2023 59th Annual Allerton Conference on Communication, Control, and Computing (Allerton) | 979-8-3503-2814-1/23/$31.00 ©2023 IEEE | DOI: 10.1109/Allerton58177.2023.10313425

Learning Fair Classifiers via Min-Max F-divergence Regularization

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

***Abstract*—As machine learning (ML) based systems are adopted in domains such as law enforcement, criminal justice, finance, hiring and admissions, ensuring the fairness of ML aided decision-making is becoming increasingly important. In this paper, we focus on the problem of fair classification, and introduce a novel min-max F-divergence regularization framework for learning fair classification models while preserving high accuracy.**

Our framework consists of two trainable networks, namely, a classifier network and a bias/fairness estimator network, where the fairness is measured using the statistical notion of F-divergence. We show that F-divergence measures possess convexity and differentiability properties, and their variational representation makes them widely applicable in practical gradient based training methods. The proposed framework can be readily adapted to multiple sensitive attributes and for high dimensional datasets. We study the F-divergence based training paradigm for two types of group fairness constraints, namely, demographic parity and equalized odds. We present a comprehensive set of experiments for several real-world data sets arising in multiple domains (including COMPAS, Law Admissions, Adult Income, and CelebA datasets).

To quantify the fairness-accuracy tradeoff, we introduce the notion of fairness-accuracy receiver operating characteristic (FA-ROC) and a corresponding *low-bias* FA-ROC, which we argue is an appropriate measure to evaluate different classifiers. In comparison to several existing approaches for learning fair classifiers (including pre-processing, post-processing and other regularization methods), we show that the proposed F-divergence based framework achieves state-of-the-art performance with respect to the trade-off between accuracy and fairness.

***Index Terms*—Fair Machine Learning, Regularization.**

I. Introduction

Machine learning based solutions are being increasingly deployed and adopted in various sectors of society, such as criminal justice, law enforcement, hiring and admissions. Despite their impressive predictive performance, there is a large body of recent evidence [1]–[3] which shows a flip side of using data driven solutions: bias in decision making, which is often attributed to the inherent bias present in training data. For instance, in criminal justice, risk assessment algorithms are often used to assess the risk of recidivism (reoffence), which is then used together with human input for decision making [4]. A classic example is that of Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) algorithm [4] which measures the recidivism risk and is used by judges for pretrial detention and release decisions. It was shown in [4] that COMPAS often falsely predicts a higher risk for some racial groups (specifically, african americans) compared to others. Another prominent example of bias with respect to gender is when a job advertisement system tends to show less STEM related job advertisements to women [5]. With the proliferation of data driven algorithms, ensuring fairness becomes crucial in the process of designing ML based decision making systems.

Notions of Fairness. There is a vast literature on the notions of fairness [6]–[8] which mainly falls into three categories: (1) Group Fairness [9]–[11] which requires that the subjects in the protected and unprotected groups have an equal probability of being assigned to the positive predicted class. (2) Individual Fairness [6], [8], [12] which requires that *similar individuals* (measured by a domain specific similarity metric) should be treated similarly. (3) Causality-based Fairness [8], [13]: which requires that using causality-based tools to design fair algorithms. In this paper, we consider group fairness, and focus on the notions of demographic parity (DP) and equalized odds (EO). The techniques for achieving group fairness can be mainly divided into (1) Pre-processing methods [2], which reduce bias by processing the training data before being used for training; (2) In-processing methods [3], [14], which add fairness constraints via regularization in the training process. (3) Post-processing methods [15], which appropriately modify the model parameters post training.

Related Work on Learning Fair Classifiers. The prominent in-processing method for learning fair classifiers is via regularization methods, where the key idea is to train a classifier which minimizes the classification error regularized by a bias penalty (which measures the discrepancy of the classifier across population sub groups). We remark that there have been several other approaches on regularization based training for fair classification, which include: a) using a covariance proxy [3] to measure the bias between predictions and sensitive attributes. Unfortunately, ensuring small correlation does not necessarily satisfy the stronger requirement of statistical independence.; b) kernel density estimation (KDE) based methods which first estimate conditional probability distribution of classifier predictions for each population group and use these as a fairness regularizer [16]. KDE methods are appropriate when the data dimensionality is relatively small but are not

This work was supported by NSF grants CCF 2100013, CAREER 1651492, CNS 2209951 and CNS 2317192.

scalable for high dimensional problems; c) another approach is to balance the TPR and FPR (true- and false-positive rates) across population sub-groups while training [17]; and d) measuring bias by the mean of Hirschfeld-Gebelein-Rényi (HGR) Maximum Correlation Coefficient [18] or mutual information [14] between predictions and sensitive attributes.

Admittedly, there are numerous choices for adding fairness constraints, and this opens up the following key questions: a) What is the optimal choice of a fairness regularization for a given notion of fairness, as well as the dataset and sensitive attributes? b) How does the regularization procedure impact the resulting tradeoff between accuracy and bias? c) How should one design a flexible procedure for learning fair classifiers which can work for high-dimensional datasets and is compatible with gradient based optimization?

Main Contributions. To deal with the above challenges, we propose a general min-max F-divergence regularization framework for learning fair classifiers. F-divergence, denoted by $D_f(P||Q)$ measures the difference between two probability distributions P, Q and different choices of the function $f(\cdot)$ lead to well-known divergence measures, such as KL divergence, Hellinger distance and Total Variation (TV) distance. Specifically, we propose to measure the bias using F-divergence between the classifier probability distributions across protected and unprotected groups. We next summarize the main contributions of this paper.

- By leveraging the variational representation of F-divergence, we cast the training process as a min-max optimization problem, which is suitable for commonly used gradient based optimization methods. The flexibility of the framework is two-fold: a) it can be readily applied in high-dimensional datasets, and b) by varying the choice of f, one can test and validate different proxies of achieving fairness within a single rubric.
- To quantify the fairness-accuracy tradeoff, we introduce the notion of fairness-accuracy receiver operating characteristic (FA-ROC) and also provide some interesting theoretical properties when using Total Variation distance as the measure of bias. Within this context, we also introduce the notion of *low-bias* FA-ROC, which we argue is an appropriate measure to evaluate different classifiers.
- We present a comprehensive set of results on multiple real world datasets (namely, COMPAS, Adult Census, Law School admissions and CelebA), and show the superiority of the proposed approach versus existing regularization, pre- and post-processing methods as discussed above. As an example, for the Adult census dataset, F-divergence regularization leads to $\approx$ 13% increase in FA-AUC (area under the curve) compared to the state-of-the-art regularization, pre- and post-processing methods for Demographic parity (we achieve a gain of $\approx$ 10% in FA-AUC for Equalized odds). For the high dimensional dataset (CelebA), our method consistently achieves better performance and receives a gain of 6% w.r.t EO constraints in the Low-bias region.

II. Preliminaries on Fair Classification

We consider a supervised classification problem, where we are given a dataset of N users: $\{X_n,\ Y_n,\ Z_n\}_{n=1}^N$, where X_n denotes the set of features of user n; Y_n represents the true label of user n; Z_n denotes the set of sensitive attributes of user n, which depends on the dataset and underlying context. For instance, in predicting recidivism risk in criminal justice, X_n includes features such as prior criminal history, demographic information, charge (type of crime); Z_n represents sensitive attributes, for instance, race or gender[1]; Y_n denotes ground truth like whether a user will re-offend in two years. Our goal is to build a fair binary classifier[2], which yields the estimate of the probability of the true label defined as follows:

$$\pi(\hat{Y}|X) \triangleq \begin{cases} \pi(0|x) = P(\hat{Y}=0|X=x) \\ \pi(1|x) = P(\hat{Y}=1|X=x). \end{cases} \tag{1}$$

Note that we do not use sensitive attributes Z as an input to the classifier. However, it is well known [8] that excluding the sensitive attributes alone does not necessarily lead to a fair classifier due to possible correlation between the sensitives attributes and features. For the scope of this paper, we focus on two statistical notions of group fairness: demographic parity (DP) and equalized odds (EO). Before we present group fairness notions, we first introduce the definition of F-divergence:

Definition 1. *(F-divergence) Let function $f : \mathbb{R}_+ \to \mathbb{R}$ be a convex, lower-semicontinuous function satisfying $f(1) = 0$. Given two probability distributions P and Q on a measurable space$(\mathcal{X}, \mathcal{F})$, F-divergence is defined as:*

$$D_f(P \parallel Q) = E_Q\left[f\left(\frac{dP}{dQ}\right)\right]$$

For instance, when $f(x) = x\log(x)$, this reduces to the KL divergence; $f(x) = (x-1)^2$ corresponds to χ^2 divergence; $f(x) = (1-\sqrt{x})^2$ corresponds to Squared Hellinger (SH) distance. Next, we now define the notion of Demographic Parity (DP) as follows:

Definition 2. *(Demographic Parity) A binary classifier π satisfies Demographic Parity (DP) if its prediction $\hat{Y}$ is independent of the sensitive attribute Z:*

$$\pi(1|Z=i) = \pi(1|Z=j), \quad \forall i \neq j.$$

Following from previous works [7], [15], [16], the standard measurement of DP is the difference between the conditional output probability distribution of the classifier given sensitive group i and j:

$$\Delta_{DP} := \sum_{i \neq j} |\pi(1|Z=i) - \pi(1|Z=j)|. \tag{2}$$

Note that our notion of Δ_{DP} is the same as Total Variation

[1]More generally, Z_n can take values from a discrete set, i.e., $Z_n \in \mathcal{G}$ and our formulation allows for the number of sensitive groups, i,e,. $|\mathcal{G}| \geq 2$ to be any finite number.

[2]We note that while our discussion in the paper is for binary classification, the proposed framework can be readily adopted for multi-class settings, as we discuss later in this section.

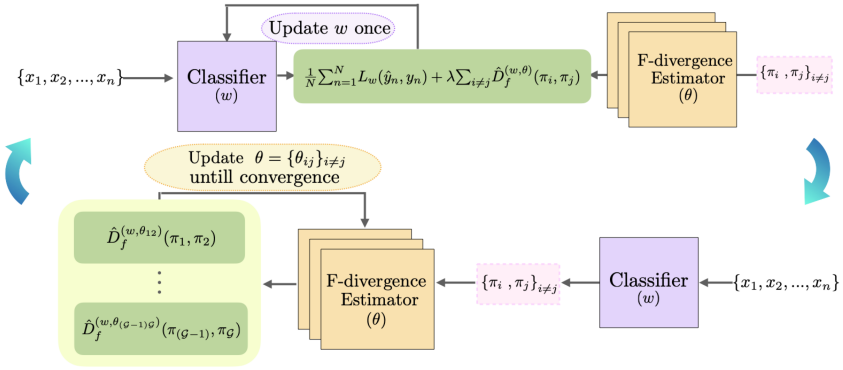


Fig. 1: Schematic of the Min-Max F-divergence regularization framework for fair training. The classifier (parameterized by w) is trained to minimize the regularized objective (containing classification loss + F-divergence regularization term). The estimator (parameterized by θ) estimates the F-divergence between distribution π_i and distribution π_j, measuring the bias in classification across groups (i, j). For DP, the distribution π_i for a group i is given by $\pi_i \sim \pi(\hat{Y}\,|Z = i)$, whereas, for EO, $\pi_i \sim \pi(\hat{Y}\,|Z = i, Y = 1)$. The two networks are trained alternatively, where for a fixed classifier, the F-divergence estimation is performed using a maximization problem leveraging the variational representation of F-divergence.

distance. For the special case when $\Delta_{DP} = 0$, this reduces to the notion of *perfect* demographic parity [9]. We note that for the case of multi-class classification, the above notion generalizes by considering $\sum_{\hat{y}} \sum_{i \neq j} |\pi(\hat{Y} = \hat{y}|Z = i) - \pi(\hat{Y} = \hat{y}|Z = j)|$. In a similar manner, we can define the notion of equalized odds as follows:

Definition 3. *(Equalized Odds) A binary classifier π satisfies Equalized Odds (EO) if its prediction $\hat{Y}$ is conditionally independent of its sensitive attribute Z given the label Y.*

$$\pi(1|Z = i, Y = y) = \pi(1|Z = j, Y = y),$$

where $\forall i \neq j$ and $y \in \{0, 1\}$. Same as above, we define the standard measurement of EO as follows:

$$\Delta_{EO} := \sum_{y} \sum_{i \neq j} |\pi(1|Z = i, Y = y) - \pi(1|Z = j, Y = y)|. \quad (3)$$

Motivation for F-divergence based Regularization– One prominent approach for learning fair classifiers is that of *fairness regularization*, i.e., adding a penalty term in the training loss function, which acts as a proxy to capture the fairness constraints (either DP or EO). In this work, we propose to use F-Divergence (as defined above) as the fairness regularization term in the loss function. F-divergence family has natural benefits like convexity and differentiability, which makes it an ideal candidate for gradient based optimization algorithms. In addition, as we discussed in the introduction, compared to other approaches, such as correlation/covariance between sensitive attributes and classifier outputs, the F-divergence notions are stronger notions to capture dependence and provide stronger fairness guarantees. Furthermore, F-divergence notions often also have an operational interpretation; for instance, the Kullback-Leibler (KL) and Chernoff divergences control the decay rates of error probabilities [19], [20]. As a member of the F-divergence family, prior work which uses mutual information [14] between classifier predictions and sensitive attributes is therefore a special case of our framework.

III. F-DIVERGENCE REGULARIZED FAIR TRAINING

We consider a classifier described by trainable parameters w, where $L_w(\hat{y}_n; y_n)$ is the loss function[3] between the output of the classifier (w) and the ground truth of user n; and the fairness constrained learning can be formulated as the following optimization problem (denoted by **OPT**):

$$\min_{w} \frac{1}{N} \sum_{n=1}^{N} L_w(\hat{y}_n; y_n) + \lambda \sum_{i \neq j} D_f^{(w)}(\pi_i \parallel \pi_j), \quad (4)$$

where $D_f^{(w)}(\pi_i \parallel \pi_j)$ is the F-divergence between group i and group j for the classifier parameterized by w; λ is a hyperparameter that can be tuned to balance the trade-off between accuracy and fairness. Solving **OPT** requires the estimation of F-divergence for a classifier. To this end, we propose three F-divergence estimators to compare their performances on fair classification problem. We first leverage the variational representation of F-divergence which allows us to estimate F-divergence in an efficient manner. We show a schematic of the F-divergence based framework in Fig 1.

A. Variational Representation of F-divergences (NN)

It is well known [19] that F-divergence between two distributions admits a variational representation, given as

$$D_f(P \parallel Q) = \sup_{T(\cdot)} E_{X \sim P}\left[T(X)\right] - E_{X \sim Q}\left[f^*(T(X))\right], \quad (5)$$

where the function $f^*(t) = \sup_{x \in dom_f} \{xt - f(x)\}$ 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 F-divergence 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

[3]For our experiments, we use binary cross-entropy loss function for training.

Algorithm 1 F-divergence based Fair Training

Input: Training set $\{x_n,\ y_n\}_{n=1}^N$, samples from distribution π_i and samples from distribution π_j, classifier w_t, F-divergence estimator $\hat{D}_f$.
Output: Fair classifier (w^*) and corresponding F-divergence estimator $(\hat{D}_f^*)$.
for training iterations $(t = 1, 2, \ldots, T_1)$ **do**
  F-divergence estimation for a fixed classifier w_t: Update F-divergence estimator θ for each pair (i, j) of groups for T_2 steps (or until convergence) to maximize $\hat{D}_f^{(w_t,\theta)}(\pi_i, \pi_j)$
  Classifier update: Update classifier w_t to minimize the regularized loss: $\frac{1}{N}\sum_{n=1}^N \mathcal{L}_{w_t}(\hat{y}_n; y_n) + \lambda \sum_{i\neq j} \hat{D}_f^{(w_t,\theta_t)}(\pi_i, \pi_j)$
end for
For demographic parity (DP): $\pi_i \sim \pi(\hat{Y}\,|Z_i)$; for equalized odds (EO): $\pi_i \sim \pi(\hat{Y}\,|Z_i, Y)$.

θ, then taking the supremum over the parameters θ yields a lower bound on F-divergence in (5) as stated next:

$$D_f(P \parallel Q) \geq \sup_{\theta} E_{X\sim P}\left[T_\theta(X)\right] - E_{X\sim Q}\left[f^*(T_\theta(X))\right]. \quad (6)$$

We use the above variational lower bound to estimate the F-divergence for fair classification as described next. For enforcing fairness constraints (DP/EO), we need to estimate F-divergence between joint distributions in different groups, i.e., $D_f(\pi_i \parallel \pi_j) \ \ \forall\ i \neq j, \ \ i, j \in \mathcal{G}$ where for Demographic parity (DP), $\pi_i \sim \pi(\hat{Y}\,|Z = i\,)$, and for Equalized Odds (EO), $\pi_i \sim \pi(\hat{Y}\,|Z = i, Y = y)$. The variational lower bound on F-divergence in (6) can then be estimated as:

$$\max_{\theta} \frac{1}{M}\left(\sum_{m=1}^{M} T_\theta\left(x_i^{(m)}\right) - \sum_{m=1}^{M} f^*\left(T_\theta\left(x_j^{(m)}\right)\right)\right), \quad (7)$$

where in (7), we have replaced the expectation operators with their empirical estimates, and $\{x_i^{(m)}\}$ denote i.i.d. samples drawn from the distribution π_i. Together with the above estimate, from **OPT** we arrive at the following min-max optimization problem (denoted by **MIN-MAX-OPT**):

$$\min_{w}\max_{\theta} \frac{1}{N}\sum_{n=1}^{N}\mathcal{L}_w(\hat{y}_n; y_n) + \lambda \sum_{i\neq j} \underbrace{\left(\frac{1}{M}\sum_{m=1}^{M}\left(T_\theta\left(x_i^{(m)}\right) - f^*\left(T_\theta\left(x_j^{(m)}\right)\right)\right)\right)}_{\triangleq \hat{D}_f^{(w,\theta)}(\pi_i,\pi_j)} \quad (8)$$

MIN-MAX-OPT can be solved by alternatively training the classifier w and F-divergence estimator θ. Specifically, we update the classifier weight w while fixing the F-divergence estimator parameter θ^*, and then we update F-divergence estimator parameter θ fixing the classifier weight w^*. The F-divergence estimator updates each time as the classifier changes. We show the training details in Algorithm 1.

IV. Evaluation Metrics, Results and Discussion

In this Section, we first discuss evaluation metrics (Section IV-A) to evaluate the tradeoffs between fairness and predictive test accuracy. Speficically, we propose the fairness-accuracy receiver operating characteristic (FA-ROC) and a corresponding low-bias version of FA-ROC as a quantitative measure of this tradeoff. We next present the details of our experimental setup in Section IV-B, specifically, the datasets, details of classifier architectures and F-divergence estimators, training hyperparameters, as well as other existing approaches for learning fair classifiers which we compare against the proposed framework. Finally, in Section IV-C, we present and discuss the results and show the improved performance of the proposed F-divergence against various other existing approaches, including regularization, pre- and post-processing methods for four real world datasets (COMPAS dataset [4] for recidivism prediction, Adult Census dataset [21] for income level prediction, Law School admissions dataset [22] for law students' scores prediction, CelebA dataset [23] for facial image classification) and one synthetic dataset (Moon dataset [24]), which are widely used for the assessment of fairness in classification problems. The comprehensive additional experiments are provided in the full version of the paper [25] (including multiple sensitive attributes, groupwise tradeoff(s) between fairness and accuracy, and comparison of different F-divergence estimators). Our code is available at https://github.com/MeiyuZhong/FairnessProject.

A. Fairness-Accuracy Tradeoff & Evaluation Metrics

In addition to imposing approximate fairness constraints, one is simultaneously interested in learning classifiers with high predictive accuracy. To this end, we next define the notion of fairness-vs-accuracy receiver operating characteristic, i.e., FA-ROC. For a classifier $\pi(\hat{Y}|X)$, we denote it's accuracy as $\text{Acc}(\pi) = \Pr(\hat{Y} = Y)$.

Definition 4. *A fairness-accuracy pair (ϵ, a) w.r.t. demographic parity is achievable if there exists a classifier $\pi(\cdot)$ with $Acc(\pi) \geq a$ and $\Delta_{DP}(\pi) \leq \epsilon$. Similarly, one can define the achievability of (ϵ, a) w.r.t. equalized odds.*

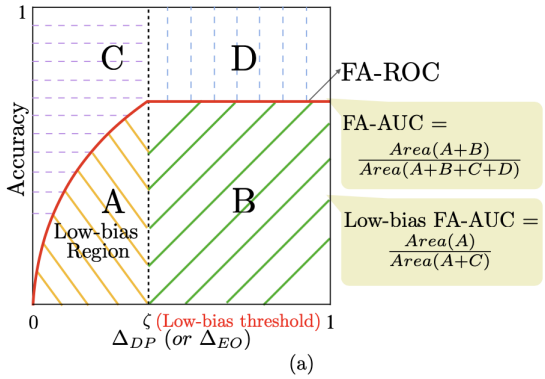
We next state a Proposition which reveals an interesting property of the achievable fairness-accuracy tuples.

Proposition 1. *Suppose that the fairness-accuracy tuples $(\epsilon_1, a_1), (\epsilon_2, a_2), \ldots, (\epsilon_k, a_k)$ are achievable. Then any convex combination $(\sum_{i=1}^k \beta_i\epsilon_i, \sum_{i=1}^k \beta_i a_i)$, for $\beta_i \in [0, 1], \sum_i \beta_i = 1$ of these pairs is also achievable. Consequently, the convex hull of these pairs is also achievable.*

Proof. Suppose that the fairness-accuracy pairs $(\epsilon_i, a_i), i = 1, \ldots, k$ are achievable, i.e., there exists classifiers $\pi_i, i = 1, \ldots, k$ for which $a_i = \pi_i(\hat{Y} = Y)$ and

$$\epsilon_i = \Delta_{DP}(p_i, q_i) = |p_i - q_i| = |\pi_i(\hat{Y}|Z = 0) - \pi_i(\hat{Y}|Z = 1)|$$

We need to show that any convex combination of these tuples, i.e., $(\epsilon, a) = (\sum_i \beta_i\epsilon_i, \sum_i \beta_i a_i)$ is achievable. To this end, we can construct a new classifier as follows: $\pi = \pi_i$ with prob. β_i,



(a)

	Unconstrained Accuracy	Low-bias Threshold	
		w.r.t DP	w.r.t EO
COMPAS	0.6802	0.1257	0.1015
Adult	0.8427	0.0874	0.0165
Law School	0.7641	0.4380	0.2839
CelebA	0.7983	0.4429	0.2376

(b)

Fig. 2: (a) Quantification of Fairness-Accuracy (FA) tradeoffs via notions of FA-AUC and Low-Bias FA-AUC. The red curve (areas A and B) represents the convex hull of the achievable fairness-accuracy pairs, i.e., (ϵ, a) obtained by varying the fairness regularization parameter. The *low-bias* FA-AUC measures the AUC when the bias is less than a prescribed threshold, $\Delta_{DP} \leq \zeta$. We pick this low-bias threshold as the bias in the classifier when it is optimized without any fairness constraint. (b) Low-bias Threshold(s) w.r.t DP/EO and unconstrained test accuracy for four real world datasets.

$i = 1, 2, \ldots, k$, for $\beta_i \in [0, 1], \sum_i \beta_i = 1$. We now prove that the above classifier achieves the desired (ϵ, a). The accuracy of π can be lower bounded as

$$\pi(\hat{Y} = Y) = \sum_{i=1}^{k} \beta_i \pi_i(\hat{Y} = Y) \geq \sum_{i=1}^{k} \beta_i a_i \triangleq a. \quad (9)$$

We next show the bound on the fairness constraint for the classifier π, where $\Delta_{DP}(p, q)$ can be upper bounded by:

$$\left| \sum_i \beta_i p_i - \sum_i \beta_i q_i \right| \overset{(a)}{\leq} \sum_{i=1}^{k} \beta_i |p_i - q_i| \leq \sum_{i=1}^{k} \beta_i \epsilon_i \triangleq \epsilon$$

where (a) follows from the fact that the norm $\Delta_{DP}(p, q)$ is convex in the pair (p, q) followed by Jensen's inequality. This completes the proof of Proposition 1. □

The above Proposition uses the convexity property of $\Delta_{DP}(p, q)$. An important consequence of this Proposition is the following: as the fairness constraint (quantified by ϵ) is varied, one achieves different (fairness, accuracy) operating points. The above Proposition shows that the convex hull of these pairs is also achievable (essentially, in our proof we show that to achieve the convex combination of the original tuples, one can construct a new classifier by combining these classifiers). We now define the fairness-accuracy receiver operating characteristic (FA-ROC) with respect to demographic parity (resp. equalized odds) as follows:

$$\text{FA-ROC}_{\text{DP}} = \{(\epsilon, a) : (\epsilon, a) \text{ is achievable w.r.t. DP}\}.$$
$$\text{FA-ROC}_{\text{EO}} = \{(\epsilon, a) : (\epsilon, a) \text{ is achievable w.r.t. EO}\}.$$

The performance of different regularization techniques can be compared by calculating the area under FA-ROC curve (denoted by FA-AUC) as shown in Fig 2(a).

Low-bias FA-ROC– One shortcoming of the FA-ROC is that it will **mask** the performance of the fair classifier when the original bias of a classifier is small as we explain next. Suppose that we do not impose any fairness constraint, and let ζ denote the bias of the resulting unconstrained classifier. Then, the *low-bias* FA-ROC consists of all achievable (ϵ, a) pairs such that $\epsilon \leq \zeta$. Intuitively, since ζ is the natural bias one would obtain when not imposing any fairness penalty, the non-trivial portion of the trade-off is the one corresponding to $\epsilon \leq \zeta$. Therefore, we introduce the notion of *low-bias FA-ROC*, and argue that this is a more intuitive and justifiable measure of comparing the performance of different regularization techniques. The low-bias region and corresponding AUC are shown in Fig. 2(a). In Fig. 2(b), we also show the low-bias threshold(s) (ζ) as well as the unconstrained classification accuracy for the four real-world datasets used in our experiments (COMPAS, Adult Income, Law School admissions and CelebA datasets).

B. Experimental Setup

1) Datasets: We consider four real-world datasets and one synthetic dataset in our experiments as described next: 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 Census Dataset (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 Admissions Dataset (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. d) *CelebA dataset*: This high-dimensional image dataset contains $N = 202,599$ ($N_{train} = 162,770$, $N_{validation} = 19,867$, $N_{test} = 19,962$) face images of celebrities with 40 binary attributes, which are cropped and resized to 64 × 64 pixels images. We use *attractive/not-attractive* as the binary classification label and *gender* as the sensitive attribute. The original CelebA contains training, validation and testing data. We use the CelebA testing

Datasets	Moon	COMPAS	Adult	Law School	CelebA
Learning Rate	2e-3 / 2e-3	6e-4 / 6e-4	1e-2 / 1e-2	1e-4 / 1e-4	1e-3 / 1e-3
Batch Size	2048 / 2048	2048 / 2048	2048 / 2048	2048 / 2048	256/256
Range of λ (Regularization parameter)	0-9 / 0-9	0-9 / 0-9	0-9 / 0-9	0-9 / 0-9	0-500 / 0-500
Number of Epochs	200 / 200	200 / 200	200 / 200	200 / 200	10/10
Number of seeds	5 / 5	5 / 5	5 / 5	5 / 5	5/5
Optimizer	Adam / Adam	Adam / Adam	Adam / Adam	Adam / Adam	Adam / Adam
Original (unconstrained) Accuracy	0.9728 / 0.9728	0.6802 / 0.6802	0.8427 / 0.8427	0.7641 / 0.7641	0.7983 / 0.7983
Original Bias (i.e., Low-bias Threshold)	0.2706 / 0.0105	0.1257 / 0.1015	0.0874 / 0.0165	0.4380 / 0.2839	0.4429 / 0.2376
Threshold for classifier prediction	0.5 / 0.5	0.5 / 0.5	0.5 / 0.5	0.5 / 0.5	0.5 / 0.5
Steps of F-divergence estimator (per classifier update)	100 / 100	100 / 100	10 / 1	100 / 100	3/3

TABLE I: Hyperparameters for training process. Each entry represents the hyperparameter w.r.t DP / EO.

data to report test accuracy and fairness measurements. e) *Moon Dataset*: This synthetic dataset contains $N = 15,000$ examples ($N_{train} = 10,000$, $N_{test} = 5,000$) with two features and one sensitive attribute.

For all the above datasets, our goal will be to build a fair binary classifier for two scenarios: a) $|Z| = 2$, when the sensitive attribute is race (COMPAS/Law School/Adult datasets), i.e., $Z \in \{C, O\}$, where $C =$ "Caucasian" or $O =$ "Other race", corresponding to two groups; or when the sensitive attribute is gender (CelebA dataset). We also study another scenario, when $|Z| = 4$, when the sensitive attribute(s) are both race and gender (COMPAS/Law School/Adult datasets), i.e., $Z \in \{(C, M), (C, F), (O, M), (O, F), (C, O), (M, F)\}$. Experimental results on multiple sensitive attributes are presented in the full version of the paper [25].

2) Other Methods on Learning Fair Classifiers: We compare and demonstrate the superiority of our proposed F-divergence framework with several existing regularization methods, pre-processing, and post-processing techniques as described next: (1) *KDE based Regularization*: Cho et al. [16]: A kernel density estimation (KDE) based fair training, which directly estimates the conditional distribution $p(\hat{Y}|Z)$ (or $p(\hat{Y}|Z, Y)$) using KDE and uses it for fairness regularization. (2) *Correlation based Regularization*: Mary et al. [18]: A correlation coefficient based fair training, which use correlation coefficient $\rho(\hat{Y}, Z)$ $(or\ \rho(\hat{Y}, Z\,|Y))$ as the regularization term added in the loss function. (3) *Covariance based Regularization*: Zafar et al. [3], [7]: For demographic parity (DP), [3] uses the covariance between the sensitive attributes and the signed distance from the feature vectors to the decision boundary. For equalized odds (EO), [7] uses the covariance between sensitive attributes and the signed distance between the feature vectors of misclassified data and the classifier decision boundary as the regularization term. (4) *FNR-FPR based Regularization*: Yahav et al. [17], which proposes the difference of FNR across different groups and the difference of FPR across different groups as the regularization term. (5) *Pre-processing Method (LFR)*: Zemel et al. [2], which proposes a pre-processing method to learn a fair representation of the data and then use the fair representation to train the model. (6) *Post-processing Method (ROC)* : Hardt et al. [11], which proposes a post-processing method to learn a fair predictor from a discriminatory binary predictor by finding the optimal threshold between TPR and FPR for equalized odds (EO) constraints. For all comparison methods, we use authors' original codes or codes for these papers adapted from AI Fairness 360[4].

3) Model Architectures, Training Methodology and Hyper-parameters: : For low/medium dimensional datasets, namely Moon, COMPAS, Adult and Law School, we consider neural network classifiers with three fully connected layers, where each hidden layer has 200 nodes and is followed by an $SeLU$ [26] non-linear layer. According to the experimental results, three fully connected layers are sufficient to obtain state-of-art accuracy; we report the original accuracy (without fairness constraints) in Table I. We obtained the best accuracy with SeLU activation (in comparison to ReLU, Sigmoid activations). For the proposed F-divergence based fair training, the F-divergence estimator is modeled using two-layer neural networks, each with 5 hidden nodes and a sigmoid non-linearity (see additional discussion on the choice of the model architecture in the full version of the paper [25]). For high dimensional image classification dataset (CelebA), we use ResNet 18 [27] as the classifier for predicting attractive/unattractive. For the F-divergence estimator, we use three linear layers, each with 10 hidden nodes and a sigmoid non-linearity. Note that this architecture is sufficient for CelebA dataset since we only take outputs of the classifier and sensitive attributes as the input of the F-divergence estimator.

For all regularization based training methods, we use the following training loss function: $\mathcal{L}_{\text{Error}} + \lambda \mathcal{R}_{\text{Fairness}}$, where $\mathcal{L}_{\text{Error}}$ denotes the binary cross-entropy loss (for the classification error) together with the fairness related regularization term $\mathcal{R}_{\text{Fairness}}$. The trade-off parameter λ is used to adjust the proportion of classification loss and fairness constraints; varying λ give us the set of points reflecting the trade-off between fairness and accuracy. To compare with pre / post - processing methods, we vary the classification threshold(s) of the corresponding algorithms to form a FA-ROC. By constructing the convex hull w.r.t these points, we form the FA-ROC as shown in Fig 2. For each λ, we train over 5 runs with different random seeds. To obtain the Low-bias FA-ROC, we set $\lambda = 0$ (i.e., no fairness constraint), and find the corresponding bias of the learned classifier. This yields the low-bias threshold ζ, and the low-bias FA-ROC is the convex hull of all points for which the bias is no more than ζ (ζ w.r.t DP / EO are shown in Table I). For all methods, the models were trained using the Adam optimizer. For fair comparison, we use same training steps and the same classification model for other mechanisms. We report the accuracy and fair measurements on the test dataset. All the training hyperparameters are summarized in Table I. For a fair

[4]https://aif360.mybluemix.net/

FA-AUC w.r.t Δ_{DP}						
Datasets:		Moon	COMPAS	Adult	Law School	CelebA
Regularization based mechanisms:	Correlation [23]	0.91 ± 0.02	0.67 ± 0.01	0.84 ± 0.00	0.70 ± 0.02	0.75 ±0.01
	KDE [8]	0.95 ± 0.02	0.66 ± 0.01	0.83 ± 0.02	0.70 ± 0.01	0.73 ± 0.03
	Covariance [39]	0.94 ± 0.00	0.67 ± 0.00	0.84 ± 0.00	0.71 ± 0.01	0.74 ± 0.00
	FNR-FPR [3]	0.95 ± 0.00	0.67 ± 0.01	0.84 ± 0.00	0.72 ± 0.02	0.73 ± 0.00
Pre-processing method:	LFR [40]	0.83 ± 0.02	0.66 ± 0.01	0.81 ± 0.02	0.65 ± 0.02	/
Proposed method:	F-divergence	0.95 ± 0.01 (KL&χ^2)	0.67 ± 0.00 (All)	**0.85 ± 0.01** (SH& χ^2)	**0.73 ± 0.01** (χ^2)	**0.76 ± 0.01** (χ^2)
FA-AUC w.r.t Δ_{EO}						
Datasets:		Moon	COMPAS	Adult	Law School	CelebA
Regularization based mechanisms:	Correlation [23]	0.97 ± 0.00	0.67 ± 0.01	0.84 ± 0.01	0.72 ± 0.02	0.75 ± 0.01
	KDE [8]	0.97 ± 0.00	0.67 ± 0.01	0.83 ± 0.01	**0.73 ± 0.01**	0.74 ± 0.02
	Covariance[38]	0.97 ± 0.00	0.66 ± 0.01	0.84 ± 0.01	0.72 ± 0.01	0.75 ± 0.01
	FNR-FPR [3]	0.97 ± 0.00	0.67 ± 0.01	0.84 ± 0.01	0.72 ± 0.02	0.75 ± 0.01
Post-processing method:	ROC [16]	0.97 ± 0.01	0.67 ± 0.01	0.84 ± 0.01	0.72 ± 0.01	0.70 ± 0.03
Proposed method:	F-divergence	0.97 ± 0.00 (All)	**0.68 ± 0.01** (χ^2)	**0.85 ± 0.01** (χ^2)	**0.73 ± 0.01** (SH)	**0.77 ± 0.00** (SH)

(a)

Low-bias FA-AUC w.r.t Δ_{DP}						
Datasets:		Moon	COMPAS	Adult	Law School	CelebA
Regularization based mechanisms:	Correlation [23]	0.7435 ± 0.0306	0.6140 ± 0.0062	0.7364 ± 0.0073	0.6211 ± 0.0144	0.6991 ± 0.0482
	KDE [8]	**0.8953 ± 0.0029**	0.5750 ± 0.0025	0.7225 ± 0.0532	0.6455 ± 0.0113	0.6659 ± 0.0103
	Covariance [39]	0.8299 ± 0.0110	0.6146 ± 0.0273	0.7778 ± 0.0170	0.6624 ± 0.0330	0.6751 ± 0.0194
	FNR-FPR [3]	0.8865 ± 0.0165	0.6357 ± 0.0367	0.7817 ± 0.0126	0.6671 ± 0.0315	0.6636 ± 0.0117
Pre-processing method:	LFR [40]	0.7705 ± 0.0531	0.5965 ± 0.0119	0.7482 ± 0.0220	0.6043 ± 0.0123	/
Proposed method:	F-divergence	0.8835 ± 0.0206 (χ^2)	**0.6502 ± 0.0210** (χ^2)	**0.8825 ± 0.0259** (χ^2)	**0.7007 ± 0.0026** (χ^2)	**0.7117 ± 0.0215** (χ^2)
Low-bias FA-AUC w.r.t Δ_{EO}						
Datasets:		Moon	COMPAS	Adult	Law School	CelebA
Regularization based mechanisms:	Correlation [23]	0.6832 ± 0.0798	0.6279 ± 0.0076	0.8011 ± 0.0613	0.6374 ± 0.0343	0.6254 ± 0.0132
	KDE [8]	0.7211 ± 0.0002	0.6356 ± 0.0027	0.8203 ± 0.0033	0.6759 ± 0.0119	0.6397 ± 0.0143
	Covariance[38]	0.7209 ± 0.0422	0.6358 ± 0.0025	0.7035 ± 0.0683	0.6236 ± 0.0127	0.6395 ± 0.0166
	FNR-FPR [3]	0.8009 ± 0.0323	0.6355 ± 0.0055	0.6787 ± 0.0115	0.6331 ± 0.0188	0.6401 ± 0.0073
Post-processing method:	ROC [16]	0.7358 ± 0.0636	0.6316 ± 0.0039	0.7331± 0.0510	0.6336±0.0086	0.6024 ± 0.0232
Proposed method:	F-divergence	**0.8085 ± 0.0105** (KL)	**0.6433 ± 0.0017** (χ^2)	**0.9035 ± 0.0172** (χ^2)	**0.6803 ± 0.0041** (SH)	**0.6764 ± 0.0121** (SH)

(b)

Fig. 3: (a) FA-AUC w.r.t Δ_{DP} (top) / Δ_{EO} (bottom) and Accuracy; (b) FA-AUC in the **low-bias** region w.r.t Δ_{DP} (top) / Δ_{EO} (bottom) and Accuracy: We compare our best results with regularization based methods (correlation [18], KDE [16], covariance [3], [7], FNR-FPR [17]), the pre-processing method (LFR [2]) and post-processing method (ROC [11]) method. We show the overall Area-under-the-curve (AUCs) (mean AUC $\pm$ standard deviation) for different techniques under both DP and EO fairness notions. Bracket All represents that all KL, SH, χ^2 divergence can achieve the value. Bracket SH represents that only SH divergence can achieve the value.

comparison of different regularization techniques (including the F-divergence techniques and prior works) pre- and post-processing methodologies, we compare all of the techniques by measuring fairness using the well accepted notion of TV distance (Δ_{DP} as defined in (2) for DP or Δ_{EO} (3) for EO), which are the same metrics as defined in previous works.

C. Comparison of F-divergence Regularization with other approaches for learning fair classifiers

In Fig 3 (a), we report the overall FA-AUC for both DP and EO fairness notions for six other existing methods and the proposed F-divergence approach. The numbers in the table are the mean AUC $\pm$ standard deviation over 5 independent trials. We notice that our proposed method achieves better trade-off than other compared methods for all the real world datasets. We also report the low-bias FA-AUC for all the datasets and the different regularization techniques in Fig 3 (b). In contrast to the overall FA-AUC (where the performance improvement given by F-divergence is modest), when we zoom in the low-bias region, we see a significant performance improvement in the fairness-accuracy tradeoffs. As an example, for the Adult dataset with the DP fairness objective, our proposed method achieves **12.89%** higher low-bias FA-AUC than other compared state-of-art mechanisms (the gain for EO fairness objective is **10.14%**). Our method also performs well on high dimensional dataset (CelebA) where we achieve a gain of **2.67%** / **5.67%** w.r.t the EO constraint for FA-AUC / Low-bias FA-AUC. We show the plots of FA-ROC in Fig 4 on Law School and COMPAS datasets (similar figures for other datasets are provided in the full version of the paper [25]). Another interesting observation is that the optimal choice of F-divergence regularization is dependent on both the dataset as well as the notion of fairness. For instance, Pearson χ^2 divergence regularization yields the best trade-off than other F-divergence based methods w.r.t DP on all real world datasets. For the equalized odds (EO) notion, SH divergence regularization shows better performance on both COMPAS and Law School datasets while χ^2 divergence achieves better results on Adult dataset.

V. Concluding Remarks

In this work, we introduced a general min-max F-divergence regularization framework for fair classifiers, which is readily adaptable for high-dimensional problems and compatible with gradient based optimization methods. In contrast to existing regularization methods such as correlation/covariance/TPR/FPR based methods, F-divergence notions for quantifying fairness/bias are stronger notions to

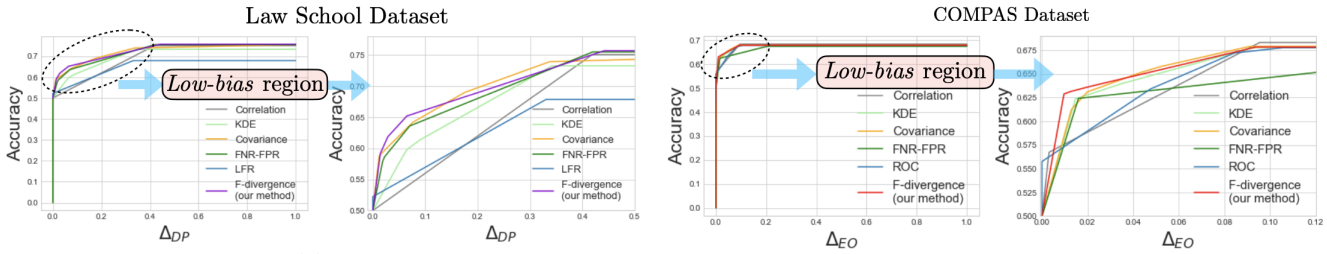


Fig. 4: (a) Trade-off between fairness and accuracy (FA-ROC) on the Law School dataset w.r.t **DP** constraints (left) and corresponding low-bias region FA-ROC (right); (b) Trade-off between fairness and accuracy (FA-ROC) on the COMPAS dataset w.r.t **EO** constraints (left) and corresponding low-bias region FA-ROC (right). As shown in the figure, for both DP and EO constraints, our method outperforms other regularization based methods.

capture dependence between classifier decisions and sensitive attributes and provide stronger fairness guarantees.

We also proposed the notion of Fairness-Accuracy ROC (FA-ROC) and a corresponding low-bias FA-ROC, which we argue as the correct approach to compare different mechanisms for learning fair classifiers and quantifing the tradeoff between fairness and accuracy. Through an extensive set of experiments on four real-world datasets (COMPAS, Adult Income, Law School Admissions and CelebA), we demonstrated the improvement in the fairness-accuracy tradeoffs compared to prior works on regularization as well as pre- and post-processing methods. For instance, for the Adult census dataset, F-divergence regularization leads to $\approx 13\%$ increase in FA-AUC (area under the curve) compared to the state-of-the-art regularization methods for Demographic parity (we achieve a gain of $\approx 10\%$ in FA-AUC for Equalized odds). For the high dimensional dataset (CelebA), our method consistently achieves better performance and achieved a gain of 6% w.r.t EO constraints in the Low-bias region.

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

[5] A. Lambrecht and C. Tucker, "Algorithmic bias? an empirical study of apparent gender-based discrimination in the display of stem career ads," *Management science*, vol. 65, no. 7, pp. 2966–2981, 2019.

[6] 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.

[7] 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.

[8] 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.

[9] S. Barocas and A. D. Selbst, "Big data's disparate impact," *Calif. L. Rev.*, vol. 104, p. 671, 2016.

[10] 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.

[11] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," *Advances in neural information processing systems*, vol. 29, 2016.

[12] M. Yurochkin, A. Bower, and Y. Sun, "Training individually fair ml models with sensitive subspace robustness," in *International Conference on Learning Representations*, 2020.

[13] 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.

[14] 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.

[15] 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.

[16] 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.

[17] Y. Bechavod and K. Ligett, "Learning fair classifiers: A regularization-inspired approach," *Workshop on Fairness, Accountability, and Transparency in Machine Learning (FAT/ML)*, 2017.

[18] J. Mary, C. Calauzenes, and N. El Karoui, "Fairness-aware learning for continuous attributes and treatments," in *International Conference on Machine Learning*, pp. 4382–4391, PMLR, 2019.

[19] 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.

[20] S. Adiga and R. Tandon, "Unsupervised change detection using drecusum," in *2022 56th Asilomar Conference on Signals, Systems, and Computers*, pp. 1103–1110, IEEE, 2022.

[21] D. Dua and C. Graff, "Adult data set," *UCI Machine Learning Repository*, 2017.

[22] L. F. Wightman, *LSAC National Longitudinal Bar Passage Study. LSAC research report series*. 1998.

[23] Z. Liu, P. Luo, X. Wang, and X. Tang, "Deep learning face attributes in the wild," in *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[24] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *the Journal of machine Learning research*, vol. 12, pp. 2825–2830, 2011.

[25] M. Zhong and R. Tandon, "Learning fair classifiers via min-max f-divergence regularization," *arXiv preprint arXiv:2306.16552*, 2023. Available: https://arxiv.org/pdf/2306.16552.pdf.

[26] G. Klambauer, T. Unterthiner, A. Mayr, and S. Hochreiter, "Self-normalizing neural networks," *Advances in neural information processing systems*, vol. 30, 2017.

[27] K. He, X. Zhang, S. Ren, and J. Sun, "Deep residual learning for image recognition," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.