Optimization-Based Approaches for Enforcing Fairness in Machine Learning

Amil Merchant [*] [1] Alexander Lin [*] [1]

Abstract

Machine learning methods are increasingly being used to make decisions of large social consequence, including credit scoring and recidivism prediction. Although the predictive power of these models is impressive, the data used to train them often contains historical prejudices. This leads to unfair predictions and discrimination based on sensitive attributes such as race or sex. In this paper, we consider two flexible mechanisms to minimize the disparities in predictions between groups. The first extends previous literature on fairness constraints, whereas the second uses an adversarial approach to equalize score distributions. We test both methods on the adult income dataset and find that they lead to fair classifiers without sacrificing significant accuracy.

1. Introduction

Over the past few years, machine learning (ML) and artificial intelligence (AI) have become increasingly more common for high-stakes decision making. Researchers have proposed machine learning algorithms for applications such as credit scoring (Huang et al., 2007), personalized medicine (Poplin et al., 2018), and redicivism prediction (Tollenaar & Van der Heijden, 2013).

In light of the increased adoption of ML/AI methods, it is important that we do not allow these technologies to foster unfairness within our society. Machine learning algorithms fundamentally rely on past data in order to function. They attempt to generalize patterns found in the data and apply these patterns to make predictions in future scenarios. However, in certain situations, historical injustices against presently protected subgroups of a population may have led to the recording of biased data. Naively training a model on this biased data may lead to a biased algorithm that discriminates against these protected subgroups. Subsequently using this algorithm for high-stakes decision making may lead to further injustices and bias the collection of future data, thereby leading to a dangerous positive feedback loop.

Thus, finding ways to enforce fair predictions for machine learning algorithms is a problem of utmost importance. In this paper, we propose and test two methods that strive to achieve this goal. These methods are primarily optimization-based, meaning that they each involve augmenting the objective function of machine learning methods in some manner and can be seen as a form of regularization. We employ our methods in neural networks, models that have garnered a great deal of popularity in recent years due to empirical success across many domains. Our empirical results are presented on the *adult income dataset*[1], which was collected from 1994 census data (Kohavi, 1996). We show that our proposed approaches can significantly reduce model bias defined in the form of *disparate impact* and uphold desired levels of *demographic parity* without sacrificing a prohibitive amount of accuracy.

1.1. Related Work

Recently, there has been an increasing focus on satisfying fairness in a machine learning context. The rise of this field corresponded to widespread concern about a model known as COMPAS. Used in the justice system of the United States to forecast recidivism probabilities, the predictions were used by judges to determine whether defendants receive bail, lighter sentences, or even early parole. While the predictive power of the model was impressive, Angwin et al. (2016) showed that the model has a discriminatory impact on African-American defendants compared to their Caucasian counterparts.

Within the computer science literature, a number of papers have introduced new fairness criterion to quantitatively measure the differential effects of these machine learning models. For example, Hardt et al. (2016) argued for equalizing the True Positive and False Positive Rates between protected groups in a metric known as equality of odds. Kusner et al. (2017) proposed a more causal definition of fairness known as counterfactual fairness. In a recent review, Narayanan (2018) summarized 21 different definitions of fairness and

[*]Equal contribution [1]Applied Mathematics 221, Harvard University, Cambridge, Massachusetts, USA. Correspondence to: Amil Merchant <amilmerchant@college.harvard.edu>, Alexander Lin <alexanderlin01@college.harvard.edu>.

[1] This dataset is publicly available at `https://archive.ics.uci.edu/ml/datasets/adult`.

the social implications of each. In this paper, we focus on one of the first proposed and arguably one of the simplest fairness criteria, known as demographic parity. This definition is particularly interesting due to its legal implications for showing that a model is discriminatory.

Similarly, a number of approaches have been developed to minimize the differences between individuals based on race, gender, etc. Some of the earliest strategies involved preprocessing the available data to learn a representation space where the differences are minimized. Dwork et al. (2012) and Feldman et al. (2015) utilized such approaches. However, these methods struggled with unpredictable and large losses in accuracy in order to ensure fairness. This paper follows two more recent research directions: introducing constraints and adversarial learning.

Instead of focusing on the inputs, constraint-based learning introduced regularization terms for classifiers to ensure fairness. For example, Kamishima et al. (2012) used a penalty to correct the training of logistic regression classifiers. More recently, Zafar et al. (2015) designed a generalizable method by penalizing properties of the decision boundary.

Adversarial networks have been utilized by a number of papers to enforce fairness definitions. For example, Beutel et al. (2017) shared hidden layers between various networks to enforce demographic parity. More recently, Wadsworth et al. (2018) introduced the setup of a predictor attempting to maximize accuracy and an adversary enforcing fairness.

In this paper, we extend the methods of Zafar et al. (2015) and Wadsworth et al. (2018) to create fair classifiers that are able to ensure fairness without forsaking significant accuracy.

2. Background

2.1. Adult Income Dataset

The adult income dataset (Kohavi, 1996) contains data from $N = 32,561$ respondents to the 1994 United States Census. Each person n is characterized by $J = 14$ attributes, denoted $\boldsymbol{x}^{(n)} = \{x_1^{(n)}, \ldots, x_J^{(n)}\}$, including education level, occupation type, capital gains, capital losses, and number of hours worked per week. The goal is to predict a binary variable $y^{(n)} \in \{0, 1\}$, which indicates whether or not person n makes over $50,000 a year.

In this case, the protected attributes $\boldsymbol{z}^{(n)}$ for person n are their *sex* and their *race*. Historical inequities have led to groups such as women and African Americans having significantly lower fractions of individuals making over $50,000 a year. Using a model naively trained on the adult income dataset for high stakes decision making in the present day – such as estimating a person's income for loan approval or determining how much to pay a new hire – may lead to heavily biased results. Thus, there is motivation to incorporate predictive fairness into the model training process.

2.2. Disparate Impact and Demographic Parity

Disparate impact is the notion in which a model's biased classification process leads to outcomes that disproportionately hurt (or benefit) people with sensitive attributes. It was first introduced by Zafar et al. (2015). Simply removing the sensitive attributes $\boldsymbol{z}$ from the dataset and training a model on the remaining attributes $\boldsymbol{x} \setminus \boldsymbol{z}$ may still yield biased predictions, because $\boldsymbol{z}$ may be correlated with the remaining subset (Agarwal et al., 2018).

To counter disparate impact, we wish to enforce *demographic parity*, which demands that the distribution of scores for any protected classes is the same. Let $\hat{p}(y = 1)$ be a model's prediction of the probability of class 1 in binary classification. Formally, demographic parity is defined as:

$$\hat{p}(y = 1 \mid \boldsymbol{z} = k_1) = \hat{p}(y = 1 \mid \boldsymbol{z} = k_2), \tag{1}$$

where k_1 and k_2 are different realizations of the random variable $\boldsymbol{z}$. For example, if $\boldsymbol{z}$ is sex, k_1 could be `Male` and k_2 could be `Female`. Intuitively, this means that only changing the protected attribute $\boldsymbol{z}$ should not influence the predictions in any way.

Using demographic parity as a definition of machine learning fairness offers some advantages. First and foremost, there exists legal support for this definition in the United States. In 1978, four government agencies – including the EEOC, Department of Labor, Department of Justice, and the Civil Service Commission – proposed the four-fifths (or 80%) rule as a benchmark with assessing adverse disparate impact for protected classes (Bobko & Roth, 2004). Specifically, these agencies required that

$$\min\left\{\frac{\hat{p}(y = 1 \mid \boldsymbol{z} = k_1)}{\hat{p}(y = 1 \mid \boldsymbol{z} = k_2)}, \frac{\hat{p}(y = 1 \mid \boldsymbol{z} = k_2)}{\hat{p}(y = 1 \mid \boldsymbol{z} = k_1)}\right\} \geq \frac{q}{100} \tag{2}$$

where $q = 80$ in the legal definition. Note that $q = 100$ corresponds to zero disparate impact and complete demographic parity. Recently, Hu & Chen (2018) additionally argue that short-term enforcement of demographic parity has long-term benefits for countering discrimination against minorities in the labor market.

3. Methods for Enforcing Demographic Parity

We present two optimization-based methods for enforcing demographic parity in neural networks.

A neural network is a cascade of linear and nonlinear transformations of the input vector $\boldsymbol{x}$ to yield an output vector $\boldsymbol{h}_L$ (Goodfellow et al., 2016). An L-layer neural network can be described by the equations,

$$\begin{aligned} \boldsymbol{h}_1 &= f^{(1)}(W^{(1)}\boldsymbol{x} + b^{(1)}), & \dots \quad (3) \\ \boldsymbol{h}_\ell &= f^{(\ell)}(W^{(\ell)}\boldsymbol{h}_{\ell-1} + b^{(\ell)}), & \dots \\ \boldsymbol{h}_L &= f^{(L)}(W^{(L)}\boldsymbol{h}_{L-1} + b^{(L)}), & \end{aligned}$$

where each pair $(W^{(\ell)}, b^{(\ell)})$ parameterizes an affine transformation (via matrix multiplication and bias addition), each $f^{(\ell)}$ is a nonlinear function applied element-wise, and each $\boldsymbol{h}_\ell$ denotes an intermediary hidden state representation of the input.

In binary classifiers, it is common to let $W^{(L)}$ be a row vector, $b^{(L)}$ be a single scalar, and $f^{(L)}$ be the sigmoid function $\sigma(a) = 1/(1 + \exp(-a))$. Such constraints force the final output $\hat{p} = \boldsymbol{h}_L$ to be a scalar within the range $[0, 1]$, which allows us to interpret it as the estimated probability of $y = 1$. For selected nonlinearities $\{f^{(\ell)}\}_{\ell=1}^L$, the weights $\{W^{(\ell)}\}_{\ell=1}^L$ and biases $\{b^{(\ell)}\}_{\ell=1}^L$ are trained to minimize the *binary cross-entropy loss* Q_0 over the entire dataset, which is defined as

$$Q_0 = \sum_{n=1}^{N} y^{(n)} \log \hat{p}^{(n)} + (1 - y^{(n)}) \log(1 - \hat{p}^{(n)}), \quad (4)$$

where each $\hat{p}^{(n)}$ is generated by passing $\boldsymbol{x}^{(n)}$ through the neural network.

3.1. Regularizing Decision Boundary Covariance

Zafar et al. (2015) propose regularizing the covariance between the distance to the decision boundary of a classifier and the protected classes $\boldsymbol{z}$ to enforce demographic parity. They apply their framework to logistic regression and support vector machines. We generalize this method to working with neural networks.

Using the neural network binary classifier of Equation 3, we define the *decision boundary distance* $d^{(n)}$ of training example n as the value obtained before the final nonlinearity, i.e.

$$d^{(n)} = W^{(L)}\boldsymbol{h}_{L-1}^{(n)} + b^{(L)}. \quad (5)$$

To see why $d^{(n)}$ is related to the decision boundary of the neural network classifier, observe that the estimated probability of $y^{(n)} = 1$ is $\hat{p}^{(n)} = \sigma(d^{(n)})$. Thus, if $d^{(n)} > 0$, then $\hat{p}^{(n)} > 1/2$ (so it makes more sense to classify n as class 1) and if $d^{(n)} < 0$, then $\hat{p}^{(n)} < 1/2$ (so it makes more sense to classify n as class 0). Thus, the variable d encodes a scale centered at zero and characterizes the confidence of the classifier to classify as class 0 or class 1.

If the covariance between the decision boundary distance d and the protected attribute $\boldsymbol{z}$ is zero, then knowing $\boldsymbol{z}$ should have no impact on knowing $p(y \mid \boldsymbol{x})$, which is the definition of satisfying demographic parity. We can empirically estimate this covariance by observing the following:

$$\begin{aligned} \mathbb{C}\text{ov}(\boldsymbol{z}, d) &= \mathbb{E}[(\boldsymbol{z} - \bar{\boldsymbol{z}}) \cdot (d - \bar{d})] \quad (6) \\ &= \mathbb{E}[(\boldsymbol{z} - \bar{\boldsymbol{z}}) \cdot d] - \mathbb{E}[(\boldsymbol{z} - \bar{\boldsymbol{z}})] \cdot \bar{d} \\ &= \mathbb{E}[(\boldsymbol{z} - \bar{\boldsymbol{z}}) \cdot d] - 0 \\ &\approx \frac{1}{N} \sum_{n=1}^{N} (\boldsymbol{z}^{(n)} - \hat{\boldsymbol{z}}) \cdot d^{(n)}, \end{aligned}$$

where $\hat{\boldsymbol{z}} = 1/N \cdot \sum_{n=1}^{N} \boldsymbol{z}^{(n)}$. Since Zafar et al. (2015) work with only convex classifiers, they simply add the following convex constraint to their logistic regression and support vector machine settings:

$$\left| \frac{1}{N} \sum_{n=1}^{N} (\boldsymbol{z}^{(n)} - \hat{\boldsymbol{z}}) \cdot d^{(n)} \right| \leq \boldsymbol{c}, \quad (7)$$

for some constant $\boldsymbol{c}$ corresponding to the level of desired demographic parity. In our neural network setting, we instead directly add the empirical covariance as a penalized regularization term to the binary cross entropy objective function of Equation 9. Thus, the full objective function is

$$\begin{aligned} Q_1 = \sum_{n=1}^{N} & y^{(n)} \log \hat{p}^{(n)} + (1 - y^{(n)}) \log(1 - \hat{p}^{(n)}) \quad (8) \\ & + \lambda \cdot \left| \frac{1}{N} \sum_{n=1}^{N} (\boldsymbol{z}^{(n)} - \hat{\boldsymbol{z}}) \cdot d^{(n)} \right|, \end{aligned}$$

where λ controls the degree of regularization. Increasing λ will increase the penalty of the covariance and ideally lead to greater demographic parity. We wish to adjust λ so that it is large enough to satisfy fairness constraints, yet small enough to not prohibitively affect classifier accuracy.

3.2. Regularizing with Adversarial Networks

Adversarial networks were first introduced by Goodfellow et al. (2014) in the context of generative adversarial nets, which simulate a minimax two-player game between two neural networks. The *generator* attempts to create realistic-looking fake images, while the *discriminator* attempts to distinguish real images from fake ones. The generator is trained so that the discriminator (also known as the *adversary*) performs poorly, which allows the generator to reach an equilibrium in which the distribution of its synthesized images approximates that of the training set.

Wadsworth et al. (2018) apply the idea of adversarial networks to fairness in machine learning, specifically looking at the context of criminal recidivism prediction. We use this concept in our income prediction problem.

Let G be a neural network binary classifier (Equation 3) that optimizes for binary cross-entropy loss Q_0 (Equation 9). Define the logit of the output probability for training example n as $d^{(n)} = \sigma^{-1}(\hat{p}^{(n)})$, where σ is the sigmoid function; notice that this is equivalent to the decision boundary distance of Equation 5. We train a second neural network binary classifier A, known as the discriminator (or adversarial network), that learns to classify the sensitive attributes $\boldsymbol{z}$ using d. That is, A works with the supervised training set $\{(d^{(1)}, \boldsymbol{z}^{(1)}), \dots, (d^{(N)}, \boldsymbol{z}^{(N)})\}$. Its loss function also follows the form of binary cross-entropy:

$$Q_A = \sum_{n=1}^{N} \boldsymbol{z}^{(n)} \log A(d^{(n)}) + (1 - \boldsymbol{z}^{(n)}) \log(1 - A(d^{(n)})). \tag{9}$$

Then, in the spirit of generative adversarial networks, G is trained so that A performs poorly. In other words, the augmented loss function of G is:

$$Q_2 = Q_0 - \alpha \cdot Q_A, \tag{10}$$

where $\alpha \geq 0$ controls the tradeoff between optimizing for Q_0 versus $-Q_A$. In the case where the loss function Q_A reaches its maximum, there is no way to predict the sensitive attribute $\boldsymbol{z}$ from the output of G, which implies zero disparate impact and complete demographic parity.

Goodfellow et al. (2014) provide some theoretical results that show the concept of adversarial training aims to minimize the Jensen-Shannon divergence $\mathbb{JS}$ between two probability distributions q_1 and q_2. This is given as

$$\mathbb{JS}(q_1, q_2) = \frac{1}{2}\mathbb{KL}\left(q_1 || q_{12}\right) + \frac{1}{2}\mathbb{KL}\left(q_2 || q_{12}\right), \tag{11}$$

where $q_{12} = (q_1 + q_2)/2$ and $\mathbb{KL}$ denotes the Kullback-Leibler divergence. The Jensen-Shannon divergence has some nicer properties, such as symmetry, in comparison to other similar divergences in its family.

In the case of fairness, these distributions are the conditional distributions $d \mid \boldsymbol{z} = k_1$ and $d \mid \boldsymbol{z} = k_2$ output by G for the logit output probabilities d with respect to the sensitive attribute $\boldsymbol{z}$. Through the deterministic sigmoid transformation σ, we arrive at prediction probabilities $\hat{p}(y = 1 \mid \boldsymbol{z} = k_1)$ and $\hat{p}(y = 1 \mid \boldsymbol{z} = k_2)$, respectively. In our application, it makes sense to constrain these distributions to be close to one another, because the very definition of demographic parity (Equation 2) is tied to this fact.

4. Results

Our empirical results are evaluated on the adult income dataset. We first naively train a vanilla neural network and show how it suffers from disparate impact. Then, we apply our methods for enforcing demographic parity to exhibit how this disparate impact can be mitigated. All experiments are implemented using the PyTorch deep learning library (Paszke et al., 2017).

4.1. Vanilla Neural Network

We train a simple neural network with $L = 2$ layers that performs well on the adult income dataset. The input is $\boldsymbol{x} \setminus \boldsymbol{z}$, the set of all attributes minus sex and race. The single hidden layer $\boldsymbol{h}^{(1)}$ has 64 hidden units. We let $f^{(1)}$ be the rectified linear (ReLU) function and $f^{(2)}$ be the sigmoid function. Weights and biases $\{W^{(1)}, W^{(2)}, b^{(1)}, b^{(2)}\}$ are initialized as $\mathcal{N}(0, 1)$ random variables.

We divide the dataset of $N = 32,561$ individuals into a training set $\mathcal{D}_{\text{train}}$ of $26,048$ people and test set $\mathcal{D}_{\text{test}}$ of $6,513$ people, which roughly corresponds to an 80%-20% split. The network is trained using the binary cross entropy loss function of Equation 9 on $\mathcal{D}_{\text{train}}$. For optimization, we use the ADAM stochastic optimizer (Kingma & Ba, 2014) with a minibatch of $1,024$ examples. The network is trained for 20 epochs, which are defined as passes through the entire training set.

Evaluation is performed on the test set. Test set accuracy is 85.00%, which is decent. However, there are gross violations of demographic parity.

If we observe the distributions over estimated probabilities of making over 50K divided by sex (i.e. `Male` vs. `Female`), we see that there are significant discrepancies. Figure 1 traces the distribution of $\hat{p}^{(n)} \mid \boldsymbol{z}^{(n)} = \texttt{Male}$ and $\hat{p}^{(n)} \mid \boldsymbol{z}^{(n)} = \texttt{Female}$ for all $n \in \mathcal{D}_{\text{test}}$ by using simple kernel density estimation. The shapes are quite different. Let $\mathcal{D}_{\text{test}}^{\texttt{Male}}$ and $\mathcal{D}_{\text{test}}^{\texttt{Female}}$ be partitions of $\mathcal{D}_{\text{test}}$ based on sex. We see that the largest possible q that satisfies Equation 2 is $q = 41.82\%$, where q is found empirically in this example as

$$q = \frac{|\mathcal{D}_{\text{test}}^{\texttt{Female}}|^{-1} \sum_{n \in \mathcal{D}_{\text{test}}^{\texttt{Female}}} \hat{p}^{(n)}}{|\mathcal{D}_{\text{test}}^{\texttt{Male}}|^{-1} \sum_{n \in \mathcal{D}_{\text{test}}^{\texttt{Male}}} \hat{p}^{(n)}}. \tag{12}$$

This model exhibits significant bias against females, likely because it was trained on a biased dataset. Thus, it is unsuitable for use in future high-stakes decision making, such as determining how much a female should make or estimating a female's income for loan approval.

We can repeat the same exercise for race on analogously defined datasets $\mathcal{D}_{\text{test}}^{\texttt{Minorities}}$ and $\mathcal{D}_{\text{test}}^{\texttt{White}}$. For race, we find that $q = 63.63\%$, which is less unfair, yet still violates the 80% rule used in legal settings. Figure 1 presents the corresponding plot.

Mean predicted probabilities of high-income for the aforementioned sensitive groups can be found in Table 1.

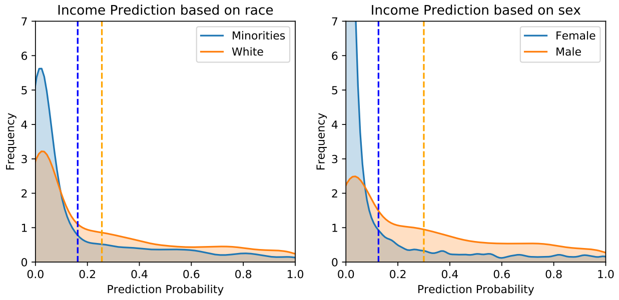

Figure 1. For the vanilla neural network, fitted kernel density estimations of test set estimated probabilities that different races (left) and different sexes (right) make over 50K a year. Dotted lines indicate the means of each distribution (Table 1).

Female	Male	Minorities	White
0.125	0.300	0.162	0.256

Table 1. For the vanilla neural network, test set mean estimated probabilities of making over 50K for various sensitive groups.

4.2. Regularizing Decision Boundary Covariance

We apply the method described in Section 3.1 to correcting disparate impact for the vanilla neural network of Section 4.1. In doing so, we keep the same general architecture and training hyperparameters described in the previous section. However, instead of training the network using normal binary cross-entropy loss Q_0 (Equation 9), we instead use the regularized objective Q_1 that penalizes decision boundary covariance (Equation 8).

In our experiments, we vary the regularization penalty λ to show corresponding effects on the final accuracy and fairness of the neural network classifier. We try λ within the set $\{3\times10^{-2}, 1\times10^{-2}, 3\times10^{-3}, 1\times10^{-3}, 3\times10^{-4}, 1\times10^{-4}\}$, which covers approximate increases in factors of three.

Graphs and a table of the results for sex on the training and test sets can be found in Figure 2 and Table 2, respectively. We see that choosing a suitable λ can satisfy demographic parity without sacrificing significant amounts of accuracy. Looking at Figure 2 and Table 2, there is a sharp bend in the curve for $\lambda = 3\times10^{-3}$, so this is an appropriate final choice.

Similar results for race can be found in Figure 3 and Table 3. Looking at these values, it appears that $\lambda = 1\times10^{-3}$ is a reasonable choice here.

Figure 4 shows the effect of regularizing the network on aligning the prediction distributions for the sensitive attributes using the aforementioned values of λ. Comparing this graph with Figure 1, we see that the gains in demographic parity are significant. The tables also show that the test accuracy for our chosen values of λ drop by a maximum of 1.2%, which is very little in comparison.

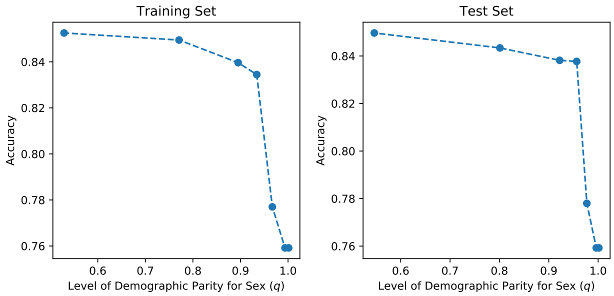

Figure 2. Tradeoff between overall neural network accuracy and level of demographic parity for *sex* by varying regularization penalty λ on the covariance between the distance to the decision boundary and sensitive attributes. Results are given on the training set (left) and the test set (right).

λ (for Sex)	Train Acc	Train q	Test Acc	Test q
3×10^{-2}	0.759	**0.994**	0.759	**0.996**
1×10^{-2}	0.777	0.967	0.778	0.977
3×10^{-3}	0.834	0.934	0.838	0.957
1×10^{-3}	0.840	0.895	0.838	0.922
3×10^{-4}	0.849	0.771	0.843	0.801
1×10^{-4}	**0.852**	0.529	**0.850**	0.547

Table 2. Numerical results of how accuracy and level of demographic parity change as functions of regularization parameter λ for constraining prediction probabilities conditioned on *sex*.

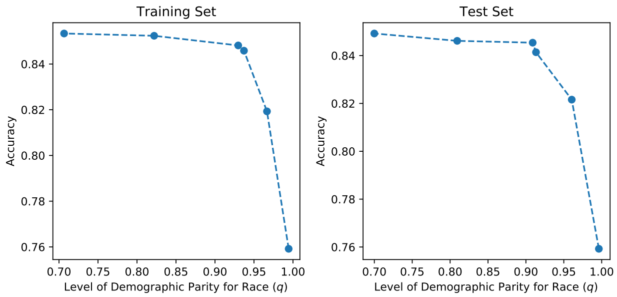

Figure 3. Tradeoff between overall neural network accuracy and level of demographic parity for *race* by varying regularization penalty λ on the covariance between the distance to the decision boundary and sensitive attributes. Results are given on the training set (left) and the test set (right).

λ (for Race)	Train Acc	Train q	Test Acc	Test q
3×10^{-2}	0.759	**0.994**	0.759	**0.996**
1×10^{-2}	0.819	0.967	0.822	0.960
3×10^{-3}	0.846	0.937	0.841	0.913
1×10^{-3}	0.848	0.930	0.845	0.910
3×10^{-4}	0.852	0.822	0.846	0.810
1×10^{-4}	**0.853**	0.707	**0.849**	0.700

Table 3. Numerical results of how accuracy and level of demographic parity change as functions of regularization parameter λ for constraining prediction probabilities conditioned on *race*.

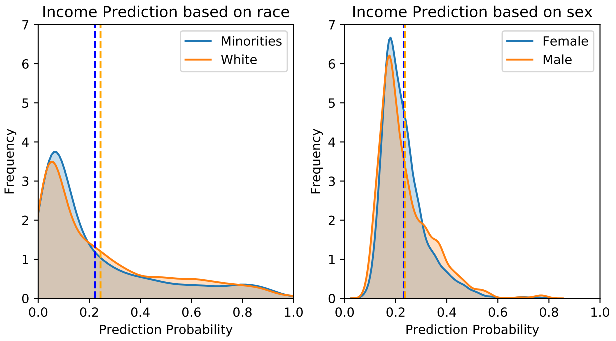

Figure 4. For the decision boundary-regularized neural network, fitted kernel density estimations of test set estimated probabilities that different races (left) and different sexes (right) make over 50K a year. Dotted lines indicate the means of each distribution.

4.3. Regularizing with Adversarial Neural Networks

Here, we adopt the adversarial model explained in Section 3.2. The predictor model G is exactly equivalent to the vanilla neural network, so after pre-training, the results are equivalent to those described in Section 4.1. The addition of the adversary A and continued training leads to improvement in the fairness of the output predictions. In our experiments, we test the adversarial model for various values of α corresponding to how heavily the adversarial loss (ability to predict the sensitive attribute) is weighted with respect to the predictor loss. We test the values of α given in the following set: $\{1, 3, 10, 30, 100, 300\}$. Note that since the adversary A is able to predict all elements of $\boldsymbol{z}$, there is only 1 trial per α value.

The resulting trends can be seen in Figures 5 and 6, with the statistics presented in Table 4. Neither figure has as clear of a sharp bend as the figures for the previous method. Nevertheless, $\alpha = 30$ appears to be a decent trade-off for both cases. For this value of α, the resulting prediction distributions are seen in Figure 7. This graph is significantly different from Figure 1 and shows significant improvement in equalizing the score distributions between individuals of different races or sexes.

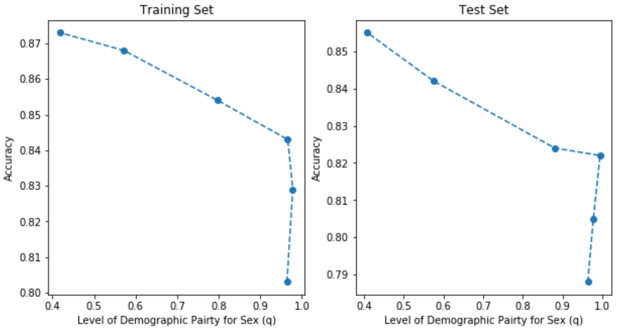

Figure 5. Tradeoff between overall neural network accuracy and level of demographic parity for *sex* by varying weight on adversarial loss α. Results are given on the training set (left) and the test set (right).

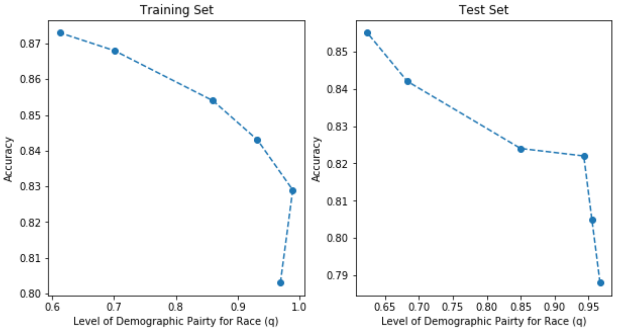

Figure 6. Tradeoff between overall neural network accuracy and level of demographic parity for *race* by varying weight on adversarial loss α. Results are given on the training set (left) and the test set (right).

α	Train Acc	Train q Race	Train q Sex	Test Acc	Test q Race	Test q Sex
300	0.803	0.969	0.964	0.788	**0.967**	0.963
100	0.829	**0.989**	**0.978**	0.805	0.955	0.977
30	0.843	0.932	0.966	0.822	0.943	**0.994**
10	0.854	0.860	0.797	0.824	0.850	0.880
3	0.868	0.701	0.572	0.842	0.683	0.575
1	**0.873**	0.613	0.418	**0.855**	0.624	0.408

Table 4. Numerical results of how accuracy and level of demographic parity change as functions of regularization parameter α for adversarial models.

5. Discussion and Conclusion

From the results, both methods clearly showed significant progress in improving the demographic parity relative to the vanilla neural network model. The plots showing the score distributions highlighted how the predictions for different

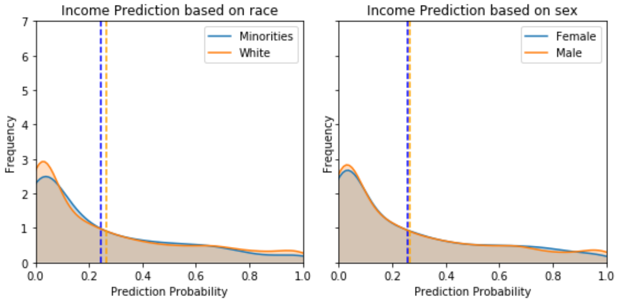

Figure 7. For the adversarial neural network approach, fitted kernel density estimations of test set estimated probabilities that different races (left) and different sexes (right) make over 50K a year. Dotted lines indicate the means of each distribution.

races or sexes are similar after the application of these methods. Furthermore, using the statistics from the tables, it was clear that the methods achieve the $q \geq 0.8$ threshold that is important in legal contexts to create defensible models.

In light of these results, there are many directions for future work. First, one could test these results on other real-world datasets to explore how effective these methods are. In addition, the results from this paper were generally empirical. Both methods were able to improve q scores without forsaking significant accuracy. It would be interesting to provide theoretical bounds for when a certain method will work and bound the loss in accuracy to achieve demographic parity. Finally, as mentioned in the introduction, demographic parity is only one definition of fairness of the more than twenty within literature. Further work could test the impact of these methods on other definitions and explore if they can be extended to meet multiple definitions of fairness.

References

Agarwal, A., Beygelzimer, A., Dudík, M., Langford, J., and Wallach, H. A reductions approach to fair classification. *arXiv preprint arXiv:1803.02453*, 2018.

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine bias. *ProPublica, May*, 23, 2016.

Beutel, A., Chen, J., Zhao, Z., and Chi, E. H. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*, 2017.

Bobko, P. and Roth, P. L. The four-fifths rule for assessing adverse impact: An arithmetic, intuitive, and logical analysis of the rule and implications for future research and practice. In *Research in personnel and human resources management*, pp. 177–198. Emerald Group Publishing Limited, 2004.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pp. 214–226. ACM, 2012.

Feldman, M., Friedler, S. A., Moeller, J., Scheidegger, C., and Venkatasubramanian, S. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 259–268. ACM, 2015.

Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Goodfellow, I., Bengio, Y., and Courville, A. *Deep learning*. MIT press, 2016.

Hardt, M., Price, E., Srebro, N., et al. Equality of opportunity in supervised learning. In *Advances in neural information processing systems*, pp. 3315–3323, 2016.

Hu, L. and Chen, Y. A short-term intervention for long-term fairness in the labor market. In *Proceedings of the 2018 World Wide Web Conference on World Wide Web*, pp. 1389–1398. International World Wide Web Conferences Steering Committee, 2018.

Huang, C.-L., Chen, M.-C., and Wang, C.-J. Credit scoring with a data mining approach based on support vector machines. *Expert systems with applications*, 33(4):847–856, 2007.

Kamishima, T., Akaho, S., Asoh, H., and Sakuma, J. Fairness-aware classifier with prejudice remover regularizer. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pp. 35–50. Springer, 2012.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kohavi, R. Scaling up the accuracy of naive-bayes classifiers: A decision-tree hybrid. In *Kdd*, volume 96, pp. 202–207. Citeseer, 1996.

Kusner, M. J., Loftus, J., Russell, C., and Silva, R. Counterfactual fairness. In *Advances in Neural Information Processing Systems*, pp. 4066–4076, 2017.

Narayanan, A. Tutorial: 21 fairness definitions and their politics. *Conference on Fairness, Accountability, and Transparency*, Mar 2018. URL https://www.youtube.com/watch?v=jIXIuYdnyyk.

Paszke, A., Gross, S., Chintala, S., and Chanan, G. Pytorch: Tensors and dynamic neural networks in python with strong gpu acceleration. *PyTorch: Tensors and dynamic neural networks in Python with strong GPU acceleration*, 6, 2017.

Poplin, R., Varadarajan, A. V., Blumer, K., Liu, Y., McConnell, M. V., Corrado, G. S., Peng, L., and Webster, D. R. Prediction of cardiovascular risk factors from retinal fundus photographs via deep learning. *Nature Biomedical Engineering*, 2(3):158, 2018.

Tollenaar, N. and Van der Heijden, P. Which method predicts recidivism best?: a comparison of statistical, machine learning and data mining predictive models. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 176(2):565–584, 2013.

Wadsworth, C., Vera, F., and Piech, C. Achieving fairness through adversarial learning: an application to recidivism prediction. *arXiv preprint arXiv:1807.00199*, 2018.

Zafar, M. B., Valera, I., Rodriguez, M. G., and Gummadi, K. P. Fairness constraints: Mechanisms for fair classification. *arXiv preprint arXiv:1507.05259*, 2015.