Fair and Robust Classification Under Sample Selection Bias

Wei Du
University of Arkansas
Fayetteville, AR, USA
wd005@uark.edu

Xintao Wu
University of Arkansas
Fayetteville, AR, USA
xintaowu@uark.edu

ABSTRACT

To address the sample selection bias between the training and test data, previous research works focus on reweighing biased training data to match the test data and then building classification models on the reweighed training data. However, how to achieve fairness in the built classification models is under-explored. In this paper, we propose a framework for robust and fair learning under sample selection bias. Our framework adopts the reweighing estimation approach for bias correction and the minimax robust estimation approach for achieving robustness on prediction accuracy. Moreover, during the minimax optimization, the fairness is achieved under the worst case, which guarantees the model's fairness on test data. We further develop two algorithms to handle sample selection bias when test data is both available and unavailable.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning algorithms**; • **Applied computing** → **Law, social and behavioral sciences**; • **Theory of computation** → **Sample complexity and generalization bounds**.

KEYWORDS

sample selection bias, algorithmic fairness, robustness

ACM Reference Format:
Wei Du and Xintao Wu. 2021. Fair and Robust Classification Under Sample Selection Bias. In *Proceedings of the 30th ACM International Conference on Information and Knowledge Management (CIKM '21), November 1–5, 2021, Virtual Event, QLD, Australia.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3459637.3482104

1 INTRODUCTION

Traditional supervised machine learning assumes that training data and test data are independently and identically distributed (iid), i.e., each example t with pairs of feature input x and label y drawn from the same distribution $Q = P(x, y)$. The conditional label distribution, $P(y|x)$, is estimated as $\hat{P}(y|x)$ (aka, a classifier $f(x)$) from the given training dataset $\mathcal{D}_s$. Similarly, in the fair machine learning, we aim to learn a fair classifier $f(x, a)$ from the training dataset drawn from $Q = P(x, a, y)$ where a is a protected attribute such as gender or race. However, when the distributions on training and test data sets do not match, we are facing sample selection bias or covariate shift. The classifier f simply learned from the training dataset is vulnerable to sample selection bias and will incur more accuracy loss over test data. Moreover, the fair classifier trained with the biased data cannot guarantee fairness over test data. This is a serious concern when it is critical and imperative to achieve fairness in many applications.

In this paper, we develop a framework for robust and fair learning under sample selection bias. We embrace the uncertainty incurred by sample selection bias by producing predictions that are both fair and robust in test data. Our framework adopts the reweighing estimation approach for bias correction and the minimax robust estimation approach for achieving robustness on prediction accuracy. Moreover, during the minimax optimization, the fairness is achieved under the worst case, which guarantees the model's fairness on test data. To address the intractable issue, we approximate the fairness constraint using the boundary fairness and combine into the classifier's loss function as a penalty. The modified loss function is minimized in view of the most adverse distribution within a Wasserstein ball centered at the empirical distribution of the training data. We present two algorithms for the scenarios where the unlabeled test dataset $\mathcal{D}$ is either available or unavailable.

Related Work. Robust classification under covariate shift has been studied recently. For example, Wen et al. [15] consider covariate shift between the training and test data and apply Gaussian kernel functions to reweigh the training examples and correct the shift. Taskesen et al. [14] study fairness from the distributionally robust perspective and assume the unknown true test distribution is contained in a Wasserstein ball centered at the empirical distribution on the observed training data. However the approach robustifies the distribution at the individual data level and overlooks the overall distribution. Rezaei et al. [12] propose the use of ambiguity set to derive the fair classifier based on the principles of distributional robustness. The proposed approach incorporates fairness criteria into a worst case logarithmic loss minimization but ignores the distribution shift. There has been research on studying fairness issues in domain adaptation, transfer learning, and federated learning [3, 5, 8, 10, 13]. However, they do not address the robustness in learning under sample selection bias.

2 FAIR CLASSIFIER UNDER SELECTION BIAS

We first define notations throughout this paper. Let X denote the feature space, A the protected attribute, and Y the label set. Let Q denote the true distribution over $X \times A \times Y$ according to which test samples $t = (x, a, y)$ are drawn. For simplicity, we assume both y and a are binary where $y = 1$ (0) denotes the favorable (unfavorable) decision and $a = 1$ (0) denotes the majority (minority) group. Under the sample selection bias setting, the learning algorithm receives a training dataset $\mathcal{D}_s$ of $N_{\mathcal{D}_s}$ labeled points $t_1, \cdots, t_{N_{\mathcal{D}_s}}$ drawn

CIKM '21, November 1–5, 2021, Virtual Event, QLD, Australia

ACM ISBN 978-1-4503-8446-9/21/11…$15.00
https://doi.org/10.1145/3459637.3482104

according to a biased distribution Q_s over $X \times A \times Y$. This sample bias can be represented by a random binary variable s that controls the selection of points, i.e., $s = 1$ for selected and $s = 0$ otherwise.

Problem Formulation With the observed $\mathcal{D}_s$, how to construct a fair classifier f that minimizes $\mathbb{E}_{(x,a,y)\in Q}[l(f(x,a),y)]$ subject to $|RD(Q)| \leq \tau$ where l is the loss function, $RD(Q)$ is the risk difference over distribution Q, i.e., $RD(Q) = |P_Q(\hat{y} = 1|a = 1) - P_Q(\hat{y} = 1|a = 0)|$, and $\tau \in [0, 1]$ is a fairness threshold.

The probability of drawing $t = (x, a, y)$ according to the true but unobserved distribution Q is related to the observed distribution Q_s. By definition of the random selection variable s, the observed biased distribution Q_s can be expressed by $P_{Q_s}(t) = P_Q(t|s = 1)$ or $P_{Q_s}(x, a, y) = P_Q(x, a, y|s = 1)$. Assuming $P(s = 1|x, a) \neq 0$ for all $t \in X \times A \times Y$, by the Bayes formula, we have $P_Q(t) = \frac{P(t|s=1)P(s=1)}{P(s=1|x,a)} = \frac{P(s=1)}{P(s=1|x,a)}P_{Q_s}(t)$. Hence, if we define and construct the new distribution $\hat{Q}_s$ as $\frac{P(s=1)}{P(s=1|t)}Q_s$, i.e., $P_{\hat{Q}_s}(x, a, y) = \frac{P(s=1)}{P(s=1|x,a,y)}P_{Q_s}(x, a, y)$, we have:

$$\begin{aligned}
\mathbb{E}_{(x,a,y)\in\hat{Q}_s}[l(f(x,a),y)] &= \sum_{x,a,y} l(f(x,a),y))P_{\hat{Q}_s}(x,a,y) \\
&= \sum_{x,a,y} l(f(x,a),y))\frac{P(s=1)}{P(s=1|x,a,y)}P_{Q_s}(x,a,y) \\
&= \sum_{x,a,y} l(f(x,a),y))\frac{P(s=1)}{P(s=1|x,a,y)}\frac{P_{Q(s=1|x,a,y)}P_Q(x,a,y)}{P_Q(s=1)} \\
&= \sum_{x,a,y} l(f(x,a),y))P_Q(x,a,y) = \mathbb{E}_{(x,a,y)\in Q}[l(f(x,a),y)]
\end{aligned}$$

Similarly we have $RD(\hat{Q}_s) = RD(Q)$. Equivalently, if we define and construct a modified training dataset $\hat{\mathcal{D}}_s$ by introducing a weight $\frac{P(s=1)}{P(s=1|x,a,y)}$ to each record $t \in \mathcal{D}_s$, we can approximate $\mathbb{E}_{(x,a,y)\in\hat{Q}_s}[l(f(x,a),y)]$ using $\mathbb{E}_{(x,a,y)\in\hat{\mathcal{D}}_s}[l(f(x,a),y)]$, which can be expressed as $\frac{1}{N_{\mathcal{D}_s}}\sum_{i=1}^{N_{\mathcal{D}_s}}\frac{P(s=1)}{P(s=1|t)}l(f(x_i,a_i),y_i)$.

Theorem 1. *Under sample selection bias, the classifier f that minimizes $\mathbb{E}_{(x,a,y,s)\in\hat{\mathcal{D}}_s}[l(f(x,a),y)]$ subject to $RD(\hat{\mathcal{D}}_s) \leq \tau$ is a fair classifier.*

The sample selection bias causes training data to be selected non-uniformly from the population to be modeled. In this paper, we assume $P(s = 1|x, a, y) = p(s = 1|x, a)$. Then the minimization of the loss on $\hat{\mathcal{D}}_s$ subject to the fairness constraint is as:

$$\begin{aligned}
&\min_{\mathbf{w}\in\mathcal{W}} L(\mathbf{w}) = \frac{1}{N_{\mathcal{D}_s}}\sum_{i=1}^{N_{\mathcal{D}_s}}\frac{P(s=1)}{P(s=1|x_i,a_i)}l(f(x_i,a_i),y_i) \\
&\text{subject to} \qquad RD(\hat{D}_s) \leq \tau
\end{aligned} \tag{1}$$

where $\mathbf{w}$ are the parameters of the classifier f, and $RD(\hat{D}_s)$ is

$$\left|\frac{\sum \mathbf{1}_{(x_i,a_i)\in\mathcal{D}_s^{11}}\frac{P(s=1)}{P(s=1|x_i,a_i)}}{\sum \mathbf{1}_{(x_i,a_i)\in\mathcal{D}_s^{1}}\frac{P(s=1)}{P(s=1|x_i,a_i)}} - \frac{\sum \mathbf{1}_{(x_i,a_i)\in\mathcal{D}_s^{10}}\frac{P(s=1)}{P(s=1|x_i,a_i)}}{\sum \mathbf{1}_{(x_i,a_i)\in\mathcal{D}_s^{0}}\frac{P(s=1)}{P(s=1|x_i,a_i)}}\right|. \tag{2}$$

In the above, $\mathbf{1}_{[.]}$ is an indicator function, $\mathcal{D}_s^{ij} = \{(x_i, a_i)|\hat{Y} = i, A = j\}$ where $i, j \in \{0, 1\}$, $\mathcal{D}_s^{\cdot j} = \{(x_i, a_i)|A = j\}$ where $j \in \{0, 1\}$ and $\cdot$ represents $\{0, 1\}$.

3 ROBUST AND FAIR CLASSIFICATION

To obtain the optimal solution of Eq. 1, we need to derive the sample selection probability $P(s = 1|t)$. However, it is rather challenging to get the true value practically because the selection mechanism is usually unknown. We estimate the sample selection probability and use the estimated probability $\hat{P}(s = 1|t)$ as the true $P(s = 1|t)$.

To take the estimation error between $P(s = 1|t)$ and $\hat{P}(s = 1|t)$ into consideration, we adopt the approach of minimax robust minimization [7, 11, 15] which advocates for the worst case of any unknown true sample selection probability. We make an assumption here that the true $P(s = 1|x, a)$ is with the ϵ range of the estimated $\hat{P}(s = 1|x, a)$. Therefore, any value of $P(s = 1|x, a)$ in this ϵ range represents the possible real unknown distribution Q. Following the standard robust optimization approaches and taking the estimation error into consideration, we reformulate Eq. 1:

$$\begin{aligned}
&\min_{\mathbf{w}\in\mathcal{W}}\max_{P(s=1|x_i,a_i)} L(\mathbf{w},P) = \frac{1}{N_{\mathcal{D}_s}}\sum_{i=1}^{N_{\mathcal{D}_s}}\frac{P(s=1)}{P(s=1|x_i,a_i)}l(f(x_i,a_i),y_i) \\
&\text{subject to } |P(s=1|x_i,a_i) - \hat{P}(s=1|x_i,a_i)| \leq \epsilon \\
&\qquad\qquad RD(\hat{D}_s) \leq \tau
\end{aligned} \tag{3}$$

In fact, $P(s = 1)$ is a constant and does not affect the problem formulation and optimization. The robust minimax optimization can be treated as an adversarial game by two players. One player selects $P(s = 1|x_i, a_i)$ within the ϵ range of the estimated $\hat{P}(s = 1|x_i, a_i)$ to maximize the loss of the objective, which can be seen as the worst case of Q. Another player minimizes the worst case loss to find the optimal $\mathbf{w}$. There are two advantages of robust minimax optimization of Eq. 3. First, it takes the worst case induced by the estimation error into consideration, thus the obtained classifier f is robust to any possible Q within the error range of the estimation. Second, during the minimax optimization, the fairness is achieved under the worst case. Therefore, we can guarantee the fairness for any possible Q within the error range of the estimation.

The computation of $RD(\hat{D}_s)$ involves the indicator function, which makes it computationally intractable to reach the optimal solution of Eq. 3. To address the intractable issue, we approximate the fairness constraint using the boundary fairness [16]. We define $C(t, \mathbf{w})$ be the covariance between the sensitive attribute and the signed distance from the non-sensitive attribute vector to the decision boundary. Then we can write the boundary fairness on $\mathcal{D}_s$ as $C_{\mathcal{D}_s}(t, \mathbf{w}) = \frac{1}{N_{\mathcal{D}_s}}\sum_{i=1}^{N_{\mathcal{D}_s}}(a_i - \bar{a})d_{\mathbf{w}(\mathbf{x}_i)}$, where a_i is the sensitive attribute value of t_i, $\bar{a} = \frac{1}{N_{\mathcal{D}_s}}\sum_{i=1}^{N_{\mathcal{D}_s}} a_i$ is the mean value of the sensitive attribute and $d_{\mathbf{w}(\mathbf{x}_i)}$ is the distance to the decision boundary of the classifier f and is formally defined as $d_{\mathbf{w}(\mathbf{x}_i)} = \mathbf{w}^T\mathbf{x}_i$. Similarly we will have the boundary fairness on $\hat{\mathcal{D}}_s$ as:

$$C_{\hat{\mathcal{D}}_s}(t, \mathbf{w}) = \frac{1}{N_{\mathcal{D}_s}}\sum_{i=1}^{N_{\mathcal{D}_s}}(a_i - \bar{a})\frac{P(s=1)}{P(s=1|x_i,a_i)}d_{\mathbf{w}(\mathbf{x}_i)}. \tag{4}$$

We enforce $C_{\hat{\mathcal{D}}_s}(t, \mathbf{w}) \leq \sigma, \sigma \in R^+$ to achieve the fair classification.

3.1 Solving Robust and Fair Optimization

The above optimization of involves two sets of parameters $\mathbf{w}$ and $P(s=1|x_i,a_i)$. According to its minimax formulation, it is preferable to obtain the optimal solution in an iterative manner by optimizing $\mathbf{w}$ and $P(s=1|x_i,a_i)$ alternatively. First, we fix $P(s=1|x_1,a_i)$ and choose to transform the fairness constraint as a penalty term and add to $L(\mathbf{w})$, which can be expressed as:

$$\begin{aligned}\min_{\mathbf{w}\in\mathcal{W}} L(\mathbf{w}) = \frac{1}{N_{\mathcal{D}_s}}\sum_{i=1}^{N_{\mathcal{D}_s}}\frac{P(s=1)}{P(s=1|x_i,a_i)}l(f(x_i,a_i),y_i)\\ +\beta(\frac{1}{N_{\mathcal{D}_s}}\sum_{i=1}^{N_{\mathcal{D}_s}}(a_i-\bar{a})\frac{P(s=1)}{P(s=1|x_i,a_i)}d_{\mathbf{w}(\mathbf{x}_i)}-\sigma)^2\end{aligned} \tag{5}$$

where β is a hyperparameter that controls the utility and fairness trade-off. By the transformation, standard optimization techniques such as stochastic gradient decent can be used to solve Eq. 5. Second, we fix $\mathbf{w}$ and derive the formula only related to $P(s=1|x_i,a_i)$ as:

$$\begin{aligned}&\max_{P(s=1|x_i,a_i)} L(P) = \frac{1}{N_{\mathcal{D}_s}}\sum_{i=1}^{N_{\mathcal{D}_s}}\frac{P(s=1)}{P(s=1|x_i,a_i)}l(f(x_i,a_i),y_i)\\ &\text{subject to } |P(s=1|x_i,a_i)-\hat{P}(s=1|x_i,a_i)|\le\epsilon\\ &\qquad|\frac{1}{N_{\mathcal{D}_s}}\sum_{i=1}^{N_{\mathcal{D}_s}}(a_i-\bar{a})\frac{P(s=1)}{P(s=1|x_i,a_i)}d_{\mathbf{w}(\mathbf{x}_i)}|\le\sigma\end{aligned} \tag{6}$$

The objective of Eq. 6 is a linear combination of $\frac{P(s=1)}{P(s=1|x_i,a_i)}$ that can be treated as one variable. $P(s=1)$ is a constant and does not affect the optimization. For the first constraint, $|P(s=1|x_i,a_i)-\hat{P}(s=1|x_i,a_i)|\le\epsilon$, we can obtain the range of $\frac{P(s=1)}{P(s=1|x_i,a_i)}$ after we estimate the range of each $P(s=1|x_i,a_i)$. The second constraint $|\frac{1}{N_{\mathcal{D}_s}}\sum_{i=1}^{N_{\mathcal{D}_s}}(a_i-\bar{a})\frac{P(s=1)}{P(s=1|x_i,a_i)}d_{\mathbf{w}(\mathbf{x}_i)}|\le\sigma$ in Eq. 6 is linear with respect to $\frac{P(s=1)}{P(s=1|x_i,a_i)}$ when $\mathbf{w}$ is fixed. Therefore, the optimization is a standard linear programming problem and can be solved without any additional relaxation.

3.2 *RFLearn*^{1}: Bias Correction with $\mathcal{D}$

Our previous formulation assumes the estimated $\hat{P}(s=1|x,a)$ is obtained. In this section, we discuss how to estimate the true $P(s=1|x,a)$ by using the unlabeled test data $\mathcal{D}$. For a particular data record t, the ratio between the number of times t in $\mathcal{D}$ and the number of times t in $\mathcal{D}_s$ in terms of (a,x) is an estimation value for $P(s=1|x,a)$. Formally, for $t\in\mathcal{D}$, let $\mathcal{D}^t$ denote the subset of $\mathcal{D}$ containing exactly all the instances of t and $n_t=|\mathcal{D}^t|$. Similarly, let $\mathcal{D}_s^t$ denote the subset of $\mathcal{D}_s$ containing exactly all the instances of t and $m_t=|\mathcal{D}_s^t|$. We then have $\hat{P}(s=1|x,a)=\frac{m_t}{n_t}$.

Lemma 1 [2] Let $\delta>0$, with probability at least $1-\delta$, the following inequality holds for all $t\in\mathcal{D}_s$:

$$|P(s=1|x,a)-\frac{m_t}{n_t}|\le\sqrt{\frac{\ln 2m'+\ln\frac{1}{\delta}}{p_0 N_{\mathcal{D}}}} \tag{7}$$

where m' denotes the number of unique points in $\mathcal{D}_s$ and $p_0=\min_{t\in\mathcal{D}}P(t)\neq 0$.

For notation convenience, we define $\theta(x_i,a_i)=P(s=1|x_i,a_i)$ and $\hat{\theta}(x_i,a_i)=\hat{P}(s=1|x_i,a_i)$, where $\hat{\theta}(x_i,a_i)$ is the empirical value based on the frequency estimation. Lemma 1 states that $|\theta(x_i,a_i)-\hat{\theta}(x_i,a_i)|$ is upper bounded by the right term in Eq. 7. Then we can apply the robust fairness aware framework by setting $\theta(x_i,a_i)$ within ϵ range of estimated $\hat{\theta}(x_i,a_i)$, where ϵ is the right term of Eq. 7. According to the Theorem 2 in [2], the generalization error between the true distribution $\theta(x_i,a_i)$ and distribution using the estimated $\hat{\theta}(x_i,a_i)$ is expressed as: $|L_\theta(\mathbf{w})-L_{\hat{\theta}}(\mathbf{w})|<\mu\sqrt{\frac{\ln 2m'+\ln\frac{1}{\delta}}{p_0 N_{\mathcal{D}}}}$ where μ is a constant determined by σ (Lemma 1) and hyperparameter β (Eq. 5). Suppose the maximum value of $L_{\hat{\theta}}(\mathbf{w})$ is defined as $L_{\hat{\theta}}(\mathbf{w})_{max}$ and our robust fairness-aware optimization is to minimize $L_{\hat{\theta}}(\mathbf{w})_{max}$ per iteration. The loss $L_{\hat{\theta}}(\mathbf{w})_{max}$ consists of both the prediction loss and fairness loss. Therefore, the minimization of upper bound of the generalization error of true distribution can provide robustness in terms of both prediction and fairness.

3.3 *RFLearn*^{2}: Bias Correction without $\mathcal{D}$

In this section, we focus on the scenario without unlabeled test data $\mathcal{D}$. The challenge is how to use $\mathcal{D}_s$ alone to estimate the true sample selection probability so that we can construct $\hat{\mathcal{D}}_s$ to resemble $\mathcal{D}$. We assume that 1) there exist K clusters in $\mathcal{D}_s$; 2) the samples in the same cluster have the same selection probability; and 3) the selection probability of each sample is within a range of the uniform selection probability. Under these assumptions, the ratio $\frac{P(s=1)}{P(s=1|x_i,a_i)}$ for each sample from the same cluster is the same. The ratio vector for K clusters is defined as $\mathbf{r}=(r_1,r_2,\cdots,r_K)$. We robustify the estimation by approximating r within a Wasserstein ball B_ρ [1] around the uniform ratio $\mathbf{r}_u$, where all of the values in $\mathbf{r}_u$ is 1. Formally we have $|\mathbf{r}-\mathbf{r}_u|\le\rho$, where ρ is the radius of the Wasserstein ball. Suppose x_i belongs to the k-th cluster where $k\in[K]$, we have:

$$\begin{aligned}&\min_{\mathbf{w}\in\mathcal{W}}\max_{\mathbf{r}\in B_\rho} L(\mathbf{w},\mathbf{r}) = \frac{1}{N_{\mathcal{D}_s}}\sum_{i=1}^{N_{\mathcal{D}_s}} r_k l(f(x_i,a_i),y_i)\\ &\text{subject to } |\mathbf{r}-\mathbf{r}_u|\le\rho,\quad |\frac{1}{N_{\mathcal{D}_s}}\sum_{i=1}^{N_{\mathcal{D}_s}}(a_i-\bar{a})r_k d_{\mathbf{w}(\mathbf{x}_i)}|\le\sigma\end{aligned} \tag{8}$$

4 EXPERIMENT

We use two benchmark datasets, Adult [6] and Dutch [17]. In both datasets, we set *gender* as the protected attribute. We follow the biased data generation approach in [9] and select the data based on the education level (married status) for Adult (Dutch). We choose the logistic regression (LR) to implement and evaluate different algorithms. We consider the following baselines: (a) LR without fairness constraint (*LR*); (b) LR with fairness constraint (*FairLR*); (c) robust LR in [15] that uses kernel functions to reweigh samples under covariate shift but ignores the fairness constraint. For *RFLearn*^{1} (*RFLearn*^{2}), we also consider its variation *RFLearn*$^{1-}$ (*RFLearn*$^{2-}$) that optimizes the robust loss with unweighted fairness constraint. For *RFLearn*$^{2-}$ and *RFLearn*^{2}, we apply K-means with K = 300 to cluster the training data. We run all experiments 20 times and report the average. The details of data preprocessing, biased data creation, and hyperparameters are included in the report [4].

Table 1: Model performance under data distribution shift (Adult and Dutch) Acc: accuracy

Methods	Adult Dataset				Dutch Dataset			
	Training Acc	Test Acc	Training *RD*	Test *RD*	Training Acc	Test Acc	Training *RD*	Test *RD*
LR (unbiased)	0.8124	0.8126	0.1562	0.1373	0.7493	0.7669	0.1498	0.1395
LR	0.8041	0.7882	0.1344	0.1228	0.7018	0.6821	0.0378	0.1012
FairLR	0.7823	0.7622	0.0231	0.0956	0.7006	0.6624	0.0289	0.0991
[15]	0.7883	0.8048	0.1348	0.1333	0.6812	0.7044	0.1421	0.1394
$RFLearn^{1-}$	0.7412	0.7875	0.0351	0.1048	0.6501	0.6879	0.0315	0.0809
$RFLearn^{1}$	0.7484	0.7816	0.0281	0.0416	0.6673	0.6910	0.0317	0.0405
$RFLearn^{2-}$	0.7473	0.7771	0.0321	0.0963	0.6457	0.6809	0.0411	0.0973
$RFLearn^{2}$	0.7336	0.7678	0.0197	0.0238	0.6479	0.6755	0.0321	0.0373

Table 2: Model performance of $RFLearn^{1}$ under sample selection bias with different δ (Adult and Dutch). Acc: accuracy

δ	Adult Dataset				Dutch Dataset			
	Training Acc	Test Acc	Training *RD*	Test *RD*	Training Acc	Test Acc	Training *RD*	Test *RD*
0.025	0.7181	0.7601	0.0189	0.0219	0.6521	0.6812	0.0291	0.0326
0.05	0.7217	0.7673	0.0239	0.0398	0.6521	0.6812	0.0321	0.0326
0.1	0.7484	0.7816	0.0307	0.0416	0.6673	0.6910	0.0378	0.0405
0.15	0.7239	0.7768	0.0277	0.0333	0.6701	0.6994	0.0275	0.0379

Table 3: Model performance of $RFLearn^{2}$ under sample selection bias with different ρ (Adult and Dutch). Acc: accuracy

ρ	Adult Dataset				Dutch Dataset			
	Training Acc	Test Acc	Training *RD*	Test *RD*	Training Acc	Test Acc	Training *RD*	Test *RD*
0.2	0.7229	0.7558	0.0178	0.0114	0.6401	0.6543	0.0175	0.0214
0.4	0.7336	0.7628	0.0197	0.0238	0.6479	0.6755	0.0321	0.0373
0.6	0.7428	0.7724	0.0269	0.0361	0.6544	0.6792	0.0301	0.0314

4.1 Accuracy vs. Fairness

Table 1 shows prediction accuracy and risk difference of each model on training and test data for both Adult and Dutch.

Accuracy. First, the testing accuracy of *LR* is lower than that of *LR (unbiased)*, which demonstrates that the model prediction performance degrades under sample selection bias. Second, with the use of robust learning, the prediction accuracy of $RFLearn^{1-}$ and $RFLearn^{2-}$ outperforms *FairLR*, which demonstrates that the robust learning can provide robust prediction under the sample selection bias. Third, the testing accuracy from robust learning methods is higher than the training accuracy, which further demonstrates the advantage of robust learning under the sample selection bias. Fourth, the accuracy of $RFLearn^{1}$ is higher than that of $RFLearn^{2}$. It is reasonable as we leverage the unlabeled test data in our $RFLearn^{1}$.

Fairness. First, all of *FairLR*, $RFLearn^{1-}$ and $RFLearn^{2-}$ can only achieve fairness on the training data with $RD \leq 0.05$, but none of these approaches can guarantee the fairness on the test data. The method proposed by [15] only considers the robustness of prediction error but ignores the fairness, so that it cannot achieve the fairness on the test data as well. Second, $RFLearn^{1}$ and $RFLearn^{2}$ can achieve fairness on both training and test data as they enforce the fairness under any possible adverse distribution.

4.2 Effects of the Hyperparameters

Table 2 shows the performance of $RFLearn^{1}$ with different δ values. Note that in Lemma 1 the estimation error of the sample selection probability is upper bounded with the probability greater than $1-\delta$. A larger upper bound indicates that the adversary can generate more possible distributions during the robust optimization, hence helping achieve better prediction accuracy on test data. However, when the upper bound is too large, excessive possible distributions may reduce the prediction accuracy on test data. Table 3 shows the result of $RFLearn^{2}$ under different radius ρ. We can see that the testing accuracy increases with the increasing ρ. Larger ρ indicates more possible generated distributions which are more likely to cover the test distribution and improve the model performance. Moreover, the proposed $RFLearn^{1}$ and $RFLearn^{2}$ can achieve both fairness on the training and test data with different δ and ρ.

5 CONCLUSION

In this paper we have developed a robust and fair learning framework with two algorithms to deal with the sample selection bias. Our framework adopts the reweighing estimation approach for bias correction and the minimax robust estimation for achieving robustness on prediction accuracy and fairness on test data. In our future work, we will investigate the sample selection bias under missing not at random, i.e., the sample selection probability also depends on the label, and study how to enforce other fairness notions.

ACKNOWLEDGMENTS

This work was supported in part by NSF 1920920, 1940093, 1946391, and 2137335.

REFERENCES

[1] Jose Blanchet, Yang Kang, and Karthyek Murthy. 2019. Robust Wasserstein profile inference and applications to machine learning. *Journal of Applied Probability* (2019).

[2] Corinna Cortes, Mehryar Mohri, Michael Riley, and Afshin Rostamizadeh. 2008. Sample Selection Bias Correction Theory. In *ALT*.

[3] Amanda Coston, Karthikeyan Natesan Ramamurthy, Dennis Wei, Kush R Varshney, Skyler Speakman, Zairah Mustahsan, and Supriyo Chakraborty. 2019. Fair transfer learning with missing protected attributes. In *AIES*.

[4] Wei Du and Xintao Wu. 2021. Robust Fairness-aware Learning Under Sample Selection Bias. *CoRR* abs/2105.11570 (2021). arXiv:2105.11570 https://arxiv.org/abs/2105.11570

[5] Wei Du, Depeng Xu, Xintao Wu, and Hanghang Tong. 2021. Fairness-aware Agnostic Federated Learning. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*. SIAM, 181–189.

[6] Dheeru Dua and Casey Graff. 2017. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[7] Weihua Hu, Gang Niu, Issei Sato, and Masashi Sugiyama. 2018. Does distributionally robust supervised learning give robust classifiers?. In *ICML*.

[8] Nathan Kallus and Angela Zhou. 2018. Residual unfairness in fair machine learning from prejudiced data. In *ICML*.

[9] Pierre Laforgue and Stephan Clémençon. 2019. Statistical Learning from Biased Training Samples. *arXiv preprint arXiv:1906.12304* (2019).

[10] Preethi Lahoti, Alex Beutel, Jilin Chen, Kang Lee, Flavien Prost, Nithum Thain, Xuezhi Wang, and Ed H Chi. 2020. Fairness without Demographics through Adversarially Reweighted Learning. In *NeurIPS*.

[11] Anqi Liu and Brian Ziebart. 2014. Robust classification under sample selection bias. In *NeurIPS*.

[12] Ashkan Rezaei, Rizal Fathony, Omid Memarrast, and Brian D. Ziebart. 2020. Fairness for Robust Log Loss Classification. In *AAAI*.

[13] Candice Schumann, Xuezhi Wang, Alex Beutel, Jilin Chen, Hai Qian, and Ed H Chi. 2019. Transfer of machine learning fairness across domains. *arXiv preprint arXiv:1906.09688* (2019).

[14] Bahar Taskesen, Viet Anh Nguyen, Daniel Kuhn, and Jose Blanchet. 2020. A Distributionally Robust Approach to Fair Classification. *arXiv preprint arXiv:2007.09530* (2020).

[15] Junfeng Wen, Chun-Nam Yu, and Russell Greiner. 2014. Robust Learning under Uncertain Test Distributions: Relating Covariate Shift to Model Misspecification.. In *ICML*.

[16] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *AISTATS*.

[17] Indre Žliobaite, Faisal Kamiran, and Toon Calders. 2011. Handling conditional discrimination. In *IEEE ICDM*.