Optimized Score Transformation for Fair Classification

Dennis Wei
IBM Research

Karthikeyan Natesan Ramamurthy
IBM Research

Flavio P. Calmon
Harvard University

Abstract

This paper considers fair probabilistic classification where the outputs of primary interest are predicted probabilities, commonly referred to as scores. We formulate the problem of transforming scores to satisfy fairness constraints while minimizing the loss in utility. The formulation can be applied either to post-process classifier outputs or to pre-process training data, thus allowing maximum freedom in selecting a classification algorithm. We derive a closed-form expression for the optimal transformed scores and a convex optimization problem for the transformation parameters. In the population limit, the transformed score function is the fairness-constrained minimizer of cross-entropy with respect to the optimal unconstrained scores. In the finite sample setting, we propose to approach this solution using a combination of standard probabilistic classifiers and ADMM. Comprehensive experiments comparing to 10 existing methods show that the proposed FairScoreTransformer has advantages for score-based metrics such as Brier score and AUC while remaining competitive for binary label-based metrics such as accuracy.

1 INTRODUCTION

Recent years have seen a surge of interest in *fair classification*, which is concerned with disparities in classification output or performance when conditioned on a protected attribute such as race or gender. Many measures of fairness and fairness-enhancing interventions have been introduced (see supplemental material (SM) for citations). Roughly categorized, these interventions either (i) change data used to train a classifier (pre-processing), (ii) change a classifier's output (post-processing), or (iii) directly change a classification model to ensure fairness (in-processing).

This paper is distinguished by its greater emphasis on probabilistic classification, where the outputs of interest are predicted probabilities of belonging to one of the classes, as opposed to binary predictions. The predicted probabilities are often referred to as *scores* and are desirable because they indicate confidences in predictions. We propose an optimization formulation for transforming scores to satisfy fairness constraints while minimizing the loss in utility. The formulation accommodates any fairness criteria that can be expressed as linear inequalities involving conditional means of scores, including variants of statistical parity (SP) (Pedreschi et al., 2012) and equalized odds (EO) (Hardt et al., 2016; Zafar et al., 2017a).

We make the following contributions beyond a novel problem formulation: We derive a closed-form expression for the optimal transformed scores and a convex dual optimization problem for the Lagrange multipliers that parametrize the transformation. In the population limit, the transformed scores minimize cross-entropy with respect to the conditional distribution $p_{Y \mid X}$ of the outcome Y given features X (i.e. the unconstrained optimal score) subject to the fairness constraints. In the finite sample setting, we propose a method called FairScoreTransformer (FST) that uses standard probabilistic classifiers (e.g. logistic regression) to approximate $p_{Y \mid X}$ and the alternating direction method of multipliers (ADMM) to solve the dual problem. The closed-form expression for the transformed scores and the low dimension of the dual problem (a small multiple of the number of protected groups) make FST computationally lightweight.

FST lends itself naturally to post-processing and can also be applied in pre-processing. As such, we envision that FST will be particularly beneficial in situations that make post- and pre-processing attractive (also discussed by Hajian and Domingo-Ferrer (2013); Calmon et al. (2017); Agarwal et al. (2018);

Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

Madras et al. (2018); Salimi et al. (2019)): a) when it is not possible or desirable to modify an existing classifier (post); b) when freedom is desired to select the most suitable classifier for an application (post, pre); and c) when standard training algorithms are used without the additional complexity of accounting for fairness (post, pre). In-processing meta-algorithms (Agarwal et al. 2018; Celis et al. 2019) can also support situation b) but not a) or c). Compared to existing post- and pre-processing methods, FST is considerably more flexible in handling more cases (see Table 1).

Comprehensive experiments compare FST to 10 existing methods, a number that compares favorably to recent meta-studies (Friedler et al., 2019). On score-based metrics such as Brier score and AUC, FST achieves better fairness-utility trade-offs and hence is indeed advantageous when scores are of interest. At the same time, it remains competitive on binary label-based metrics such as accuracy.

In summary, it is shown that FairScoreTransformer enables fairness-ensuring post- and pre-processing that

- is theoretically grounded and optimal in the population limit (Sections 2 and 3),
- is computationally lightweight (Section 4),
- performs favorably compared to the state-of-the-art (Section 5 and Supplementary Material).

1.1 Related Work

Existing post-processing methods take predicted scores as input but most (Kamiran et al., 2012; Fish et al., 2016; Hardt et al., 2016; Chzhen et al. 2019) produce only binary output and not scores. Pleiss et al. (2017) aim to maintain calibrated probability estimates, a requirement that we do not enforce herein. Furthermore, Kamiran et al. (2012); Fish et al. (2016); Hardt et al. (2016); Pleiss et al. (2017) all assume knowledge of the protected attribute at test time. Kamiran et al. (2012); Fish et al. (2016); Jiang et al. (2019) address only SP, Hardt et al. (2016); Pleiss et al. (2017) address disparities in error rates, and Chzhen et al. (2019) address only equal opportunity. Our approach does not have these limitations.

Existing pre-processing methods (Kamiran and Calders, 2012; Hajian and Domingo-Ferrer, 2013; Calmon et al., 2017) only address SP or the related notion of disparate impact (Feldman et al., 2015). Learning representations that are invariant to protected attributes (Zemel et al., 2013; Louizos et al., 2016; Edwards and Storkey, 2016; Xie et al., 2017; Xu et al. 2018) can also be seen as pre-processing, at the cost of losing the original data domain and its semantics. Recent adversarial approaches (Beutel et al., 2017; Zhang et al., 2018; Madras et al., 2018) target EO as well as SP but can be computationally challenging.

Several works have technical similarities but focus on binary outputs with 0-1 risk (Celis et al., 2019; Agarwal et al., 2018) or cost-sensitive risk (Menon and Williamson, 2018; Corbett-Davies et al., 2017) as the objective function. The closest is Celis et al. (2019), which also solves a fairness-constrained classification problem via the dual problem. Celis et al. (2019); Agarwal et al. (2018) propose in-processing algorithms that solve multiple instances of a subproblem whereas we solve only one instance. Menon and Williamson (2018); Corbett-Davies et al. (2017) also characterize optimal fair classifiers in the population limit in which probability distributions are known; however, they do not propose algorithms for computing the solution.

2 PROBLEM FORMULATION

We represent one or more protected attributes by a random variable A and an outcome variable by Y. We make the common assumption that $Y \in \{0, 1\}$ is binary-valued. It is assumed that A takes a finite number of values in a set $\mathcal{A}$, corresponding to protected groups. Let $X \in \mathcal{X}$ denote features used to predict Y in a supervised classification setting. We consider two scenarios in which X either includes or does not include A, like in other works (e.g. Agarwal et al. (2018); Donini et al. (2018)). While the former scenario can achieve better trade-offs between utility and fairness, the latter is needed in applications where disparate treatment laws and regulations forbid the explicit use of A. To develop our approach in this section and Section 3, we work in the population limit and make use of probability distributions involving A, X, Y. Section 4 discusses how these distributions are approximated using a training sample.

As stated earlier, we focus more heavily on probabilistic classification in which the output of interest is the predicted probability of being in the positive class $Y = 1$ rather than a binary prediction. The optimal probabilistic classifier is the conditional probability $r(x) \triangleq p_{Y \mid X}(1 \mid x)$, which we refer to as the *original score*. Bayes-optimal binary classifiers can be derived from $r(x)$ by thresholding.

We propose a mathematical formulation and method called FairScoreTransformer (FST) that can be applied to both post-processing and pre-processing. In both cases, the goal is to transform $r(x)$ into a *transformed score* $r'(x)$ that satisfies fairness conditions while minimizing the loss in optimality compared to $r(x)$. We elaborate on the utility and fairness measures consid-

ered in Sections 2.1 and 2.2. The application of FST to post-processing is straightforward: $r'(x)$ is used directly as the classification output and can be thresholded to provide a binary prediction.

In the pre-processing case, we additionally define a *transformed outcome* $Y' \in \{0,1\}$ and let $r'(x) \triangleq p_{Y'\,|\,X}(1\,|\,x)$ be its conditional probability. The overall procedure consists of two steps, performed in general by two different parties: 1) The *data owner* transforms the outcome variable from Y to Y'; 2) The *modeler* trains a classifier with Y' as target variable and X as input, without regard for fairness. The transformed score $r'(x)$ plays two roles in this procedure. The first is to specify the (randomized) mapping from X to Y' in step 1). We will see that this mapping depends only indirectly on Y through the original score $r(x)$. The second role stems from the main challenge faced by pre-processing methods, namely that the predominant fairness metrics depend on the output of the classifier trained in step 2) but this classifier is not under direct control of the pre-processing. In recognition of this challenge, we make the following assumption, also discussed by Madras et al. (2018); Salimi et al. (2019):

Assumption 1 (pre-processing)**.** The classifier trained by the modeler approximates the transformed score $r'(x)$ if it is a probabilistic classifier or a thresholded version of $r'(x)$ if it is a binary classifier.

Assumption 1 is satisfied for modelers who are indeed learning to predict Y' from X since the optimal classifier in this case is $r'(x)$ or a function thereof. Given the assumption, we will use $r'(x)$ as a surrogate for the actual classifier output. Assumption 1 is not satisfied if the modeler is not competent or, worse, malicious in discriminating against certain protected groups.

2.1 Utility Measure

We propose to measure the loss in optimality, i.e. utility, between the transformed score $r'(x)$ and original score $r(x)$ using the following cross-entropy:

$$\mathbb{E}\big[-\log p_{Y'\,|\,X}(Y\,|\,X)\big] = \mathbb{E}\left[H_b\left(r(X), r'(X)\right)\right], \quad (1)$$

where $H_b(p,q) \triangleq -p\log q - (1-p)\log(1-q)$ is the binary cross-entropy function. The right-hand side of (1) results from expanding the expectation over Y conditioned on X. On the left-hand side, $p_{Y'\,|\,X}$ is used only as notational shorthand in the post-processing case since Y' is not generated.

One way to arrive at (1) is to assume that $r'(x)$, which is the classifier output in the post-processing case and a surrogate thereof in the pre-processing case, is evaluated against the original outcomes $y_1, \ldots, y_n$ in a training set using the cross-entropy a.k.a. log loss. This yields the empirical version of the left-hand side of (1), i.e., $-\frac{1}{n}\sum_{i=1}^{n} \log p_{Y'\,|\,X}(y_i\,|\,x_i)$. The use of log loss is well-motivated by the desire for $r'(x)$ to be close to the true conditional probability $r(x)$.

An equivalent way to motivate (1) in the pre-processing context is to measure the utility loss by the Kullback-Leibler (KL) divergence between the original and transformed distributions $p_{X,Y}$, $p_{X,Y'}$:

$$\begin{aligned} D_{\mathrm{KL}}\big(p_{X,Y}\,\|\,p_{X,Y'}\big) &= \mathbb{E}_{p_{X,Y}}\left[\log\frac{p_{X,Y}}{p_{X,Y'}}\right] \quad (2)\\ &= \mathbb{E}_{p_{X,Y}}[\log p_{Y\,|\,X}] - \mathbb{E}_{p_{X,Y}}[\log p_{Y'\,|\,X}]. \end{aligned}$$

The first term depends on the data distribution but not $r'(x)$ and the second term is exactly (1).

Starting from a different premise, Jiang and Nachum (2019) proposed a similar formulation in which the arguments of the KL divergence are reversed from those in (2). The form of the solution in Jiang and Nachum (2019) is therefore different from the one presented herein. The order of arguments in (2) is justified by the connection to log loss discussed above.

2.2 Fairness Measures

We consider fairness criteria expressible as linear inequalities in conditional means of scores,

$$\sum_{j=1}^{J} b_{lj}\mathbb{E}\big[r'(X)\,|\,\mathcal{E}_{lj}\big] \le c_l, \quad l = 1,\ldots,L, \quad (3)$$

where $\{b_{lj}\}$ and $\{c_l\}$ are real-valued coefficients and the conditioning events $\mathcal{E}_{lj}$ are defined in terms of (A,X,Y) but do not depend on r'. Special cases of (3) correspond to the well-studied notions of statistical parity (SP) and equalized odds (EO). More precisely, we focus on the following variant of SP:

$$-\epsilon \le \mathbb{E}[r'(X)\,|\,A=a] - \mathbb{E}[r'(X)] \le \epsilon \qquad \forall a \in \mathcal{A}, \quad (4)$$

which we refer to as *mean score parity* (MSP) following Coston et al. (2019). Similar notions can also be put in the form of (3), for example bounds on the ratio $\mathbb{E}[r'(X)\,|\,A=a]/\mathbb{E}[r'(X)]$ referred to as *disparate impact* (Feldman et al., 2015).

For EO, we add the condition $Y=y$ to the conditioning events in (4), resulting in

$$\begin{aligned} -\epsilon \le \mathbb{E}[r'(X)\,|\,A=a, Y=y] - \mathbb{E}[r'(X)\,|\,Y=y] \le \epsilon \\ \forall a \in \mathcal{A},\ y \in \{0,1\}. \quad (5) \end{aligned}$$

For $y=0$ (respectively $y=1$), $\mathbb{E}[r'(X)\,|\,Y=y]$ is the false (true) positive rate (FPR, TPR) generalized for a probabilistic classifier, and $\mathbb{E}[r'(X)\,|\,A=a, Y=$

$y]$ is the corresponding group-specific rate. Following Pleiss et al. (2017), we refer to (5) for $y = 0$ or $y = 1$ alone as approximate equality in generalized FPRs or TPRs, and to (5) for $y = 0$ and $y = 1$ together as generalized EO (GEO). The SM specifies the exact correspondences between (4), (5) and (3).

The fairness measures (3) in our formulation are defined in terms of probabilistic scores. Parallel notions defined for binary predictions, i.e. by replacing $r'(X)$ with a thresholded version $\mathbf{1}(r'(X) > t)$, are more common in the literature. For example, the counterpart to (5) is (non-generalized) EO while the counterpart to (4) is called *thresholded score parity* by Coston et al. (2019). While our formulation does not optimize for these binary prediction measures, we nevertheless use them for evaluation in Section 5.

The form of (3) is inspired by but is less general than the constraints of Agarwal et al. (2018), which replace $r'(X)$ in (3) by an arbitrary bounded function $g_j(A, X, Y, r'(X))$. We have restricted ourselves to (3) to derive a closed-form optimal solution in Section 3. We note however that in the examples in Agarwal et al. (2018) and many fairness measures, $g_j(A, X, Y, r'(X)) = r'(X)$ and the additional generality is not required.

2.3 Optimization Problem

The transformed score $r'(x)$ is obtained by minimizing the cross-entropy in (1) (equivalently maximizing its negative) subject to fairness constraints (3):

$$\begin{aligned} \max_{r'} \quad & -\mathbb{E}\left[H_b\left(r(X), r'(X)\right)\right] \\ \text{s.t.} \quad & \sum_{j=1}^{J} b_{lj}\mathbb{E}\left[r'(X) \mid \mathcal{E}_{lj}\right] \le c_l, \quad l = 1, \ldots, L. \end{aligned} \tag{6}$$

The next section characterizes the optimal solution to this problem. In the SM, we elaborate on the fact that when utility and fairness are measured according to the objective and constraints in (6), it suffices to transform scores and not also transform features X into X', as proposed by Hajian and Domingo-Ferrer (2013); Feldman et al. (2015); Calmon et al. (2017).

3 CHARACTERIZATION OF OPTIMAL FAIR SCORE

We derive a closed-form expression for the optimal solution to (6) using the method of Lagrange multipliers. We then state the dual optimization problem that determines the Lagrange multipliers. These results are specialized to the cases of MSP (4) and GEO (5).

Define Lagrange multipliers $\lambda_l \ge 0$, $l = 1, \ldots, L$ for the constraints in (6), and let $\lambda \triangleq (\lambda_1, \ldots, \lambda_L)$. Then the Lagrangian function is given by

$$\begin{aligned} L(r', \lambda) = & -\mathbb{E}\left[H_b\left(r(X), r'(X)\right)\right] \\ & - \sum_{l=1}^{L}\sum_{j=1}^{J} \lambda_l b_{lj}\mathbb{E}\left[r'(X) \mid \mathcal{E}_{lj}\right] + \sum_{l=1}^{L} c_l\lambda_l. \end{aligned} \tag{7}$$

The dual optimization problem corresponding to (6) is $\min_{\lambda\ge 0}\max_{r'} L(r', \lambda)$.

Note that $L(r', \lambda)$ is a strictly concave function of r' and the fairness constraints in (6) are affine functions of r'. Consequently, as long as the constraints in (6) are feasible, the optimal transformed score r^* can be found by (i) maximizing $L(r', \lambda)$ with respect to r', resulting in an optimal solution r^* that is a function of λ, and then (ii) minimizing $L(r^*, \lambda)$ with respect to λ (Boyd and Vandenberghe, 2004, Section 5.5.5). Substituting the optimal λ^* into the solution for r^* found in the first step then yields the optimal transformed score. Note that this procedure would not necessarily be correct if a linear objective function were considered (e.g., 0-1 loss in Celis et al. (2019)) due to lack of strict concavity. The next proposition states the general form of the solution to the inner maximization (i) above. Its proof is in the SM.

Proposition 1. Let $L(r', \lambda)$ be as given in (7). Then for fixed λ, $r^*(\lambda) = \arg\max_{r'} L(r', \lambda)$ is given by

$$r^*\big(\mu(x); r(x)\big) = \tag{8}$$
$$\begin{cases} \dfrac{1 + \mu(x) - \sqrt{(1+\mu(x))^2 - 4r(x)\mu(x)}}{2\mu(x)}, & \mu(x) \ne 0 \\ r(x), & \mu(x) = 0, \end{cases}$$

$$\text{where} \quad \mu(x) \triangleq \sum_{l=1}^{L}\sum_{j=1}^{J} \lambda_l b_{lj} \frac{\Pr(\mathcal{E}_{lj} \mid X = x)}{\Pr(\mathcal{E}_{lj})}. \tag{9}$$

We can interpret the optimal primal solution (8) as a prescription for *score transformation* controlled by $\mu(x)$, which is in turn a linear function of λ. When $\mu(x) = 0$, the score is unchanged from $r(x)$, and as $\mu(x)$ increases or decreases from zero, the score $r^*(\mu(x); r(x))$ decreases or increases smoothly from $r(x)$ (as can be seen by plotting the function). It can also be shown by differentiating r^* with respect to r that r^* has a rank-preserving property for fixed μ in the sense that if $r_1 < r_2$ then $r^*(\mu; r_1) < r^*(\mu; r_2)$.

It is shown in the proof of Proposition 1 that the result of substituting the optimal primal solution (8) into the first two terms of the Lagrangian (7) is the expectation of the function $g(\mu(x); r(x)) \triangleq -H_b\big(r(x), r^*(\mu(x); r(x))\big) - \mu(x)r^*(\mu(x); r(x))$. The

dual problem is therefore

$$\begin{aligned} \min_{\lambda \geq 0} \quad & \mathbb{E}\left[g\big(\mu(X); r(X)\big)\right] + \sum_{l=1}^{L} c_l \lambda_l \\ \text{s.t.} \quad & \mu(X) = \sum_{l=1}^{L} \sum_{j=1}^{J} \lambda_l b_{lj} \frac{\Pr(\mathcal{E}_{lj} \mid X)}{\Pr(\mathcal{E}_{lj})}. \end{aligned} \tag{10}$$

The solution to (10) provides the values of λ^* for the optimal transformed score (8). Like all Lagrangian duals, (10) is a convex optimization (although it is no longer apparent from (10) that this is the case). Furthermore, (10) is typically low-dimensional in cases where the number of dual variables L is a small multiple of the number of protected groups $|\mathcal{A}|$.

We now specialize and simplify (10) for MSP (4) and GEO (5) constraints, utilizing their correspondences with (3) as shown in the SM.

Proposition 2. Under the MSP constraint (4), the dual optimization (10) reduces to

$$\begin{aligned} \min_{\lambda} \quad & \mathbb{E}\left[g\big(\mu(X); r(X)\big)\right] + \epsilon \|\lambda\|_1 \\ \text{s.t.} \quad & \mu(X) = \sum_{a \in \mathcal{A}} \lambda_a \left(\frac{p_{A \mid X}(a \mid X)}{p_A(a)} - 1 \right). \end{aligned} \tag{11}$$

For the GEO constraint (5), (10) reduces to

$$\begin{aligned} \min_{\lambda} \quad & \mathbb{E}\left[g\big(\mu(X); r(X)\big)\right] + \epsilon \|\lambda\|_1, \\ \text{s.t.} \quad & \mu(X) = \sum_{y \in \{0,1\}} \frac{p_{Y \mid X}(y \mid X)}{p_Y(y)} \times \\ & \qquad \sum_{a \in \mathcal{A}} \lambda_{a,y} \left(\frac{p_{A \mid X,Y}(a \mid X, y)}{p_{A \mid Y}(a \mid y)} - 1 \right). \end{aligned} \tag{12}$$

In (11), (12), there is no longer a non-negativity constraint on λ but instead an ℓ_1 norm, and the problem dimension is only $|\mathcal{A}|$ in (11) and $2|\mathcal{A}|$ in (12). Moreover, both dual formulations are well-suited for decomposition using the alternating direction method of multipliers (ADMM), as discussed in Section 4.2. In the case where X includes A, the constraints in (11) and (12) simplify as shown in the proof of Proposition 2 and, importantly, eliminate the need to estimate A.

4 FairScoreTransformer PROCEDURE

We now consider the finite sample setting in which the probability distributions of A, X, Y are not known and we have instead a training set $\mathcal{D}_n \triangleq \{(a_i, x_i, y_i), i = 1, \ldots, n\}$. This section presents the proposed FairScoreTransformer (FST) procedure that approximates the optimal fairness-constrained score in Section 3. We focus on the cases of MSP and GEO.

The procedure consists of the following steps: 1) Estimate the original score and other probabilities required to define the dual problem (11) or (12); 2) Solve the dual problem to obtain dual variables λ^* (the "fit" step); 3) Transform scores using (8) and (9) ("transform" step); 4) For pre-processing, modify the training data; 5) For binary-valued predictions, binarize scores. The following subsections elaborate on steps 1), 2), and 4). Step 5) is done simply by selecting a threshold $t \in [0, 1]$ to maximize accuracy on the training set.

4.1 Estimation of Original Score and Other Probabilities

In some post-processing applications, original scores $r(x)$ may already be estimated by an existing base classifier. If no suitable base classifier exists, any probabilistic classification algorithm may be used to estimate $r(x)$. We experiment with logistic regression and gradient boosting machines in Section 5. We naturally recommend selecting a model and any hyperparameter values to maximize performance in this regard, i.e. to yield accurate and calibrated probabilities.

In the case where A is one of the features in X, the other probabilities required are $p_A(a)$ for MSP (11) and $p_Y(y)$, $p_{A \mid Y}(a \mid y)$ for GEO (12) ($p_{Y \mid X}(y \mid x)$ is already given by $r(x)$ and $p_{A \mid X}$, $p_{A \mid X,Y}$ are delta functions). Since Y is binary and $|\mathcal{A}|$ is typically small, it suffices to use the empirical probabilities. If A is not included in X, then it is also necessary to estimate it using $p_{A \mid X}(a \mid X)$ for MSP (11) and $p_{A \mid X,Y}(a \mid X, y)$ for GEO (12). Again, any probabilistic classification algorithm can be used, provided that it can handle more than two classes if $|\mathcal{A}| > 2$. We highlight that FST translates the effort of ensuring fair classification into training well-calibrated models for predicting Y and, if necessary, A. This echoes the plug-in approach advocated by Menon and Williamson (2018).

4.2 ADMM for Optimizing Dual Variables

Both optimizations in Proposition 2 are of the form

$$\begin{aligned} \min_{\lambda \in \mathbb{R}^d} \quad & \frac{1}{n} \sum_{i=1}^{n} g\big(\mu(x_i); r(x_i)\big) + \epsilon \|\lambda\|_1 \\ \text{s.t.} \quad & \mu(x_i) = \lambda^T \mathbf{f}(x_i), \quad i = 1, \ldots, n, \end{aligned} \tag{13}$$

where (i) we approximate the expectation in the objective by the average over the training dataset, (ii) d is the dimension of λ, and (iii) $\mathbf{f} : \mathcal{X} \to \mathbb{R}^d$ is defined by the expression for $\mu(x)$ in (11) or (12) and uses the probabilities estimated in Section 4.1.

Formulation (13) is well-suited for ADMM because the objective function is separable between $\mu(x)$ and λ, which are linearly related through the constraint. We

present one ADMM decomposition here and alternatives in the SM. Under the first decomposition, application of the scaled ADMM algorithm (Boyd et al. 2011, Section 3.1.1) to (13) yields the following three steps in each iteration $k = 0, 1, \ldots$:

$$\mu^{(k+1)}(x_i) = \arg\min_{\mu} \frac{1}{n} g\big(\mu; r(x_i)\big) + \frac{\rho}{2} \times \Big(\mu - (\lambda^{(k)})^T \mathbf{f}(x_i) + c^{(k)}(x_i)\Big)^2 \quad \forall i = 1, \ldots, n \quad (14)$$

$$\lambda^{(k+1)} = \arg\min_{\lambda} \epsilon\|\lambda\|_1 + \frac{\rho}{2} \times \sum_{i=1}^{n} \Big(\mu^{(k+1)}(x_i) - \lambda^T \mathbf{f}(x_i) + c^{(k)}(x_i)\Big)^2 \quad (15)$$

$$c^{(k+1)}(x_i) = c^{(k)}(x_i) + \mu^{(k+1)}(x_i) - \Big(\lambda^{(k+1)}\Big)^T \mathbf{f}(x_i) \quad \forall i = 1, \ldots, n. \quad (16)$$

The first update (14) can be computed in parallel for each sample x_i in the dataset. Given an x_i, finding $\mu(x_i)$ is a single-parameter optimization where the objective possesses closed-form derivatives, provided in the SM. The second update (15) reduces to an ℓ_1-penalized quadratic minimization over (at most) $2|A|$ variables. Details on this reduction are also in the SM.

From (14)–(16), it is seen that the complexity of each ADMM iteration is linear in n. We have fixed the ADMM penalty parameter $\rho = 1$ and have not attempted to tune it for faster convergence.

4.3 Additional Steps for Pre-Processing

In pre-processing, the transformed score $r'(x)$ is used to generate samples of a transformed outcome Y'. Since $r'(x) = p_{Y'\,|\,X}(1 \mid x)$ is a probabilistic mapping, we propose generating a *weighted* dataset $\mathcal{D}' = \{(x_i, y'_i, w_i)\}$ with weights w_i that reflect $p_{Y'\,|\,X}$. Specifically, $\mathcal{D}' = \mathcal{D}'_0 \cup \mathcal{D}'_1$ with $\mathcal{D}'_0 = \{(x_i, 0, 1 - r'(x_i)), i = 1, \ldots, n\}$ and $\mathcal{D}'_1 = \{(x_i, 1, r'(x_i)), i = 1, \ldots, n\}$ so that $\mathcal{D}'$ has $2n$ samples. The data owner passes $\mathcal{D}'$ to the modeler, who uses it to train a classifier for Y' without fairness constraints.

5 EMPIRICAL EVALUATION

This section discusses experimental evaluation of the proposed FST methods for MSP and GEO constraints.

Datasets Four datasets were used, the first three of which are standard in the fairness literature: 1) adult income, 2) ProPublica's COMPAS recidivism, 3) German credit risk, 4) Medical Expenditure Panel Survey (MEPS). We used versions pre-processed by AI Fairness 360 (Bellamy et al., 2019). To facilitate comparison with other methods, we used binary-valued protected attributes and consider gender and race for both adult and COMPAS, age and gender for German, and race for MEPS. Each dataset was randomly split 10 times into training (75%) and test (25%) sets.

Methods Compared Since FST is intended for post- and pre-processing, comparisons to other such methods are most natural as they accommodate situations a)–c) in Section 1. For post-processing, we have chosen Hardt et al. (2016) (HPS) and the reject option method of Kamiran et al. (2012), both as implemented in AI Fairness 360, as well as the Wass-1 Post-Process $\hat{p}_S$ method (WPP) of Jiang et al. (2019). For pre-processing, the massaging and reweighing methods of Kamiran and Calders (2012) and the optimization method (OPP) of Calmon et al. (2017) were chosen. Among in-processing methods, meta-algorithms that work with essentially any base classifier can handle situation b). The reductions method ('red') (Agarwal et al. 2018) was selected from this class. We also compared to in-processing methods specific to certain base classifiers, thus precluding any of a)–c): fairness constraints (FC) (Zafar et al. 2017b), disparate mistreatment (DM) (Zafar et al. 2017a), and fair empirical risk minimization (FERM) (Donini et al., 2018).

The methods in the previous paragraph have various limitations summarized by Table 1. In particular, the three post-processing methods require knowledge of the protected attribute A at test time. Accordingly, the experiments presented in this section include A in the features X to make it available to all methods; experiments without A at test time are in the SM. We also encountered computational problems with OPP and DM and thus perform separate comparisons with FST on reduced feature sets, also reported in the SM.

Three versions of FST were evaluated: post-processing (FSTpost), pre-processing (FSTpre), and a second post-processing version (FSTbatch) that assumes that test instances can be processed in a batch rather than one by one. In this case, the fit step (Section 4.2) can actually be done on test data since it does not depend on labels y_i (and uses only predicted probabilities for A if A is unavailable).

Base Classifiers We used ℓ_1-regularized logistic regression (LR) and gradient boosted classification trees (GBM) from scikit-learn (Pedregosa et al., 2011). Post-processing methods operate on the scores produced by the base classifier, pre-processing methods train the base classifier after modifying the training data, and the reductions method repeatedly calls the base classification algorithm. In the SM, we used linear SVMs (with Platt scaling (Platt, 1999) to output prob-

Table 1: Capabilities of Methods in Comparison ($\star$ refers to an extension implemented in Bellamy et al. (2019))

method	pre	in	post	SP	EO	no A at test time	scores	approx fairness	any classifier
massage, reweigh	✓			✓		✓	✓		✓
OPP	✓			✓		✓	✓	✓	✓
HPS			✓		✓				✓
reject			✓	✓	$\star$			✓	✓
WPP			✓	✓			✓		✓
FC		✓		✓		✓	✓	✓	
DM		✓			✓	✓	✓	✓	
FERM		✓			✓	✓		✓	
reductions		✓		✓	✓	✓	✓	✓	✓
proposed FST	✓		✓	✓	✓	✓	✓	✓	✓

abilities) to compare with FERM. We found it impractical to train nonlinear SVMs on the larger datasets for reductions and FERM since reductions needs to do so repeatedly and FERM uses a slower specialized algorithm. For a similar reason, 5-fold cross-validation to select parameters for LR (regularization parameter C from $[10^{-4}, 10^{4}]$) and GBM (min samples/leaf from $\{5, 10, 15, 20, 30\}$) was done only once per training set. All other parameters were set to defaults. The base classifier was then instantiated with the best parameter value for use by all methods.

Results Figure 1 shows the trade-offs between classification performance and fairness obtained in a subset of the experiments. The full set with other dataset-protected attribute combinations, etc. is in the SM. Each dataset occupies two rows with the first showing score-based measures (Brier score vs. differences in mean scores (MSP) or GEO, AUC is in the SM) and the second showing binary label-based measures (accuracy vs. differences in mean binary predictions (SP) or non-generalized EO). The columns correspond to combinations of base classifier (LR, GBM) and fairness measure targeted (SP, EO). Markers indicate mean values over the 10 splits, error bars indicate standard errors in the means, and Pareto-optimal points have been connected to ease visualization.

Considering first the score-based plots (odd rows), FSTpost and FSTbatch achieve trade-offs that are at least as good as all other methods, with the slight exception of the GBM case on MEPS. In all cases, the advantage of FST lies in extending the Pareto frontiers farther to the left, attaining smaller MSP or GEO differences; this is especially apparent for GEO. FSTpre sometimes performs less well, e.g. with GBM on adult and MEPS, likely due to the loss incurred in approximating the transformed score $r'(x)$ with the output of a classifier fit to the pre-processed data.

Turning to the binary label-based plots (even rows), the trade-offs for FSTpost and FSTbatch generally coincide with or are close to the trade-offs of the best method, and are even sometimes the best, despite not optimizing for binary metrics beyond tuning the binarization threshold for accuracy. Again FSTpre with GBM is worse on adult, but FSTpre with LR is the best performer on COMPAS. The main disadvantage of FST is that its trade-off curves may not extend as far to the left as other methods, in particular on adult. This is the converse of its advantage for score-based metrics. We discuss other limitations in the SM.

Among the existing methods, reductions is the strongest and also the most versatile, handling all cases that FST does. However, it is an in-processing method and far more computationally expensive, requiring an average of nearly 30 calls to the base classification algorithm compared to one for FSTpost, FSTbatch and two for FSTpre. Reductions also returns a randomized classifier, which may be undesirable in some applications. Reject option and HPS do not output scores and hence are omitted from the score-based plots. Reject option performs close to the best except on MEPS and at small unfairness values. HPS is limited to EO, does not have a parameter to vary the trade-off, and is less competitive. Massaging, reweighing, and WPP likewise do not have a trade-off parameter and are limited to SP. As also observed by Agarwal et al. (2018), massaging is often dominated by other methods while reweighing lies on the Pareto frontier but with substantial disparity. WPP results in low disparity but its Brier score or accuracy is sometimes less competitive. FC applies only to the LR-SP column and could not substantially reduce unfairness, possibly due to the larger feature dimension.

6 CONCLUSION

We proposed FairScoreTransformer for transforming scores to satisfy fairness constraints and optimize cross-entropy. FST is theoretically optimal in the population limit, has a computationally attractive implementation, and allows flexibility as a post- and pre-processing method. Via a comprehensive set of exper-

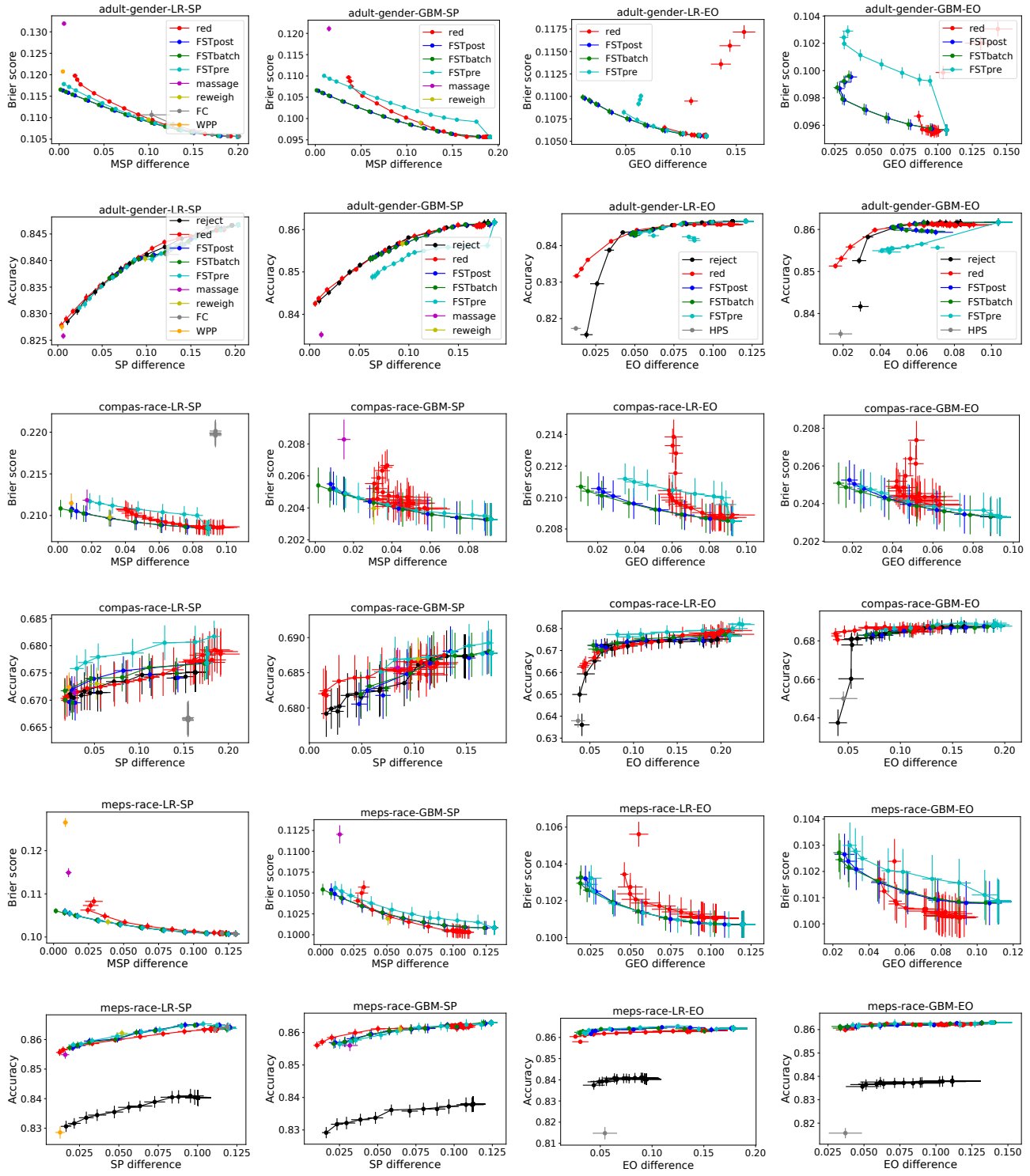

Figure 1: Trade-Offs Between Fairness and Classification Performance (see SM for full set; in each column of plots, legend in first row applies to all odd rows within that column, legend in second row applies to even rows)

iments, we demonstrated that FST is either as competitive or outperforms 10 existing fairness interventions over a range of constraints and datasets. Future directions include characterizing convergence rates for the ADMM iterations and adapting FST to fairness criteria that are not based on conditional expectations of scores (e.g., calibration across groups (Pleiss et al. 2017)) as well as to non-binary outcomes Y.

Acknowledgements

This work was supported in part by IBM Open Collaborative Research Award W1771055. F. P. Calmon was partly supported by the National Science Foundation under Grant No. CIF-1845852.

References

A. Agarwal, A. Beygelzimer, M. Dudík, J. Langford, and H. Wallach. A reductions approach to fair classification. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 60–69, July 2018.

R. K. E. Bellamy, K. Dey, M. Hind, S. C. Hoffman, S. Houde, K. Kannan, P. Lohia, J. Martino, S. Mehta, A. Mojsilović, S. Nagar, K. N. Ramamurthy, J. Richards, D. Saha, P. Sattigeri, M. Singh, K. R. Varshney, and Y. Zhang. AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias. *IBM Journal of Research and Development*, 63(4/5):4:1–4:15, Sept. 2019. doi: 10.1147/JRD.2019.2942287.

A. Beutel, J. Chen, Z. Zhao, and E. H. Chi. Data decisions and theoretical implications when adversarially learning fair representations. In *Workshop on Fairness, Accountability, and Transparency in Machine Learning (FATML)*, pages 1–5, Aug. 2017.

S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, Cambridge, UK, 2004.

S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine learning*, 3(1):1–122, 2011.

F. Calmon, D. Wei, B. Vinzamuri, K. N. Ramamurthy, and K. R. Varshney. Optimized pre-processing for discrimination prevention. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3992–4001. Dec. 2017.

L. E. Celis, L. Huang, V. Keswani, and N. K. Vishnoi. Classification with fairness constraints: A meta-algorithm with provable guarantees. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 319–328, Jan. 2019. doi: 10.1145/3287560.3287586.

E. Chzhen, C. Denis, M. Hebiri, L. Oneto, and M. Pontil. Leveraging labeled and unlabeled data for consistent fair binary classification. In *Advances in Neural Information Processing Systems (NeurIPS)*, Dec. 2019.

S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 797–806, Aug. 2017. doi: 10.1145/3097983.3098095.

A. Coston, K. N. Ramamurthy, D. Wei, K. R. Varshney, S. Speakman, Z. Mustahsan, and S. Chakraborty. Fair transfer learning with missing protected attributes. In *Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*, pages 1–8, Jan. 2019.

M. Donini, L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil. Empirical risk minimization under fairness constraints. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2791–2801. Dec. 2018.

H. Edwards and A. Storkey. Censoring representations with an adversary. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pages 1–14, May 2016.

M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21st ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 259–268, Aug. 2015. doi: 10.1145/2783258.2783311.

B. Fish, J. Kun, and Ádám D. Lelkes. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining (SDM)*, pages 144–152, May 2016.

S. A. Friedler, C. Scheidegger, S. Venkatasubramanian, S. Choudhary, E. P. Hamilton, and D. Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 329–338. ACM, 2019.

S. Hajian and J. Domingo-Ferrer. A methodology for direct and indirect discrimination prevention in data mining. *IEEE Transactions on Knowledge and Data Engineering*, 25(7):1445–1459, July 2013. doi: 10.1109/TKDE.2012.72.

M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3323–3331, Dec. 2016.

H. Jiang and O. Nachum. Identifying and correcting label bias in machine learning. arXiv e-print `https://arxiv.org/abs/1901.04966`, Jan. 2019.

R. Jiang, A. Pacchiano, T. Stepleton, H. Jiang, and S. Chiappa. Wasserstein fair classification. In *Conference on Uncertainty in Artificial Intelligence (UAI)*, pages 1–10, July 2019.

F. Kamiran and T. Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, Oct. 2012. doi: 10.1007/s10115-011-0463-8.

F. Kamiran, A. Karim, and X. Zhang. Decision theory for discrimination-aware classification. In *IEEE 12th International Conference on Data Mining (ICDM)*, pages 924–929, Dec 2012. doi: 10.1109/ICDM.2012.45.

C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel. The variational fair encoder. In *Proceedings of the International Conference on Learning Representations (ICLR)*, pages 1–11, May 2016.

D. Madras, E. Creager, T. Pitassi, and R. Zemel. Learning adversarially fair and transferable representations. In *Proceedings of the 35th International Conference on Machine Learning (ICML)*, pages 3384–3393, July 2018.

A. K. Menon and R. C. Williamson. The cost of fairness in binary classification. In *Proceedings of the 1st Conference on Fairness, Accountability and Transparency (FAT*)*, pages 107–118, Feb. 2018.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

D. Pedreschi, S. Ruggieri, and F. Turini. A study of top-k measures for discrimination discovery. In *Proceedings of the 27th Annual ACM Symposium on Applied Computing*, pages 126–131. ACM, 2012.

J. C. Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in Large Margin Classifiers*, pages 61–74, 1999.

G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger. On fairness and calibration. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 5680–5689. Dec. 2017.

B. Salimi, L. Rodriguez, B. Howe, and D. Suciu. Interventional fairness: Causal database repair for algorithmic fairness. In *Proceedings of the ACM International Conference on Management of Data (SIGMOD)*, pages 793–810, June 2019. doi: 10.1145/3299869.3319901.

Q. Xie, Z. Dai, Y. Du, E. Hovy, and G. Neubig. Controllable invariance through adversarial feature learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 585–596. Dec. 2017.

D. Xu, S. Yuan, L. Zhang, and X. Wu. FairGAN: Fairness-aware generative adversarial networks. In *IEEE International Conference on Big Data (Big Data)*, pages 570–575, Dec. 2018. doi: 10.1109/BigData.2018.8622525.

M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web (WWW)*, pages 1171–1180. International World Wide Web Conferences Steering Committee, 2017a.

M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi. Fairness Constraints: Mechanisms for Fair Classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 962–970, Apr. 2017b.

R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *Proceedings of the 30th International Conference on Machine Learning (ICML)*, pages 325–333, Atlanta, Georgia, USA, June 2013.

B. H. Zhang, B. Lemoine, and M. Mitchell. Mitigating unwanted biases with adversarial learning. In *Proceedings of the AAAI/ACM Conference on Artificial Intelligence, Ethics, and Society (AIES)*, pages 335–340, Feb. 2018. doi: 10.1145/3278721.3278779.