A Geometric Solution to Fair Representations

Yuzi He
Department of Physics & Astronomy
Information Sciences Institute
University of Southern California
yuzihe@usc.edu

Keith Burghardt
Information Sciences Institute
University of Southern California
keithab@isi.edu

Kristina Lerman
Information Sciences Institute
University of Southern California
lerman@isi.edu

ABSTRACT

To reduce human error and prejudice, many high-stakes decisions have been turned over to machine algorithms. However, recent research suggests that this *does not* remove discrimination, and can perpetuate harmful stereotypes. While algorithms have been developed to improve fairness, they typically face at least one of three shortcomings: they are not interpretable, their prediction quality deteriorates quickly compared to unbiased equivalents, and they are not easily transferable across models. To address these shortcomings, we propose a geometric method that removes correlations between data and any number of protected variables. Further, we can control the strength of debiasing through an adjustable parameter to address the trade-off between prediction quality and fairness. The resulting features are interpretable and can be used with many popular models, such as linear regression, random forest, and multilayer perceptrons. The resulting predictions are found to be more accurate and fair compared to several state-of-the-art fair AI algorithms across a variety of benchmark datasets. Our work shows that debiasing data is a simple and effective solution toward improving fairness.

CCS CONCEPTS

• **Security and privacy → Privacy-preserving protocols**; **Social aspects of security and privacy**; **Privacy protections**; • **Computing methodologies → Linear algebra algorithms**; **Machine learning**; **Supervised learning by classification**; **Supervised learning by regression**; **Machine learning approaches**; • **Applied computing → Law, social and behavioral sciences**.

KEYWORDS

fair AI, debiased features, sensitive information, fair classification, geometric method, interpretable method, projection, orthogonal space

ACM Reference Format:
Yuzi He, Keith Burghardt, and Kristina Lerman . 2020. A Geometric Solution to Fair Representations. In *Proceedings of the 2020 AAAI/ACM Conference on AI, Ethics, and Society (AIES '20), February 7–8, 2020, New York, NY, USA.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3375627.3375864

AIES '20, February 7–8, 2020, New York, NY, USA

ACM ISBN 978-1-4503-7110-0/20/02…$15.00
https://doi.org/10.1145/3375627.3375864

1 INTRODUCTION

Machine learning (ML) models sift through mountains of data to make decisions on matters big and small: e.g., who should be shown a product, hired for a job, or given a home loan. Machine inference can systematize decision processes to take into account orders of magnitude more information, produce accurate decisions, and avoid the common pitfalls of human judgment, such as belief in a just world or selective attention [18]. Moreover, unlike people, machines will never make poor decisions when tired [9], pressed for time or distracted by other matters [22, 30].

Recent research suggests, however, that discrimination remains pervasive [1, 6, 10, 26]: for example, a model used to evaluate criminal defendants for recidivism assigned systematically higher risk scores to African Americans than to Caucasians [1]. As a result, reformed African American defendants, who would never commit another crime, were deemed by the model to present a higher risk to society—as much as twice as high [1, 10]—as reformed white defendants, with potentially grave consequences on how they were treated by the justice system.

The emerging field of AI fairness has suggested ways to mitigate harmful model biases [6, 7, 11], e.g., penalizing unfair inferences [2, 11], or creating representations that do not strongly depend on protected features [14, 20, 23]. These methods, however, fall short in one or more critical dimensions: interpretability, prediction quality, and generalizability. We define *interpretability* as the ability to understand how features affect—or bias—a model's predicted outcome. Interpretability is needed to improve transparency and accountability of AI systems. While models must sacrifice *prediction quality* (as measured by accuracy, mean squared error, or another metric) to improve fairness [27], the trade-off does not need to be as drastic as what current methods achieve. Finally, we define *generalizability* as the ability to easily apply fairness algorithms across multiple models and datasets. In contrast, state-of-the-art fairness methods are specialized to linear regressions or random forests [2, 16, 35]. Similarly, methods that create fair latent features for neural networks (NN) [14, 23] cannot be easily applied to improve fairness in non-NN models. These fair AI algorithms were not meant to be generalizable because there does not seem to be adequate meta-algorithms that debias a whole host of ML models. One might naively expect that we can just create a single fair model and apply it to all datasets. The problem is that model performance varies greatly on different datasets. While NNs are critical for, e.g., image recognition [8], other methods perform better for small data [25], especially when the number of dimensions is high and the sample size low [19]. There is no one-size-fits-all model and there is no one-size-fits-all model debiasing method. Is there an easier way to create fairer predictions other than specialized methods for

specialized ML models? Chen et al. offer some clues to addressing this fundamental issue in fair AI [5]: by addressing data biases, we can potentially improve fair AI across the spectrum of models, and achieve fairness without greatly sacrificing prediction quality.

Inspired by these ideas, we describe a geometric method for *debiasing features*. Depending on the hyperparameter we choose, these features are mathematically guaranteed to be uncorrelated with specified sensitive, or *protected*, features. This method is exceedingly fast and the debiased features are highly correlated with the original features (average Pearson correlations are between 0.993–0.994 across the three datasets studied in this paper). These debiased features are as interpretable as the original features when applied to any model. When applied to linear regression, for example, the coefficients are the same or similar to the coefficients of the original features when controlling for protected variables (see Methods). These debiased features serve as a fair representation of data that can be used with a number of NN and non-NN ML models, such as linear regression, random forest, support vector machines (SVMs), and multilayer perceptrons (MLPs). While previous methods have created fair representations [14, 23, 24, 29], these methods create representations that are either not very interpretable, like PCA components, or the relationship between these fair representations and the original features have not been established. We evaluate the proposed approach on several benchmark datasets. We show that models using these debiased features are more accurate for almost any level of fairness we desire.

In the rest of the paper, we first review recent advances in fair AI to highlight the novelty of our method. Next, we describe in the Methods section our methodology to improve data fairness, and the definitions of fairness we use in the paper. In Results, we describe how our method improves fairness in both synthetic data and empirical benchmark data. We compare to several competing methods and demonstrate the advantages of our method. Finally, we summarize our results and discuss future work in the Conclusion section.

2 RELATED WORK

Social scientists use linear regression for data analysis due to its simplicity and interpretability. Interpretability comes from regression coefficients, which specify how the outcome, or response, changes when features change by one unit. However, regression creates unfair outcomes, even when protected features are excluded from the model, because other features may be correlated with them.

To make regression models fair, researchers introduced a loss function to penalize regression for unfair outcomes [2]. Similarly, [33] created fair logistic regression by introducing fairness constraints that limit the covariance between protected features and the outcome. An alternate method achieved fairness by constraining false positive or false negative rates [34]. There are some issues in these works, however. First, protected features are not included in the logistic model with fairness constraints. While this improves privacy, it forces the parameters of logistic models to take certain combinations which will minimize the correlation with the protected features. This can reduce the accuracy when the constraints are strict. The issue for the second method is mainly numeric. The algorithm requires an optimization of a convex loss function on a non-convex parameter space. While these models are generally interpretable, the approaches do not transfer to other models.

Researchers have explored a variety of fair data representation methods [14, 21, 23, 24, 29, 32, 36]. Some of those works use NNs to embed raw features in a lower-dimensional space, such that the embedding will contain the information about the outcome variable, but at the same time, contain little information about the protected feature. Fair logistic models or fair scoring, on the other hand, can be regarded as a one dimensional embedding of data, which makes sure that the predictions, $\hat{y}$, are independent of the protected features. They are mainly used with NNs, which are accurate but often lack interpretability. Two methods were instead developed to improve fairness of PCA features [24, 29]. While they can be applied to many ML models, they lack interpretability compared to the original features.

Johndrow and Lum (2017) proposed an algorithm which removes sensitive information about protected groups based on inverse transform sampling. The algorithm transforms individual features such that the transformed features satisfy the marginal distribution. Although this method can guarantee that predictions are fair in a probabilistic sense, it has a critical disadvantage — as the number of protected features n_p increases, the number of protected groups increases as $O(2^{n_p})$. This means that in order to properly estimate conditional and marginal distribution of features, one needs exponentially increasing population size. Our method overcomes these difficulties by using linear algebra as the basis for learning unbiased representations. This allows our algorithm to only take $O(n_p^2)$ time to debias data. Moreover, our method is a white box: it is interpretable and can be fully scrutinized, unlike a black box method.

3 METHODS

We describe a geometric method for constructing fair interpretable representations. These representations can be used with a variety of ML methods to create fairer accurate models of data.

3.1 Fair Interpretable Representations

We consider tabular data with n entries and m features. The features are vectors in the n-dimensional space, denoted as $\boldsymbol{x}_i$ where $i = 1, 2, \cdots, m$, and one of the columns corresponds to the outcome, or target variable $\boldsymbol{y}$. Among the features, there are also n_p protected features, $\boldsymbol{p}_i, i = 1, \ldots, n_p$. As a pre-processing step, all features are centered around the mean: $\langle x_i \rangle = 0$.

We describe a procedure to debias the data so as to create linearly fair features. We aim to construct a representation $\boldsymbol{r}_j$ of a feature $\boldsymbol{x}_j$, that is uncorrelated with n_p protected columns $\boldsymbol{p}_i, i = 1, \ldots n_p$, but highly correlated to feature $\boldsymbol{x}_j$. We recall that Pearson correlation between the representation r_j and any feature x_k is defined as

$$\mathrm{Corr}(r_j, x_k) = (\mathrm{E}[r_j \cdot x_k] - \mathrm{E}[r_j]\mathrm{E}[x_k])/(\sigma_{r_j}\sigma_{x_k}),$$

where $\mathrm{E}[.]$ is the expectation, and $\sigma_{r_j} = \sqrt{\mathrm{E}[r_j^2] - \mathrm{E}[r_j]^2}$ and $\sigma_{x_k} = \sqrt{\mathrm{E}[x_k^2] - \mathrm{E}[x_k]^2}$. Because all the features are centered (and we also assume that r_j is centered), $\mathrm{E}[r_j] = \mathrm{E}[x_k] = 0$, we have

$$\sigma_{r_j} = \sqrt{\mathrm{E}[r_j^2]} = \|r_j\|/\sqrt{n},$$

$$\sigma_{x_k} = \sqrt{\mathrm{E}[x_k^2]} = \|x_k\|/\sqrt{n}$$

and

$$\mathrm{E}[r_j \cdot x_k] = \boldsymbol{r}_j \cdot \boldsymbol{x}_k/n.$$

Therefore

$$\mathrm{Corr}(r_j, p_i) = \boldsymbol{r}_j \cdot \boldsymbol{p}_i/(\|r_j\| \cdot \|p_i\|)$$

and

$$\mathrm{Corr}(r_j, x_j) = \boldsymbol{r}_j \cdot \boldsymbol{x}_j/(\|r_j\| \cdot \|x_j\|).$$

Zero correlations between r_j and n_p protected columns requires that r_j lives in the solution space of $\boldsymbol{r}_j \cdot \boldsymbol{p}_i = 0, i = 1 \ldots n_p$. Maximizing correlations between r_j and x_j under this constraint is equivalent to projecting $\boldsymbol{x}_j$ into the solution space of $\boldsymbol{r}_j \cdot \boldsymbol{p}_i = 0, i = 1 \ldots n_p$.

To calculate $\boldsymbol{r}_j$, we can first create an orthonormal basis of vectors $\boldsymbol{p}_i$, which we can label as $\bar{\boldsymbol{p}}_i$. We then construct a projector $P_f = \sum_{i=1}^{n_p} \bar{\boldsymbol{p}}_i \bar{\boldsymbol{p}}_i^T$. The representation $\boldsymbol{r}$ is given as

$$\boldsymbol{r}_j = \boldsymbol{x}_j - P_f \boldsymbol{x}_j = (I - P_f)\boldsymbol{x}_j. \tag{1}$$

Using the Gram–Schmidt process, the orthonormal basis can be constructed in $O(n \times n_p^2)$ time and for every fair representation of features, the projection takes $O(n \times n_p)$ time. Given n_f features, the total time of the algorithm is $O(n \times n_f \times n_p^2)$ Therefore our method scales linearly with respect to the size of the data and the number of features. In practice, this is exceedingly fast. For example, this algorithm only takes less than 200 milliseconds to run on the Adult dataset described below, which has 45K rows, 103 unprotected features, and 1 protected feature.

While the previous discussion was on how to create linearly fair features, one can make linearly fair outcome variables, $\boldsymbol{r}_y$ through the same process. In prediction tasks, however, we do not have access to the outcome data. While our method does not guarantee that every model's estimate of the outcome variable, $\hat{y}$ is fair, we find that it can significantly improve the fairness compared to competing methods. Moreover, in the special case of linear regression, it can be shown that the resulting estimate, $\hat{y}$, is uncorrelated with the protected variables.

Inevitably, the prediction quality of a model using such linearly fair features will drop compared to using the original features, because the solution is more constrained. To address this issue, we introduce a parameter $\lambda \in [0, 1]$, which indicates the fairness level. We define the parameterized latent variable as

$$\boldsymbol{r}'_j(\lambda) = \boldsymbol{r}_j + \lambda \cdot (\boldsymbol{x}_j - \boldsymbol{r}_j). \tag{2}$$

Here, $\lambda = 0$ corresponds to $\boldsymbol{r}'_j(\lambda) = \boldsymbol{r}_j$, which is strictly orthogonal to the protected features $\boldsymbol{p}_i$; while $\lambda = 1$ gives $\boldsymbol{r}'(\lambda) = \boldsymbol{x}_j$.

The protected features can be both real valued and cardinal. The fair representation method can also handle categorical protected features by introducing dummy variables. Specifically, if a variable X has k categories $x_1, x_2, ..., x_k$, we can convert them to $k-1$ binary variables where the i^{th} variable is 1 if the variable is category x_i, and otherwise 0. If all variables are 0, then the category is x_k. As a simple example, if a feature X has 3 categories, x_1, x_2, and x_3, then the dummy variables would be $\tilde{x}_1$ and $\tilde{x}_2$. If $\tilde{x}_1 = 1$, the category is x_1, if $\tilde{x}_2 = 1$, then the category is x_2, and otherwise is x_3. The condition of fairness in this case is interpreted as same mean value of the latent variables in different categorical groups.

3.2 Fair Models

Using the procedure described above, we can construct a fair representation of every feature, and use the fair features to model the outcome variable. Consider a linear regression model that includes all features: n_p protected features $p_i, i = 1, \cdots, n_p$ and $n_f = m - n_p$ non-protected features features $x_i, i = 1, \ldots, n_f$.

$$\hat{y} = \beta_0 + \sum_{i=1}^{n_f} \beta_i x_i + \sum_{i=1}^{n_p} \gamma_i p_i. \tag{3}$$

After transforming the features to fair features x'_i, the fair regression model reduces to:

$$\hat{y}' = \beta'_0 + \sum_{i=1}^{n_f} \beta'_i r_i. \tag{4}$$

Here, r_i corresponds to the fair versions of x_i. We can prove that $\beta_i = \beta'_i, i = 1, \ldots, n_f$, but the predicted value $\hat{y}'$ is uncorrelated with protected features $p_i, i = 1, \ldots, n_p$. In general linear regression, such as logistic regression, this proof does not hold, but we numerically find that coefficients are similar.

We should take a step back at this point. The fair latent features are close approximations of the original features, therefore we expect that, and in certain cases can prove, that the regression coefficients of the fair features should be approximately the coefficients of the original features. The fair features can, by this definition, be considered almost as interpretable as the original features.

In addition to regression, fair representations could be used with other ML models, such as AdaBoost [12], NuSVM [4], random forest [3], and multilayer perceptrons [28].

3.3 Measuring Fairness

While there exists no consensus for measuring fairness, researchers have proposed a variety of metrics, some focusing on representations and some on the predicted outcomes [13, 31]. We will therefore compare our method to competing methods using the following metrics: Pearson correlation, mutual information, discrimination, calibration, balance of classes, and accuracy of the inferred protected features. Due to space limitations, we leave mutual information out of our analysis in this paper, and do not compare calibration and balance of classes to model accuracy. Results in all cases are similar.

3.3.1 Fairness of Outcomes. One can argue that outcomes are fair if they do not depend on the protected features. If this is the case, a malicious adversary won't be able to guess the protected features from the model's predictions. One way to quantify the dependence is through *Pearson correlation* between (real valued or cardinal) predictions and protected features. For models making binary predictions, fairness can be measured using the *mutual information* between predictions and the protected features, given that protected features are discrete. We find mutual information and Pearson correlations create qualitatively similar findings, despite mutual information being a non-linear metric, therefore we focus on Pearson correlations in this paper. Previous work [36] has also defined a *discrimination metric* for binary predictions as below. Consider a protected variable p_1, a binary prediction $\hat{y}$ of an outcome y. The metric measures the bias of a binary prediction $\hat{y}$ with respect to a single binary

protected feature p_1 using the difference of positive rates between the two groups.

$$y_{\text{Discrim}} = \left| \frac{\sum_{n:p_1[n]=0} \hat{y}[n]}{\sum_{n:p_1[n]=0} 1} - \frac{\sum_{n:p_1[n]=1} \hat{y}[n]}{\sum_{n:p_1[n]=1} 1} \right| \quad (5)$$

For real valued predictions ($\hat{y} \in [0, 1]$), Kleinberg et al. (2016) suggested a more nuanced way to measure fairness:

- **Calibration within groups**: Individuals assigned predicted probability $\hat{y} \in [r_0 - \delta, r_0 + \delta]$, ($\delta > 0$ and $\delta \ll 1$) should have an approximate positive rate of r. This should hold for both protected groups ($p_1 = 0$ and $p_1 = 1$).
- **Balance for the negative class**: The mean $\hat{y}$ of group $p_1 = 0, y = 0$ and group $p_1 = 1, y = 0$ should be the same.
- **Balance for the positive class**: The mean $\hat{y}$ of group $p_1 = 0, y = 1$ and group $p_1 = 1, y = 1$ should be the same.

In some cases, calibration error is difficult to calculate, as it depends on how predictions are binned. In these cases, we can measure calibration error using log-likelihood of the labels given the real valued predictions as a proxy. By definition, logistic regression maximizes the (log-)likelihood function, assuming the observations are sampled from independent Bernoulli distributions where $P(y[n]|X[n]) = \hat{y}_i[n]$. Better log-likelihood implies that the individuals assigned probabilities $\hat{y} \in [r_0 - \delta, r_0 + \delta]$ are more likely to have a positive rate r, which is better calibrated according to Kleinberg et al.

3.3.2 Fairness of Representations. Several past studies examined the fairness of representations, arguing that models using fair representations will also make fair predictions. Learned representations are considered fair if they do not reveal any information about the protected features [14, 21, 23, 31, 32]. The studies trained a discriminator to predict protected features from the learned representations—using accuracy as a measure of fairness.

Following this approach, we treat the predicted probabilities as a one-dimensional representation of data and use the *accuracy of the inferred protected features* as a measure of fairness. However, this method is not effective in situations where the protected classes are unbalanced. Let us assume the fair representation is R and the protected feature is p_1. For simplicity, we only consider the case of a single binary protected feature. The discriminator infers the protected feature in a Bayesian way, namely,

$$P(p_1 = c|R) = \frac{P(R|p_1 = c)P(p_1 = c)}{P(R)}, c = 0|1 \quad (6)$$

In the case where there is a large difference between $P(p_1 = 0)$ and $P(p_1 = 1)$, even if there is useful information in the distribution $P(R|p_1 = c)$, the discriminator will not perform significantly better than the baseline model, the majority class classifier.

4 RESULTS

We demonstrate how our method can achieve fair classification using synthetic data (see Appendix), and also compare our prediction quality and fairness to other fair AI algorithms using benchmark datasets.

German dataset has 61 features about 1,000 individuals, with a binary outcome variable denoting whether an individual has a good credit score or not. The protected feature is gender. (https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data))

COMPAS dataset contains data about 6,172 defendants. The binary outcome variable denotes whether the defendant will recidivate (commit a crime) within two years. The protected feature is race (whether the race is African American or not), and there are nine features in total. (https://github.com/propublica/compas-analysis)

Adult dataset contains data about 45,222 individuals. The outcome variable is binary, denoting whether an individual has more than $50,000. The protected feature is age, and there are 104 features in total. (https://archive.ics.uci.edu/ml/datasets/Adult)

Debiased features had mean correlations of 0.993, 0.994, and 0.994, for the German, COMPAS, and Adult data, respectively. We reserved 20% of the data in the Adult and COMPAS datasets for testing and used the remaining data to perform 5-fold cross validation. This ensured no leakage of information from the training set to the testing set. The German dataset is much smaller than the rest, so it was randomly divided into five folds of training, validation and testing sets. Each set had 50%, 20% and 30% of all the data. We measured the performance metrics on the test data.

We varied the fairness parameter λ between 0 and 1 and applied the debiased features to logistic regression, AdaBoost, NuSVM, random forest, and multilayer perceptrons. In practice, one could use a host of commercial ML models and pick the most accurate one given their fairness tolerance.

4.1 Comparison Against State-of-the-Art

We compared our method to several previous fair AI algorithms. For the models proposed by [33, 34], we vary the fairness constraints from perfect fairness to unconstrained. For the "Unified Adversarial Invariance" (UAI) model proposed by [14], we vary the δ term in the loss function from 0 (no fairness) to very large value, e.g., 9.0×10^{19} for COMPAS dataset, (large δ value corresponds to perfect fairness). The predictions of the UAI model for the German and Adult datasets are provided by the authors. We are interested in (1) how different models tradeoff between accuracy and fairness and (2) how different metrics of fairness compare to each other.

Fairness Versus Accuracy. We first investigate the trade-offs between prediction accuracy (*Acc Y*) and fairness, which we measure three different ways: (1) Pearson correlation between the protected feature and model predictions, (2) discrimination between the binary protected feature and the binarized predictions (predicted probabilities above 1/2 are given a value of 1, and are otherwise 0) and (3) the accuracy of predicting protected features from the predictions (*Acc P*). To robustly predict the protected features from the model predictions, we used both a NN with three hidden layers, which is used by former works [14, 21, 23, 32, 36] and a random forest model. We report the better accuracy of those two models. Figure 1,2 and 3 shows the resulting comparisons.

The figures show that models using the proposed fair features achieve significantly higher accuracy—for the same degree of fairness—compared to competing methods. Equivalently, we achieve greater fairness with equivalent accuracy. In Fig. 3, we find *Acc P* shows little difference from the baseline majority class classifier for the German and Adult datasets. The reason is explained in Eq.(6). On the other hand, *Acc P* of COMPAS dataset shows a clear trend

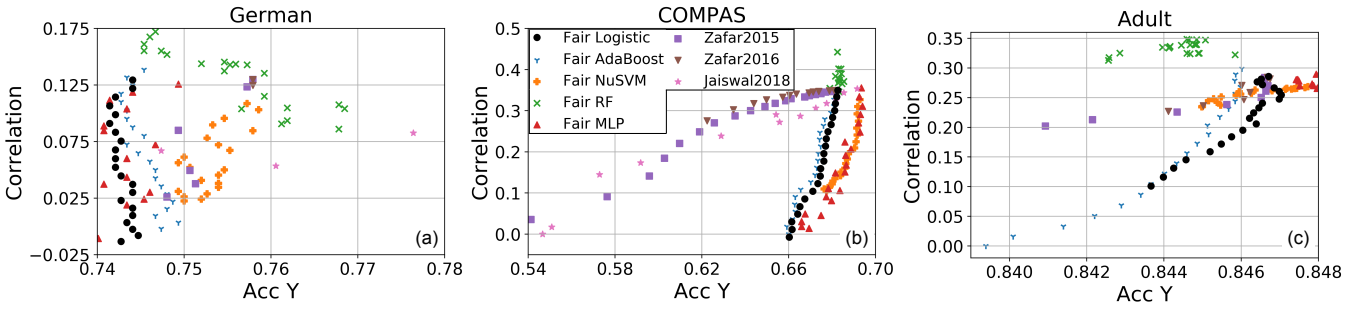

Figure 1: Fairness versus accuracy. Plots show Pearson correlation versus accuracy of predictions (*Acc Y*) for the German, COMPAS and Adult datasets. For each plot, *Zafar2015* stands for [33], *Zafar2016* for [34] and *Jaiswal2018* for [14]. *Fair NuSVM*, *Fair RF*, *Fair AdaBoost*, and *Fair MLP* results are produced using the fair representations constructed by our proposed method with NuSVM [4], random forest [3], AdaBoost [12], and multilayer perceptrons [28] models, respectively. The results of UAI are not shown for the Adult dataset, since its best accuracy (0.83) lies outside of the boundary of the plot. (Same for Figure 2 and 3.)

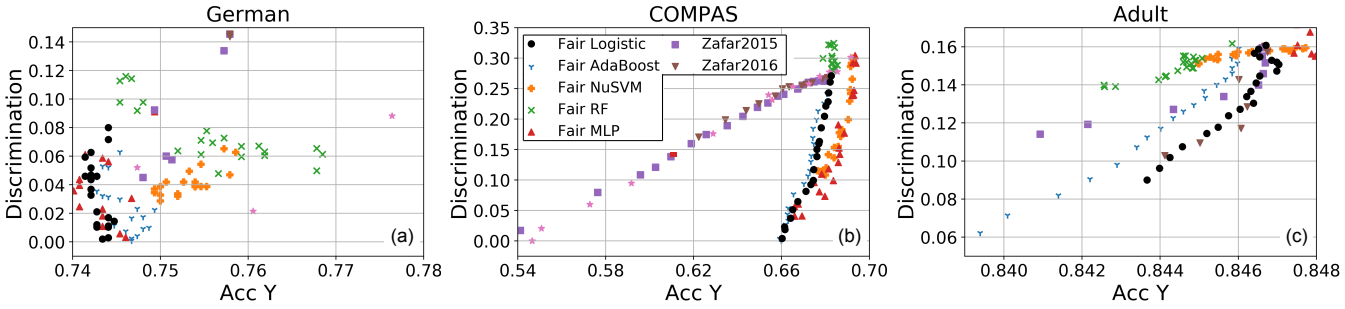

Figure 2: Discrimination versus accuracy plots for the three datasets.

because the majority baseline is around 0.51, which is consistent with the Eq.(6). For the Adult dataset, the fair logistic regression cannot achieve perfect fairness but the situation is improved by AdaBoost. We discover, in other words, that there is no single ML model that achieves greater accuracy for a given value of fairness, but our method allows us to choose suitable models to achieve greater accuracy.

4.1.1 Fairness of Representations. We compared our method to earlier works using fair representations. Previous works used NNs to encode the features into a high dimensional embedding space and then separately trained discriminators to infer the protected feature and the outcome variables. The accuracy of inferring protected feature and outcome are reported. Ideally, the accuracy for the outcome should be high and the accuracy of inferring the protected features should be close to the majority class baseline. We set the fairness level to $\lambda = 0$ (perfect fairness). We show *Acc P* and *Acc Y* for various methods in Table **??** (Appendix) and Fig. 3. Our method applied to a logistic model has similar fairness to the best existing methods but is very fast, easy to understand, and creates more interpretable features.

4.1.2 Balance Versus Calibration. Finally, we use another measure of fairness that captures the degree to which each model makes mistakes. Figure 4 shows delta score (i.e., balance) versus negative log-likelihood (i.e., calibration error). Fairer predictions are located in the lower left corner of each figure, meaning that there are fewer differences in outcomes for the different classes. We only compare the logistic model with fair features to the models proposed by Zafar et al. [33, 34], because these models maximize the log-likelihood function (minimize calibration error) when selecting parameters. For all datasets, our method generally achieves greater fairness.

5 CONCLUSION

We show that our algorithm simultaneously achieves three advances over many previous fair AI algorithms. First, it is interpretable; the features we construct are minimally affected by our fair transform. While this does not mean the models trained on these features are interpretable (they could be a black box), it does mean that any method used to interpret features could easily be used for these fairer features as well. Next, the features better preserve model prediction quality. Namely, models using these features were more accurate than competing methods when the value of the fairness metric was held fixed. This is in part due to the third principle: that our method can be applied to any number of commercial models; it merely acts as a pre-processing step. Different models have different strengths and weaknesses; while some are more accurate, others are fairer. We can pick and choose particular models that achieve both high fairness and accuracy, whether it is

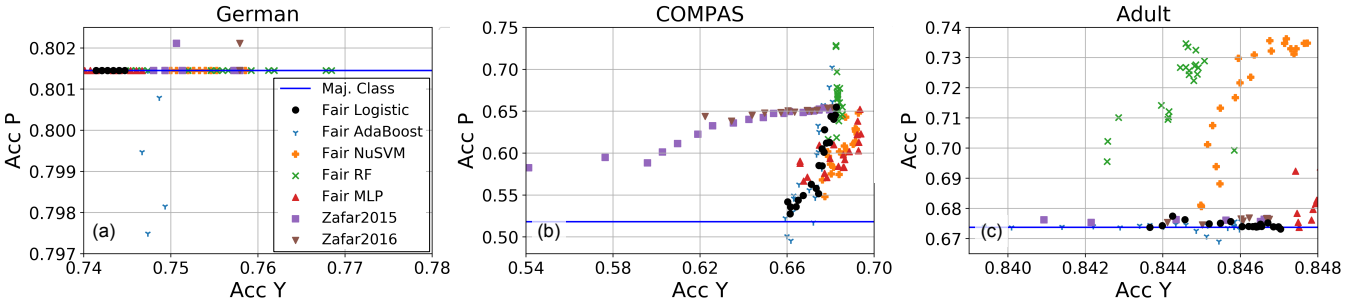

Figure 3: Accuracy of inferring the protected variable from the model's predictions (*Acc P*) versus the accuracy of predicting the outcome (*Acc Y*) for the three datasets.

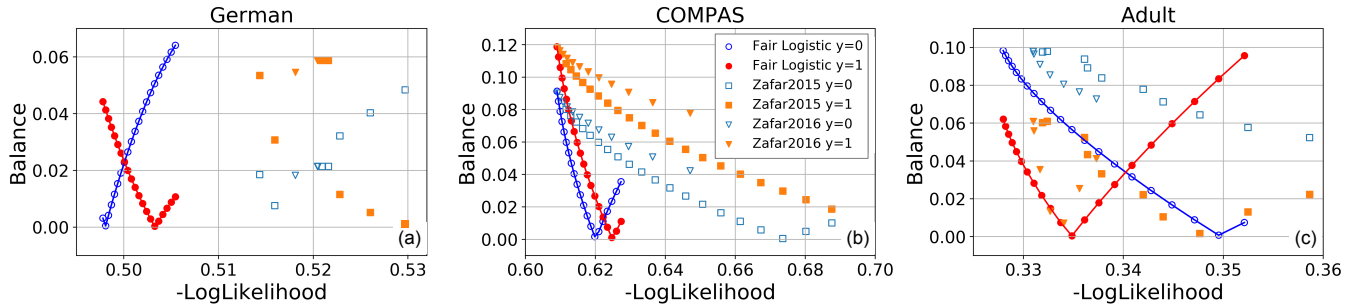

Figure 4: Balance vs. negative log-likelihood (calibration error) for the German, COMPAS and Adult datasets. In the plot, there are two sets of curves for every model, labeled $y = 0$ and $y = 1$. $y = 0$ stands for the difference of mean $\hat{y}$ (between different protected classes) given to the individuals with negative $y = 0$, and $y = 1$ stands for individuals with positive outcomes $y = 1$. (These differences are called balance of negative or positive class by [17].) Fairer models are those in the lower left corner of each plot.

a linear model like logistic regression or a non-linear model like a multilayer perceptron, as shown in Figs. 1, 2., & 3.

We propose some ideas for future work. First, while making linearly fair features works very well in practice, the fairness could be improved by removing non-linear correlations. Second, we can extend our method to more easily address categorical protected variables. In the present method, a categorical variable with alphabet size n becomes a set of $n - 1$ bivariate variables. It would be ideal, however, if a method reduced the mutual information between the categorical variable directly, rather than first creating $n - 1$ variables, and removed correlations.

ACKNOWLEDGMENTS

Authors would like to thank Ayush Jaiswal for providing the code for learning adversarial models and feedback on results. Authors also thank Daniel Moyer and Greg Ver Steeg for insightful discussions about the approach. This material is based upon work supported in part by the Defense Advanced Research Projects Agency (DARPA)) under Contracts No. W911NF-18-C-0011 and HR00111990114. This research is also based upon work supported in part by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via 2017-17071900005. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

REFERENCES

[1] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing
[2] Richard Berk, Hoda Heidari, Shahin Jabbari, Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. 2017. A Convex Framework for Fair Regression. (2017), 1–15. arXiv:1706.02409 http://arxiv.org/abs/1706.02409
[3] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.
[4] Chih-Chung Chang and Chih-Jen Lin. 2011. LIBSVM: A library for support vector machines. *ACM TIST* 2, 3 (2011), 27.
[5] Irene Chen, Fredrik D. Johansson, and David Sontag. 2018. Why Is My Classifier Discriminatory? (2018). https://doi.org/arXiv:1805.12002v2 arXiv:1805.12002
[6] Alexandra Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data* 5, 2 (2017), 153–163.
[7] Alexandra Chouldechova and Aaron Roth. 2018. The frontiers of fairness in machine learning. *arXiv preprint arXiv:1810.08810* (2018).
[8] D. Ciregan, U. Meier, and J. Schmidhuber. 2012. Multi-column deep neural networks for image classification. In *CVPR*. 3642–3649. https://doi.org/10.1109/CVPR.2012.6248110
[9] Shai Danziger, Jonathan Levav, and Liora Avnaim-Pesso. 2011. Extraneous factors in judicial decisions. *PNAS* 108, 17 (2011), 6889–6892. https://doi.org/10.1073/pnas.1018033108
[10] Julia Dressel and Hany Farid. 2018. The accuracy, fairness, and limits of predicting recidivism. *Science advances* 4, 1 (2018), eaao5580.

[11] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *ITCS*. ACM, 214–226.

[12] Yoav Freund and Robert E Schapire. 1997. A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of computer and system sciences* 55, 1 (1997), 119–139.

[13] Ben Hutchinson and Margaret Mitchell. 2019. 50 Years of Test (Un) fairness: Lessons for Machine Learning. In *FAT**. ACM, 49–58.

[14] Ayush Jaiswal, Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. 2018. Unsupervised Adversarial Invariance. In *NIPS*. Curran Associates, Inc., 5092–5102. arXiv:1809.10083 http://arxiv.org/abs/1809.10083

[15] James E. Johndrow and Kristian Lum. 2017. An algorithm for removing sensitive information: application to race-independent recidivism prediction. (2017), 1–25. arXiv:1703.04957 http://arxiv.org/abs/1703.04957

[16] F Kamiran, T Calders, and M Pechenizkiy. 2010. Discrimination Aware Decision Tree Learning. In *ICDM*. 869–874. https://doi.org/10.1109/ICDM.2010.50

[17] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. 2016. Inherent Trade-Offs in the Fair Determination of Risk Scores. (sep 2016), 1–23. arXiv:1609.05807 http://arxiv.org/abs/1609.05807

[18] Arie W. Kruglanski and Icek Ajzen. 1983. Bias and error in human judgment. *European Journal of Social Psychology* 13, 1 (1983), 1–44. https://doi.org/10.1002/ejsp.2420130102 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/ejsp.2420130102

[19] Bo Liu, Ying Wei, and Yu Zhang amd Qiang Yang. 2017. Deep Neural Networks for High Dimension, Low Sample Size Data. In *IJCAI*. 2287–2293.

[20] Francesco Locatello, Gabriele Abbati, Tom Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. 2019. On the Fairness of Disentangled Representations. *arXiv preprint arXiv:1905.13662* (2019).

[21] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard Zemel. 2015. The Variational Fair Autoencoder. (2015), 1–11. arXiv:1511.00830 http://arxiv.org/abs/1511.00830

[22] Anandi Mani, Sendhil Mullainathan, Eldar Shafir, and Jiaying Zhao. 2013. Poverty impedes cognitive function. *science* 341, 6149 (2013), 976–980.

[23] Daniel Moyer, Shuyang Gao, Rob Brekelmans, Greg Ver Steeg, and Aram Galstyan. 2018. Invariant Representations without Adversarial Training. Nips (2018). arXiv:1805.09458 http://arxiv.org/abs/1805.09458

[24] Matt Olfat and Anil Aswani. 2018. Convex Formulations for Fair Principal Component Analysis. (2018). arXiv:1802.03765 http://arxiv.org/abs/1802.03765

[25] Matthew Olson, Abraham J. Wyner, and Richard Berk. 2018. Modern Neural Networks Generalize on Small Data Sets. In *NIPS*. 3623–3632.

[26] Cathy O'Neil. 2016. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Broadway Books.

[27] Emma Pierson, Camelia Simoiu, Jan Overgoor, Sam Corbett-Davies, Vignesh Ramachandran, Cheryl Phillips, and Sharad Goel. 2017. A large-scale analysis of racial disparities in police stops across the United States. *preprint arXiv:1706.05678* (2017).

[28] Frank Rosenblatt. 1961. *Principles of Neurodynamics: Perceptrons and the Theory of Brain Mechanisms*. Spartan Books, Washington DC.

[29] Samira Samadi, Uthaipon Tantipongpipat, Jamie Morgenstern, Mohit Singh, and Santosh Vempala. 2018. The Price of Fair PCA: One Extra Dimension. Nips (2018). https://doi.org/10.1152/ajprenal.00633.2017 arXiv:1811.00103

[30] Anuj K Shah, Sendhil Mullainathan, and Eldar Shafir. 2012. Some consequences of having too little. *Science* 338, 6107 (2012), 682–685.

[31] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE, 1–7.

[32] Qizhe Xie, Zihang Dai, Yulun Du, Eduard Hovy, and Graham Neubig. 2017. Controllable invariance through adversarial feature learning. *NIPS* 2017-December, Mmd (2017), 586–597. arXiv:arXiv:1705.11122v3

[33] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2015. Fairness Constraints: Mechanisms for Fair Classification. 54 (2015). https://doi.org/10.1109/TRO.2009.2019886 arXiv:1507.05259

[34] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P. Gummadi. 2016. Fairness Beyond Disparate Treatment & Disparate Impact: Learning Classification without Disparate Mistreatment. (2016). https://doi.org/10.1145/3038912.3052660 arXiv:1610.08452

[35] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P. Gummadi, and Adrian Weller. 2017. From Parity to Preference-based Notions of Fairness in Classification. In *NIPS*. arXiv:1707.00010 http://arxiv.org/abs/1707.00010

[36] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. 2013. Learning Fair Representations. In *ICML*. 325–333. http://proceedings.mlr.press/v28/zemel13.html