Towards Threshold Invariant Fair Classification

Mingliang Chen and Min Wu
Electrical and Computer Engineering Department and Institute for Advanced Computer Studies
University of Maryland, College Park, MD, USA
{mchen126, minwu}@umd.edu

Abstract

Effective machine learning models can automatically learn useful information from a large quantity of data and provide decisions in a high accuracy. These models may, however, lead to unfair predictions in certain sense among the population groups of interest, where the grouping is based on such sensitive attributes as race and gender. Various fairness definitions, such as demographic parity and equalized odds, were proposed in prior art to ensure that decisions guided by the machine learning models are equitable. Unfortunately, the "fair" model trained with these fairness definitions is threshold sensitive, i.e., the condition of fairness may no longer hold true when tuning the decision threshold. This paper introduces the notion of threshold invariant fairness, which enforces equitable performances across different groups independent of the decision threshold. To achieve this goal, this paper proposes to equalize the risk distributions among the groups via two approximation methods. Experimental results demonstrate that the proposed methodology is effective to alleviate the threshold sensitivity in machine learning models designed to achieve fairness.

1 INTRODUCTION

Machine learning is being used across a wide variety of practical tasks in recent years, influencing many aspects of our daily life. Many companies and government departments have deployed machine learning based decision-making systems to facilitate decision making in business operations. The U.S. courts use a software known as Correctional Offender Management Profiling for Alternative Sanctions (COMPAS), to measure the risk for a defendant to recommit another crime, which helps judges make parole decisions. Banks turn to machine learning models to evaluate the prospective customers' credits and to help decide the approval of loan applications. Like humans, machine learning algorithms can suffer from bias that render their decisions "unfair" (Barocas & Selbst, 2016; Mehrabi et al., 2019). A number of investigations have shown that the existing machine learning systems have fairness issues in certain sense. For instance, the works by (Angwin et al., 2016; Larson et al., 2016) studied the COMPAS software and found a bias against African-Americans: it is more likely to assign a higher risk score to an African-American defendants than to a Caucasian with a similar profile.

In the context of decision making, fairness typically means the equivalent (non-discriminatory) decision performances across different groups of people based on their inherent or acquired characteristics. From this point of view, a number of recent studies introduced *demographic parity* (DP) (Dwork et al., 2012; Feldman et al., 2015) and *equalized odds* (EO) (Hardt et al., 2016) to characterize and evaluate the fairness level in the machine learning models. Since the goal of these definitions is to equalize certain probability statistics computed from the confusion matrix of classification, this type of definitions were named as *classification parity* in (Corbett-Davies & Goel, 2018). To take the COMPAS case as an example, DP means that parole granting rates should be equal across the race groups; and EO enforces that among defendants who would or would not have gone on to commit a violent crime if released, the rates are equal across the race groups.

The fairness algorithms in the literature mainly fall into three categories: 1) Preprocessing: the studies by (Zemel et al., 2013; Calmon et al., 2017) map the data to a fair representation in a latent space satisfying the defined fairness or independent of the protected attribute. Ideally, removing correlated features and finding independent representation to the protected attribute may lead

Proceedings of the 36th Conference on Uncertainty in Artificial Intelligence (UAI), PMLR volume 124, 2020.

to undesirable decision outcomes. Imagine a company scores randomly on the applicants over two groups. Even though the scoring function is independentof the groups, it is likely that a considerable number of unqualified applicants are selected at last. 2) Constraints on training: the works by (Zafar et al., 2017a,b) constrain the classifier on DP and EO, respectively, during the training. Fair classification can also be reduced to a set of subproblems (Agarwal et al., 2018). The tradeoff is required between accuracy and fairness constraints. 3) Postprocessing: the work by (Hardt et al., 2016) searches proper thresholds on scores over different groups. The studies by (Crowson et al., 2016; Pleiss et al., 2017) calibrate on the prediction score such that the probability of positive label is equal to the prediction score. The postprocessing requires access to protected attributes in the test phase.

Classification parity can help machine learning models achieve equitable decision performances across the groups from a probability point of view, such as positive proportion of decisions, true positive rate, false positive rate, and other statistics alike. However, such parity of the group-wise performances is generally not retained, when we tune the decision threshold. Trained under the constraint of classification parity, the classifiers only "delicately" achieve fairness requirement in the default decision thresholds, and are sensitive and vulnerable to the change of decision thresholds. We will illustrate the limitation of classification parity in Section 2.2.

To alleviate the limitation of classification parity, we introduce a new notion of *threshold invariant fairness* in this paper. Unlike classification parity, threshold invariant fairness enforces a stronger condition on the classifier such that it has a consistent fairness level of classification results against the change of decision threshold.

To develop our proposed notion, we find that it is sufficient to design the scoring function of the classifier to equalize the distributions of the risk scores over all groups. From classification parity to threshold invariant fairness, the equalization focus shifts from statistical attributes at a given decision rule to the probability distribution of the risk scores, suggesting that our proposed notion imposes a stronger constraint than that by the classification parity. In this paper, we propose two approximation approaches to equalize the distributions of risk scores over all groups and incorporate them into a variety of differentiable classifiers, such as logistic regression and support vector machine. The problem can be solved by a gradient-based optimizer. Our new methods have two advantages compared with the prior art: 1) we do not require the protected attributes in the test phase; 2) we can tune the threshold to modify the positive rate of the classifier predictions and maintain the fairness as much as possible.

2 BACKGROUND

2.1 CLASSIFICATION PARITY

Classification parity is an approach to alleviate the machine bias by equalizing performance measure of classification across the groups with the protected attributes to achieve a group-wise fair classification. As early reviewed, *demographic parity* (Dwork et al., 2012; Feldman et al., 2015) and *Equalized Odds* (Hardt et al., 2016) are two well-known fairness definitions in classification parity. We assume that one input sample with observed m-dimensional features $X \in \mathbb{R}^m$ has a protected attribute $A \in \{0, 1\}$, and a groud truth label $Y \in \{0, 1\}$; and $\hat{Y} \in \{0, 1\}$ is the classifier's prediction. In a simplified recidivism example, A may represent the defendant's race as African-American versus Caucasian, Y represents whether the defendant recommits another crime after being released. The definition of two existing classification parities are described as follows.

Definition 1. *(Demographic Parity). A predictor $\hat{Y}$ satisfies demographic parity if*

$$P(\hat{Y} = 1|A = 0) = P(\hat{Y} = 1|A = 1). \tag{1}$$

Definition 2. *(Equalized Odds). A predictor $\hat{Y}$ achieves equalized odds if the true positive rates and false positive rates across the groups are equal, respectively, i.e.,*

$$\begin{cases} P(\hat{Y} = 1|A = 0, Y = 0) = P(\hat{Y} = 1|A = 1, Y = 0), \\ P(\hat{Y} = 1|A = 0, Y = 1) = P(\hat{Y} = 1|A = 1, Y = 1). \end{cases} \tag{2}$$

We see that DP enforces that the selection rates are equal across all the protected attributes; EO enforces that the prediction accuracy is equally high across all the protected attributes. Based on the above parity definition, many papers have proposed algorithms to achieve classification parity via pre-processing (Kamiran & Calders, 2012; Zemel et al., 2013), post-processing (Hardt et al., 2016), and regularization techniques (Dwork et al., 2012; Feldman et al., 2015; Zafar et al., 2017a,b; Corbett-Davies et al., 2017; Agarwal et al., 2018).

2.2 LIMITATION OF THE PRIOR ART

As pointed out by (Corbett-Davies & Goel, 2018), classification parity is a problematic measure of fairness. We consider a risk score $s(X) = P(\hat{Y} = 1|X)$, describing the probability that an input sample X will be predicted positive. Given a group $\mathcal{D}_a$ with protected attribute $A = a$, we can compute the distribution of $s(X)$

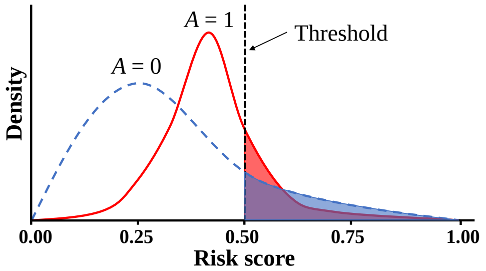


Figure 1: Hypothetical risk distributions of two groups (e.g., $A = 0$ and $A = 1$). When classification parity is satisfied, i.e., the areas of the highlighted regions are equal, the risk distributions can still differ from each other. Classification parity does not achieve equitable decision performance in terms of risk distribution.

over the group $\mathcal{D}_a$, which we refer to as risk distribution $r(\mathcal{D}_a)$. Figure 1 shows hypothetical risk distributions for two groups with different protected attributes. In the model inference stage, an input sample is classified by comparing its risk score with a decision threshold. The proportion of the positive predictions in one group is the fraction of the risk distribution that is to the right of the decision threshold, highlighted in Figure 1.

According to the definitions reviewed in Section 2.1, to achieve classification parity is to equalize the positive (or negative) proportions among different groups. However, when the classifier achieves classification parity at a specific decision threshold, i.e., the areas of the highlighted regions are equal, the risk distributions can still differ from each other. This suggests that the classification parity can be easily destroyed by changing the decision threshold. The "fair" machine learning model under classification parity does not show equitable performances among different groups in terms of the risk distribution. The goal of this paper is to propose a stronger condition on fairness to alleviate the above issue, and to develop systematic method by designing the scoring function during the classifier training to achieve it.

3 THRESHOLD INVARIANT FAIRNESS

3.1 NOTATIONS

Recall that an input sample with m-D features $X \in \mathbb{R}^m$ has protected attribute $A \in \{0, 1\}$, its corresponding ground truth label $Y \in \{0, 1\}$, and the classifier prediction is $\hat{Y} \in \{0, 1\}$. Define the scoring function in classifier $g : \mathbb{R}^m \mapsto \mathbb{R}$, mapping the m-D features to a scalar, the raw score. A *sigmoid* activation $\sigma(\cdot)$ is applied to confine the raw score in the range of $[0, 1]$, which can be interpreted as the risk score defined in Section 2.2, i.e., $s(X) = \sigma(g(X))$. Given the threshold $t \in [0, 1]$, the prediction of the sample X is determined as

$$\hat{Y} = \begin{cases} 0, & s(X) \le t, \\ 1, & s(X) > t. \end{cases} \tag{3}$$

3.2 FAIRNESS DEFINITION

We define threshold invariant fairness, a stronger condition on classification parity, in the context of DP and EO.

Definition 3. *(Threshold Invariant Fairness). Threshold Invariant Demographic Parity (TIDP) or Threshold Invariant Equalized Odds (TIEO) is achieved when DP or EO is satisfied, respectively, independent of the decision threshold t.*

We use the Calder-Verwer (CV) score (Calders & Verwer, 2010) to measure classification parity. For simplicity, we denote that $P_a(\hat{Y}) := P(\hat{Y}|A = a)$ and $P_{a,y}(\hat{Y}) := P(\hat{Y}|A = a, Y = y)$. The CV scores for DP and EO are defined as

$$\begin{aligned} \Delta\text{DP} =& \big|P_0(\hat{Y} = 1) - P_1(\hat{Y} = 1)\big|, \\ \Delta\text{EO} =& \frac{1}{2}\big(\big|P_{0,0}(\hat{Y} = 1) - P_{1,0}(\hat{Y} = 1)\big| \\ &+ \big|P_{0,1}(\hat{Y} = 1) - P_{1,1}(\hat{Y} = 1)\big|\big). \end{aligned} \tag{4}$$

Note that the smaller the CV score is, the better the classifier achieves a fair classification. In the most ideal cases, the classification parity, DP or EO, is satisfied when ΔDP or ΔEO is zero.

In the following, we investigate the relationship between threshold invariant fairness and the risk distribution. Denote $f_a(s) := f(s|A = a)$ and $f_{a,y}(s) := f(s|A = a, Y = y)$ as the risk distributions over the group with the protected attribute A and label Y. Recall that the support of the risk score s is $[0, 1]$ and the decision threshold t ranges within $[0, 1]$. For the DP constraint, we have

$$\begin{aligned} P_a(\hat{Y} = 1) = P_a(s(X) > t) = \textstyle\int_t^1 f_a(s)\mathrm{d}s, \\ P_a(\hat{Y} = 0) = P_a(s(X) \le t) = \textstyle\int_0^t f_a(s)\mathrm{d}s. \end{aligned} \tag{5}$$

Then,

$$\begin{aligned} \Delta\text{DP} &= \big|P_0(\hat{Y} = 1) - P_1(\hat{Y} = 1)\big| \\ &= \big|\int_t^1 \big(f_0(s) - f_1(s)\big)\mathrm{d}s\big| \\ &\le \int_t^1 \big|f_0(s) - f_1(s)\big|\mathrm{d}s \le \epsilon_{DP}(1 - t). \end{aligned} \tag{6}$$

where we define the upper bound of the difference between two risk distributions $\big|f_0(s) - f_1(s)\big| \le \epsilon_{DP}, \forall s \in [0, 1]$. Since $P_a(\hat{Y} = 1) - P_a(\hat{Y} = 0) = 1, \forall a \in \{0, 1\}$, we have

$$\Delta\text{DP} = \left|P_1(\hat{Y}=0) - P_0(\hat{Y}=0)\right| \\ \leq \int_0^t \left|f_1(s) - f_0(s)\right| \mathrm{d}s \leq \epsilon_{DP} t. \tag{7}$$

Combining (6) and (7), we have

$$\Delta\text{DP} \leq \epsilon_{DP} \cdot \min\{t, 1-t\} \leq \frac{1}{2}\epsilon_{DP}. \tag{8}$$

Remark 1. *(The sufficient condition of TIDP). When ϵ_{DP} reaches 0, DP is satisfied regardless of the decision threshold. Hence, equalizing the risk distributions between the positive and negative protected attributes, i.e., $f_0(s) = f_1(s)$, is the sufficient condition of TIDP.*

Similarly, for EO constraint, we have

$$\Delta\text{EO} \leq \frac{1}{2}(\epsilon_0 + \epsilon_1) \cdot \min\{t, 1-t\} \leq \frac{1}{4}(\epsilon_0 + \epsilon_1). \tag{9}$$

where we define the upper bounds $\left|f_{0,0}(s) - f_{1,0}(s)\right| \leq \epsilon_0$ and $\left|f_{0,1}(s) - f_{1,1}(s)\right| \leq \epsilon_1, \forall s \in [0,1]$. The derivation is given in Appendix A.

Remark 2. *(The sufficient condition of TIEO). When ϵ_0 and ϵ_1 reach 0, EO is satisfied regardless of the decision threshold. Hence, equalizing the risk distributions of the positive/negative samples between positive and negative protected feature attributes, i.e., $f_{0,y}(s) = f_{1,y}(s)$, $\forall y \in \{0,1\}$, respectively, is the sufficient condition of TIEO.*

Instead of equalizing statistical attributes, threshold invariant fairness requires equalization of risk distributions, which is a stricter condition than the family of classification parity.

4 PROPOSED FAIR CLASSIFIER

So far, we know that equalizing risk distributions across different groups is a sufficient condition to achieve threshold-invariant fairness. One straightforward post-processing approach is to find a mapping function to equalize the histogram of risk scores over one group to another group, a strategy similar to histogram matching in image processing (Gonzalez et al., 2001). Here, we focus on holistic systematic approaches to equalize risk distributions during the classifier training by formulating an appropriate fairness regularization. We start with the formulation of the proposed fairness in Section 4.1, and then show how to incorporate the regularizer into the classifier design in Section 4.2.

4.1 EQUALIZATION OF RISK DISTRBUTION

In this subsection, we present two approaches to design the regularizer between two groups $\mathcal{D}_0$ and $\mathcal{D}_1$ in the whole sample set $\mathcal{D}$.

Approach 1: Histogram Approximation (HA)

Define $C = \{c_j\}_{j=1}^N$ as the bin centers of N-bin histogram with equal interval and the bin width $\delta = c_j - c_{j-1}$. The count of the samples from the group $\mathcal{D}_i$ in the bin $(c - \frac{\delta}{2}, c + \frac{\delta}{2})$ is expressed as

$$n_c = \sum_{X \in \mathcal{D}_i} \Pi_c(s(X)), \tag{10}$$

where $\Pi_c(\cdot)$ is a rectangular function:

$$\Pi_c(x) = \begin{cases} 1, & x \in (c - \frac{\delta}{2}, c + \frac{\delta}{2}), \\ 0, & \text{otherwise}. \end{cases} \tag{11}$$

Since $\Pi_c(\cdot)$ is not differentiable, we approximate it with a Gaussian kernel,

$$n_c = \sum_{X \in \mathcal{D}_i} G_c(s(X)) = \sum_{X \in \mathcal{D}_i} \exp(-\frac{(s(X) - c)^2}{2\sigma_c^2}). \tag{12}$$

Hence, the histogram of the risk score in $\mathcal{D}_i$ can be expressed as

$$h(\mathcal{D}_i) = normalize\big([n_{c_1}, n_{c_2}, ..., n_{c_N}]\big), \tag{13}$$

where $normalize(\cdot)$ scales the histogram such that the histogram entries sum to one. Note that $h(\mathcal{D}_i)$ is an N-D vector since the histogram has N bins. The HA method formulates a differentiable histogram via a Gaussian kernel to construct the risk distribution.

We use the symmetric combination of Kullback–Leibler (KL) divergence to evaluate the distance of the risk distributions between two groups $\mathcal{D}_0$ and $\mathcal{D}_1$.

$$\begin{aligned} & d\big(r(\mathcal{D}_0), r(\mathcal{D}_1)\big) \\ =& D_{KL}\big(h(\mathcal{D}_0)||h(\mathcal{D}_1)\big) + D_{KL}\big(h(\mathcal{D}_1)||h(\mathcal{D}_0)\big) \\ =& \sum_{i=1}^N \big(h_{(i)}(\mathcal{D}_0) - h_{(i)}(\mathcal{D}_1)\big) \log \frac{h_{(i)}(\mathcal{D}_0)}{h_{(i)}(\mathcal{D}_1)}. \end{aligned} \tag{14}$$

where subscript (i) denotes the i-th entry of the histogram. It can be seen that the distance d is non-negative and approaches zero when two risk distributions are equal, i.e., $r(\mathcal{D}_0) = r(\mathcal{D}_1)$.

Approach 2: Gaussian Assumption (GA)

Since the risk scores are the sigmoid of the raw scores, i.e., $s(X) = \sigma(g(X))$, equalizing the distribution of raw scores would lead to the equalization of the risk distribution. Given two groups $\mathcal{D}_0$ and $\mathcal{D}_1$, we assume that the distributions of raw scores obey the same Gaussian distribution over $\mathcal{D}_0$ and $\mathcal{D}_1$, and the mean and variance are sufficient to characterize the distributions. We estimate means and variances of the distributions of raw score over $\mathcal{D}_0$ and $\mathcal{D}_1$ using maximum likelihood estimate:

$$\mathcal{D}_i : \begin{cases} \hat{\mu}_i = \frac{1}{|\mathcal{D}_i|} \sum_{X \in \mathcal{D}_i} g(X) \\ \hat{\sigma}_i^2 = \frac{1}{|\mathcal{D}_i|} \sum_{X \in \mathcal{D}_i} \big(g(X) - \hat{\mu}_i\big)^2 \end{cases}, \ i \in \{0, 1\}. \tag{15}$$

Recall that KL divergence of two 1-D Gaussian distributions $\mathcal{N}_0(\mu_0, \sigma_0^2)$ and $\mathcal{N}_1(\mu_1, \sigma_1^2)$ is

$$D_{KL}\big(\mathcal{N}_0 || \mathcal{N}_1\big) = \frac{1}{2}\big(\log \frac{\sigma_1^2}{\sigma_0^2} + \frac{(\mu_0 - \mu_1)^2 + \sigma_0^2}{\sigma_1^2} - 1\big). \tag{16}$$

Based on (16), we define the distance between two risk distributions $r(\mathcal{D}_0)$ and $r(\mathcal{D}_1)$ as

$$\begin{aligned} d\big(r(\mathcal{D}_0), r(\mathcal{D}_1)\big) &= D_{KL}\big(\mathcal{N}_0||\mathcal{N}_1\big) + D_{KL}\big(\mathcal{N}_1||\mathcal{N}_0\big) \\ &= \frac{1}{2}\big(\frac{(\hat{\mu}_0 - \hat{\mu}_1)^2 + \hat{\sigma}_0{}^2}{\hat{\sigma}_1^2} + \frac{(\hat{\mu}_1 - \hat{\mu}_0)^2 + \hat{\sigma}_1^2}{\hat{\sigma}_0^2} - 2\big). \end{aligned} \tag{17}$$

Similar to (14), the distance d is non-negative and approaches zero when the means and variances from $\mathcal{D}_0$ and $\mathcal{D}_1$ are equal, i.e., $\hat{\mu}_0 = \hat{\mu}_1$ and $\hat{\sigma}_0^2 = \hat{\sigma}_1^2$.

Note that the distance d in both approaches discussed in this section are differentiable, which can facilitate the classifier optimization using gradient-based methods.

Fairness Regularizer E_f

For simplicity, we assume that the sample set $\mathcal{D}$ has a binary protected attribute, i.e., $|A| = 2$. To distinguish the DP and EO cases, we name the regularizers as $E_{f,DP}$ and $E_{f,EO}$, respectively. For TIDP case, $\mathcal{D}$ is splited into two groups: negative protected attribute $\mathcal{D}_0$ and positive protected attribute $\mathcal{D}_1$. The regularizer fir this case $E_{f,DP}$ is defined as

$$E_{f,DP}(\mathcal{D}) = d(r(\mathcal{D}_0), r(\mathcal{D}_1)). \tag{18}$$

For TIEO case, $\mathcal{D}$ is splited into four groups: negative and positive samples with negative protected attribute, $\mathcal{D}_{0,n}$ and $\mathcal{D}_{0,p}$, respectively, and negative and positive samples with positive protected attribute, $\mathcal{D}_{1,n}$ and $\mathcal{D}_{1,p}$, respectively. The regularizer for this case $E_{f,EO}$ is defined as

$$E_{f,EO}(\mathcal{D}) = \sum_{y \in \{n,p\}} d\big(r(\mathcal{D}_{0,y}), r(\mathcal{D}_{1,y})\big). \tag{19}$$

Note that the regularizer can be extended to the case of multiple protected attributes, i.e., $|A| > 2$.

4.2 CLASSIFIER FORMULATION

As a proof-of-concept, we consider two kinds of classifiers: logistic regression (LR) and support vector machine (SVM). It is worthwhile to note that our proposed method can be easily extended to other complex differentiable classifiers. Define the classifier $g(x) = w^{\mathrm{T}}x + b$, and the risk score is computed as $s(x) = \sigma(g(x))$.

LR Model: The loss function L_{LR} for logistic regression can be expressed as

$$L_{\text{LR}}(w, b) = \frac{1}{|\mathcal{D}|} \sum_{(X,Y) \in \mathcal{D}} L_{ce}(s(X), Y) + \eta E_f(\mathcal{D}), \tag{20}$$

where $L_{ce}(\cdot)$ denotes the cross-entropy loss of the risk score and the label, and η is a positive hyperparameter. Since (20) is differentialable, we can use gradient-based methods to minimize (20) with respect to w and b.

Linear SVM Model: Pegasos was proposed to solve SVM with the hinge loss version using a sub-gradient solver (Shalev-Shwartz et al., 2011), i.e.,

$$L_{\text{Pegasos}}(w, b) = \frac{\lambda}{2}||w||^2 + \frac{1}{|\mathcal{D}|} \sum_{(X,Y) \in \mathcal{D}} \max\{0, 1 - Y \cdot g(X)\}, \tag{21}$$

where λ is a positive hyperparameter, and the label $Y \in \{-1, 1\}$ in an SVM denotes negative/positive label.

By incorporating the fairness regularizer into (21), we obtain the loss function as

$$\begin{aligned} L_{\text{LSVM}}(w, b) = &\frac{\lambda}{2}||w||^2 + \eta E_f(\mathcal{D}) \\ &+ \frac{1}{|\mathcal{D}|} \sum_{(X,Y) \in \mathcal{D}} \max\{0, 1 - Y \cdot g(X)\}, \end{aligned} \tag{22}$$

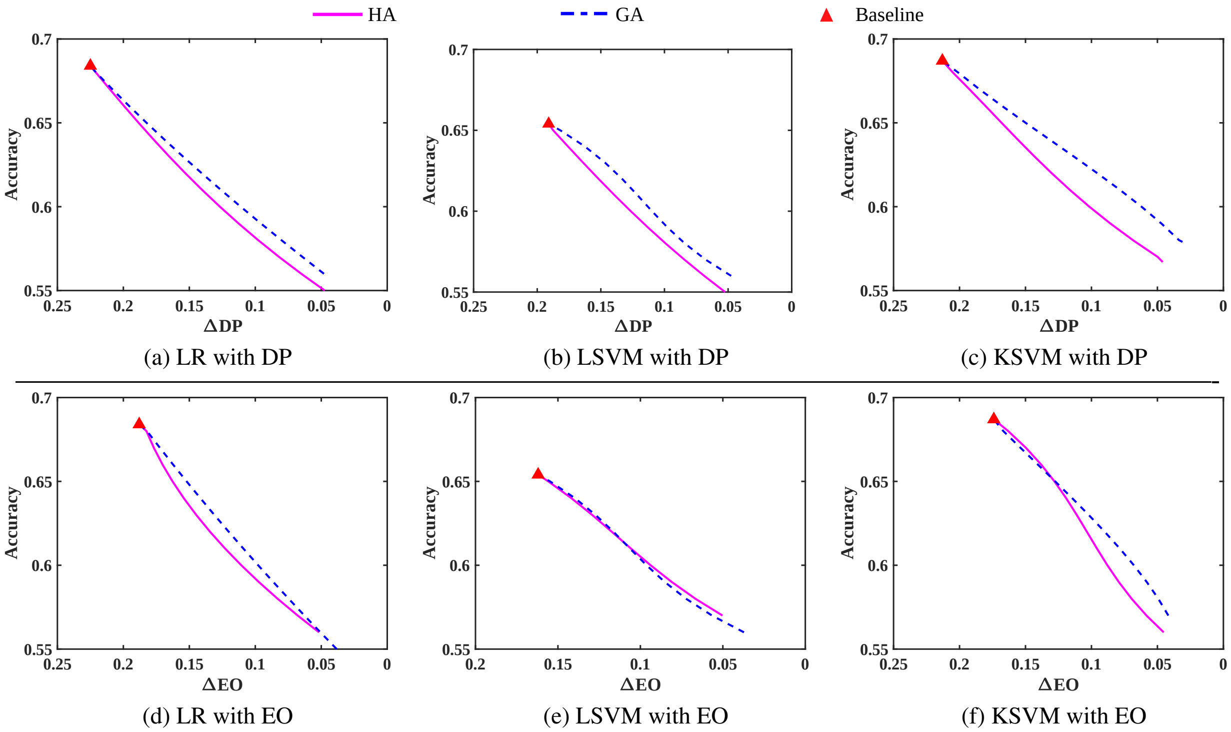
where λ and η are positive hyperparameters. Similar to Pegasos in (21), we can apply sub-gradient solvers, due to the sub-differentiability of (22).

Kernel SVM Model: In a kernel SVM, we use a mapping function $\phi(\cdot)$ to transform the original features to a higher dimensional subspace. Substituting original features X with the mapped features $\phi(X)$, we can rewrite (22) as

$$\begin{aligned} L_{\text{KSVM}}(w, b) = &\frac{1}{|\mathcal{D}|} \sum_{(X,Y) \in \mathcal{D}} \max\{0, 1 - Y \cdot g(\phi(X))\} \\ &+ \frac{\lambda}{2}||w||^2 + \eta E_f(\phi(\mathcal{D})). \end{aligned} \tag{23}$$

According to the Representer Theorem (Kimeldorf & Wahba, 1971), the optimal solution of (23) can be spanned by the training samples, i.e., it is of the form $w = \sum_{i=1}^{|\mathcal{D}|} \alpha_i \phi(X_i)$. Hence, the kernel SVM predictor becomes $\tilde{g}(x) = \sum_{i=1}^{|\mathcal{D}|} \alpha_i \phi(X_i)^{\mathrm{T}} \phi(x) + b = \sum_{i=1}^{|\mathcal{D}|} \alpha_i K(X_i, x) + b$, where we define the kernel operator $K(X_i, x) = \phi(X_i)^{\mathrm{T}} \phi(x)$. Equation (23) can be rewritten with respect to α as

$$\begin{aligned} L_{\text{KSVM}}(\alpha, b) = &\frac{\lambda}{2} \sum_{i,j=1}^{|\mathcal{D}|} \alpha_i \alpha_j K(X_i, X_j) + \eta E_f(\phi(\mathcal{D})) \\ &+ \frac{1}{|\mathcal{D}|} \sum_{(X,Y) \in \mathcal{D}} \max\{0, 1 - Y \cdot \tilde{g}(X)\}, \end{aligned} \tag{24}$$



(a) LR with DP (b) LSVM with DP (c) KSVM with DP

(d) LR with EO (e) LSVM with EO (f) KSVM with EO

Figure 2: Tradeoff between model accuracy and classification parity index: (a) LR with the DP constraint, (b) LSVM with the DP constraint, (c) KSVM with the DP constraint, (d) LR with the EO constraint, (e) LSVM with the EO constraint, and (f) KSVM with EO constraint. The red filled triangles indicate the baseline cases in each classifier.

where λ and η are positive hyperparameters. To calculate the fairness regularizer, we can use the kernel trick to compute the risk scores and the risk distributions over the mapped sample set $\phi(\mathcal{D})$. Similarly, we can use subgradient-based methods to optimize the loss function (24) with respect to α.

5 EXPERIMENTS

In this section, we carry out experiments on the COMPAS risk assessment dataset compiled by ProPublica (Larson et al., 2016) and evaluate how well our proposed method can alleviate the problem of threshold variant fairness in the prior art of classification parity.

5.1 DATASET AND FEATURE PROCESSING

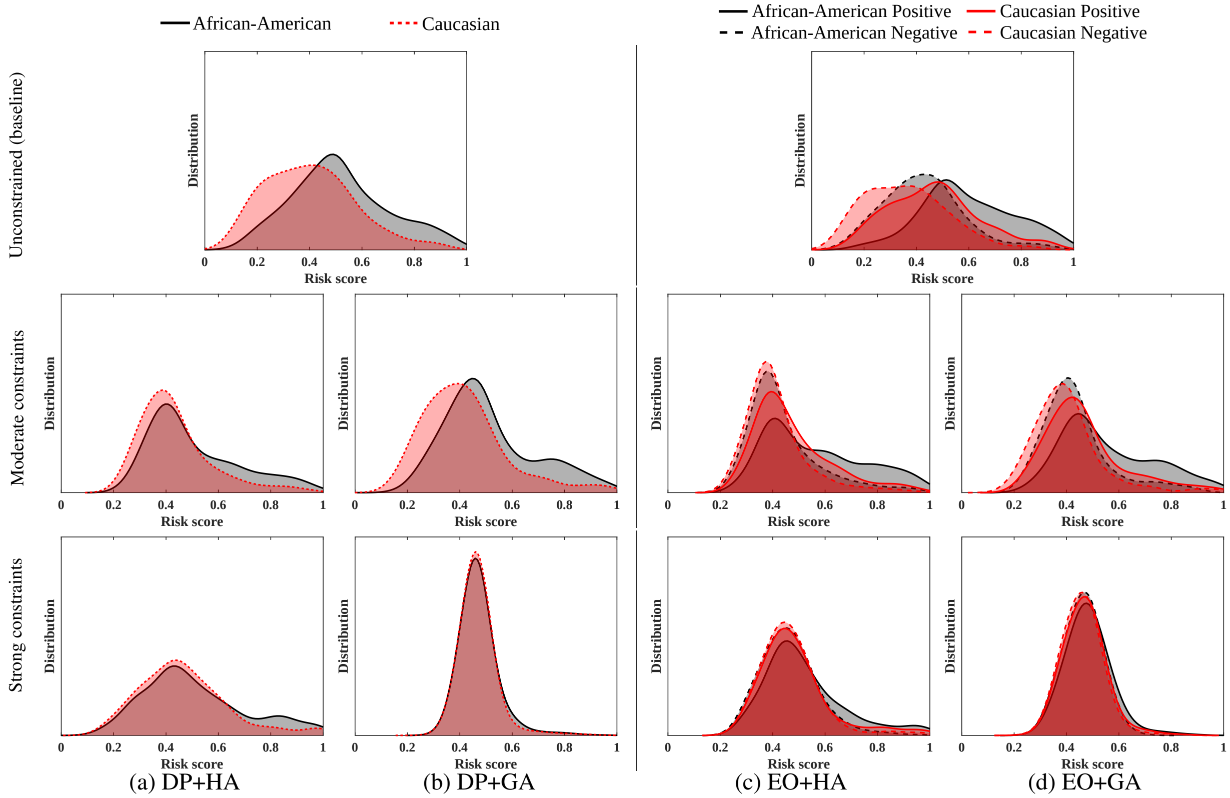
ProPublica compiled a list of all criminal defendants screened through the COMPAS tool in Broward County, Florida, during 2013 to 2014.[1] The dataset contains defendants' records of prison times, demographics (e.g., gender, race, and age), criminal histories (current charge type, charge degree, and number of prior crimes), the COMPAS risk scores, and the ground truth of recidivism within two years after the screening. Previous discussions on bias decisions from machine learning on this dataset can be found in (Angwin et al., 2016; Dieterich et al., 2016; Dressel & Farid, 2018).

For simplicity, we only consider a subset of defendants whose race was either African-American or Caucasian, which is the protected attribute in our study. The set of features used in the classification task is summarized in Table 1. We only use the protected attribute, "race", as an indicator of the groups in the training and exclude it from the features for the classification task. The features are a combination of continuous and categorical features. For continuous features, we subtract their mean and scale them to unit variance; for categorical features, we use $0/1$ to encode the features, expect for the labels in SVM where $-1/+1$ encoding is used. The dataset was randomly split into training (70%) and test (30%).

5.2 TRADEOFF BETWEEN ACCURACY AND FAIRNESS

We have trained three types of classifiers: LR, linear SVM (LSVM), and kernel SVM (KSVM). For SVM, we set hyperparameter $\lambda = \frac{1}{10\cdot|\mathcal{D}|}$. We used radial ba-

[1] The dateset can be downloaded from `https://github.com/propublica/compas-analysis`.



(a) DP+HA (b) DP+GA (c) EO+HA (d) EO+GA

Figure 3: Risk distributions of LR classifiers with different levels of fairness regularizers. The first row is the baseline, i.e., unconstrained classifier; the second row is the classifier with moderate constraints; the third row is the classifier with strong constraints. (a) DP+HA, (b) DP+GA, (c) EO+HA, and (d) EO+GA. The figures are best viewed in color.

Table 1: Preprocessed Features Used in COMPAS Data

Feature	Type	Preprocessing
Age	Continuous	Normalization
Gender	Binary	0-1 encoding
Race (Protected)	Binary	0-1 encoding
Priors count	Continuous	Normalization
Charge degree	Binary	0-1 encoding
Recidivism (Label)	Binary	0-1 encoding

sis function (RBF) kernel with $\gamma = 0.5$ in KSVM, i.e., $K(x_1, x_2) = \exp(-\gamma||x_1 - x_2||_2^2)$. We applied gradient descent to optimize the models with momentum 0.9 and learning rate starting from 0.1 to 0.0001.

We tuned η from 0 to 5 in each classifier to show the tradeoff curve of accuracy versus fairness in Figure 2 via the two proposed equalization approaches: histogram approximation (HA) and gaussian assumption (GA). Note that the classifiers with $\eta = 0$ are the unconstrained classifiers, considered as *baseline*.

As shown in Figure 2, we can see that our proposed methods are effective at alleviating disparate learning performances across the groups. The CV scores of ΔDP and ΔEO reduce from around 0.2 to 0.05, with the help of the proposed fairness regularizers, whereas the resulting accuracies drop by approximately 0.1 on a $[0, 1]$ scale. Generally, our proposed methods can be used in different types of classifiers (e.g., LR, LSVM, and KSVM) and different types of classification parity indices (e.g., DP and EO). Comparing two equalization methods of HA and GA, the models with GA based fairness regularizer have slightly better performances than those with the HA approach in terms of the accuracy cost to achieve the same fairness level.

5.3 VISUALIZATION OF RISK DISTRIBUTION

In this section, we use linear regression classifiers as an example to visualize the influence of the proposed fairness regularizers on risk distributions. Figure 3 shows the risk distributions of this type of classifiers in three example cases: with no constraints (baseline), medium constraints, and large constraints. For the DP constraint, we present the risk distributions of the African-American group and the Caucasian group; for the EO constraint,

Table 2: Performance Under DP Constraint

	Accuracy	**ΔDP**	**Int.**	**STD**
Baseline	68.4%	0.225	0.145	0.044
Zafar-DP	57.4%	0.060	0.079	0.024
LR-HA	56.6%	0.075	0.089	0.024
LR-GA	56.9%	0.066	0.093	0.023
LSVM-HA	56.2%	0.035	0.073	0.022
LSVM-GA	56.5%	0.048	0.097	0.020
KSVM-HA	57.0%	0.059	0.057	0.014
KSVM-GA	58.4%	0.064	0.098	0.024

Table 3: Performance Under EO Constraint

	Accuracy	**ΔEO**	**Int.**	**STD**
Baseline	68.4%	0.188	0.123	0.037
Zafar-EO	63.2%	0.120	0.154	0.047
LR-HA	62.7%	0.117	0.077	0.024
LR-GA	62.9%	0.117	0.098	0.028
LSVM-HA	63.3%	0.137	0.109	0.029
LSVM-GA	63.0%	0.129	0.062	0.018
KSVM-HA	63.2%	0.112	0.067	0.012
KSVM-GA	62.9%	0.096	0.040	0.009

we show the risk distributions of African-American recidivists, African-American non-recidivists, Caucasian recidivists, and Caucasian non-recidivists, respectively.

Several observations can be made from Figure 3. First, the distributions from the baseline classifiers are distinct among the groups, suggesting the discriminating decision performances for different groups. The unconstrained classifiers suffer from some bias in their predictions. Second, our proposed methods progressively equalize the risk distributions among the groups as we weigh more on fairness regularizers. When we impose medium constraints, the corresponding risk distributions become closer in shape than that in the baseline cases. When we keep increasing the constraints to a significant level, the corresponding risk distributions almost overlap with each other, indicating the effectiveness of the regularizer to equalize the risk distributions.

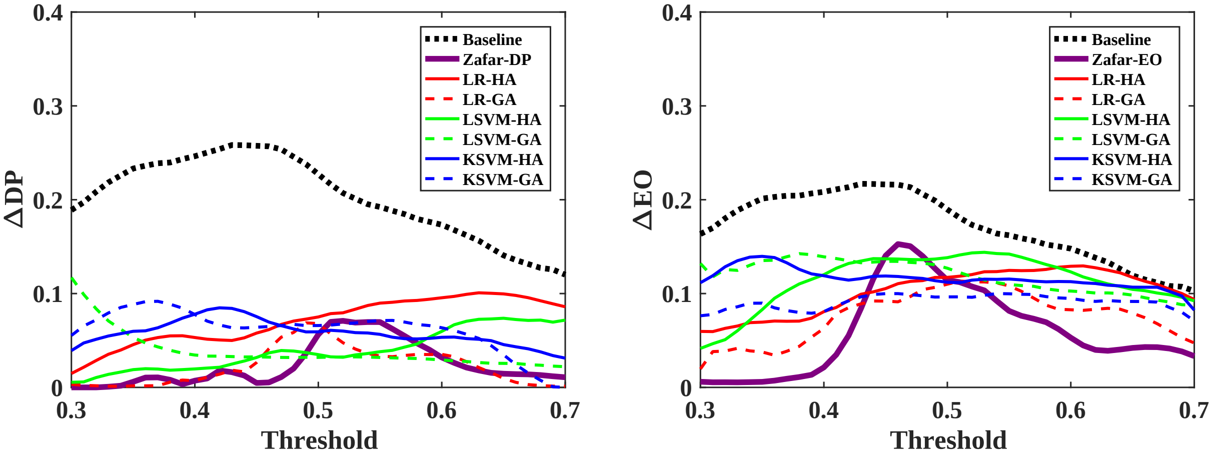
5.4 PARITY VARIATION VERSUS THRESHOLD

We conducted experiments on three types of classifiers (e.g., LR, LSVM, and KSVM) with 2 equalization approaches (e.g., HA, GA). In this section, we name the proposed classifiers as "{classifier type}-{equalization}". For instance, "LR-HA" refers to LR classifier with HA approach. We compare our proposed approaches with the prior works based on DP constraint (Zafar et al., 2017b) and EO constraint (Zafar et al., 2017a). We refer to them as "Zafar-DP" and "Zafar-EO" in the paper, respectively. The baseline refers to the unconstrained LR classifier. Note that all these methods do not use protected attributes to make predictions in the test stage. We show the variation of classification parity versus the decision threshold of the risk score for a given classifier. For fair comparison, we tune the parameter η to make the proposed classifiers have comparable performances to the prior works, in terms of prediction accuracy. Performance accuracy and the CV scores of ΔDP and ΔEO come from the classifiers with the default decision threshold.

To investigate the parity variation versus the decision threshold, we present the variation of the classification parity by changing the decision threshold ranging from 0.3 to 0.7. Table 2 presents the performances of the classifiers with DP constraint and Table 3 presents the performances of the classifiers with EO constraint. "Int." refers to the interval length of the parity variation and "STD" denotes the standard deviation. Figure 4 illustrates the variations of classification parity versus the risk score threshold ranging from 0.3 to 0.7. Less fluctuation and more flatness of the curve suggest that the classifier achieves higher threshold-invariant parity.

As we expect, the baseline classifier has the best performance in terms of accuracy (68.4%), but worst in CV scores of classification parity (ΔDP$=$ 0.225 and ΔEO$=$ 0.188). It spans a wider range and has a larger variance in parity when decision threshold ranges from 0.3 to 0.7. The variation curves in Figure 4 are also above the classifiers with fairness constraints. With the fairness constraints during the training, the CV scores are both reduced in Table 2 and 3 at the cost of the accuracy performances, and variation ranges and variances decrease to a lower level, indicating the better equitable decision performances among the groups in terms of classification parity. For the DP constraint in Table 2, the proposed classifiers have slightly better DP consistency against the change of decision threshold than Zafar-DP, according to the SD of ΔDP. For the EO constraint in Table 3, the proposed classifiers have a noticeable advantage in invariant EO against the change of decision threshold. Compared with Zafar-EO, the proposed classifiers have smaller interval ranges and smaller variances of ΔEO.

From Figure 4, we can observe that our proposed methods have minor fluctuations on the classification parity curves, compared with Zafar-DP (Zafar et al., 2017b) and Zafar-EO (Zafar et al., 2017a), respectively. Slightly modifying the decision threshold do not affect the classification parity to a large extent in the proposed classifiers. This suggests that our proposed constraints are effective



(a) (b)

Figure 4: Variations of classification parity versus decision threshold: (a) classifiers with the DP constraint and (b) classifiers with the EO constraint. The figures are best viewed in color.

in reducing the sensitivity of classifiers to the decision threshold in terms of classification parity. The noticeable drops in the CV scores in the three types of the classifiers demonstrate that our methods can be applied to a variety of classifiers.

5.5 DISCUSSIONS

From the experimental results, we can see that two proposed methods are effective in equalizing the risk distributions so as to achieve threshold invariant fairness. In this section, we discuss and summarize the difference between the two methods.

The HA method uses an N-bin histogram to compute the risk distribution. To make the histogram operation differentiable, a Gaussian kernel is used to approximate the accumulation process in each histogram bin. The HA method requires two hyperparameters: the number of bins N and the variance σ_c^2 in (12). Their settings influence the computational cost and the precision of the distribution approximation. The HA method computes the weight of every sample to each histogram bin via a Gaussian kernel and then sum up the weights in each bin to construct the risk distribution. The overall computational complexity of computing risk distribution in HA is $O(MN)$, where M is the number of training samples.

The GA method assumes the Gaussian distribution and uses only the statistics of the first and second moments (mean and variance) to characterize the risk distribution. The estimation of means and variances has a computational complexity of $O(M)$, which is lower than the HA method. Considering the similar performances between the two methods and the difference in computational complexities, we can see that the GA method is a preferred approach to achieve the equalization of risk distributions.

In addition, our proposed methods can be extended to the case of multiple protected attributes and other complex differentiable classifiers. For example, the proposed methods are compatible with neural network classifiers. To make our methods adaptive to the batch-based training of neural networks, we can modify gradient descent optimizer to batch-based optimizer (e.g., stochastic gradient descent) and calculate the risk distributions over one training batch instead of over the whole training set.

6 CONCLUSION

In this paper, we have introduced a novel fairness notion in machine learning models. Different from the prior definition of classification parity, the new notion aims to build fair classifiers that are not sensitive to the decision thresholds, and achieves this by equalizing the risk distributions among different groups. We perform distribution equalization using histogram approximation and Gaussian assumption, respectively, and incorporate them into logistic regression and SVM classifiers. Experimental results show that our fairness notion allows better control of the fairness level against the decision threshold of the classifier, and has broad compatibility to multiple types of machine learning classifiers.

References

Agarwal, A., Beygelzimer, A., Dudik, M., Langford, J., & Wallach, H. (2018). A reductions approach to fair classification.

In *International Conference on Machine Learning*, (pp. 60–69).

Angwin, J., Larson, J., Mattu, S., & Kirchner, L. (2016). Machine bias: There's software used across the country to predict future criminals and it's biased against blacks. *ProPublica*.

Barocas, S., & Selbst, A. D. (2016). Big data's disparate impact. *California Law Review*, *104*, 671.

Calders, T., & Verwer, S. (2010). Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, *21*(2), 277–292.

Calmon, F., Wei, D., Vinzamuri, B., Ramamurthy, K. N., & Varshney, K. R. (2017). Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems*, (pp. 3992–4001).

Corbett-Davies, S., & Goel, S. (2018). The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv preprint arXiv:1808.00023*.

Corbett-Davies, S., Pierson, E., Feller, A., Goel, S., & Huq, A. (2017). Algorithmic decision making and the cost of fairness. In *ACM International Conference on Knowledge Discovery and Data Mining*, (pp. 797–806).

Crowson, C. S., Atkinson, E. J., & Therneau, T. M. (2016). Assessing calibration of prognostic risk scores. *Statistical Methods in Medical Research*, *25*(4), 1692–1706.

Dieterich, W., Mendoza, C., & Brennan, T. (2016). *COMPAS risk scales: Demonstrating accuracy equity and predictive parity*. Northpointe Inc.

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science Advances*, *4*(1), eaao5580.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., & Zemel, R. (2012). Fairness through awareness. In *Innovations in Theoretical Computer Science Conference*, (pp. 214–226). ACM.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., & Venkatasubramanian, S. (2015). Certifying and removing disparate impact. In *ACM International Conference on Knowledge Discovery and Data Mining*, (pp. 259–268).

Gonzalez, R. C., Woods, R. E., & Eddins, S. L. (2001). *Digital Image Processing*. Pearson Education.

Hardt, M., Price, E., & Srebro, N. (2016). Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, (pp. 3315–3323).

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

Kimeldorf, G., & Wahba, G. (1971). Some results on Tchebycheffian spline functions. *Journal of Mathematical Analysis and Applications*, *33*(1), 82–95.

Larson, J., Mattu, S., Kirchner, L., & Angwin, J. (2016). How we analyzed the COMPAS recidivism algorithm. *ProPublica*.

Mehrabi, N., Morstatter, F., Saxena, N., Lerman, K., & Galstyan, A. (2019). A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*.

Pleiss, G., Raghavan, M., Wu, F., Kleinberg, J., & Weinberger, K. Q. (2017). On fairness and calibration. In *Advances in Neural Information Processing Systems*, (pp. 5680–5689).

Shalev-Shwartz, S., Singer, Y., Srebro, N., & Cotter, A. (2011). Pegasos: Primal estimated sub-gradient solver for SVM. *Mathematical Programming*, *127*(1), 3–30.

Zafar, M. B., Valera, I., Rodriguez, M. G., & Gummadi, K. P. (2017a). Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *International Conference on World Wide Web*, (pp. 1171–1180).

Zafar, M. B., Valera, I., Rogriguez, M. G., & Gummadi, K. P. (2017b). Fairness constraints: Mechanisms for fair classification. In *Artificial Intelligence and Statistics*, (pp. 962–970).

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., & Dwork, C. (2013). Learning fair representations. In *International Conference on Machine Learning*, (pp. 325–333).

A Relation of Equalized Odds and Risk Distributions

In this section, we show the derivation of (9) in the main paper. We denote $P_{a,y,\hat{Y}} := P(\hat{Y} = \hat{y} | A = a, Y = y)$ and $f_{a,y}(s) := f(s|A = a, Y = y)$ as the risk score of one sample and risk distributions over the group with protected attribute A and label Y, respectively. Given the decision threshold $t \in [0, 1]$, we have

$$\begin{aligned}\Delta\text{EO} &= \frac{1}{2}\Big(\big|P_{0,0,1} - P_{1,0,1}\big| + \big|P_{0,1,1} - P_{1,1,1}\big|\Big)\\ &= \frac{1}{2}\Big(\big|\int_t^1 f_{0,0}(s) - f_{1,0}(s)\mathrm{d}s\big| + \big|\int_t^1 f_{0,1}(s) - f_{1,1}(s)\mathrm{d}s\big|\Big)\\ &\leq \frac{1}{2}\Big(\int_t^1 \big|f_{0,0}(s) - f_{1,0}(s)\big|\mathrm{d}s + \int_t^1 \big|f_{0,1}(s) - f_{1,1}(s)\big|\mathrm{d}s\Big)\\ &\leq \frac{1}{2}(\epsilon_0 + \epsilon_1)(1 - t),\end{aligned} \tag{25}$$

where we define the upper bounds $\big|f_{0,0}(s) - f_{1,0}(s)\big| \leq \epsilon_0$ and $\big|f_{0,1}(s) - f_{1,1}(s)\big| \leq \epsilon_1, \forall s \in [0, 1]$. Also, we have

$$\begin{aligned}\Delta\text{EO} &= \frac{1}{2}\Big(\big|P_{0,0,1} - P_{1,0,1}\big| + \big|P_{0,1,1} - P_{1,1,1}\big|\Big)\\ &= \frac{1}{2}\Big(\big|\big(1 - P_{0,0,1}\big) - \big(1 - P_{1,0,0}\big)\big|\\ &\qquad + \big|\big(1 - P_{0,1,0}\big) - \big(1 - P_{1,1,0}\big)\big|\Big)\\ &= \frac{1}{2}\Big(\big|P_{1,0,0} - P_{0,0,0}\big| + \big|P_{1,1,0} - P_{0,1,0}\big|\Big)\\ &= \frac{1}{2}\Big(\big|\int_0^t f_{1,0}(s) - f_{0,0}(s)\mathrm{d}s\big| + \big|\int_0^t f_{1,1}(s) - f_{0,1}(s)\mathrm{d}s\big|\Big)\\ &\leq \frac{1}{2}\Big(\int_0^t \big|f_{1,0}(s) - f_{0,0}(s)\big|\mathrm{d}s + \int_0^t \big|f_{1,1}(s) - f_{0,1}(s)\big|\mathrm{d}s\Big)\\ &\leq \frac{1}{2}(\epsilon_0 + \epsilon_1)t.\end{aligned} \tag{26}$$

Combining (25) and (26), we have

$$\Delta\text{EO} \leq \frac{1}{2}(\epsilon_0 + \epsilon_1) \cdot \min\{t, 1 - t\} \leq \frac{1}{4}(\epsilon_0 + \epsilon_1). \tag{27}$$