Mitigating Unwanted Biases with Adversarial Learning

Brian Hu Zhang
Stanford University
Stanford, California
bhz@cs.stanford.edu

Blake Lemoine
Google
Mountain View, California
lemoine@google.com

Margaret Mitchell
Google
Mountain View, California
mmitchellai@google.com

ABSTRACT

Machine learning is a tool for building models that accurately represent input training data. When undesired biases concerning demographic groups are in the training data, well-trained models will reflect those biases. We present a framework for mitigating such biases by including a variable for the group of interest and simultaneously learning a predictor and an adversary. The input to the network X, here text or census data, produces a prediction Y, such as an analogy completion or income bracket, while the adversary tries to model a protected variable Z, here gender or zip code.

The objective is to maximize the predictorâĂŹs ability to predict Y while minimizing the adversary's ability to predict Z. Applied to analogy completion, this method results in accurate predictions that exhibit less evidence of stereotyping Z. When applied to a classification task using the UCI Adult (Census) Dataset, it results in a predictive model that does not lose much accuracy while achieving very close to equality of odds (Hardt, et al., 2016). The method is flexible and applicable to multiple definitions of fairness as well as a wide range of gradient-based learning models, including both regression and classification tasks.

CCS CONCEPTS

• **Theory of computation** → **Adversarial learning**; *Nonconvex optimization*; • **Computing methodologies** → **Multi-task learning**; **Neural networks**; *Natural language processing*; *Supervised learning by classification*; • **Mathematics of computing** → *Maximum likelihood estimation*; *Convex optimization*; • **Social and professional topics** → *User characteristics*; *Race and ethnicity*; *Gender*;

KEYWORDS

debiasing; unbiasing; adversarial learning; multi-task learning

ACM Reference Format:
Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating Unwanted Biases with Adversarial Learning. In *2018 AAAI/ACM Conference on AI, Ethics, and Society (AIES '18), February 2–3, 2018, New Orleans, LA, USA.* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/3278721.3278779

AIES '18, February 2–3, 2018, New Orleans, LA, USA

ACM ISBN 978-1-4503-6012-8/18/02…$15.00
https://doi.org/10.1145/3278721.3278779

1 INTRODUCTION

Machine learning leverages data to build models capable of assessing the labels and properties of novel data. Unfortunately, the available training data frequently contains biases with respect to things that we would rather not use for decision making. Machine learning builds models faithful to training data and can lead to perpetuating these undesirable biases. For example, systems designed to predict creditworthiness and systems designed to perform analogy completion have been demonstrated to be biased against racial minorities and women respectively. Ideally we would be able to build a model which captures exactly those generalizations from the data which are useful for performing some task which are not discriminatory in a way which the people building those models consider unfair.

Work on training machine learning systems that output fair decisions has defined several useful measurements for *fairness*: Demographic Parity, Equality of Odds, and Equality of Opportunity. These can be imposed as constraints or incorporated into a loss function in order to mitigate disproportional outcomes in the system's output predictions regarding a protected demographic, such as sex.

In this paper, we examine these fairness measures in the context of *adversarial debiasing*. We consider supervised deep learning tasks in which the task is to predict an output variable Y given an input variable X, while remaining unbiased with respect to some variable Z. We refer to Z as the *protected variable*. For these learning systems, the predictor $\hat{Y} = f(X)$ is learned from a training set of (input, output, protected) tuples (X, Y, Z). The predictor f is usually given access to the protected variable Z, though this is not strictly necessary. This construction allows the determination of which types of bias are considered undesirable for a particular application to be chosen through the specification of the protected variable.

We speak to the concept of *mitigating bias* using the known term *debiasing*[1], following definitions provided by Hardt et al. [5] and refined by Beutel et al. [2].

Definition 1.1. Demographic Parity. A predictor $\hat{Y}$ satisfies *demographic parity* if $\hat{Y}$ and Z are independent.

This means that $P(\hat{Y} = \hat{y})$ is equal for all values of the protected variable Z: $P(\hat{Y} = \hat{y}) = P(\hat{Y} = \hat{y}|Z = z)$.

Definition 1.2. Equality of Odds. A predictor $\hat{Y}$ satisfies *equality of odds* if $\hat{Y}$ and Z are conditionally independent given Y.

This means that, for all possible values of the true label Y, $P(\hat{Y} = \hat{y})$ is the same for all values of the protected variable: $P(\hat{Y} = \hat{y}|Y = y) = P(\hat{Y} = \hat{y}|Z = z, Y = y)$

[1]Note that "debias" may not be quite the right word, as all bias is not necessarily removed.

Definition 1.3. Equality of Opportunity. If the output variable Y is discrete, a predictor $\hat{Y}$ satisfies *equality of opportunity* with respect to a class y if $\hat{Y}$ and Z are independent conditioned on $Y = y$.

This means that, for a *particular* value of the true label Y, $P(\hat{Y} = \hat{y})$ is the same for all values of the protected variable: $P(\hat{Y} = \hat{y}|Y = y) = P(\hat{Y} = \hat{y}|Z = z, Y = y)$

We present an adversarial technique for achieving whichever one of these definitions is desired.[2] A predictor f will be trained to model Y as accurately as possible while satisfying one of the above equality constraints. Demographic parity will be achieved by introducing an adversary g which will attempt to predict a value for Z from $\hat{Y}$. The gradient of g will then be incorporated into the weight update rule of f so as to reduce the amount of information about Z transmitted through $\hat{Y}$. Equality of odds will be achieved by also giving g access to the true label Y, thereby limiting any information about Z which $\hat{Y}$ contains beyond the information already contained in Y.

We consider the case where the protected variable is a discrete feature present in the training set as well as the case in which the protected variable must be inferred from latent semantics (in particular, gender from word embeddings). In order to accomplish the latter we adapt a technique presented by Bolukbasi et al. [3] to define a subspace capturing the semantics of the protected variable, and then train a model to perform a word analogies task accurately, while unbiased on this protected variable. A consequence of this technique is that the network learns *"debiased"* embeddings, embeddings that have the semantics of the protected variable removed. These embeddings are still able to perform the analogy task well, but are better at avoiding problematic examples such as those shown in Bolukbasi et al. [3].

Results on the UCI Adult Dataset demonstrate the technique we introduce allows us to train a model that achieves equality of odds to within 1% on both protected groups.

We also compare with the related previous work of Beutel et al. [2], and find we are able to better equalize the differences between the two groups, measured by both False Positive Rate and False Negative Rate (1 - True Positive Rate), although note that the previous work performs better overall for False Negative Rate.

We provide some discussion on caveats pertaining to this approach, difficulties in training these models that are shared by many adversarial approaches, as well as some discussion on difficulties that the fairness constraints introduce.

2 RELATED WORK

There has been significant work done in the area of debiasing various specific types of data or predictor.

Debiasing word embeddings: Bolukbasi et al. [3] devises a method to remove gender bias from word embeddings. The method relies on a lot of human input; namely, it needs a large "training set" of gender-specific words.

Simple models: Lum and Johndrow [8] demonstrate that removing the protected variable from the training data fails to yield a debiased model (since other variables can be highly correlated with the protected variable), and devise a method for learning fair predictive models in cases when the learning model is simple (e.g. linear regression). Hardt et al. [5] discuss the shortcomings of focusing solely on demographic parity, present alternate definitions of fairness, and devise a method for deriving an unbiased predictor from a biased one, in cases when both the output variable and the protected variable are discrete.

Adversarial training: Goodfellow et al. [4] pioneered the technique of using multiple networks with competing goals to force the first network to "deceive" the second network, applying this method to the problem of creating real-life-like pictures. Beutel et al. [2] apply an adversarial training method to achieve equality of opportunity in cases when the output variable is discrete. They also discuss the ability of the adversary to be powerful enough to enforce a fairness constraint even when it has access to a very small training sample.

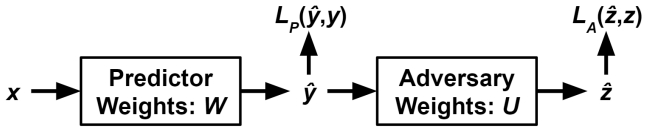

Figure 1: The architecture of the adversarial network.

3 ADVERSARIAL DEBIASING

We begin with a model, which we call the *predictor*, trained to accomplish the task of predicting Y given X. As in Figure 1, we assume that the model is trained by attempting to modify weights W to minimize some loss $L_P(\hat{y}, y)$, using a gradient-based method such as stochastic gradient descent.

The output layer of the predictor is then used as an input to another network called the *adversary* which attempts to predict Z. This is part of the network corresponds to the *discriminator* in a typical GAN [4]. We will suppose the adversary has loss term $L_A(\hat{z}, z)$ and weights U. Depending on the definition of fairness being achieved, the adversary may have other inputs.

- For Demographic Parity, the adversary gets the predicted label $\hat{Y}$. Intuitively, this allows the adversary to try to predict the protected variable using nothing but the predicted label. The goal of the predictor is to prevent the adversary from doing this.
- For Equality of Odds, the adversary gets $\hat{Y}$ and the true label Y.
- For Equality of Opportunity on a given class y, we can restrict the training set of the adversary to training examples where $Y = y$.[3]

In order for gradients to propagate correctly, $\hat{Y}$ above refers to the output layer of the network, not to the discrete prediction; for example, for a classification problem, $\hat{Y}$ could refer to the output of the softmax layer.

[2] Achieving equality of odds and demographic parity are generally incongruent goals. See also Kleinberg, Mullainathan, and Raghavan [7] for incongruency between calibration and equalized odds.

[3] This last technique of restricting the training set is discussed at length by Beutel et al. [2], so we only mention it here.

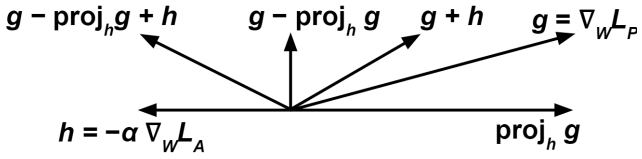

Figure 2: Diagram illustrating the gradients in Eqn. 1 and the relevance of the projection term $\text{proj}_h g$. Without the projection term, in the pictured scenario, the predictor would move in the direction labelled $g + h$ in the diagram, which actually *helps* the adversary. With the projection term, the predictor will never move in a direction that helps the adversary.

We update U to minimize L_A at each training time step, according to the gradient $\nabla_U L_A$. We modify W according to the expression:

$$\nabla_W L_P - \text{proj}_{\nabla_W L_A} \nabla_W L_P - \alpha \nabla_W L_A \tag{1}$$

where α is a tuneable hyperparameter that can vary at each time step and we define $\text{proj}_v x = 0$ if $v = 0$.

The middle term $\text{proj}_{\nabla_W L_A} \nabla_W L_P$ prevents the predictor from moving in a direction that helps the adversary decrease its loss while the last term, $\alpha \nabla_W L_A$, attempts to increase the adversary's loss. Without the projection term, it is possible for the predictor to end up helping the adversary (see Fig. 2). Without the last term, the predictor will never try to *hurt* the adversary, and, due to the stochastic nature of many gradient-based methods, will likely end up helping the adversary anyway. The result is that when training is completed the desired definition of equality should be satisfied.

Notice that our definitions and method make no assumptions about the nature of the output and protected variables: in particular, they work with both regression and classification models, as well as with both discrete and continuous protected variables.

4 PROPERTIES

We note several properties of the above method that we believe distinguish it from past work.

(1) **Generality:** The above method can be used to enforce DEMOGRAPHIC PARITY, EQUALITY OF ODDS, or EQUALITY OF OPPORTUNITY as described in Hardt et al. [5]. Further, it applies without modification to the cases when the output variable and/or protected variable are continuous instead of discrete.
(2) **Model-agnostic:** The adversarial approach described can be applied regardless of how simple or complex the predictor's model is, as long as the model is trained using a gradient-based method, as many modern learning models are. Further, as we will discuss later, at least in some situations, we suggest that the adversary does not need to be nearly as complex as the predictor—a simple adversary can be used with a complex predictor.
(3) **Optimality:** Under certain conditions, we show that if the predictor converges, it must converge to a model that satisfies the desired fairness definition. Since the predictor also attempts to decrease the prediction loss L_P, the predictor should still perform well on the target task.

5 THEORETICAL GUARANTEES

PROPOSITION 5.1. *Let the predictor, the adversary, and their weights W, U be defined according to Section 3 Let $L_A(W, U)$ be the adversary's loss, convex in U, concave in W,[4] and continuously differentiable everywhere.*

Suppose that:

(1) *When the predictor's weights are W_0, the predictor gives the same output $\hat{Y}$ regardless of input X. (For example, when $W_0 = 0$).*
(2) *There are some weights U_0 that minimize L_A when the weights for $\hat{Y}$ have no effect on the output: For all W, $L_A(W, U_0) = \min_U L_A(W_0, U)$.*
(3) *Predictor and adversary converge to W^* and U^* respectively.*

Then, $L_A(W^*, U^*) = L_A(W^*, U_0)$. That is, the adversary gains no advantage from using the weights for $\hat{Y}$.

PROOF. Since the adversary converges, $L_A(W^*, U^*) \leq L_A(W^*, U_0)$: otherwise, since L_A is convex in U, the adversary's weights would move toward U_0. In other words, the adversary's minimum is the point at which the adversary gains an advantage from using $\hat{Y}$. Similarly, since the predictor converges, $L_A(W^*, U^*) \geq L_A(W_0, U^*)$: Otherwise, the predictor would be able to increase the adversary's loss by moving toward W_0, and the projection term and negative weight on $\nabla_W L_A$ in Eqn. 1 would push the predictor to move towards 0. Then:

$$\begin{aligned}
L_A(W^*, U_0) &\geq L_A(W^*, U^*) && \text{(as stated above)} \\
&\geq L_A(W_0, U^*) && \text{(as stated above)} \\
&\geq L_A(W_0, U_0) && \text{(by definition of } U_0\text{)} \\
&= L_A(W^*, U_0) && \text{(by definition of } U_0\text{)}
\end{aligned}$$

so we must have $L_A(W^*, U^*) = L_A(W^*, U_0)$. □

Note that, in this proof, the adversary can be operating in a few different ways, as long as it is given $\hat{Y}$ as one of its inputs; for example, for demographic parity, it could be given only $\hat{Y}$; for equality of odds, it can be given both $\hat{Y}$ and Y.

We will show in the next propositions that the adversary gaining no advantage from information about $\hat{Y}$ is exactly the condition needed to guarantee that desired definitions of equality are satisfied.

PROPOSITION 5.2. *Let the training data be comprised of triples (X, Y, Z) drawn according to some distribution D. Suppose:*

(1) *The protected variable Z is discrete.*
(2) *The adversary is trained for* DEMOGRAPHIC PARITY*; i.e. the adversary is given only the prediction $\hat{y}$.*
(3) *The adversary is strong enough that, at convergence, it has learned a randomized function A that minimizes the cross-entropy loss $\mathbb{E}_{(x,y,z)\sim D}[-\log P(A(\hat{y}) = z)]$; i.e. the adversary in fact achieves the optimal accuracy with which you can predict Z from $\hat{Y}$*
(4) *The predictor completely fools the adversary; in particular, the adversary achieves loss $H(Z)$, the entropy of Z.*

[4] We understand that these assumptions are not satisfied in most use cases involving neural networks; however, as with most theoretical analyses of machine learning models (see, for example, Goodfellow et al. [4] or Kingma and Ba [6]; the former makes even stronger assumptions), assumptions of concavity are necessary for any proofs to work

Then the predictor satisfies DEMOGRAPHIC PARITY*; i.e.,* $\hat{Y} \perp Z$.

PROOF. Notice that if the adversary draws $A(\hat{y})$ according to the distribution $Z|\hat{Y} = \hat{y}$, then its loss is exactly the conditional entropy

$$\begin{aligned} H(Z|\hat{Y}) &= \mathbb{E}[-\log P(Z = z|\hat{Y} = \hat{y})] \\ &= \mathbb{E}[-\log P(A(\hat{y}) = z|\hat{Y} = \hat{y})] \end{aligned}$$

where the expectation is taken over $(x, y, z) \sim D$. Now suppose for contradiction that $\hat{Y}$ is dependent on Z. Then $H(Z|\hat{Y}) < H(Z)$, so the adversary can achieve loss less than $H(Z)$, contradicting assumption (4). □

PROPOSITION 5.3. *If assumptions (2)-(4) above are replaced with the analogous equality of odds assumptions; in particular, that the adversary is given* $\hat{y}$ *and* y*, and the adversary cannot achieve loss better than* $H(Z|Y)$ *then the predictor will satisfy* EQUALITY OF ODDS*; i.e.,* $(\hat{Y} \perp Z)|Y$

PROOF. Analogous to the above. Notice that if the adversary draws $A(\hat{y}, y) \sim (Z|\hat{Y} = \hat{y}, Y = y)$, then its loss is exactly the conditional entropy

$$\begin{aligned} H(Z|\hat{Y}, Y) &= \mathbb{E}[-\log P(Z = z|\hat{Y} = \hat{y}, Y = y)] \\ &= \mathbb{E}[-\log P(A(\hat{y}) = z|\hat{Y} = \hat{y}, Y = y)] \end{aligned}$$

where the expectation is again taken over $(x, y, z) \sim D$. But if $\hat{Y}$ is conditionally dependent on Z given Y, then $H(Z|\hat{Y}, Y) < H(Z|Y)$, so the adversary can achieve loss less than $H(Z|Y)$. □

Note that Propositions 5.2 and 5.3 work analogously in the case of continuous Y and Z, with the probability mass function P replaced with the probability density function p, and the discrete entropy H replaced by the differential entropy $h(X) = \mathbb{E}[-\log p(x)]$, since the relevant property ($h(A) = h(A|B)$ iff $A \perp B$) holds for differential entropy as well. They also work analogously when the adversary A is restricted to a limited set of predictors.

For example, an adversary using least-squares regression trying to enforce equality of odds can be thought of as one that outputs $A(\hat{y}, y) \sim N(\mu(\hat{y}, y), \sigma^2)$ where $\mu(\hat{y}, y)$ is the output of the regressor, and $\sigma^2 > 0$ is a fixed constant. Note now that the differential entropy $h(Z|\hat{Y}, Y) = \mathbb{E}[-\log p(z|\hat{y}, y)]$ is nothing more than the expected log-likelihood, and so the function μ that minimizes this quantity is the optimal least-squares regressor. Thus, for example, if we restrict μ to be a linear function of $(\hat{y}, y)$, and the other conditions of Proposition 5.3 hold, then an analogous argument to the above propositions shows that $\hat{Y}$ has no linear relationship with Z after conditioning on Y.

These claims together illustrate that a sufficiently powerful adversary trained on a sufficiently large training set can indeed accurately enforce the demographic parity or equality of odds constraints on the predictor, if the adversary and predictor converge. Guaranteed convergence is harder to achieve, both in theory and practice. In the practical scenarios below we discuss methods to encourage the training algorithm to converge, as well as reasonable choices of the adversary model that are both powerful and easy to train.

6 EXPERIMENTS

All models were trained using the Adam optimizer [6] for both predictor and adversary.

6.1 Toy Scenario

We generate a training sample $(x^{(i)}, y^{(i)}, z^{(i)})_{i=1}^{n}$ (where z is the protected variable) as follows. For each i, let $r \in 0, 1$ be picked uniformly at random, and let $v \sim N(r_i, 1)$. Let $u, w \sim N(v_i, 1)$ vary independently. Then $x^{(i)} = (r, u), y^{(i)} = [w > 0], z^{(i)} = r$. (where [ ] denotes an indicator function). Intuitively, the variable that we are trying to predict, y, depends directly on v and r. We are given as inputs the protected variable r, and a noisy measurement of v. The end goal would be to train a model that predicts y while being unbiased on r, effectively removing the direct signal for r from the learned model.

If one trains generically a logistic regression model to predict y given x, it outputs something like $y = \sigma(0.7u + 0.7r)$, which is a reasonable model, but heavily incorporates the protected variable r. To debias, We now train a model that achieves DEMOGRAPHIC PARITY. Note that removing the variable r from the training data is insuffucient for debiasing: the model will still learn to use u to predict y, and u is correlated with r. If we use the described technique and add in another logistic model that tries to predict z given y, we find that the predictor model outputs something like $y = \sigma(0.6u - 0.6r + 0.6)$. Notice that not only is r not included with a positive weight anymore, the model actually learns to use a negative weight on r in order to balance out the effect of r on u Notice that $u - r \sim N(0, 2)$; i.e., it is not dependent on r, so we have successfully trained a model to predict y independently of r.

6.2 Word Embeddings

We train a model to perform the analogy task (i.e., fill in the blank: `man : woman :: he : ?`).

It is known that word embeddings reflect or amplify problematic biases from the data they are trained on, for example, gender [3]. We seek to train a model that can still solve analogies well, but is less prone to these gender biases. We first calculate a "gender direction" g using a method based on Bolukbasi et al. [3] which gives a method for defining the protected variable. We will use this technique in the context of defining gender for word embeddings, but, as discussed in Bolukbasi et al. [3], the technique generalizes to other protected variables and other forms of embeddings. Following Bolukbasi et al. [3], we pick 10 (male, female) word pairs, and define the and define the *bias subspace* to be the space spanned by the top k principal components of the differences, where k is a tuneable parameter. In our experiments, we find that $k = 1$ gives reasonable results, so we did not experiment further.

We use embeddings trained from Wikipedia to generate input data from the Google analogy data set [9]. For each analogy in the dataset, we let $x = (x_1, x_2, x_3) \in \mathbb{R}^{3d}$ comprise the word vectors for the first three words, y be the word vector of the fourth word, and z be $\text{proj}_g y$. It is worth noting that these word vectors computed from the original embeddings are never updated nor is there projection onto the *bias subspace* and therefore the original word embeddings are never modified. What is learned is a tranform from a biased embedding space to a debiased embedding space.

biased		debiased	
neighbor	similarity	neighbor	similarity
nurse	1.0121	nurse	0.7056
nanny	0.9035	obstetrician	0.6861
fiancée	0.8700	pediatrician	0.6447
maid	0.8674	dentist	0.6367
fiancé	0.8617	surgeon	0.6303
mother	0.8612	physician	0.6254
fiance	0.8611	cardiologist	0.6088
dentist	0.8569	pharmacist	0.6081
woman	0.8564	hospital	0.5969

Table 1: Completions for `he : she :: doctor : ?`

As a model, we use the following: let $v = x_2 + x_3 - x_1$, and output $\hat{y} = v - ww^T v$, where our model parameter is w. Intuitively, v is the "generic" analogy vector as is commonly[5] used for the analogy task. If left to its own devices (*i.e.*, if not told to be unbiased on anything), the model should either learn $w = 0$ or else learn w as a useless vector.

By contrast, if we add the adversarial discriminator network (here, simply $\hat{z} = w_2^T \hat{y}$), we expect the debiased prediction model to learn that w should be something close to g (or $-g$), so that the discriminator cannot predict $z = proj_g y$. Indeed, both of these expectations hold: Without debiasing, the trained vector w is approximately a unit vector nearly perpendicular to $g : w^T g = 0.08, ||w|| = 0.82$; with debiasing, w is approximately a unit vector pointing in a direction highly correlated with $g : w^T g = 0.55, ||w|| = 0.96$. Even after debiasing, gendered analogies such as `man : woman :: he : she` are still preserved; however, many biased analogies go away, suggesting that the adversarial training process was indeed successful. An example of the kinds of changes in analogy completions observed after debiasing are illustrated in Table 1[6].

6.3 UCI Adult Dataset

To better align with the work in Beutel et al. [2], we attempt to enforce EQUALITY OF ODDS on a model for the task of predicting the income of a person – in particular, predicting whether the income is $> \$50k$ – given various attributes about the person, as made available in the UCI Adult dataset [1].

Details on the features that the dataset provides are available in Table 2. We use both categorical and continuous columns as given, with exception to the `fnlwgt` feature, which we discard. We convert the remaining columns into tensors where the categorical columns are sparse tensors, age is bucketized at boundaries $[18, 25, 30, 35, 40, 45, 50, 55, 60, 65]$, and the rest of the continuous columns are real-valued.

As discussed before, to enforce equality of odds, we give the adversary access to the true label y. The adversary will learn the relationship between y and z regardless of what the predictor does; further, if the predictor's predictions $\hat{y}$ give more information about

Feature	Type	Description
age	Cont	Age of the individual
capital_gain	Cont	Capital gains recorded
capital_loss	Cont	Capital losses recorded
education_num	Cont	Highest education level (numerical form)
fnlwgt	Cont	# of people census takers believe that observation represents
hours_per_week	Cont	Hours worked per week
education	Cat	Highest level of education achieved
income	Cat	Whether individual makes > \$50K annually
marital_status	Cat	Marital status
native_country	Cat	Country of origin
occupation	Cat	Occupation
race	Cat	White, Asian-Pac-Islander, Amer-Indian-Eskimo, Other, Black
relationship	Cat	Wife, Own-child, Husband, Not-in-family, Other-relative, Unmarried
sex	Cat	Female, Male
workclass	Cat	Employer type

Table 2: Features in the UCI dataset per individual. Features are either continuous (Cont) or Categorical (Cat). Categorical features are converted to sparse tensors for the model.

z than is already contained in y, the adversary will be able to improve its loss. Thus, the predictor, in attempting to fool the adversary, will move toward making sure that $\hat{y}$ does not give such additional information; in other words, toward equality of odds.

Our protected variable z is a binary-valued variable for the two sexes annotated, MALE and FEMALE.

Our predictor model is straightforward logistic regression: $\hat{y} = \sigma(w_1 \cdot x + b)$, where σ is the sigmoid function. Our adversary model takes the form of the following logistic-regression-like model:

$$s = \sigma\Big((1 + |c|)\sigma^{-1}(\hat{y})\Big) \quad \hat{z} = w_2 \cdot [s, sy, s(1-y)] + b$$

where c and b are learnable scalars, w_2 is a learnable vector, and σ^{-1} is the inverse of the sigmoid function (logit function) $\sigma^{-1}(t) = \log t - \log(1-t)$. Intuitively, we want our adversary to be able to learn functions of the form $\hat{z} = f(y, [\hat{y} > 0.5])$ (i.e. dependent only on the boolean predicted value $[\hat{y} > 0.5]$), and thus enforce equality of odds. Here, the adversary would learn such a function by making c extremely large. We add 1 to $|c|$ to make sure the adversary never tries to ignore $\hat{y}$ by setting $c = 0$, which could be a difficult local minimum for the adversary to escape[7]. This adversary is both general enough to be used whenever y and z are both discrete[8], and powerful enough that deviation from true equality of odds should cause the adversary to be able to decrease its loss.

Without tweaking, this algorithm ran into issues with local minima, and the resulting models were often closer to demographic parity than equality of odds. We implemented a technique that helped: by increasing the hyperparameter α in Eqn. 1 over time, the predictor had a much easier time learning to deceive the adversary

[5] see e.g. Mikolov et al. [9]

[6] The presence of *nurse* in the second position may seem worrying, but it should be noted that in this particular set of word embeddings, *nurse* is the nearest neighbor to *doctor*; no amount of debiasing will change this.

[7] This value added to $|c|$ is an adjustable hyperparameter; we found reasonable results using the value 1 and thus not feel the need to experiment further.

[8] If y and z are multi-class, then the sigmoid becomes a softmax, but everything else remains the same.

and therefore more strictly enforce equality of odds. We set $\alpha = \sqrt{t}$ (where t is the step counter), and to avoid divergence we set the predictor's step size to $\eta \propto 1/t$, so that $\alpha\eta \to 0$ as is preferred for stochastic gradient-based methods such as Adam.

We train the model twice, once with debiasing and once without, and present side-by-side confusion matrices on the test set for income bracket with respect to the protected variable values *Male* and *Female*, shown in Table 3, and we present the false positive rates (FPR) and false negative rates (FNR) in Table 4. Note that false negative rate is equal to 1 − true positive rate, so the trade-offs are directly comparable to the (x, y) values of an ROC curve.

Without Debiasing			With Debiasing		
Female	Pred 0	Pred 1	*Female*	Pred 0	Pred 1
True 0	4711	120	True 0	4518	313
True 1	265	325	True 1	263	327
Male	Pred 0	Pred 1	*Male*	Pred 0	Pred 1
True 0	6907	697	True 0	7071	533
True 1	1194	2062	True 1	1416	1840

Table 3: Confusion matrices on the UCI Adult dataset, with and without equality of odds enforcement.

We notice that debiasing has only a small effect on overall accuracy (86.0% vs 84.5%), and that the debiased model indeed (nearly) obeys equality of odds: as shown in Table 4, with debiasing, the FNR and FPR values are approximately equal across sex subgroups: $0.0647 \approx 0.0701$ and $0.4458 \approx 0.4349$.

Although the values don't exactly reach equality, neither difference is statistically significant: a two-proportion two-tail large sample z-test yields p-values 0.25 for $y = 0$ and 0.62 for $y = 1$.

7 CONCLUSION

In this work, we demonstrate a general and powerful method for training unbiased machine learning models. We state and prove theoretical guarantees for our method under reasonable assumptions, demonstrating in theory that the method can enforce the constraints that we claim, across multiple definitions of fairness, regardless of the complexity of the predictor's model, or the nature (discrete or continuous) of the predicted and protected variables in question. We apply the method in practice to two very different scenarios: a standard supervised learning task, and the task of debiasing word embeddings while still maintaining ability to perform a certain task (analogies). We demonstrate in both cases the ability to train a model that is demonstrably less biased than the original one, and yet still performs extremely well on the task at hand. We discuss difficulties in getting these models to converge. We propose, in the common case of discrete output and protected variables, a simple adversary that is usable regardless of the complexity of the underlying model.

		Female		Male	
		Without	With	Without	With
Beutel et al. [2]	FPR	0.1875	0.0308	0.1200	0.1778
	FNR	0.0651	0.0822	0.1828	0.1520
Current work	FPR	0.0248	0.0647	0.0917	0.0701
	FNR	0.4492	0.4458	0.3667	0.4349

Table 4: False Positive Rate (FPR) and False Negative Rate (FNR) for income bracket predictions for the two sex subgroups, with and without adversarial debiasing.

8 FUTURE WORK

This process yields many questions that require further work to answer.

(1) The debiased word embeddings we have trained are still useful in analogies. Are they still useful in other, more complex tasks?
(2) The adversarial training method is hard to get right and often touchy, in that getting the hyperparameters wrong results in quick divergence of the algorithm. What ways can be used to stabilize training and ensure convergence, and thus ensure that the theoretical guarantees presented here can work?
(3) There is a body of existing work for image recognition using adversarial networks. Image recognition in general can sometimes be subject to various biases such as being more or less successful at recognizing the faces of people of different races. Can multiple adversaries be combined to create high accuracy image recognition systems which do not exhibit such biases?
(4) In general, do more complex predictors require more complex adversaries? It would appear that in the case of y and z discrete, a very simple adversary suffices no matter how complex the predictor. Does this also apply to continuous cases, or would a simple adversary be too easy to deceive for a complex predictor?

REFERENCES

[1] Asuncion, A., and Newman, D. 2007. Uci machine learning repository.
[2] Beutel, A.; Chen, J.; Zhao, Z.; and Chi, E. H. 2017. Data decisions and theoretical implications when adversarially learning fair representations. *arXiv preprint arXiv:1707.00075*.
[3] Bolukbasi, T.; Chang, K.-W.; Zou, J. Y.; Saligrama, V.; and Kalai, A. T. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems*, 4349–4357.
[4] Goodfellow, I.; Pouget-Abadie, J.; Mirza, M.; Xu, B.; Warde-Farley, D.; Ozair, S.; Courville, A.; and Bengio, Y. 2014. Generative adversarial nets. In *Advances in neural information processing systems*, 2672–2680.
[5] Hardt, M.; Price, E.; Srebro, N.; et al. 2016. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, 3315–3323.
[6] Kingma, D., and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
[7] Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2016. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*.
[8] Lum, K., and Johndrow, J. 2016. A statistical framework for fair predictive algorithms. *arXiv preprint arXiv:1610.08077*.
[9] Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G. S.; and Dean, J. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, 3111–3119.