FAIREE: FAIR CLASSIFICATION WITH FINITE-SAMPLE AND DISTRIBUTION-FREE GUARANTEE

Anonymous authors
Paper under double-blind review

ABSTRACT

Algorithmic fairness plays an increasingly critical role in machine learning research. Several group fairness notions and algorithms have been proposed. However, the fairness guarantee of existing fair classification methods mainly depend on specific data distributional assumptions, often requiring large sample sizes, and fairness could be violated when there is a modest number of samples, which is often the case in practice. In this paper, we propose FaiREE, a fair classification algorithm which can satisfy group fairness constraints with finite-sample and distribution-free theoretical guarantees. FaiREE can be adapted to satisfying various group fairness notions (e.g., Equality of Opportunity, Equalized Odds, Demographic Parity, etc.) and achieve the optimal accuracy. These theoretical guarantees are further supported by experiments on both synthetic and real data. FaiREE is shown to have favorable performance over state-of-the-art algorithms.

1 INTRODUCTION

As machine learning algorithms have been increasingly used in consequential domains such as college admission Chouldechova & Roth (2018), loan application Ma et al. (2018), and disease diagnosis Fatima et al. (2017), there are emerging concerns about the algorithmic fairness in recent years. When standard machine learning algorithms are directly applied to the biased data provided by human, the output are sometimes found to be biased towards certain sensitive attribute that we want to protect (race, gender, etc). To address this fairness problem, many fairness notions have been proposed, including the individual fairness notion Biega et al. (2018), group fairness notions such as Demographic Parity, Equality of Opportunity, Predictive Parity, Equalized Odds (Dieterich et al., 2016; Hardt et al., 2016; Gajane & Pechenizkiy, 2017; Verma & Rubin, 2018), and multi-group fairness notions including multi-calibration Hébert-Johnson et al. (2018) and multi-accuracy Kim et al. (2019). Based on these fairness notions or constraints, corresponding algorithms were designed to help satisfy the fairness constraints.

Among these fairness algorithms, post-processing is a popular type of algorithm which modifies the output of the model to satisfy fairness constraints. However, recent post-processing algorithms are found to lack the ability to realize accuracy–fairness trade-off and perform poorly when the sample size is limited (Hardt et al., 2016; Pleiss et al., 2017). In addition, since most fairness constraints are non-convex, some papers propose convex relaxation-based methods Zafar et al. (2017b); Krishnaswamy et al. (2021). This type of algorithms generally do not have the theoretical guarantee how the output satisfies the exact original fairness constraint. Another line of research consider recalibrating the Bayes classifier by a group-dependent threshold (Valera et al., 2018; Chzhen et al., 2019; Zeng et al., 2022). However, their results require either some distributional assumptions or infinite sample size, which is hard to verify/satisfy in practice.

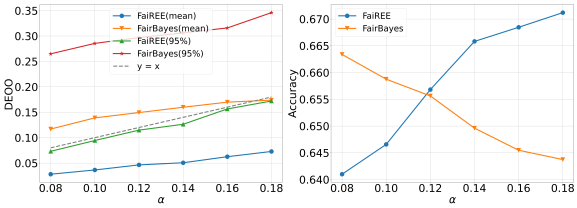

Figure 1: Comparison of FairBayes and FaiREE on the synthetic data with sample size = 1000. See Table 2 for detailed numerical results. Left: $DEOO$ v.s. α, Right: Test accuracy v.s. α. Here, $DEOO$ is the degree of violation to fairness constraint Equality of Opportunity and α is the pre-specified upper bound to control $DEOO$. See Eq. (1) in Section 2 for the detailed definition.

In this paper, we propose a post-processing algorithm that achieves group fairness notions with finite-sample and distribution free guarantee (FaiREE). To the best of our knowledge, this is the first algorithm in fairness classification with a finite-sample and distribution-free guarantee. A brief pipeline of FaiREE is to first score the dataset with the given classifier, and select candidate set based on these scores which can fit the fairness constraint with theoretical guarantee. As there are possibly multiple classifiers that can satisfy this constraint, we further develop a distribution-free estimate of the test mis-classification error, resulting in an algorithm that produces the optimal mis-classification error given the fairness constraints. As a motivating example, Figure 1 shows that applying state-of-the-art FairBayes method in Zeng et al. (2022) on a dataset with 1000 samples results in substantial fairness violation on the test data due to lack of fairness generalization. Our proposed FaiREE improved fairness generalization in these finite sample settings.

Additional Related Works. The fairness algorithms in the literature can be roughly categorized into three types: 1). Pre-processing algorithms that learn a fair representation to improve fairness (Zemel et al., 2013; Louizos et al., 2015; Lum & Johndrow, 2016; Adler et al., 2018; Calmon et al., 2017; Gordaliza et al., 2019) 2). The algorithms that optimize during training time (Calders et al., 2009; Woodworth et al., 2017; Zafar et al., 2017b;a; Agarwal et al., 2018; Russell et al., 2017) 3). The post-processing algorithms that try to modify the output of the original method to fit fairness constraints (Kamiran et al., 2012; Feldman, 2015; Hardt et al., 2016; Pleiss et al., 2017; Zeng et al., 2022; Chzhen et al., 2019; Fish et al., 2016; Valera et al., 2018; Corbett-Davies et al., 2017; Menon & Williamson, 2018; Hébert-Johnson et al., 2018; Kim et al., 2019).

The design of post-processing algorithms with distribution-free and finite-sample guarantees gains much attention recently due to its flexibility in practice, as it can be applied to any given algorithm (eg. a black-box neural network), and achieve desired theoretical guarantee with almost no assumption. One of the research area that satisfies this property is conformal prediction (Shafer & Vovk, 2008; Lei et al., 2018; Romano et al., 2019) whose aim is to construct prediction intervals that cover a future response with high probability. In this paper, we extend this line of research beyond prediction intervals, by designing classification algorithms that satisfy certain grup fairness with distribution-free and finite-sample guarantees.

Paper Organization. Section 2 provides the definitions and notations we use in the paper. Section 3 provides the general pipeline of FaiREE. In Section 4, we further extend the results to other fairness notions. Finally, Section 5 conducts experiments on both synthetic and real data and compare with several state-of-art algorithms to show that FaiREE has desirable performance.

2 PRELIMINARY

In this paper, we consider two types of features in a classification problem: the standard feature vector $X \in \mathcal{X}$, and the sensitive attribute, which we want the output to be fair on, is denoted as $A \in \mathcal{A} = \{0, 1\}$. For the simplicity of presentation, we consider the binary classification problem with labels in $\mathcal{Y} = \{0, 1\}$. We note that our analysis can be similarly extended to the multi-class and multi-attribute setting. Under the binary classification setting, we use the score-based classifier that outputs a prediction $\widehat{Y} = \widehat{Y}(X, A) \in \{0, 1\}$ based on a score function $f(X, A) \in [0, 1]$ that depends on X and A:

Definition 1. *(Score-based classifier) A score-based classifier is an indication function* $\hat{Y} = \phi(x, a) = \mathbb{1}\{f(x, a) > c\}$ *for a measurable score function* $f : \mathcal{X} \times \{0, 1\} \to [0, 1]$ *and some threshold* $c > 0$.

To address the algorithmic fairness problem, several group fairness notions have been developed in the literature. In the following, we introduce two of the popular notions, Equality of Opportunity and Equalized Odds. We will discuss other fairness notions in Section A.8 of the Appendix.

Equality of Opportunity asks that the true positive rates are comparable across different protected groups.

Definition 2. *(Equality of Opportunity (Hardt et al., 2016)) A classifier satisfies Equality of Opportunity if it satisfies the same true positive rate among protected groups:* $\mathbb{P}_{X|A=1,Y=1}(\widehat{Y} = 1) = \mathbb{P}_{X|A=0,Y=1}(\widehat{Y} = 1)$.

Equalized Odds is an extension of Equality of Opportunity, requiring both false positive rate and false negative rate are similar across different attributes.

Definition 3. *(Equalized Odds (Hardt et al., 2016)) A classifier satisfies Equalized Odds if it satisifes the following equality:* $\mathbb{P}_{X|A=1,Y=1}(\widehat{Y}=1) = \mathbb{P}_{X|A=0,Y=1}(\widehat{Y}=1)$ *and* $\mathbb{P}_{X|A=1,Y=0}(\widehat{Y}=0) = \mathbb{P}_{X|A=0,Y=0}(\widehat{Y}=0)$.

Sometimes it is too strict to require the classifier to satisfy Equality of Opportunity or Equalized Odds exactly, which may sacrifice a lot of accuracy (as a very simple example is $f(x,a) \equiv 1$). In practice, to strike a balance between fairness and accuracy, it makes sense to relax the equality above to an inequality with a small error bound. We use the difference with respect to Equality of Opportunity, denoted by $DEOO$, to measure the disparate impact:

$$DEOO = \mathbb{P}_{X|A=1,Y=1}(\widehat{Y}=1) - \mathbb{P}_{X|A=0,Y=1}(\widehat{Y}=1). \tag{1}$$

For a classifier ϕ, following (Zeng et al., 2022; Cho et al., 2020), we use $|DEOO(\phi)| \leq \alpha$ to denote a α-tolerance fairness constraint that controls the difference between the true positive rates below α.

Similarly, we define the following difference with for Equalized Odds. Since Equalized Odds, the difference is a two-dimensional vector:

$$DEO = (\mathbb{P}_{X|A=1,Y=1}(\widehat{Y}=1) - \mathbb{P}_{X|A=0,Y=1}(\widehat{Y}=1), \mathbb{P}_{X|A=1,Y=0}(\widehat{Y}=1) - \mathbb{P}_{X|A=0,Y=0}(\widehat{Y}=1)).$$

For notational simplicity, we use the notation $\preceq$ for the element-wise comparison between vectors, that is, $DEO \preceq (\alpha_1, \alpha_2)$ if and only if $\mathbb{P}_{X|A=1,Y=1}(\widehat{Y}=1) - \mathbb{P}_{X|A=0,Y=1}(\widehat{Y}=1) \leq \alpha_1$ and $\mathbb{P}_{X|A=1,Y=0}(\widehat{Y}=1) - \mathbb{P}_{X|A=0,Y=0}(\widehat{Y}=1) \leq \alpha_2$.

Additional Notation. We denote the proportion of group a by $p_a := \mathbb{P}(A=a)$ for $a \in \{0,1\}$; the proportion of group $Y=1$ conditioned on A for $p_{Y,a} := \mathbb{P}(Y=1 \mid A=a)$; the proportion of group $Y=1$ conditioned on A and X for $\eta_a(x) := \mathbb{P}(Y=1 \mid A=a, X=x)$. Also, we denote by $\mathbb{P}_X(x)$ and $\mathbb{P}_{X|A=a,Y=y}(x)$ respectively the distribution function of X and the distribution function of X conditioned on A and Y. The score function of standard Bayes-optimal classifier without fairness constraint is defined as $\phi^*(X,A) = \mathbb{1}\{f^*(X,A) > 1/2\}$, where $f^* \in \arg\min_f[\mathbb{P}(Y \neq \mathbb{1}\{f(X,A) > 1/2\})]$. We denote $v_{(k)}$ as the kth ordered value of sequence v in non-decreasing order. For a set T, we denote $sort(T)$ as a function that returns T in non-decreasing order. For a number $a \in \mathbb{R}$, we use $\lceil a \rceil$ to denote the ceiling function that maps a to the least integer greater than or equal to a. For a positive integer n, we use $[n]$ to denote the set $\{1, 2, ..., n\}$.

3 FAIREE: A FINITE SAMPLE BASED ALGORITHM

In this section, we propose FaiREE, a general post-processing algorithm that produces a Fair classifier in a finite-sample and distribution-fREE manner, and can be applied to a wide range of group fairness notions. We will illustrate its use in Equality of Opportunity as an example in this section, and discuss more application in later sections.

3.1 THE GENERAL PIPELINE OF FAIREE

Suppose we have dataset $S = S^{0,0} \cup S^{0,1} \cup S^{1,0} \cup S^{1,1}$, where $S^{y,a} = \{x_1^{y,a}, \ldots, x_{n^{y,a}}^{y,a}\}$ is the set of features associated with label $Y=y \in \{0,1\}$ and protected attribute $A=a \in \{0,1\}$. We denote the size of $S^{y,a}$ by $n^{y,a}$. Throughout the paper, we assume that $x_i^{y,a}, i \in \{1, \ldots, n^{y,a}\}$ are independently and identically distributed given $Y=y, A=a$. We define $n = n^{0,0} + n^{0,1} + n^{1,0} + n^{1,1}$ to be the total number of samples. Our goal is to post-process any given classifier to make it satisfy certain group fairness constraints.

FaiREE is a post-processing algorithm that can transform any pre-trained classification function score f in order to satisfy fairness constraints. In particular, FaiREE consists of three main steps, *scoring*, *candidate set construction* and *candidate selection*. See Figure 2 for an illustration.

Step 1: Scoring. FaiREE takes input as 1). a given fairness guarantee $\mathcal{G}$, such as Equality of Opportunity or Equalized Odds; 2). an error bound α, which controls the violation with respect to

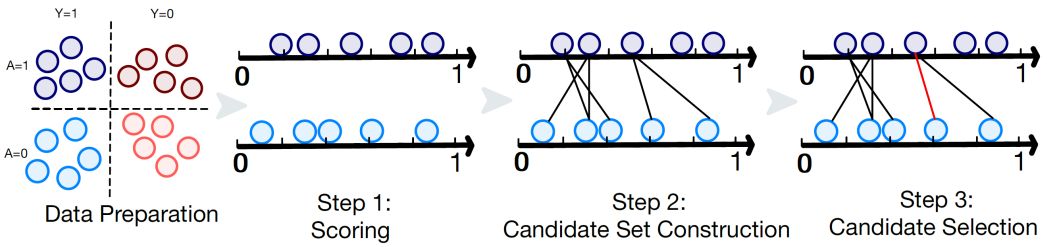

Figure 2: A concrete pipeline of FaiREE for Equality of Opportunity. Edges in Step 2 represent the selected candidate pair, and the red edge in Step 3 represents the final optimal candidate selected from all the edges. Each pair represents two different thresholds of a single classifier.

our given fairness notion; 3). a small tolerance level δ, which makes sure our final classifier satisfies our requirement with probability at least $1-\delta$; 4). a dataset S.

For *scoring*, we first apply the given classifier f to $S^{y,a}$ and denote the outcome $t_i^{y,a} := f(x_i^{y,a})$ as scores for each sample. These scores are then sorted within each subset in non-decreasing order respectively and obtain $T^{y,a} = \{t_{(1)}^{y,a}, \ldots, t_{(n^{y,a})}^{y,a}\}$.

Step 2: Candidate Set Construction. We first present a key observation for this step, which holds for many group fairness notions such as Equality of Opportunity, Equalized Odds (see details of more fairness notions in Section 3.2):

Any classifier can fit the fairness constraint with high probability by setting the decision threshold appropriately, regardless of the data distribution.

The insight of this observation comes from recent literature of post-processing algorithm and Neyman-Pearson classification algorithm (Tong et al., 2018; Chzhen et al., 2019; Fish et al., 2016; Valera et al., 2018; Corbett-Davies et al., 2017; Menon & Williamson, 2018). Under Equality of Opportunity, this observation is formalized in Proposition 1. We also establish similar results under other fairness notions beyond Equality of Opportunity in Section 3.2. From this observation we can build an algorithm to calculate the probability that a classifier f with certain threshold will satisfy the fairness constraint: $\text{Diff}_{\mathcal{G}}(f) \leq \alpha$, where $\text{Diff}_{\mathcal{G}}(f)$ is a generic notation to denote the violation rate of f under some fairness notion $\mathcal{G}$. Then we choose the classifiers with the probability $\mathbb{P}(\text{Diff}_{\mathcal{G}}(f) > \alpha) \leq \delta$ as our candidate set C. This candidate set consists of a set of threshold values, with potentially different thresholds for different subpopulation.

Step 3: Candidate Selection. Furthermore, as there might be multiple classifiers that satisfy the given fairness constraints, we aim to choose a classifier a small mis-classification error. To do this, FaiREE estimates the mis-classification error $err(f)$ of the classifier f, and choose the one with the smallest error among the candidate set constructed in the second step.

In the rest part of the section, as an example, we consider Equality of Opportunity as our target group fairness constraint and provide our algorithm in detail.

3.2 Application to Equality of Opportunity

In this section, we apply FaiREE to the fairness notion Equality of Opportunity. The following two subsections explain the steps **Candidate Set Construction** and **Candidate Selection** in detail.

3.2.1 Candidate Set Construction

We first formalize our observation in the following proposition. Using property of order statistics, the following proposition states that it is sufficient to choose the threshold of the score-based classifier from the sorted scores to control the fairness violation in a distribution-free and finite-sample manner. Here, $k^{1,a}$ is the index from which we select the threshold in $T^{1,a}$.

Proposition 1. *Consider* $k^{1,a} \in \{1, \ldots, n^{1,a}\}$ *for* $a \in \{0,1\}$*, and the score-based classifier* $\phi(x,a) = \mathbb{1}\{f(x,a)) > t_{(k^{1,a})}^{1,a}\}$*. Let* $g_1(k,a) = \mathbb{E}[\sum_{j=k}^{n^{1,a}} \binom{n^{1,a}}{j} (Q^{1,1-a} - \alpha)^j (1 - (Q^{1,1-a} -$

$\alpha))^{n^{1,a}-j}]$ *with* $Q^{1,a} \sim Beta(k, n^{1,a}-k+1)$*, then we have:*

$$\mathbb{P}(|DEOO(\phi)| > \alpha) \leq g_1(k^{1,1}, 1) + g_1(k^{1,0}, 0).$$

Additionally, if $t^{1,a}_{(k^{1,a})}$ *is a continuous random variable, the inequality above becomes tight equality.*

Here, g_1 is a function constructed using the property of order statistics so that $g_1(k^{1,1}, 1)$ and $g_1(k^{1,0}, 0)$ upper bound $\mathbb{P}(DEOO(\phi) > \alpha)$ and $\mathbb{P}(DEOO(\phi) < -\alpha)$ respectively. This proposition ensures that the $DEOO$ of a given classifier can be controlled with high probability if we choose an appropriate threshold value when post-processing.

Based on the above proposition, we then build our classifiers for an arbitrarily given score function f as below. We define $L(k^{1,0}, k^{1,1}) = g_1(k^{1,1}, 1) + g_1(k^{1,0}, 0)$. Recall that the error tolerance is α, and δ is the tolerance level. Our candidate set is then constructed as $K = \{(k^{1,0}, k^{1,1}) \mid L(k^{1,0}, k^{1,1}) \leq \delta\}$.

Before we proceed to the theoretical guarantee for this candidate set, we introduce a bit more notation. Let us denote the size of the candidate set K by M, and the elements in the set K by $(k^{1,0}_1, k^{1,1}_1), \ldots, (k^{1,0}_M, k^{1,1}_M)$. Additionally, we let $\hat{\phi}_i(x, a) = \mathbb{1}\{f(x, a) > t^{1,a}_{(k^{1,a}_i)}\}$, for $i = 1, \ldots, M$.

To ensure that there exists at least one valid classifier (i.e. $M \geq 1$), we should have $\mathbb{E}[(Q^{1,0} - \alpha)^{n^{1,0}}] + \mathbb{E}[(Q^{1,1} - \alpha)^{n^{1,1}}] \leq \delta$, which requires a necessary and sufficient lower bound requirement on the sample size, as formulated in the following proposition:

Theorem 1. *If* $\min\{n^{1,0}, n^{1,1}\} \geq \lceil \frac{\log \frac{\delta}{2}}{\log(1-\alpha)} \rceil$*, for each* $i \in \{1, \ldots, M\}$ *in the candidate set, we have* $|DEOO(\hat{\phi}_i)| < \alpha$ *with probability* $1 - \delta$*.*

As there are at most $n^{1,0}n^{1,1}$ elements in the candidate set, the size of K, M, can be as large as $O(n^2)$. To further reduce the computational complexity, in the following part, we provide a method to shrink the candidate set.

Our construction is inspired by the following lemma, which gives the analytical form of the fair Bayes-optimal classifier under the Equality of Opportunity constraint. This Bayes-optimal classifier is defined as $\phi^*_\alpha = \arg\min_{|DEOO(\phi)| \leq \alpha} \mathbb{P}(\phi(X, A) \neq Y)$.

Lemma 1 (Adapted from Theorem E.4 in Zeng et al. (2022))**.** *The fair Bayes-optimal classifier under Equality of Opportunity can be explicitly written as* $\phi^*_\alpha(x, a) = \mathbb{1}\{f^*(x, a) > t^*_a\}$*, then* $t^*_1 = \frac{p_1 p_{Y,1}}{2p_1 p_{Y,1} - (1/t^{1,0}_{(k)} - 2) \cdot p_0 p_{Y,0}}$.

Note that in practice, the input classifier f can be the classifier trained by a classification algorithm on the training set, which means it is close to f^*. Thus from this observation, we can adopt a new way of building a much smaller candidate set. Note that our original candidate set is defined as : $K = \{(k^{1,0}, k^{1,1}) \mid L(k^{1,0}, k^{1,1}) \leq \delta\} = \{(k^{1,0}_1, k^{1,1}_1), \ldots, (k^{1,0}_M, k^{1,1}_M)\}$. Now, for every $1 \leq k \leq n^{1,0}$, from Lemma 1 we denote $u_1(k) = \arg\min_u |t^{1,1}_{(u)} - \frac{\hat{p}_1 \hat{p}_{Y,1}}{2\hat{p}_1 \hat{p}_{Y,1} - (1/t^{1,0}_{(k)} - 2) \cdot \hat{p}_0 \hat{p}_{Y,0}}|$, where $\hat{p}_a = \frac{n^{1,a} + n^{0,a}}{n^{0,0} + n^{0,1} + n^{1,0} + n^{1,1}}$ and $\hat{p}_{y,a} = \frac{n^{1,a}}{n^{0,a} + n^{1,a}}$. We then build our candidate set as below:

$$K' = \{(k^{1,0}, u_1(k^{1,0})) \mid L(k^{1,0}, u_1(k^{1,0})) \leq \delta\}. \tag{2}$$

This candidate set K' has cardinality at most n. Since our next step, Candidate Selection, has computational complexity that is linear in the size of candidate set, using the new set K' would help us reduce the computational complexity from $O(n^2)$ to $O(n)$.

3.2.2 Candidate Selection

In this subsection, we explain in detail how we choose the classifier with the smallest misclassification error from the candidate set constructed in the last step. For a given pair $(k^{1,0}_i, k^{1,1}_i)$ in the candidate set of index ($i \in [M]$), we need to know the rank of $t^{1,0}_{(k^{1,0}_i)}$ and $t^{1,1}_{(k^{1,1}_i)}$ in the sorted set

$T^{0,0}$ and $T^{0,1}$ respectively in order to compute the test error where we need to consider both $y = 0$ and 1. Specifically, we find the $k_i^{0,a}$ such that $t^{0,a}_{(k_i^{0,a})} \leq t^{1,a}_{(k_i^{1,a})} < t^{0,a}_{(k_i^{0,a}+1)}$ for $a \in \{0, 1\}$.

To estimate the test mis-classification error of $\hat{\phi}_i = \mathbb{1}\{f(x, a) > t^{1,a}_{(k_i^{1,a})}\}$, we divide the error into four terms by different values of y and a. We then estimate each part using the property of order statistics respectively, and obtain the following proposition:

Proposition 2. *Let* $\hat{e}_i = \frac{k_i^{1,0}}{n^{1,0}+1}\frac{n^{1,0}}{n} + \frac{k_i^{1,1}}{n^{1,1}+1}\frac{n^{1,1}}{n} + \frac{n^{0,0}+\frac{1}{2}-k_i^{0,0}}{n^{0,0}+1}\frac{n^{0,0}}{n} + \frac{n^{0,1}+\frac{1}{2}-k_i^{0,1}}{n^{0,1}+1}\frac{n^{0,1}}{n}$*, for* $i = 1, 2, ..., M$*. Then, we have* $\mid \mathbb{P}(\hat{\phi}_i \neq Y) - \hat{e}_i \mid = O(1/\min(n^{0,0}, n^{0,1}))$ *with probability* $1 - O(\exp(-\min(n^{0,0}, n^{0,1})))$.

The above proposition enables us to efficiently estimate the test error of the M classifiers $\hat{\phi}_i$'s defined above, from which we can choose a classifier with the lowest test error $\hat{\phi}$. The algorithm is summarized in Algorithm 1.

Algorithm 1: FaiREE for Equality of Opportunity

Input: Training data $S = S^{0,0} \cup S^{0,1} \cup S^{1,0} \cup S^{1,1}$; the error bound α; the tolerance level δ; a given pre-trained classifier f

$T^{y,a} = \{f(x_1^{y,a}), \ldots, f(x_{n_{y,a}}^{y,a})\}$

$\{t^{y,a}_{(1)}, \ldots, t^{y,a}_{(n_{y,a})}\}$ =sort($T^{y,a}$)

Define $g_1(k, a)$ as in Proposition 1, and let $L(k^{1,0}, k^{1,1}) = g_1(k^{1,1}, 1) + g_1(k^{1,0}, 0)$

Build candidate set K' as in Eq. 2, and write $K' = \{(k_1^{1,0}, k_1^{1,1}), \ldots, (k_{M'}^{1,0}, k_{M'}^{1,1})\}$

Find $k_i^{0,0}, k_i^{0,1}$: $t^{0,0}_{(k_i^{0,0})} \leq t^{1,0}_{(k_i^{1,0})} < t^{0,0}_{(k_i^{0,0}+1)}$, $t^{0,1}_{(k_i^{0,1})} \leq t^{1,1}_{(k_i^{1,1})} < t^{0,1}_{(k_i^{0,1}+1)}$

$i_* \leftarrow \arg\min_{i \in [M']}\{\hat{e}_i\}$ ($\hat{e}_i$ is defined in Proposition 2)

Output: $\hat{\phi}(x, a) = \mathbb{1}\{f(x, a) > t^{1,a}_{(k_{i_*}^{1,a})}\}$

In the following, we provide the theory showing that the output of Algorithm 1 is approaching the optimal mis-classification error under Equality of Opportunity. The following theorem states that the final output of FaiREE has both controlled $DEOO$, and achieved almost minimum mis-classification error when the input classifier is properly chosen.

Theorem 2. *Given any* $\alpha' < \alpha$*. If* $\min\{n^{1,0}, n^{1,1}\} \geq \lceil \frac{\log \frac{\delta}{2}}{\log(1-\alpha)} \rceil$*. Set* $\delta = c_0/M$ *for some small* $c_0 > 0$*, where* M *is the size of the candidate. Suppose* $\hat{\phi}$ *is the final output of FaiREE, we have:*

(1). $|DEOO(\hat{\phi})| < \alpha$ *with probability* $1 - c_0$.

(2). Suppose the density functions of f^* *under* $A = a, Y = 1$ *are continuous. For any* $\delta' > 0$*, there exists* $\epsilon > 0$ *and* $0 < c < 1$ *such that when the input classifier* f *is* f^* *and the constructed candidate set is* K'*, we have* $\mid \mathbb{P}(\hat{\phi} \neq Y) - \mathbb{P}(\phi^*_{\alpha'} \neq Y) \mid \leq \delta'$ *with probability* $1 - O(c^{\min\{n^{1,0}, n^{0,0}, n^{0,1}\}})$.

Theorem 2 ensures that our classifier will be close to fair Bayes-optimal classifier if the input classifier is f^*. Here, $\alpha' < \alpha$ is any positive constant.

Remark: We remark that FaiREE requires no assumption on data distribution except for a minimum sample size. It has advantage over existing literature which generally imposes different assumptions to data distribution. For example, Chzhen et al. (2019) assumes $\eta(x, a)$ must surpass the level $\frac{1}{2}$ on a set of non-zero measure. Valera et al. (2018) assumes that the shifting threshold of the classifier follows beta distribution. Also, Zeng et al. (2022)'s result only holds for population-level, and the finite-sample version is not studied.

4 APPLICATION TO MORE FAIRNESS NOTIONS

4.1 EQUALIZED ODDS

In this section, we apply our algorithm to the fairness notion Equalized Odds, who has two fairness constraints simultaneously. To ensure the two constraints, the algorithm of equalized odds should be different from Algorithm 1. We should consider all $S^{y,a}$ instead of just $S^{1,a}$ when estimating the violation to fairness constraint in the step **Candidate Set Construction**. Thus we add a function g_0 that deals with data with protected attribute $A = 1$, to perfect our algorithm together with g_1 defined in the last section.

Similar to Proposition 1, we have the following proposition, which promises that DEO of a given classifier can be controlled with high probability if we choose the proper threshold when post-processing.

Proposition 3. *Given $k^{1,0}, k^{1,1}$ satisfying $k^{1,a} \in \{1, \ldots, n^{1,a}\}$ $(a = 0, 1)$. Define $\phi(x, a) = \mathbb{1}\{f(x, a)) > t^{1,a}_{(k^{1,a})}\}$, $g_y(k, a) = \mathbb{E}[\sum_{j=k}^{n^{y,a}} \binom{n^{y,a}}{j} (Q^{y,1-a} - \alpha)^j (1 - (Q^{y,1-a} - \alpha))^{n^{y,a}-j}]$ with $Q^{y,a} \sim Beta(k + 1 - y, n^{y,a} - k + y)$, then we have:*

$$\mathbb{P}(|DEO(\phi)| \preceq (\alpha, \alpha)) \geq 1 - g_1(k^{1,1}, 1) - g_1(k^{1,0}, 0) - g_0(k^{0,1}, 1) - g_0(k^{0,0}, 0).$$

Similar to Proposition 1, g_0 and g_1 together upper bound the probability that ϕ violates the DEO constraint.

Algorithm 2: FaiREE for Equalized Odds

Input: Training data $S = S^{0,0} \cup S^{0,1} \cup S^{1,0} \cup S^{1,1}$; the error bound α; the tolerance level δ; a given pre-trained classifier f

1 $T^{y,a} = \{f(x_1^{y,a}), \ldots, f(x_{n_{y,a}}^{y,a})\}$

2 $\{t^{y,a}_{(1)}, \ldots, t^{y,a}_{(n_{y,a})}\}$ =sort($T^{y,a}$)

3 Define $g_0(k, a)$ and $g_1(k, a)$ as in Proposition 3, $L_1(k^{1,0}, k^{1,1}) = g_1(k^{1,1}, 1) + g_1(k^{1,0}, 0)$, and $L_0(k^{0,0}, k^{0,1}) = g_0(k^{0,1}, 1) + g_0(k^{0,0}, 0)$. Define

4 For every $k^{1,0}, k^{1,1}$, there exists $k^{0,0}, k^{0,1}$ such that $t^{0,0}_{(k^{0,0})} \leq t^{1,0}_{(k^{1,0})} < t^{0,0}_{(k^{0,0}+1)}$, $t^{0,1}_{(k^{0,1})} \leq t^{1,1}_{(k^{1,1})} < t^{0,1}_{(k^{0,1}+1)}$.

5 Build the candidate set as $K = \{(k^{1,0}, k^{1,1}) \mid L_1(k^{1,0}, k^{1,1}) + L_0(k^{0,0}, k^{0,1}) \leq \delta\} = \{(k_1^{1,0}, k_1^{1,1}), \ldots, (k_M^{1,0}, k_M^{1,1})\}$.

6 Compute $\hat{e_i}$ as in Proposition 2), and let $i_* = \arg\min_{i \in [M]} \{\hat{e_i}\}$.

Output: $\hat{\phi}(x, a) = \mathbb{1}\{f(x, a) > t^{1,a}_{(k^{1,a}_{i_*})}\}$

From Proposition 3, we directly get the following proposition of the DEO of classifiers in the candidate set.

Theorem 3. *If $\min\{n^{0,0}, n^{0,1}, n^{1,0}, n^{1,1}\} \geq \lceil \frac{\log \frac{\delta}{4}}{\log(1-\alpha)} \rceil$, then for each $i \in \{1, \ldots, M\}$, we have $|DEO(\hat{\phi}_i)| \preceq (\alpha, \alpha)$ with probability $1 - \delta$.*

The part for the theoretical analysis of test error is similar to the algorithm for Equality of Opportunity.

Theorem 4. *Given $\alpha' < \alpha$. If $\min\{n^{0,0}, n^{0,1}, n^{1,0}, n^{1,1}\} \geq \lceil \frac{\log \frac{\delta}{4}}{\log(1-\alpha)} \rceil$. Set $\delta = c_0/M$ for some small $c_0 > 0$, where M is the size of the candidate. Suppose $\hat{\phi}$ is the final output of FaiREE, we have:*

(1). $|DEO(\hat{\phi})| \preceq (\alpha, \alpha)$ with probability $1 - c_0$.

(2). Suppose the density functions of f^* under $A = a, Y = 1$ are continuous. $\phi^*_{\alpha'} = \arg\min_{|DEO(\phi)| \preceq (\alpha', \alpha')} \mathbb{P}(\phi(X, A) \neq Y)$. For any $\delta' > 0$, there exist $\epsilon > 0$ and $0 < c < 1$

such that when the input classifier f is f^*, we have $\mid \mathbb{P}(\hat{\phi} \neq Y) - \mathbb{P}(\phi^*_{\alpha',\alpha'} \neq Y) \mid \leq \delta'$ *with probability* $1 - O(c^{\min\{n^{1,0},n^{1,1},n^{0,0},n^{0,1}\}})$.

4.2 On comparing different fairness constraints

In this subsection, we further extend our algorithms to more fairness notions. We defer the detailed technical results and derivations in Section A.8 of the appendix, and compare the sample size requirement for different fairness notions. We note that our algorithm is almost assumption-free, except for the $i.i.d.$ assumption and a necessary and sufficient condition of the sample size. Therefore, we make a chart below to recommend different fairness notions used in practice when the sample sizes are limited. Specifically, we have the following table:

Table 1: Sample complexity requirements for FaiREE of different fairness notions. We consider the following fairness notions: DP (Demographic Parity), EOO (Equality of Opportunity), EO (Equalized Odds), PE (Predictive Equality), EA (Equalized Accuracy), and $n^a = n^{0,a} + n^{1,a}$.

DP	EOO	PE	EO	EA
$n^a \geq \lceil \frac{\log \frac{\delta}{2}}{\log(1-\alpha)} \rceil$	$n^{1,a} \geq \lceil \frac{\log \frac{\delta}{2}}{\log(1-\alpha)} \rceil$	$n^{0,a} \geq \lceil \frac{\log \frac{\delta}{2}}{\log(1-\alpha)} \rceil$	$n^{y,a} \geq \lceil \frac{\log \frac{\delta}{4}}{\log(1-\alpha)} \rceil$	$n^{y,a} \geq \lceil \frac{\log \frac{\delta}{4}}{\log(\frac{1-y+(2y-1)p_{Y,\mid y-a\mid}-\alpha}{y(2p_{Y,a}-1)+1-p_{Y,a})}} \rceil$

From this table, we find that Demographic Parity requires the least sample size, Equality of Opportunity and Predictive Equality need a lightly larger sample size, and Equalized Odds is the notion that requires the largest sample size among the first four fairness notions. The sample size requirement for Equalized Accuracy is similar to that of Equalized Odds, but does not have a strict dominance.

5 Experiments

In this section, we conduct experiments to test and understand the effectiveness of FaiREE. For both synthetic data and real data analysis, we compare FaiREE with the following representative methods for fair classification: Reject-Option-Classification (ROC) method in Kamiran et al. (2012), Eqodds-Postprocessing (Eq) method in Hardt et al. (2016), Calibrated Eqodds-Postprocessing (C-Eq) method in Pleiss et al. (2017) and FairBayes method in Zeng et al. (2022).The first three baselines are designed to cope with Equalized Odds and the last one is for Equality of Opportunity.

5.1 Synthetic data

To show the distribution-free and finite-sample guarantee of FaiREE, we generate the synthetic data from mixed distributions. Real world data are generally heavy-tailed (Resnick (1997)). Thus, we consider the following models with various heavy-tailed distributions for generating synthetic data:

Model 1. We generate the protected attribute A and label Y with probability $p_1 = \mathbb{P}(A = 1) = 0.7, p_0 = \mathbb{P}(A = 0) = 0.3, p_{y,1} = \mathbb{P}(Y = y \mid A = 1) = 0.7$ and $p_{y,0} = \mathbb{P}(Y = y \mid A = 0) = 0.4$ for $y \in \{0, 1\}$. The dimension of features is set to 60, and we generate features with $x^{0,0}_{i,j} \overset{i.i.d.}{\sim} t(3)$, where $t(k)$ denotes the t-distribution with degree of freedom k, $x^{0,1}_{i,j} \overset{i.i.d.}{\sim} \chi^2_1$, $x^{1,0}_{i,j} \overset{i.i.d.}{\sim} \chi^2_3$ and $x^{1,1}_{i,j} \overset{i.i.d.}{\sim} N(\mu, 1)$, where $\mu \sim U(0, 1)$ and the scale parameter is fixed to be 1, for $j = 1, 2, ..., 60$.

Model 2. We generate the protected attribute A and label Y with the probability, location parameter and scale parameter the same as Model 1. The dimension of features is set to 80, and we generate features with $x^{0,0}_{i,j} \overset{i.i.d.}{\sim} t(4)$, $x^{0,1}_{i,j} \overset{i.i.d.}{\sim} \chi^2_2$, $x^{1,0}_{i,j} \overset{i.i.d.}{\sim} \chi^2_4$ and $x^{1,1}_{i,j} \overset{i.i.d.}{\sim} Laplace(\mu, 1)$, for $j = 1, 2, ..., 80$.

For each model, we generate 1000 $i.i.d.$ samples, and the experimental results are summarized in Tables 2 and 3.

From these two tables, we find that our proposed FaiREE, when applied to different fairness notions Equality of Opportunity and Equality of Opportunity, is able to control the required fairness violation respectively with high probability, while all the other methods cannot. In addition, although

Table 2: Experimental studies under Model 1. Here $\overline{|DEOO|}$ denotes the sample average of the absolute value of $DEOO$ defined in Eq. (1), and $|DEOO|_{95}$ denotes the sample upper 95% quantile. $\overline{|DPE|}$ and $|DPE|_{95}$ are defined similarly for DPE defined in Eq. (13). $\overline{ACC}$ is the sample average of accuracy. We use "/" in the DPE line because FairBayes and FaiREE-EOO are not designed to control DPE.

	Eq	C-Eq	ROC	FairBayes			FaiREE-EOO			FaiREE-EO		
α	/	/	/	0.08	0.12	0.16	0.08	0.12	0.16	0.08	0.12	0.16
$\overline{\lvert DEOO\rvert}$	0.061	0.132	0.255	0.117	0.149	0.170	0.028	0.046	0.063	0.025	0.031	0.042
$\lvert DEOO\rvert_{95}$	0.146	0.307	0.500	0.265	0.297	0.316	0.073	0.115	0.157	0.079	0.108	0.133
$\overline{\lvert DPE\rvert}$	0.051	0.029	0.511	/	/	/	/	/	/	0.039	0.042	0.045
$\lvert DPE\rvert_{95}$	0.110	0.091	0.850	/	/	/	/	/	/	0.075	0.084	0.106
$\overline{ACC}$	0.472	0.606	0.637	0.663	0.656	0.646	0.621	0.657	0.669	0.552	0.562	0.615

Table 3: Experimental studies under Model 2, with the same notation as Table 2.

	Eq	C-Eq	ROC	FairBayes			FaiREE-EOO			FaiREE-EO		
α	/	/	/	0.08	0.12	0.16	0.08	0.12	0.16	0.08	0.12	0.16
$\overline{\lvert DEOO\rvert}$	0.063	0.105	0.255	0.251	0.320	0.324	0.027	0.047	0.073	0.028	0.035	0.047
$\lvert DEOO\rvert_{95}$	0.080	0.137	0.500	0.676	0.742	0.765	0.075	0.112	0.153	0.077	0.114	0.143
$\overline{\lvert DPE\rvert}$	0.043	0.107	0.511	/	/	/	/	/	/	0.041	0.044	0.056
$\lvert DPE\rvert_{95}$	0.066	0.144	0.850	/	/	/	/	/	/	0.071	0.090	0.127
$\overline{ACC}$	0.380	0.600	0.637	0.606	0.598	0.589	0.595	0.627	0.639	0.575	0.589	0.606

satisfying stronger constraints, the mis-classification error FaiREE is comparable to, and sometimes better than the state-of-the-art methods.

5.2 Real Data analysis

In this section, we apply FaiREE to a real data set, Adult Census dataset (Dua et al., 2017), whose task is to predict whether a person's income is greater than $50,000. The protected attribute is gender, and the sample size is 45,222, including 32561 training samples and 12661 test samples. To facilitate the numerical study, we randomly split data into training set, calibration set and test set at each repetition and repeat for 500 times. FaiREE is compared the existing methods described in the last subsection. Again, as shown in Table 4, the proposed FaiREE method controls the fairness constraints at the desired level, and achieve small mis-classification error. More implementation details and experiments on other benchmark datasets are presented in A.9.

Table 4: Result of different methods on Adult Census dataset

	Eq	C-Eq	ROC	FairBayes			FaiREE-EOO			FaiREE-EO		
α	/	/	/	0.07	0.1	0.14	0.07	0.1	0.14	0.07	0.1	0.14
$\overline{\lvert DEOO\rvert}$	0.043	0.087	0.031	0.107	0.101	0.104	0.034	0.039	0.066	0.002	0.039	0.067
$\lvert DEOO\rvert_{95}$	0.112	0.154	0.097	0.140	0.153	0.153	0.065	0.090	0.124	0.008	0.094	0.125
$\overline{\lvert DPE\rvert}$	0.027	0.048	0.044	/	/	/	/	/	/	0.030	0.066	0.074
$\lvert DPE\rvert_{95}$	0.058	0.105	0.117	/	/	/	/	/	/	0.056	0.078	0.086
$\overline{ACC}$	0.815	0.823	0.691	0.847	0.847	0.847	0.845	0.846	0.846	0.512	0.846	0.846

6 Conclusion and discussion

In this paper, we propose FaiREE, a post-processing algorithm for fair classification with theoretical guarantees in a finite-sample and distribution-free manner. FaiREE can be applied to a wide range of group fairness notions, and is shown to achieve small mis-classification error while the fairness constraints. Numerical studies on both synthetic and real data shows the practical value of FaiREE in achieving a superior fairness-accuracy trade-off than the state-of-the-art methods. One interesting direction of future work is to extend the FaiREE techniques to multi-group fairness notions such as multi-calibration (Hébert-Johnson et al., 2018) and multi-accuracy Kim et al. (2019).

REFERENCES

Philip Adler, Casey Falk, Sorelle A Friedler, Tionney Nix, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. Auditing black-box models for indirect influence. *Knowledge and Information Systems*, 54(1):95–122, 2018.

Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *International Conference on Machine Learning*, pp. 60–69. PMLR, 2018.

Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: There's software used across the country to predict future criminals. *And it's biased against blacks. ProPublica*, 23: 77–91, 2016.

Asia J Biega, Krishna P Gummadi, and Gerhard Weikum. Equity of attention: Amortizing individual fairness in rankings. In *The 41st international acm sigir conference on research & development in information retrieval*, pp. 405–414, 2018.

Toon Calders, Faisal Kamiran, and Mykola Pechenizkiy. Building classifiers with independency constraints. In *2009 IEEE International Conference on Data Mining Workshops*, pp. 13–18. IEEE, 2009.

Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. *Advances in neural information processing systems*, 30, 2017.

Jaewoong Cho, Gyeongjo Hwang, and Changho Suh. A fair classifier using kernel density estimation. *Advances in Neural Information Processing Systems*, 33:15088–15099, 2020.

Alexandra Chouldechova and Aaron Roth. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810*, 2018.

Evgenii Chzhen, Christophe Denis, Mohamed Hebiri, Luca Oneto, and Massimiliano Pontil. Leveraging labeled and unlabeled data for consistent fair binary classification. *Advances in Neural Information Processing Systems*, 32, 2019.

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*, pp. 797–806, 2017.

William Dieterich, Christina Mendoza, and Tim Brennan. Compas risk scales: Demonstrating accuracy equity and predictive parity. *Northpointe Inc*, 7(4), 2016.

Dheeru Dua, Casey Graff, et al. Uci machine learning repository. 2017.

Meherwar Fatima, Maruf Pasha, et al. Survey of machine learning algorithms for disease diagnostic. *Journal of Intelligent Learning Systems and Applications*, 9(01):1, 2017.

Michael Feldman. *Computational fairness: Preventing machine-learned discrimination*. PhD thesis, 2015.

Benjamin Fish, Jeremy Kun, and Ádám D Lelkes. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM international conference on data mining*, pp. 144–152. SIAM, 2016.

Pratik Gajane and Mykola Pechenizkiy. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184*, 2017.

Paula Gordaliza, Eustasio Del Barrio, Gamboa Fabrice, and Jean-Michel Loubes. Obtaining fairness using optimal transport theory. In *International Conference on Machine Learning*, pp. 2357–2365. PMLR, 2019.

Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

Ursula Hébert-Johnson, Michael Kim, Omer Reingold, and Guy Rothblum. Multicalibration: Calibration for the (computationally-identifiable) masses. In *International Conference on Machine Learning*, pp. 1939–1948. PMLR, 2018.

Faisal Kamiran and Toon Calders. Classifying without discriminating. In *2009 2nd international conference on computer, control and communication*, pp. 1–6. IEEE, 2009.

Faisal Kamiran, Asim Karim, and Xiangliang Zhang. Decision theory for discrimination-aware classification. In *2012 IEEE 12th International Conference on Data Mining*, pp. 924–929. IEEE, 2012.

Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 247–254, 2019.

Anilesh Krishnaswamy, Zhihao Jiang, Kangning Wang, Yu Cheng, and Kamesh Munagala. Fair for all: Best-effort fairness guarantees for classification. In *International Conference on Artificial Intelligence and Statistics*, pp. 3259–3267. PMLR, 2021.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113 (523):1094–1111, 2018.

Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. The variational fair autoencoder. *arXiv preprint arXiv:1511.00830*, 2015.

Kristian Lum and James Johndrow. A statistical framework for fair predictive algorithms. *arXiv preprint arXiv:1610.08077*, 2016.

Xiaojun Ma, Jinglan Sha, Dehua Wang, Yuanbo Yu, Qian Yang, and Xueqi Niu. Study on a prediction of p2p network loan default based on the machine learning lightgbm and xgboost algorithms according to different high dimensional data cleaning. *Electronic Commerce Research and Applications*, 31:24–39, 2018.

Aditya Krishna Menon and Robert C Williamson. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency*, pp. 107–118. PMLR, 2018.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. *Advances in neural information processing systems*, 30, 2017.

Sidney I Resnick. Heavy tail modeling and teletraffic data: special invited paper. *The Annals of Statistics*, 25(5):1805–1869, 1997.

Yaniv Romano, Evan Patterson, and Emmanuel Candes. Conformalized quantile regression. *Advances in neural information processing systems*, 32, 2019.

Chris Russell, Matt J Kusner, Joshua Loftus, and Ricardo Silva. When worlds collide: integrating different counterfactual assumptions in fairness. *Advances in neural information processing systems*, 30, 2017.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Xin Tong, Yang Feng, and Jingyi Jessica Li. Neyman-pearson classification algorithms and np receiver operating characteristics. *Science advances*, 4(2):eaao1659, 2018.

Isabel Valera, Adish Singla, and Manuel Gomez Rodriguez. Enhancing the accuracy and fairness of human decision making. *Advances in Neural Information Processing Systems*, 31, 2018.

Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pp. 1–7. IEEE, 2018.

Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *Conference on Learning Theory*, pp. 1920–1953. PMLR, 2017.

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*, pp. 1171–1180, 2017a.

Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Artificial intelligence and statistics*, pp. 962–970. PMLR, 2017b.

Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *International conference on machine learning*, pp. 325–333. PMLR, 2013.

Xianli Zeng, Edgar Dobriban, and Guang Cheng. Bayes-optimal classifiers under group fairness. *arXiv preprint arXiv:2202.09724*, 2022.

A Appendix

A.1 Proof of Proposition 1

To prove Proposition 1, we first introduce the following lemma.

Lemma 2. $F^{1,0}(t^{1,0}_{(k^{1,0})}) \sim Beta(k^{1,0}, n^{1,0}-k^{1,0}+1)$, $F^{1,1}(t^{1,1}_{(k^{1,1})}) \sim Beta(k^{1,1}, n^{1,1}-k^{1,1}+1)$.

Proof. Since $F^{1,0}, F^{1,1}$ are the cumulative distribution functions of the $t^{1,0}$'s and $t^{1,1}$'s, we have $F^{1,0}(t^{1,0}), F^{1,1}(t^{1,1}) \sim U(0,1)$, thus $F^{1,0}(t^{1,0}_{(k^{1,0})})$ is the $k^{1,0}{}^{th}$ order statistic of $n^{1,0}$ i.i.d samples from $U(0,1)$ and $F^{1,1}(t^{1,1}_{(k^{1,1})})$ is the $k^{1,1}{}^{th}$ order statistic of $n^{1,1}$ i.i.d samples from $U(0,1)$.

Thus, from the well known fact of the ordered statistics, we have $F^{1,0}(t^{1,0}_{(k^{1,0})}) \sim Beta(k^{1,0}, n^{1,0} - k^{1,0}+1)$ and $F^{1,1}(t^{1,1}_{(k^{1,1})}) \sim Beta(k^{1,1}, n^{1,1} - k^{1,1}+1)$.

□

Now we come back to the proof of the proposition.

Proof. The classifier is

$$\phi = \begin{cases} \mathbb{1}\{f(x,0) > t^{1,0}_{(k^{1,0})}\}, a = 0 \\ \mathbb{1}\{f(x,1) > t^{1,1}_{(k^{1,1})}\}, a = 1 \end{cases}$$

we have:

$$\begin{aligned} |DEOO(\phi)| &= |\mathbb{P}(\hat{Y}=1 \mid A=0, Y=1) - \mathbb{P}(\hat{Y}=1 \mid A=1, Y=1)| \\ &= |\mathbb{P}(f(x,0) > t^{1,0}_{(k^{1,0})} \mid A=0, Y=1) - \mathbb{P}(f(x,1) > t^{1,1}_{(k^{1,1})} \mid A=1, Y=1)| \\ &= |1 - F^{1,0}(t^{1,0}_{(k^{1,0})}) - [1 - F^{1,1}(t^{1,1}_{(k^{1,1})})]| \\ &= |F^{1,1}(t^{1,1}_{(k^{1,1})}) - F^{1,0}(t^{1,0}_{(k^{1,0})})| \end{aligned}$$

Hence,

$$\begin{aligned} \mathbb{P}(|DEOO(\phi)| > \alpha) &= \mathbb{P}(|F^{1,1}(t^{1,1}_{(k^{1,1})}) - F^{1,0}(t^{1,0}_{(k^{1,0})})| > \alpha) \\ &= \mathbb{P}(F^{1,1}(t^{1,1}_{(k^{1,1})}) - F^{1,0}(t^{1,0}_{(k^{1,0})}) > \alpha) + \mathbb{P}(F^{1,1}(t^{1,1}_{(k^{1,1})}) - F^{1,0}(t^{1,0}_{(k^{1,0})}) < -\alpha) \\ &\triangleq A + B. \end{aligned}$$

We then have

$$
\begin{aligned}
A &= \mathbb{P}(F^{1,1}(t^{1,1}_{(k^{1,1})}) - F^{1,0}(t^{1,0}_{(k^{1,0})}) > \alpha) \\
&= \mathbb{P}(F^{1,0}(t^{1,0}_{(k^{1,0})}) < F^{1,1}(t^{1,1}_{(k^{1,1})}) - \alpha) \\
&= \mathbb{E}[\mathbb{P}(t^{1,0}_{(k^{1,0})} < F^{1,0^{-1}}(F^{1,1}(t^{1,1}_{(k^{1,1})}) - \alpha))\mathbb{1}\{F^{1,1}(t^{1,1}_{(k^{1,1})}) - \alpha > 0\} \mid t^{1,1}_{(k^{1,1})}] \\
&= \mathbb{E}\{\mathbb{P}[\text{at least } k^{1,0} \text{ of } t^{1,0}\text{'s are less than } F^{1,0^{-1}}(F^{1,1}(t^{1,1}_{(k^{1,1})}) - \alpha)]\mathbb{1}\{F^{1,1}(t^{1,1}_{(k^{1,1})}) - \alpha > 0\} \mid t^{1,1}_{(k^{1,1})}\}
\end{aligned}
$$

Following this, we obtain

$$
\begin{aligned}
A &= \mathbb{E}\{\sum_{j=k^{1,0}}^{n^{1,0}} \mathbb{P}[\text{exactly j of the } t^{1,0}\text{'s are less than } F^{1,0^{-1}}(F^{1,1}(t^{1,1}_{(k^{1,1})}) - \alpha)]\mathbb{1}\{F^{1,1}(t^{1,1}_{(k^{1,1})}) - \alpha > 0\} \mid t^{1,1}_{(k^{1,1})}\} \\
&= \mathbb{E}\{\sum_{j=k^{1,0}}^{n^{1,0}} \binom{n^{1,0}}{j} \mathbb{P}[t^{1,0} < F^{1,0^{-1}}(F^{1,1}(t^{1,1}_{(k^{1,1})}) - \alpha)]^j (1 - \mathbb{P}[t^{1,0} < F^{1,0^{-1}}(F^{1,1}(t^{1,1}_{(k^{1,1})}) - \alpha)])^{n^{1,0}-j} \\
&\quad \mathbb{1}\{F^{1,1}(t^{1,1}_{(k^{1,1})}) - \alpha > 0\} \mid t^{1,1}_{(k^{1,1})}\} \\
&\leq \mathbb{E}[\sum_{j=k^{1,0}}^{n^{1,0}} \binom{n^{1,0}}{j} (F^{1,1}(t^{1,1}_{(k^{1,1})}) - \alpha)^j (1 - (F^{1,1}(t^{1,1}_{(k^{1,1})}) - \alpha))^{n^{1,0}-j} \mid t^{1,1}_{(k^{1,1})}]
\end{aligned}
$$

Similarly, we have

$$
B \leq \mathbb{E}[\sum_{j=k^{1,1}}^{n^{1,1}} \binom{n^{1,1}}{j} (F^{1,0}(t^{1,0}_{(k^{1,0})}) - \alpha)^j (1 - (F^{1,0}(t^{1,0}_{(k^{1,0})}) - \alpha))^{n^{1,1}-j} \mid t^{1,0}_{(k^{1,0})}]
$$

Hence, we have

$$
A + B \leq \mathbb{E}[\sum_{j=k^{1,0}}^{n^{1,0}} \binom{n^{1,0}}{j} (F^{1,1}(t^{1,1}_{(k^{1,1})}) - \alpha)^j (1 - (F^{1,1}(t^{1,1}_{(k^{1,1})}) - \alpha))^{n^{1,0}-j} \mid t^{1,1}_{(k^{1,1})}] + \mathbb{E}[\sum_{j=k^{1,1}}^{n^{1,1}} \binom{n^{1,1}}{j} (F^{1,0}(t^{1,0}_{(k^{1,0})}) - \alpha)^j (1 - (F^{1,0}(t^{1,0}_{(k^{1,0})}) - \alpha))^{n^{1,1}-j} \mid t^{1,0}_{(k^{1,0})}].
$$

If $t^{1,a}$ is continuous random variable, the equality holds.

From Lemma 2, we have $F^{1,0}(t^{1,0}_{(k^{1,0})}) \sim Beta(k^{1,0}, n^{1,0} - k^{1,0} + 1)$ and $F^{1,1}(t^{1,1}_{(k^{1,1})}) \sim Beta(k^{1,1}, n^{1,1} - k^{1,1} + 1)$, and we complete the proof. □

A.2 Proof of Lemma 1

We first introduce the lemma (theorem E.4 in Zeng et al. (2022)):

Lemma 3. *(Fair Bayes-optimal Classifiers under Equality of Opportunity). Let $E^\star = \mathrm{DEOO}(f^\star)$. For any $\alpha > 0$, all fair Bayes-optimal classifiers $f^\star_{E,\alpha}$ under the fairness constraint $|\mathrm{DEOO}(f)| \leq \alpha$ are given as follows:*
- When $|E^\star| \leq \alpha$, $f^\star_{E,\alpha} = f^\star$
- When $|E^\star| > \alpha$, suppose $\mathbb{P}_{X|A=1,Y=1}\left(\eta_1(X) = \frac{p_1 p_{Y,1}}{2(p_1 p_{Y,1} - t^\star_{E,\alpha})}\right) = 0$, then for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$,

$$
f^\star_{E,\alpha}(x,a) = I\left(\eta_a(x) > \frac{p_a p_{Y,a}}{2p_a p_{Y,a} + (1-2a)t^\star_{E,\alpha}}\right)
$$

where $t^\star_{E,\alpha}$ is defined as

$$
t^\star_{E,\alpha} = \sup\left\{t : \mathbb{P}_{Y|A=1,Y=1}\left(\eta_1(X) > \frac{p_1 p_{Y,1}}{2p_1 p_{Y,1} - t}\right) > \mathbb{P}_{Y|A=0,Y=1}\left(\eta_0(X) > \frac{p_0 p_{Y,0}}{2p_0 p_{Y,0} + t}\right) + \frac{E^\star}{|E^\star|}\alpha\right\}.
$$

Now we come back to prove Proposition 1.

Proof. From Lemma 3, we have

$$t_0^* = \frac{p_0 p_{Y,0}}{2p_0 p_{Y,0} + t_{E,\alpha}^*} \tag{3}$$

$$t_1^* = \frac{p_1 p_{Y,1}}{2p_1 p_{Y,1} - t_{E,\alpha}^*} \tag{4}$$

Combine Eq. (3) and (4) together and we complete the proof. □

A.3 Proof of Proposition 2

We first provide a lemma for the mis-classification error of the classifier in the candidate set.

Lemma 4. $\mid \mathbb{P}(\hat{\phi}_i \neq Y) - [\frac{k_i^{1,0}}{n^{1,0}+1} p_0 p_{Y,0} + \frac{k_i^{1,1}}{n^{1,1}+1} p_1 p_{Y,1} + \frac{n^{0,0}+\frac{1}{2}-\mathbb{E}(k_i^{0,0})}{n^{0,0}+1} p_0(1-p_{Y,0}) + \frac{n^{0,1}+\frac{1}{2}-\mathbb{E}(k_i^{0,1})}{n^{0,1}+1} p_1(1-p_{Y,1})] \mid \leq \frac{p_0(1-p_{Y,0})}{2(n^{0,0}+1)} + \frac{p_1(1-p_{Y,1})}{2(n^{0,1}+1)}$

Similar to Lemma 2, we have the following lemma:

Lemma 5. $F^{0,0}(t^{0,0}_{(k^{1,0})}) \sim Beta(k^{0,0}, n^{0,0}-k^{0,0}+1)$, $F^{0,1}(t^{0,1}_{(k^{0,1})}) \sim Beta(k^{0,1}, n^{0,1}-k^{0,1}+1)$.

Now we come back to the proof of Lemma 4:

Proof of Lemma 4. The classifier is:

$$\hat{\phi} = \begin{cases} \mathbb{1}\{f(x,0) > t^{1,0}_{(k^{1,0})}\}, A=0 \\ \mathbb{1}\{f(x,1) > t^{1,1}_{(k^{1,1})}\}, A=1 \end{cases}$$

So we have the mis-classification error:

$$\begin{aligned}
\mathbb{P}(Y \neq \hat{Y}) &= \mathbb{P}(Y=1, \hat{Y}=0) + \mathbb{P}(Y=0, \hat{Y}=1) \\
&= \mathbb{P}(Y=1, \hat{Y}=0, A=0) + \mathbb{P}(Y=1, \hat{Y}=0, A=1) \\
&\quad + \mathbb{P}(Y=0, \hat{Y}=1, A=0) + \mathbb{P}(Y=0, \hat{Y}=1, A=1) \\
&= \mathbb{P}(\hat{Y}=0 | Y=1, A=0)\mathbb{P}(Y=1, A=0) + \mathbb{P}(\hat{Y}=0 | Y=1, A=1)\mathbb{P}(Y=1, A=1) \\
&\quad + \mathbb{P}(\hat{Y}=1 | Y=0, A=0)\mathbb{P}(Y=0, A=0) + \mathbb{P}(\hat{Y}=1 | Y=0, A=1)\mathbb{P}(Y=0, A=1) \\
&= \mathbb{E}[\mathbb{P}(f(x,0) \leq t^{1,0}_{(k^{1,0})} \mid Y=1, A=0) \mid t^{1,0}_{(k^{1,0})}] p_0 p_{Y,0} \\
&\quad + \mathbb{E}[\mathbb{P}(f(x,1) \leq t^{1,1}_{(k^{1,1})} \mid Y=1, A=1) \mid t^{1,1}_{(k^{1,1})}] p_1 p_{Y,1} \\
&\quad + \mathbb{E}[\mathbb{P}(f(x,0) \geq t^{1,0}_{(k^{1,0})} \mid Y=0, A=0) \mid t^{1,0}_{(k^{1,0})}] p_0 (1-p_{Y,0}) \\
&\quad + \mathbb{E}[\mathbb{P}(f(x,1) \geq t^{1,1}_{(k^{1,1})} \mid Y=0, A=1) \mid t^{1,1}_{(k^{1,1})}] p_1 (1-p_{Y,1}) \\
&\leq \mathbb{E}[\mathbb{P}(f(x,0) \leq t^{1,0}_{(k^{1,0})} \mid Y=1, A=0) \mid t^{1,0}_{(k^{1,0})}] p_0 p_{Y,0} \\
&\quad + \mathbb{E}[\mathbb{P}(f(x,1) \leq t^{1,1}_{(k^{1,1})} \mid Y=1, A=1) \mid t^{1,1}_{(k^{1,1})}] p_1 p_{Y,1} \\
&\quad + \mathbb{E}[\mathbb{P}(f(x,0) \geq t^{0,0}_{(k^{0,0})} \mid Y=0, A=0) \mid t^{1,0}_{(k^{1,0})}] p_0 (1-p_{Y,0}) \\
&\quad + \mathbb{E}[\mathbb{P}(f(x,1) \geq t^{0,1}_{(k^{0,1})} \mid Y=0, A=1) \mid t^{1,1}_{(k^{1,1})}] p_1 (1-p_{Y,1}) \\
&= \mathbb{E}[F^{1,0}(t^{1,0}_{(k^{1,0})}) \mid t^{1,0}_{(k^{1,0})}] p_0 p_{Y,0} + \mathbb{E}[F^{1,1}(t^{1,1}_{(k^{1,1})}) \mid t^{1,1}_{(k^{1,1})}] p_1 p_{Y,1} \\
&\quad + \mathbb{E}[1 - \mathbb{E}[F^{0,0}(t^{0,0}_{(k^{0,0})}) \mid t^{0,0}_{(k^{0,0})}] \mid t^{1,0}_{(k^{1,0})}] p_0 (1-p_{Y,0}) + \mathbb{E}[1 - \mathbb{E}[F^{0,1}(t^{0,1}_{(k^{0,1})}) \mid t^{0,1}_{(k^{0,1})}] \mid t^{1,1}_{(k^{1,1})}] p_1 (1-p_{Y,1}) \\
&= \frac{k_i^{1,0}}{n^{1,0}+1} p_0 p_{Y,0} + \frac{k_i^{1,1}}{n^{1,1}+1} p_1 p_{Y,1} + \frac{n^{0,0}+1-\mathbb{E}(k_i^{0,0})}{n^{0,0}+1} p_0(1-p_{Y,0}) + \frac{n^{0,1}+1-\mathbb{E}(k_i^{0,1})}{n^{0,1}+1} p_1(1-p_{Y,1})
\end{aligned}$$

The last equality comes from Lemma 2 and 5, and from the fact that $\mathbb{E}(Beta(\alpha,\beta)) = \frac{\alpha}{\alpha+\beta}$.

Similarly, we have

$$\mathbb{P}(Y \neq \hat{Y}) \geq \frac{k_i^{1,0}}{n^{1,0}+1} p_0 p_{Y,0} + \frac{k_i^{1,1}}{n^{1,1}+1} p_1 p_{Y,1} + \frac{n^{0,0} - \mathbb{E}(k_i^{0,0})}{n^{0,0}+1} p_0(1-p_{Y,0}) + \frac{n^{0,1} - \mathbb{E}(k_i^{0,1})}{n^{0,1}+1} p_1(1-p_{Y,1})$$

Thus, we have $\mid \mathbb{P}(\hat{\phi}_i \neq Y) - [\frac{k_i^{1,0}}{n^{1,0}+1} p_0 p_{Y,0} + \frac{k_i^{1,1}}{n^{1,1}+1} p_1 p_{Y,1} + \frac{n^{0,0}+\frac{1}{2}-\mathbb{E}(k_i^{0,0})}{n^{0,0}+1} p_0(1-p_{Y,0}) + \frac{n^{0,1}+\frac{1}{2}-\mathbb{E}(k_i^{0,1})}{n^{0,1}+1} p_1(1-p_{Y,1})] \mid \leq \frac{p_0(1-p_{Y,0})}{2(n^{0,0}+1)} + \frac{p_1(1-p_{Y,1})}{2(n^{0,1}+1)}$.

Now we complete the proof of Lemma 4.

□

Next, we come to prove Proposition 2.

Lemma 6 (Hoeffding's inequality)**.** *Let $X_1, \ldots, X_n$ be independent random variables. Assume that $X_i \in [m_i, M_i]$ for every i. Then, for any $t > 0$, we have*

$$\mathbb{P}\left\{\sum_{i=1}^{n}(X_i - \mathbb{E}X_i) \geq t\right\} \leq e^{-\frac{2t^2}{\sum_{i=1}^{n}(M_i - m_i)^2}}$$

Proof of Proposition 2. First, we notice that $k_i^{0,a}$ is the number of $t^{0,a}$'s such that $t^{0,a} < t^{1,a}_{(k_i^{1,a})}$, i.e. $k_i^{0,a} = \sum_{j=1}^{n^{0,1}} \mathbb{1}\{t_j^{0,a} < t^{1,a}_{(k_i^{1,a})}\}$.

Thus, for a given $\epsilon > 0$, from Hoeffding's inequality, we have with probability $1 - e^{-2n^{0,a}\epsilon^2}$,

$$\frac{k_i^{0,a} - \mathbb{E}(k_i^{0,a})}{n^{0,a}} = \frac{\sum_{j=1}^{n^{0,1}} \mathbb{1}\{t_j^{0,a} < t^{1,a}_{(k_i^{1,a})}\} - \sum_{j=1}^{n^{0,1}} \mathbb{E}(\mathbb{1}\{t_j^{0,a} < t^{1,a}_{(k_i^{1,a})}\})}{n^{0,a}} \leq \epsilon.$$

Similarly, with probability $1 - e^{-2n^{0,a}\epsilon^2}$,

$$\frac{k_i^{0,a} - \mathbb{E}(k_i^{0,a})}{n^{0,a}} = \frac{\sum_{j=1}^{n^{0,1}} \mathbb{1}\{t_j^{0,a} < t^{1,a}_{(k_i^{1,a})}\} - \sum_{j=1}^{n^{0,1}} \mathbb{E}(\mathbb{1}\{t_j^{0,a} < t^{1,a}_{(k_i^{1,a})}\})}{n^{0,a}} \geq -\epsilon.$$

Thus, $\mid \frac{k_i^{0,a} - \mathbb{E}(k_i^{0,a})}{n^{0,a}} \mid \leq \epsilon$ with probability $1 - 2e^{-2n^{0,a}\epsilon^2}$.

Then we estimate p_a and $p_{Y,a}$ by $\hat{p}_a = \frac{n^{1,a}+n^{0,a}}{n}$ and $\hat{p}_{Y,a} = \frac{n^{Y,a}}{n}$ (n is the number of the total samples). Here, $\frac{n^{1,a}+n^{0,a}}{n} = \frac{\sum_{i=1}^{n} \mathbb{1}\{Z_i^a = 1\}}{n}$ and $\frac{n^{Y,a}}{n} = \frac{\sum_{i=1}^{n} \mathbb{1}\{Z_i^{Y,a} = 1\}}{n}$, where $Z_i^a \sim B(1, p_a)$ and $Z_i^{Y,a} \sim B(1, p_{Y,a})$. From Hoeffding's inequality (Lemma 6), we have:

$$\mathbb{P}(\mid \hat{p}_a - p_a \mid \geq \sqrt{\frac{n^{0,a}}{n}}\epsilon) \leq 2e^{-2n^{0,a}\epsilon^2},$$

$$\mathbb{P}(\mid \hat{p}_{Y,a} - p_{Y,a} \mid \geq \sqrt{\frac{n^{0,a}}{n}}\epsilon) \leq 2e^{-2n^{0,a}\epsilon^2}$$

Thus, with probability $1 - 6e^{-2n^{0,a}\epsilon^2}$, we have:

$$\begin{cases} \mid \hat{p}_a - p_a \mid \leq \sqrt{\frac{n^{0,a}}{n}}\epsilon \\ \mid \hat{p}_{Y,a} - p_{Y,a} \mid \leq \sqrt{\frac{n^{0,a}}{n}}\epsilon \\ \mid \frac{k_i^{0,a} - \mathbb{E}(k_i^{0,a})}{n^{0,a}} \mid \leq \epsilon \end{cases}$$

Hence, we have with probability $1-6(e^{-2n^{0,0}h^2(\epsilon)}+e^{-2n^{0,1}h^2(\epsilon)})$,

$$
\begin{aligned}
&|\mathbb{P}(\hat{\phi}_i \neq Y) - \hat{\mathbb{P}}(\hat{\phi}_i \neq Y)| \\
\leq &|\frac{k_i^{1,0}}{n^{1,0}+1}p_0p_{Y,0} + \frac{k_i^{1,1}}{n^{1,1}+1}p_1p_{Y,1} + \frac{n^{0,0}+0.5-\mathbb{E}(k_i^{0,0})}{n^{0,0}+1}p_0(1-p_{Y,0}) + \frac{n^{0,1}+0.5-\mathbb{E}(k_i^{0,1})}{n^{0,1}+1}p_1(1-p_{Y,1}) \\
&- [\frac{k_i^{1,0}}{n^{1,0}+1}\hat{p}_0\hat{p}_{Y,0} + \frac{k_i^{1,1}}{n^{1,1}+1}\hat{p}_1\hat{p}_{Y,1} + \frac{n^{0,0}+0.5-k_i^{0,0}}{n^{0,0}+1}\hat{p}_0(1-\hat{p}_{Y,0}) + \frac{n^{0,1}+0.5-k_i^{0,1}}{n^{0,1}+1}\hat{p}_1(1-\hat{p}_{Y,1})]| \\
&+ \frac{p_0(1-p_{Y,0})}{2(n^{0,0}+1)} + \frac{p_1(1-p_{Y,1})}{2(n^{0,1}+1)}
\end{aligned}
$$

Then we have

$$
\begin{aligned}
&|\mathbb{P}(\hat{\phi}_i \neq Y) - \hat{\mathbb{P}}(\hat{\phi}_i \neq Y)| \\
\leq &\epsilon[\sqrt{\frac{n^{0,0}}{n}}\frac{k_i^{1,0}}{n^{1,0}+1}(p_0+p_{Y,0}) + \sqrt{\frac{n^{0,1}}{n}}\frac{k_i^{1,1}}{n^{1,1}+1}(p_1+p_{Y,1})] + \epsilon^2(\frac{n^{0,0}}{n}\frac{k_i^{1,0}}{n^{1,0}+1} + \frac{n^{0,1}}{n}\frac{k_i^{1,1}}{n^{1,1}+1}) \\
&+ \epsilon[\frac{n^{0,0}}{n^{0,0}+1}[p_0+p_0p_{Y,0} + \sqrt{\frac{n^{0,0}}{n}}\epsilon(\sqrt{\frac{n^{0,0}}{n}}\epsilon + p_0 + p_{Y,0} + 1)] \\
&+ \frac{n^{0,1}}{n^{0,1}+1}[p_1+p_1p_{Y,1} + \sqrt{\frac{n^{0,1}}{n}}\epsilon(\sqrt{\frac{n^{0,1}}{n}}\epsilon + p_1 + p_{Y,1} + 1)]] \\
&+ \frac{n^{0,0}+0.5-\mathbb{E}(k_i^{0,0})}{n^{0,0}+1}\sqrt{\frac{n^{0,0}}{n}}\epsilon[\sqrt{\frac{n^{0,0}}{n}}\epsilon + p_0 + p_{Y,0} + 1] \\
&+ \frac{n^{0,1}+0.5-\mathbb{E}(k_i^{0,1})}{n^{0,1}+1}\sqrt{\frac{n^{0,1}}{n}}\epsilon[\sqrt{\frac{n^{0,1}}{n}}\epsilon + p_1 + p_{Y,1} + 1] + \frac{p_0(1-p_{Y,0})}{2(n^{0,0}+1)} + \frac{p_1(1-p_{Y,1})}{2(n^{0,1}+1)} \\
\leq &\epsilon[\sqrt{\frac{n^{0,0}}{n}}(p_0+p_{Y,0}) + \sqrt{\frac{n^{0,1}}{n}}(p_1+p_{Y,1})] + \epsilon^2(\frac{n^{0,0}}{n} + \frac{n^{0,1}}{n}) \\
&+ \epsilon[\frac{n^{0,0}}{n^{0,0}+1}[p_0+p_0p_{Y,0} + \sqrt{\frac{n^{0,0}}{n}}\epsilon(\sqrt{\frac{n^{0,0}}{n}}\epsilon + p_0 + p_{Y,0} + 1)] \\
&+ \frac{n^{0,1}}{n^{0,1}+1}[p_1+p_1p_{Y,1} + \sqrt{\frac{n^{0,1}}{n}}\epsilon(\sqrt{\frac{n^{0,1}}{n}}\epsilon + p_1 + p_{Y,1} + 1)]] \\
&+ \sqrt{\frac{n^{0,0}}{n}}\epsilon[\sqrt{\frac{n^{0,0}}{n}}\epsilon + p_0 + p_{Y,0} + 1] + \sqrt{\frac{n^{0,1}}{n}}\epsilon[\sqrt{\frac{n^{0,1}}{n}}\epsilon + p_1 + p_{Y,1} + 1] \\
&+ \frac{p_0(1-p_{Y,0})}{2(n^{0,0}+1)} + \frac{p_1(1-p_{Y,1})}{2(n^{0,1}+1)} \\
\leq &2\epsilon + \epsilon^2 + \epsilon[\epsilon^2 + 4\epsilon + 2] + \epsilon^2 + 4\epsilon + \frac{p_0(1-p_{Y,0})}{2(n^{0,0}+1)} + \frac{p_1(1-p_{Y,1})}{2(n^{0,1}+1)} \\
= &\epsilon^3 + 6\epsilon^2 + 8\epsilon + \frac{p_0(1-p_{Y,0})}{2(n^{0,0}+1)} + \frac{p_1(1-p_{Y,1})}{2(n^{0,1}+1)}
\end{aligned}
$$

Thus we complete the proof. □

A.4 THEOREM FOR THE ORIGINAL CANDIDATE SET K

Sometimes we would use the original candidate set K instead of the small set K' to achieve the optimal accuracy more precisely. Now we provide a theorem for candidate set K.

Theorem 5. *Given $\alpha' < \alpha$. If $\min\{n^{1,0}, n^{1,1}\} \geq \lceil \frac{\log \frac{\delta}{2}}{\log(1-\alpha)} \rceil$. For any $\epsilon > 0$ such that $g(\epsilon) \leq \frac{\alpha-\alpha'}{2}$. Suppose $\hat{\phi}$ is the final output of FaiREE, we have:*
(1) $|DEOO(\hat{\phi})| < \alpha$ with probability $(1-\delta)^M$, where M is the size of the candidate set.
(2) Suppose the density distribution functions of f^* under $A=a, Y=1$ are continuous. When the input classifier f is f^*,

$$
|\ \mathbb{P}(\hat{\phi} \neq Y) - \mathbb{P}(\phi^* \neq Y)\ | \leq 2g(\epsilon) + 2h^3(\epsilon) + 12h^2(\epsilon) + 16h(\epsilon) + \frac{p_0(1-p_{Y,0})}{n^{0,0}+1} + \frac{p_1(1-p_{Y,1})}{n^{0,1}+1}
$$

with probability $1-(2M+4)(e^{-2n^{0,0}h^2(\epsilon)}+e^{-2n^{0,1}h^2(\epsilon)})-(1-h(2\epsilon))^{n^{1,0}}-(1-h(2\epsilon))^{n^{1,1}}$.

The (1) of the theorem is a direct corollary from Theorem 1, now we prove the (2) of the theorem.

To facilitate the theoretical analysis, we first introduce the following lemma which implies that the difference between the output of the function can be controlled by the difference between the input. (i.e. the cumulative distribution function won't increase drastically.)

Lemma 7. *For a distribution F with a continuous density function, suppose $q(x)$ denotes the quantile of x under F, then for $x > y$, we have $h(x-y) \leq q(x)-q(y) \leq g(x-y)$, where $h(x)$ and $g(x)$ are two monotonically increasing functions, $h(\epsilon) > 0, g(\epsilon) > 0$ for any $\epsilon > 0$ and $\lim_{\epsilon\to 0} h(\epsilon) = \lim_{\epsilon\to 0} g(\epsilon) = 0$.*

Proof of lemma 7. Since the domain of $q(x)$ is a closed set and $q(x)$ is continuous, we know that $q(x)$ is uniformly continuous. Thus we can easily find g to satisfy the RHS. For h, we simply define $h(t) = \inf_x\{q(x+t)-q(t)\}$. Since $q(x+t)-q(t) > 0$ for $t > 0$ and the domain of x is a closed set, we have $h(\epsilon) > 0$ for $\epsilon > 0$ and $\lim_{\epsilon\to 0} h(\epsilon) = 0$. Now we complete the proof. □

Now we come to prove Theorem 5.

Proof of Theorem 5. The proof can be divided into two parts. The first part is to prove that there exist classifiers in our candidate set that are close to fair Bayes-optimal classifier. The second part is to prove that our algorithm can successfully choose one of these classifiers with high probability.

We suppose the fair Bayes optimal classifier has the form $\phi^*_\alpha(x,a) = \mathbb{1}\{f^*(x,a) > \frac{1}{2}+\lambda^*_a\}$. And the output classifier of our algorithm is of the form $\hat{\phi} = \mathbb{1}\{f(x,a) > \frac{1}{2}+\lambda_a\}$.

For the first part, for any $\epsilon > 0$, from Assumption 7, $t^{1,a}$ has a positive probability $h(2\epsilon)$ to fall in the interval $[\lambda^*_a-\epsilon, \lambda^*_a+\epsilon]$, which implies that the probability that there exists $a \in \{0,1\}$ such that all $t^{1,a}$'s fall out of $[\lambda^*_a-\epsilon, \lambda^*_a+\epsilon]$ is less than $(1-h(2\epsilon))^{n^{1,0}}+(1-h(2\epsilon))^{n^{1,1}}$. So with probability $1-(1-h(2\epsilon))^{n^{1,0}}-(1-h(2\epsilon))^{n^{1,1}}$, there will exist $t^{1,a}$ in $[\lambda^*_a-\epsilon, \lambda^*_a+\epsilon]$, which we denote as $\phi_0 = \mathbb{1}\{f^*(x,a) > t^{1,a}_*\}$. Hence the gap between the classifier ϕ_0 and the Bayes-optimal classifier will be very close. In detail, we have

$$\begin{aligned}
\mid \mathbb{P}(\phi_0 \neq Y) - \mathbb{P}(\phi^*_{\alpha'} \neq Y) \mid &\leq \mathbb{P}(\min\{t^{1,a}_*, \lambda^*_a\} \leq f^*(x,a) \leq max\{t^{1,a}_*, \lambda^*_a\}) \\
\text{(Assumption 7)} \quad &\leq g(max\{t^{1,a}_*, \lambda^*_a\} - \min\{t^{1,a}_*, \lambda^*_a\}) \\
&\leq 2g(\epsilon)
\end{aligned}$$

so we complete the first part of the proof.

Now we come to the second part. First, we notice that $DEOO(\phi_0)$ and $DEOO(\phi^*_{\alpha'})$ are close to each other.

$$\begin{aligned}
&\mid\mid DEOO(\phi_0) \mid - \mid DEOO(\phi^*_{\alpha'}) \mid\mid \\
=&\mid\mid \mathbb{P}(f^* > t^{1,0}_* \mid Y=1, A=0) - \mathbb{P}(f^* > t^{1,1}_* \mid Y=1, A=1) \mid \\
&- \mid \mathbb{P}(f^* > \lambda^*_0 \mid Y=1, A=0) - \mathbb{P}(f^* > \lambda^*_1 \mid Y=1, A=1) \mid\mid \\
\leq&\mid \mathbb{P}(f^* > t^{1,0}_* \mid Y=1, A=0) - \mathbb{P}(f^* > t^{1,1}_* \mid Y=1, A=1) \\
&- \mathbb{P}(f^* > \lambda^*_0 \mid Y=1, A=0) + \mathbb{P}(f^* > \lambda^*_1 \mid Y=1, A=1) \mid \\
\leq&\mathbb{P}(\min\{t^{1,a}_*, \lambda^*_a\} \leq f^*(x,a) \leq max\{t^{1,a}_*, \lambda^*_a\}) \\
\text{(Assumption 7)} \quad &\leq g(max\{t^{1,a}_*, \lambda^*_a\} - \min\{t^{1,a}_*, \lambda^*_a\}) \\
&\leq 2g(\epsilon)
\end{aligned}$$

Thus, $\mid DEOO(\phi_0) \mid \leq \mid DEOO(\phi_{\alpha'}) \mid + 2g(\epsilon) = \alpha' + 2g(\epsilon)$. If $g(\epsilon) \leq \frac{\alpha-\alpha'}{2}$, then there will exist at least one feasible classifier in the candidate set.

From Lemma 4, we have the mis-classification error

$| \mathbb{P}(\hat{\phi}_i \neq Y) - [\frac{k_i^{1,0}}{n^{1,0}+1} p_0 p_{Y,0} + \frac{k_i^{1,1}}{n^{1,1}+1} p_1 p_{Y,1} + \frac{n^{0,0}+\frac{1}{2}-\mathbb{E}(k_i^{0,0})}{n^{0,0}+1} p_0(1-p_{Y,0}) + \frac{n^{0,1}+\frac{1}{2}-\mathbb{E}(k_i^{0,1})}{n^{0,1}+1} p_1(1-p_{Y,1})] | \leq \frac{p_0(1-p_{Y,0})}{2(n^{0,0}+1)} + \frac{p_1(1-p_{Y,1})}{2(n^{0,1}+1)}$.

If we can accurately estimate the mis-classification error, than the second part is almost done. For the estimation of $\mathbb{E}(k_i^{0,0})$, we can easily use $k_i^{0,0}$. We notice that $k_i^{0,a}$ is the number of $t^{0,a}$'s such that $t^{0,a} < t^{1,a}_{(k_i^{1,a})}$, i.e. $k_i^{0,a} = \sum_{i=1}^{n^{0,1}} \mathbb{1}\{t_i^{0,a} < t^{1,a}_{(k_i^{1,a})}\}$.

Thus, from Hoeffding's inequality, we have with probability $1 - e^{-2n^{0,a}h^2(\epsilon)}$,

$$\frac{k_i^{0,a} - \mathbb{E}(k_i^{0,a})}{n^{0,a}} = \frac{\sum_{j=1}^{n^{0,1}} \mathbb{1}\{t_j^{0,a} < t^{1,a}_{(k_i^{1,a})}\} - \sum_{j=1}^{n^{0,1}} \mathbb{E}(\mathbb{1}\{t_j^{0,a} < t^{1,a}_{(k_i^{1,a})}\})}{n^{0,a}} \leq h(\epsilon).$$

Similarly, with probability $1 - e^{-2n^{0,a}h^2(\epsilon)}$,

$$\frac{k_i^{0,a} - \mathbb{E}(k_i^{0,a})}{n^{0,a}} = \frac{\sum_{j=1}^{n^{0,1}} \mathbb{1}\{t_j^{0,a} < t^{1,a}_{(k_i^{1,a})}\} - \sum_{j=1}^{n^{0,1}} \mathbb{E}(\mathbb{1}\{t_j^{0,a} < t^{1,a}_{(k_i^{1,a})}\})}{n^{0,a}} \geq -h(\epsilon) - \mathbb{P}(h_{\frac{k_i^{1,a}}{n^{1,a}}} - \epsilon < t_i^{0,a} < h_{\frac{k_i^{1,a}}{n^{1,a}}} + \epsilon).$$

Thus, $| \frac{k_i^{0,a} - \mathbb{E}(k_i^{0,a})}{n^{0,a}} | \leq h(\epsilon)$ with probability $1 - 2e^{-2n^{0,a}h^2(\epsilon)}$.

Then, it's easy to estimate p_a and $p_{Y,a}$ with $\hat{p}_a = \frac{n^{1,a}+n^{0,a}}{n}$ and $\hat{p}_{Y,a} = \frac{n^{Y,a}}{n}$ (n is the number of the total samples). Here, $\frac{n^{1,a}+n^{0,a}}{n} = \frac{\sum_{i=1}^{n} \mathbb{1}\{Z_i^a = 1\}}{n}$ and $\frac{n^{Y,a}}{n} = \frac{\sum_{i=1}^{n} \mathbb{1}\{Z_i^{Y,a} = 1\}}{n}$, where $Z_i^a \sim B(1, p_a)$ and $Z_i^{Y,a} \sim B(1, p_{Y,a})$. From Hoeffding's inequality, we have:

$$\mathbb{P}(| \hat{p}_a - p_a | \geq \sqrt{\frac{n^{0,a}}{n}} h(\epsilon)) \leq 2e^{-2n^{0,a}h^2(\epsilon)},$$

$$\mathbb{P}(| \hat{p}_{Y,a} - p_{Y,a} | \geq \sqrt{\frac{n^{0,a}}{n}} h(\epsilon)) \leq 2e^{-2n^{0,a}h^2(\epsilon)}$$

Thus, with probability $1 - 6e^{-2n^{0,a}h^2(\epsilon)}$, we have:

$$\begin{cases} | \hat{p}_a - p_a | \leq \sqrt{\frac{n^{0,a}}{n}} h(\epsilon) \\ | \hat{p}_{Y,a} - p_{Y,a} | \leq \sqrt{\frac{n^{0,a}}{n}} h(\epsilon) \\ | \frac{k_i^{0,a} - \mathbb{E}(k_i^{0,a})}{n^{0,a}} | \leq h(\epsilon) \end{cases}$$

Hence, we have with probability $1-(2M+4)(e^{-2n^{0,0}h^2(\epsilon)}+e^{-2n^{0,1}h^2(\epsilon)})$, for each $i \in \{1, \ldots, M\}$,

$$\begin{aligned} & |\mathbb{P}(\hat{\phi}_i \neq Y) - \hat{\mathbb{P}}(\hat{\phi}_i \neq Y)| \\ \leq & |\frac{k_i^{1,0}}{n^{1,0}+1} p_0 p_{Y,0} + \frac{k_i^{1,1}}{n^{1,1}+1} p_1 p_{Y,1} + \frac{n^{0,0}+0.5-\mathbb{E}(k_i^{0,0})}{n^{0,0}+1} p_0(1-p_{Y,0}) + \frac{n^{0,1}+0.5-\mathbb{E}(k_i^{0,1})}{n^{0,1}+1} p_1(1-p_{Y,1}) \\ & - [\frac{k_i^{1,0}}{n^{1,0}+1} \hat{p}_0 \hat{p}_{Y,0} + \frac{k_i^{1,1}}{n^{1,1}+1} \hat{p}_1 \hat{p}_{Y,1} + \frac{n^{0,0}+0.5-k_i^{0,0}}{n^{0,0}+1} \hat{p}_0(1-\hat{p}_{Y,0}) + \frac{n^{0,1}+0.5-k_i^{0,1}}{n^{0,1}+1} \hat{p}_1(1-\hat{p}_{Y,1})]| \\ & + \frac{p_0(1-p_{Y,0})}{2(n^{0,0}+1)} + \frac{p_1(1-p_{Y,1})}{2(n^{0,1}+1)} \end{aligned}$$

Then we have

$$
\begin{aligned}
&|\mathbb{P}(\hat{\phi}_i \neq Y) - \hat{\mathbb{P}}(\hat{\phi}_i \neq Y)| \\
\leq & h(\epsilon)[\sqrt{\frac{n^{0,0}}{n}}\frac{k_i^{1,0}}{n^{1,0}+1}(p_0+p_{Y,0}) + \sqrt{\frac{n^{0,1}}{n}}\frac{k_i^{1,1}}{n^{1,1}+1}(p_1+p_{Y,1})] + h^2(\epsilon)(\frac{n^{0,0}}{n}\frac{k_i^{1,0}}{n^{1,0}+1} + \frac{n^{0,1}}{n}\frac{k_i^{1,1}}{n^{1,1}+1}) \\
& + h(\epsilon)[\frac{n^{0,0}}{n^{0,0}+1}[p_0 + p_0 p_{Y,0} + \sqrt{\frac{n^{0,0}}{n}}h(\epsilon)(\sqrt{\frac{n^{0,0}}{n}}h(\epsilon) + p_0 + p_{Y,0} + 1)] \\
& + \frac{n^{0,1}}{n^{0,1}+1}[p_1 + p_1 p_{Y,1} + \sqrt{\frac{n^{0,1}}{n}}h(\epsilon)(\sqrt{\frac{n^{0,1}}{n}}h(\epsilon) + p_1 + p_{Y,1} + 1)]] \\
& + \frac{n^{0,0} + 0.5 - \mathbb{E}(k_i^{0,0})}{n^{0,0}+1}\sqrt{\frac{n^{0,0}}{n}}h(\epsilon)[\sqrt{\frac{n^{0,0}}{n}}h(\epsilon) + p_0 + p_{Y,0} + 1] \\
& + \frac{n^{0,1} + 0.5 - \mathbb{E}(k_i^{0,1})}{n^{0,1}+1}\sqrt{\frac{n^{0,1}}{n}}h(\epsilon)[\sqrt{\frac{n^{0,1}}{n}}h(\epsilon) + p_1 + p_{Y,1} + 1] \\
& + \frac{p_0(1-p_{Y,0})}{2(n^{0,0}+1)} + \frac{p_1(1-p_{Y,1})}{2(n^{0,1}+1)}
\end{aligned}
$$

Further, we obtain

$$
\begin{aligned}
&|\mathbb{P}(\hat{\phi}_i \neq Y) - \hat{\mathbb{P}}(\hat{\phi}_i \neq Y)| \\
\leq & h(\epsilon)[\sqrt{\frac{n^{0,0}}{n}}(p_0+p_{Y,0}) + \sqrt{\frac{n^{0,1}}{n}}(p_1+p_{Y,1})] + h^2(\epsilon)(\frac{n^{0,0}}{n} + \frac{n^{0,1}}{n}) \\
& + h(\epsilon)[\frac{n^{0,0}}{n^{0,0}+1}[p_0 + p_0 p_{Y,0} + \sqrt{\frac{n^{0,0}}{n}}h(\epsilon)(\sqrt{\frac{n^{0,0}}{n}}h(\epsilon) + p_0 + p_{Y,0} + 1)] \\
& + \frac{n^{0,1}}{n^{0,1}+1}[p_1 + p_1 p_{Y,1} + \sqrt{\frac{n^{0,1}}{n}}h(\epsilon)(\sqrt{\frac{n^{0,1}}{n}}h(\epsilon) + p_1 + p_{Y,1} + 1)]] \\
& + \sqrt{\frac{n^{0,0}}{n}}h(\epsilon)[\sqrt{\frac{n^{0,0}}{n}}h(\epsilon) + p_0 + p_{Y,0} + 1] \\
& + \sqrt{\frac{n^{0,1}}{n}}h(\epsilon)[\sqrt{\frac{n^{0,1}}{n}}h(\epsilon) + p_1 + p_{Y,1} + 1] \\
& + \frac{p_0(1-p_{Y,0})}{2(n^{0,0}+1)} + \frac{p_1(1-p_{Y,1})}{2(n^{0,1}+1)} \\
\leq & 2h(\epsilon) + h^2(\epsilon) + h(\epsilon)[h^2(\epsilon) + 4h(\epsilon) + 2] + h^2(\epsilon) + 4h(\epsilon) \\
& + \frac{p_0(1-p_{Y,0})}{2(n^{0,0}+1)} + \frac{p_1(1-p_{Y,1})}{2(n^{0,1}+1)} \\
= & h^3(\epsilon) + 6h^2(\epsilon) + 8h(\epsilon) + \frac{p_0(1-p_{Y,0})}{2(n^{0,0}+1)} + \frac{p_1(1-p_{Y,1})}{2(n^{0,1}+1)}
\end{aligned}
$$

Combining two parts together, we have:
with probability $1 - (2M+4)(e^{-2n^{0,0}h^2(\epsilon)} + e^{-2n^{0,1}h^2(\epsilon)}) - (1-h(2\epsilon))^{n^{1,0}} - (1-h(2\epsilon))^{n^{1,1}}$,

$$
| \mathbb{P}(\hat{\phi} \neq Y) - \mathbb{P}(\phi^* \neq Y) | \leq 2g(\epsilon) + 2h^3(\epsilon) + 12h^2(\epsilon) + 16h(\epsilon) + \frac{p_0(1-p_{Y,0})}{n^{0,0}+1} + \frac{p_1(1-p_{Y,1})}{n^{0,1}+1}.
$$

Now we complete the proof. □

A.5 Proof of Theorem 2

Proof. The (1) of the theorem is a direct corollary from Theorem 1, now we prove the (2) of the theorem.

It's sufficient to modify the first part of the proof of Theorem 5 and the second part simply follows Theorem 5.

For the first part, for any $\epsilon > 0$, from Assumption 7, $t^{1,0}$ has a positive probability $h(2\epsilon)$ to fall in the interval $[\lambda_0^* - \epsilon, \lambda_0^* + \epsilon]$, which implies that the probability that all $t^{1,0}$'s fall out of $[\lambda_0^* - \epsilon, \lambda_0^* + \epsilon]$ is less than $(1-h(2\epsilon))^{n^{1,0}}$. So with probability $1-(1-h(2\epsilon))^{n^{1,0}}$, there will exist $t^{1,0}$ in $[\lambda_0^*-\epsilon, \lambda_0^*+\epsilon]$, which we denote as λ_0. We denote the corresponding classifier as $\mathbb{1}\{f^*(x,a) > \lambda_a\}$. From the proof of Theorem 2, we have with probability $1 - 4e^{-2n^{0,a}h^2(\epsilon)} - (1-h(2\epsilon))^{n^{1,0}}$,

$$\begin{cases} \mid \hat{p}_a - p_a \mid \leq \sqrt{\dfrac{n^{0,a}}{n}} h(\epsilon) \\ \mid \hat{p}_{Y,a} - p_{Y,a} \mid \leq \sqrt{\dfrac{n^{0,a}}{n}} h(\epsilon) \\ \mid \lambda_0 - \lambda_0^* \mid \leq \epsilon \end{cases}$$

We have the following equalities:

$$\begin{cases} \lambda_1^* = \dfrac{1}{2 - \frac{(\frac{1}{\lambda_0^*}-2)p_0 p_{Y,0}}{p_1 p_{Y,1}}} \\ \lambda_1 = \dfrac{1}{2 - \frac{(\frac{1}{\lambda_0}-2)\hat{p}_0 \hat{p}_{Y,0}}{\hat{p}_1 \hat{p}_{Y,1}}} \end{cases}$$

Hence,

$$\begin{cases} \dfrac{p_0 p_{Y,0}}{\lambda_0^*} + \dfrac{p_1 p_{Y,1}}{\lambda_1^*} = 2(p_0 p_{Y,0} + p_1 p_{Y,1}) \\ \dfrac{\hat{p}_0 \hat{p}_{Y,0}}{\lambda_0} + \dfrac{\hat{p}_1 \hat{p}_{Y,1}}{\lambda_1} = 2(\hat{p}_0 \hat{p}_{Y,0} + \hat{p}_1 \hat{p}_{Y,1}) \end{cases}$$

By subtracting, we have

$$(2 - \frac{1}{\lambda_1})\hat{p}_1 \hat{p}_{Y,1} - (2 - \frac{1}{\lambda_1^*})p_1 p_{Y,1} = (\frac{1}{\lambda_0} - 2)\hat{p}_0 \hat{p}_{Y,0} - (\frac{1}{\lambda_0^*} - 2)p_0 p_{Y,0}.$$

We have

$$\begin{aligned} &(\frac{1}{\lambda_0} - 2)\hat{p}_0 \hat{p}_{Y,0} - (\frac{1}{\lambda_0^*} - 2)p_0 p_{Y,0} \\ \leq &(\frac{1}{\lambda_0} - 2)(p_0 + \sqrt{\frac{n^{0,0}}{n}} h(\epsilon))(p_{Y,0} + \sqrt{\frac{n^{0,0}}{n}} h(\epsilon)) - (\frac{1}{\lambda_0^*} - 2)p_0 p_{Y,0} \\ \leq &\frac{\epsilon}{\lambda_0^*(\lambda_0^* - \epsilon)} p_0 p_{Y,0} + \sqrt{\frac{n^{0,0}}{n}} h(\epsilon)(\frac{1}{\lambda_0^* - \epsilon} - 2)(p_0 + p_{Y,0} + \sqrt{\frac{n^{0,0}}{n}} h(\epsilon)) \end{aligned}$$

Similarly, we have

$$\begin{aligned} &(\frac{1}{\lambda_0} - 2)\hat{p}_0 \hat{p}_{Y,0} - (\frac{1}{\lambda_0^*} - 2)p_0 p_{Y,0} \\ \geq &-\frac{\epsilon}{\lambda_0^*(\lambda_0^* + \epsilon)} p_0 p_{Y,0} + \sqrt{\frac{n^{0,0}}{n}} h(\epsilon)(\frac{1}{\lambda_0^* + \epsilon} - 2)(-p_0 - p_{Y,0} + \sqrt{\frac{n^{0,0}}{n}} h(\epsilon)); \\ &(2 - \frac{1}{\lambda_1})\hat{p}_1 \hat{p}_{Y,1} - (2 - \frac{1}{\lambda_1^*})p_1 p_{Y,1} \\ \leq &(\frac{1}{\lambda_1^*} - \frac{1}{\lambda_1})p_1 p_{Y,1} + \sqrt{\frac{n^{0,1}}{n}} h(\epsilon)(2 - \frac{1}{\lambda_1})(p_1 + p_{Y,1} + \sqrt{\frac{n^{0,1}}{n}} h(\epsilon)); \\ &(2 - \frac{1}{\lambda_1})\hat{p}_1 \hat{p}_{Y,1} - (2 - \frac{1}{\lambda_1^*})p_1 p_{Y,1} \\ \geq &(\frac{1}{\lambda_1^*} - \frac{1}{\lambda_1})p_1 p_{Y,1} + \sqrt{\frac{n^{0,1}}{n}} h(\epsilon)(2 - \frac{1}{\lambda_1})(-p_1 - p_{Y,1} + \sqrt{\frac{n^{0,1}}{n}} h(\epsilon)). \end{aligned}$$

Now, we get:

$$\begin{cases} (\frac{1}{\lambda_1^*}-\frac{1}{\lambda_1})p_1p_{Y,1}+\sqrt{\frac{n^{0,1}}{n}}h(\epsilon)(2-\frac{1}{\lambda_1})(-p_1-p_{Y,1}+\sqrt{\frac{n^{0,1}}{n}}h(\epsilon)) \\ \leq \frac{\epsilon}{\lambda_0^*(\lambda_0^*-\epsilon)}p_0p_{Y,0}+\sqrt{\frac{n^{0,0}}{n}}h(\epsilon)(\frac{1}{\lambda_0^*-\epsilon}-2)(p_0+p_{Y,0}+\sqrt{\frac{n^{0,0}}{n}}h(\epsilon)) \\ (\frac{1}{\lambda_1^*}-\frac{1}{\lambda_1})p_1p_{Y,1}+\sqrt{\frac{n^{0,1}}{n}}h(\epsilon)(2-\frac{1}{\lambda_1})(p_1+p_{Y,1}+\sqrt{\frac{n^{0,1}}{n}}h(\epsilon)) \\ \geq -\frac{\epsilon}{\lambda_0^*(\lambda_0^*+\epsilon)}p_0p_{Y,0}+\sqrt{\frac{n^{0,0}}{n}}h(\epsilon)(\frac{1}{\lambda_0^*+\epsilon}-2)(-p_0-p_{Y,0}+\sqrt{\frac{n^{0,0}}{n}}h(\epsilon)) \end{cases}$$

Hence, we have:

$$\begin{aligned} &\lambda_1-\lambda_1^* \\ \leq &\frac{\lambda_1\lambda_1^*}{p_1p_{Y,1}}[\frac{\epsilon}{\lambda_0^*(\lambda_0^*-\epsilon)}p_0p_{Y,0}+\sqrt{\frac{n^{0,0}}{n}}h(\epsilon)(\frac{1}{\lambda_0^*-\epsilon}-2)(p_0+p_{Y,0}+\sqrt{\frac{n^{0,0}}{n}}h(\epsilon))-2\sqrt{\frac{n^{0,1}}{n}}h(\epsilon)(-p_1-p_{Y,1}+\sqrt{\frac{n^{0,1}}{n}}h(\epsilon))] \end{aligned}$$

and

$$\begin{aligned} &\lambda_1-\lambda_1^* \\ \geq &\frac{\lambda_1\lambda_1^*}{p_1p_{Y,1}}[-\frac{\epsilon}{\lambda_0^*(\lambda_0^*+\epsilon)}p_0p_{Y,0}+\sqrt{\frac{n^{0,0}}{n}}h(\epsilon)(\frac{1}{\lambda_0^*+\epsilon}-2)(-p_0-p_{Y,0}+\sqrt{\frac{n^{0,0}}{n}}h(\epsilon))-2\sqrt{\frac{n^{0,1}}{n}}h(\epsilon)(p_1+p_{Y,1}+\sqrt{\frac{n^{0,1}}{n}}h(\epsilon))] \end{aligned}$$

Thus, we have:

$$\mid \lambda_1-\lambda_1^* \mid \leq \frac{\frac{\epsilon}{\lambda_0^*(\lambda_0^*-\epsilon)}p_0p_{Y,0}+h(\epsilon)(\frac{1}{\lambda_0^*-\epsilon}-2)(2+h(\epsilon))+4h(\epsilon)}{p_1p_{Y,1}}$$

Now, combined with above, we have with probability $1-4e^{-2n^{0,a}h^2(\epsilon)}-(1-h(2\epsilon))^{n^{1,0}}$,

$$\begin{cases} \mid \hat{p}_a-p_a \mid \leq \sqrt{\frac{n^{0,a}}{n}}h(\epsilon) \\ \mid \hat{p}_{Y,a}-p_{Y,a} \mid \leq \sqrt{\frac{n^{0,a}}{n}}h(\epsilon) \\ \mid \lambda_0-\lambda_0^* \mid \leq \epsilon \\ \mid \lambda_1-\lambda_1^* \mid \leq \frac{\frac{\epsilon}{\lambda_0^*(\lambda_0^*-\epsilon)}p_0p_{Y,0}+h(\epsilon)(\frac{1}{\lambda_0^*-\epsilon}-2)(2+h(\epsilon))+4h(\epsilon)}{p_1p_{Y,1}} \end{cases}$$

From the proof of Theorem 2, we have:
If $g(\epsilon) \leq \frac{\alpha-\alpha'}{2}$, then with probability $1-(2M+4)(e^{-2n^{0,0}h^2(\epsilon)}+e^{-2n^{0,1}h^2(\epsilon)})-(1-h(2\epsilon))^{n^{1,0}}$,

$$\begin{aligned} &\mid \mathbb{P}(\hat{\phi}\neq Y)-\mathbb{P}(\phi^*\neq Y) \mid \\ \leq &g(\epsilon)+g(\frac{\frac{\epsilon}{\lambda_0^*(\lambda_0^*-\epsilon)}p_0p_{Y,0}+h(\epsilon)(\frac{1}{\lambda_0^*-\epsilon}-2)(2+h(\epsilon))+4h(\epsilon)}{p_1p_{Y,1}})+2h^3(\epsilon)+12h^2(\epsilon)+16h(\epsilon). \end{aligned}$$

Now we complete the proof. □

A.6 Fair Bayes-optimal Classifiers under Equalized Odds

Theorem 6 (Fair Bayes-optimal Classifiers under Equalized Odds)**.** *Let $EO^\star = \mathrm{DEO}(f^\star) = (E^\star, P^\star)$. For any $\alpha > 0$, there exist $0 < \alpha_1 \leq \alpha$ and $0 < \alpha_2 \leq \alpha$ such that all fair Bayes-optimal classifiers $f^\star_{EO,\alpha}$ under the fairness constraint $|\mathrm{DEO}(f)| \preceq (\alpha_1, \alpha_2)$ are given as below:*

- *When $|EO^\star| \preceq (\alpha_1, \alpha_2)$, $f^\star_{EO,\alpha} = f^\star$.*
- *When $E^* > \alpha_1$ or $P^* > \alpha_2$, for all $x \in \mathcal{X}$ and $a \in \mathcal{A}$, there exist $t^\star_{1,EO,\alpha}$ and $t^\star_{2,EO,\alpha}$ such that*

$$f^\star_{EO,\alpha}(x,a) =$$
$$I\left(\eta_a(x) > \frac{p_a p_{Y,a} + (2a-1)\frac{P_{Y,a}}{1-P_{Y,a}} t^\star_{2,EO,\alpha}}{2p_a p_{Y,a} + (2a-1)(\frac{P_{Y,a}}{1-P_{Y,a}} t^\star_{2,EO,\alpha} - t^\star_{1,EO,\alpha})}\right)$$
$$+ a\tau^\star_{EO,\alpha} I\left(\eta_a(x) = \frac{p_a p_{Y,a} + (2a-1)\frac{P_{Y,a}}{1-P_{Y,a}} t^\star_{2,EO,\alpha}}{2p_a p_{Y,a} + (2a-1)(\frac{P_{Y,a}}{1-P_{Y,a}} t^\star_{2,EO,\alpha} - t^\star_{1,EO,\alpha})}\right),$$

Here, we assume $P_{X|A=1,Y=1}\left(\eta_1(X) = \frac{p_1 p_{Y,1} + \frac{p_{Y,1}}{1-p_{Y,1}} t^\star_{2,EO,\alpha}}{2p_1 p_{Y,1} + \frac{p_{Y,1}}{1-p_{Y,1}} t^\star_{2,EO,\alpha} - t^\star_{1,EO,\alpha}}\right)$ $= P_{X|A=0,Y=0}\left(\eta_1(X) = \frac{p_1 p_{Y,1} + \frac{p_{Y,1}}{1-p_{Y,1}} t^\star_{2,EO,\alpha}}{2p_1 p_{Y,1} + \frac{p_{Y,1}}{1-p_{Y,1}} t^\star_{2,EO,\alpha} - t^\star_{1,EO,\alpha}}\right) = 0$ *and thus* $\tau^\star_{EO,\alpha} \in [0,1]$ *can be an arbitrary constant.*

To prove Theorem 6, we first introduce the Neyman-Pearson Lemma.

Lemma 8. *(Generalized Neyman-Pearson lemma). Let* $f_0, f_1, \ldots, f_m$ *be* $m+1$ *real-valued functions defined on a Euclidean space* $\mathcal{X}$*. Assume they are* ν *-integrable for a* σ *-finite measure* ν*. Let* ϕ_0 *be any function of the form*

$$\phi_0(x) = \begin{cases} 1, & f_0(x) > \sum_{i=1}^m c_i f_i(x) \\ \gamma(x) & f_0(x) = \sum_{i=1}^m c_i f_i(x) \\ 0, & f_0(x) < \sum_{i=1}^m c_i f_i(x) \end{cases}$$

where $0 \le \gamma(x) \le 1$ *for all* $x \in \mathcal{X}$*. For given constants* $t_1, \ldots, t_m \in \mathbb{R}$*, let* $\mathcal{T}$ *be the class of Borel functions* $\phi : \mathcal{X} \mapsto \mathbb{R}$ *satisfying*

$$\int_{\mathcal{X}} \phi f_i d\nu \le t_i, \quad i = 1, 2, \ldots, m \tag{5}$$

and $\mathcal{T}_0$ *be the set of* ϕ *s in* $\mathcal{T}$ *satisfying (5) with all inequalities replaced by equalities. If* $\phi_0 \in \mathcal{T}_0$*, then* $\phi_0 \in \operatorname*{argmax}_{\phi \in \mathcal{T}_0} \int_{\mathcal{X}} \phi f_0 d\nu$*. Moreover, if* $c_i \ge 0$ *for all* $i = 1, \ldots, m$*, then* $\phi_0 \in \operatorname*{argmax}_{\phi \in \mathcal{T}} \int_{\mathcal{X}} \phi f_0 d\nu$*.*

Then we come to prove the theorem.

Proof. If $|EO^\star| \preceq (\alpha, \alpha)$, we are done since $f^\star$ is just our target classifier. Now, we assume $|EO^\star| \preceq (\alpha, \alpha)$ does not hold. Let f be a classifier that gives output $\widehat{Y} = 1$ with probability $f(x,a)$ under $X = x$ and $A = a$. The mis-classification error for f is

$$\begin{aligned} R(f) &= \mathbb{P}(\widehat{Y} \neq Y) = 1 - \mathbb{P}(\widehat{Y} = 1, Y = 1) - \mathbb{P}(\widehat{Y} = 0, Y = 0) \\ &= \mathbb{P}(\widehat{Y} = 1, Y = 0) - \mathbb{P}(\widehat{Y} = 1, Y = 1) + \mathbb{P}(Y = 1) \end{aligned}$$

Thus, to minimize the mis-classification error is just equivalent to maximize $\mathbb{P}(\widehat{Y} = 1, Y = 0) - \mathbb{P}(\widehat{Y} = 1, Y = 1)$, which can be expressed as:

$$\begin{aligned} &\mathbb{P}(\widehat{Y} = 1, Y = 1) - \mathbb{P}(\widehat{Y} = 1, Y = 0) \\ =&\mathbb{P}_{X|A=1,Y=1}(\widehat{Y} = 1) p_1 p_{Y,1} + \mathbb{P}_{X|A=0,Y=1}(\widehat{Y} = 1)\,(1 - p_1)\, p_{Y,0} \\ &- \mathbb{P}_{X|A=1,Y=0}(\widehat{Y} = 1) p_1\,(1 - p_{Y,1}) - \mathbb{P}_{X|A=0,Y=0}(\widehat{Y} = 1)\,(1 - p_1)\,(1 - p_{Y,0}) \\ =&p_1 \left[p_{Y,1} \int_{\mathcal{X}} f(x,1) d\mathbb{P}_{X|1,1}(x) - (1 - p_{Y,1}) \int_{\mathcal{X}} f(x,1) d\mathbb{P}_{X|A=1,Y=0}(x)\right] \\ &+ (1 - p_1) \left[p_{Y,0} \int_{\mathcal{X}} f(x,0) d\mathbb{P}_{X|A=0,Y=1}(x) - (1 - p_{Y,0}) \int_{\mathcal{X}} f(x,0) d\mathbb{P}_{X|A=0,Y=0}(x)\right] \\ =&\int_{\mathcal{A}} \int_{\mathcal{X}} f(x,a) M(x,a) d\mathbb{P}_X(x) d\mathbb{P}(a) \end{aligned}$$

with

$$\begin{aligned} M(x,a) = ap_1 \left[p_{Y,1} \frac{d\mathbb{P}_{X|A=1,Y=1}(x)}{d\mathbb{P}_X(x)} - (1-p_{Y,1}) \frac{d\mathbb{P}_{X|A=1,Y=0}(x)}{d\mathbb{P}_X(x)} \right] \\ +(1-a)p_0 \left[p_{Y,0} \frac{d\mathbb{P}_{X|A=0,Y=1}(x)}{d\mathbb{P}_X(x)} - (1-p_{Y,0}) \frac{d\mathbb{P}_{X|A=0,Y=0}(x)}{d\mathbb{P}_X(x)} \right]. \end{aligned} \tag{6}$$

Next, for any classifier f, we have,

$$\begin{aligned} DEO(f) &= (\mathbb{P}_{X|A=1,Y=1}(\widehat{Y}=1) - \mathbb{P}_{X|A=0,Y=1}(\widehat{Y}=1), \mathbb{P}_{X|A=1,Y=0}(\widehat{Y}=1) - \mathbb{P}_{X|A=0,Y=0}(\widehat{Y}=1)) \\ &= (\int_{\mathcal{X}} f(x,1) d\mathbb{P}_{X|A=1,Y=1}(x) - \int_{\mathcal{X}} f(x,0) d\mathbb{P}_{X|A=0,Y=1}(x), \\ &\quad \int_{\mathcal{X}} f(x,1) d\mathbb{P}_{X|A=1,Y=0}(x) - \int_{\mathcal{X}} f(x,0) d\mathbb{P}_{X|A=0,Y=0}(x)) \\ &= (\int_{\mathcal{A}} \int_{\mathcal{X}} f(x,a) H_E(x,a) d\mathbb{P}_X(x) d\mathbb{P}(a), \int_{\mathcal{A}} \int_{\mathcal{X}} f(x,a) H_P(x,a) d\mathbb{P}_X(x) d\mathbb{P}(a)) \end{aligned}$$

with

$$\begin{cases} H_E(x,a) = \dfrac{a d\mathbb{P}_{X|A=1,Y=1}(x)}{p_1 d\mathbb{P}_X(x)} - \dfrac{(1-a) d\mathbb{P}_{X|A=0,Y=1}(x)}{p_0 d\mathbb{P}_X(x)} \\ H_P(x,a) = \dfrac{a d\mathbb{P}_{X|A=1,Y=0}(x)}{p_1 d\mathbb{P}_X(x)} - \dfrac{(1-a) d\mathbb{P}_{X|A=0,Y=0}(x)}{p_0 d\mathbb{P}_X(x)}. \end{cases} \tag{7}$$

Since $\lim_{t_2 \to \infty} \frac{p_1^2 p_{Y,1} + \frac{p_{Y,1}}{1-p_{Y,1}} t_2}{2p_1^2 p_{Y,1} + \frac{p_{Y,1}}{1-p_{Y,1}} t_2 - t_1} = \lim_{t_2 \to \infty} \frac{p_0^2 p_{Y,0} - \frac{p_{Y,0}}{1-p_{Y,0}} t_2}{2p_0^2 p_{Y,0} - \frac{p_{Y,0}}{1-p_{Y,0}} t_2 + t_1} = 1$, we have:

$\lim_{t_2 \to \infty} \mathbb{P}_{X|A=1,Y=1}(\eta_1(x) > \frac{p_1^2 p_{Y,1} + \frac{p_{Y,1}}{1-p_{Y,1}} t_2}{2p_1^2 p_{Y,1} + \frac{p_{Y,1}}{1-p_{Y,1}} t_2 - t_1}) = \lim_{t_2 \to \infty} \mathbb{P}_{X|A=0,Y=1}(\eta_0(x) > \frac{p_0^2 p_{Y,0} - \frac{p_{Y,0}}{1-p_{Y,0}} t_2}{2p_0^2 p_{Y,0} - \frac{p_{Y,0}}{1-p_{Y,0}} t_2 + t_1})$

$= \lim_{t_2 \to \infty} \mathbb{P}_{X|A=1,Y=0}(\eta_1(x) > \frac{p_1^2 p_{Y,1} + \frac{p_{Y,1}}{1-p_{Y,1}} t_2}{2p_1^2 p_{Y,1} + \frac{p_{Y,1}}{1-p_{Y,1}} t_2 - t_1}) = \lim_{t_2 \to \infty} \mathbb{P}_{X|A=0,Y=0}(\eta_0(x) > \frac{p_0^2 p_{Y,0} - \frac{p_{Y,0}}{1-p_{Y,0}} t_2}{2p_0^2 p_{Y,0} - \frac{p_{Y,0}}{1-p_{Y,0}} t_2 + t_1})$

$= 0.$

So there exist $t^\star_{1,EO,\alpha}$ and $t^\star_{2,EO,\alpha}$, such that:

$t^\star_{1,EO,\alpha} \frac{E^*}{|E^*|} > 0$, $t^\star_{2,EO,\alpha} \frac{P^*}{|P^*|} > 0$, and

$$\begin{cases} \dfrac{E^\star}{|E^\star|} [\mathbb{P}_{X|A=1,Y=1}(\eta_1(x) > \dfrac{p_1^2 p_{Y,1} + \frac{p_{Y,1}}{1-p_{Y,1}} t^\star_{2,EO,\alpha}}{2p_1^2 p_{Y,1} + \frac{p_{Y,1}}{1-p_{Y,1}} t^\star_{2,EO,\alpha} - t^\star_{1,EO,\alpha}}) + \tau \mathbb{P}_{X|A=1,Y=1}(\eta_1(x) = \dfrac{p_1^2 p_{Y,1} + \frac{p_{Y,1}}{1-p_{Y,1}} t^\star_{2,EO,\alpha}}{2p_1^2 p_{Y,1} + \frac{p_{Y,1}}{1-p_{Y,1}} t^\star_{2,EO,\alpha} - t^\star_{1,EO,\alpha}}) \\ \qquad\qquad - \mathbb{P}_{X|A=0,Y=1}(\eta_0(x) > \dfrac{p_0^2 p_{Y,0} - \frac{p_{Y,0}}{1-p_{Y,0}} t^\star_{2,EO,\alpha}}{2p_0^2 p_{Y,0} - \frac{p_{Y,0}}{1-p_{Y,0}} t^\star_{2,EO,\alpha} + t^\star_{1,EO,\alpha}})] = \alpha_1 < \alpha \\ \dfrac{P^\star}{|P^\star|} [\mathbb{P}_{X|A=1,Y=0}(\eta_1(x) > \dfrac{p_1^2 p_{Y,1} + \frac{p_{Y,1}}{1-p_{Y,1}} t^\star_{2,EO,\alpha}}{2p_1^2 p_{Y,1} + \frac{p_{Y,1}}{1-p_{Y,1}} t^\star_{2,EO,\alpha} - t^\star_{1,EO,\alpha}}) + \tau \mathbb{P}_{X|A=1,Y=0}(\eta_1(x) = \dfrac{p_1^2 p_{Y,1} + \frac{p_{Y,1}}{1-p_{Y,1}} t^\star_{2,EO,\alpha}}{2p_1^2 p_{Y,1} + \frac{p_{Y,1}}{1-p_{Y,1}} t^\star_{2,EO,\alpha} - t^\star_{1,EO,\alpha}}) \\ \qquad\qquad - \mathbb{P}_{X|A=0,Y=0}(\eta_0(x) > \dfrac{p_0^2 p_{Y,0} - \frac{p_{Y,0}}{1-p_{Y,0}} t^\star_{2,EO,\alpha}}{2p_0^2 p_{Y,0} - \frac{p_{Y,0}}{1-p_{Y,0}} t^\star_{2,EO,\alpha} + t^\star_{1,EO,\alpha}})] = \alpha_2 < \alpha \end{cases}$$

We consider the constraint,

$$\begin{cases} \dfrac{E^\star}{|E^\star|} \displaystyle\int_{\mathcal{A}} \int_{\mathcal{X}} f(x,a) H_E(x,a) d\mathbb{P}_X(x) d\mathbb{P}(a) \leq \alpha_1 \\ \dfrac{F^\star}{|F^\star|} \displaystyle\int_{\mathcal{A}} \int_{\mathcal{X}} f(x,a) H_P(x,a) d\mathbb{P}_X(x) d\mathbb{P}(a) \leq \alpha_2. \end{cases} \tag{8}$$

Let f be the classifier of the form:

$$f_{s_1,s_2,\tau}(x,a) = \begin{cases} 1, & M(x,a) > s_1 \frac{E^\star}{|E^\star|} H_E(x,a) + s_2 \frac{P^\star}{|P^\star|} H_P(x,a); \\ a\tau, & M(x,a) = s_1 \frac{E^\star}{|E^\star|} H_E(x,a) + s_2 \frac{P^\star}{|P^\star|} H_P(x,a); \\ 0, & M(x,a) < s_1 \frac{E^\star}{|E^\star|} H_E(x,a) + s_2 \frac{P^\star}{|P^\star|} H_P(x,a), \end{cases} \tag{9}$$

From (6) & (7), $M(x,a) > s_1 \frac{E^\star}{|E^\star|} H_E(x,a) + s_2 \frac{P^\star}{|P^\star|} H_P(x,a)$ is equal to

$$\begin{aligned}
& a p_1 \left[p_{Y,1} \frac{d\mathbb{P}_{X|A=1,Y=1}(x)}{d\mathbb{P}_X(x)} - (1-p_{Y,1}) \frac{d\mathbb{P}_{X|A=1,Y=0}(x)}{d\mathbb{P}_X(x)} \right] \\
& + (1-a) p_0 \left[p_{Y,0} \frac{d\mathbb{P}_{X|A=0,Y=1}(x)}{d\mathbb{P}_X(x)} - (1-p_{Y,0}) \frac{d\mathbb{P}_{X|A=0,Y=0}(x)}{d\mathbb{P}_X(x)} \right] \\
> & \frac{E^\star}{|E^\star|} s_1 \left(\frac{a d\mathbb{P}_{X|A=1,Y=1}(x)}{p_1 d\mathbb{P}_X(x)} - \frac{(1-a) d\mathbb{P}_{X|A=0,Y=1}(x)}{p_0 d\mathbb{P}_X(x)} \right) \\
& + \frac{P^\star}{|P^\star|} s_2 \left(\frac{a d\mathbb{P}_{X|A=1,Y=0}(x)}{p_1 d\mathbb{P}_X(x)} - \frac{(1-a) d\mathbb{P}_{X|A=0,Y=0}(x)}{p_0 d\mathbb{P}_X(x)} \right)
\end{aligned}$$

which is equal to

$$\begin{cases}
p_1 \left[p_{Y,1} \frac{d\mathbb{P}_{X|A=1,Y=1}(x)}{d\mathbb{P}_X(x)} - (1-p_{Y,1}) \frac{d\mathbb{P}_{X|A=1,Y=0}(x)}{d\mathbb{P}_X(x)} \right] > \frac{\frac{E^\star}{|E^\star|} s_1 d\mathbb{P}_{X|A=1,Y=1}(x) + \frac{P^\star}{|P^\star|} s_2 d\mathbb{P}_{X|A=1,Y=0}(x)}{p_1 d\mathbb{P}_X(x)} & , a = 1 \\
p_0 \left[p_{Y,0} \frac{d\mathbb{P}_{X|A=0,Y=1}(x)}{d\mathbb{P}_X(x)} - (1-p_{Y,0}) \frac{d\mathbb{P}_{X|A=0,Y=0}(x)}{d\mathbb{P}_X(x)} \right] > -\frac{\frac{E^\star}{|E^\star|} s_1 d\mathbb{P}_{X|A=0,Y=1}(x) + \frac{P^\star}{|P^\star|} s_2 d\mathbb{P}_{X|A=0,Y=0}(x)}{p_0 d\mathbb{P}_X(x)} & , a = 0
\end{cases}$$

Thus,

$$\begin{aligned}
& M(x,a) > s_1 \frac{E^\star}{|E^\star|} H_E(x,a) + s_2 \frac{P^\star}{|P^\star|} H_P(x,a) \\
\iff & p_a \left[p_{Y,a} \frac{d\mathbb{P}_{X|A=a,Y=1}(x)}{d\mathbb{P}_X(x)} - (1-p_{Y,a}) \frac{d\mathbb{P}_{X|A=a,Y=0}(x)}{d\mathbb{P}_X(x)} \right] > (2a-1) \frac{\frac{E^\star}{|E^\star|} s_1 d\mathbb{P}_{X|A=a,Y=1}(x) + \frac{P^\star}{|P^\star|} s_2 d\mathbb{P}_{X|A=a,Y=0}(x)}{p_a d\mathbb{P}_X(x)}. \\
\iff & \frac{p_{Y,a} d\mathbb{P}_{X|A=a,Y=1}(x)}{p_{Y,a} d\mathbb{P}_{X|A=a,Y=1}(x) + (1-p_{Y,a}) d\mathbb{P}_{X|A=a,Y=0}(x)} > \frac{p_a^2 p_{Y,a} + (2a-1) \frac{p_{Y,a}}{1-p_{Y,a}} t_2}{2p_a^2 p_{Y,a} + (2a-1)(\frac{p_{Y,a}}{1-p_{Y,a}} t_2 - t_1)}. \\
& \text{where } t_1 = 2 \frac{E^\star}{|E^\star|} s_1, \quad t_2 = 2 \frac{P^\star}{|P^\star|} s_2. \\
\iff & \eta_a(x) > \frac{p_a^2 p_{Y,a} + (2a-1) \frac{p_{Y,a}}{1-p_{Y,a}} t_2}{2p_a^2 p_{Y,a} + (2a-1)(\frac{p_{Y,a}}{1-p_{Y,a}} t_2 - t_1)}.
\end{aligned}$$

As a result, $f_{s_1,s_2,\tau}(x,a)$ in (9) can be written as

$$f_{t_1,t_2,\tau}(x,a) = \mathbb{1}\{\eta_a(x) > \frac{p_a^2 p_{Y,a} + (2a-1) \frac{p_{Y,a}}{1-p_{Y,a}} t_2}{2p_a^2 p_{Y,a} + (2a-1)(\frac{p_{Y,a}}{1-p_{Y,a}} t_2 - t_1)}\} + a\tau \mathbb{1}\{\eta_a(x) = \frac{p_a^2 p_{Y,a} + (2a-1) \frac{p_{Y,a}}{1-p_{Y,a}} t_2}{2p_a^2 p_{Y,a} + (2a-1)(\frac{p_{Y,a}}{1-p_{Y,a}} t_2 - t_1)}\}. \tag{10}$$

Further, the constraint (8) for f in (10) is equivalent to

$$\begin{cases}
\frac{E^\star}{|E^\star|} [\mathbb{P}_{X|A=1,Y=1}(\eta_1(x) > \frac{p_1^2 p_{Y,1} + \frac{p_{Y,1}}{1-p_{Y,1}} t_2}{2p_1^2 p_{Y,1} + \frac{p_{Y,1}}{1-p_{Y,1}} t_2 - t_1}) + \tau \mathbb{P}_{X|A=1,Y=1}(\eta_1(x) = \frac{p_1^2 p_{Y,1} + \frac{p_{Y,1}}{1-p_{Y,1}} t_2}{2p_1^2 p_{Y,1} + \frac{p_{Y,1}}{1-p_{Y,1}} t_2 - t_1}) \\
\qquad - \mathbb{P}_{X|A=0,Y=1}(\eta_0(x) > \frac{p_0^2 p_{Y,0} - \frac{p_{Y,0}}{1-p_{Y,0}} t_2}{2p_0^2 p_{Y,0} - \frac{p_{Y,0}}{1-p_{Y,0}} t_2 + t_1})] \le \alpha_1 \\
\frac{P^\star}{|P^\star|} [\mathbb{P}_{X|A=1,Y=0}(\eta_1(x) > \frac{p_1^2 p_{Y,1} + \frac{p_{Y,1}}{1-p_{Y,1}} t_2}{2p_1^2 p_{Y,1} + \frac{p_{Y,1}}{1-p_{Y,1}} t_2 - t_1}) + \tau \mathbb{P}_{X|A=1,Y=0}(\eta_1(x) = \frac{p_1^2 p_{Y,1} + \frac{p_{Y,1}}{1-p_{Y,1}} t_2}{2p_1^2 p_{Y,1} + \frac{p_{Y,1}}{1-p_{Y,1}} t_2 - t_1}) \\
\qquad - \mathbb{P}_{X|A=0,Y=0}(\eta_0(x) > \frac{p_0^2 p_{Y,0} - \frac{p_{Y,0}}{1-p_{Y,0}} t_2}{2p_0^2 p_{Y,0} - \frac{p_{Y,0}}{1-p_{Y,0}} t_2 + t_1})] \le \alpha_2
\end{cases} \tag{11}$$

Now, let $\mathcal{T}_{\alpha_1,\alpha_2}$ be the class of Borel functions f that satisfy (8) and $\mathcal{T}_{\alpha_1,\alpha_2,0}$ be the set of f-s in $\mathcal{T}_\alpha$ that satisfy (8) with all the inequalities being replaced by equalities.

From the definition of $t^\star_{1,EO,\alpha}$ and $t^\star_{2,EO,\alpha}$, clearly $f_{t^\star_{1,EO,\alpha}, t^\star_{2,EO,\alpha}, \tau}(x,a) \in \mathcal{T}_{\alpha_1,\alpha_2,0}$. Further, we have $s_1^* = t^\star_{1,EO,\alpha} \frac{E^*}{2|E^*|} > 0$ and $s_2^* = t^\star_{2,EO,\alpha} \frac{P^*}{2|P^*|} > 0$.

Hence, from Generalized Neyman-Pearson lemma, we have:

$$f_{t^\star_{1,EO,\alpha}, t^\star_{2,EO,\alpha}, \tau}(x,a) \in \underset{f \in \mathcal{T}_{\alpha_1,\alpha_2}}{argmax} \int_{\mathcal{A}} \int_{\mathcal{X}} f_{t_1,t_2,\tau}(x,a) M(x,a) d\mathbb{P}_X(x) d\mathbb{P}(a)$$

Now we complete our proof. □

A.7 PROOF OF PROPOSITION 3

Proof. The classifier is

$$\phi = \begin{cases} \mathbb{1}\{f(x,0) > t^{1,0}_{(k^{1,0})}\}, a = 0 \\ \mathbb{1}\{f(x,1) > t^{1,1}_{(k^{1,1})}\}, a = 1 \end{cases}$$

we have:

$$|DEO(\phi)| = (|F^{1,1}(t^{1,1}_{(k^{1,1})}) - F^{1,0}(t^{1,0}_{(k^{1,0})})|, |F^{0,1}(t^{1,1}_{(k^{1,1})}) - F^{0,0}(t^{1,0}_{(k^{1,0})})|)$$

From Proposition 1, we have $\mathbb{P}(|F^{1,1}(t^{1,1}_{(k^{1,1})}) - F^{1,0}(t^{1,0}_{(k^{1,0})})| > \alpha) \leq \mathbb{E}[\sum_{j=k^{1,0}}^{n^{1,0}} \binom{n^{1,0}}{j} (Q^{1,1} - \alpha)^j (1 - (Q^{1,1} - \alpha))^{n^{1,0}-j}] + \mathbb{E}[\sum_{j=k^{1,1}}^{n^{1,1}} \binom{n^{1,1}}{j} (Q^{1,0} - \alpha)^j (1 - (Q^{1,0} - \alpha))^{n^{1,1}-j}]$

Also,

$$\begin{aligned} &\mathbb{P}(|F^{0,1}(t^{1,1}_{(k^{1,1})}) - F^{0,0}(t^{1,0}_{(k^{1,0})})| > \alpha) \\ =&\mathbb{P}(F^{0,1}(t^{1,1}_{(k^{1,1})}) - F^{0,0}(t^{1,0}_{(k^{1,0})}) > \alpha) + \mathbb{P}(F^{0,1}(t^{1,1}_{(k^{1,1})}) - F^{0,0}(t^{1,0}_{(k^{1,0})}) < -\alpha) \\ \triangleq& A + B \end{aligned}$$

And we have

$$\begin{aligned} A &= \mathbb{P}(F^{0,1}(t^{1,1}_{(k^{1,1})}) - F^{0,0}(t^{1,0}_{(k^{1,0})}) > \alpha) \\ &= \mathbb{P}(F^{0,0}(t^{1,0}_{(k^{1,0})}) < F^{0,1}(t^{1,1}_{(k^{1,1})}) - \alpha) \\ &\leq \mathbb{P}(F^{0,0}(t^{0,0}_{(k^{0,0})}) < F^{0,1}(t^{0,1}_{(k^{0,1}+1)}) - \alpha) \\ &= \mathbb{E}[\mathbb{P}(t^{0,0}_{(k^{0,0})} < F^{0,0^{-1}}(F^{0,1}(t^{0,1}_{(k^{0,1}+1)}) - \alpha))\mathbb{1}\{F^{0,1}(t^{0,1}_{(k^{0,1}+1)}) - \alpha > 0\} \mid t^{0,1}_{(k^{0,1}+1)}] \\ &= E\{\mathbb{P}[\text{at least } k^{0,0} \text{ of } t^{0,0}\text{'s are less than } F^{0,0^{-1}}(F^{0,1}(t^{0,1}_{(k^{0,1}+1)}) - \alpha)]\mathbb{1}\{F^{0,1}(t^{0,1}_{(k^{0,1}+1)}) - \alpha > 0\} \mid t^{0,1}_{(k^{0,1}+1)}\} \\ &= E\{\sum_{j=k^{0,0}}^{n^{0,0}} \mathbb{P}[\text{exactly j of the } t^{0,0}\text{'s are less than } F^{0,0^{-1}}(F^{0,1}(t^{0,1}_{(k^{0,1}+1)}) - \alpha)]\mathbb{1}\{F^{0,1}(t^{0,1}_{(k^{0,1}+1)}) - \alpha > 0\} \mid t^{0,1}_{(k^{0,1}+1)}\} \\ &= E\{\sum_{j=k^{0,0}}^{n^{0,0}} \binom{n^{0,0}}{j} \mathbb{P}[t^{0,0} < F^{0,0^{-1}}(F^{0,1}(t^{0,1}_{(k^{0,1}+1)}) - \alpha)]^j (1 - \mathbb{P}[t^{0,0} < F^{0,0^{-1}}(F^{0,1}(t^{0,1}_{(k^{0,1}+1)}) - \alpha)])^{n^{0,0}-j} \\ &\quad \mathbb{1}\{F^{0,1}(t^{0,1}_{(k^{0,1}+1)}) - \alpha > 0\} \mid t^{0,1}_{(k^{0,1}+1)}\} \\ &\leq \mathbb{E}[\sum_{j=k^{0,0}}^{n^{0,0}} \binom{n^{0,0}}{j} (F^{0,1}(t^{0,1}_{(k^{0,1}+1)}) - \alpha)^j (1 - (F^{0,1}(t^{0,1}_{(k^{0,1}+1)}) - \alpha))^{n^{0,0}-j} \mid t^{0,1}_{(k^{0,1}+1)}] \end{aligned}$$

Similarly, we have

$$B \leq \mathbb{E}[\sum_{j=k^{0,1}}^{n^{0,1}} \binom{n^{0,1}}{j} (F^{0,0}(t^{0,0}_{(k^{0,0}+1)}) - \alpha)^j (1 - (F^{0,0}(t^{0,0}_{(k^{0,0}+1)}) - \alpha))^{n^{0,1}-j} \mid t^{0,0}_{(k^{0,0}+1)}]$$

Hence, we have

$A + B \leq \mathbb{E}[\sum_{j=k^{0,0}}^{n^{0,0}} \binom{n^{0,0}}{j} (F^{0,1}(t^{0,1}_{(k^{0,1}+1)}) - \alpha)^j (1 - (F^{0,1}(t^{0,1}_{(k^{0,1}+1)}) - \alpha))^{n^{0,0}-j} \mid t^{0,1}_{(k^{0,1}+1)}] + \mathbb{E}[\sum_{j=k^{0,1}}^{n^{0,1}} \binom{n^{0,1}}{j} (F^{0,0}(t^{0,0}_{(k^{0,0}+1)}) - \alpha)^j (1 - (F^{0,0}(t^{0,0}_{(k^{0,0}+1)}) - \alpha))^{n^{0,1}-j} \mid t^{0,0}_{(k^{0,0}+1)}]$.

From Lemma 2, we have $F^{0,0}(t^{0,0}_{(k^{0,0}+1)}) \sim Beta(k^{0,0}+1, n^{0,0}-k^{0,0})$ and $F^{0,1}(t^{0,1}_{(k^{0,1}+1)}) \sim Beta(k^{0,1}+1, n^{0,1}-k^{0,1})$, and we complete the proof. □

A.8 Algorithms for other group fairness constraints

In addition to Equality of Opportunity and Equalized Odds, there are other common fairness constraints and we can extend FaiREE to them.

Definition 4 (Demographic Parity)**.** *A classifier satisfies Demographic Parity if its prediction $\widehat{Y}$ is statistically independent of the sensitive attribute A :*

$$\mathbb{P}(\widehat{Y}=1 \mid A=1) = \mathbb{P}(\widehat{Y}=1 \mid A=0)$$

Definition 5 (Predictive Equality)**.** *A classifier satisfies Predictive Equality if it achieves the same TNR (or FPR) among protected groups:*

$$\mathbb{P}_{X|A=1,Y=0}(\widehat{Y}=1) = \mathbb{P}_{X|A=0,Y=0}(\widehat{Y}=1)$$

Definition 6 (Equalized Accuracy)**.** *A classifier satisfies Equalized Accuracy if its mis-classification error is statistically independent of the sensitive attribute A:*

$$\mathbb{P}(\widehat{Y} \neq Y \mid A=1) = \mathbb{P}(\widehat{Y} \neq Y \mid A=0)$$

Similar to $DEOO$, we can define the following measures:

$$\text{DDP} = \mathbb{P}_{X|A=1}(\widehat{Y}=1) - \mathbb{P}_{X|A=0}(\widehat{Y}=1) \tag{12}$$

$$\text{DPE} = \mathbb{P}_{X|A=1,Y=0}(\widehat{Y}=1) - \mathbb{P}_{X|A=0,Y=0}(\widehat{Y}=1) \tag{13}$$

$$\text{DEA} = \mathbb{P}(\widehat{Y} \neq Y \mid A=1) - \mathbb{P}(\widehat{Y} \neq Y \mid A=0). \tag{14}$$

A.8.1 FaiREE for Demographic Parity

Algorithm 3: FaiREE for Demographic Parity

Input:
Training data: $S = S^{0,0} \cup S^{0,1} \cup S^{1,0} \cup S^{1,1}$
α: error bound
δ: small tolerance level
f: a classifier

$T^{y,a} = \{f(x_1^{y,a}), \ldots, f(x_{n_{y,a}}^{y,a})\}$

$\{t^{y,a}_{(1)}, \ldots, t^{y,a}_{(n_{y,a})}\}$ =sort($T^{y,a}$)

$T^y = T^{y,0} \cup T^{y,1}$

$\{t^{y}_{(1)}, \ldots, t^{y}_{(n_y)}\}$ =sort(T^{y})

Define $g(k,a) = \mathbb{E}[\sum_{j=k}^{n^a} \binom{n^a}{j} (Q^{1-a}-\alpha)^j (1-(Q^{1-a}-\alpha))^{n^a-j}]$ with $Q^a \sim Beta(k, n^a-k+1)$, $L(k^0,k^1) = g(k^0,0) + g(k^1,1)$

Build candidate set $K = \{(k^0,k^1) \mid L(k^0,k^1) \leq \delta\} = \{(k_1^0,k_1^1), \ldots, (k_M^0,k_M^1)\}$

Find $k_i^{y,0}$: $t^{y,0}_{(k_i^{y,0})} \leq t^0_{(k_i^0)} < t^{y,0}_{(k_i^{y,0}+1)}$, $t^{y,1}_{(k_i^{y,1})} \leq t^1_{(k_i^1)} < t^{y,1}_{(k_i^{y,1}+1)}$, $y \in \{0,1\}$

$i_* \leftarrow \arg\min_{i\in[M]}\{\hat{e}_i\}$ ($\hat{e}_i$ is defined in Proposition 2)

Output: $\hat{\phi}(x,a) = \mathbb{1}\{f(x,a) > t^a_{(k^a_{i_*})}\}$

Similar to the algorithm for Equality of Opportunity, we have the following propositions and assumption:

Proposition 4. *Given k^0, k^1 satisfying $k^a \in \{1, \dots, n^a\}$ $(a = 0, 1)$. Define $\phi(x, a) = \mathbb{1}\{f(x, a) > t^a_{(k^a)}\}$, $g(k, a) = \mathbb{E}[\sum_{j=k}^{n^a} \binom{n^a}{j} (Q^{1-a} - \alpha)^j (1 - (Q^{1-a} - \alpha))^{n^a - j}]$ with $Q^a \sim Beta(k, n^a - k + 1)$, then we have:*

$$\mathbb{P}(|DDP(\phi)| > \alpha) \leq g(k^0, 0) + g(k^1, 1).$$

If t^a is continuous random variable, the equality holds.

Theorem 7. *If $\min\{n^0, n^1\} \geq \lceil \frac{\log \frac{\delta}{2}}{\log(1-\alpha)} \rceil$, we have $|DDP(\hat{\phi}_i)| < \alpha$ with probability $1 - \delta$, for each $i \in \{1, \dots, M\}$.*

Theorem 8. *Given $\alpha' < \alpha$. If $\min\{n^0, n^1\} \geq \lceil \frac{\log \frac{\delta}{2}}{\log(1-\alpha)} \rceil$. For any $\epsilon > 0$ such that $g(\epsilon) \leq \frac{\alpha - \alpha'}{2}$. Suppose $\hat{\phi}$ is the final output of FaiREE, we have:*
(1) $|DDP(\hat{\phi})| \leq \alpha$ with probability $(1 - \delta)^M$, where M is the size of the candidate set.
(2) Suppose the density distribution functions of f^* under $A = a, Y = 1$ are continuous. $\phi^*_{DDP,\alpha} = \arg\min_{|DDP(\phi)| \leq \alpha} \mathbb{P}(\phi(X, A) \neq Y)$. When the input classifier f is f^*, we have

$$|\ \mathbb{P}(\hat{\phi} \neq Y) - \mathbb{P}(\phi^*_{\alpha'} \neq Y)\ | \leq 2g(\epsilon) + 2h^3(\epsilon) + 12h^2(\epsilon) + 16h(\epsilon)$$

with probability $1 - (2M + 4)(e^{-2n^{1,0}h^2(\epsilon)} + e^{-2n^{1,1}h^2(\epsilon)} + e^{-2n^{0,0}h^2(\epsilon)} + e^{-2n^{0,1}h^2(\epsilon)})$.

A.8.2 FAIREE FOR PREDICTIVE EQUALITY

Algorithm 4: FaiREE for Predictive Opportunity

Input:
Training data: $S = S^{0,0} \cup S^{0,1} \cup S^{1,0} \cup S^{1,1}$
α: error bound
δ: small tolerance level
f: a classifier

1 $T^{y,a} = \{f(x^{y,a}_1), \dots, f(x^{y,a}_{n_{y,a}})\}$
2 $\{t^{y,a}_{(1)}, \dots, t^{y,a}_{(n_{y,a})}\}$ =sort($T^{y,a}$)
3 Define $g_0(k, a) = \mathbb{E}[\sum_{j=k}^{n^{0,a}} \binom{n^{0,a}}{j} (Q^{0,1-a} - \alpha)^j (1 - (Q^{0,1-a} - \alpha))^{n^{0,a} - j}]$ with $Q^{0,a} \sim Beta(k, n^{0,a} - k + 1)$, $L(k^{0,0}, k^{0,1}) = g_0(k^{0,0}, 0) + g_0(k^{0,1}, 1)$
4 Build candidate set $K = \{(k^{0,0}, k^{0,1}) \mid L(k^{0,0}, k^{0,1}) \leq \delta\} = \{(k^{0,0}_1, k^{0,1}_1), \dots, (k^{0,0}_M, k^{0,1}_M)\}$
5 Find $k^{1,0}_i, k^{1,1}_i$: $t^{1,0}_{(k^{1,0}_i)} \leq t^{0,0}_{(k^{0,0}_i)} < t^{1,0}_{(k^{1,0}_i + 1)}$, $t^{1,1}_{(k^{1,1}_i)} \leq t^{0,1}_{(k^{0,1}_i)} < t^{1,1}_{(k^{1,1}_i + 1)}$
6 $i_* \leftarrow \arg\min_{i \in [M]} \{\hat{e}_i\}$ ($\hat{e}_i$ is defined in Proposition 2)

Output: $\hat{\phi}(x, a) = \mathbb{1}\{f(x, a) > t^{0,a}_{(k^{0,a}_{i_*})}\}$

Similar to the algorithm for Equality of Opportunity, we have the following propositions and assumption:

Proposition 5. *Given $k^{0,0}, k^{0,1}$ satisfying $k^{1,a} \in \{1, \dots, n^{0,a}\}$ $(a = 0, 1)$. Define $\phi(x, a) = \mathbb{1}\{f(x, a) > t^{0,a}_{(k^{0,a})}\}$, $g_0(k, a) = \mathbb{E}[\sum_{j=k}^{n^{0,a}} \binom{n^{0,a}}{j} (Q^{0,1-a} - \alpha)^j (1 - (Q^{0,1-a} - \alpha))^{n^{0,a} - j}]$ with $Q^{0,a} \sim Beta(k, n^{0,a} - k + 1)$, then we have:*

$$\mathbb{P}(|DPE(\phi)| > \alpha) \leq g_0(k^{0,0}, 0) + g_0(k^{0,1}, 1).$$

If $t^{0,a}$ is continuous random variable, the equality holds.

Theorem 9. *If $\min\{n^{0,0}, n^{0,1}\} \geq \lceil \frac{\log \frac{\delta}{2}}{\log(1-\alpha)} \rceil$, we have $|DPE(\hat{\phi}_i)| < \alpha$ with probability $1 - \delta$, for each $i \in \{1, \dots, M\}$.*

Theorem 10. *Given $\alpha' < \alpha$. If $\min\{n^{0,0}, n^{0,1}\} \geq \lceil \frac{\log \frac{\delta}{2}}{\log(1-\alpha)} \rceil$. For any $\epsilon > 0$ such that $g(\epsilon) \leq \frac{\alpha-\alpha'}{2}$. Suppose $\hat{\phi}$ is the final output of FaiREE, we have:*
(1) $|DPE(\hat{\phi})| \leq \alpha$ with probability $(1-\delta)^M$, where M is the size of the candidate set.
(2) Suppose the density distribution functions of f^* under $A = a, Y = 1$ are continuous. $\phi^*_{DPE,\alpha} = \arg\min_{|DPE(\phi)| \leq \alpha} \mathbb{P}(\phi(X, A) \neq Y)$. When the input classifier f is f^*, we have

$$| \mathbb{P}(\hat{\phi} \neq Y) - \mathbb{P}(\phi^*_{\alpha'} \neq Y) | \leq 2g(\epsilon) + 2h^3(\epsilon) + 12h^2(\epsilon) + 16h(\epsilon)$$

with probability $1 - (2M+4)(e^{-2n^{1,0}h^2(\epsilon)} + e^{-2n^{1,1}h^2(\epsilon)}) - (1-h(2\epsilon))^{n^{0,0}} - (1-h(2\epsilon))^{n^{0,1}}$.

A.8.3 FAIREE FOR EQUALIZED ACCURACY

Algorithm 5: FaiREE for Equalized Accuracy

Input:
Training data: $S = S^{0,0} \cup S^{0,1} \cup S^{1,0} \cup S^{1,1}$
α: error bound ($\alpha > |p_{Y,1} - p_{Y,0}|$)
δ: small tolerance level
f: a classifier

1 $T^{y,a} = \{f(x_1^{y,a}), \ldots, f(x_{n_{y,a}}^{y,a})\}$
2 $\{t_{(1)}^{y,a}, \ldots, t_{(n_{y,a})}^{y,a}\}$ =sort($T^{y,a}$)
3 Define $g_1(k,a) = \mathbb{E}[\sum_{j=k}^{n^{1,a}} \binom{n^{1,a}}{j} (\frac{p_{Y,1-a}Q^{1,1-a} - \alpha}{p_{Y,a}})^j (1 - \frac{p_{Y,1-a}Q^{1,1-a} - \alpha}{p_{Y,a}})^{n^{1,a}-j}]$ with $Q^{1,a} \sim Beta(k, n^{1,a} - k + 1)$,$L_1(k^{1,0}, k^{1,1}) = g_1(k^{1,1}, 1) + g_1(k^{1,0}, 0)$
4 Define $g_0(k,a) = \mathbb{E}[\sum_{j=k}^{n^{0,a}} \binom{n^{0,a}}{j} (\frac{(1-p_{Y,1-a})Q^{0,1-a} + p_{Y,1-a} - p_{Y,a} - \alpha}{1-p_{Y,0}})^j (1 - \frac{(1-p_{Y,1-a})Q^{0,1-a} + p_{Y,1-a} - p_{Y,a} - \alpha}{1-p_{Y,a}})^{n^{0,a}-j}]$ with $Q^{0,a} \sim Beta(k+1, n^{0,a} - k)$, $L_0(k^{0,0}, k^{0,1}) = g_0(k^{0,1}, 1) + g_0(k^{0,0}, 0)$
5 Find $k^{0,0}, k^{0,1}$: $t_{(k^{0,0})}^{0,0} \leq t_{(k^{1,0})}^{1,0} < t_{(k^{0,0}+1)}^{0,0}$, $t_{(k^{0,1})}^{0,1} \leq t_{(k^{1,1})}^{1,1} < t_{(k^{0,1}+1)}^{0,1}$.
6 Build candidate set $K = \{(k^{1,0}, k^{1,1}) \mid L_1(k^{1,0}, k^{1,1}) + L_0(k^{0,0}, k^{0,1}) \leq \delta\} = \{(k_1^{1,0}, k_1^{1,1}), \ldots, (k_M^{1,0}, k_M^{1,1})\}$
7 $i_* \leftarrow \arg\min_{i \in [M]} \{\hat{e}_i\}$ ($\hat{e}_i$ is defined in Proposition 2)

Output: $\hat{\phi}(x,a) = \mathbb{1}\{f(x,a) > t_{(k_{i_*}^{1,a})}^{1,a}\}$

Similar to the algorithm for Equalized Odds, we have the following propositions and assumption:

Proposition 6. *Given $k^{1,0}, k^{1,1}$ satisfying $k^{1,a} \in \{1, \ldots, n^{1,a}\}$ ($a = 0, 1$) and $\alpha > |p_{Y,1} - p_{Y,0}|$. Define $\phi(x,a) = \mathbb{1}\{f(x,a) > t_{(k^{1,a})}^{1,a}\}$, $g_1(k,a) = \mathbb{E}[\sum_{j=k}^{n^{1,a}} \binom{n^{1,a}}{j} (\frac{p_{Y,1-a}Q^{1,1-a} - \alpha}{p_{Y,a}})^j (1 - \frac{p_{Y,1-a}Q^{1,1-a} - \alpha}{p_{Y,a}})^{n^{1,a}-j}]$ with $Q^{1,a} \sim Beta(k, n^{1,a} - k + 1)$, $q_{Y,a}(\alpha) = \frac{(1-p_{Y,1-a})Q^{0,1-a} + p_{Y,1-a} - p_{Y,a} - \alpha}{1-p_{Y,a}}$, and $g_0(k,a) = \mathbb{E}[\sum_{j=k}^{n^{0,a}} \binom{n^{0,a}}{j} (q_{Y,a}(\alpha))^j (1 - q_{Y,a}(\alpha))^{n^{0,a}-j}]$ with $Q^{0,a} \sim Beta(k+1, n^{0,a} - k)$. Then we have:*

$$\mathbb{P}(|DPE(\phi)| > \alpha) \leq g_1(k^{1,1}, 1) + g_1(k^{1,0}, 0) + g_0(k^{0,1}, 1) + g_0(k^{0,0}, 0).$$

Assumption 1. $n^{0,0} \geq \lceil \frac{\log \frac{\delta}{4}}{\log(1 - \frac{\alpha}{1-p_{Y,0}})} \rceil$, $n^{0,1} \geq \lceil \frac{\log \frac{\delta}{4}}{\log(1 - \frac{\alpha}{1-p_{Y,1}})} \rceil$, $n^{1,0} \geq \lceil \frac{\log \frac{\delta}{4}}{\log(\frac{p_{Y,1} - \alpha}{p_{Y,0}})} \rceil$, $n^{1,1} \geq \lceil \frac{\log \frac{\delta}{4}}{\log(\frac{p_{Y,0} - \alpha}{p_{Y,1}})} \rceil$, *in which $\lceil \cdot \rceil$ denotes the ceiling function.*

Theorem 11. *Under Assumption 1, we have $|DEA(\hat{\phi}_i)| \leq \alpha$ with probability $1 - \delta$, for each $i \in \{1, \ldots, M\}$.*

Corollary 1. *Under Assumption 1, we have $|DEA(\hat{\phi})| \leq \alpha$ with probability $(1-\delta)^M$, where M is the size of the candidate set.*

A.9 IMPLEMENTATION DETAILS AND OTHER EXPERIMENTS

From Lemma 3, we adopt a new way of building a much smaller candidate set. Note that our shrunk candidate set for Equality of Opportunity is:

$$K' = \{(k^{1,0}, u_1(k^{1,0})) \mid L_1(k^{1,0}, u_1(k^{1,0})) \leq \delta\}.$$

Since Equalized Odds constraint is an extension of Equality of Opportunity, our target classifier should be in K'.

To select our target classifier, it's sufficient to add a condition of similar false negative rate between privileged and unprivileged groups. Specifically, we choose our final candidate set as below:

$$K'' = \{(k^{1,0}, u_1(k^{1,0})) \mid L_1(k^{1,0}, u_1(k^{1,0})) \leq \delta, L_0(k^{0,0}, k^{1,0}) \leq \delta\}.$$

We also did experiments on other benchmark datasets.

First, we apply FaiREE to German Credit dataset (Kamiran & Calders, 2009), whose task is to predict whether a bank account holder's credit is *good* or *bad*. The protected attribute is gender, and the sample size is 1000, with 800 training samples and 200 test samples. To facilitate the numerical study, we randomly split data into training set, calibration set and test set at each repetition and repeat for 500 times.

Table 5: Result of different methods on German Credit dataset

	Eq	C-Eq	ROC	FairBayes			FaiREE-EOO			FaiREE-EO		
α	/	/	/	0.07	0.1	0.14	0.07	0.1	0.14	0.07	0.1	0.14
$\overline{\lvert DEOO \rvert}$	0.078	0.093	0.100	0.126	0.125	0.126	0.020	0.029	0.034	0.001	0.025	0.048
$\lvert DEOO \rvert_{95}$	0.179	0.127	0.267	0.160	0.182	0.186	0.066	0.097	0.130	0.004	0.084	0.120
$\overline{\lvert DPE \rvert}$	0.109	0.072	0.138	/	/	/	/	/	/	0.041	0.063	0.092
$\lvert DPE \rvert_{95}$	0.235	0.114	0.334	/	/	/	/	/	/	0.059	0.097	0.133
$\overline{ACC}$	0.707	0.720	0.591	0.722	0.723	0.723	0.717	0.729	0.745	0.702	0.721	0.722

Then we apply FaiREE to Compas Score dataset (Angwin et al., 2016), whose task is to predict whether a person will conduct crime in the future. The protected attribute is gender, and the sample size is 5278, with 4222 training samples and 1056 test samples. To facilitate the numerical study, we randomly split data into training set, calibration set and test set at each repetition and repeat for 500 times.

Table 6: Result of different methods on Compas Score dataset

	Eq	C-Eq	ROC	FairBayes			FaiREE-EOO			FaiREE-EO		
α	/	/	/	0.07	0.1	0.14	0.07	0.1	0.14	0.07	0.1	0.14
$\overline{\lvert DEOO \rvert}$	0.083	0.642	0.070	0.077	0.109	0.136	0.026	0.033	0.042	0.017	0.049	0.098
$\lvert DEOO \rvert_{95}$	0.101	0.684	0.174	0.145	0.186	0.250	0.068	0.097	0.131	0.068	0.090	0.140
$\overline{\lvert DPE \rvert}$	0.025	0.291	0.067	/	/	/	/	/	/	0.024	0.048	0.060
$\lvert DPE \rvert_{95}$	0.047	0.332	0.148	/	/	/	/	/	/	0.053	0.092	0.114
$\overline{ACC}$	0.629	0.664	0.652	0.658	0.658	0.659	0.654	0.660	0.672	0.553	0.654	0.669