Fairness-aware Model-agnostic Positive and Unlabeled Learning

ZIWEI WU, University of Illinois at Urbana-Champaign, USA
JINGRUI HE, University of Illinois at Urbana-Champaign, USA

With the increasing application of machine learning in high-stake decision-making problems, potential algorithmic bias towards people from certain social groups poses negative impacts on individuals and our society at large. In the real-world scenario, many such problems involve positive and unlabeled data such as medical diagnosis, criminal risk assessment and recommender systems. For instance, in medical diagnosis, only the diagnosed diseases will be recorded (positive) while others will not (unlabeled). Despite the large amount of existing work on fairness-aware machine learning in the (semi-)supervised and unsupervised settings, the fairness issue is largely under-explored in the aforementioned Positive and Unlabeled Learning (PUL) context, where it is usually more severe. In this paper, to alleviate this tension, we propose a fairness-aware PUL method named FairPUL. In particular, for binary classification over individuals from two populations, we aim to achieve similar true positive rates and false positive rates in both populations as our fairness metric. Based on the analysis of the optimal fair classifier for PUL, we design a model-agnostic post-processing framework, leveraging both the positive examples and unlabeled ones. Our framework is proven to be statistically consistent in terms of both the classification error and the fairness metric. Experiments on the synthetic and real-world data sets demonstrate that our framework outperforms state-of-the-art in both PUL and fair classification.

Additional Key Words and Phrases: Fairness, Machine Learning, Positive and Unlabeled Learning

ACM Reference Format:
Ziwei Wu and Jingrui He. 2022. Fairness-aware Model-agnostic Positive and Unlabeled Learning. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22), June 21–24, 2022, Seoul, Republic of Korea.* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3531146.3533225

1 INTRODUCTION

Nowadays, machine learning systems are assisting, or in some cases even replacing, human decision making in an increasing number of application domains. Due to the profound impacts of these systems on individuals and our society at large, traditional performance metrics such as accuracy and precision are no longer the sole measure of success. In applications such as credit approval [44], medical diagnosis [7], criminal risk assessment [38] and recommender systems [43], fairness must be carefully taken into account to ensure the absence of discrimination against certain social groups (e.g., women, blacks). Recent years have witnessed a growing interest in fairness-aware machine learning to study the fairness issue. Various metrics of fairness for a predictive model have been studied in the literature [11], including group fairness [10, 25, 49], individual fairness [19, 20, 29] and causal fairness [32, 47]. Depending on the amount of label information available during training, researchers have developed a variety of algorithms to address the unfairness issue. For example, [25, 42, 49, 50] focused on the traditional supervised learning setting where the learning algorithms have access to the class labels of all training examples; [8, 9, 31] studied the unsupervised learning

Manuscript submitted to ACM

setting where no label information is available in the training set; and [12, 51] mainly designed algorithms for the semi-supervised learning setting where only a small portion of the training examples contain class labels.

In many real-world applications, the training data consists of only positive labeled examples and unlabeled ones, which cannot naturally fit in any of the aforementioned learning settings. For example, medical records usually only list which diseases a patient has been diagnosed with (i.e., positive samples) and they usually do not include which diseases a patient does not have. However, the absence of a diagnosis does not mean that the patient does not have the disease. Another example is scoring recidivism. When predicting criminal defendants' likelihood of re-offending, it is easy for us to collect part of the positive samples from the criminal records. For those who have not yet appeared in the criminal records, it is wrong to assume that they will not re-offend. They should therefore not be treated as negative examples but as unlabeled ones. In general, positive and unlabeled learning (PUL) [34] attempts to learn a classifier from this type of data. Existing (semi-)supervised and unsupervised methods can neither deal with positive-only labeled data nor make a full use of unlabeled data, and thus different PUL methods have been proposed [4].

Despite the large amount of existing work on PUL [16, 18, 21, 30], the fairness issue in this setting has been largely under-explored. In applications such as medical diagnosis and recidivism scoring mentioned before, existing techniques can easily lead to inequality among different groups (e.g., black and white, female and male). The difference between fairness in PUL and in the other settings can be summarized in the following three folds: (1) Intuitively, the label imbalance will exacerbate the unfairness problem in PUL. (2) Empirically, methods proposed from supervised and semi-supervised settings cannot handle the unfairness issue in PUL as shown in Table 1. (3) Theoretically, the properties (e.g. consistency) of other methods no longer hold in PUL due to the lack of negatively labeled data. Therefore, in this paper, we aim to bridge this gap and study the fairness-aware binary classification problem in PUL. To be more specific, we seek a classifier which minimizes the misclassification risk in PUL while satisfying the Equalized Odds / Equal Opportunity [25] fairness constraint. We derive the optimal fair classifier via recalibration of the Bayes regressor. This theoretical result motivates us to devise a generic post-processing framework named FairPUL. Based on both positive and unlabeled examples, FairPUL estimates the regression function and the unknown threshold to achieve the fairness criterion in the PUL setting. It enjoys the consistency property where it asymptotically satisfies the fairness criterion and its risk converges to the one of the theoretical optimal fair classifier. Extensive experiments on both the synthetic and the real-world data sets demonstrate FairPUL's effectiveness. Our main contributions can be summarized as follows:

- To the best of our knowledge, this is the first work to systematically study the fairness issue in the positive and unlabeled learning setting.
- We derive the optimal fair classifier in PUL and propose a model-agnostic post-processing framework, which can accommodate different base models and enjoys the consistency property.
- Experiments on the synthetic and real-world data sets show that our framework performs favorably against state-of-the-art in both PUL and fair classification.

The rest of the paper is organized as follows. After a brief review of the related work in Section 2, we present the problem definition in Section 3. Section 4 describes our proposed framework. The experimental results are discussed in Section 5. Finally, we conclude the paper in Section 6.

2 RELATED WORK

2.1 Positive and Unlabeled Learning

Positive and Unlabeled Learning is a variant of the classical classification setup where the training data consists of only positive and unlabeled (PU) examples. It fits within the long-standing interest in developing learning algorithms that do not require fully supervised data. The PU data can originate from two scenarios. One is called the single-training-set scenario [21] where the data come from one single training set; the other is called the case-control scenario [46] where the data come from two independently drawn datasets, one with all positive examples and one with all unlabeled examples. The single-training-set scenario has received substantially more attention in the literature [4]. Most PUL methods can be divided into the following three categories. Two-step techniques [26, 28] first identify reliable negative examples and then conduct classical classification. Biased learning [13, 27, 41] considers PU data as fully labeled data with class label noise for the negative class. Class prior incorporation [18, 21, 30] modifies standard learning methods using the class prior and learns weights for all examples. However, none of the existing work in PUL explores the fairness issue. As shown in our experiments (Subsection 5.2), current PUL methods exhibit unfairness against subgroups, which will result in potential discrimination in real-world applications.

2.2 Fairness in Classification

The existing studies on fairness in classification have focused on two key issues: how to formalize the fairness metric in classification, and how to design efficient algorithms that strike a desirable trade-off between classification performance and fairness. To seek equality between different populations, a number of works have been proposed to control group fairness. Equal Opportunity [25] requires the same true positive rates among groups and Equalized Odds [5] simultaneously considers false positive rates. Equalizing Disincentives [23] requires the difference of two metrics to be equal across the groups. Using these metrics, a variety of fairness-aware algorithms in classification have been proposed under different learning settings. [1, 25, 42, 49, 50] focused on the supervised learning setting where the learning algorithms have access to the class labels of all training examples. Since they highly rely on supervised data, even if we adjust their decision boundary with a standard PUL technique, the unlabeled data still cannot contribute to algorithmic fairness explicitly. In a more realistic scenario where both labeled and unlabeled examples are given, [6] devised a linear transformation pre-processing technique to remove the underlying discrimination. [51] incorporated the fairness constraint into the original training process. [12] post-processed the output conditional probability to improve fairness by recalibration with the unlabeled data. Although these methods do not rely on labeled data as much as the fully supervised methods, both positive and negative labeled samples are necessary for their algorithms. The most related work to us is [12] where they limit the fairness criterion to Equal Opportunity. However, the unlabeled examples can only be used to estimate the fairness threshold in their method. How to properly use the positive-only and unlabeled data in fair classification is still under-explored. In this paper, we instead explore how to make the best use of PU data to boost both classification performance and algorithmic fairness.

3 PROBLEM DEFINITION

We consider a fairness-aware binary classification task in the single-training-set scenario [21] as explained in Subsection 2.1. The training examples consist of tuples (X, S, L) where $X \in \mathbb{R}^d$ is a feature vector, $S \in \{0, 1\}$ is a binary sensitive attribute, and $L \in \{0, 1\}$ represents whether the example is labeled ($L = 1$) or not. Notice that we focus on one binary sensitive attribute in this paper, although the proposed technique can be naturally extended to multiple

multi-class sensitive attributes (details in Subsection 4.2). The training examples are drawn randomly from distribution $p(X, S, L, Y)$, where $Y \in \{0, 1\}$ is the ground-truth label. But for each tuple that is drawn, only (X, S, L) is observed. A classifier g receives a pair (X, S) as input, and outputs a binary prediction for the label. The set of all such functions from $\mathbb{R}^d \times \{0, 1\}$ to $\{0, 1\}$ is denoted by $\mathcal{G}$. For any classifier g, we denote its associated misclassification risk as $\mathcal{R}(g)$. An optimal fair classifier is then defined as:

$$g^* = \arg\min_{g \in \mathcal{G}} \{\mathcal{R}(g) : g \text{ is fair}\}$$

Various definitions of fairness have been proposed so far [11], but there is no consensus regarding which definition is universally the most appropriate. In this work, we employ the following group fairness metrics Equalized Odds and Equal Opportunity introduced in [25]:

Definition 1 (Equalized Odds (EO))**.** We say that a binary classifier $g(X, S)$ satisfies equalized odds with respect to S and Y if $\mathbb{P}\left(g(X, S) = 1 | Y = y, S = 1\right) = \mathbb{P}\left(g(X, S) = 1 | Y = y, S = 0\right), y \in \{0, 1\}$.

Definition 2 (Equal Opportunity (EOP))**.** We say that a binary classifier $g(X, S)$ satisfies equal opportunity with respect to S and Y if $\mathbb{P}\left(g(X, S) = 1 | Y = 1, S = 1\right) = \mathbb{P}\left(g(X, S) = 1 | Y = 1, S = 0\right)$.

In the two fairness metrics, EO requires the same true positive rates and the same false positive rates across the sensitive groups. Compared to EO, EOP is a weaker notion which only requires the same true positive rates. The two fairness metrics have been used extensively in the literature either as a post-processing step [25] on a learned classifier or directly during training [17]. The motivation of these metrics and more discussion regarding the comparison with other fairness metrics can be found in [1, 25, 36].

4 THE PROPOSED FAIRPUL FRAMEWORK

In this section, we introduce the model-agnostic post-processing framework FairPUL for fairness-aware PUL.

4.1 Labeling Mechanism in PUL

As stated before, the unlabeled samples in PUL mainly come from the following two sources: (1) It is truly a negative example; (2) It is a positive example, but simply was not selected by the labeling mechanism. In order to enable learning with PU data, it is necessary [4] to make assumptions about either the labeling mechanism, the class distributions in the data, or both. We base our work on the most frequently used assumption [33] in PUL:

Assumption 1 (Selected Completely At Random (SCAR))**.** Labeled examples are selected completely at random, independent from their attributes, from the positive distribution. Formally speaking, $\mathbb{P}(L = 1 | X, S, Y = 1) = \mathbb{P}(L = 1 | Y = 1) \triangleq c$, where c is called label frequency.

Under this assumption, the set of labeled examples are i.i.d. samples from the positive distribution. The label frequency c plays an important role in the single-training-set scenario [21] we consider here. Define the regression function $f(X, S) := \mathbb{P}(L = 1 \mid X, S)$, i.e., the probability of an example (X, S) being labeled. Then we have the following lemma regarding the constant c:

Lemma 1. Suppose that the SCAR assumption holds. Then $\mathbb{P}(Y = 1 \mid X, S) = f(X, S)/c$.

The proof can be found in Appendix A. This lemma shows how we can obtain the probability of an example being positive from f. Next, we further incorporate fairness into the classification problem and show how to get the optimal fair classifier theoretically and empirically.

4.2 Optimal Fair Classifier in PUL

To obtain the optimal fair classifier, we study the following problem:

$$\min_{g\in\mathcal{G}}\{\mathcal{R}(g) : \mathbb{P}\left(g(X,S)=1|Y=y,S=1\right)=\mathbb{P}\left(g(X,S)=1|Y=y,S=0\right), y\in\mathcal{Y}\} \tag{1}$$

For EO, $\mathcal{Y}=\{0,1\}$. When $y=1$, the constraint requires the same true positive rates (TPR), that is $TPR^{(1)}=TPR^{(0)}$, where we use the superscript to identify the sensitive attribute S. Similarly, when $y=0$, the constraint requires the same false positive rates (FPR), that is $FPR^{(1)}=FPR^{(0)}$. For EOP, $\mathcal{Y}=\{1\}$ and it only requires $TPR^{(1)}=TPR^{(0)}$. In our later derivation, we will focus on EO while discussing how our results can easily accommodate EOP.

Using the misclassification risk $\mathcal{R}(g)=\mathbb{P}(g\neq Y)$ and based on Lemma 1, we can solve problem (1) and get the following result:

Lemma 2. The minimizer $g^*_{\boldsymbol{\lambda}}$ for every Lagrange multipliers $\boldsymbol{\lambda}=(\lambda_1,\lambda_2)\in\mathbb{R}^2$ is:

$$\begin{aligned} g^*_{\boldsymbol{\lambda}}(X,1) =& \mathbb{1}_{\{\frac{f(X,1)}{c}(1-\frac{\lambda_1}{\mathbb{P}(Y=1,S=1)})+(1-\frac{f(X,1)}{c})(1-\frac{\lambda_2}{\mathbb{P}(Y=0,S=1)})\geq 0\}} \\ g^*_{\boldsymbol{\lambda}}(X,0) =& \mathbb{1}_{\{\frac{f(X,0)}{c}(1+\frac{\lambda_1}{\mathbb{P}(Y=1,S=0)})-(1-\frac{f(X,0)}{c})(1+\frac{\lambda_2}{\mathbb{P}(Y=0,S=0)})\geq 0\}} \end{aligned} \tag{2}$$

Here, λ_1 is the Lagrange multiplier for the same TPR constraint, and λ_2 is for the same FPR constraint. Since $g(X,S)\in\{0,1\}$, we can rewrite the Lagrange function of Problem (1) into expressions of $g(X,S)$ and then get the above equations. The detailed derivation is shown in Appendix B. Note that by incorporating c explicitly, our designed minimizer can boost performance and meanwhile reduce discrimination for PU data.

Remark. We consider one binary sensitive attribute in this problem for derivation simplicity. This can be naturally extended to accommodate multiple multi-class sensitive attributes (such as White, Black, Asian for race). We can break down the fairness into pairs of equations, add more equality constraints of EOD/EOP with respect to all sensitive attributes to the optimization problem in Equation (1), and obtain a similar solution as Equation (2).

With Lemma 2, the problem of finding the optimal fair classifier in PUL is equivalent to obtaining the optimal value of $\boldsymbol{\lambda}$ in Equation (2). To do this, we introduce an assumption on the regression function $\mathbb{P}(Y=1\mid X,S)$, i.e., the probability of an example (X,S) being positive. For notation simplicity, we abbreviate it as $p^+_{X,S}$.

Assumption 2. The mapping $t\mapsto\mathbb{P}(p^+_{X,S}\leq t|S=s)$ is continuous on $(0,1)$.

This Assumption requires that the random variable $p^+_{X,s}$ does not have atoms for each $s\in\{0,1\}$. This is proved achievable by many distributions [14, 40, 48]. Based on this assumption, we introduce the optimal λ^*:

Lemma 3. The optimal $\boldsymbol{\lambda}^*$ for g^* satisfies

$$\begin{aligned} \frac{\mathbb{E}_{X|S=1}[g^*_{\boldsymbol{\lambda}^*}(X,1)f(X,1)]}{\mathbb{P}(Y=1\mid S=1)} &= \frac{\mathbb{E}_{X|S=0}[g^*_{\boldsymbol{\lambda}^*}(X,0)f(X,0)]}{\mathbb{P}(Y=1\mid S=0)} \\ \frac{\mathbb{E}_{X|S=1}[(c-f(X,1))(1-g^*_{\boldsymbol{\lambda}^*}(X,1))]}{\mathbb{P}(Y=0\mid S=1)} &= \frac{\mathbb{E}_{X|S=0}[(c-f(X,0))(1-g^*_{\boldsymbol{\lambda}^*}(X,0))]}{\mathbb{P}(Y=0\mid S=0)} \end{aligned} \tag{3}$$

Proof Sketch. Equation (3) is obtained with first-order optimality conditions, and its optimality can be proved based on Assumption 2. The detailed proof is shown in Appendix C. □

Remark. The results in Lemma 2 are the minimizer under the fairness metric of EO, and Equations (3) show the constraints for the optimal $\boldsymbol{\lambda}^*$ under EO. To accommodate EOP, since we do not need to consider TNR, we can simply let $\lambda_2 = 0$, and the optimal $\boldsymbol{\lambda}^*$ only needs to satisfy the first equation in Equations (3). Notice that since $\lambda_2 = 0$ for EOP, the first equation is only the constraint for the optimal λ_1.

4.3 FairPUL for Fairness-aware PUL

Based on the theoretical optimal classifier in Equations (2) and (3), we are now ready to introduce the proposed FairPUL framework for empirically estimating the optimal fair classifier. Given an existing classifier, FairPUL will modify its output for PUL data while reducing unfairness, and output the fair version of the classifier with a low misclassification risk. In the next subsections, any notation (e.g., c) with a hat denotes its estimated counterpart (e.g., $\hat{c}$).

In the PUL setting, the training data consists of two parts, the labeled (positive) data set D_L and the unlabeled data set D_U. A classifier $\hat{f}$ can be constructed to estimate the probability of an example being labeled by treating D_L as positive examples and D_U as negative ones. Let V be the validation set that is drawn from the overall distribution in the same manner as the training set and P be the subset of examples in V that are labeled (and positive). We can estimate the label frequency c as follows:

$$\hat{c} = \frac{1}{n_P} \sum_{(x,s)\in P} \hat{f}(x,s) \tag{4}$$

where n_P is the cardinality of P. If $\hat{f}$ is trained well enough, $\hat{f}(x,s)$ provides a precise estimate of labeling probability. Since every example in P is positive, we have $\hat{f}(x,s) \approx c$ based on Total Probability Theorem and SCAR assumption. Although this ideal case hardly holds in practice, the average estimation is a good choice for estimating the label frequency c with low variance.

Since the ideal $\boldsymbol{\lambda}^*$ should satisfy Equations (3), we minimize the difference between its two sides. Based on the rule of conditional probability, this is equivalent to minimizing unfairness defined as follows:

Definition 3 (Unfairness)**.** For a binary classifier g, its unfairness Δ under different fairness metrics can be defined as:

Under EO:

$$\Delta_{EO}(g) = AOD(g) = \frac{1}{2}\left[|\mathrm{TPR}^{(1)} - \mathrm{TPR}^{(0)}| + |\mathrm{FPR}^{(1)} - \mathrm{FPR}^{(0)}|\right]$$

Under EOP:

$$\Delta_{EOP}(g) = EOD(g) = |\mathrm{TPR}^{(1)} - \mathrm{TPR}^{(0)}|$$

We can get their empirical versions by substituting all the unknown terms with their empirical estimators:

$$\begin{aligned}
\hat{\Delta}_{EO}(g) &= \tfrac{1}{2}\Big[\Big|\frac{\hat{\mathbb{E}}_{X|S=1}[\hat{f}(X,1)g(X,1)]}{\hat{\mathbb{E}}_{X|S=1}[\hat{f}(X,1)]} - \frac{\hat{\mathbb{E}}_{X|S=0}[\hat{f}(X,0)g(X,0)]}{\hat{\mathbb{E}}_{X|S=0}[\hat{f}(X,0)]}\Big| \\
&+\Big|\frac{\hat{\mathbb{E}}_{X|S=1}[(\hat{c}-\hat{f}(X,1))(1-g(X,1))]}{\hat{c}-\hat{\mathbb{E}}_{X|S=1}[\hat{f}(X,1)]} - \frac{\hat{\mathbb{E}}_{X|S=0}[(\hat{c}-\hat{f}(X,0))(1-g(X,0))]}{\hat{c}-\hat{\mathbb{E}}_{X|S=0}[\hat{f}(X,0)]}\Big|\Big] \\
\hat{\Delta}_{EOP}(g) &= \Big|\frac{\hat{\mathbb{E}}_{X|S=1}[\hat{f}(X,1)g(X,1)]}{\hat{\mathbb{E}}_{X|S=1}[\hat{f}(X,1)]} - \frac{\hat{\mathbb{E}}_{X|S=0}[\hat{f}(X,0)g(X,0)]}{\hat{\mathbb{E}}_{X|S=0}[\hat{f}(X,0)]}\Big|
\end{aligned} \tag{5}$$

Then we can obtain the estimator of $\boldsymbol{\lambda}^*$ as:

$$\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}} \hat{\Delta}(\hat{g}^*_{\boldsymbol{\lambda}}) \tag{6}$$

Algorithm 1 FairPUL

Input: Labeled data set D_L, unlabeled dataset D_U; Validation set V; Base model $\hat{f}$;
Output: $\hat{g}_{\hat{\lambda}}$

1: Estimate $\hat{c}$ in V with Equation (4).
2: Compute terms in $\hat{\Delta}(\hat{g}_\lambda)$ defined in Equations (5) and (7) with D_L and D_U.
3: Find $\hat{\lambda}$ to minimize $\hat{\Delta}(\hat{g}_\lambda)$ with Simulated Annealing.
4: Compute $\hat{g}_{\hat{\lambda}}$ in Equation (7).

where $\hat{\Delta}$ can be either $\hat{\Delta}_{EO}$ or $\hat{\Delta}_{EOP}$ to satisfy EO or EOP respectively. $\hat{g}^*_{\boldsymbol{\lambda}}$ is the estimated optimal fair classifier defined in a similar way as Equation (2) by replacing all the unknown terms with their empirical estimators.:

$$\begin{aligned}\hat{g}^*_{\boldsymbol{\lambda}}(X,1) &= \mathbb{1}_{\{\frac{\hat{f}(X,1)}{\hat{c}}(1-\frac{\lambda_1}{\hat{\mathbb{P}}(Y=1,S=1)})+(1-\frac{\hat{f}(X,1)}{\hat{c}})(1-\frac{\lambda_2}{\hat{\mathbb{P}}(Y=0,S=1)})\geq 0\}}\\ \hat{g}^*_{\boldsymbol{\lambda}}(X,0) &= \mathbb{1}_{\{\frac{\hat{f}(X,0)}{\hat{c}}(1+\frac{\lambda_1}{\hat{\mathbb{P}}(Y=1,S=0)})-(1-\frac{\hat{f}(X,0)}{\hat{c}})(1+\frac{\lambda_2}{\hat{\mathbb{P}}(Y=0,S=0)})\geq 0\}}\end{aligned} \tag{7}$$

We adopt the Simulated Annealing strategy to search for the optimal $\hat{\boldsymbol{\lambda}}$.

Alg. 1 summarizes our proposed FairPUL framework. It takes the training and validation sets as well as a base model $\hat{f}$ as input, and outputs a fairness-aware classifier with a low misclassification risk. The base regressor $\hat{f}$ is trained with D_L and D_U. It can be any classifier that outputs the conditional probability of an example being labeled. First, an estimator of c can be learned on the validation set with $\hat{f}$ in Step 1. Then we use Simulated Annealing to search for the best $\hat{\lambda}$ to minimize unfairness for PUL data in Step 2 and 3. We finally compute the empirical optimal fair classifier accordingly in Step 4.

4.4 Consistency of FairPUL

Consistency is a desired property for a classifier in asymptotic theory. It guarantees that with the increase in the amount of data, the estimation will converge to the true value. In this subsection, we show that the proposed framework FairPUL is consistent, that is, it asymptotically satisfies the fairness criterion and its risk converges to the one of the theoretical optimal fair classifier.

First of all, following [12], we make the following realistic assumptions on the estimator of $p^+_{X,S}$:

Assumption 3. The estimator $\hat{p}^+_{X,S}$ satisfies that, $\forall s \in \{0,1\}$,

- $\mathbb{E}_D\mathbb{E}_{X|S=s}|p^+_{X,S} - \hat{p}^+_{X,S}| \to 0$ as $n_U, n_L \to \infty$, where n_U and n_L denote the number of examples in D_U and D_L respectively.
- There exists a sequence $c_{U,L} > 0$ satisfying $\frac{1}{c_{U,L}\sqrt{N}} = o_{U,L}(1)$ and $c_{U,L} = o_{U,L}(1)$ such that $\mathbb{E}_{X|S=s}[\hat{p}^+_{X,S}] \geq c_{U,L}$ almost surely.
- The mapping $t \mapsto \mathbb{P}(\hat{p}^+_{X,s} \leq t|S=s)$ is continuous on $(0,1)$ almost surely.

The first part of the assumption requires the estimator to be consistent in l_1 norm. This can be achieved by a variety of estimations for different regression functions as shown in the literature [3, 15, 45]. The second part means that $\mathbb{E}_{X|S=s}[\hat{p}^+_{X,S}]$ is lower bounded by a certain positive term which vanishes as n_U and n_L go to infinity. This can be easily achieved by slightly modifying any existing consistent estimator. The last part is similar to Assumption 2.

Based on these realistic assumptions on $p^+_{X,S}$, next we establish the statistical consistency of FairPUL.

Theorem 1 (Asymptotic properties)**.** FairPUL satisfies:

$\lim_{n_U,n_L\to\infty} \mathbb{E}\left[\Delta(\hat{g})\right] = 0, \lim_{n_U,n_L\to\infty} \mathbb{E}\left[\mathcal{R}(\hat{g})\right] = \mathcal{R}(g*)$

Proof Sketch. To prove asymptotic optimality (the second part), we introduce an intermediate estimator and use it to upper bound the excess risk with triangle inequality. The upper bound converges to zero based on the first two parts in Assumption 3. For asymptotic fairness (the first part), we upper bound the unfairness with its empirical version which converges to zero. Detailed proof can be found in Appendix D. Recall that $\hat{p}^+_{X,S} = \hat{f}(X,S)/\hat{c}$. Its consistency as $n_U, n_L \to \infty$ is vital to the consistency of our framework in PUL. □

5 EXPERIMENTS

To evaluate the effectiveness of FairPUL, we conduct extensive experiments to answer the following research questions:

- **RQ1**: How does FairPUL perform compared with state-of-the-art in fair classification and PUL?
- **RQ2**: How does FairPUL for post-processing compare with in-processing and pre-processing methods?
- **RQ3**: How do unlabeled examples affect FairPUL?

5.1 Experiments on Synthetic Data.

5.1.1 Experimental setup. The aim of the synthetic experiment is to study the behavior of FairPUL in comparison with other methods with the base model of linear logistic regression (Lin.LR), in terms of both classification performance and fairness. To this end, we generate a synthetic binary classification data set with two sensitive groups following Donini et al. [17]. For each group in the class 0 and for the group a in the class 1, we generate 1,000 examples for training and the same number for testing. For the group b in the class 1, we generate 200 examples for training and the same number for testing. Each set of examples is sampled from a 2-dimensional isotropic Gaussian distribution with different mean μ and variance σ^2 : (i) Group a, Label 1: μ = (-1, -1), σ^2 = 0.8; (ii) Group a, Label 0: μ = (1, 1), σ^2 = 0.8; (iii) Group b, Label 1: μ = (-0.5, -0.5), σ^2 = 0.5; (iv) Group b, Label 0: μ = (0.5, 0.5), σ^2 = 0.5. When a standard machine learning method is applied to this synthetic data set, the generated model is unfair with respect to the group b, in that the classifier tends to negatively classify the examples in this group. We search in $[0.01, 0.1, 1, 10, 100]$ for the best regularization parameter C. We generate the validation set from the training set via holdout validation and the holdout ratio is set to 0.2.

5.1.2 Baselines. To study and compare the performance of FairPUL, we first adopt two classic methods as baselines:

- **Oracle:** A method using the fully labeled training set (all examples are labeled either positive or negative). Its results correspond to the fully supervised setting.
- **Naïve:** A method using the labeled examples as positive ones and treating unlabeled examples as negative.

Since this is the first work on the fairness issue in the PUL setting, we further compare FairPUL with the following two types of existing work. For fairness work:

- **Agarwal [1]:** An in-processing method reducing fair classification to cost-sensitive classification problems and yielding a randomized classifier with the lowest error subject to the desired constraints.
- **Hardt [25]:** A post-processing method which takes as input an existing classifier and the sensitive feature, and derives a monotone transformation of the prediction to enforce the specified fairness constraints.
- **Chzhen [12]:** A post-processing method which recalibrates the Bayes classifier by a group-dependent threshold to minimize unfairness.

All the above fairness baselines adopt the similar fairness metric as we do. For PUL work:

- **uPU/wPU [21]:** Unbiased PUL methods under the SCAR assumption by reweighting the examples. Both unweighted (uPU) and weighted (wPU) versions were proposed.
- **nnPU [30]:** A designed non-negative risk estimator for PUL which can be trained on flexible neural networks when minimized.
- **BaggingSVM [37]:** A bootstrap-bagging-based model which iteratively trains multiple classifiers to discriminate the known positive examples from random subsamples of the unlabeled set, and averages their predictions.

5.1.3 Results (RQ1). We compare the test classification error and the fairness metrics under different labeling rates in $[1.0, 0.8, 0.6, 0.4, 0.2]$ of the baselines and our proposed FairPUL. For FairPUL with EO constraint, the fairness metric is AOD. For FairPUL with EOP constraint, the fairness metric is EOD. The base model for all the methods is Lin.LR. We report the average results of misclassification error, AOD and EOD of 5 independent trials in Fig. 1. The closer the dot is to the origin, the more fair and accurate the model is. From the figures we can see that for all the methods, their classification performances drop as the labeling rate decreases. Under high labeling rates (≥ 0.8), our framework achieves much lower AOD and EOD (i.e., higher level of fairness) while maintaining a good level of accuracy; under low labeling rates, our framework achieves much better performance in both classification accuracy and fairness. Note that even though some methods achieve zero AOD/EOD when the labeling rate is 0.2, it is not an ideal model we are seeking. In this case, the model simply predicts every example to be negative, showing "fake" fairness. What we are seeking is maintaining task-specific performance and reducing discrimination simultaneously.

5.2 Experiments on Real Data.

5.2.1 Data sets and experimental setup. To further study how FairPUL performs, we conduct experiments on three publicly available real-world data sets. In all the experiments, to obtain reliable estimates of classification performance and fairness, we repeatedly randomly split each data set into training (70%) and test (30%) sets 10 times, and report the averages and standard deviations of the metrics over different independent runs. Following [24], to realize the PUL setting, for each training set, we randomly select 90% positive examples as labeled and leave the remaining 10% positive examples as well as all negative examples as unlabeled. We split up the training set for holdout validation in FairPUL and the holdout ratio is set to 0.2.

- COMPAS recidivism data [2]. It includes 5278 records with 47% positive examples. The task is to predict recidivism from someone's criminal history, jail and prison time, demographics, and COMPAS risk scores, with race as the protected sensitive attribute restricted to black (about 40%) and white defendants.
- German Credit [35]. It includes 1000 examples where 30% are positive. The task is to classify people as good or bad credit risks by features related to the economical situation, with gender as the sensitive attribute restricted to female (about 31%) and male.
- Drug [22]. It comprises 1885 records of human subjects, and for each subject, it provides five demographic features, seven features measuring personality traits and 18 features each of which describes the subject's last use of a certain drug. We choose the use of heroin here, where 80% of the subjects have never used heroin. We restrict the sensitive attribute race to black (about 9%) and white.

We compare FairPUL with the baselines described in Subsection 5.1.2 with the base model of Linear Support Vector Machine (Lin.SVM). Since previous fairness works cannot be naturally adapted to the PUL setting, the fairness baselines we use here all follow the naïve method to transform PUL to the traditional learning setting. The hyper-parameters of

Table 1. Average results and standard deviations on real data of 10 runs. The labeling rate is 90%. The top 2 results under each metric are marked bold.

Models	German			COMPAS			Drug		
	F1	AOD	EOD	F1	AOD	EOD	F1	AOD	EOD
Oracle	0.565±0.051	0.053±0.033	0.061±0.042	0.604±0.018	0.194±0.020	0.219±0.032	0.777±0.021	0.130±0.024	0.041±0.036
Naïve	0.471±0.048	0.068±0.024	0.099±0.052	0.529±0.058	0.180±0.031	0.237±0.045	0.753±0.019	0.142±0.031	0.086±0.049
Agarwal	0.470±0.033	0.053±0.026	0.077±0.051	0.467±0.042	0.032±0.021	0.040±0.036	0.743±0.031	0.123±0.019	0.065±0.033
Hardt	0.449±0.058	0.059±0.025	0.093±0.046	0.409±0.062	0.034±0.023	0.034±0.029	0.756±0.022	0.111±0.033	0.075±0.042
Chzhen	0.393±0.069	0.048±0.018	**0.053±0.028**	0.486±0.047	0.036±0.012	0.032±0.020	0.755±0.020	0.108±0.024	0.041±0.027
uPU	0.574±0.022	0.097±0.045	0.054±0.027	0.644±0.012	0.031±0.024	0.016±0.009	**0.871±0.026**	0.143±0.060	0.066±0.043
wPU	0.534±0.131	0.050±0.022	0.066±0.039	0.612±0.071	0.188±0.035	0.245±0.048	0.766±0.027	0.147±0.016	0.074±0.049
Bagging	0.544±0.056	0.054±0.034	0.062±0.052	**0.649±0.013**	0.215±0.031	0.252±0.038	0.813±0.020	0.196±0.041	0.074±0.051
FairPUL-EO	**0.578±0.023**	**0.045±0.018**	0.056±0.031	0.646±0.012	**0.005±0.003**	**0.004±0.002**	**0.871±0.021**	**0.056±0.023**	**0.024±0.013**
FairPUL-EOP	**0.580±0.020**	**0.042±0.029**	**0.034±0.027**	**0.648±0.013**	**0.006±0.004**	**0.003±0.002**	**0.897±0.023**	**0.061±0.048**	**0.022±0.010**

every algorithm have been carefully tuned to achieve the best classification performance, and the details can be found in Appendix E.

5.2.2 Metrics. We compare our framework with the baselines using the metrics of F1 score for evaluating the model performance, and AOD and EOD for evaluating the fairness level. It is worth noting that most previous works on fairness simply use classification accuracy to evaluate the models' performance. However, higher accuracy does not necessarily mean a better classification ability, especially on imbalanced data sets like German Credit and drug. Therefore, we use the F1 score instead.

5.2.3 Results (RQ1). The comparison of baselines and FairPUL is summarized in Table 1. From the results we have the following observations:

- The naïve and PUL baselines show high AOD and EOD, and most often higher than the oracle method. It demonstrates the unfairness problem in the PUL setting, which is usually more severe than in the supervised learning setting as we discussed before.
- Fair classification methods often obtain lower AOD and EOD, and PUL methods often obtain higher F1 scores. However, none of them can perform well on both metrics. That is, existing fairness works cannot ensure model performance in the PUL setting, while existing PUL methods cannot ensure fairness.
- FairPUL methods achieve both high F1 scores and low AOD and EOD. This demonstrates that our framework strikes a good trade-off between classification performance and fairness. FairPUL even beats the oracle model. This may be due to the imbalance of the data sets, where FairPUL resolves it via proper estimation of the label frequency.
- FairPUL-EOP often achieves higher F1 scores than FairPUL-EO under comparable AOD and EOD. This is because EOP is a weaker fairness constraint compared with EO, and thus it typically allows for stronger task performance.

We report the results of the methods with different base models on COMPAS in Table 4 in Appendix F. Similar observations can be drawn, which demonstrate that our proposed FairPUL can always strike a good trade-off between classification performance and fairness, and generalize well to different base models.

5.3 Post-processing vs. In-/Pre-processing (RQ2)

In this section, we compare the post-processing FairPUL with the in-processing method proposed by [1] and the pre-processing method in [6]. The in-processing method is introduced in Section 5.2. The pre-processing method is a widely-used technique which transforms the non-sensitive features to remove their correlation with the sensitive feature while retaining as much information as possible. We report the average results of 10 independent runs on

Table 2. Post-processing vs In-processing/Pre-processing. The best 2 results for each metric are marked bold.

Rates	Models		F1	AOD	EOD
100%	Lin.SVM	Naïve	0.604±0.018	0.194±0.020	0.219±0.032
		In	0.588±0.014	0.030±0.026	0.039±0.021
		Pre	0.598±0.015	0.076±0.025	0.098±0.027
		FairPUL-EO	**0.649±0.013**	**0.004±0.003**	**0.005±0.004**
		FairPUL-EOP	**0.650±0.014**	**0.008±0.004**	**0.004±0.002**
	Lin.LR	Naïve	0.621±0.015	0.236±0.033	0.293±0.045
		In	0.595±0.011	0.042±0.022	0.038±0.026
		Pre	0.590±0.016	0.054±0.029	0.047±0.027
		FairPUL-EO	**0.657±0.024**	**0.014±0.010**	**0.013±0.006**
		FairPUL-EOP	**0.660±0.013**	**0.018±0.005**	**0.012±0.009**
90%	Lin.SVM	Naïve	0.529±0.058	0.180±0.031	0.237±0.045
		In	0.467±0.042	0.032±0.021	0.040±0.036
		Pre	0.429±0.054	0.028±0.016	0.038±0.029
		FairPUL-EO	**0.646±0.012**	**0.005±0.003**	**0.004±0.002**
		FairPUL-EOP	**0.648±0.013**	**0.006±0.004**	**0.003±0.002**
	Lin.LR	Naïve	0.552±0.023	0.240±0.048	0.311±0.054
		In	0.507±0.019	0.032±0.021	0.037±0.036
		Pre	0.495±0.019	0.049±0.030	0.053±0.024
		FairPUL-EO	**0.656±0.016**	**0.015±0.008**	**0.013±0.010**
		FairPUL-EOP	**0.658±0.012**	**0.016±0.007**	**0.009±0.008**

Table 3. Comparison results of different labeling rates.

Rates	Models		F1	AOD	EOD
100%	Lin.SVM	Naïve	0.604±0.018	0.194±0.020	0.219±0.032
		FairPUL	0.650±0.014	0.008±0.004	0.004±0.002
	Lin.LR	Naïve	0.621±0.015	0.236±0.033	0.293±0.045
		FairPUL	0.660±0.013	0.018±0.005	0.012±0.009
90%	Lin.SVM	Naïve	0.529±0.058	0.180±0.031	0.237±0.045
		FairPUL	0.648±0.013	0.006±0.004	0.003±0.002
	Lin.LR	Naïve	0.552±0.023	0.240±0.048	0.311±0.054
		FairPUL	0.658±0.012	0.016±0.007	0.009±0.008
80%	Lin.SVM	Naïve	0.274±0.024	0.090±0.011	0.126±0.023
		FairPUL	0.646±0.012	0.006±0.005	0.004±0.003
	Lin.LR	Naïve	0.401±0.018	0.142±0.021	0.205±0.031
		FairPUL	0.655±0.011	0.015±0.007	0.007±0.007
50%	Lin.SVM	Naïve	0.000±0.000	0.000±0.000	0.000±0.000
		FairPUL	0.644±0.012	0.009±0.007	0.006±0.005
	Lin.LR	Naïve	0.068±0.015	0.022±0.009	0.037±0.016
		FairPUL	0.652±0.014	0.017±0.007	0.008±0.004

COMPAS data set under labeling rates of 1.0 and 0.9 in Table 2. Results on German can be found in Appendix F. In the fully labeled setting, the in-processing and pre-processing methods sacrifice much classification performance for better fairness. FairPUL achieves a good balance instead. Under labeling rate 90%, FairPUL significantly outperforms all the baselines in both metrics. Compared with the in-processing method, FairPUL only needs black-box access to the predictions and sensitive attribute information without requiring access to the actual algorithms and ML models. While the pre-processing method only needs to transform the data set before the actual model takes effect, it leads to classifiers that still exhibit substantial unfairness in practice. So our proposed FairPUL is more flexible and also effective in fairness-aware PUL.

5.4 In-depth Study

5.4.1 Effect of labeling rates (RQ3). We test labeling rates of [100%, 90%, 80%, 50%] on the base models Lin.SVM and Lin.LR. Changing the labeling rates affects the number of positive and unlabeled examples simultaneously. As we

have observed before, FairPUL-EOP often achieves a better trade-off between fairness and classification compared with FairPUL-EO, so we only report the results of FairPUL-EOP here. Average results of 10 runs on COMPAS data set are shown in Table 3. Additional results on German are in Appendix F. From the table we can see that FairPUL always achieves higher F1 scores and lower AOD and EOD under different labeling rates compared with the naïve method. Although FairPUL's classification performance drops with the decrease in labeling rates, it shows much more stable performance and manages to always maintain a good level of fairness. Even under the labeling rate as low as 50%, where the naïve method fails and simply predicts every example as negative, FairPUL can achieve even comparable performance with naïve method in the fully labeled setting. This means that FairPUL can achieve satisfying performance with much fewer labeled examples. It is a very desirable property in practice where labeled data is often time-consuming and expensive to obtain.

5.4.2 Effect of unlabeled samples (RQ3). To further study how unlabeled examples affect our framework, we fix the number of positive examples and compare FairPUL's performance with different numbers of unlabeled examples. Since the benchmark data sets are not provided with additional unlabeled data, we deploy the following data generation procedure: we randomly select 50% examples in the original training set. The remaining 50% will serve as a pool of unlabeled examples. We test our model leveraging $[0\%, 10\%, 20\%, 30\%, 40\%, 50\%]$ examples in the pool. Note that to ensure the SCAR assumption holds, we randomly label examples of the fixed number from the positive distribution in the current training set. Average results of 10 runs of FairPUL-EOP on COMPAS are shown in Fig. 2. We can see that with more unlabeled examples, our framework achieves better F1 scores and lower AOD and EOD. From the perspective of our framework design, unlabeled examples are leveraged in two aspects: to help predict the labeling probability for building the traditional classifier; to help reduce unfairness in constructing the empirical optimal fair classifier. Adding more unlabeled examples to the training set will benefit both aspects, which is proved by this experiment. This demonstrates FairPUL's advantage in improving both classification performance and fairness simply with unlabeled examples, which may be otherwise useless in other methods.

6 CONCLUSION

In this paper, motivated by real applications such as medical diagnosis and recidivism scoring, we study the important and yet less studied problem of the optimal fair binary classifier in the positive and unlabeled learning using the notion of equalized odds and equal opportunity. In particular, we bridge the gap between fair classification and PUL by first providing the theoretical analysis, and then designing a model-agnostic post-processing framework which preserves favorable consistency properties under mild assumptions. We highlight our framework's flexibility of being easily generalized to any base classifier which outputs conditional labeling probabilities. Extensive experiments demonstrate that our framework outperforms state-of-the-art in both PUL and fair classification.

ACKNOWLEDGMENTS

This work is supported by National Science Foundation under Award No. IIS-1947203, IIS-2117902, and IIS-2002540. The views and conclusions are those of the authors and should not be interpreted as representing the official policies of the funding agencies or the government.

REFERENCES

[1] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. 2018. A reductions approach to fair classification. In *ICML*. PMLR, 60–69.

[2] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[3] Jean-Yves Audibert, Alexandre B Tsybakov, et al. 2007. Fast learning rates for plug-in classifiers. *The Annals of statistics* 35, 2 (2007), 608–633.

[4] Jessa Bekker and Jesse Davis. 2020. Learning from positive and unlabeled data: A survey. *Machine Learning* (2020), 719–760.

[5] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2018. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research* (2018), 0049124118782533.

[6] Sarah Bird, Miro Dudík, Richard Edgar, Brandon Horn, Roman Lutz, Vanessa Milan, Mehrnoosh Sameki, Hanna Wallach, and Kathleen Walker. 2020. *Fairlearn: A toolkit for assessing and improving fairness in AI.* Technical Report MSR-TR-2020-32. Microsoft. https://www.microsoft.com/en-us/research/publication/fairlearn-a-toolkit-for-assessing-and-improving-fairness-in-ai/

[7] Danton S Char, Nigam H Shah, and David Magnus. 2018. Implementing machine learning in health care—addressing ethical challenges. *The New England journal of medicine* 378, 11 (2018), 981.

[8] Xingyu Chen, Brandon Fain, Liang Lyu, and Kamesh Munagala. 2019. Proportionally fair clustering. In *ICML*. PMLR, 1032–1041.

[9] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. 2017. Fair Clustering Through Fairlets. In *NeurIPS*. 5029–5037.

[10] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.

[11] Alexandra Chouldechova and Aaron Roth. 2018. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810* (2018).

[12] Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. 2019. Leveraging Labeled and Unlabeled Data for Consistent Fair Binary Classification. In *NeurIPS*, Vol. 32.

[13] Marc Claesen, Frank De Smet, Johan AK Suykens, and Bart De Moor. 2015. A robust ensemble approach to learn from positive and unlabeled data using SVM base models. *Neurocomputing* 160 (2015), 73–84.

[14] Christophe Denis and Mohamed Hebiri. 2020. Consistency of plug-in confidence sets for classification in semi-supervised learning. *Journal of Nonparametric Statistics* 32, 1 (2020), 42–72.

[15] Luc Devroye. 1978. The uniform convergence of nearest neighbor regression function estimators and their application in optimization. *IEEE Transactions on Information Theory* 24, 2 (1978), 142–151.

[16] Amit Dhurandhar and Karthik S Gurumoorthy. 2020. Classifier Invariant Approach to Learn from Positive-Unlabeled Data. In *ICDM*. IEEE, 102–111.

[17] Michele Donini, Luca Oneto, Shai Ben-David, John S Shawe-Taylor, and Massimiliano Pontil. 2018. Empirical risk minimization under fairness constraints. In *NeurIPS*. 2791–2801.

[18] Marthinus Christoffel du Plessis, Gang Niu, and Masashi Sugiyama. 2014. Analysis of Learning from Positive and Unlabeled Data. In *NeurIPS*.

[19] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*. 214–226.

[20] Cynthia Dwork, Christina Ilvento, and Meena Jagadeesan. 2020. Individual fairness in pipelines. *arXiv preprint arXiv:2004.05167* (2020).

[21] Charles Elkan and Keith Noto. 2008. Learning classifiers from only positive and unlabeled data. In *SIGKDD*. 213–220.

[22] Elaine Fehrman, Awaz K Muhammad, Evgeny M Mirkes, Vincent Egan, and Alexander N Gorban. 2017. The five factor model of personality and evaluation of drug consumption risk. In *Data science*. Springer, 231–242.

[23] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and removing disparate impact. In *SIGKDD*. 259–268.

[24] Chen Gong, Tongliang Liu, Jian Yang, and Dacheng Tao. 2019. Large-margin label-calibrated support vector machines for positive and unlabeled learning. *IEEE transactions on neural networks and learning systems* 30, 11 (2019), 3471–3483.

[25] Moritz Hardt, Eric Price, Eric Price, and Nati Srebro. 2016. Equality of Opportunity in Supervised Learning. In *NeurIPS*.

[26] Fengxiang He, Tongliang Liu, Geoffrey I Webb, and Dacheng Tao. 2018. Instance-dependent pu learning by bayesian optimal relabeling. *arXiv preprint arXiv:1808.02180* (2018).

[27] Cho-Jui Hsieh, Nagarajan Natarajan, and Inderjit Dhillon. 2015. PU learning for matrix completion. In *ICML*. PMLR, 2445–2453.

[28] Dino Ienco and Ruggero G Pensa. 2016. Positive and unlabeled learning in categorical data. *Neurocomputing* 196 (2016), 113–124.

[29] Christopher Jung, Michael Kearns, Seth Neel, Aaron Roth, Logan Stapleton, and Zhiwei Steven Wu. 2019. Eliciting and enforcing subjective individual fairness. *arXiv preprint arXiv:1905.10660* (2019).

[30] Ryuichi Kiryo, Gang Niu, Marthinus Christoffel du Plessis, and Masashi Sugiyama. 2018. Positive-Unlabeled Learning with Non-Negative Risk Estimator. In *NeurIPS*.

[31] Matthäus Kleindessner, Pranjal Awasthi, and Jamie Morgenstern. 2019. Fair k-center clustering for data summarization. In *ICML*. PMLR, 3448–3457.

[32] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. 2017. Counterfactual Fairness. In *NeurIPS*.

[33] Wee Sun Lee and Bing Liu. 2003. Learning with positive and unlabeled examples using weighted logistic regression. In *ICML*, Vol. 3. 448–455.

[34] Xiao-Li Li and Bing Liu. 2005. Learning from positive and unlabeled examples with different data distributions. In *European conference on machine learning*. Springer, 218–229.

[35] M. Lichman. 2013. UCI Machine Learning Repository. http://archive.ics.uci.edu/ml

[36] Aditya Krishna Menon and Robert C Williamson. 2018. The cost of fairness in binary classification. In *FAccT*. PMLR, 107–118.

[37] Fantine Mordelet and J-P Vert. 2014. A bagging SVM to learn from positive and unlabeled examples. *Pattern Recognition Letters* 37 (2014), 201–209.

[38] Walt L Perry. 2013. *Predictive policing: The role of crime forecasting in law enforcement operations*. Rand Corporation.

[39] David Pollard. 1990. Empirical processes: theory and applications. In *NSF-CBMS regional conference series in probability and statistics*. JSTOR, i–86.
[40] Mauricio Sadinle, Jing Lei, and Larry Wasserman. 2019. Least ambiguous set-valued classifiers with bounded error levels. *J. Amer. Statist. Assoc.* 114, 525 (2019), 223–234.
[41] Yuan-Hai Shao, Wei-Jie Chen, Li-Ming Liu, and Nai-Yang Deng. 2015. Laplacian unit-hyperplane learning from positive and unlabeled examples. *Information Sciences* 314 (2015), 152–168.
[42] Jiaming Song, Pratyusha Kalluri, Aditya Grover, Shengjia Zhao, and Stefano Ermon. 2019. Learning controllable fair representations. In *The 22nd International Conference on Artificial Intelligence and Statistics*. PMLR, 2164–2173.
[43] Latanya Sweeney. 2013. Discrimination in online ad delivery. *Commun. ACM* 56, 5 (2013), 44–54.
[44] Bhekisipho Twala. 2010. Multiple classifier application to credit risk assessment. *Expert Systems with Applications* 37, 4 (2010), 3326–3336.
[45] Sara A Van de Geer et al. 2008. High-dimensional generalized linear models and the lasso. *Annals of Statistics* 36, 2 (2008), 614–645.
[46] Gill Ward, Trevor Hastie, Simon Barry, Jane Elith, and John R Leathwick. 2009. Presence-only data and the EM algorithm. *Biometrics* 65, 2 (2009), 554–563.
[47] Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. 2019. PC-Fairness: A Unified Framework for Measuring Causality-based Fairness. In *NeurIPS*, Vol. 32.
[48] Bowei Yan, Sanmi Koyejo, Kai Zhong, and Pradeep Ravikumar. 2018. Binary classification with karmic, threshold-quasi-concave metrics. In *ICML*. PMLR, 5531–5540.
[49] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*. PMLR, 962–970.
[50] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning fair representations. In *ICML*. PMLR, 325–333.
[51] Tao Zhang, Tianqing Zhu, Mengde Han, Jing Li, Wanlei Zhou, and Philip S Yu. 2020. Fairness Constraints in Semi-supervised Learning. *arXiv preprint arXiv:2009.06190* (2020).

A PROOF OF LEMMA 1

Proof. Using the conditional probabilities, we have

$$\begin{aligned} f &= \mathbb{P}(L=1, Y=1 \mid X, S) \\ &= \mathbb{P}(Y=1 \mid X, S)\mathbb{P}(L=1 \mid Y=1, X, S) \\ &= \mathbb{P}(Y=1 \mid X, S) \cdot c \end{aligned}$$

The result follows by dividing both sides with c. □

B DERIVATION OF THE OPTIMAL FAIR CLASSIFIER

For EO, using weak duality we can write Equation (1) as:

$$\begin{aligned} &\min_{g\in\mathcal{G}} \max_{\lambda\in\mathbb{R}} \{\mathcal{R}(g) + \lambda_1(\mathbb{P}(g(X,S)=1 \mid Y=1, S=1) \\ &- \mathbb{P}(g(X,S)=1 \mid Y=1, S=0)) \\ &+ \lambda_2(\mathbb{P}(g(X,S)=0 \mid Y=0, S=1) - \mathbb{P}(g(X,S)=0 \mid Y=0, S=0))\} \\ &\geq \max_{\lambda\in\mathbb{R}} \min_{g\in\mathcal{G}} \{\mathcal{R}(g) + \lambda_1(\mathbb{P}(g(X,S)=1 \mid Y=1, S=1) \\ &- \mathbb{P}(g(X,S)=1 \mid Y=1, S=0)) \\ &+ \lambda_2(\mathbb{P}(g(X,S)=0 \mid Y=0, S=1) - \mathbb{P}(g(X,S)=0 \mid Y=0, S=0))\} \end{aligned} \tag{8}$$

In the PUL setting, based on Lemma 1, we can write:

$$\begin{aligned}
\mathbb{P}(g(X,S)=1 \mid Y=1,S=1) &= \frac{\mathbb{P}(g(X,S)=1,Y=1 \mid S=1)}{\mathbb{P}(Y=1 \mid S=1)} \\
&= \frac{\mathbb{E}_{X|S=1}[g(X,1)f(X,1)]}{c\mathbb{P}(Y=1 \mid S=1)} \\
\mathbb{P}(g(X,S)=1 \mid Y=1,S=0) &= \frac{\mathbb{P}(g(X,S)=1,Y=1 \mid S=0)}{\mathbb{P}(Y=1 \mid S=0)} \\
&= \frac{\mathbb{E}_{X|S=0}[g(X,0)f(X,0)]}{c\mathbb{P}(Y=1 \mid S=0)} \\
\mathbb{P}(g(X,S)=0 \mid Y=0,S=1) &= \frac{\mathbb{P}(g(X,S)=0,Y=0 \mid S=1)}{\mathbb{P}(Y=0 \mid S=1)} \\
&= \frac{\mathbb{E}_{X|S=1}[(1-g(X,1))(1-f(X,1)/c)]}{\mathbb{P}(Y=0 \mid S=1)} \\
\mathbb{P}(g(X,S)=0 \mid Y=0,S=0) &= \frac{\mathbb{P}(g(X,S)=0,Y=0 \mid S=0)}{\mathbb{P}(Y=0 \mid S=0)} \\
&= \frac{\mathbb{E}_{X|S=0}[(1-g(X,0))(1-f(X,0)/c)]}{\mathbb{P}(Y=0 \mid S=0)}
\end{aligned} \tag{9}$$

The risk function is:

$$\begin{aligned}
\mathcal{R}(g) =& \mathbb{P}(g(X,S) \neq Y) \\
=& \mathbb{P}(g(X,S)=0,Y=1) + \mathbb{P}(g(X,S)=1,Y=0) \\
=& \mathbb{P}(g(X,S)=1) + \mathbb{P}(Y=1) - 2\mathbb{P}(g(X,S)=1,Y=1) \\
=& \mathbb{P}(Y=1) + \mathbb{E}[g(X,S)] \\
& - 2\mathbb{E}\left[\mathbf{1}_{\{g(X,S)=1,Y=1\}} \mid S=1\right]\mathbb{P}(S=1) \\
& - 2\mathbb{E}\left[\mathbf{1}_{\{g(X,S)=1,Y=1\}} \mid S=0\right]\mathbb{P}(S=0) \\
=& \mathbb{P}(Y=1) + \mathbb{E}[g(X,S)] \\
& - \frac{2}{c}\mathbb{E}_{X|S=1}[g(X,1)f(X,1)]\mathbb{P}(S=1) \\
& - \frac{2}{c}\mathbb{E}_{X|S=0}[g(X,0)f(X,0)]\mathbb{P}(S=0) \\
=& \mathbb{P}(Y=1) - \mathbb{E}_{X|S=1}[g(X,1)(\frac{2}{c}f(X,1)-1)]\mathbb{P}(S=1) \\
& - \mathbb{E}_{X|S=0}[g(X,0)(\frac{2}{c}f(X,0)-1)]\mathbb{P}(S=0)
\end{aligned}$$

So the objective function can be simplified as:

$$\begin{aligned}
&\mathbb{P}(Y=1)\\
&+\mathbb{E}_{X|S=1}[g(X,1)(f(X,1)\left(\frac{\lambda_1}{c\mathbb{P}(Y=1\mid S=1)}-\frac{2}{c}\mathbb{P}(S=1)\right)+\mathbb{P}(S=1)\\
&+\frac{\lambda_2(f(X,1)/c-1)}{\mathbb{P}(Y=0\mid S=1)})]\\
&+\mathbb{E}_{X|S=1}\left[\frac{\lambda_2(1-c(X,1)/c)}{\mathbb{P}(Y=0\mid S=1)}\right]\\
&+\mathbb{E}_{X|S=0}[g(X,0)(f(X,0)\left(-\frac{\lambda_1}{c\mathbb{P}(Y=1\mid S=0)}-\frac{2}{c}\mathbb{P}(S=0)\right)+\mathbb{P}(S=0)\\
&+\frac{\lambda_2(1-f(X,0)/c)}{\mathbb{P}(Y=0\mid S=0)})]\\
&+\mathbb{E}_{X|S=0}\left[\frac{-\lambda_2(1-f(X,0)/c)}{\mathbb{P}(Y=0\mid S=0)}\right]
\end{aligned}$$

Since $g(X,S)\in\{0,1\}$, we can get the minimizer $g^*_{\boldsymbol{\lambda}}$:

$$\begin{aligned}
g^*_{\boldsymbol{\lambda}}(X,1)=&\mathbb{1}_{\{\frac{f(X,1)}{c}(1-\frac{\lambda_1}{\mathbb{P}(Y=1,S=1)})+(1-\frac{f(X,1)}{c})(1-\frac{\lambda_2}{\mathbb{P}(Y=0,S=1)})\geq 0\}}\\
g^*_{\boldsymbol{\lambda}}(X,0)=&\mathbb{1}_{\{\frac{f(X,0)}{c}(1+\frac{\lambda_1}{\mathbb{P}(Y=1,S=0)})-(1-\frac{f(X,0)}{c})(1+\frac{\lambda_2}{\mathbb{P}(Y=0,S=0)})\geq 0\}}
\end{aligned}\tag{10}$$

C PROOF OF LEMMA 2

Substituting the minimizer g^*_λ into the Lagrange function, we could see that the mappings of λ_1 and λ_2 are convex. We can write the first order optimality conditions of the objective function as:

$$\begin{aligned}
0\in&\partial_{\boldsymbol{\lambda}}\mathbb{E}_{X|S=1}[g(X,1)(f(X,1)\left(\frac{\lambda_1}{c\mathbb{P}(Y=1\mid S=1)}-\frac{2}{c}\mathbb{P}(S=1)\right)\\
&+\mathbb{P}(S=1)+\frac{\lambda_2(f(X,1)/c-1)}{\mathbb{P}(Y=0\mid S=1)})]\\
&+\partial_{\boldsymbol{\lambda}}\mathbb{E}_{X|S=1}\left[\frac{\lambda_2(1-c(X,1)/c)}{\mathbb{P}(Y=0\mid S=1)}\right]\\
&+\partial_{\boldsymbol{\lambda}}\mathbb{E}_{X|S=0}[g(X,0)(f(X,0)\left(-\frac{\lambda_1}{c\mathbb{P}(Y=1\mid S=0)}-\frac{2}{c}\mathbb{P}(S=0)\right)\\
&+\mathbb{P}(S=0)+\frac{\lambda_2(1-f(X,0)/c)}{\mathbb{P}(Y=0\mid S=0)})]\\
&+\partial_{\boldsymbol{\lambda}}\mathbb{E}_{X|S=0}\left[\frac{-\lambda_2(1-f(X,0)/c)}{\mathbb{P}(Y=0\mid S=0)}\right]
\end{aligned}$$

Based on Assumption 2, this subgradient is reduced to the gradient almost surely. So we have Equations (3):

$$\begin{aligned}
\frac{\mathbb{E}_{X|S=1}[g^*_{\boldsymbol{\lambda}^*}(X,1)f(X,1)]}{\mathbb{P}(Y=1\mid S=1)}&=\frac{\mathbb{E}_{X|S=0}[g^*_{\boldsymbol{\lambda}^*}(X,0)f(X,0)]}{\mathbb{P}(Y=1\mid S=0)}\\
\frac{\mathbb{E}_{X|S=1}[(c-f(X,1))(1-g^*_{\boldsymbol{\lambda}^*}(X,1))]}{\mathbb{P}(Y=0\mid S=1)}&=\frac{\mathbb{E}_{X|S=0}[(c-f(X,0))(1-g^*_{\boldsymbol{\lambda}^*}(X,0))]}{\mathbb{P}(Y=0\mid S=0)}
\end{aligned}$$

Combining it with Equation (9), we can see that $\mathbb{P}(g^*_{\boldsymbol{\lambda}^*}(X,S)=1\mid Y=y,S=1)=\mathbb{P}(g^*_{\boldsymbol{\lambda}^*}(X,S)=1\mid Y=y,S=0)$. In other words, $g^*_{\boldsymbol{\lambda}^*}$ satisfies Equalized Odds with respect to S and is thus fair. Therefore, we have $\mathcal{R}(g^*_{\boldsymbol{\lambda}^*})\geq\mathcal{R}(g^*)$

because g^* is defined as the optimal fair classifier to minimize the risk. Furthermore, since $(\boldsymbol{\lambda}^*, g^*_{\boldsymbol{\lambda}^*})$ is a solution to the dual problem, we have $\mathcal{R}(g^*_{\boldsymbol{\lambda}^*}) \leq \mathcal{R}(g^*)$ according to Equation (8). Therefore, we can conclude that $g^* = g^*_{\boldsymbol{\lambda}^*}$.

D PROOF OF THEOREM 1

Proof. Following the strategy of [12, 14], we first introduce an intermediate pseudo-estimator $\tilde{g}$ as follows:

$$\begin{aligned}
\tilde{g}_{\tilde{\boldsymbol{\lambda}}}(X,1) =& \mathbb{1}_{\{\hat{p}^+_{X,1}(1-\frac{\tilde{\lambda_1}}{\mathbb{E}_{X|S=1}[\hat{p}^+_{X,1}]\mathbb{P}(S=1)})+(1-\hat{p}^+_{X,1})(1-\frac{\tilde{\lambda_2}}{(1-\mathbb{E}_{X|S=1}[\hat{p}^+_{X,1}])\mathbb{P}(S=1)})\geq 0\}} \\
\tilde{g}_{\tilde{\boldsymbol{\lambda}}}(X,0) =& \mathbb{1}_{\{\hat{p}^+_{X,0}(1+\frac{\tilde{\lambda_1}}{\mathbb{E}_{X|S=0}[\hat{p}^+_{X,0}]\mathbb{P}(S=0)})-(1-\hat{p}^+_{X,0})(1+\frac{\tilde{\lambda_2}}{(1-\mathbb{E}_{X|S=0}[\hat{p}^+_{X,0}])\mathbb{P}(S=0)})\geq 0\}}
\end{aligned} \tag{11}$$

where $\tilde{\boldsymbol{\lambda}}$ satisfies:

$$\begin{aligned}
\frac{\mathbb{E}_{X|S=1}[\tilde{g}(X,1)\hat{f}(X,1)]}{\mathbb{E}_{X|S=1}[\hat{p}^+_{X,1}]} &= \frac{\mathbb{E}_{X|S=0}[\tilde{g}(X,0)\hat{f}(X,0)]}{\mathbb{E}_{X|S=0}[\hat{p}^+_{X,0}]} \\
\frac{\mathbb{E}_{X|S=1}[(\hat{c}-\hat{f}(X,1))(1-\tilde{g}(X,1))]}{1-\mathbb{E}_{X|S=1}[\hat{p}^+_{X,1}]} &= \frac{\mathbb{E}_{X|S=0}[(\hat{c}-\hat{f}(X,0))(1-\tilde{g}(X,0))]}{1-\mathbb{E}_{X|S=0}[\hat{p}^+_{X,0}]}
\end{aligned} \tag{12}$$

Comparing this pseudo-estimator $\tilde{g}$ with the theoretical ideal g^* in Equation (2) and our estimator $\hat{g}$ in Equation (7), we can see that $\tilde{g}$ knows the marginal distribution of (X, S). That is, it has precise information on the distributions p_S and $p_{X|S}$. It can be seen as a nearly-idealized version of $\hat{g}$ where the uncertainty in it is only induced by the estimator $\hat{p}^+_{X,S}$.

To demonstrate our proposed method is asymptotically optimal, we can upper bound the excess risk by expressing it as a sum of two terms, $\mathbb{E}[\mathcal{R}(\tilde{g})] - \mathcal{R}(g^*) + \mathbb{E}[\mathcal{R}(\hat{g}) - \mathcal{R}(\tilde{g})]$. The first term can be bounded by the l_1 distance between $\hat{p}^+_{X,S}$ and $p^+_{X,S}$. Based on the first part of Assumption 3 that $\hat{p}^+_{X,S}$ is consistent as $n_U, n_L \to \infty$, it converges to zero. For the second term, comparing the upper bound on $\mathcal{R}(\hat{g})$ and the lower bound on $\mathcal{R}(\tilde{g})$, we can upper bound the difference $\mathcal{R}(\hat{g}) - \mathcal{R}(\tilde{g})$ and show that $\lim_{n_U,n_L\to\infty}\mathbb{E}[\mathcal{R}(\hat{g}) - \mathcal{R}(\tilde{g})] \to 0$ based on the law of large numbers and the second part of Assumption 3.

To demonstrate our proposed framework is asymptotically fair, we can first upper bound the unfairness with the triangle inequality by considering $TPR^{(1)}$ and $TPR^{(0)}$ respectively and their estimators:

$$\begin{aligned}
|TPR^{(1)} - TPR^{(0)}| \leq& |\frac{\mathbb{E}_{X|S=1}[\hat{f}(X,1)\hat{g}(X,1)]}{\mathbb{E}_{X|S=1}[\hat{f}(X,1)]} - \frac{\mathbb{E}_{X|S=0}[\hat{f}(X,0)\hat{g}(X,0)]}{\mathbb{E}_{X|S=0}[\hat{f}(X,0)]}| \\
&+ |\frac{\mathbb{E}_{X|S=1}[\hat{f}(X,1)\hat{g}(X,1)]}{\mathbb{E}_{X|S=1}[\hat{f}(X,1)]} - \frac{\hat{\mathbb{E}}_{X|S=1}[\hat{f}(X,1)\hat{g}(X,1)]}{\hat{\mathbb{E}}_{X|S=1}[\hat{f}(X,1)]}| \\
&+ |\frac{\mathbb{E}_{X|S=0}[\hat{f}(X,0)\hat{g}(X,0)]}{\mathbb{E}_{X|S=0}[\hat{f}(X,0)]} - \frac{\hat{\mathbb{E}}_{X|S=0}[\hat{f}(X,0)\hat{g}(X,0)]}{\hat{\mathbb{E}}_{X|S=0}[\hat{f}(X,0)]}|
\end{aligned}$$

$|FPR^{(1)} - FPR^{(0)}|$ can be processed in a similar way. We can then prove that $\mathbb{E}[\Delta(\hat{g})] \leq \mathbb{E}\left[\hat{\Delta}(\hat{g})\right] + o_{U,L}(1)$ using the second part of Assumption 3. Using the consistency and continuity of $\hat{p}^+_{X,S}$ in Assumption 3, and means of theory of empirical processes [39], we can have $\lim \mathbb{E}\left[\hat{\Delta}(\hat{g})\right]$ converges to zero almost surely, which concludes the proof. □

E IMPLEMENTATION DETAILS

The three data sets we use can be obtained as follows:

- COMPAS is available at https://github.com/propublica/

compas-analysis.

- German Credit is available at https://archive.ics.uci.edu/ml/machine-learning-databases/statlog/german/.
- Drug is available at https://archive.ics.uci.edu/ml/datasets/Drug+Review+Dataset+%28Drugs.com%29

The hyper-parameters are found by grid search and we report the best results in terms of classification performance. More details of different models (Lin.LR, Support Vector Machine with polynomial kernel (SVM) and Multilayer Perceptron (MLP)) are shown as follows:

- For SVM, C = 0.1, degree = 2, γ = 2.
- For Lin.SVM, C = 10, tolerance = 1e-4.
- For Lin.LR, C = 1, solver = 'lbfgs', max iteration=1000.
- For MLP, we use the ReLU activation function and Adam optimizer. For hyper-parameters, regularization term parameter=1e-4, learning rate=1e-3. The hidden layers sizes in MLP are set to $(8, 16, 2)$ for COMPAS, $(24, 48, 2)$ for German Credit and $(12, 24, 2)$ for Drug.

For baselines, we use the code provided by original authors and Fairlearn (https://fairlearn.org/), and carefully tune all the hyper-parameters for the best classification performance.

F AUXILIARY EXPERIMENTAL RESULTS

Table 4. Average results and standard deviation on COMPAS of 10 runs. Labeling rate is 90%.

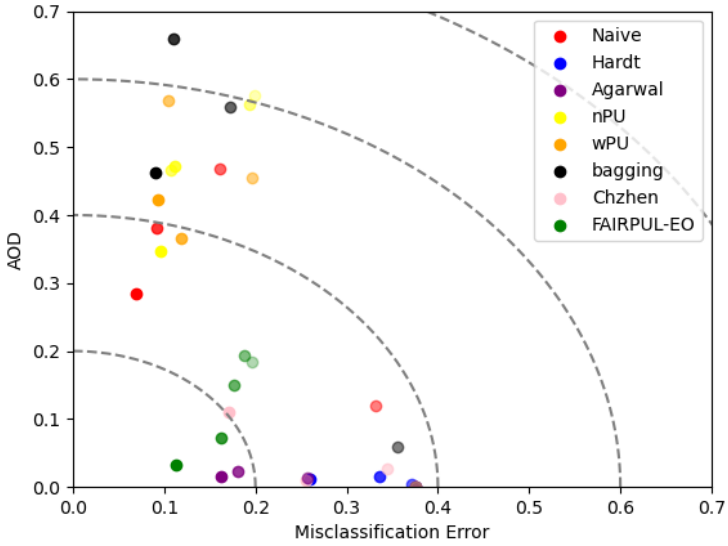
Models		F1	AOD	EOD
Lin.LR	Oracle	0.621±0.015	0.236±0.033	0.293±0.045
	Naïve	0.552±0.023	0.240±0.048	0.311±0.054
	+Agarwal	0.507±0.019	0.032±0.021	0.037±0.036
	+Hardt	0.357±0.042	0.022±0.012	0.027±0.019
	+Chzhen	0.491±0.021	0.018±0.015	0.014±0.010
	uPU	0.635±0.010	0.075±0.048	0.096±0.028
	wPU	0.588±0.036	0.278±0.164	0.306±0.051
	Bagging	0.559±0.020	0.293±0.223	0.335±0.037
	FairPUL-EO	0.656±0.016	0.015±0.008	0.013±0.010
	FairPUL-EOP	0.658±0.012	0.016±0.007	0.009±0.008
SVM	Oracle	0.628±0.019	0.287±0.127	0.246±0.038
	Naïve	0.553±0.022	0.299±0.102	0.324±0.073
	+Agarwal	0.491±0.037	0.046±0.032	0.035±0.032
	+Hardt	0.358±0.047	0.028±0.021	0.030±0.023
	+Chzhen	0.478±0.033	0.032±0.012	0.040±0.034
	uPU	0.644±0.012	0.186±0.086	0.134±0.023
	wPU	0.565±0.027	0.298±0.076	0.315±0.055
	Bagging	0.648±0.013	0.249±0.048	0.286±0.035
	FairPUL-EO	0.646±0.013	0.012±0.006	0.015±0.013
	FairPUL-EOP	0.648±0.012	0.016±0.012	0.012±0.004
MLP	Oracle	0.625±0.008	0.239±0.068	0.264±0.051
	Naïve	0.592±0.018	0.248±0.098	0.299±0.066
	+Hardt	0.457±0.038	0.036±0.016	0.029±0.024
	+Chzhen	0.548±0.027	0.032±0.029	0.024±0.017
	uPU	0.660±0.012	0.198±0.087	0.230±0.067
	wPU	0.606±0.021	0.267±0.138	0.295±0.064
	nnPU	0.659±0.023	0.248±0.243	0.289±0.062
	FairPUL-EO	0.663±0.014	0.032±0.017	0.040±0.034
	FairPUL-EOP	0.666±0.023	0.036±0.018	0.034±0.012

Table 5. Post-processing FairPUL-EOP vs In-processing/Pre-processing on German. The best results for each metric are marked bold.

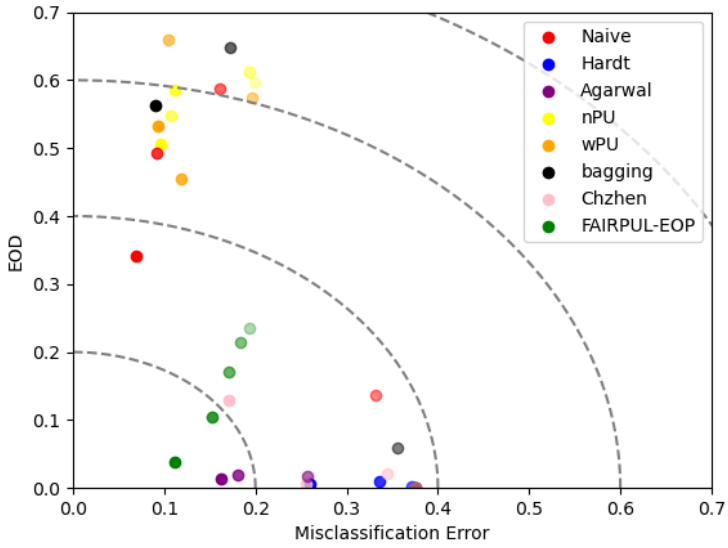
Rates	Models		F1	EOD
100%	Lin.SVM	Naïve	0.565±0.051	0.061±0.042
		In	0.541±0.050	0.051±0.044
		Pre	0.520±0.057	0.093±0.069
		FairPUL	0.591±0.035	0.056±0.049
	Lin.LR	Naïve	0.557±0.054	0.064±0.053
		In	0.540±0.057	0.071±0.064
		Pre	0.533±0.057	0.062±0.036
		FairPUL	0.609±0.038	0.051±0.041
90%	Lin.SVM	Naïve	0.471±0.048	0.099±0.052
		In	0.470±0.033	0.077±0.051
		Pre	0.414±0.062	0.078±0.074
		FairPUL	0.580±0.020	0.034±0.027
	Lin.LR	Naïve	0.507±0.045	0.132±0.037
		In	0.486±0.057	0.069±0.044
		Pre	0.479±0.052	0.125±0.072
		FairPUL	0.606±0.034	0.062±0.047

Table 6. Comparison results of different labeling rates on German.

Rates	Models		F1	EOD
100%	Lin.SVM	Naïve	0.565±0.051	0.061±0.042
		FairPUL	0.591±0.035	0.056±0.049
	Lin.LR	Naïve	0.557±0.054	0.064±0.053
		FairPUL	0.609±0.038	0.051±0.041
90%	Lin.SVM	Naïve	0.557±0.054	0.064±0.053
		FairPUL	0.580±0.020	0.034±0.027
	Lin.LR	Naïve	0.507±0.045	0.132±0.037
		FairPUL	0.606±0.034	0.062±0.047
80%	Lin.SVM	Naïve	0.193±0.136	0.044±0.051
		FairPUL	0.527±0.047	0.033±0.029
	Lin.LR	Naïve	0.401±0.046	0.083±0.055
		FairPUL	0.605±0.027	0.056±0.040
50%	Lin.SVM	Naïve	0.000±0.000	0.000±0.000
		FairPUL	0.481±0.042	0.019±0.018
	Lin.LR	Naïve	0.106±0.036	0.053±0.026
		FairPUL	0.593±0.036	0.044±0.019



(a) FairPUL-EO



(b) FairPUL-EOP

Fig. 1. Test classification error, AOD and EOD under different labeling rates. The dots with higher transparency correspond to the results under lower labeling rates.

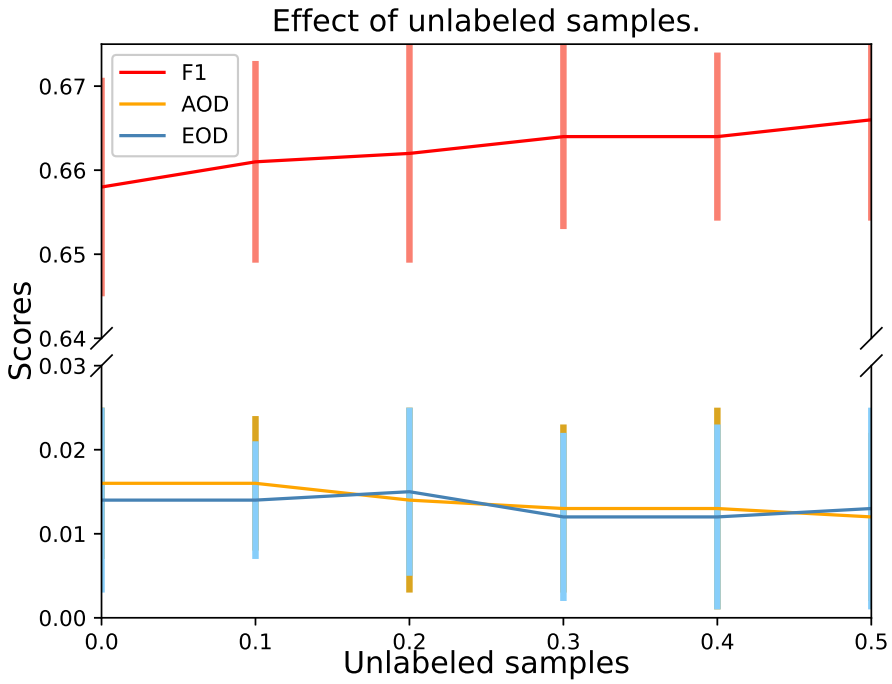


Fig. 2. F1 scores, AOD and EOD using different numbers of unlabeled examples. Results are averaged over 10 runs on COMPAS.