# One-network Adversarial Fairness

**Tameem Adel**
University of Cambridge, UK
tah47@cam.ac.uk

**Isabel Valera**
MPI-IS, Germany
isabel.valera@tuebingen.mpg.de

**Zoubin Ghahramani**
University of Cambridge, UK
Uber AI Labs, USA
zoubin@eng.cam.ac.uk

**Adrian Weller**
University of Cambridge, UK
The Alan Turing Institute, UK
aw665@cam.ac.uk

**Abstract**

There is currently a great expansion of the impact of machine learning algorithms on our lives, prompting the need for objectives other than pure performance, including fairness. Fairness here means that the outcome of an automated decision-making system should not discriminate between subgroups characterized by sensitive attributes such as gender or race. Given any existing differentiable classifier, we make only slight adjustments to the architecture including adding a new hidden layer, in order to enable the concurrent adversarial optimization for fairness and accuracy. Our framework provides one way to quantify the tradeoff between fairness and accuracy, while also leading to strong empirical performance.

## 1 Introduction

Automated decision support has become widespread, raising concerns about potential unfairness. Here, following earlier work, unfairness means discriminating against a particular group of people due to sensitive group characteristics such as gender or race (Grgic-Hlaca et al. 2018b; Hardt, Price, and Srebro 2016; Kusner et al. 2017; Louizos et al. 2016; Zafar et al. 2017c; 2017a; Zemel et al. 2013). Fairness concerns have been considered in applications including predictive policing (Brennan, Dieterich, and Ehret 2009), recidivism prediction (Chouldechova 2017) and credit scoring (Khandani, Kim, and Lo 2010). The current trend in the literature on fairness is to begin with a notion (definition) of fairness, and then to construct a model which automates the detection and/or eradication of unfairness accordingly. Common definitions of fairness consider whether or not a decision is related to sensitive attributes, such as gender.

Most state-of-the-art machine learning algorithms for fairness build from scratch a model's architecture tailored for a specific fairness notion. In contrast, we propose a method that slightly modifies the architecture of the model which was to be optimized solely for accuracy. We learn a fair representation together with a new performance function acting on it, with the goal of concurrently optimizing for both fairness and performance (accuracy). Our method is based on an adversarial framework, which allows explicitly measuring the tradeoff between fairness and accuracy. Our approach is general, in that it may be applied to any differentiable discriminative model. We establish a fairness paradigm where the architecture of a deep discriminative model, optimized for accuracy, is modified such that fairness is imposed (the same paradigm could be applied to a deep generative model in future work). Beginning with an ordinary neural network optimized for prediction accuracy of the class labels in a classification task, we propose an adversarial fairness framework performing a change to the network architecture, leading to a neural network that is maximally uninformative about the sensitive attributes of the data as well as predictive of the class labels. In this adversarial learning framework, there is no need for a separate network architecture representing the adversary, thus we avoid the well-known difficulties which may arise from double-network adversarial learning (Goodfellow 2016; Goodfellow et al. 2014).

We make the following contributions: 1) We propose a fairness algorithm by modifying the architecture of a potentially unfair model to simultaneously optimize for both accuracy and fairness with respect to sensitive attributes (Section 2); 2) We quantify the tradeoff between the accuracy and fairness objectives (Section 2); 3) We propose a variation of the adversarial learning procedure to increase diversity among elements of each minibatch of the gradient descent training, in order to achieve a representation with higher fidelity. This variation may be of independent interest since it is applicable generally to adversarial frameworks (Section 2.2); 4) We develop a novel generalization bound for our framework illustrating the theoretical grounding behind the relationship between the label classifier and the adversary's ability to predict the sensitive attribute value (Section 3); and 5) We experiment on two fairness datasets comparing against many earlier approaches to demonstrate state-of-the-art effectiveness of our methods (Section 4).

## 2 Fair Learning by Modifying an Unfair Model

We aim to adapt a classifier that was to learn solely to predict the data labels into a fair classifier without reconstructing its architecture from scratch. Our strategy is to add a term to the optimization objective to simultaneously maximize both fairness and accuracy. This term imposes fairness by establishing a fair representation of the input data which

is invariant to changes in the sensitive attribute values. To optimize this invariance, an adversary is included at the top of the model in the form of a classifier which tries to predict the sensitive attribute value. For a deep model, this can be implemented via: (i) adding another classifier at the top of the network such that we have two classifiers: the original label predictor and the new predictor of sensitive attribute value(s); and (ii) adding a network layer just under the classifiers (now the top hidden layer) that aims at maximizing the performance of the label predictor, while minimizing the performance of the sensitive attribute predictor by reversing the gradients of the latter in the backward phase of the gradient descent training. See Figure 1 for an illustration.

## 2.1 Fair adversarial discriminative (FAD) model

We consider data $D = \{(\mathbf{x_i}, \mathbf{y_i}, \mathbf{s_i})\}_{\mathbf{i=1}}^{\mathbf{n}}$, where $\mathbf{x_i}$, $\mathbf{y_i}$ and $\mathbf{s_i}$ refer to the input features, the ground truth label and the sensitive attribute(s), respectively. Denote by $\hat{\mathbf{y}}_\mathbf{i}$ the corresponding label estimate. Typically speaking, $\hat{\mathbf{y}}_\mathbf{i}$ is predicted by a potentially unfair discriminative classifier possibly represented by a neural network whose input is features $\mathbf{x}$ and sensitive attributes $\mathbf{s}$. The goal of fairness is to ensure that the prediction of $\mathbf{y}$ is not dependent on these attributes $\mathbf{s}$ (Zafar et al. 2017c). A naive approach is to simply discard $\mathbf{s}$ and let classifier $\mathbf{f}$ have input $\mathbf{x}$ - this has some proponents in terms of process (Grgic-Hlaca et al. 2018a; 2018b) but suffers since $\mathbf{x}$ might include significant information about $\mathbf{s}$. For example, an address attribute in $\mathbf{x}$ might give a strong indication about race $\mathbf{s}$.

We achieve fairness through the adversarial construction illustrated in Figures 1 and 2, as follows. Define the empirical classification loss of the label classifier $\mathbf{f}'(\mathbf{x}')$ as $L(\hat{\mathbf{y}}, \mathbf{y}) = L(\mathbf{f}'(\mathbf{x}'), \mathbf{y}) = \frac{1}{|D|}\sum_{i=1}^{|D|} \mathbf{Err}(\mathbf{f}'(\mathbf{x}'_\mathbf{i}), \mathbf{y_i})$, and similarly the classification loss of the classifier $\boldsymbol{\phi}'$, predicting the value of the sensitive attribute $\mathbf{s}$, as $L(\boldsymbol{\phi}'(\mathbf{x}'), \mathbf{s}) = \frac{1}{|D|}\sum_{i=1}^{|D|} \mathbf{Err}(\boldsymbol{\phi}'(\mathbf{x}'_\mathbf{i}), \mathbf{s_i})$. A simplified view of the architecture of our fairness adversarial discriminative (FAD) framework is displayed in Figure 1. We extend the original architecture of the potentially unfair predictor, by adding a new layer $\mathbf{g}$ that produces a fair data representation $\mathbf{x}'$, and adding a new predictor $\boldsymbol{\phi}'$ aiming to predict the sensitive attribute $\mathbf{s}$. Denoting by $\mathbf{w}$ the weight vector of a network layer, the backpropagation forward pass proceeds normally in FAD. During the backward pass (dotted lines), however, the gradient from the loss of the $\mathbf{s}$ classifier $\frac{\partial L(\boldsymbol{\phi}'(\mathbf{x}'),\mathbf{s})}{\partial w_{\boldsymbol{\phi}'}}$ is multiplied with a negative sign so that $\mathbf{g}$ adversarially aims at increasing the loss of $\boldsymbol{\phi}'$, resulting in a representation $\mathbf{x}'$ maximally invariant to the change in values of $\mathbf{s}$. Since the rest of the FAD training proceeds normally, the classifier $\mathbf{f}'$ should also make $\mathbf{x}'$ maximally informative about the original prediction task, depicted by $\mathbf{y}$. The idea of reversing a layer's gradient with respect to the layer below in the backward pass was used in domain adaptation to develop invariant representations to changes between the source and target domains (Ajakan et al. 2014; Ganin and Lempitsky 2015; Ganin et al. 2016).

In our fairness paradigm, we aim at the following: i) achieving fairness; ii) quantitatively evaluating how much fairness we achieve, and; iii) quantifying the impact of such fairness on accuracy, i.e. computing the difference in accuracy between the proposed fair model and a corresponding (potentially unfair) model. In Figure 2, another schematic diagram of the proposed modifications is displayed. The predictor $\boldsymbol{\phi}'(\mathbf{x}')$ depicts a classifier predicting the $\mathbf{s}$ value from $\mathbf{x}'$. The ability to accurately predict $\mathbf{s}$ given $\mathbf{x}'$ signifies a high risk of unfairness since this means that the sensitive attributes may be influential in any decision making process based on $\mathbf{x}'$. Adversarially through $\mathbf{g}$, the non-sensitive features $\mathbf{x}$ are transformed into a representation $\mathbf{x}'$ which attempts to break any dependence on the sensitive attributes $\mathbf{s}$. As such, the optimization objective can have many trivial solutions, e.g. $\mathbf{g}$ that maps every $\mathbf{x}$ into 0 which provides no information to predict $\mathbf{s}$. In order to prevent that, and to ensure that fairness is rather aligned with the accuracy objective, a classifier $\mathbf{f}'$ predicts the labels $\mathbf{y}$ from $\mathbf{x}'$. Here we consider classification accuracy to be the metric of the initial model. However, our adversarial paradigm can also be adopted in models with other metrics to achieve fairness.

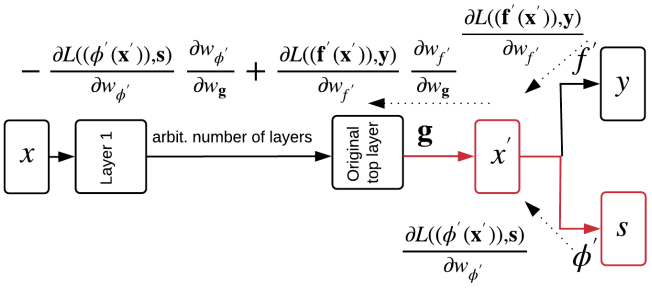



Figure 1: Architecture of the proposed fair adversarial discriminative model. The parts added, due to FAD, to a potentially unfair deep architecture with input $\mathbf{x}$ are (shown in red): i) the layer $\mathbf{g}$ where $\mathbf{x}'$ is learned and; 2) the sensitive attribute $\mathbf{s}$ predictor $\boldsymbol{\phi}'$ at the top of the network. Denote by $\mathbf{w}$ the weight vector of the respective layer. The forward pass of backpropagation proceeds normally in FAD. During the backward pass (dotted lines), the gradient from the loss of the $\mathbf{s}$ classifier $\frac{\partial L((\phi'(\mathbf{x}')),\mathbf{s})}{\partial w_{\phi'}}$ is multiplied with a negative sign so that $\mathbf{g}$ adversarially aims at increasing the loss of $\boldsymbol{\phi}'$, resulting in a representation $\mathbf{x}'$ maximally invariant to the change in values of $\mathbf{s}$. The rest of the FAD training proceeds normally, i.e. gradient from the labeling classifier $\mathbf{f}'$, $\frac{\partial L((\mathbf{f}'(\mathbf{x}')),\mathbf{y})}{\partial w_{f'}}$, is normally (with a positive sign) imported to $\mathbf{g}$, ultimately making $\mathbf{x}'$ maximally informative about $\mathbf{y}$.

To summarize: $\mathbf{g}$, $\boldsymbol{\phi}'$ and $\mathbf{f}'$ are all involved in the concurrent optimization for fairness and accuracy. The parameters of $\mathbf{g}$ and $\boldsymbol{\phi}'$ adversarially optimize for fairness while $\mathbf{f}'$ guarantees that accuracy is not fully sacrificed for the sake of fairness. The labeling accuracy difference between the proposed fair model and a corresponding potentially unfair model may be quantified via the discrepancy between

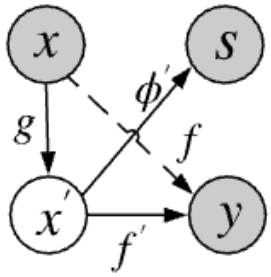



Figure 2: The proposed FAD paradigm. We establish an adversarial objective where $\mathbf{g}$ minimizes the ability of the predictor $\phi'$ to correctly predict values of the sensitive attributes $\mathbf{s}$, whereas $\phi'$ on the other hand maximizes its own accuracy in predicting $\mathbf{s}$. Another predictor $\mathbf{f}'$ predicts the labels $\mathbf{y}$ from $\mathbf{x}'$.

$\mathbf{f}'(\mathbf{x}')$ obtained after learning $\mathbf{g}$, and $\mathbf{f}(\mathbf{x})$, respectively.

**Fairness notions** We focus on two common notions (definitions) of fairness, `disparate impact` (Barocas and Selbst 2016; Feldman et al. 2015; Primus 2010) and `disparate mistreatment` (Zafar et al. 2017b). We first address the former. Disparate impact refers to a decision making process that, in aggregate, leads to different outcomes for subpopulations with different sensitive attribute values. According to this notion, and assuming WLOG a binary label and sensitive attribute, fairness is achieved (in other words, there is no disparate impact) when:

$$\mathbf{p}(\hat{\mathbf{y}} = 1|\mathbf{s} = 0) = \mathbf{p}(\hat{\mathbf{y}} = 1|\mathbf{s} = 1). \tag{1}$$

For our proposed model, define the part of the data $D$ where $\mathbf{s} = 0$ as $D_{\mathbf{s}=0}$, then $\mathbf{p}(\hat{\mathbf{y}} = 1|\mathbf{s} = 0) = \frac{1}{|D_{\mathbf{s}=0}|}\sum_{i=1}^{|D_{\mathbf{s}=0}|}\mathbf{p}(\mathbf{f}'(\mathbf{x}') = 1)$. Accordingly, disparate impact is avoided when:

$$\tfrac{1}{|D_{\mathbf{s}=0}|}\textstyle\sum_{i=1}^{|D_{\mathbf{s}=0}|}\mathbf{p}(\mathbf{f}'(\mathbf{x}') = 1) = \tfrac{1}{|D_{\mathbf{s}=1}|}\sum_{i=1}^{|D_{\mathbf{s}=1}|}\mathbf{p}(\mathbf{f}'(\mathbf{x}') = 1) \tag{2}$$

The disparity (Disp) between both sides of (2) quantifies unfairness in terms of disparate impact. Larger disparity values signify more unfairness:

$$\text{Disp}_{\text{DI}} = \tag{3}$$
$$\left|\tfrac{1}{|D_{\mathbf{s}=0}|}\textstyle\sum_{i=1}^{|D_{\mathbf{s}=0}|}\mathbf{p}(\mathbf{f}'(\mathbf{x}') = 1) - \tfrac{1}{|D_{\mathbf{s}=1}|}\sum_{i=1}^{|D_{\mathbf{s}=1}|}\mathbf{p}(\mathbf{f}'(\mathbf{x}') = 1)\right|$$

The other notion of fairness we inspect is referred to as `disparate mistreatment` (Zafar et al. 2017b). One major difference between the two notions is that disparate mistreatment depends on the ground truth labels $\mathbf{y}$, and therefore it can be considered only when such information is available. Assuming a binary label $\mathbf{y}$ with ground truth values 1 and $-1$, disparate mistreatment arises when the rate of erroneous decisions (false positives, false negatives or both) is different for subpopulations with different values of sensitive attributes. Unfairness in terms of disparate mistreatment is thus avoided when the false positive rate (FPR) and false negative rate (FNR) are as in (4) and (5), respectively.

$$\mathbf{p}(\hat{\mathbf{y}} \neq \mathbf{y}|\mathbf{y} = -1, \mathbf{s} = 0) = \mathbf{p}(\hat{\mathbf{y}} \neq \mathbf{y}|\mathbf{y} = -1, \mathbf{s} = 1) \tag{4}$$
$$\mathbf{p}(\hat{\mathbf{y}} \neq \mathbf{y}|\mathbf{y} = 1, \mathbf{s} = 0) = \mathbf{p}(\hat{\mathbf{y}} \neq \mathbf{y}|\mathbf{y} = 1, \mathbf{s} = 1) \tag{5}$$

A larger disparity between both sides in either of these two equations signifies more disparate mistreatment. For our model, disparate mistreatment, as a whole, is defined in (6).

$$\tfrac{1}{|D_{\mathbf{s}=0}|}\textstyle\sum_{i=1}^{|D_{\mathbf{s}=0}|}\mathbf{Err}(\mathbf{f}'(\mathbf{x}'_\mathbf{i}), \mathbf{y_i}) = \tfrac{1}{|D_{\mathbf{s}=1}|}\sum_{i=1}^{|D_{\mathbf{s}=1}|}\mathbf{Err}(\mathbf{f}'(\mathbf{x}'_\mathbf{i}), \mathbf{y_i}) \tag{6}$$

Decomposing (6) into FPR and FNR is straightforward. The higher the following disparity values the more unfair the model is, in terms of disparate mistreatment:

$$\text{Disp}_{\text{FPR}} = \tag{7}$$
$$\left|\tfrac{1}{|D_{\mathbf{s}=0}|}\textstyle\sum_{i=1}^{|D_{\mathbf{s}=0}|}\mathbf{Err}(\mathbf{f}'(\mathbf{x}'_\mathbf{i}), \mathbf{y_i}|\mathbf{y_i}{=}{-1}) - \tfrac{1}{|D_{\mathbf{s}=1}|}\sum_{i=1}^{|D_{\mathbf{s}=1}|}\mathbf{Err}(\mathbf{f}'(\mathbf{x}'_\mathbf{i}), \mathbf{y_i}|\mathbf{y_i}{=}{-1})\right|$$

$$\text{Disp}_{\text{FNR}} = \tag{8}$$
$$\left|\tfrac{1}{|D_{\mathbf{s}=0}|}\textstyle\sum_{i=1}^{|D_{\mathbf{s}=0}|}\mathbf{Err}(\mathbf{f}'(\mathbf{x}'_\mathbf{i}), \mathbf{y_i}|\mathbf{y_i}{=}1) - \tfrac{1}{|D_{\mathbf{s}=1}|}\sum_{i=1}^{|D_{\mathbf{s}=1}|}\mathbf{Err}(\mathbf{f}'(\mathbf{x}'_\mathbf{i}), \mathbf{y_i}|\mathbf{y_i}{=}1)\right|$$

**Modeling objective** The overall training objective of the proposed model, for a dataset of size $\mathbf{n}$, is stated as follows:

$$\min_{\mathbf{g},\mathbf{f}'}\left[L(\mathbf{f}'(\mathbf{g}(\mathbf{x})), \mathbf{y}) - \beta\ \max_{\phi'}(L(\phi'(\mathbf{g}(\mathbf{x})), \mathbf{s}))\right] \equiv \tag{9}$$
$$\min_{\mathbf{g},\mathbf{f}'}\left[\tfrac{1}{\mathbf{n}}\textstyle\sum_{i=1}^{\mathbf{n}}\mathbf{Err}(\mathbf{f}'(\mathbf{g}(\mathbf{x}_i)), \mathbf{y}_i) - \beta\max_{\phi'}\left(\tfrac{1}{\mathbf{n}}\sum_{i=1}^{\mathbf{n}}\mathbf{Err}(\phi'(\mathbf{g}(\mathbf{x}_i)), \mathbf{s}_i)\right)\right],$$

where the fairness hyperparameter $\beta > 0$ controls the degree of fairness induced in the model.

The model, as such, is based on optimizing for disparate impact. Recall that the whole proposed model is ultimately implemented as one neural network in which the adversarial layer $\mathbf{g}$ is injected as a hidden layer right beneath the top layer consisting of the classifiers $\phi'$ and $\mathbf{f}'$.

In order to optimize for disparate mistreatment, a modification to (9) is needed. The function $\phi'(\mathbf{x}') = \phi'(\mathbf{g}(\mathbf{x}))$ should be multiplied by a term reflecting what is needed to be equal for subpopulations with different sensitive attribute values such that disparate mistreatment is achieved. Multiplying $\phi'(\mathbf{g}(\mathbf{x}))$ with the ratio of $\mathbf{Err}(\mathbf{f}'(\mathbf{x}'_\mathbf{i}), \mathbf{y_i})$ over $\mathbf{n}$, which signifies the misclassification rate, leads to the objective becoming an interpretation of disparate mistreatment. The overall training objective when optimizing for disparate mistreatment therefore becomes:

$$\min_{\mathbf{g},\mathbf{f}'}\left[\tfrac{1}{\mathbf{n}}\textstyle\sum_{i=1}^{\mathbf{n}}\mathbf{Err}(\mathbf{f}'(\mathbf{g}(\mathbf{x}_i)), \mathbf{y}_i) - \beta\max_{\phi'}\left(\tfrac{1}{\mathbf{n}}\sum_{i=1}^{\mathbf{n}}\mathbf{Err}(\phi'(\mathbf{g}(\mathbf{x}_i)), \mathbf{s}_i)\mathbf{I}(\hat{y}_i \neq y_i)\right)\right], \tag{10}$$

where $\mathbf{I}(\mathbf{z})$ is the indicator function: $I(\mathbf{z}) = 1$, when $\mathbf{z}$ is true and 0 otherwise. However, since $\hat{y}$ is the classification output, which depends on both $\mathbf{g}$ and $\mathbf{f}'$, the indicator function might be problematic to optimize by gradient descent. Hence, we can relax this optimization problem and rewrite it as follows to express the labeling misclassification probability:

$$\min_{\mathbf{g},\mathbf{f}'}\left[\tfrac{1}{\mathbf{n}}\textstyle\sum_{i=1}^{\mathbf{n}}\mathbf{Err}(\mathbf{f}'(\mathbf{g}(\mathbf{x}_i)), \mathbf{y}_i) - \beta\max_{\phi'}\left(\tfrac{1}{\mathbf{n}}\sum_{i=1}^{\mathbf{n}}\mathbf{Err}(\phi'(\mathbf{g}(\mathbf{x}_i)), \mathbf{s}_i)p(\hat{y}_i \neq y_i)\right)\right] \equiv$$
$$\min_{\mathbf{g},\mathbf{f}'}\left[\tfrac{1}{\mathbf{n}}\textstyle\sum_{i=1}^{\mathbf{n}}\mathbf{Err}(\mathbf{f}'(\mathbf{g}(\mathbf{x}_i)), \mathbf{y}_i) - \beta\max_{\phi'}\left(\sum_{i=1}^{\mathbf{n}}\frac{\mathbf{Err}(\mathbf{f}'(\mathbf{x}'_\mathbf{i}), \mathbf{y_i})}{\mathbf{n}}\mathbf{Err}(\phi'(\mathbf{g}(\mathbf{x}_i)), \mathbf{s}_i)\right)\right] \tag{11}$$

For the decomposition of disparate mistreatment into FPR and FNR, the term $\mathbf{Err}(\mathbf{f}'(\mathbf{x}'_\mathbf{i}), \mathbf{y_i}) = \mathbf{Err}(\mathbf{f}'(\mathbf{g}(\mathbf{x}_i)), \mathbf{y_i})$ in (11) shall be divided into $\mathbf{Err}(\mathbf{f}'(\mathbf{x}'_\mathbf{i}), \mathbf{y_i})_{\mathbf{y_i}=0}$ and $\mathbf{Err}(\mathbf{f}'(\mathbf{x}'_\mathbf{i}), \mathbf{y_i})_{\mathbf{y_i}=1}$, respectively.

## 2.2 Fair adversarial discriminative model with increased minibatch diversity (FAD-MD)

Although our approach avoids two 'dueling' networks, still it is prone to some of the common problems of adversarial learning, including 'mode collapse' (Goodfellow 2016;

Goodfellow et al. 2014). Imagine if the input to the discriminator is MNIST images, and, instead of covering the whole space of the ten digits, the comparisons are restricted to two or three digits that are ultimately being represented quite similarly by data of both classes introduced as input to the discriminator. The latter therefore becomes maximally confused and a convergence of the whole adversarial learning framework is reached although in reality only a small subset of the data is well represented by the inferred representation. Since we want to avoid imposing too many variations on the original settings of the system, we need to develop a computationally efficient solution to mode collapse. Most of the previously proposed solutions are either tailored to the generative versions of adversarial learning (Arjovsky, Chintala, and Bottou 2017; Kim et al. 2017; Metz et al. 2017; Rosca et al. 2017) or computationally inefficient for our purpose (Srivastava et al. 2017). We experimented with the heuristics proposed in (Salimans et al. 2016) but they did not improve results in our one-network adversarial framework.

Therefore we propose a method based on making elements of a minibatch as diverse as possible from one another. Each minibatch is formed as follows: i) Beginning with very few data points randomly chosen to belong to the minibatch, there is a pool of points from which some are to be selected for the addition to the minibatch based on the criterion of ending up with minibatch elements as diverse as possible without losing too much in terms of computational runtime; ii) A point is selected from the pool via the score resulting from a one-class support vector machine (SVM) classifier where the class consists of the current elements of the minibatch. The next added data point from the pool to the minibatch is the point with the lowest score, i.e. the point believed to be the least likely to belong to the class formed by the current minibatch elements, or the most dissimilar point to the minibatch elements; iii) This process continues until greedily reaching the prespecified size of the minibatch. The size of each pool of points to begin with is a hyperparameter to be specified by e.g. cross-validation. We claim that using the efficient one-class SVM to form minibatches is a step in the direction of establishing adversarial learners with a better coverage of all the regions of the data space.

## 3 Generalization Bound

We illustrate the theoretical foundation of the relationship between the label classifier and the adversary's ability to predict the sensitive attribute value. We begin by deriving a generalization upper bound for the framework (Theorem 1). We shed light on the interpretation of the sensitive attribute classifier as an adversary in the concurrent adversarial optimization for fairness and accuracy, in terms of the distance between data distributions given different sensitive attribute values.

Assume the label $\mathbf{y}$ (as well as its estimate $\hat{\mathbf{y}}$) and the sensitive attribute $\mathbf{s}$ values are binary. Recall that a data sample $D = \{(\mathbf{x_i}, \mathbf{y_i}, \mathbf{s_i})\}_{\mathbf{i}=1}^{\mathbf{n}}$ is given. Also, recall that, for the $\mathbf{i}^{th}$ instance, the classification loss between $\hat{\mathbf{y}}$ and $\mathbf{y}$ (equivalently defined between any two labelings) is denoted by $\mathbf{Err}(\hat{\mathbf{y}}_\mathbf{i}, \mathbf{y_i})$. Denote the expected loss between two labelings $\hat{\mathbf{y}}$ and $\mathbf{y}$ with respect to a distribution $\mathbf{P}$ by $L_\mathbf{P}(\hat{\mathbf{y}}, \mathbf{y}) = \mathbb{E}_\mathbf{P}[\mathbf{Err}(\hat{\mathbf{y}}, \mathbf{y})]$. When abbreviated as $\mathbf{Err}(\hat{\mathbf{y}})$, this means that the other side is the ground truth label $\mathbf{y}$.

Next we describe the Rademacher complexity, which measures the richness of a class of real-valued functions with respect to a probability distribution (Shalev-Shwartz and Ben-David 2014).

Rademacher complexity facilitates the derivation of general learning guarantees for problems with infinite hypothesis sets, like ours. Thanks to the introduction of a Rademacher random variable $\sigma$, Rademacher complexity directly maps the measurement of accuracy (or inversely error rate) of the hypothesis into the richness or expressiveness of the hypothesis w.r.t. the probability distribution resulting from the introduction of $\sigma$. More formally, let $\mathbf{H}$ be a hypothesis set, where $\mathbf{h} \in \mathbf{H}$ is an arbitrary hypothesis of the set. Define the expected loss of a hypothesis class as $\mathbf{Err}(\mathbf{H})$[1]. When each $\mathbf{h}$ is a real-valued function, the *empirical Rademacher complexity* (Mansour, Mohri, and Rostamizadeh 2009; Mohri and Afshin 2008a; Shalev-Shwartz and Ben-David 2014) of $\mathbf{H}$ given $D$ can be defined as:

$$\mathbf{Rad}_D(\mathbf{H}) = \frac{1}{\mathbf{n}}\mathbb{E}_\sigma[\sup_{\mathbf{h}\in\mathbf{H}} \sum_{\mathbf{i}=1}^{\mathbf{n}} \sigma_i \mathbf{h}(\mathbf{x_i})], \tag{12}$$

where $\sigma_1$, $\sigma_2$, $\cdots$, $\sigma_n$ are independent random variables drawn from the Rademacher distribution, which notes that $Pr(\sigma = 1) = Pr(\sigma = -1) = 0.5$. The *Rademacher complexity* of a hypothesis set $\mathbf{H}$ is thence denoted by the expectation of $\mathbf{Rad}_D(\mathbf{H})$ over all samples of size $\mathbf{n}$:

$$\mathbf{Rad_n}(\mathbf{H}) = \mathbb{E}_D[\mathbf{Rad}_D(\mathbf{H})] \tag{13}$$

For deriving our bound, we use a notion of distance between distributions, proposed by Mansour, Mohri, and Rostamizadeh (2009), and referred to as the discrepancy distance. It is symmetric and it satisfies the triangle inequality. Denote by $\mathbf{H}$ a hypothesis set, where each $\mathbf{h} \in \mathbf{H}$ is a classifier $\mathbf{X} \to \mathbf{Y}$. The discrepancy distance between two arbitrary distributions $\mathbf{P}$ and $\mathbf{Q}$ is defined as:

$$\mathbf{disc}(\mathbf{P}, \mathbf{Q}) = \max_{\mathbf{h},\mathbf{h}'\in\mathbf{H}} |L_\mathbf{P}(\mathbf{h}, \mathbf{h}') - L_\mathbf{Q}(\mathbf{h}, \mathbf{h}')| \tag{14}$$

We also state the following bound (Bartlett and Mendelson 2002; Koltchinskii and Panchenko 2000), whose proof is in (Bartlett and Mendelson 2002). For a probability distribution $\mathbf{P}$ (estimated by $\hat{\mathbf{P}}$) defined on $\mathbf{X} \times \{\pm 1\}$, let $\mathbf{H}$ be a hypothesis set where each $\mathbf{h} \in \mathbf{H}$ is a $\{\pm 1\}$-valued function mapping $\mathbf{X}$ to a binary $\mathbf{Y}$. Then for a data sample $D$ with size $\mathbf{n}$, With probability at least $1 - \delta$, and for a 0-1 classification loss, every hypothesis $\mathbf{h} \in \mathbf{H}$ satisfies:

$$\mathbb{E}_\mathbf{P}(\mathbf{Err}(\mathbf{Y}, \mathbf{h}(\mathbf{X}))) \leq \mathbb{E}_{\hat{\mathbf{P}}}(\mathbf{Err}(\mathbf{Y}, \mathbf{h}(\mathbf{X}))) + \frac{\mathbf{Rad}_D(\mathbf{H})}{2} + \sqrt{\frac{\log(1/\delta)}{2\mathbf{n}}} \tag{15}$$

[1]Note that this is the expectation of a class of functions with respect to a specific data sample $D$. This differs from $L_\mathbf{P}(\mathbf{h}(\mathbf{x}), \mathbf{y})$ since the latter is the expectation of a specific function $\mathbf{h}$ with respect to a probability distribution $\mathbf{P}$ from which $D$ is drawn.

**Theorem 1.** *Denote by* $\mathbf{P_0}$ *and* $\mathbf{P_1}$ *the distributions* $\mathbf{P}(\mathbf{x}|\mathbf{s}=0)$ *and* $\mathbf{P}(\mathbf{x}|\mathbf{s}=1)$*, respectively. Assuming training data* $D$ *with size* $\mathbf{n}$*, which, without loss of generality (WLOG), is equally divided into data points with* $\mathbf{s}=0$*,* $D_\mathbf{0}$*, and those with* $\mathbf{s}=1$*,* $D_\mathbf{1}$*. For a class* $\mathbf{H}$ *of binary classifiers, and for* $\delta > 0$ *with probability greater than or equal to* $1-\delta$*, the following holds:*

$$\begin{aligned}\mathbf{disc}(\mathbf{P_0},\mathbf{P_1}) \leq & \mathbf{disc}(\hat{\mathbf{P}}_\mathbf{0},\hat{\mathbf{P}}_\mathbf{1}) + 2\sqrt{\frac{\log(1/\delta)}{\mathbf{n}}} \\ & + \frac{1}{2}(\mathbf{Rad}_{D_\mathbf{0}}(\mathbf{Err}(\mathbf{H}))+\mathbf{Rad}_{D_\mathbf{1}}(\mathbf{Err}(\mathbf{H})))\end{aligned} \tag{16}$$

*Proof.* Using the triangle inequality, the L.H.S. of (16) can turn into:

$$\mathbf{disc}(\mathbf{P_0},\mathbf{P_1}) \leq \mathbf{disc}(\mathbf{P_0},\hat{\mathbf{P}}_\mathbf{0}) + \mathbf{disc}(\mathbf{P_1},\hat{\mathbf{P}}_\mathbf{1}) + \mathbf{disc}(\hat{\mathbf{P}}_\mathbf{0},\hat{\mathbf{P}}_\mathbf{1}) \tag{17}$$

For a 0-1 classification error, we can use (15) with the class of real-valued functions being a class where each element is the loss of $\mathbf{h}$, $\mathbf{Err}(\mathbf{h})$, on a specific training sample $D$, instead of $\mathbf{h}$ itself. We can as well replace $\mathbf{y}$ in (15) with an arbitrary $\mathbf{h}$, and since we are upper bounding we can assume we replace it with the worst case $\mathbf{h}$, i.e. $\mathbf{h}$ resulting in the highest upper bound. This leads to:

$$\begin{aligned}\mathbb{E}_\mathbf{P}(\mathbf{Err}(\mathbf{h}(\mathbf{x}),\mathbf{h}'(\mathbf{x}))) \leq & \mathbb{E}_{\hat{\mathbf{P}}}(\mathbf{Err}(\mathbf{h}(\mathbf{x}),\mathbf{h}'(\mathbf{x}))) \\ & + \frac{\mathbf{Rad}_D(\mathbf{Err}(\mathbf{H}))}{2} + \sqrt{\frac{\log(1/\delta)}{2\mathbf{n}}}\end{aligned} \tag{18}$$

Since $\mathbb{E}_\mathbf{P}(\mathbf{Err}(\mathbf{h}(\mathbf{x}),\mathbf{h}'(\mathbf{x}))) = L_\mathbf{P}(\mathbf{h}(\mathbf{x}),\mathbf{h}'(\mathbf{x}))$, and from (14), then (18) turns into:

$$\mathbf{disc}(\mathbf{P},\hat{\mathbf{P}}) \leq \frac{\mathbf{Rad}_D(\mathbf{Err}(\mathbf{H}))}{2} + \sqrt{\frac{\log(1/\delta)}{2\mathbf{n}}} \tag{19}$$

Using (19) to describe the terms $\mathbf{disc}(\mathbf{P_0},\hat{\mathbf{P}}_\mathbf{0})$ and $\mathbf{disc}(\mathbf{P_1},\hat{\mathbf{P}}_\mathbf{1})$ in (17), we get:

$$\begin{aligned}\mathbf{disc}(\mathbf{P_0},\mathbf{P_1}) \leq & \mathbf{disc}(\hat{\mathbf{P}}_\mathbf{0},\hat{\mathbf{P}}_\mathbf{1}) \\ & + \frac{\mathbf{Rad}_{D_\mathbf{0}}(\mathbf{Err}(\mathbf{H}))}{2} + \sqrt{\frac{\log(1/\delta)}{\mathbf{n}}} \\ & + \frac{\mathbf{Rad}_{D_\mathbf{1}}(\mathbf{Err}(\mathbf{H}))}{2} + \sqrt{\frac{\log(1/\delta)}{\mathbf{n}}}\end{aligned} \tag{20}$$

$$\begin{aligned}\mathbf{disc}(\mathbf{P_0},\mathbf{P_1}) \leq & \mathbf{disc}(\hat{\mathbf{P}}_\mathbf{0},\hat{\mathbf{P}}_\mathbf{1}) + 2\sqrt{\frac{\log(1/\delta)}{\mathbf{n}}} \\ & + \frac{1}{2}(\mathbf{Rad}_{D_\mathbf{0}}(\mathbf{Err}(\mathbf{H})) + \mathbf{Rad}_{D_\mathbf{1}}(\mathbf{Err}(\mathbf{H})))\end{aligned} \tag{21}$$

which concludes the proof. □

This provides an interpretation of our modeling objective in (9) in the main document, since: The sensitive attribute classifier $\boldsymbol{\phi}'(\mathbf{g}(\mathbf{x}))$ aims at minimizing the first term in the bound on the right of (21), $\mathbf{disc}(\hat{\mathbf{P}}_\mathbf{0},\hat{\mathbf{P}}_\mathbf{1})$. The label classifier $\mathbf{f}'(\mathbf{g}(\mathbf{x}))$ aims at minimizing the second term on the right of (21). The third term on the right of (21) tends to zero as the sample size $\mathbf{n}$ goes to infinity.

To further clarify that: As noted in (Mansour, Mohri, and Rostamizadeh 2009), for a 0-1 classification error and for our hypothesis class consisting of the loss on each hypothesis, $\mathbf{Err}(\mathbf{h})$, the discrepancy distance $\mathbf{disc}(\hat{\mathbf{P}}_\mathbf{0},\hat{\mathbf{P}}_\mathbf{1})$ is equivalent to the following notion of distance, referred to as $\mathcal{H}$-divergence (Ben-David et al. 2007; 2010; Devroye, Gyorfi, and Lugosi 1996; Kifer, Ben-David, and Gehrke 2004):

$$\mathbf{d}_\mathcal{H}(\hat{\mathbf{P}}_\mathbf{0},\hat{\mathbf{P}}_\mathbf{1}) = \sup_{\mathbf{a}\in|\mathbf{h}-\mathbf{h}'|} |\hat{\mathbf{P}}_\mathbf{0}(\mathbf{a}) - \hat{\mathbf{P}}_\mathbf{1}(\mathbf{a})| \tag{22}$$

As proved in (Ben-David et al. 2007; 2010) (Lemma 2 in (Ben-David et al. 2010)), the $\mathcal{H}$-divergence in such case can be approximated by performing a classification task of the data $D$ into points belonging to $D_\mathbf{0}$ or to $D_\mathbf{1}$:

$$\mathbf{d}_\mathcal{H}(\hat{\mathbf{P}}_\mathbf{0},\hat{\mathbf{P}}_\mathbf{1}) = 2(1 - \min_\mathbf{h}\Big[\frac{2}{\mathbf{n}}\mathbf{I}(\mathbf{x}\in D_\mathbf{0}) + \frac{2}{\mathbf{n}}\mathbf{I}(\mathbf{x}\in D_\mathbf{1})\Big]) \tag{23}$$

where the distance ($\mathcal{H}$-divergence, which is equivalent in this case to the discrepancy distance) between $\hat{\mathbf{P}}_\mathbf{0}$ and $\hat{\mathbf{P}}_\mathbf{1}$ is inversely proportional to the performance of the classifier. From (21) and (23), for an arbitrary $\mathbf{h}$:

$$\begin{aligned}\mathbf{disc}(\mathbf{P_0},\mathbf{P_1}) \ \leq & \ 2 - \Big[\frac{4}{\mathbf{n}}\mathbf{I}(\mathbf{x}\in D_\mathbf{0}) + \frac{4}{\mathbf{n}}\mathbf{I}(\mathbf{x}\in D_\mathbf{1})\Big] \\ & + \frac{1}{2}(\mathbf{Rad}_{D_\mathbf{0}}(\mathbf{Err}(\mathbf{H})) + \mathbf{Rad}_{D_\mathbf{1}}(\mathbf{Err}(\mathbf{H}))) \\ & + 2\sqrt{\frac{\log(1/\delta)}{\mathbf{n}}}\end{aligned} \tag{24}$$

The term $\Big[\frac{4}{\mathbf{n}}\mathbf{I}(\mathbf{x}\in D_\mathbf{0}) + \frac{4}{\mathbf{n}}\mathbf{I}(\mathbf{x}\in D_\mathbf{1})\Big]$ is what is estimated in our proposed formulation in (9) in the main document by $-\beta\, max_{\boldsymbol{\phi}'}(L(\boldsymbol{\phi}'(\mathbf{g}(\mathbf{x}_i)),\mathbf{s}_i))$. Due to the negative sign preceding the latter term in (9) and the equivalent former term in (24), it is inversely proportional to the maximization of our objective, and to the overall distance between estimated distributions of data with different $\mathbf{s}$ values, respectively. Also, the larger the value of the fairness hyperparameter $\beta$, the higher the impact of this term on the optimization of our objective.

The label classifier $\mathbf{f}'(\mathbf{g}(\mathbf{x}_i))$ aims at minimizing the error of the classifier, i.e. minimizing the error in predicting the label $\hat{\mathbf{y}}$ of the data points $D$ (both $D_\mathbf{0}$ and $D_\mathbf{1}$). It is therefore straightforward to see that $\mathbf{f}'(\mathbf{g}(\mathbf{x}_i))$ naturally aims at minimizing the second term in the bound in (21), $(\mathbf{Rad}_{D_\mathbf{0}}(\mathbf{Err}(\mathbf{H})) + \mathbf{Rad}_{D_\mathbf{1}}(\mathbf{Err}(\mathbf{H})))$.

From the latter note and (24), the two classifiers of the adversarial formulation proposed in (9) in the main document can be interpreted w.r.t. the first two terms of the upper bound on the right of (21); minimizing the classifier's losses can be interpreted as minimizing the generalization upper bound.

# 4 Experiments

On two datasets, we perform experiments to evaluate the following: (i) the difference in classification accuracy due to optimizing for fairness–Table 1 and Figure 3. Comparisons among the unfair and fair versions of the proposed frameworks, FAD and FAD−MD, as well as previous state-of-the-art algorithms, show that the quest for fairness with FAD and FAD−MD leads to minimal loss in accuracy; (ii) (un)fairness, in terms of both disparate impact and disparate mistreatment–Table 2. Results demonstrate state-of-the-art effectiveness of FAD and FAD−MD; and (iii) an MMD 2-sample test to assess the fidelity of the learned fair representation $\mathbf{x}'$–Figure 4. This also shows the effectiveness of increasing the minibatch diversity in FAD−MD. Moreover, the increase in training time due to optimizing for fairness is not considerable for DAF compared to its unfair version.

We test our framework on two popular real-world datasets in the fairness literature, the Propublica COMPAS dataset (Larson et al. 2016) and the Adult dataset (Dheeru and Taniskidou 2017). The task in the COMPAS dataset is a binary classification task with two classes depicting whether or not a criminal defendant is to recidivate within two years. We use the black vs. white values of the race feature as our sensitive attribute. The total number of the COMPAS data instances we work on is 5,278 instances and 12 features. On the other hand, the Adult dataset consists of 32,561 complete instances denoting adults from the US Census in 1994. The task is to predict, from 14 features, whether an adult's income is higher or lower than 50K USD. Gender is the sensitive attribute in the Adult data (male or female). Each experiment is repeated ten times where, in each run, data is randomly split into three partitions, training, validation (to identify the value of the fairness hyperparameter $\beta$) and test. A portion of $60\%$ of the data is reserved for training, $20\%$ for validation and $20\%$ for testing. Statistics reported are the averages of the ten repetitions. For FAD−MD, we use the one-class SVM introduced in (Scholkopf et al. 2000) with $\nu = 0.5$ (fractions of support vectors and outliers).

On both datasets, and in addition to the unfair versions of the proposed algorithms, of (Zafar et al. 2017b) and of (Zafar et al. 2017a), we compare (where applicable) FAD and FAD−MD to the following state-of-the-art fairness algorithms: Zafar et al. (2017b), Zafar et al. (2017a), Zafar et al. (2017c), Hardt, Price, and Srebro (2016), Feldman et al. (2015), Kamishima et al. (2012), Fish, Kun, and Lelkes (2016), Bechavod and Ligett (2017), Komiyama et al. (2018), Agarwal et al. (2018), Narasimhan (2018). Classification results are displayed in Table 1. A 2-layer neural network is utilized to obtain the results of the unfair classification, whereas the adversarial layer is added under the top layer to accomplish fairness via FAD and its variation FAD−MD. Values of the fairness hyperparameter $\beta$ selected by cross-validation are $0.3$ and $0.8$ for the COMPAS and Adult datasets, respectively. Classification results on both datasets demonstrate that, among fairness algorithms, FAD and FAD−MD achieve state-of-the-art classification accuracy. In addition, the impact on classification accuracy due to the optimization for fairness by the proposed algorithms, FAD and FAD−MD, is rather minimal. This is quantified via the difference in accuracy between the unfair and fair versions of the proposed framework compared to such difference in the cases of (Zafar et al. 2017b) and (Zafar et al. 2017a).

Table 1: Results of the classification accuracy on the COMPAS and Adult datasets. In addition to its impact on fairness, increasing the minibatch diversity with FAD−MD improves the classification accuracy. Classification accuracy values achieved by FAD and FAD−MD on both datasets are higher than previous state-of-the-art results. Loss in accuracy due to fairness (difference in classification accuracy between the unfair and fair versions) is not big; with FAD−MD, it is as minimal as state-of-the-art by Zafar et al. (2017b) for the COMPAS data, and uniquely minimal for the Adult data. Bold refers to an accuracy value that is significantly better than the other fair (all apart from the first three entries) competitors. To test significance, we perform a paired t-test with significance level at $5\%$.

| | | | | |
|---|---|---|---|---|
| COMPAS | Unfair ($\beta = 0$) 89.3% | Unfair (Zafar et al. 2017b) 66.8% | Unfair (Zafar et al. 2017a) 69.0% | FAD 88.4% |
| | FAD−MD **88.7**% | Zafar et al. (2017b) 66.2% | Zafar et al. (2017a) 67.5% | Hardt et al. (2016) 64.4% |
| | Feldman et al. (2015) 86.8% | Kamishima et al. (2012) 72.4% | Fish et al. (2016) 81.2% | Bechavod (2017) 66.4% |
| | Komiyama et al. (2018) 86.6% | Agarwal et al. (2018) 71.2% | Narasimhan (2018) 77.7% | |
| Adult | Unfair ($\beta = 0$) 90.1% | Unfair (Zafar et al. 2017b) 85.8% | Unfair (Zafar et al. 2017a) 87.0% | FAD 88.6% |
| | FAD−MD **89**% | Zafar et al. (2017b) 83.1% | Zafar et al. (2017a) 84.0% | Hardt et al. (2016) 84.6% |
| | Feldman et al. (2015) 82.1% | Kamishima et al. (2012) 84.3% | Fish et al. (2016) 84.0% | Bechavod (2017) 78.3% |
| | Komiyama et al. (2018) 85.7% | Agarwal et al. (2018) 86.2% | Narasimhan (2018) 81.5% | |

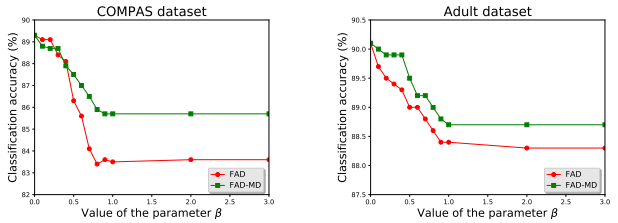
In Figure 3, we vary the classification accuracy as a function of the fairness hyperparameter $\beta$. FAD−MD leads to a slightly higher classification accuracy than FAD. Classification accuracy initially decreases when optimizing for fairness until it rather saturates with larger values of $\beta$.

Fairness results, in the form of empirical values of the disparity notions described in (3), (7) and (8), are displayed in Table 2. Such values are shown for FAD, FAD−MD and the same competitors as in Table 1, where applicable. The proposed algorithms, FAD and FAD−MD, minimize the disparity values (unfairness) when optimizing for fairness, and achieve the (at times joint) best results, in terms of the fairness metrics, in five out of six cases (three disparity values -$\text{Disp}_{\text{DI}}$, $\text{Disp}_{\text{FPR}}$ and $\text{Disp}_{\text{FNR}}$- for each dataset).

We move on now to a more rigorous evaluation of the minibatch diversity augmentation by FAD−MD. One of the problems of most current frameworks of adversarial learning is the fact that the adversary bases its optimization on comparing data points instead of distributions or, at least, of sets of points. We aim at evaluating this here by performing a nonparametric two-sample test between representations of data points belonging to different values of the sensitive attribute $\mathbf{s}$. The null hypothesis denotes that the distributions of the learned representation $\mathbf{x}'$ given different $\mathbf{s}$ values are equal, $H_0 : \mathbf{p}(\mathbf{x}'|\mathbf{s} = 0) = \mathbf{p}(\mathbf{x}'|\mathbf{s} = 1)$, and its alternative is $H_1 : \mathbf{p}(\mathbf{x}'|\mathbf{s} = 0) \neq \mathbf{p}(\mathbf{x}'|\mathbf{s} = 1)$. Failing to reject (i.e. accepting) the null hypothesis $H_0$ is the favorable outcome since it means that the model has learned a representation $\mathbf{x}'$ through which different values

Table 2: Unfairness of different algorithms measured by disparate impact -$\text{Disp}_{\text{DI}}$ in Eq. (3)- and disparate mistreatment -$\text{Disp}_{\text{FPR}}$, $\text{Disp}_{\text{FNR}}$ in Eqs. (7,8)- on the COMPAS and the Adult datasets. Smaller values are more favorable since they denote less unfairness. For each competitor, we report their best value achieved throughout their different settings. In five out of six cases (three disparity values for each dataset), the proposed algorithms, FAD and FAD−MD, (at times jointly) achieve the best results in terms of the fairness metrics. Bold refers to a value that is significantly less (better) than its non-bold competitors. To test significance, we perform a paired t-test with significance level at 5%. Empty cells indicate non-applicable experiments.

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| **COMPAS** | Unfair ($\beta = 0$) | Unfair (Zafar et al. 2017b) | Unfair (Zafar et al. 2017a) | FAD | FAD−MD | Zafar et al. (2017b) | Zafar et al. (2017a) | Hardt et al. (2016) |
| | $\text{Disp}_{\text{DI}}$: 0.6 | — | 0.62 | **0.08** | 0.11 | — | 0.38 | — |
| | $\text{Disp}_{\text{FPR}}$: 0.21 | 0.18 | — | **0.01** | **0.01** | 0.03 | — | **0.01** |
| | $\text{Disp}_{\text{FNR}}$: 0.29 | 0.3 | — | **0.01** | 0.02 | 0.1 | — | **0.01** |
| | Feldman et al. (2015) | Kamishima et al. (2012) | Fish et al. (2016) | Bechavod (2017) | Komiyama et al. (2018) | Agarwal et al. (2018) | Narasimhan (2018) | |
| | 0.95 | 0.9 | 0.15 | — | 0.2 | 0.09 | 0.1 | |
| | 0.4 | 0.2 | 0.03 | **0.01** | — | 0.05 | 0.09 | |
| | 0.45 | 0.15 | 0.03 | 0.03 | — | 0.05 | 0.11 | |
| **Adult** | Unfair ($\beta = 0$) | Unfair (Zafar et al. 2017b) | Unfair (Zafar et al. 2017a) | FAD | FAD−MD | Zafar et al. (2017b) | Zafar et al. (2017a) | Hardt et al. (2016) |
| | $\text{Disp}_{\text{DI}}$: 0.71 | — | 0.68 | 0.14 | **0.13** | — | 0.29 | — |
| | $\text{Disp}_{\text{FPR}}$: 0.36 | 0.35 | — | 0.02 | 0.01 | 0.12 | — | 0.04 |
| | $\text{Disp}_{\text{FNR}}$: 0.32 | 0.4 | — | **0.01** | 0.02 | 0.09 | — | 0.03 |
| | Feldman et al. (2015) | Kamishima et al. (2012) | Fish et al. (2016) | Bechavod (2017) | Komiyama et al. (2018) | Agarwal et al. (2018) | Narasimhan (2018) | |
| | 0.25 | 0.3 | 0.16 | — | 0.28 | **0.13** | 0.19 | |
| | 0.3 | 0.07 | 0.02 | **0.0** | — | 0.04 | 0.14 | |
| | 0.4 | 0.08 | 0.03 | 0.04 | — | 0.05 | 0.08 | |



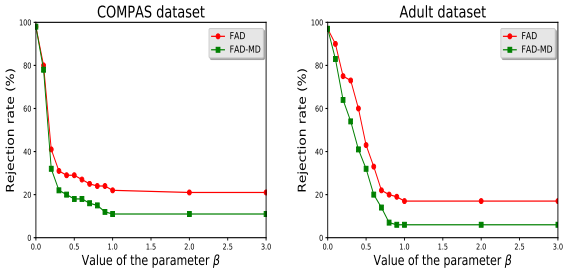

(a) COMPAS dataset (b) Adult dataset

Figure 3: Classification accuracy as a function of the fairness hyperparameter $\beta$, where $\beta = 0$ indicates no adjustment for fairness. The COMPAS data is balanced; 47% of the instances have recidivated and 53% have not. Hence, a random guessing classifier's accuracy would be as low as around 53%. For the Adult data, a random guessing classifier would result in 75.9% accuracy, since there are many more records for people whose income is less than or equal to 50K. Classification accuracy initially decreases with increasing $\beta$ until it saturates. With both datasets, classification accuracy has not considerably changed for $\beta > 1$. Accuracy achieved by FAD−MD is slightly higher.

of $\mathbf{s}$ are indistinguishable. We perform a two-sample maximum mean discrepancy (MMD) test. Let $\mathbf{x}'_0$ and $\mathbf{x}'_1$ refer to data points sampled from $\mathbf{x}'$ given $\mathbf{s} = 0$ and $\mathbf{s} = 1$, respectively. We compute the unbiased estimate $\text{MMD}(\mathbf{x}'_0, \mathbf{x}'_1)$ as a two-sample test between $\mathbf{x}'_0$ and $\mathbf{x}'_1$ (Gretton et al. 2006; Lloyd and Ghahramani 2015) by the following expression:

$$\frac{1}{{n_0}^2}\sum_{i=1}^{n_0}\sum_{j=1}^{n_0}\mathbf{k}(\mathbf{x}'_{0(i)}, \mathbf{x}'_{0(j)}) + \frac{1}{{n_1}^2}\sum_{i=1}^{n_1}\sum_{j=1}^{n_1}\mathbf{k}(\mathbf{x}'_{1(i)}, \mathbf{x}'_{1(j)}) - \frac{2}{n_0 n_1}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1}\mathbf{k}(\mathbf{x}'_{0(i)}, \mathbf{x}'_{1(j)}) \quad (25)$$

We use a Gaussian kernel $\mathbf{k}(\mathbf{x}'_0, \mathbf{x}'_1) = e^{-\gamma\|\mathbf{x}'_0 - \mathbf{x}'_1\|^2}$. Cross-validation has been used to indicate $\gamma$. The threshold used (the given allowable probability of false rejection (Gretton et al. 2012)) is 0.05. Results of running an MMD two-sample test for 100 times are displayed in Figure 4. Smaller values are better since they signify a more similar representation for instances with different sensitive attribute, $\mathbf{s}$, values. Hence, FAD−MD achieves better results with lower rejection rates than FAD for both datasets.




(a) COMPAS dataset (b) Adult dataset

Figure 4: MMD 2-sample test results as a function of the fairness hyperparameter $\beta$. The lower the value the better. The threshold used, i.e. given allowable probability of false rejection, is 0.05. FAD−MD leads to less distinguishable (less unfair) representations of data with different $\mathbf{s}$, than FAD.

Details of the model architectures are listed in Table 3. Adam (Kingma and Ba 2015) is the optimizer used to compute the gradients.

Table 3: Architecture of the neural network used in the introduced models. There are two layers, in addition to the adversarial layer $\mathbf{g}$. FC stands for fully connected.

| Dataset | Architecture |
|---|---|
| **COMPAS** | FC 16 ReLU, FC 32 ReLU, $\mathbf{g}$ : FC 16 ReLU, FC output. |
| **Adult** | FC 32 ReLU, FC 32 ReLU, $\mathbf{g}$ : FC 16 ReLU, FC output. |

## 5 Related Work

We first focus on fairness frameworks that are based on adversarial learning. We begin with a comparison with the framework of (Madras et al. 2018). Differences between the latter and our proposed framework include:

• Our proposed platform can be applied to an existing neural network architecture with rather slender modifications. We do not have to reconstruct everything from scratch to obtain a potentially fair model.
• The whole optimization in our framework is implemented within one neural network, compared to up to four neural networks in (Madras et al. 2018). This leads to avoiding some of the well documented problems of adversarial learning that appear when two or more neural networks are involved in the optimization (Goodfellow 2016; Goodfellow et al. 2014).
• Performing the training with diverse minibatches, i.e. minibatches whose elements are maximally diverse from one another, leads to improved results.
• Our framework is capable of quantitatively evaluating the achieved degree of fairness as well as the difference (potential loss) in accuracy due to imposing fairness.

The work in (Edwards and Storkey 2016) learns a representation that is concurrently fair and discriminative w.r.t. the prediction task. Similar to (Madras et al. 2018), it is based on more than one neural network since each adversary consists of a separate network, leading to difficulties in reaching stability among adversaries. The work in (Edwards and Storkey 2016) also sheds light on a mapping between fairness, distances between distributions and the adversary's ability to predict the sensitive attribute value. In addition to having a different approach, we extend beyond that by linking all these notions to the original labeling classification accuracy as well. Zhang, Lemoine, and Mitchell (2018) utilize an adversarial learning framework to achieve fairness via directly comparing the labeling classification outcomes, i.e. without learning an intermediate fair representation. The work in (Beutel et al. 2017) analyzes the impact of the data distribution on the fairness notion adopted by the adversary. Another adversarial framework is the one introduced by Louppe, Kagan, and Cranmer (2017) which permits two-network based adversarial frameworks to act on a continuous sensitive attribute. Although they note that the algorithm is applicable for fairness, the experiments performed are on one real-world dataset that is not fairness-related. As a result, implementing the same idea in (Louppe, Kagan, and Cranmer 2017) using continuous sensitive attributes on fairness datasets in a monolithic network within our framework is an interesting direction for future work. Other fairness-aware adversarial works include (Wadsworth, Vera, and Piech 2018; Xu et al. 2018).

Looking more broadly (beyond adversarial frameworks), and in addition to those mentioned elsewhere in the paper, other fairness algorithms include Goel, Rao, and Shroff (2015) that defines disparities as a function of false positive rates for people from different races, i.e. similar to disparate mistreatment. In (Celis et al. 2018), more than one notion of fairness can be jointly enforced on a meta-algorithm via fairness constraints. A comparative study of some fairness algorithms has been provided in (Friedler et al. 2018). An interpolation between statistical notions of fairness, with the aim of obtaining the good properties of each definition, has been presented in (Kearns et al. 2018). The work in (Hajian et al. 2015) imposes fairness via a post-processing approach of the frequent patterns in the data.

## 6 Conclusion

We introduced a fair adversarial framework applicable to differentiable discriminative models. Instead of having to establish the architecture from scratch, we make slight adjustments to an existing differentiable classifier by adding a new hidden layer and a new classifier above it, to concurrently optimize for fairness and accuracy. We analyzed and evaluated the resulting tradeoff between fairness and accuracy. We proposed a minibatch diversity variation of the learning procedure which may be of independent interest for other adversarial frameworks. We provided a theoretical interpretation of the two classifiers (adversaries) constituting the model. We demonstrated strong empirical performance of our methods compared to previous leading approaches. Our approach applies to existing architectures; hence, it will be interesting to study how a pre-trained network adapts to the new dual objective.

## Acknowledgements

TA, ZG and AW acknowledge support from the Leverhulme Trust via the CFI. AW acknowledges support from the David MacKay Newton research fellowship at Darwin College and The Alan Turing Institute under EPSRC grant EP/N510129/1 & TU/B/000074. IV acknowledges support from the MPG Minerva Fast Track program.

## References

Agarwal, A.; Beygelzimer, A.; Dudik, M.; Langford, J.; and Wallach, H. 2018. A reductions approach to fair classification. *ICML*.

Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; and Marchand, M. 2014. Domain adversarial neural networks. *arXiv preprint arXiv:1412.4446*.

Arjovsky, M.; Chintala, S.; and Bottou, L. 2017. Wasserstein generative adversarial networks. *ICML*.

Barocas, S., and Selbst, A. 2016. Big Data's Disparate Impact. *California Law Review*.

Bartlett, P., and Mendelson, S. 2002. Rademacher and Gaussian complexities: Risk bounds and structural results. *JMLR*.

Bechavod, Y., and Ligett, K. 2017. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044*.

Ben-David, S.; Blitzer, J.; Crammer, K.; and Pereira, F. 2007. Analysis of representations for domain adaptation. *NIPS* 21:137–144.

Ben-David, S.; Blitzer, S.; Crammer, K.; Kulesza, A.; Pereira, F.; and Vaughan, J. 2010. A theory of learning from different domains. *Machine learning* 79(2):151–175.

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

Brennan, T.; Dieterich, W.; and Ehret, B. 2009. Evaluating the predictive validity of the COMPAS risk and needs assessment system. *Criminal Justice and Behavior* 36:21–40.

Celis, E.; Huang, L.; Keswani, V.; and Vishnoi, N. 2018. Classification with fairness constraints: A meta-algorithm with provable guarantees. *arXiv preprint arXiv:1806.06055*.

Chouldechova, A. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data* 2.

Devroye, L.; Gyorfi, L.; and Lugosi, G. 1996. *A probabilistic theory of pattern recognition*. Springer.

Dheeru, D., and Taniskidou, E. K. 2017. UCI ML Repository.

Edwards, H., and Storkey, A. 2016. Censoring representations with an adversary. *ICLR*.

Feldman, M.; Friedler, S.; Moeller, J.; Scheidegger, C.; and Venkatasubramanian, S. 2015. Certifying and removing disparate impact. *KDD*.

Fish, B.; Kun, J.; and Lelkes, A. 2016. A confidence-based approach for balancing fairness and accuracy. *SDM*.

Friedler, S.; Scheidegger, C.; Venkatasubramanian, S.; Choudhary, S.; Hamilton, E.; and Roth, D. 2018. A comparative study of fairness-enhancing interventions in machine learning. *arXiv preprint arXiv:1802.04422*.

Ganin, Y., and Lempitsky, V. 2015. Unsupervised domain adaptation by backpropagation. *ICML* 32.

Ganin, Y.; Ustinova, E.; Ajakan, H.; Germain, P.; Larochelle, H.; Laviolette, F.; Marchand, M.; and Lempitsky, V. 2016. Domain-adversarial training of neural networks. *JMLR* 17(59):1–35.

Goel, S.; Rao, J.; and Shroff, R. 2015. Precinct or prejudice? Understanding racial disparities in New York City's stop-and-frisk policy. *Annals of Applied Statistics*.

Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. *NIPS* 2672–2680.

Goodfellow, I. 2016. NIPS 2016 tutorial: Generative adversarial networks. *arXiv preprint arXiv:1701.00160*.

Gretton, A.; Borgwardt, K.; Rasch, M.; Scholkopf, B.; and Smola, A. 2006. A kernel method for the two-sample-problem. *NIPS*.

Gretton, A.; Borgwardt, K.; Rasch, M.; Scholkopf, B.; and Smola, A. 2012. A kernel two-sample test. *JMLR* 13.

Grgic-Hlaca, N.; Redmiles, E.; Gummadi, K.; and Weller, A. 2018a. Human perceptions of fairness in algorithmic decision making: A case study of criminal risk prediction. *WWW*.

Grgic-Hlaca, N.; Zafar, M.; Gummadi, K.; and Weller, A. 2018b. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. *AAAI*.

Hajian, S.; Domingo-Ferrer, J.; Monreale, A.; Pedreschi, D.; and Giannotti, F. 2015. Discrimination and privacy-aware patterns. *Data Mining and Knowledge Discovery* 1733–1782.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *NIPS*.

Kamishima, T.; Akaho, S.; Asoh, H.; and Sakuma, J. 2012. Fairness-aware classifier with prejudice remover regularizer.

Kearns, M.; Neel, S.; Roth, A.; and Wu, Z. 2018. Preventing fairness gerrymandering: Auditing and learning for subgroup fairness. *ICML*.

Khandani, A.; Kim, A.; and Lo, A. 2010. Consumer credit-risk models via machine-learning algorithms. *JBF* 34:2767–2787.

Kifer, D.; Ben-David, S.; and Gehrke, J. 2004. Detecting change in data streams. *VLDB* 180–191.

Kim, T.; Cha, M.; Kim, H.; Lee, J.; and Kim, J. 2017. Learning to discover cross-domain relations with generative adversarial networks. *ICML*.

Kingma, D., and Ba, J. 2015. Adam: A Method for Stochastic Optimization. *ICLR*.

Koltchinskii, V., and Panchenko, D. 2000. Rademacher processes and bounding the risk of function learning. *HDP*.

Komiyama, J.; Takeda, A.; Honda, J.; and Shimao, H. 2018. Nonconvex optimization for regression with fairness constraints. *ICML*.

Kusner, M.; Loftus, J.; Russell, C.; and Silva, R. 2017. Counterfactual fairness. *NIPS*.

Larson, J.; Mattu, S.; Kirchner, L.; and Angwin, J. 2016. https://github.com/propublica/compas-analysis.

Lloyd, J., and Ghahramani, Z. 2015. Statistical model criticism using kernel two sample tests. *NIPS*.

Louizos, C.; Swerky, K.; Li, Y.; Welling, M.; and Zemel, R. 2016. The variational fair autoencoder. *ICLR*.

Louppe, G.; Kagan, M.; and Cranmer, K. 2017. Learning to pivot with adversarial networks. *NIPS*.

Madras, D.; Creager, E.; Pitassi, T.; and Zemel, R. 2018. Learning adversarially fair and transferable representations. *ICML*.

Mansour, Y.; Mohri, M.; and Rostamizadeh, A. 2009. Domain adaptation: Learning bounds and algorithms. *COLT*.

Metz, L.; Poole, B.; Pfau, D.; and Sohl-Dickstein, J. 2017. Unrolled generative adversarial networks. *ICLR*.

Mohri, M., and Afshin, R. 2008a. Rademacher complexity bounds for non-IID processes. *NIPS*.

Narasimhan, H. 2018. Learning with complex loss functions and constraints. *AISTATS* 1646–1654.

Primus, R. 2010. The future of disparate impact. *Mich. Law Rev.*

Rosca, M.; Lakshminarayanan, B.; Warde-Farley, D.; and Mohamed, S. 2017. Variational approaches for auto-encoding generative adversarial networks. *arXiv preprint arXiv:1706.04987*.

Salimans, T.; Goodfellow, I.; Zaremba, W.; Cheung, V.; and Radford, A. 2016. Improved techniques for training GANs. *NIPS*.

Scholkopf, B.; Williamson, R.; Smola, A.; Shawe-Taylor, J.; and Platt, J. 2000. Support vector method for novelty detection. *NIPS*.

Shalev-Shwartz, S., and Ben-David, S. 2014. *Understanding machine learning: From theory to algorithms*. Cambridge Univ. Press.

Srivastava, A.; Valkov, L.; Russell, C.; Gutmann, M.; and Sutton, C. 2017. VEEGAN: Reducing mode collapse in GANs using implicit variational learning. *NIPS*.

Wadsworth, C.; Vera, F.; and Piech, C. 2018. Achieving fairness through adversarial learning: an application to recidivism prediction. *FAT/ML Workshop*.

Xu, D.; Yuan, S.; Zhang, L.; and Wu, X. 2018. FairGAN: Fairness-aware generative adversarial networks. *arXiv preprint arXiv:1805.11202*.

Zafar, M.; Valera, I.; Rodriguez, M.; Gummadi, K.; and Weller, A. 2017a. From parity to preference-based notions of fairness in classification. *NIPS*.

Zafar, M.; Valera, I.; Rodriguez, M.; and Gummadi, K. 2017b. Fairness beyond disparate treatment and disparate impact: Learning classification without disparate mistreatment. *WWW*.

Zafar, M.; Valera, I.; Rodriguez, M.; and Gummadi, K. 2017c. Fairness constraints: Mechanisms for fair classification. *AISTATS*.

Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning fair representations. *ICML* 325–333.

Zhang, B.; Lemoine, B.; and Mitchell, M. 2018. Mitigating unwanted biases with adversarial learning. *arXiv:1801.07593*.