

Fair Selection through Kernel Density Estimation

Abstract—With the prevalence of machine learning in many high-stakes decision-making processes, e.g., hiring and admission, it is important to take fairness into consideration when practitioners design and deploy machine learning models. Although many approaches have been developed for fair machine learning, most of them focus on classification. In this paper, we target a notable but under-explored task, selection, where the number of selected individuals cannot exceed a pre-defined budget, such as employee hiring or university admission with limited positions or capabilities. In the selection task, existing fairness notions designed for classification are not suitable. In particular, our experimental results show that the selection models subject to common fairness notions may still make biased predictions against the underrepresented group. Hence, we propose a novel fairness notion, Selection Parity, which captures the demographic diversity among the selected groups in this restricted selection problem. Since the selection of qualified individuals with a fixed budget is non-differentiable, existing fairness regularization terms cannot be directly integrated with the selection task. To close the gap, we develop a novel in-processing framework named Fair Selection with the Differentiable Distribution Difference constraint (FS-DD), which incorporates a differentiable constraint into the training process and produces fair decisions for selection problems. Our theoretical analysis shows that common fairness metrics are bounded by the proposed Distribution Difference measurement. In other words, the FS-DD framework can guarantee fairness with regard to the common existing fairness metrics. We evaluate the performance of our method as well as several baselines on four real-world datasets. The experimental results demonstrate that the proposed method achieves fairness in various selection settings. In addition, the proposed method has a better fairness-accuracy trade-off compared with existing baseline methods.

Index Terms—fairness, selection, kernel density estimation, algorithmic bias.

I. INTRODUCTION

Machine learning algorithms have been widely used in many fields, which often encounter concerns about bias incurred by algorithms. Since the classic machine learning models aim to maximize the predictive performance, e.g., accuracy, the models may produce unwanted adverse bias when they are deployed in many high-stake scenarios [34], e.g., hiring, admission, banking, etc. As the algorithmic bias becomes a public concern [29], it is imperative to ensure the machine learning model is accurate as well as subject to the fairness requirements. In order to measure and eliminate the unfair/discriminatory effects brought by the machine learning models, many classification-based fairness notions and techniques have been proposed, using various statistical concepts to customize classic classification algorithms. Existing fairness notions can be mainly categorized into two sets: group fair-

ness and individual fairness. Group fairness [5], [19], [36], [37] requires that sensitive information, e.g., race, gender, or disabilities, should be independent of the algorithmic decisions at the population level. Various metrics have been derived from the concept of group fairness, e.g., *risk difference*, *risk ratio*, *relative change*, *odds ratio*, and so on [39]. In addition, individual fairness has been explored in [33], [49], where similar individuals should receive similar decisions. We refer the readers to surveys, e.g., [34], [52], for details.

Existing methods for bias elimination are categorized into three types: pre-processing, in-processing, and post-processing. Pre-processing methods [14], [16], [20], [53] modify the historical training data to remove the potential prejudice and discrimination based on the defined fairness notions before the data are leveraged to train machine learning models. The in-processing methods [6], [9], [21], [22], [24], [25], [43], [47], [48] tweak the machine learning algorithms to ensure fair predictions. The methods for post-processing [4], [17], [23] correct the predictions produced by the vanilla machine learning models.

Different from the traditional classification tasks, the selection tasks attempt to identify the most qualified individuals from the candidates with a pre-defined selection budget. These tasks are ubiquitous, such as employee recruiting, college admission, and jury selection, where the positions or capacities are constrained and only a fixed number of individuals can be selected. The constraints on the number of selected individuals, known as **the selection budget**, make the selection tasks fundamentally different from the classification tasks which has no constraints on the positive decisions. Recently, several works [7], [11], [26]–[28], [35] have been proposed for achieving fairness in the selection tasks. Generally, they assume that a pre-trained score function is provided to quantify the qualification of an individual. After the qualification scores are acquired, a set of strategies, e.g., the Rooney Rule, have been proposed to ensure that individuals are equally treated and selected. However, the score function is hard to accurately developed and estimated. [12], [13], [28] investigated a setting the qualification scores have group-dependent bias and variance and partially addressed the challenge with novel selection strategies. Another challenge is the strategy-based bias elimination solutions for selection are intuitive. The trade-off between performance and fairness is not guaranteed. Finally, the selection task is highly connected with the classification problem whereas the fairness relationship between selection and classification is not implicitly explored. Hence, designing a fair selection model still remains an open problem.

In this paper, we formulate the fair selection problem and propose a general framework, namely Fair Selection with the Differentiable Distribution Difference constraint (FS-DD) which closes the gaps posed by the unique selection tasks. The framework formulates the fairness notion as the difference in score distributions, referred to as *distribution difference*, between the favorable group and the unfavorable group. Further, it shows this *distribution difference* is differentiable, which means it could be naturally incorporated into existing soft decision classifiers, e.g., logistic regression, support vector machine, or deep neural networks. Within this framework, we connect the prominent fairness metrics used in classification tasks, e.g., *risk difference* [37] or *difference w.r.t demographic parity* [8], with the proposed *distribution difference*. Our theoretical results show that those existing metrics are bounded by the proposed distribution difference. In other words, the constrained selection model guarantees that the predictions meet the fairness requirements defined by *risk difference* or *difference w.r.t. demographic parity*.

The contributions of this paper are summarized as follows.

- We show that the classic models that satisfy common fairness notions may still make biased predictions.
- To fill the gap, we introduce a novel fairness notion, *Selection Parity*, for the selection tasks, which captures the demographic diversity among the selected individuals. We further propose a novel framework where a fairness regularizer, *Distribution Difference*, is integrated with existing selection algorithms. We leverage kernel density estimation to approximate *distribution difference* with finite predictions from a score function. The estimated regularizer is differential and easy to solve when it is integrated with a loss function.
- We also investigate the connections between the *distribution difference*, *selection parity*, and other common fairness notions, e.g., *risk difference*, *difference w.r.t. demographic disparity*. The theoretical analysis shows that the common fairness notions are bounded by *distribution difference*.
- The experimental results on various real-world datasets show that the proposed framework outperforms the baselines by achieving fairness with similar accuracy performance and having better fairness-accuracy trade-offs with varying hyper-parameters.

II. RELATED WORK

Fair machine learning has been studied extensively in the literature. There are two tasks in fair machine learning in general, bias assessment and bias elimination. The task of bias assessment aims to evaluate adverse bias in existing machine learning models leveraging fairness notions. Various fairness notions have been proposed for uneven settings, including group fairness [5], [19], [36], [37], [42], [45], [50], [51], individual fairness [3], [33], [49], etc. The majority of notions are developed in the language of statistical independence between sensitive information and decisions. Based on those notions, a set of quantitative metrics have been developed, among which

demographic parity, *equalized odds*, and *equalized opportunity* have attracted increasing attention.

For the bias elimination task, there are three dominant approaches to building fair models, *pre-processing* which removes the prejudice by modifying the historical training data before model learning, *in-processing* which imposes fairness constraints in the training process, and *post-processing* which mitigates the adverse bias by tweaking the prediction of classic machine learning algorithms. Common pre-processing methods include *Massaging* [19], which changes the labels of some individuals around the decision boundaries to remove discrimination, *Reweighting* [5], which assigns weights to individuals to balance the majority and minority groups, and *Preferential Sampling* [20], which resamples subgroups to make the dataset discrimination-free. Some in-processing research [9], [22], [24], [25], [43], [48], [48] incorporates fairness constraints or regularizers into the objective functions in machine learning tasks. The post-processing methods correct predictions made by machine learning algorithms. Kamiran et al. [23] exploited the decision theory to adjust classifiers with soft decisions for fair decision-making. Hardt et al. [17] developed an optimization solution to adjust any learned predictor to remove discrimination according to *equalized odds*. Awasthi et al. [4] extended the optimized solution in [17] into the setting with imperfect sensitive information.

In addition to fair classification, there are several works studying algorithmic bias in other machine learning applications, e.g., recommendation, ranking [44], etc. Our work targets the selection problem. Unlike the well-studied classification problem, the number of acceptances (or the selected) in a selection task is limited due to constrained resources or restricted opening positions. Concerning algorithmic bias in selection problems, Kleinberg and Raghavan [28] have investigated the implicit bias problem in a selection task where the underrepresented group received implicitly unequal treatment and shown the *Rooney Rule*, a requirement that at least one member of an underrepresented group should be selected, can not only improve the representation of the affected group but also leads to higher payoffs. Emelianov et al. [11] have studied the implicit bias in selection problems where the underrepresented candidates have higher estimation variance. The authors study the γ -rule in two settings where the variance is known or unknown and show the connection between utility and fairness with the γ -rule. Dwork et al. [10] have studied the selection problem under sequential pipeline composition, where a ranking stage is followed by a selection stage and provided a rigorous framework for evaluating different types of fairness guarantees for pipelines. Khalili et al. [27] have studied the compatibility of fairness and privacy in the selection problem and shown the differentially private exponential mechanism is able to improve both fairness and privacy as a post-processing step. Emelianov et al. [11] have studied fairness in k -stage selection problem where additional features are observed at each stage. A probabilistic model is proposed to achieve local (per stage) and global (final stage) fairness. Khalili et al. [26] have considered a sequential

selection problem where the sequentially arrived applicants apply for a limited number of positions and the decision-maker accepts or declines an applicant using a pre-trained supervised model at each time step. Their results have shown that a fair pre-trained cannot guarantee the selection outcomes are fair. A post-processing approach has been proposed for sequential selection problems with privacy and fairness concerns simultaneously. Celis et al. [7] have also focused on the sequential selection problem where the selection decisions repeatedly occur over time. Their results have shown that the Rooney Rule reduces implicit bias with theoretical guarantees in sequential settings.

III. PRELIMINARIES

In this section, we present some preliminaries for the fair selection problem. We start with the fairness notions used in the traditional classification tasks, then present the formulation of traditional classification and selection. We briefly introduce the in-processing practice of fair classification where the bias is eliminated through differential constraints. Finally, we introduce Kernel Density Estimation which will be leveraged to build differential constraints for alleviating adverse bias in selection tasks.

Throughout this paper, we use \mathbf{X} to denote the feature and Y to denote the labels. The sensitive attribute is denoted by S . The value domains of \mathbf{X}, Y, S are represented by $\mathcal{X}, \mathcal{Y}, \mathcal{S}$. For the sake of simplicity, we only consider the binary case $S \in \mathcal{S} = \{s^+, s^-\}$ and $Y \in \mathcal{Y} = \{y^+, y^-\}$ where the favorable group is denoted by s^+ , the unfavorable group is denoted by s^- , the positive decision is denoted by y^+ , and the negative decision is denoted by y^- . It is worth pointing out that our method can be readily extended to multi-class and multi-sensitive feature cases, motivated by the definition of intersectional fairness [15]. The training data with N samples are denoted by $\mathcal{D} = \{(s_i, \mathbf{x}_i, y_i)\}_{i=1}^N$, where \mathbf{x}_i, y_i, s_i is the instance of features, labels, and sensitive attribute for the i -th individual.

A. Classification and Selection

Fair classification is well-studied in the machine learning field, where it tackles the challenge of learning a fair classifier while satisfying fairness requirements. The learning goal of classification is to find a classifier $\hat{Y} = f(\mathbf{X})$ that minimizes the empirical classification errors, given by

$$\min \mathbf{L}(f) = \min \sum_{i=1}^N \ell(f(\mathbf{x}_i), y_i), \quad (1)$$

where ℓ is a loss function, e.g., a sign function [31]. For the margin-based classifiers, e.g., logistic regression, support vector machine, and neural networks, one defines a classifier $Z = h(\mathbf{X})$ where Z indicates the marginal distance from the data samples to the decision boundary. By letting $f = \text{sign}(h)$, the classification loss function in Eq. 1 is formulated as follows

$$\ell(f(\mathbf{x}_i), y_i) = \ell(h(\mathbf{x}_i), y_i) = \phi(y_i l(\mathbf{x}_i)), \quad (2)$$

where ϕ is the margin-based differentiable loss function, e.g., the cross entropy function, the hinge function, or the softmax function. Eq. 2 is differentiable, which can be efficiently solved using existing optimization solvers, e.g., CVXPY, Gurobi, or deep learning frameworks, e.g., PyTorch and TensorFlow. f is called a hard-decision classifier as it directly produces the discrete labels, and h is called a soft-decision classifier, as the distance can be used to estimate the conditional probabilities and further make predictions based on probabilities [32].

The classic classifiers make independent predictions for new unseen data samples without any budget constraints for positive decisions. However, the decisions made in the real world have to be subject to budgets, e.g., the universities admit students based on their facility capacity, which changes the problem fundamentally. Thus, the classic classifier cannot meet the real-world requirement as the budget is not taken into consideration. In this paper, we consider a selection task where the decision-maker selects a subset of individuals from a candidate pool, subject to selection budgets (i.e., the maximal number of positive decisions is no larger than a pre-defined threshold m). The selection task is ubiquitous in the real world, such as employee hiring, college admission, and jury selection, where the maximum amount of selected individuals is fixed or bounded due to resource or position constraints. The selection task can be formulated as $g : \mathbf{X}, m \rightarrow \tilde{Y}$ where m candidates are selected (denoted by $\tilde{Y} = y^+$), and the rest are not selected (denoted by $\tilde{Y} = y^-$). Specifically, the selection model g has two concatenating components [28]. (1) A score function $r : \mathbf{X} \rightarrow Z, Z \in \mathbb{R}$ maps the sample features to a score, e.g., the qualification score for university admission. A candidate with a higher score implies he/she is more qualified than that one a lower score. Traditionally, the soft classifier $f : \mathbf{X} \rightarrow R$ is leveraged to learn a supervised score function from the historical data. (2) A selection criterion $c : Z, m \rightarrow \tilde{Y}, \tilde{Y} \in \{y^+, y^-\}$ selects the candidates based on the score Z_i of the i -th individual and makes prediction \tilde{Y}_i based on the threshold m . An intuitive criterion [41] is that candidates with higher qualification scores should be selected in higher priority than those with lower qualification scores. In particular, when a decision-maker attempt to select a fixed number m , the top- m candidates ranked based on the qualification score are selected and assigned positive decisions $\tilde{Y} = y^+$. The rest are assigned negative decisions $\tilde{Y} = y^-$. To this end, a selection model is defined as $\tilde{Y} = g(\mathbf{X}, m) = c(r(\mathbf{X}), m)$.

B. Fairness Notions

The traditional classifiers aim to minimize the misclassification rate (known as empirical risk minimization) given a dataset \mathcal{D} while the adverse bias might be incurred. To assess the discrimination incurred in classification tasks, abundant fairness notions have been proposed in the literature, such as *demographic parity*, *disparity treatment*, *disparity impact*, *disparity mistreatment*, *equalized odds*, and so on. In this section, we introduce the most widely-used notion, *demographic parity*, for classification tasks.

Definition 1. (Demographic Parity) A classifier $\hat{Y} = f(\mathbf{X})$ is said to satisfy demographic parity if its prediction \hat{Y} is independent of the sensitive attribute S .

Usually, *demographic parity* is quantified with regard to *risk difference (RD)* [38] or *difference w.r.t. demographic parity (DP)* [8] (originally known as elift_d in [37]). *Risk difference (RD)* measures the portion difference of the positive decisions between the favorable group ($S = s^+$) and the unfavorable group ($S = s^-$). Based on the classification formulation in Eq. 1, the risk difference of a classifier $\hat{Y} = f(\mathbf{X})$ is expressed as

$$\text{RD}(f) := P(\hat{Y} = y^+ | S = s^+) - P(\hat{Y} = y^+ | S = s^-). \quad (3)$$

The classifier f is considered fair if the value of **RD** is smaller than a threshold, e.g., 0.05 in the Equality and Human Rights Commission (EHRC) [1].

Difference w.r.t. demographic parity (DP) is a metric capturing the difference between a demographic group, the favorable group as an example, and the population with regard to the positive decision rate. It can be formulated as

$$\text{DP}(f) := P(\hat{Y} = y^+ | S = s) - P(\hat{Y} = y^+), \forall s \in \mathcal{S}. \quad (4)$$

The measures of *demographic parity*, e.g., **RD** and **DP**, have been used to quantify the magnitude of bias in classification predictions. In addition, their differentiable variants are used as a fairness regularizer to eliminate the adverse bias in classification tasks. Zafar et al. [47], [48] have leveraged surrogate functions, e.g., linear and hinge functions, to apply differential fairness regularizers to a logistic regression classifier. Cho et al. [8] have adopted Kernel Density Estimator (KDE) to express the probabilities in **DP** as a differentiable function. However, existing differentiable regularizers for classification cannot guarantee fairness for the selection tasks as they only pose weak constraints, i.e., the expectation of difference between two groups should be small. In addition, the selection criterion $\tilde{Y} = c(Z, m)$ is non-differentiable as it selects the top- m individuals based on the qualification score derived from a machine learning model. Thus, the existing fairness regularization terms are incompatible with the selection tasks. To illustrate that existing metrics, e.g., *risk difference* and *difference w.r.t. demographic parity*, are incompatible with selection, we consider the **Law School** dataset and build several “fair” models (**FS-RD** and **FS-DP**) as the score function. Then, the selection threshold varies from $[0, 1]$ in the selection process. The result in Fig. 1 shows that both **FS-RD** and **FS-DP** indeed achieve fairness for the selection budget threshold range $[0.4, 0.55]$. However, they incur significant bias if the selection budget threshold is smaller than 0.4 or larger than 0.55 (as highlighted in Fig. 1).

C. Kernel Density Estimation (KDE)

Kernel Density Estimation (KDE) is a non-parametric method for continuous probability density estimation from finite data samples [40]. Let z_1, \dots, z_n be a set of independent and identically distributed examples drawn from a univariate

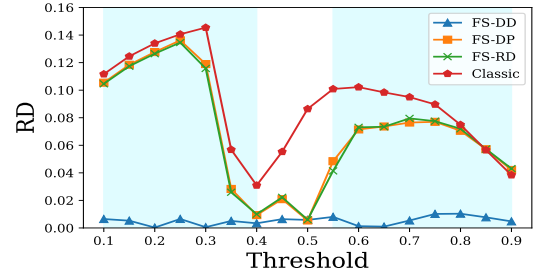


Fig. 1: Conventional fair methods (**FS-RD** and **FS-DP**) cannot guarantee fairness of selection in the **Law School Admission** dataset.

distribution with an unknown density p . One can estimate the continuous density function $\hat{p}(\cdot)$ for a variable Z using:

$$\hat{p}(z) := \frac{1}{nh} \sum_{i=1}^n K\left(\frac{z - z_i}{h}\right), \quad (5)$$

where $h > 0$ is a smoothing parameter called bandwidth and K is a non-negative kernel. A prominent kernel is a Gaussian function:

$$K(z) := \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{z^2}{2}\right). \quad (6)$$

IV. PROPOSED APPROACH

We propose a new fairness notion, *selection parity*, which captures the demographic diversity among the selected individuals given a fixed budget. As existing fairness metrics and derived regularization terms are incompatible with the selection tasks, we design a new metric, *distribution difference*, and its differentiable regularizer computed from the finite samples using kernel density estimation. Finally, we incorporate the differential distribution difference constraint into the selection objective and solve this problem efficiently.

A. Selection Parity

Although demographic parity has been widely used to evaluate and eliminate adverse bias in classification tasks, it is still inapplicable for selection tasks, due to the fluctuating selection budgets and the non-differentiable selection criterion. Inspired by the demographic parity, we propose a new metric, **selection parity**, for evaluating the fairness strength of the selection model $\tilde{Y} = g(\mathbf{X}, m)$. Selection parity requires the portion of the favorable group among the selected should be similar to the portion of that among the population, i.e.,

$$\text{SP}(g) = P(S = s^+ | \tilde{Y} = y) - P(S = s^+) \leq \epsilon, \forall \tilde{Y} \in \mathcal{Y}, \quad (7)$$

where ϵ is a small user-defined threshold. A selection model is considered fair if the selection parity (Eq. 7) is met.

B. Problem Formulation

Given the proposed fair selection metric, selection parity, we formulate the fair selection problem. Technically, we aim to learn a selection model $g : \mathbf{X}, m \rightarrow \tilde{Y}$ with a selection budget

m , which is also subject to selection disparity in Eq 7. Since the selection budget is fluctuating and the selection criterion m is non-differentiable, we attempt to learn a fair selection model g by enforcing fairness in the score function r . To this end, we aim to learn a fair score function r is expected to be fair w.r.t any selection budget m .

C. Fair Selection Using Kernel Density Estimation

To eliminate the gap between the fair score function and fairness requirement on the selection decisions, we define a fairness metric, **distribution difference** (DD), for the score function r . Technically, DD capture the maximal gaps of the qualification score distributions between the favorable group ($S = s^+$) and the unfavorable group ($S = s^-$), i.e.,

$$\mathbf{DD}(r) =: \max_z \left| p_{s^+}(z) - p_{s^-}(z) \right|, \quad (8)$$

where p_{s^+} and p_{s^-} are the distributions of two demographic groups and can be estimated using kernel density estimation. We leverage the KDE trick to estimate $p_{s^+}(z)$ and $p_{s^-}(z)$ for three-fold reasons. (1) The score variable $Z = r(\mathbf{X})$ is a finite set of samples. Z is an output of a soft classifier, e.g., a vector of probabilities of being positive in logistic regression for n samples. KDE is able to infer the continuous distribution of p_{s^+} and p_{s^-} from the finite samples. (2) The distribution functions estimated by KDE are differential and thus compatible with existing soft classification models, e.g. logistic regression, support vector machine, and deep neural networks. The proposed fairness metric, distribution difference, can be easily integrated with those prevalent machine learning models and efficiently implemented using PyTorch or TensorFlow. (3) Our theoretical results show that DD designed for the score function has a guarantee for ensuring SP, RD, and DP if these metrics are evaluated on the selection prediction \hat{Y} . The theoretical results are presented in the next section.

To this end, we aim to train a fair score function r from \mathcal{D} with the DD constraint:

$$\min \left\{ \lambda \mathbf{L} \left[c(r(\mathbf{X}), m), Y \right] + \mathbf{DD}(r(\mathbf{X})) \right\} \quad (9)$$

where λ is a hyper-parameter to balance the utility loss and fairness, $\mathbf{L}(\cdot)$ is the empirical loss of the selection, and $\mathbf{DD}(\cdot)$ is the distribution difference among two groups, given in Eq. 8.

V. THEORETICAL DISCUSSION

The selection model g makes predictions based on the features \mathbf{X} and the budget m . Without loss of generalization, we simply the selection criterion that a positive decision $\hat{Y}_\tau = y^+$ is given if the score Z is larger than a threshold value τ . It is worth pointing out that the threshold τ can be derived from the budget m , i.e., τ is the m -th individual's score z_m if the score S is sorted in descending order. Note that the sort operation is not differentiable in general. As we are evaluating the selection models rather than training the model, the differentiable property is not required.

A. Connection between RD and DD

One can assess \hat{Y}_τ with *risk difference* and denotes the magnitude as $\mathbf{RD}(\hat{Y}_\tau)$. Obviously, if we traverse the values of $\tau \in [0, 1]$, the maximal value $\max_\tau \mathbf{RD}(\hat{Y}_\tau)$ of this magnitude for the selection decision is equivalent to the value of $\mathbf{DD}(s)$ for the score variable, i.e.,

$$\mathbf{DD}(r) =: \max_\tau \mathbf{RD} \left(c(r(\mathbf{X}), \tau) \right). \quad (10)$$

Theorem 1. *Given a dataset and a selection function g with any arbitrary threshold τ , the value of risk difference of the decisions \hat{Y} is bounded by the distribution difference of the score function $r : \mathbf{X} \rightarrow Z$.*

Proof. Given a selection function, we calculate the *risk difference* of its decision:

$$\begin{aligned} \mathbf{RD}(\hat{Y}_\tau) &= \int_\tau^1 (p_+(z) - p_-(z)) dz \\ &\leq \int_\tau^1 |p_+(z) - p_-(z)| dz \\ &\leq \int_\tau^1 \mathbf{DD} dz \\ &= \mathbf{DD}(1 - \tau) \end{aligned}$$

where τ is a threshold-based selection budget. \square

Based on Thm. 1, we have the following remarks.

Remark.

- If distribution difference is achieved on the score function, risk difference is guaranteed on the decisions made by a selection function.
- The feasible solution of Eq. 9 satisfies the fairness requirement defined by risk difference.

B. Connection between SP and DD

Theorem 2. *Given a dataset and a selection function g with any arbitrary threshold τ , the value of selection parity of the decisions \hat{Y} is bounded by the distribution difference of the score function $r : \mathbf{X} \rightarrow Z$.*

Proof. Given Thm. 1 and the definition of SP in Eq. 7, we have

$$\begin{aligned} SP &= P(s^+ | \hat{Y}_\tau^+) - P(s^+) \\ &= (1 - P(s^+)) P(s^+ | \hat{y}_\tau^+) - P(s^+) (1 - P(s^+ | \hat{Y}_\tau^+)) \\ &= \frac{P(s^+) P(s^-)}{P(\hat{y}_\tau^+)} \left[P(\hat{y}_\tau^+ | P(s^+)) - P(\hat{y}_\tau^+ | s^-) \right] \\ &= \frac{P(s^+) P(s^-)}{P(\hat{y}_\tau^+)} \mathbf{RD}(\hat{y}_\tau^+) \\ &\leq \frac{\mathbf{DD}(1 - \tau)}{c_1} \end{aligned}$$

where $c_1 = \frac{P(s^+) P(s^-)}{\int_\tau^1 p(z) dz}$ is a constant for a given dataset. \square

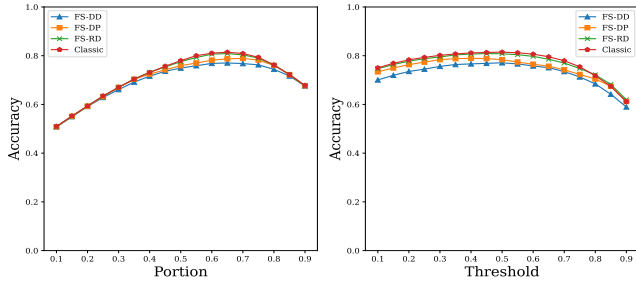


Fig. 2: Selection models with varying portion-based/threshold-based budgets have similar accuracy on **Census Adult**.

C. Connection between DP and DD

Theorem 3. Given a dataset and a selection function g with any arbitrary threshold τ , the value of difference w.r.t demographic parity of the decisions \hat{Y} is bounded by the distribution difference of the score function $r : \mathbf{X} \rightarrow \mathbb{Z}$.

Proof. Let us start with $P(\tilde{y}^+|s^+) - P(\tilde{y}^+)$:

$$\begin{aligned} P(\tilde{y}^+|s^+) - P(\tilde{y}^+) &= \frac{P(s^+|\tilde{y}^+)P(\tilde{y}^+)}{P(s^+)} - P(\tilde{y}^+) \\ &= P(\tilde{y}^+) \frac{P(s^+|\tilde{y}^+) - P(s^+)}{P(s^+)} \\ &= P(\tilde{y}^+) \mathbf{SP} \end{aligned}$$

We have similar result for $P(\tilde{y}^+|s^-) - P(\tilde{y}^+)$

$$P(\tilde{y}^+|s^-) - P(\tilde{y}^+) = -\frac{P(\tilde{y}^+)P(s^+)}{P(s^-)} \mathbf{SP}$$

For the sake of simplicity, we assume \mathbf{SP} is non-negative, thus

$$\mathbf{DP} \leq c_2 \mathbf{DD}(1 - \tau)$$

where $c_2 = \frac{P(\tilde{y}^+)(P(s^+) - P(s^-))}{P(s^-)}$ is a constant for a given dataset. \square

VI. EXPERIMENTS

In this section, we implement the proposed in-processing fair selection algorithm in Eq. 9, namely **FS-DD**, and compare it with several baseline methods. We evaluate those methods using four real-world datasets. All the methods are implemented in PyTorch.

A. Datasets

Census Adult Dataset [30] is extracted from the 1994 *Census database*, consisting of 48,842 samples (32,561 training examples and 16,281 testing examples) with 11 attributes including *age*, *education*, *sex*, *occupation*, *income*, *marital status* etc. This dataset is for classification and the decision attribute is *income*, i.e., whether the income is larger than 50k. **COMPAS Dataset** [29] is a database containing 5,278 criminal records from Broward County between 2013 and 2014, including features like *age*, *criminal history*, and more. The

task of interest is to predict the recidivism of criminals. We consider *race* as the sensitive attribute.

Law School Admission Dataset [2] provides information on over 100,000 individual applications. *LSAT score*, *undergraduate GPA*, *gender*, and more are included. The associated task is to predict whether applicants get admission from a law school. We choose *race* to be the sensitive attribute.

Credit Card Default [46] contains 30,000 records of information on *default payments*, *demographic factors*, *credit data*, *history of payment*, and *bill statements* of credit card clients in Taiwan. The task is to predict whether a credit card user declares a default in the coming month. We choose *race* and *gender* to be the sensitive attribute.

B. Baselines and the Proposed Method

We implement three baseline methods and compare them with the proposed method. The three baselines are a set of classifiers with different constraints in the score function. 1) **Classic** is unconstrained, 2) **FS-RD** is with a **RD** constraint [43], 3) **FS-DP** is with a **DP** constraint [8]. Then the selection criterion is applied to these score functions, i.e., the top- m qualified individuals are selected with positive decisions while the rest receive negative decisions.

The proposed fair selection method is named **FS-DD** where the **DD** constraint is integrated into a soft classifier depicted as Eq. 9. In our implementation, we adopt the Huber Loss [18] for the trackability of the absolute function in Eq. 8.

In the baselines and the proposed method, we adopt a three-layer neural network as the score function r . For the sake of simplicity, we map the qualification score Z into a range of $[0, 1]$. In the selection process g , we consider two types of budgets, portion-based and threshold-based. The portion-based selection budget requires that a specified percent, e.g., 10%, of individuals with top qualification scores are selected regardless of the total number of the application pool. The threshold-based selection budget sets up a hard cut-off point within the range of $[0, 1]$ for the qualification score such that any individual with a larger score than this threshold would be selected.

C. Experimental Results

We evaluate the performance of our method in terms of fairness and accuracy. For a fair comparison, we first find a set of regularization coefficients of the proposed approach and the baselines such that they have similar accuracy. We compare their fairness performance with varying portion-based budgets and threshold-based budgets. Then, we tune the regularization coefficients to compare the fairness-accuracy trade-offs of the methods.

1) *The comparison of fairness with similar accuracy:* We select a set of the regularization coefficients for three baselines and **FS-RD** such that the models' accuracy performance is similar, as shown in Fig. 2. Then, we evaluate the fairness magnitude of the models either with a varying threshold-based budget or a varying portion-based budget, with regard to various fairness metrics. The results in Fig. 3 and Fig. 4 show

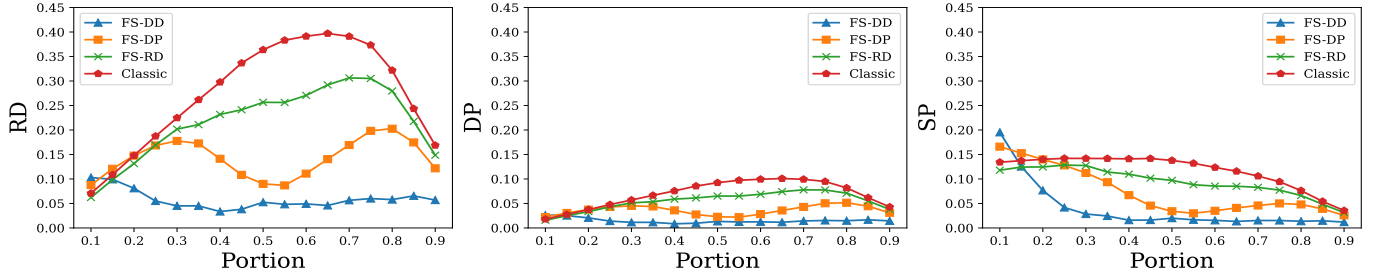


Fig. 3: Fairness assessment on portion-based selection models w.r.t various metrics.

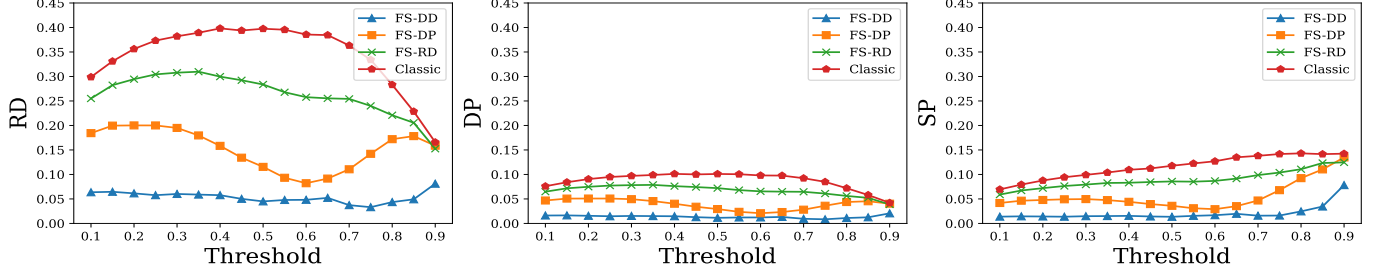


Fig. 4: Fairness assessment on threshold-based selection models w.r.t various metrics.

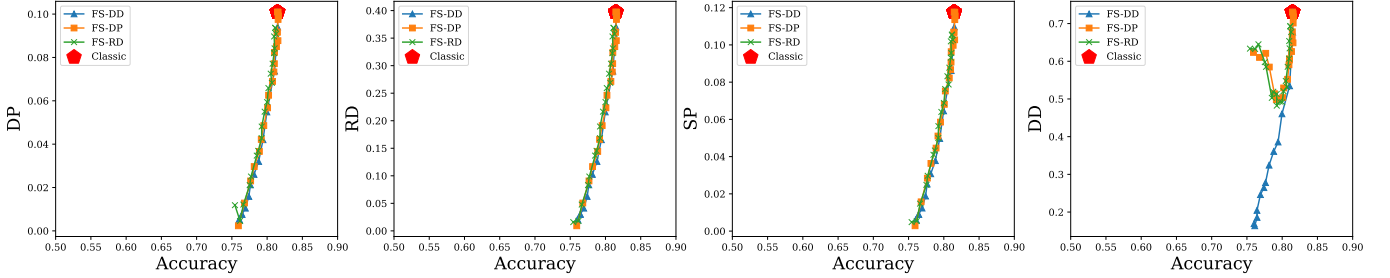


Fig. 5: Trade-off between accuracy and fairness with a fixed budget.

that the proposed method **FS-DD** has small values than other baselines in three fairness metrics, indicating the proposed method achieves better fairness at the same accuracy level.

2) *Fairness-accuracy trade-offs*: We fix the selection budget, i.e., the threshold $\tau = 0.5$ for four methods. Then, we tune the hyper-parameters of methods to investigate the relationship between the obtained fairness and accuracy. As shown in Fig. 5, four methods are able to achieve a good trade-off in existing metrics. With regard to **DD**, only the proposed method **FS-DD** achieves good trade-offs between accuracy and fairness.

We conduct experiments on other real datasets, e.g., **COMPAS**, **Law School Admission**, **Credit Card Default**. The experimental results are similar to **Census Adult** and skipped due to space limits.

VII. CONCLUSION

We investigated the fair selection problem where the selected are restricted due to hard requirements, e.g., positions or capabilities. We showed that existing fairness notions and fair learning methods could not guarantee fairness in the

selection tasks. To tackle this challenge, we proposed an in-processing framework, referred to as **FS-DD**. It incorporates the the differential distribution difference (**DD**) metric as a constraint such that it could be efficiently solved with gradient descent-based solvers. We theoretically analyzed the connections between **DD** and the existing fairness metrics, e.g., **RD** and **DP**, then illuminated their quantitative relationships. We proved that the proposed **FS-DD** framework has fairness guarantees with regard to existing fairness metrics. The experimental results showed the proposed method outperforms the three baselines. One limitation of this framework is the sensitivity of kernel density estimation. In future work, we will compare various kernel functions and develop robust and efficient in-processing fair selection models.

REFERENCES

- [1] Equality and Human Rights Commission.
- [2] LSAC national longitudinal bar passage study. LSAC research report author = Wightman, Linda F, date = 1998,.
- [3] Situation testing-based discrimination discovery: A causal inference approach. In *IJCAI*.

- [4] Pranjal Awasthi, Matthäus Kleindessner, and Jamie Morgenstern. Equalized odds postprocessing under imperfect group information. In *AISTATS'20*, 2020.
- [5] Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *ICDM Workshops 2009, IEEE International Conference on Data Mining Workshops, Miami, Florida, USA, 6 December 2009*.
- [6] Toon Calders and Sicco Verwer. Three naive Bayes approaches for discrimination-free classification.
- [7] L. Elisa Celis, Chris Hays, Anay Mehrotra, and Nisheeth K. Vishnoi. The effect of the rooney rule on implicit bias in the long term. In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*.
- [8] Jaewoong Cho, Gyeongjo Hwang, and Changho Suh. A fair classifier using kernel density estimation. In *NeurIPS'20*, 2020.
- [9] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Halifax, NS, Canada, August 13 - 17, 2017*.
- [10] Cynthia Dwork, Christina Ilvento, and Meena Jagadeesan. Individual fairness in pipelines. In *FORC'20*, 2020.
- [11] Vitalii Emelianov, George Arvanitakis, Nicolas Gast, Krishna P. Gummadi, and Patrick Loiseau. The price of local fairness in multistage selection. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI 2019, Macao, China, August 10-16, 2019*.
- [12] Vitalii Emelianov, Nicolas Gast, Krishna P. Gummadi, and Patrick Loiseau. On fair selection in the presence of implicit and differential variance.
- [13] Vitalii Emelianov, Nicolas Gast, Krishna P. Gummadi, and Patrick Loiseau. On fair selection in the presence of implicit variance. In *EC '20: The 21st ACM Conference on Economics and Computation, Virtual Event, Hungary, July 13-17, 2020*.
- [14] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *SIGKDD'15, 2015*. ACM, 2015.
- [15] James R. Foulds, Rashidul Islam, Kamrun Naher Keya, and Shimei Pan. An intersectional definition of fairness. In *ICDE'20*, 2020.
- [16] Sara Hajian and Josep Domingo-Ferrer. A Methodology for Direct and Indirect Discrimination Prevention in Data Mining.
- [17] Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, 2016.
- [18] Peter J Huber. Robust estimation of a location parameter.
- [19] Faisal Kamiran and Toon Calders. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*.
- [20] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 2012.
- [21] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Discrimination aware decision tree learning. In *ICDM 2010, the 10th IEEE International Conference on Data Mining, Sydney, Australia, 14-17 December 2010*.
- [22] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Discrimination aware decision tree learning. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*. IEEE, 2010.
- [23] Faisal Kamiran, Asim Karim, and Xiangliang Zhang. Decision theory for discrimination-aware classification. In *Proceedings of the 12nd IEEE International Conference on Data Mining (ICDM 2012)*, December 2012.
- [24] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2012, Bristol, UK, September 24-28, 2012. Proceedings, Part II*.
- [25] Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on, Vancouver, BC, Canada, December 11, 2011*.
- [26] Mohammad Mahdi Khalili, Xueru Zhang, and Mahed Abroshan. Fair sequential selection using supervised learning models. In *NeurIPS'21*, 2021.
- [27] Mohammad Mahdi Khalili, Xueru Zhang, Mahed Abroshan, and Somayeh Sojoudi. Improving fairness and privacy in selection problems. In *AAAI'21*, 2021.
- [28] Jon M. Kleinberg and Manish Raghavan. Selection problems in the presence of implicit bias. In *ITCS'18*, 2018.
- [29] Jeff Larson, Surya Mattu, Lauren Kirchner, and Julia Angwin. How We Analyzed the COMPAS Recidivism Algorithm - ProPublica.
- [30] M Lichman. UCI Machine Learning Repository.
- [31] Yi Lin. A note on margin-based loss functions in classification.
- [32] Yufeng Liu, Hao Helen Zhang, and Yichao Wu. Hard or Soft Classification? Large-Margin Unified Machines.
- [33] Binh Thanh Luong, Salvatore Ruggieri, and Franco Turini. K-NN as an implementation of situation testing for discrimination discovery and prevention. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '11*, New York, New York, USA, 2011.
- [34] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning.
- [35] Anay Mehrotra, Bary S. R. Pradelski, and Nisheeth K. Vishnoi. Selection in the presence of implicit bias: The advantage of intersectional constraints. In *FAccT '22: 2022 ACM Conference on Fairness, Accountability, and Transparency, Seoul, Republic of Korea, June 21 - 24, 2022*.
- [36] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Las Vegas, Nevada, USA, August 24-27, 2008*.
- [37] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. Measuring discrimination in socially-sensitive decision records. In *Proceedings of the SIAM International Conference on Data Mining, SDM 2009, April 30 - May 2, 2009, Sparks, Nevada, USA*.
- [38] Dino Pedreschi, Salvatore Ruggieri, and Franco Turini. A study of top-k measures for discrimination discovery. In *Proceedings of the ACM Symposium on Applied Computing, SAC 2012, Riva, Trento, Italy, March 26-30, 2012*.
- [39] Andrea Romei and Salvatore Ruggieri. A multidisciplinary survey on discrimination analysis.
- [40] Murray Rosenblatt. Remarks on some nonparametric estimates of a density function.
- [41] Sima Wolgast, Martin Bäckström, and Fredrik Björklund. Tools for fairness: Increased structure in the selection process reduces discrimination.
- [42] Yongkai Wu, Lu Zhang, and Xintao Wu. Counterfactual fairness: Unidentification, bound and algorithm. In *IJCAI'19*, 2019.
- [43] Yongkai Wu, Lu Zhang, and Xintao Wu. On convexity and bounds of fairness-aware classification. In *WWW'19*, 2019.
- [44] Yongkai Wu, Lu Zhang, and Xintao Wu. On discrimination discovery and removal in ranked data using causal graph. In *SIGKDD'18*, 2018.
- [45] Yongkai Wu, Lu Zhang, Xintao Wu, and Hanghang Tong. PC-Fairness: A unified framework for measuring causality-based fairness. In *NeurIPS'19*, 2019.
- [46] I-Cheng Yeh and Che-hui Lien. The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients.
- [47] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*.
- [48] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness constraints: Mechanisms for fair classification. In *AISTATS'17*, 2017.
- [49] Richard S Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. Learning fair representations. *ICML*, 2013.
- [50] Lu Zhang, Yongkai Wu, and Xintao Wu. A causal framework for discovering and removing direct and indirect discrimination. In *IJCAI'17*, 2017.
- [51] Lu Zhang, Yongkai Wu, and Xintao Wu. Causal modeling-based discrimination discovery and removal: Criteria, bounds, and algorithms. 31.
- [52] Indre Zliobaite. Measuring discrimination in algorithmic decision making.
- [53] Indre Zliobaite, Faisal Kamiran, and Toon Calders. Handling conditional discrimination. In *2011 IEEE 11th International Conference On Data Mining (ICDM)*. IEEE, 2011.