FNNC: Achieving Fairness through Neural Networks

Manisha Padala and **Sujit Gujar**
International Institute of Information and Technology, Hyderabad
manisha.padala@research.iiit.ac.in, sujit.gujar@iiit.ac.in

Abstract

In classification models, fairness can be ensured by solving a constrained optimization problem. We focus on fairness constraints like Disparate Impact, Demographic Parity, and Equalized Odds, which are non-decomposable and non-convex. Researchers define convex surrogates of the constraints and then apply convex optimization frameworks to obtain fair classifiers. Surrogates serve as an upper bound to the actual constraints, and convexifying fairness constraints is challenging.

We propose a neural network-based framework, *FNNC*, to achieve fairness while maintaining high accuracy in classification. The above fairness constraints are included in the loss using Lagrangian multipliers. We prove bounds on generalization errors for the constrained losses which asymptotically go to zero. The network is optimized using two-step mini-batch stochastic gradient descent. Our experiments show that FNNC performs as good as the state of the art, if not better. The experimental evidence supplements our theoretical guarantees. In summary, we have an automated solution to achieve fairness in classification, which is easily extendable to many fairness constraints.

1 Introduction

In recent years machine learning models have been popularized as prediction models to supplement the process of decision making. Such models are used for criminal risk assessment, credit approvals, online advertisements. These machine learning models unknowingly introduce a societal bias through their predictions [Barocas and Selbst, 2016; Berk *et al.*, 2018; Chouldechova, 2017]. E.g., ProPublica conducted its study of the risk assessment tool, which was widely used by the judiciary system in the USA. ProPublica observed that the risk values for recidivism estimated for African-American defendants were on an average higher than for Caucasian defendants. Since then, researchers started looking at fairness in machine learning, especially quantifying the notion of fairness and achieving it.

Broadly fairness measures are divided into two categories. *Individual fairness* [Dwork *et al.*, 2012], requires similar decision outcomes for two individuals belonging to two different groups concerning the sensitive feature and yet sharing similar non-sensitive features. The other notion is of *group fairness* [Zemel *et al.*, 2013], which requires different sensitive groups to receive beneficial outcomes in similar proportions. We are concerned with group fairness and specifically: *Demographic Parity* (DP) [Dwork *et al.*, 2012], *Disparate Impact* (DI) [Feldman *et al.*, 2015] and *Equalized odds* (EO) [Hardt *et al.*, 2016]. DP ensures that the fraction of the positive outcome is the same for all the groups. DI ensures the ratio of the fractions is above a threshold. However, both the constraints fail when the base rate itself differs, hence EO is the more useful notion of fairness, which ensures even distribution of false-positive rates and false-negative rates among the groups. All these definitions make sense only when the classifier is well calibrated. That is, if a classifier predicts an instance belongs to a class with a probability of 0.8, then there should be 80% of samples belonging to that class. [Pleiss *et al.*, 2017; Chouldechova, 2017] show that it is impossible to achieve EO with calibration unless we have perfect classifiers. Hence, the major challenge is to devise an algorithm that guarantees the best predictive accuracy while satisfying the fairness constraints to a certain degree.

Towards designing such algorithms, one approach is preprocessing the data. The methods under this approach treat the classifier as a black box and focus on learning fair representations. The fair representations learned may not result in optimal accuracy. The other approach models achieving fairness as constrained optimization, [Bilal Zafar *et al.*, 2015; Kamishima *et al.*, 2011; Wu *et al.*, 2018]. In [Wu *et al.*, 2018], the authors have provided a generalized convex optimization framework with theoretical guarantees. The fairness constraints are upper-bounded by convex surrogate functions and then directly incorporated into classification models.

There are several limitations in the existing approaches which ensure fairness in classification models. Surrogate constraints may not be a reasonable estimate of the original fairness constraint. Besides, coming up with good surrogate losses for the different definitions of fairness is challenging. In this paper, we study how to achieve fairness in classification. In doing so, we do not aim to propose a new fairness measure or new optimization technique. As opposed to the above approaches, we propose to use neural networks for

implementing non-convex complex measures like DP, DI, or EO. The network serves as a simple classification model that achieves fairness. One need not define surrogates or do rigorous analysis to design the model. Mainly, it is adaptable to any definition of fairness.

Typically, one cannot evaluate fairness measures per sample as these measures make sense only when calculated across a batch, which contains data points from all the sensitive groups. Given that at every iteration, the network processes mini-batch of data, we can approximate the fairness measure given an appropriate batch size. Hence, we use mini-batch stochastic gradient descent (SGD) for optimizing the network. We empirically find that it is possible to train a network using the Lagrangian Multiplier method, which ensures these constraints and achieves accuracy at par with the other complex frameworks. Likewise, it is also possible to incorporate other complex measures like F1-score, H-mean loss, and Q-mean loss, – not related to fairness. We have included an experiment on training a network to minimize Q-mean loss with DP as a constraint.

Our Contribution: i) We propose to design a fair neural network classifier (FNNC) to achieve fairness in classification. ii) We provide generalization bounds for the different losses and fairness constraints DP and EO (Theorem 3) in FNNC. iii) We show that, in some instances, it may be difficult to approximate DI constraint by another surrogate DI constraint (Theorem 4). iv) We empirically show that FNNC can achieve the state of the art performance, if not better.

2 Related Work

In [Zemel *et al.*, 2013], the notion of fairness is discussed. DP, EO, and DI are few of its types. It is a major challenge to enforce these in any general machine learning framework. Widely there are three primary approaches to deal with the challenge:

i) The first body of work focuses on pre-processing i.e., coming up with fair representations as opposed to fair classification e.g., [Feldman *et al.*, 2015; Dwork *et al.*, 2012; Kamiran and Calders, 2009]. Neural networks have been extensively used in such pursuit. E.g., [Louizos *et al.*, 2015] gives a method for learning fair representations with a variational auto-encoder by using maximum mean discrepancies between the two sensitive groups.[Edwards and Storkey, 2016; Madras *et al.*, 2018; Beutel *et al.*, 2017] explore the notion of adversarially learning a classifier that achieves DP, EO or DI.

ii) The second approach focuses on analytically designing convex surrogates for the fairness definitions [Calders and Verwer, 2010; Kamishima *et al.*, 2011; Bechavod and Ligett, 2017] introduce penalty functions to penalize unfairness. [Bilal Zafar *et al.*, 2015; Wu *et al.*, 2018] gives a generalized convex framework that incorporates all possible surrogates and gives appropriate bounds. [Zhang *et al.*, 2018] uses neural network-based adversarial learning, which attempts to predict the sensitive attribute based on the classifier output, to learn an equal opportunity classifier.

iii) The third is the reductionist approach, in which the task of fair classification is reduced to a sequence of cost-sensitive classification [Narasimhan, 2018], and [Agarwal *et al.*, 2018] which can then be solved by a standard classifier. [Agarwal *et al.*, 2018] allows for fairness definitions that can be characterized as linear inequalities under conditional moments like DP and EO (DI does not qualify for the same). FNNC does not have such restrictions and hence performs reasonably for DI as well. We are easily able to include complex and non-decomposable loss functions like Q-mean loss, whereas [Agarwal *et al.*, 2018] aims to improve only the accuracy of the model.

3 Preliminaries and Background

In this section, we introduce the notation used and state the definitions of the fairness measures and the performance measures that we have analyzed.

We consider a binary classification problem with no assumption on the instance space. X is our (d-dimensional) instance space s.t. $X \in \mathbb{R}^d$ and output space $Y \in \{0, 1\}$. We also have a protected attribute $\mathcal{A}$ associated with each individual instance, which for example could be age, sex or caste information. For each $a \in \mathcal{A}$, a could be a particular category of the sensitive attribute like male or female.

Definition 1 (Demographic Parity (DP))**.** *A classifier h satisfies demographic parity under a distribution over $(X, \mathcal{A}, Y)$ if its predictions $h(X)$ is independent of the protected attribute A. That is, $\forall a \in \mathcal{A}$ and $p \in \{0, 1\}$*

$$\mathbf{P}[h(X) = p|\mathcal{A} = a] = \mathbf{P}[h(X) = p]$$

Given that $p \in \{0, 1\}$, we can say $\forall\ a$

$$\mathbb{E}[h(X)|\mathcal{A} = a] = \mathbb{E}[h(X)]$$

Definition 2 (Equalized Odds (EO))**.** *A classifier h satisfies equalized odds under a distribution over $(X, \mathcal{A}, Y)$ if its predictions $h(X)$ are independent of the protected attribute $\mathcal{A}$ given the label Y. That is, $\forall a \in \mathcal{A},\ p \in \{0, 1\}$ and $y \in Y$*

$$\mathbf{P}[h(X) = p|\mathcal{A} = a, Y = y] = \mathbf{P}[h(X) = p|Y = y]$$

Given that $p \in \{0, 1\}$, we can say $\forall\ a, y$

$$\mathbb{E}[h(X)|\mathcal{A} = a, Y = y] = \mathbb{E}[h(X)|Y = y]$$

Definition 3 (Disparate Impact (DI))**.** *The outcomes of a classifier h disproportionately hurt people with certain sensitive attributes. The following is the definition for completely removing DI,*

$$min\left(\frac{\mathbf{P}(h(x) > 0|a = 1)}{\mathbf{P}(h(x) > 0|a = 0)}, \frac{\mathbf{P}(h(x) > 0|a = 0)}{\mathbf{P}(h(x) > 0|a = 1)}\right) = 1$$

[Pleiss *et al.*, 2017] strongly claim that the above mentioned measures are rendered useless, if the classifier is not calibrated, in which case the probability estimate p of the classifier could carry different meanings for the different groups.

Definition 4 (Calibration)**.** *A classifier h is perfectly calibrated if $\forall\ p \in [0, 1],\ \mathbf{P}(y = 1|h(x) = p) = p$.*

Given the definition the authors prove the following impossibility of calibration with equalized odds.

Theorem 1 (Impossibility Result [Pleiss *et al.*, 2017])**.** *Let h_1, h_2 be two classifiers for groups a_1 and $a_2 \in \mathcal{A}$ with $\mathbf{P}(y=1|a_1=1) \neq \mathbf{P}(y=1|a_2=1)$. Then h_1 and h_2 satisfy the equalized odds and calibration constraints if and only if h_1 and h_2 are perfect predictors.*

Given the above result, we cannot guarantee to ensure the fairness constraints perfectly, hence we relax the conditions while setting up our optimization problem as follows,

3.1 Problem Framework

We have used the cross-entropy loss or the Q-mean loss as our performance measures, defined in the next section. We denote this loss by $l(h_\theta(X), Y)$ parameterized by θ, the weights of the network. Our aim is to minimize the loss under the additional constraints of fairness. Below we state the ϵ-relaxed fairness constraints that we implement in our model. $\forall\ a, y$,
DP:

$$|\mathbb{E}[h(X=x)|\mathcal{A}=a] - \mathbb{E}[h(X=x)]| \leq \epsilon \tag{1}$$

EO:

$$|\mathbb{E}[h(X=x)|\mathcal{A}=a, Y=y] - \mathbb{E}[h(X=x)|Y=y]| \leq \epsilon \tag{2}$$

DI: It is not possible to completely remove DI but one has to ensure least possible DI specified by the $p\% - rule$,

$$min\left(\frac{\mathbf{P}(h(x)>0|a=1)}{\mathbf{P}(h(x)>0|a=0)}, \frac{\mathbf{P}(h(x)>0|a=0)}{\mathbf{P}(h(x)>0|a=1)}\right) \geq \frac{p}{100} \tag{3}$$

We have the following generic optimization framework. Both the loss and the constraints can be replaced according to the need,

$$\boxed{\begin{array}{l} \min\limits_{\theta} l_\theta \\ s.t.\ Eq\ 1\ or\ 2\ or\ 3 \end{array}} \tag{4}$$

4 FNNC and Loss Functions

In this section, we discuss how we use the neural network for solving the optimization problem framework in Eq. 4.

4.1 Network Architecture

Our network is a two-layered feed-forward neural network. We only consider binary classification in all our experiments, although this method and the corresponding definitions are easily extendable to multiple classes. Let $h_\theta(.)$ be the function parameterized by θ that the neural network learns. In the last layer of this network we have a softmax function which gives the prediction probability p_i, where p_i is the predicted probability that the i^{th} data sample belongs to one class and $1-p_i$ is the probability for that it belongs to the other. Hence $p := h_\theta(.)$. We use the output probabilities to define the loss and the fairness measure.

4.2 Loss Function and Optimizer

Given the constrained optimization defined by Eq. 4, we use the Lagrangian Multiplier method to incorporate the constraints within a single overall loss. Since the constraints are non-convex, we can only guarantee that the optimizer converges to a local minima. Nevertheless, our experiments show that the model has at par or better performance compared to the existing approaches. We now describe the different loss functions that we have used in the experiments.

Fairness Constraints with Cross Entropy Loss
The fairness constraint DP as in the Def. 1 is given by $\forall a \in \mathcal{A}$,

$$\mathbb{E}[h(X=x)|\mathcal{A}=a] = \mathbb{E}[h(X=x)]$$
$$\mathbb{E}[h(X=x)|\mathcal{A}=1-a] = \mathbb{E}[h(X=x)]$$
$$\therefore\ \mathbb{E}[h(X=x)|\mathcal{A}=a] = \mathbb{E}[h(X=x)|\mathcal{A}=1-a]$$

Hence the constraint for a fixed batch size S of samples given by $z_S = (h(x_S), a_S, y_S)$ and $p_i = h(x_i) \in [0,1]$, can be defined as follows,

$$const^{DP}(z_S) = \left| \frac{\sum_{i=1}^S p_i a_i}{\sum_{i=1}^S a_i} - \frac{\sum_{i=1}^S p_i(1-a_i)}{\sum_{i=1}^S 1-a_i} \right|$$

For the next constraint EO, we first define the difference in false-positive rate between the two sensitive attributes,

$$fpr(z_S) = \left| \frac{\sum_{i=1}^S p_i(1-y_i)a_i}{\sum_{i=1}^S a_i} - \frac{\sum_{i=1}^S p_i(1-y_i)(1-a_i)}{\sum_{i=1}^S 1-a_i} \right|$$

The difference in false-negative rate between the two sensitive attributes,

$$fnr(z_S) = \left| \frac{\sum_{i=1}^S (1-p_i)y_i a_i}{\sum_{i=1}^S a_i} - \frac{\sum_{i=1}^S (1-p_i)y_i(1-a_i)}{\sum_{i=1}^S 1-a_i} \right|$$

Following a similar argument as before the empirical version of EO as defined by Eq. 2 and also used by [Madras *et al.*, 2018] in the experiments is,

$$const^{EO}(z_S) = fpr + fnr$$

EO as defined in [Agarwal *et al.*, 2018] is,

$$const^{EO}(z_S) = \max\{fpr, fnr\}$$

Empirical version of DI for a batch of S samples as defined in Eq. 3 as a constraint for binary classes is given by,

$$const^{DI}(z_S) = -min\left(\frac{\frac{\sum_{i=1}^S a_i p_i}{\sum_{i=1}^S a_i}}{\frac{\sum_{i=1}^S (1-a_i)p_i}{\sum_{i=1}^S 1-a_i}}, \frac{\frac{\sum_{i=1}^S (1-a_i)p_i}{\sum_{i=1}^S 1-a_i}}{\frac{\sum_{i=1}^S a_i p_i}{\sum_{i=1}^S a_i}} \right)$$

The tolerance for each constraint is given by ϵ, which gives the following inequality constraints, for $const^k$, $\forall k \in \{DP, EO, DI\}$ the empirical loss for B batches of samples with each batch having S samples denoted by z_S,

$$l_k(h(X), \mathcal{A}, Y) : \frac{1}{B}\sum_{l=1}^{B} const^k(z_S^{(l)}) - \epsilon \leq 0 \tag{5}$$

Specifically, for $const^{DI}(z_S)$, $\epsilon = -\frac{p}{100}$, where p is typically set to 80.

For maximizing the prediction accuracy, we use cross-entropy loss which is defined as follows for each sample,

$$l_{CE}(h(x_i), y_i) = -y_i \log(p_i) - (1-y_i)\log(1-p_i)$$

The empirical loss,

$$\hat{l}_{CE}(h(X), Y) = \frac{1}{SB}\sum_{i=1}^{SB} l_{CE}(h(x_i), y_i)$$

Hence, the overall loss by the Lagrangian method is,

$$L_{NN}(h(X), \mathcal{A}, Y) = \hat{l}_{CE}(h(X), Y) + \lambda\, l_k(h(X), \mathcal{A}, Y) \tag{6}$$

Satisfying DP with Q-mean Loss

The loss due to DP as already defined before is given by Eq. 5, when $k = DP$. The empirical version of Q-mean loss for binary classes that is for $\forall i \in \{0, 1\}$ is defined as follows,

$$\sqrt{\frac{1}{2}\sum_{i=0}^{1}\left(1 - \frac{\mathbf{P}(h(x) = i, y = i)}{\mathbf{P}(y = i)}\right)^2} \tag{7}$$

The corresponding constraint is given by,

$$l_Q(h(x_S), y_S) = \frac{1}{\sqrt{2}}\sqrt{\left(1 - \frac{\sum_{i=1}^{S} y_i p_i}{\sum_{i=1}^{S} y_i}\right)^2 + \left(1 - \frac{\sum_{i=1}^{S}(1-y_i)(1-p_i)}{\sum_{i=1}^{S}(1-y_i)}\right)^2}$$

The empirical Q-mean loss is,

$$\hat{l}_Q(h(X), Y) = \frac{1}{B}\sum_{l=1}^{B} l_Q(h(x_S^{(l)}), y_S^{(l)})$$

Hence, the overall loss by the Lagrangian method is,

$$L_{NN}(h(X), \mathcal{A}, Y) = \hat{l}_Q(h(X),Y) + \lambda\, l_{DP}(h(X), \mathcal{A}, Y) \tag{8}$$

Lagrangian Multiplier Method

The combination of losses and constraints mentioned above are not exhaustive. The generic definition of the loss could thus be given by, $\forall k \in \{DP, EO, DI\}$

$$\boxed{L_{NN} = l_\theta + \lambda l_k} \tag{9}$$

In the equation above, λ is the Lagrangian multiplier. Any combination can be tried by changing l_θ and l_k as defined in Eq. 6 and Eq. 8. The overall optimization of Eq. 9 is as follows,

$$\min_{\theta} \max_{\lambda} L_{NN}$$

The above optimization is carried by performing SGD twice, once for minimizing the loss w.r.t. θ and again for maximizing the loss w.r.t. λ at every iteration [Eban *et al.*, 2016].

4.3 Generalization Bounds

In this subsection, we provide uniform convergence bounds using Rademacher complexity [Shalev-Shwartz and Ben-David, 2014] for the loss functions and the constraints discussed above. We assume the class of classifiers learned by the neural network has a finite capacity and we use covering numbers to get this capacity. Given the class of neural network, $\mathcal{H}$, for any $h, \hat{h} \in \mathcal{H},\ h : \mathbb{R}^d \to [0, 1]$, we define the following l_∞ distance: $\max_x |h(x) - \hat{h}(x)|$. $\mathcal{N}_\infty(\mathcal{H}, \mu)$ is the minimum number of balls of radius μ required to cover $\mathcal{H}$ under the above distance for any $\mu > 0$.

Theorem 2. *For each of $k \in \{DP, EO\}$, the relation between the statistical estimate of the constraint given batches of samples, z_S, $\mathbb{E}_{z_S}[const^k(z_S)]$, and the empirical estimate for B batches of samples is listed below. Given that $const^k(z_S) \leq 1$, for a fixed $\delta \in (0, 1)$ with a probability at least $1 - \delta$ over a draw of B batches of samples from $(h(X), \mathcal{A}, Y)$, where $h \in \mathcal{H}$,*

$$\mathbb{E}\left[const^k(z_s)\right] \leq \frac{1}{B}\sum_{\ell=1}^{B} const^k\left(z_S^{(\ell)}\right) + 2\Omega_k + C\sqrt{\frac{\log(\frac{1}{\delta})}{B}}$$

$$\Omega_{DP,EO} = \inf_{\mu>0}\left\{\mu + \sqrt{\frac{2\log\left(\mathcal{N}_\infty(\mathcal{H}, \mu/2S)\right)}{B}}\right\}$$

Similarly for cross entropy loss l_{CE} and Q- mean loss l_Q we get the following bounds.

CE loss: consider $h(x) = \phi(f(x))$ where ϕ is the softmax over the neural network output $f(x)$ where $f \in \mathcal{F}$, assuming $f(x) \leq L$

$$\mathbb{E}[l_{CE}(f(x), y)] \leq \frac{1}{B}\sum_{i=1}^{B} l_{CE}(f(x_i), y_i) + 2\Omega_L + CL\sqrt{\frac{\log(\frac{1}{\delta})}{B}}$$

$$\Omega_L = \inf_{\mu>0}\left\{\mu + L\sqrt{\frac{2\log\left(\mathcal{N}_\infty(\mathcal{F}, \mu/S)\right)}{B}}\right\}$$

Q-mean loss:

$$\mathbb{E}\left[l_Q(h(x_S), y_S)\right] \leq \frac{1}{B}\sum_{\ell=1}^{B} l_Q\left(h(x_S^{(\ell)}), y_S^{\ell}\right) + 2\Omega_Q + C\sqrt{\frac{\log(\frac{1}{\delta})}{B}}$$

$$\Omega_Q = \inf_{\mu>0}\left\{\mu + \sqrt{\frac{2\log\left(\mathcal{N}_\infty(\mathcal{H}, \mu/S)\right)}{B}}\right\}$$

C is the distribution independent constant.

The theorem below gives the bounds for the covering numbers for the class of neural networks that we use for our experiments

Theorem 3. *[Dütting* et al.*, 2017] For the network with R hidden layers, D parameters, and vector of all model parameters $\| w \|_1 \leq W$. Given that w_l is bounded, the output of the network is bounded by some constant L.*

$$\mathcal{N}_\infty(\mathcal{F}, \mu/S) = \mathcal{N}_\infty(\mathcal{H}, \mu/S) \leq \left\lceil \frac{DLS(2W)^{R+1}}{\mu} \right\rceil^D$$

Hence, on choosing $\mu = \frac{1}{\sqrt{B}}$ we get,

$$\Omega_{DP} = \Omega_{EO} = \Omega_Q \leq \mathcal{O}\left(\sqrt{RD\frac{\log(WBSDL)}{B}}\right)$$

$$\Omega_L = \mathcal{O}\left(L\sqrt{RD\frac{\log(WBSDL)}{B}}\right)$$

Theorem 4. *Given $h(x) : X \to [0, 1]$, for any $\hat{h}(x) : X \to [0, 1]$ such that $h(x) \neq \hat{h}(x)$, we cannot define a $\widehat{const}_{DI} : (\hat{h}(X), \mathcal{A}, Y) \to \mathbb{R}$ for a $const_{DI} : (h(X), \mathcal{A}, Y) \to \mathbb{R}$ such that $|const_{DI} - \widehat{const}_{DI}| \leq \gamma$ is guaranteed, for any $\gamma > 0$. Thus, $\mathcal{N}_\infty(\mathcal{DI}, \mu)$ is unbounded for any $\mu > 0$ where $\mathcal{DI}$ is set of all possible $const_{DI}$.*

We emphasize that, Theorem 4 indicates that if we approximate DI by a surrogate constraint, however close the learnt classifier is to a desired classifier, the actual DI constraint may get unbounded under specific instances. That is, even two close classifiers (i.e., $|h(x) - \hat{h}(x)| < \mu$ for any $\mu \in (0, 1)$) may have arbitrarily different DI. For our problem, due to this negative results, we cannot give generalization guarantees by using $\mathcal{N}_\infty(\mathcal{DI}, \mu)$ as an upper bound. The few cases where, DI becomes unbounded may not occur in practice as we observe in our experiments that DI results are also comparable. We provide the proofs for all the above theorems in [Manisha and Gujar, 2018]

While training the network, in the loss we use the ϵ-relaxed fairness constraints as defined in Eq.5. We believe that, given the above generalization bounds for the constraints, the trained model will be ϵ-fair with the same bounds.

5 Experiments and Results

In this section, we discuss the network parameters and summarize the results.

5.1 Hyperparameters

The architecture that we used is a simple two-layered feed-forward network. The number of hidden neurons in both the layers was one of the following $(100, 50), (200, 100), (500, 100)$. As fairness constraint has no meaning for a single sample, SGD optimizer cannot be used. Hence we use batch sampling. We fix the batch size to be either 1000 or 500 depending on the dataset, to get proper estimates of the loss while training. It is to be noted that batch processing is mandatory for this network to be trained efficiently. For training, we have used the Adam Optimizer with a learning rate of 0.01 or 0.001 and the training typically continues for a maximum of 5000 epochs for each experiment before convergence. The results are averaged over 5-fold cross-validation performance on the data.

5.2 Performance across Datasets

We have conducted experiments on the six most common datasets used in fairness domain. In Adult, Default, and German dataset, we use gender as the sensitive attribute while predicting income, crime rate, and quality of the customer, respectively, in each of the datasets. In Default and Compass datasets that we used, the race was considered as the sensitive attribute while predicting default payee and recidivism respectively. In the Bank dataset, age is the sensitive attribute while predicting the income of the individual.

In Fig. 1a we observe the inherent biases corresponding to each fairness measure within the datasets considered. In order to obtain the values, we set $\lambda = 0$ in Eq. 9 while training. We compare the baseline accuracy, that is obtained by setting $\lambda = 0$, and accuracy using FNNC. In Fig. 1b, we observe a drop in accuracy when the model is trained to limit DP violations within 0.01, i.e., $\epsilon = 0.01$. There is a significant drop in the accuracy of the Crimes dataset, where the DP is violated the most. Similarly, in Fig. 1c and Fig. 1d, we study the effect of training the models to limit EO and DI, respectively. We observe that the drop in accuracy is more

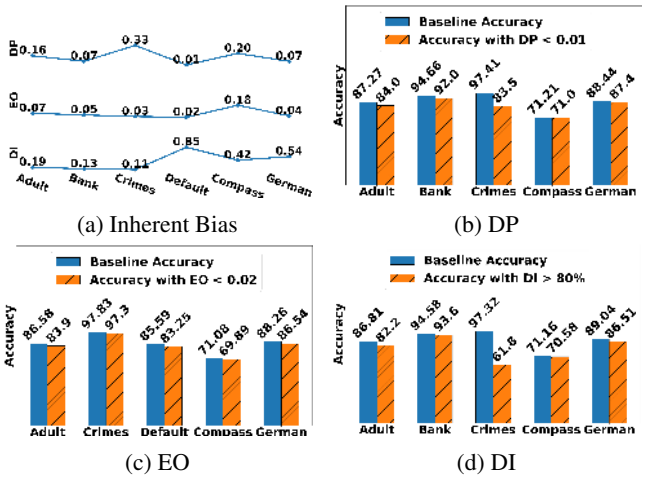


Figure 1: Comparison across datasets

		Female	Male
Zhang *et al.*	FPR	0.0647	0.0701
	FNR	0.4458	0.4349
FNNC	FPR	0.1228	0.1132
	FNR	0.0797	0.0814

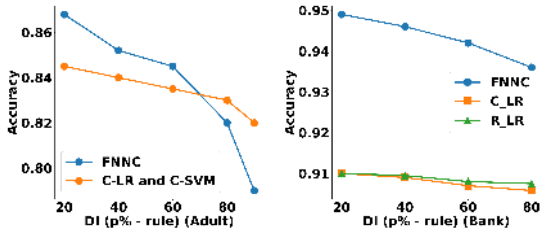
Table 1: False-Positive Rate (FPR) and False-Negative Rate (FNR) for income prediction for the two sex groups in Adult dataset

for datasets that are inherently more biased. In the following section, we compare with other papers and all the results are mostly reported on Adult and Compass dataset. Although all the experiments have single attribute, the approach is easily extendable to multiple attributes.

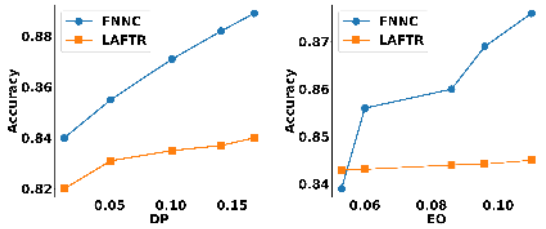
5.3 Comparative Results

In this subsection, we compare our results with related work.

- [Bilal Zafar *et al.*, 2015]: In this paper, the authors propose C-SVM and C-LR to maintain DI while maximizing accuracy. We compare our results with theirs on Adult and bank datasets as observed in the Fig. 2a. We can see that FNNC obtains higher accuracy for ensuring $p\%$ DI rule for upto $p = 80$, for $p > 80$, the accuracy reduces by 2 %. For obtaining the results we train our network using the loss given in Eq. 5 with $const^{DI}$.
- [Madras *et al.*, 2018]: In this work, the authors propose LAFTR to ensure DP and EO while maximizing accuracy on Adult dataset. We have compared our results with theirs in Fig. 2b. For this, we have used loss defined in Eq. 5 with $const^{DP}, const^{EO}$.
- [Zhang *et al.*, 2018]: The authors have results for EO on Adult Dataset as can be found in Table 1. Less violation of EO implies that the FPR and FNR values are almost same across different attributes. We get FPR (female) $0.1228 \sim$ FPR (male) 0.1132 and FNR values for female and male are $0.0797 \sim 0.0814$. The accuracy of the classifier remains at 85%. We carry out similar experiments on Compass dataset and compare FNNC with the baseline i.e., trained without fairness constraints in Fig. 3
- [Agarwal *et al.*, 2018]: We compare our results with theirs on Adult and Compass Dataset both for DP and EO as



(a) Accuracy vs $p\% - rule$ comparison of results with Zafar *et al.* on Adult dataset in the left subplot and Bank dataset in the right subplot



(b) Accuracy vs ϵ (ϵ is tolerance for DP and EO respectively) and compare with Madras *et al.* on Adult dataset

Figure 2: Comparative Results

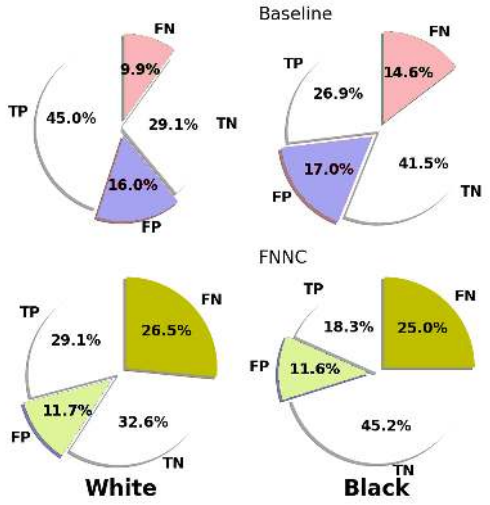


Figure 3: Compass dataset:The FPR and FNR is comparable across race in FNNC as observed in the bottom left and right pie charts

given in Fig. 4. On observing the plots we find our performance is better for Compass dataset but worse for Adult dataset. The violation of EO in Compass dataset is less compared to the Adult dataset as observed in Fig. 1a. Hence, the cost of maintaining fairness is higher in Adult dataset. We can observe in Figs. 2a 2b, 4, that as the fairness constraint is too strict, i.e., ϵ is very small or $p > 80$, the accuracy reduce or error increases.

- [Narasimhan, 2018]: The authors propose COCO and FRACO algorithm for fractional convex losses with convex constraints. In Table 2, we have results for Q-mean loss with DP as the constraint, whose loss function is given by Eq. 8. In the table the values inside the parenthesis

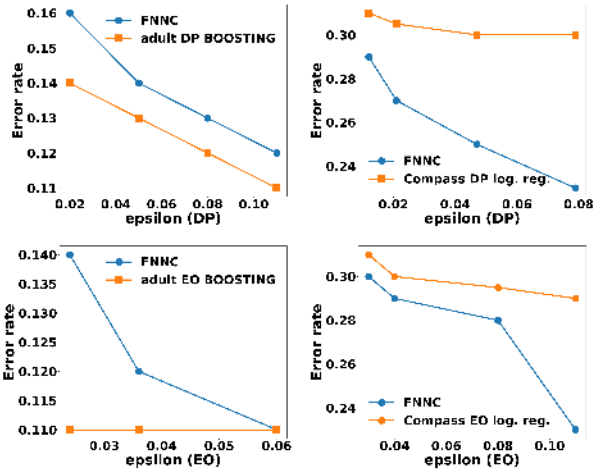


Figure 4: We compare our results with Agrawal *et al.* for Error rate vs (ϵ) tolerance of DP in top row and EO in bottom row

Dataset	ϵ	FNNC	COCO	LinCon
adult	0.05	0.28 (0.027)	0.33 (0.035)	0.39 (0.027)
compass	0.20	0.32 (0.147)	0.41 (0.206)	0.57 (0.107)
crimes	0.20	0.28 (0.183)	0.32 (0.197)	0.52 (0.190)
default	0.05	0.29 (0.011)	0.37 (0.032)	0.54 (0.015)

Table 2: Q-mean loss s.t. DP is within ϵ (actual DP in parentheses)

correspond to the DP obtained during testing and the values outside the parenthesis is the Q-mean loss. We achieve lower Q-mean loss when compared on 4 datasets while DP stays within ϵ.

6 Discussion and Conclusion

The results prove that neural networks perform remarkably well on complex and non-convex measures using batch training. From the analysis on generalization bounds, in Theorem 3, we see that, as $B \to \infty, \Omega \to 0$. As the number of batches of samples increase, the generalization error asymptotically reduces to zero. The batch size S that we use during the training of the network is a crucial parameter. The generalization error increases in $\sqrt{\log S}$ and also increasing S would reduce B (for fixed data-set). Thus, a smaller value of S is preferable for better generalization. On the other hand, having a very small S, would not give a good estimate of the fairness constraint itself. We may end up with sub-optimal classifiers with high loss and less generalization error. Hence, the right balance between S and B is needed to get optimal results.

We believe that the neural networks can learn optimal feature representation of the data to ensure fairness while maintaining accuracy in an end-to-end manner. Hence, our method, FNNC, combines the traditional approach which learns fair representations by pre-processing the data and the approach for training a fair classifier using surrogate losses. One could consider implementing other non-decomposable performance measures like F1-score, Precision, recall, etc., using this approach, and we leave this for future work.

References

[Agarwal *et al.*, 2018] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudik, John Langford, and Hanna Wallach. A reductions approach to fair classification. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 60–69, Stockholmsmässan, Stockholm Sweden, 10–15 Jul 2018. PMLR.

[Barocas and Selbst, 2016] Solon Barocas and Andrew D Selbst. Big data's disparate impact. *Cal. L. Rev.*, 104:671, 2016.

[Bechavod and Ligett, 2017] Yahav Bechavod and Katrina Ligett. Learning fair classifiers: A regularization-inspired approach. *CoRR*, abs/1707.00044, 2017.

[Berk *et al.*, 2018] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, page 0049124118782533, 2018.

[Beutel *et al.*, 2017] Alex Beutel, Jilin Chen, Zhe Zhao, and Ed Huai hsin Chi. Data decisions and theoretical implications when adversarially learning fair representations. *CoRR*, abs/1707.00075, 2017.

[Bilal Zafar *et al.*, 2015] M. Bilal Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. *ArXiv e-prints*, July 2015.

[Calders and Verwer, 2010] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21:277–292, 2010.

[Chouldechova, 2017] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5 2:153–163, 2017.

[Dütting *et al.*, 2017] Paul Dütting, Zhe Feng, Harikrishna Narasimhan, David C Parkes, and Sai Srivatsa Ravindranath. Optimal auctions through deep learning. *arXiv preprint arXiv:1706.03459*, 2017.

[Dwork *et al.*, 2012] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, pages 214–226, New York, NY, USA, 2012. ACM.

[Eban *et al.*, 2016] Elad ET Eban, Mariano Schain, Alan Mackey, Ariel Gordon, Rif A Saurous, and Gal Elidan. Scalable learning of non-decomposable objectives. *arXiv preprint arXiv:1608.04802*, 2016.

[Edwards and Storkey, 2016] Harrison Edwards and Amos Storkey. Censoring representations with an adversary. In *International Conference in Learning Representations (ICLR2016)*, 2 2016.

[Feldman *et al.*, 2015] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '15, pages 259–268, New York, NY, USA, 2015. ACM.

[Hardt *et al.*, 2016] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *NIPS*, 2016.

[Kamiran and Calders, 2009] F. Kamiran and T. Calders. Classifying without discriminating. In *2009 2nd International Conference on Computer, Control and Communication*, pages 1–6, Feb 2009.

[Kamishima *et al.*, 2011] T. Kamishima, S. Akaho, and J. Sakuma. Fairness-aware learning through regularization approach. In *2011 IEEE 11th International Conference on Data Mining Workshops*, pages 643–650, Dec 2011.

[Louizos *et al.*, 2015] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S. Zemel. The variational fair autoencoder. *CoRR*, abs/1511.00830, 2015.

[Madras *et al.*, 2018] David Madras, Elliot Creager, Toniann Pitassi, and Richard S. Zemel. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 3381–3390, 2018.

[Manisha and Gujar, 2018] P. Manisha and Sujit Gujar. A neural network framework for fair classifier. *CoRR*, abs/1811.00247, 2018.

[Narasimhan, 2018] Harikrishna Narasimhan. Learning with complex loss functions and constraints. In *International Conference on Artificial Intelligence and Statistics, AISTATS 2018, 9-11 April 2018, Playa Blanca, Lanzarote, Canary Islands, Spain*, pages 1646–1654, 2018.

[Pleiss *et al.*, 2017] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5680–5689. Curran Associates, Inc., 2017.

[Shalev-Shwartz and Ben-David, 2014] Shai Shalev-Shwartz and Shai Ben-David. *Rademacher Complexities*, page 325–336. Cambridge University Press, 2014.

[Wu *et al.*, 2018] Yongkai Wu, Lu Zhang, and Xintao Wu. Fairness-aware classification: Criterion, convexity, and bounds. *CoRR*, abs/1809.04737, 2018.

[Zemel *et al.*, 2013] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In Sanjoy Dasgupta and David McAllester, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 325–333, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR.

[Zhang *et al.*, 2018] Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. Mitigating unwanted biases with adversarial learning. *CoRR*, abs/1801.07593, 2018.