

Intrinsic Fairness-Accuracy Tradeoffs under Equalized Odds

Meiyu Zhong Ravi Tandon
Department of Electrical and Computer Engineering
University of Arizona, Tucson, USA
E-mail: {meiyuzhong, tandonr}@arizona.edu

Abstract—With the growing adoption of machine learning (ML) systems in areas like law enforcement, criminal justice, finance, hiring, and admissions, it is increasingly critical to guarantee the fairness of decisions assisted by ML. In this paper, we study the tradeoff between fairness and accuracy under the statistical notion of equalized odds. We present a new upper bound on the accuracy (that holds for *any* classifier), as a function of the fairness budget. In addition, our bounds also exhibit dependence on the underlying statistics of the data, labels and the sensitive group attributes. We validate our theoretical upper bounds through empirical analysis on three real-world datasets: COMPAS, Adult, and Law School. Specifically, we compare our upper bound to the tradeoffs that are achieved by various existing fair classifiers in the literature. Our results show that achieving high accuracy subject to a low-bias could be fundamentally limited based on the statistical disparity across the groups.

I. INTRODUCTION

Machine learning-based solutions are increasingly being implemented across various sectors, including criminal justice, law enforcement, hiring, and admissions. These systems have demonstrated remarkable predictive capabilities. However, recent studies [1]–[3] indicate a significant downside to data-driven approaches: bias in decision-making. To address this problem, there is a vast amount of research focused on various concepts of fairness [4]–[6], which mainly falls into three categories: (1) Group Fairness [7]–[9] which requires that the subjects in the subgroups have an equal probability of being assigned to the same predicted class. (2) Individual Fairness [4], [6], [10] which requires that *similar individuals* (measured by a domain specific similarity metric) should be treated similarly. (3) Causality-based Fairness [6], [11]: which uses causality-based tools to design notions of fairness.

In this paper, we focus on group fairness (also known as statistical fairness). There are three types of statistical fairness notions which have been widely studied: demographic parity (DP) [3], [4]; equalized odds (EO) [5], [9] and predictive rate parity (PP) [5]. DP requires that the classifier’s decision be independent of the sensitive group attribute. The notion of DP however suffers from two drawbacks: first, when the sensitive group attribute is correlated with the class labels, this may rule out the perfect predictor (and hurt accuracy) [12]; second, the fairness notion of DP does not take into account the true label into account; if the distribution of data across subgroups is uneven, then enforcing DP may be unfair to those individuals which were worthy of a positive outcome. To avoid such disparity, the notion of Equalized Odds (EO) requires

the classifier’s prediction should be independent of sensitive attributes given the true class label. In addition, Predictive Rate Parity (PP) is defined as the condition where different groups have equal predictive values, meaning that the probability of a true positive (or negative) result is consistent across all groups.

The techniques for learning a fair classifier subject to one of the group fairness notions can be mainly divided into three categories: (1) Pre-processing methods [2], which focus on mitigating bias by altering the training data (e.g., by creating a more balanced or fair dataset) before it is utilized in the training phase. (2) In-processing methods [3], [13], which involve integrating fairness constraints directly into the model training process (for instance, via explicit fairness aware regularization). (3) Post-processing methods [14], which entail adjusting the model’s parameters after training. This technique involves fine-tuning the trained model to rectify any unfair biases that may have been introduced during the training process. However, experimental results [3], [5], [9], [15] have demonstrated, and recent theoretical evidence [16], [17] has further confirmed, that there is often a drop in accuracy when enforcing fairness constraints, when compared to unconstrained training.

Overview of recent work on Fairness-Accuracy Tradeoffs: The above observations motivate a theoretical treatment of the tradeoff between fairness and accuracy. Recent works [16]–[18] have studied this tradeoff for the case of demographic parity (DP), and have obtained bounds which quantify the minimal drop in accuracy as a function of the fairness budget. Additionally, Dutta et al. [19] employ mismatched hypothesis testing to demonstrate that there exist distributions for which there is no trade-off between fairness and accuracy. However, they also show that these trade-offs do exist in real-world datasets. Wang et al. [20] suggest that randomized prediction methods might more consistently meet equalized odds in classification (also see [21]–[24]). Another line of works [17], [25] study fair Bayes-optimal classifiers subject to equal opportunity (which is a related, but weaker notion than EO); the idea herein is to change the Bayes-optimal classifier by designing subgroup-specific decision thresholds to satisfy the fairness constraint. Another research direction, as explored in [26], concentrates on determining whether it’s possible to establish a bound on the accuracy of a given classifier relative to its allocated fairness budget. However, these bounds are intrinsically linked to the characteristics of the classifier itself

and do not directly give an insight to the fundamental fairness-accuracy tradeoff.

Main Contributions. In this paper, we focus on the problem of binary (0/1) classification subject to equalized odds (EO) fairness constraints and present a new upper bound on the accuracy as a function of the fairness budget. Our bounds are *classifier-independent* (i.e., they must hold for any classifier) and are determined by the underlying statistics of the data, labels, and sensitive groups. Our primary technique for deriving these bounds involves adapting Le Cam's bound [27], [28], which is traditionally used for binary classification problems; encompassing the Equalized Odds (EO) fairness constraints. The original Le Cam's bound is based on the total variation distance between two class distributions (i.e., $d_{TV}(P_0, P_1)$). Our EO constrained bound depends on the total variation distances within sensitive subgroups (i.e., $d_{TV}(P_0^a, P_1^a), d_{TV}(P_0^b, P_1^b)$, where a, b denotes the two subgroups), as well as the relative proportions of the subgroups. The extent of statistical discrepancy across the subgroups plays a critical role in influencing the behavior of our bound. As the subgroup discrepancy increases, our bound becomes tighter compared to the classical Le Cam's bound.

We also present experimental results using three real-world datasets: COMPAS, Adult, and the Law School dataset. We estimate our bounds for these datasets by employing an estimator for the total variation distance. Additionally, we compare our upper bounds with the trade-offs between fairness and accuracy achieved by various fair classifiers on these datasets.

II. PRELIMINARIES AND PROBLEM STATEMENT

We consider a supervised classification problem, where we are given a dataset of n users: $\{x_i, y_i, z_i\}_{i=1}^n$, where $x_i \in \mathcal{X} \subseteq \mathbb{R}^d$ denotes the set of features of i th training sample; $y_i \in \mathcal{Y} = \{0, 1\}$ represents the corresponding binary class label; $z_i \in \mathcal{Z} = \{a, b\}$ denotes the set of binary sensitive attributes of i th training sample, which depends on the dataset and underlying context. For instance, if we consider gender as a sensitive attribute (or subgroup), we could designate a to represent females and b to represent males. Consider a classifier f which maps from input data space to output labels: $f: \mathcal{X} \rightarrow \mathcal{Y}$, i.e., $f(x) \in \{0, 1\}$ denotes the classifier's decision for an input x .

Notation: For the scope of this paper, we use capital letters to represent the random variables/vectors (X, Y, Z) and use lowercase letters (x, y, z) to denote a specific realization. We next define the following quantities which will appear in our results:

- $\alpha = \mathbb{P}(Y = 1)$ (probability of the label $Y = 1$).
- $\beta = \mathbb{P}(Z = a)$ (probability of sensitive attribute $Z = a$).
- $\phi(x) = \mathbb{P}(Y = 1|x)$ (posterior probability of class 1)
- $P(x)$ represents the unconditional PDF of X .
- $P_0(x)$ (respectively, $P_1(x)$) represents the conditional PDF of $X|Y = 0$ (respectively, $X|Y = 1$).
- $P_0^a(x)$ (respectively, $P_0^b(x)$) represents the conditional PDF of $X|(Y = 0, Z = a)$ (respectively, $X|(Y = 0, Z = b)$). Similarly, $P_1^a(x)$ (respectively, $P_1^b(x)$) represents the

conditional PDF of $X|(Y = 1, Z = a)$ (respectively, $X|(Y = 1, Z = b)$).

Problem Statement From the scope of this paper, our goal is to explore tradeoff between fairness and accuracy for any classifier. We mainly focus on the group fairness notion: equalized odds, which is defined as follows.

Definition 1. (Approximate Equalized Odds) A binary classifier f satisfies ϵ_{EO} -Equalized Odds (EO) if $\Delta_{EO} \leq \epsilon_{EO}$, where, for $y \in \{0, 1\}$, $\Delta_{EO} :=$

$$\max_y |\mathbb{P}(f(X) = 1|Z=a, Y=y) - \mathbb{P}(f(X) = 1|Z=b, Y=y)|.$$

When $\epsilon_{EO} = 0$, the input X 's prediction $f(X)$ is conditionally independent of its sensitive attribute Z given the label Y (equivalent to the Markov chain $f(X) \rightarrow Y \rightarrow Z$) which corresponds to the notion of *perfect* equalized odds [5], [7]. Following prior works, [5], [13], [29], we focus on the approximate EO setting, $\Delta_{EO} \leq \epsilon_{EO}$, and ϵ_{EO} refers to the fairness budget.

Definition 2. (Total Variation (TV) Distance) Consider P and Q as two probability distributions over a common probability space Ω . Then, the total variation distance between them, denoted $d_{TV}(P, Q)$, is defined as follows

$$d_{TV}(P, Q) = \sup_{A \subseteq \Omega} |P(A) - Q(A)| = \frac{1}{2} \int_{\Omega} |P(x) - Q(x)| dx$$

III. MAIN RESULTS AND DISCUSSION

In this section, we present our main results on the fundamental trade-off between fairness constraints (subject to EO) and accuracy for binary classification. Our results are organized as follows. First, we present an upper bound on accuracy based on Le Cam's method without any fairness constraint. This bound is attainable by the Bayes optimal classifier for three cases: when the class distribution is either balanced $\alpha = 0.5$ or extremely unbalanced $\alpha = 0/1$. We then present the main result of this paper: a new *classifier-independent* upper bound on accuracy as a function of the EO budget (ϵ_{EO}), taking into account the underlying statistics of the data, labels, and the proportions of subgroups.

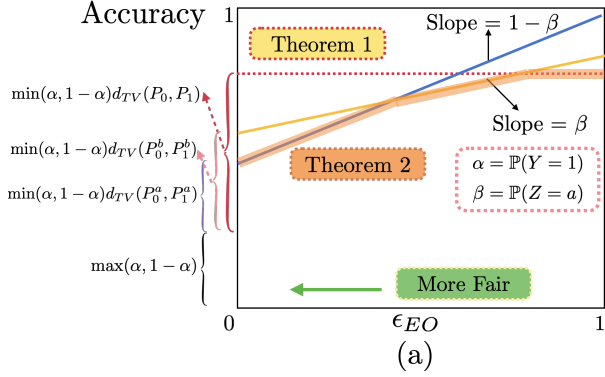
We first present an upper bound on the accuracy for any arbitrary binary classifier, which is given as follows:

Theorem 1. (Unconstrained Upper Bound on Accuracy) For any binary classifier f , its accuracy $\text{Acc}(f)$ satisfies $\text{Acc}(f) \leq \text{Acc}$, where Acc is given as follows:

$$\overline{\text{Acc}} = \max(1 - \alpha, \alpha) + \min(1 - \alpha, \alpha) \cdot d_{TV}(P_1, P_0). \quad (1)$$

Furthermore, the upper bound can be attained by the Bayes optimal classifier when $\alpha = 0, 0.5$ or 1 .

We can draw the following insights from the above result: for a fixed value of α , accuracy is directly proportional to $d_{TV}(P_1, P_0)$. On the other hand, for a fixed value of $d_{TV}(P_1, P_0)$, the lowest accuracy occurs when $\alpha = 0.5$, i.e., when both the classes 0/1 appear in a balanced manner. We next present the proof of the above result, which will allow us to contrast the corresponding bounds we obtain subject to equalized odds fairness constraints.



	COMPAS	Law School	Adult
$\alpha = \mathbb{P}(Y = 1)$	0.4425	0.5223	0.2448
$\beta = \mathbb{P}(Z = a)$	0.3418	0.7457	0.1384
$d_{TV}(P_0, P_1)$	0.5502	0.6110	0.6500
$d_{TV}(P_0^a, P_1^a)$	0.3116	0.4359	0.5512
$d_{TV}(P_0^b, P_1^b)$	0.3359	0.4216	0.6004
Theorem 2 ($\epsilon_{EO} = 0$)	0.6954	0.7236	0.8901
Theorem 2 ($\epsilon_{EO} = 0.05$)	0.7232	0.7432	0.9091
Theorem 1 (Unconstrained)	0.8009	0.8142	0.9143

(b)

Fig. 1: (a) Illustration of the relationship between Theorem 1 and Theorem 2, where the red dotted line represents the bound established in Theorem 1, while the orange fluorescent line depicts the minimum of the two functions T_1 and T_2 in Theorem 2, where T_1 and T_2 have slopes of $1 - \beta$ and β , respectively. (b) Dataset-related parameters and upper-bound-related parameters in the real world datasets: COMPAS, Adult and Law School dataset.

Proof. We expand the expression for accuracy of a binary classifier using total probability theorem as follows:

$$\begin{aligned} \text{Acc}(f) &= E_{(X,Y)}[\mathbb{1}(f(X) = Y)] = \mathbb{P}(f(X) = Y) \\ &= \mathbb{P}(Y = 1, f(X) = 1) + \mathbb{P}(Y = 0, f(X) = 0) \\ &= \alpha \mathbb{P}(f(X) = 1|Y = 1) + (1 - \alpha) \mathbb{P}(f(X) = 0|Y = 0) \end{aligned} \quad (2)$$

We can re-write the above expression for $\text{Acc}(f)$ in two different ways as follows, where $\text{Acc}(f)$

$$\begin{aligned} &\stackrel{(a)}{=} \alpha \mathbb{P}(f(X) = 1|Y = 1) + (1 - \alpha)(1 - \mathbb{P}(f(X) = 1|Y = 0)) \\ &\stackrel{(b)}{=} \alpha(1 - \mathbb{P}(f(X) = 0|Y = 1)) + (1 - \alpha) \mathbb{P}(f(X) = 0|Y = 0) \end{aligned}$$

We now consider two scenarios based on the value of α . In the first case, when $\alpha \leq 0.5$, we use the expression in (a) and upper bound it as follows:

$$\begin{aligned} \text{Acc}(f) &\leq 1 - \alpha + \alpha(\mathbb{P}(f(X) = 1|Y = 1) - \mathbb{P}(f(X) = 1|Y = 0)) \\ &\leq 1 - \alpha + \sup_f \alpha(\mathbb{P}(f(X) = 1|Y = 1) - \mathbb{P}(f(X) = 1|Y = 0)) \\ &\leq 1 - \alpha + \alpha \sup_f |\mathbb{P}(f(X) = 1|Y = 1) - \mathbb{P}(f(X) = 1|Y = 0)| \\ &= 1 - \alpha + \alpha \cdot d_{TV}(P_0, P_1). \end{aligned} \quad (3)$$

Similarly, for the case when $\alpha > 0.5$, we use the expression in (b) and bound it in a similar manner to arrive at

$$\text{Acc}(f) \leq \alpha + (1 - \alpha) \cdot d_{TV}(P_1, P_0) \quad (4)$$

Combining (3) and (4), we obtain the upper bound stated in the Theorem 1:

$$\text{Acc}(f) \leq \max(\alpha, 1 - \alpha) + \min(\alpha, 1 - \alpha) \cdot d_{TV}(P_1, P_0).$$

Let us now consider the Bayes optimal classifier:

$$f_{\text{Bayes}}(x) = \begin{cases} 1, & \text{if } \phi(x) \geq 1 - \phi(x) \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

where $\phi(x) = \mathbb{P}(Y = 1|X = x)$. The accuracy of Bayes optimal classifier can be readily computed as:

$$\text{Acc}(f_{\text{Bayes}}) = \mathbb{E}_X[\max(\phi(X), 1 - \phi(X))] \quad (6)$$

$$= \frac{1}{2} + \frac{1}{2} \int_x |\alpha \mathbb{P}(x|y = 1) - (1 - \alpha) \mathbb{P}(x|y = 0)| dx \quad (7)$$

It is straightforward to verify that the Bayes' classifier accuracy matches with our upper bound when $\alpha = 0.5, 0$ or 1 .

This completes the proof of Theorem 1. $\square$

We next present our main result, which is an upper bound on accuracy as a function of the fairness budget ϵ_{EO} .

Theorem 2. For any binary classifier which satisfies equalized odds ($\Delta_{EO} \leq \epsilon_{EO}$), its accuracy $\text{Acc}(f, \epsilon_{EO})$ satisfies $\text{Acc}(f, \epsilon_{EO}) \leq \overline{\text{Acc}}(\epsilon_{EO})$, where $\overline{\text{Acc}}(\epsilon_{EO})$ is given as follows:

$$\overline{\text{Acc}}(\epsilon_{EO}) = \max(1 - \alpha, \alpha) + \min(T_1, T_2), \quad (8)$$

where

$$T_1 \triangleq \min(1 - \alpha, \alpha) \cdot d_{TV}(P_1^b, P_0^b) + \beta \epsilon_{EO} \quad (9)$$

$$T_2 \triangleq \min(1 - \alpha, \alpha) \cdot d_{TV}(P_1^a, P_0^a) + (1 - \beta) \epsilon_{EO}. \quad (10)$$

We next present a sequence of remarks which give an operational interpretation of the above bound.

Remark 1. We note that the upper bound is a piece-wise linear function of the fairness budget ϵ_{EO} , and obtained as a minimum over two expressions related to T_1, T_2 as shown in Fig. 1(a). The expressions in T_1 depend upon the $d_{TV}(P_1^b, P_0^b)$, which reflect the difference in the class conditional probabilities for subgroup b. Similarly, the bound in T_2 depends on $d_{TV}(P_1^a, P_0^a)$, which reflects the difference in the class conditional probabilities for subgroup a. When $\epsilon_{EO} = 0$, i.e., perfect EO fairness constraint, the upper bound shows that the accuracy of any classifier will always be limited by the minimum of $d_{TV}(P_1^b, P_0^b)$ and $d_{TV}(P_1^a, P_0^a)$, i.e., the subgroup with the worst classification accuracy.

Remark 2. For the case when $\epsilon_{EO} > 0$, i.e., approximate EO fairness constraint, the bound is given by the minimum over two linear functions of ϵ_{EO} . The slopes of these two lines are given by $\beta = P(Z = a)$, and $1 - \beta = P(Z = b)$, i.e., relative proportion/size of the two subgroups shown in Fig. 1 (a). The interplay between $(\beta, 1 - \beta)$ and the subgroup wise statistical distances dictate the overall behaviour of the upper bound. For instance, suppose subgroup a is a minority group with $\beta \ll 1 - \beta$, then there always exists a threshold ϵ_{EO} , for which the classification accuracy will be dictated by the statistical distance corresponding to the majority subgroup b.

Remark 3. In the next sub-section, we discuss a methodology to estimate the upper bounds for real-world datasets by leveraging tools for statistical estimation of f -divergence (e.g., TV distance). In Fig. 1(b), we show the corresponding experimental results we obtained on three datasets (COMPAS, Law-School admissions and Adult Income prediction). Specifically, for each of these datasets, we estimate the corresponding values of $\beta, \alpha, d_{TV}(P_0, P_1), d_{TV}(P_0^a, P_1^a)$ and $d_{TV}(P_0^b, P_1^b)$ from the datasets and then show the bounds obtained from the two Theorems. We show the corresponding bounds for $\epsilon_{EO} = 0$ and for $\epsilon_{EO} = 0.05$ (more experiments are presented in the next Section). Specifically, we can observe the estimated bound from Theorem 1 (corresponding to unconstrained classifiers) is always larger than those provided by Theorem 2 (the upper bound on accuracy with EO constraints), as well as the interplay between the statistical properties, namely $\beta, 1 - \beta$ and the statistical disparity across the two subgroups.

We next present the proof of Theorem 2.

Proof. To simplify the notation used in the proof, we denote $\eta = \mathbb{P}(Z = a|Y = 1)$ and denote $\gamma = \mathbb{P}(Z = a|Y = 0)$. We start by the definition of the total accuracy of the binary classifier using total probability theorem as follows:

$$\begin{aligned} \text{Acc}(f) &= E_{(X,Y)}[\mathbb{1}(f(X) = Y)] = \mathbb{P}(f(X) = Y) \\ &= \alpha \mathbb{P}(f(X) = 1|Y = 1) + (1 - \alpha) \mathbb{P}(f(X) = 0|Y = 0) \quad (11) \\ &\stackrel{(a)}{=} \alpha \mathbb{P}(f(X) = 1|Y = 1) + (1 - \alpha)(1 - \mathbb{P}(f(X) = 1|Y = 0)) \\ &\stackrel{(b)}{=} \alpha(1 - \mathbb{P}(f(X) = 0|Y = 1)) + (1 - \alpha) \mathbb{P}(f(X) = 0|Y = 0) \end{aligned}$$

Then we further apply total probability theorem on $\mathbb{P}(f(X) = 1|Y = 1)$ with respect to Z in (a). Therefore, we can express $\mathbb{P}(f(X) = 1|Y = 1)$ as follows:

$$\eta \mathbb{P}(f(X) = 1|Y = 1, Z = a) + (1 - \eta) \mathbb{P}(f(X) = 1|Y = 1, Z = b)$$

Similarly, we can also take the advantage of total probability theorem on $\mathbb{P}(f(X) = 1|Y = 0)$ with respect to Z in (a), where $\mathbb{P}(f(X) = 1|Y = 0)$ can be written as:

$$\gamma \mathbb{P}(f(X) = 1|Y = 0, Z = a) + (1 - \gamma) \mathbb{P}(f(X) = 1|Y = 0, Z = b)$$

Since the classifier satisfies the fairness constraints $\Delta_{EO} \leq \epsilon_{EO}$, we have:

$$\begin{aligned} |\mathbb{P}(f(X) = 1|Y = 1, Z = a) - \mathbb{P}(f(X) = 1|Y = 1, Z = b)| &\leq \epsilon_{EO} \\ |\mathbb{P}(f(X) = 1|Y = 0, Z = a) - \mathbb{P}(f(X) = 1|Y = 0, Z = b)| &\leq \epsilon_{EO}. \end{aligned}$$

By incorporating the above inequalities, we can bound $\mathbb{P}(f(X) = 1|Y = 1)$ and $\mathbb{P}(f(X) = 1|Y = 0)$ as follows:

$$\mathbb{P}(f(X) = 1|Y = 1) \leq \eta \epsilon_{EO} + \mathbb{P}(f(X) = 1|Y = 1, Z = b) \quad (12)$$

$$\mathbb{P}(f(X) = 1|Y = 0) \geq -\gamma \epsilon_{EO} + \mathbb{P}(f(X) = 1|Y = 0, Z = b). \quad (13)$$

By plugging (12) and (13) into (a), the total accuracy can be bounded by:

$$\begin{aligned} \text{Acc}(f) &\leq (1 - \alpha) + \alpha \mathbb{P}(f(X) = 1|Y = 1, Z = b) \\ &\quad - (1 - \alpha) \mathbb{P}(f(X) = 1|Y = 0, Z = b) + \beta \epsilon_{EO}, \quad (14) \end{aligned}$$

where $\beta = \alpha \eta + (1 - \alpha) \gamma = P(Z = a)$. Similar to the proof of Theorem 1, we next consider two cases with respect to α . If $\alpha \leq 0.5$, then (14) can be further upper bounded by taking

a supremum over f . We thus arrive at the following:

$$\begin{aligned} \text{Acc}(f) &\leq (1 - \alpha) + \beta \epsilon_{EO} + \\ &\alpha \cdot \sup_f |\mathbb{P}(f(X) = 1|Y = 1, Z = b) - \mathbb{P}(f(X) = 1|Y = 0, Z = b)| \\ &= (1 - \alpha) + \alpha d_{TV}(P_1^b, P_0^b) + \beta \epsilon_{EO}. \quad (15) \end{aligned}$$

When $\alpha > 0.5$, we focus on (b), where we derive the upper bound subject to accuracy following the same steps above as (a). we arrive at the upper bound:

$$\text{Acc}(f) \leq \alpha + (1 - \alpha) \cdot d_{TV}(P_1^b, P_0^b) + \beta \epsilon_{EO}. \quad (16)$$

To this end, combining two inequalities (15) and (16), we have the upper bound of accuracy as a function of ϵ_{EO} with respect to the subgroup $Z = b$ and the proportion of another subgroup $Z = a$, which can be compactly written as:

$$\max(\alpha, 1 - \alpha) + \min(\alpha, 1 - \alpha) \cdot d_{TV}(P_1^b, P_0^b) + \beta \epsilon_{EO}$$

Note that we can also bound $\mathbb{P}(f(X) = 1|Y = 1)$, $\mathbb{P}(f(X) = 1|Y = 0)$ and $\mathbb{P}(f(X) = 0|Y = 1)$, $\mathbb{P}(f(X) = 0|Y = 0)$ by another subgroup $Z = a$ in (12) and (13). By doing so, we arrive at another upper bound of accuracy as a function of ϵ_{EO} with respect to the subgroup $Z = a$ and the proportion of another subgroup $Z = b$, which can be expressed as:

$$\max(\alpha, 1 - \alpha) + \min(\alpha, 1 - \alpha) \cdot d_{TV}(P_1^a, P_0^a) + (1 - \beta) \epsilon_{EO}$$

Combining the two upper bounds, we arrive at Theorem 2. $\square$

Estimation of Upper Bounds. We next describe a methodology for estimating the upper bounds for real-world datasets. For Theorem 1 and Theorem 2, estimating the fraction of the label or sensitive group (i.e., α, β) are quite straightforward. In addition, in order to obtain estimates of these bounds, we need an estimate of $d_{TV}(P_1, P_0)$, $d_{TV}(P_1^a, P_0^a)$, and $d_{TV}(P_1^b, P_0^b)$. To this end, we can leverage the fact that TV distance between two distributions is a special case of F -divergence, which is known to admit a variational representation [31], [32] expressed as follows:

$$D_f(P \parallel Q) = \sup_{T(\cdot)} E_{X \sim P}[T(X)] - E_{X \sim Q}[f^*(T(X))], \quad (17)$$

where the function $f^*(t) = \sup_{x \in \text{dom}_f} \{xt - f(x)\}$ denotes the convex conjugate (also known as the Fenchel conjugate) of the function f . The above variational representation involves a supremum over all possible functions $T(\cdot)$. We can obtain an estimate for TV distance by replacing the supremum over a restricted class of functions. Specifically, if we use a parametric model T_θ , (e.g., a neural network) with parameters θ , then taking the supremum over the parameters θ yields a lower bound on F -divergence in (17) as stated next:

$$D_f(P \parallel Q) \geq \sup_{\theta} E_{X \sim P}[T_\theta(X)] - E_{X \sim Q}[f^*(T_\theta(X))]. \quad (18)$$

We use the above variational lower bound to estimate the TV distance as described next. Take the $d_{TV}(P_1, P_0)$ as an example, we need to estimate TV distance between joint distributions of feature X in different class. The variational lower bound on F -divergence in (18) can then be estimated as:

$$\max_{\theta} \frac{1}{M} \left(\sum_{m=1}^M T_\theta(X_1^{(m)}) - \sum_{m=1}^M f^*(T_\theta(X_0^{(m)})) \right), \quad (19)$$

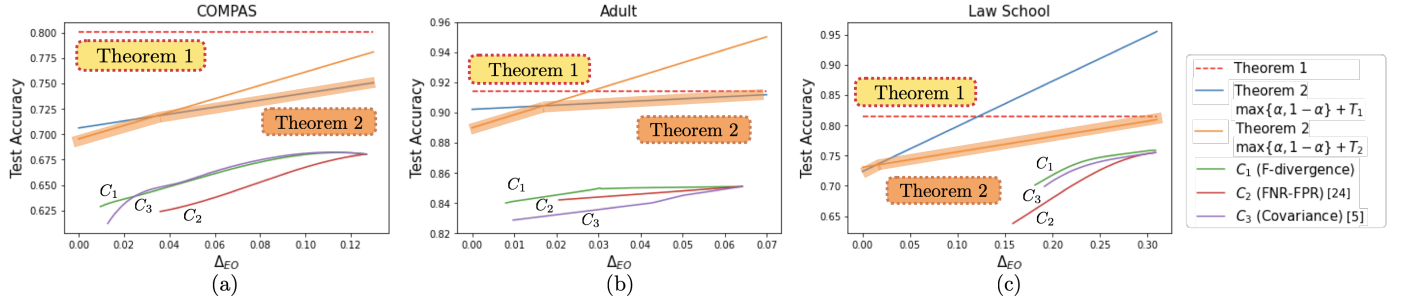


Fig. 2: Comparison of the upper bound of the test accuracy in real world datasets: (a) COMPAS dataset, (b) Adult dataset, (c) Law School dataset under three fair classifiers (C_1 [15], C_2 [30] and C_3 [5]). We can observe that Theorem 2 is consistently tighter than the unconstrained upper bound (Theorem 1), and Theorem 2 provides the tightly upper bound on the tradeoffs achieved by three classifiers for three real world datasets.

where in (19), we have replaced the expectation operators with their empirical estimates, and $\{X_1^{(m)}\}$ (respectively, $\{X_0^{(m)}\}$) denote i.i.d. samples drawn from the distribution P_1 (respectively, P_0). The consistency and convergence of the above estimator to the true divergence has been studied in [31] under some mild assumptions. For our experiments, we modeled T_θ using two-layer neural networks, each with 10 hidden nodes and followed by a sigmoid non-linearity activation layer.

IV. EXPERIMENTS

In this section, we show the experimental results to verify the tightness of our theoretical bounds. We consider three real world datasets: COMPAS, Adult and Law school admission dataset. We describe the dataset as follows: a) *COMPAS Dataset*: This dataset consists of data from $N = 7,214$ users ($N_{train} = 5,049$, $N_{test} = 2,165$), with 10 features (including age, prior criminal history, charge degree etc.) which are used for predicting the risk of recidivism in the next two years. b) *Adult Dataset*: This dataset includes income related data with 14 features (i.e., age, work class, occupation, education etc.) of $N = 45,222$ users ($N_{train} = 32,561$, $N_{test} = 12,661$) to predict whether the income of a person exceeds a threshold (e.g., \$50k) in a year. c) *Law School Dataset*: This dataset includes the admission related data with 7 features (LSAT score, gender, undergraduate GPA etc.) of $N = 4,862$ applicants ($N_{train} = 3,403$, $N_{test} = 1,459$) to predict the likelihood of passing the bar. For all the above datasets, we use race as the sensitive attribute. Specifically we consider the situation when $|Z| = 2$, $Z \in \{C, O\}$, where $C = \text{"Caucasian"}$ or $O = \text{"Other race"}$, corresponding to two groups.

For the methodology of training a fair classifier, we applied three in-processing mechanisms: (a) Zafar et al. [5] (C_3 (Covariance)) employ the covariance between sensitive attributes and the signed distance from misclassified data's feature vectors to the classifier decision boundary as a regularization term. (b) Bechavod et al. [30] (C_2 (FNR-FPR)) use the differences in False Negative Rate (FNR) and False Positive Rate (FPR) across subgroups as the regularization term. (c) Zhong et al. [15] (C_1 (F-divergence)) propose the F-divergence between the

conditional probability of predictions among subgroups as the regularization term. For the above mechanisms, they added fairness constraints (covariance, FNR-FPR, F-divergence) as a regularization in the loss function to learn a fair classifier subject to equalized odds.

Fig 2 shows the corresponding tradeoffs achieved by the three fair classifiers as the budget ϵ_{EO} is increased; we also show the upper bounds of Theorems 1 and 2 as a function of ϵ_{EO} . Notably, our upper bound incorporating EO constraints (Theorem 2) is tighter compared to the accuracy upper bound without fairness constraints (Theorem 1). Additionally, it is observed that Theorem 2 closely approximates the upper bound of the classifier's test accuracy under varying EO constraints, until it aligns with Theorem 1. Our experimental findings reinforce the validity of our theorems, demonstrating a tight correlation between the trade-offs in fairness and accuracy. These results not only align with the trends predicted by our theorems but also underscore the practical applicability of the theoretical framework in diverse real-world scenarios.

V. CONCLUSION

We presented a new upper bound on accuracy for binary classification subject to equalized odds, where the fairness budget is measured by ϵ_{EO} . Our results show that in addition to the fairness budget, relative subgroup sizes ($\beta, 1 - \beta$) as well as the statistical differences across subgroups (measured by $d_{TV}(P_1^a, P_0^a)$, $d_{TV}(P_1^b, P_0^b)$) impose a fundamental limit on the accuracy. We also validated these bounds using empirical estimation of TV-distance and compared them with the tradeoffs achieved by various fair classifiers. There are several directions for future work including generalization to other notions of statistical fairness (such as predictive parity); furthermore, it would be interesting to use the upper bounds as a guideline for designing fair classifiers (such as design of group-wise thresholding rules which maximize accuracy).

ACKNOWLEDGMENT

This work was supported by NSF grants CCF 2100013, CNS 2209951, CCF 1651492, CNS 2317192, CNS 1822071 and U.S. Department of Energy, Office of Science, Office of Advanced Scientific Computing under Award Number DE-SC-ERKJ422, and by NIH through Award 1R01CA261457-01A1.

REFERENCES

- [1] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, "A survey on bias and fairness in machine learning," *ACM Computing Surveys (CSUR)*, vol. 54, no. 6, pp. 1–35, 2021.
- [2] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork, "Learning fair representations," in *International Conference on Machine Learning*, pp. 325–333, PMLR, 2013.
- [3] M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi, "Fairness constraints: Mechanisms for fair classification," in *Artificial Intelligence and Statistics*, pp. 962–970, PMLR, 2017.
- [4] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, "Fairness through awareness," in *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, pp. 214–226, 2012.
- [5] M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. 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, 2017.
- [6] M. J. Kusner, J. Loftus, C. Russell, and R. Silva, "Counterfactual fairness," in *Advances in Neural Information Processing Systems* (I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, eds.), vol. 30, Curran Associates, Inc., 2017.
- [7] S. Barocas and A. D. Selbst, "Big data's disparate impact," *Calif. L. Rev.*, vol. 104, p. 671, 2016.
- [8] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian, "Certifying and removing disparate impact," in *proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268, 2015.
- [9] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in Neural Information Processing Systems*, vol. 29, 2016.
- [10] M. Yurochkin, A. Bower, and Y. Sun, "Training individually fair ML models with sensitive subspace robustness," in *International Conference on Learning Representations*, 2020.
- [11] N. Kilbertus, M. Rojas Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf, "Avoiding discrimination through causal reasoning," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [12] S. Barocas, M. Hardt, and A. Narayanan, *Fairness and machine learning: Limitations and opportunities*. MIT Press, 2023.
- [13] J. Cho, G. Hwang, and C. Suh, "A fair classifier using mutual information," in *2020 IEEE International Symposium on Information Theory (ISIT)*, pp. 2521–2526, IEEE, 2020.
- [14] G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger, "On fairness and calibration," *Advances in Neural Information Processing Systems*, vol. 30, 2017.
- [15] M. Zhong and R. Tandon, "Learning fair classifiers via min-max f-divergence regularization," in *2023 59th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1–8, IEEE, 2023.
- [16] H. Zhao and G. J. Gordon, "Inherent tradeoffs in learning fair representations," *The Journal of Machine Learning Research*, vol. 23, no. 1, pp. 2527–2552, 2022.
- [17] X. Zeng, E. Dobriban, and G. Cheng, "Bayes-optimal classifiers under group fairness," *arXiv preprint arXiv:2202.09724*, 2022.
- [18] A. K. Menon and R. C. Williamson, "The cost of fairness in binary classification," in *Conference on Fairness, accountability and transparency*, pp. 107–118, PMLR, 2018.
- [19] S. Dutta, D. Wei, H. Yueksel, P.-Y. Chen, S. Liu, and K. Varshney, "Is there a trade-off between fairness and accuracy? a perspective using mismatched hypothesis testing," in *International Conference on Machine Learning*, pp. 2803–2813, PMLR, 2020.
- [20] H. Wang, L. He, R. Gao, and F. Calmon, "Aleatoric and epistemic discrimination: Fundamental limits of fairness interventions," in *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.
- [21] Z. Tang and K. Zhang, "Attainability and optimality: The equalized odds fairness revisited," in *Conference on Causal Learning and Reasoning*, pp. 754–786, PMLR, 2022.
- [22] J. S. Kim, J. Chen, and A. Talwalkar, "Fact: A diagnostic for group fairness trade-offs," in *International Conference on Machine Learning*, pp. 5264–5274, PMLR, 2020.
- [23] F. Hamman and S. Dutta, "Demystifying local and global fairness trade-offs in federated learning using partial information decomposition," *arXiv preprint arXiv:2307.11333*, 2023.
- [24] S. Sabato and E. Yom-Tov, "Bounding the fairness and accuracy of classifiers from population statistics," in *International Conference on Machine Learning*, pp. 8316–8325, PMLR, 2020.
- [25] E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil, "Leveraging labeled and unlabeled data for consistent fair binary classification," *Advances in Neural Information Processing Systems*, vol. 32, 2019.
- [26] H. Zhao, A. Coston, T. Adel, and G. J. Gordon, "Conditional learning of fair representations," *arXiv preprint arXiv:1910.07162*, 2019.
- [27] L. Le Cam, "On the asymptotic theory of estimation and testing hypotheses," in *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Contributions to the Theory of Statistics*, vol. 3, pp. 129–157, University of California Press, 1956.
- [28] B. Yu, "Assouad, Fano, and Le Cam," in *Festschrift for Lucien Le Cam: research papers in probability and statistics*, pp. 423–435, Springer, 1997.
- [29] J. Cho, G. Hwang, and C. Suh, "A fair classifier using kernel density estimation," *Advances in Neural Information Processing Systems*, vol. 33, pp. 15088–15099, 2020.
- [30] Y. Bechavod and K. Ligett, "Learning fair classifiers: A regularization-inspired approach," *Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)*, 2017.
- [31] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *IEEE Transactions on Information Theory*, vol. 56, no. 11, pp. 5847–5861, 2010.
- [32] X. Nguyen, M. J. Wainwright, and M. Jordan, "Estimating divergence functionals and the likelihood ratio by penalized convex risk minimization," *Advances in Neural Information Processing Systems*, vol. 20, 2007.