (NA)

4. Does this paper include computational experiments? (yes)

If yes, please complete the list below.

- Any code required for pre-processing data is included in the appendix. (yes).
- All source code required for conducting and analyzing the experiments is included in a code appendix. (yes)
- All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes. (yes)
- All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes)
- If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results. (yes)
- This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks. (yes)
- This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics. (yes)
- This paper states the number of algorithm runs used to compute each reported result. (yes)
- Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information. (yes)
- The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank). (no)
- This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments. (yes)
- This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting. (partial)

Refining Selection Function

In this section, we detail the derivation of equations related to the new optimization goal, which focuses on refining the selection function to consider fair distribution.

Derivation of the optimization problem in Equation 5

$$
\begin{aligned}
&\max_{(x,y)\in B_{t+1}} \frac{1}{m}\sum_{i=1}^{m}\sum_{s\in[S]} \bar{p}(s)[\log p(y_i^* \mid x_i^*, \mathcal{D}_t \cup (x,y), s)] \\
\Leftrightarrow &\max_{(x,y)\in B_{t+1}} \sum_{s\in[S]} \bar{p}(s)[\log p(\mathcal{D}^* \mid \mathcal{D}_t \cup (x,y), s)].
\end{aligned}
$$

Derivation of the Bayes' rule in Equation 6

$$
\begin{aligned}
p(\mathcal{D}^* \mid \mathcal{D}_t \cup (x,y), s) &= \frac{p(y \mid x, \mathcal{D}^*, \mathcal{D}_t, s)}{p(y \mid x, \mathcal{D}_t, s)} \cdot p(\mathcal{D}^* \mid \mathcal{D}_t, x, s) \\
&= \frac{p(y \mid x, \mathcal{D}^*, \mathcal{D}_t, s)}{p(y \mid x, \mathcal{D}_t, s)} \cdot p(\mathcal{D}^* \mid \mathcal{D}_t).
\end{aligned}
$$

Derivation of the lower bound in Equation 8

$$
\begin{aligned}
&\log p(y \mid x, \mathcal{D}^*, \mathcal{D}_t, s) \\
=&\log \int p(\theta \mid \mathcal{D}^*, \mathcal{D}_t) p(y \mid x, \theta, s) d\theta \\
=&\log \int \frac{p(\mathcal{D}_t \mid \theta) p(\theta \mid \mathcal{D}^*)}{p(\mathcal{D}_t \mid \mathcal{D}^*)} p(y \mid x, \theta, s) d\theta \\
=&\log \int p(\mathcal{D}_t \mid \theta) p(\theta \mid \mathcal{D}^*) p(y \mid x, \theta, s) d\theta - \log p(\mathcal{D}_t \mid \mathcal{D}^*) \\
\geq& \int p(\theta \mid \mathcal{D}^*) \log[p(\mathcal{D}_t \mid \theta) p(y \mid x, \theta, s)] d\theta - \log p(\mathcal{D}_t \mid \mathcal{D}^*) \\
=&\mathbb{E}_{p(\theta|\mathcal{D}^*)} \log p(\mathcal{D}_t \mid \theta) + \mathbb{E}_{p(\theta|\mathcal{D}^*)} \log p(y \mid x, \theta, s) - \log p(\mathcal{D}_t \mid \mathcal{D}^*)
\end{aligned}
$$

Drop the terms irrelevant to (x, y), we get:

$$
\log p(y \mid x, \mathcal{D}^*, \mathcal{D}_t, s) \geq \mathbb{E}_{p(\theta|\mathcal{D}^*)} \log p(y \mid x, \theta, s).
$$

Derivation of the equivalent expectation version of the peer loss in Equation 12

We denote $\mathcal{D}_s$ as a shorthand for $\mathcal{D}_{S=s}$ for $s \in [S]$. For each x_i belonging to the demographic group $S = s$, let $|\mathcal{D}_s| = N_s$, and $[I]$ be the set of sample index in $\mathcal{D}_s$. By taking the expectation over (X_{i_1}, Y_{i_2}), we have:

$$
\begin{aligned}
&\frac{1}{N_s}\sum_{i\in[I]]} \mathbb{E}_{X_{i_1}|\mathcal{D}_s, Y_{i_2}|\mathcal{D}_{s'}} \Big[L[y_i, \tilde{f}(x_i), s] - \gamma L[Y_{i_2}, \tilde{f}(X_{i_1}), s]\Big] \\
=&\frac{1}{N_s}\sum_{i\in[I]} \left[L[y_i, \tilde{f}(x_i), s] - \sum_{i'\in[I], i'\neq i} P(X_{i_1} = x_{i'} \mid S = s) \mathbb{E}_{Y|\mathcal{D}_{s'}} \gamma L[Y, \tilde{f}(x_{i'}), s]\right] \\
=&\frac{1}{N_s}\sum_{i\in[I]} \left[L[y_i, \tilde{f}(x_i), s] - \sum_{i'\in[I], i'\neq i} \frac{1}{N_s} \mathbb{E}_{Y|\mathcal{D}_{s'}} \gamma L[Y, \tilde{f}(x_{i'}), s]\right] \\
=&\frac{1}{N_s}\sum_{i\in[I]} \Big[L[y_i, \tilde{f}(x_i), s] - \mathbb{E}_{Y|\mathcal{D}_{s'}} \gamma L[Y, \tilde{f}(x_i), s]\Big]
\end{aligned}
$$

So, the result of the fairness-regularized loss on the zero-shot predictor will be:

$$
L[y, \tilde{f}(x), s] - \mathbb{E}_{Y|\mathcal{D}_{s'}} \gamma L[Y, \tilde{f}(x), s]
$$

Is is worth noting that the last term is taking the expectation of Y over s', while x is drawn from s, we designed this intentionally: in the peer prediction mechanism, we sample x_{i_1} from D_s and y_{i_2} from $D_{s'}$.

Decompose the Loss

In this section, we demonstrate the effectiveness of the loss proposed in Equation 12 by decomposing its expected version.

$$\underbrace{L[y, \tilde{f}(x), s]}_{\text{A}} - \underbrace{\gamma \mathbb{E}_{Y|\mathcal{D}_{s'}}[L[Y, \tilde{f}(x), s]]}_{\text{B}}. \tag{15}$$

Decomposition of A

To clarify the notation, we use $\tilde{D}$ to denote the observed distribution, which contains label bias, and D to denote the underlying clean fair distribution. In the context of label bias, the flip does not depend on X; that is, $P(Y = j \mid Z = i, S = s, X) = P(Y = j \mid Z = i, S = s)$. For brevity, we refer to this simply as T^s_{ij}. We use $\theta^+_s = P(Y = 1 \mid Z = 0, S = s)$ and $\theta^-_s = P(Y = 0 \mid Z = 1, S = s)$ to denote the specific flip rate. The following formulation incorporates the independence between $P(S)$ and $P(Z)$. For Y, Z and S, we only consider the binary case and we use K to denote the number of class ($[K] = \{0, 1\}$ in our case). Part A in Eq. (15) can be decomposed into:

$$\begin{aligned}
&\mathbb{E}_{\tilde{D}}\left[L[Y, \tilde{f}(X), S]\right] \\
&= \sum_{j\in[K]} \sum_{i\in[K]} \sum_{s\in[S]} P(Z = i)P(S = s)\mathbb{E}_{D|Z=i,S=s}\left[T^s_{ij} L[j, \tilde{f}(X), s]\right] \\
&= \sum_{j\in[K]} \sum_{i\in[K]} \sum_{s\in[S]} P(Z = i)P(S = s)\mathbb{E}_{D|Z=i,S=s}T^s_{ij}\mathbb{E}_{D|Z=i,S=s}\left[L[j, \tilde{f}(X), s]\right] \\
&+ \sum_{j\in[K]} \sum_{i\in[K]} \sum_{s\in[S]} P(Z = i)P(S = s)\mathrm{Cov}_{D|Z=i,S=s}\left(T^s_{ij}, L[j, \tilde{f}(X), s]\right)
\end{aligned} \tag{16}$$

The first term can be further decomposed into:

$$\begin{aligned}
&\underbrace{\sum_{s\in[S]} P(S = s)\Big[\mathbb{E}_D(1 - \theta^-_s - \theta^+_s)\mathbb{E}_{D|S=s}L[Z, \tilde{f}(X), s]}_{\text{Term 1}} \\
&+ \underbrace{P(Z = 1)\mathbb{E}_{D|Z=1,S=s}\theta^+_s\mathbb{E}_{D|Z=1,S=s}L[1, \tilde{f}(X), s] + P(Z = 0)\mathbb{E}_{D|Z=0,S=s}\theta^-_s\mathbb{E}_{D|Z=0,S=s}L[0, \tilde{f}(X), s]}_{\text{Term 2}} \\
&+ \underbrace{\sum_{j\in[K],i\in[K],i\neq j} P(Z = i)\mathbb{E}_{D|Z=i,S=s}T^s_{ij}\mathbb{E}_{D|Z=i,S=s}L[j, \tilde{f}(X), s]\Big]}_{\text{Term 3}}
\end{aligned} \tag{17}$$

The second term will be:

$$\begin{aligned}
&\sum_{s\in[S]} P(S = s)\left[\sum_{j\in[K]} P(Z = j)\mathbb{E}_{D|Z=j,S=s}\big(T^s_{jj} - \bar{T}^s_{jj}\big)\left(L[j, \tilde{f}(X), s] - \mathbb{E}_{D|Z=j,S=s}[L[j, \tilde{f}(X), s]\right)\right. \\
&\left. + \sum_{j\in[K]} \sum_{i\in[K],i\neq j} P(Z = i)\mathbb{E}_{D|Z=i,S=s}\left(T^s_{ij} - \bar{T}^s_{ij}\right) \times \left(L[j, \tilde{f}(X), s] - \mathbb{E}_{D|Z=i,S=s}L[j, \tilde{f}(X), s]\right)\right],
\end{aligned} \tag{18}$$

where $\bar{T}^s_{ij} = \mathbb{E}_{D|Z=i,S=s}T^s_{ij}$. Then, combine Eq. (18) with Term 3, we get:

$$\begin{aligned}
&\sum_{s\in[S]} P(S = s)\Big[P(Z = 1)\mathbb{E}_{D|Z=1,S=s}\left(1 - \theta^-_s - \mathbb{E}_{D|Z=1,S=s}(1 - \theta^-_s)\right)L(1, \tilde{f}(x), s] \\
&+ P(Z = 0)\mathbb{E}_{D|Z=0,S=s}\left(1 - \theta^+_s - \mathbb{E}_{D|Z=0,S=s}(1 - \theta^+_s)\right)L[0, \tilde{f}(X), s] \\
&+ P(Z = 0)\mathbb{E}_{D|Z=0,S=s}(\theta^+_s\ell[1, \tilde{f}(X)]) + P(Z = 1)\mathbb{E}_{D|Z=1,S=s}(\theta^-_s L[0, \tilde{f}(X), s])\Big].
\end{aligned} \tag{19}$$

Integrating Eq. (19) with Term 1 and Term 2, and considering that the fair and clean distribution D remains unaffected by the sensitive variable S, we obtain:

$$\begin{aligned}
&\mathbb{E}_{\tilde{D}}[L[Y, \tilde{f}(X), S]] \\
&= \mathbb{E}_D\Delta_s L[Z, \tilde{f}(X), S] + \sum_{s\in[S]} \sum_{j\in[K]} \sum_{i\in[K]} P(S = s)P(Z = i)\mathbb{E}_{D|Z=i,S=s}U^s_{ij}L[j, \tilde{f}(X), s]
\end{aligned} \tag{20}$$

where

$$U^s_{ij} = \begin{cases} T^s_{ij} \text{ if } i \neq j, \\ T^s_{jj} - \bar{T}^s_{jj} \text{ if } i = j. \end{cases}$$

Decomposition of B

$$
\begin{aligned}
&\mathbb{E}_{\tilde{D}}\left[-\gamma \mathbb{E}_{Y|\tilde{D},S=s} L\left[Y, \tilde{f}(X), S=s'\right]\right] \\
&= -\gamma \int_{\tilde{D}} \Bigg[P(\tilde{D} \mid S=0) P(S=0) \sum_{j\in[K]} P(Y=j \mid \tilde{D}, S=0) L\left[Y=j, \tilde{f}(X), S=1\right] \\
&\qquad + P(\tilde{D} \mid S=1) P(S=1) \sum_{j\in[K]} P(Y=j \mid \tilde{D}, S=0) L\left[Y=j, \tilde{f}(X), S=1\right] \Bigg] \\
&= -\gamma \sum_{j\in[K]} \Big[P(Y=j, S=0) \mathbb{E}_{D_{X|S=1}} L[j, \tilde{f}(X), S=1] + P(Y=j, S=1) \mathbb{E}_{D_{X|S=0}} L[j, \tilde{f}(X), S=0] \Big] \\
&= -\gamma \sum_{j\in[K]} \Big[P(Y=j, S=0) \Big(\mathbb{E}_{D_{X|S=1}} L[j, \tilde{f}(X), S=1] - \mathbb{E}_{D_{X|S=0}} L[j, \tilde{f}(X), S=0] \Big) \\
&\qquad + P(Y=j, S=1) \Big(\mathbb{E}_{D_{X|S=0}} L[j, \tilde{f}(X), S=0] - \mathbb{E}_{D_{X|S=1}} L[j, \tilde{f}(X), S=1] \Big) \\
&\qquad + P(Y=j, S=1) \mathbb{E}_{D_{X|S=1}} L[j, \tilde{f}(X), S=1] + P(Y=j, S=0) \mathbb{E}_{D_{X|S=0}} L[j, \tilde{f}(X), S=0] \Big] \\
&= -\gamma \sum_{j\in[K]} d_j^s \Big(\mathbb{E}_{D_{X|S=1}} L[j, \tilde{f}(X), S=1] - \mathbb{E}_{D_{X|S=0}} L[j, \tilde{f}(X), S=0] \Big) \\
&\qquad - \gamma \sum_{j\in[K]} \sum_{S} P(Y=j, S=s) \mathbb{E}_{D_{X|S=s}} L[j, \tilde{f}(X), s],
\end{aligned}
\tag{21}
$$

where $d_j^s = (P(Y=j, S=0) - P(Y=j, S=1))$, which measures the discrepancy in the probability of observed label j between two demographic groups. For the second term in Eq. (21), we can further expand it as:

$$
\begin{aligned}
& - \gamma \sum_{j\in[K]} \sum_{s\in[S]} P(Y=j, S=s) \mathbb{E}_{D_{X|S=s}} L[j, \tilde{f}(X), s] \\
&= - \sum_{j\in[K]} \sum_{i\in[K]} \sum_{s\in[S]} P(Z=i) \mathbb{E}_{D|Z=i,S=s} \left[\gamma P(Y=j, S=s) L[j, \tilde{f}(X), s] \right]
\end{aligned}
\tag{22}
$$

Combine Eq. (22) with the second term in Eq. (20), we get:

$$
\sum_{j\in[K]} \sum_{i\in[K]} \sum_{s\in[S]} P(S=s) P(Z=i) \mathbb{E}_{D|Z=i,S=s} [(U_{ij}^s - \gamma P(Y=j \mid S=s)) L[j, \tilde{f}(X), s]] \tag{23}
$$

Finally, the original expected loss is decomposed into the following components:

$$
\begin{aligned}
& \mathbb{E}_{\tilde{D}} \left[L[Z, \tilde{f}(X), S] - \gamma \mathbb{E}_{Z|\mathcal{D}_{s'}} L[Z, \tilde{f}(x), s] \right] \\
&= \underbrace{\mathbb{E}_D \Delta_s L[Z, \tilde{f}(X), S]}_{\text{fair model}} \\
&\quad + \underbrace{\sum_{j\in[K]} \sum_{i\in[K]} \sum_{s\in[S]} P(S=s) P(Z=i) \mathbb{E}_{D|Z=i,S=s} [(U_{ij}^s - \gamma P(Y=j \mid S=s)) L[j, \tilde{f}(X), s]]}_{\text{Penalty on noisy loss}} \\
&\quad \underbrace{-\gamma \sum_{j\in[K]} p_j^s \left[\mathbb{E}_{D_{X|S=1}} L[j, \tilde{f}(X)] - \mathbb{E}_{D_{X|S=0}} L[j, \tilde{f}(X)] \right],}_{\text{Penalty for the disagreement between demographic groups}}
\end{aligned}
\tag{24}
$$

where

$$
\begin{aligned}
\Delta_s &= 1 - \theta_s^- - \theta_s^+ \\
p_j^s &= P(Y=j, S=0) - P(Y=j, S=1) \\
U_{ij}^s &= \begin{cases} T_{ij}^s \text{ if } i \neq j, \\ T_{jj}^s - \mathbb{E}_{D|Y=j,S=s} T_{jj}^s \text{ if } i = j. \end{cases}
\end{aligned}
$$