Two Simple Ways to Learn Individual Fairness Metrics from Data

Debarghya Mukherjee [1 *] **Mikhail Yurochkin** [2 *] **Moulinath Banerjee** [1] **Yuekai Sun** [1]

Abstract

Individual fairness is an intuitive definition of algorithmic fairness that addresses some of the drawbacks of group fairness. Despite its benefits, it depends on a task specific fair metric that encodes our intuition of what is fair and unfair for the ML task at hand, and the lack of a widely accepted fair metric for many ML tasks is the main barrier to broader adoption of individual fairness. In this paper, we present two simple ways to learn fair metrics from a variety of data types. We show empirically that fair training with the learned metrics leads to improved fairness on three machine learning tasks susceptible to gender and racial biases.[1] We also provide theoretical guarantees on the statistical performance of both approaches.

1. Introduction

Machine learning (ML) models are an integral part of modern decision-making pipelines. They are even part of some high-stakes decision support systems in criminal justice, lending, medicine *etc.*. Although replacing humans with ML models in the decision-making process appear to eliminate human biases, there is growing concern about ML models reproducing historical biases against certain historically disadvantaged groups. This concern is not unfounded. For example, Dastin (2018) reports gender-bias in Amazon's resume screening tool, Angwin et al. (2016) mentions racial bias in recidivism prediction instruments, Vigdor (2019) reports gender bias in the credit limits of Apple Card.

As a first step towards mitigating algorithmic bias in ML models, researchers proposed a myriad of formal definitions of algorithmic fairness. At a high-level, there are two groups of mathematical definitions of algorithmic fairness: group fairness and individual fairness. Group fairness divides the feature space into (non-overlapping) protected subsets and imposes invariance of the ML model on the subsets. Most prior work focuses on group fairness because it is amenable to statistical analysis. Despite its prevalence, group fairness suffers from two critical issues. First, it is possible for an ML model that satisfies group fairness to be blatantly unfair with respect to subgroups of the protected groups and individuals (Dwork et al., 2011). Second, there are fundamental incompatibilities between seemingly intuitive notions of group fairness (Kleinberg et al., 2016; Chouldechova, 2017).

In light of the issues with group fairness, we consider individual fairness in our work. Intuitively, individually fair ML models should treat similar users similarly. Dwork et al. (2011) formalize this intuition by viewing ML models as maps between input and output metric spaces and defining individual fairness as Lipschitz continuity of ML models. The metric on the input space is the crux of the definition because it encodes our intuition of which users are similar. Unfortunately, individual fairness was dismissed as impractical because there is no widely accepted similarity metric for most ML tasks. In this paper, we take a step towards operationalizing individual fairness by showing it is possible to learn good similarity metrics from data.

The rest of the paper is organized as follows. In Section 2, we describe two different ways to learn data-driven fair metric: one from knowledge of groups of similar inputs and another from knowledge of similar and dissimilar pairs of inputs. In Section 3, we show that (i) the methods are robust to noise in the data, and (ii) the methods leads to individually fair ML models. Finally, in Section 4, we demonstrate the effectiveness of the methods in mitigating bias on two ML tasks susceptible to gender and racial biases.

2. Learning fair metrics from data

The intuition underlying individual fairness is fair ML models should treat *comparable* users similarly. We write *comparable* instead of *similar* in the rest of this paper to emphasize that comparable samples may differ in ways that are irrelevant to the task at hand. Formally, we consider an ML model as a map $h : \mathcal{X} \to \mathcal{Y}$, where $(\mathcal{X}, d_x)$ and $(\mathcal{Y}, d_y)$ are the input and output metric spaces respectively. Individual fairness (Dwork et al., 2011; Friedler et al., 2016)

[*]Equal contribution [1]Department of Statistics, University of Michigan [2]IBM Research, MIT-IBM Watson AI Lab. Correspondence to: Debarghya Mukherjee <mdeb@umich.edu>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020.

[1]Codes are available at https://github.com/mdebumich/Fair_metric_learning.

is L-Lipschitz continuity of h:

$$d_y(h(x_1), h(x_2)) \leq L d_x(x_1, x_2) \text{ for all } x_1, x_2 \in \mathcal{X}. \tag{2.1}$$

The choice of $d_{\mathcal{Y}}$ depends on the form of the output. For example, if the ML model outputs a vector of the logits, then we may pick the Euclidean norm as $d_{\mathcal{Y}}$ (Kannan et al., 2018; Garg et al., 2018). The fair metric d_x is the crux of the definition. It encodes our intuition of which samples are comparable; *i.e.* which samples only differ in ways that are irrelevant to the task at hand. Originally, Dwork et al. (2011) deferred the choice of d_x to regulatory bodies or civil rights organizations, but we are unaware of widely accepted fair metrics for most ML tasks. This lack of widely accepted fair metrics has led practitioners to dismiss individual fairness as impractical. Our goal here is to address this issue by describing two ways to learn fair metrics from data.

We start from the premise there is generally more agreement than disagreement about what is fair in many application areas. For example, in natural language processing, there are ways of identifying groups of training examples that should be treated similarly (Bolukbasi et al., 2016; Madaan et al., 2018) or augmenting the training set with hand-crafted examples that should be treated similarly as observed training examples (Garg et al., 2019). Even in areas where humans disagree, there are attempts to summarize the cases on which humans agree in metrics by fitting metrics to human feedback (Wang et al., 2019). Our goal is similar: encode what we agree on in a metric, so that we can at least mitigate the biases that we agree on with methods for enforcing individual fairness (Kim et al., 2018; Rothblum & Yona, 2018; Yurochkin et al., 2020; Yurochkin & Sun, 2020).

To keep things simple, we focus on fitting metrics of the form

$$d_x(x_1, x_2) \triangleq \langle \varphi(x_1) - \varphi(x_2), \Sigma(\varphi(x_1) - \varphi(x_2)) \rangle, \tag{2.2}$$

where $\varphi(x) : \mathcal{X} \to \mathbf{R}^d$ is an embedding map and $\Sigma \in \mathbf{S}_+^d$. The reason behind choosing Mahalanobis distance is that the learned feature maps (*e.g.* the activations of the penultimate layer of a deep neural network) typically map non-linear structures in the raw feature space to linear structures in the learned feature space (Mikolov et al., 2013; Radford et al., 2015; Brock et al., 2018). To keep things simple, we assume φ is known and learn the matrix Σ from the embedded observations $\varphi's$. The data may consist of human feedback, hand-picked groups of similar training examples, hand-crafted examples that should be treated similarly as observed training examples, or a combination of the above. In this section, we describe two simple methods for learning fair metrics from diverse data types.

2.1. FACE: Factor Analysis of Comparable Embeddings

In this section, we consider learning Σ from groups of comparable samples. The groups may consist of hand-picked training examples (Bolukbasi et al., 2016; Madaan et al., 2018) or hand-crafted examples that differ in certain "sensitive" ways from observed training examples (Garg et al., 2019).

To motivate the approach, we posit the embedded features satisfy a factor model:

$$\varphi_i = A_* u_i + B_* v_i + \epsilon_i \tag{2.3}$$

where $\varphi_i \in \mathbf{R}^d$ is the learned representation of x_i, $u_i \in \mathbf{R}^K$ (resp. $v_i \in \mathbf{R}^L$) is the protected/sensitive (resp. discriminative/relevant) attributes of x_i for the ML task at hand, and ϵ_i is an error term. A pair of samples are comparable if their (unobserved) relevant attributes are similar. For example, Bolukbasi et al.'s method for mitigating gender bias in word embeddings relies on word pairs that only differ in their gender associations (*e.g.* (he, she), (man, woman), (king, queen) *etc.*).

The factor model (2.3) decomposes the variance of the learned representations into variance due to the sensitive attributes and variance due to the relevant attributes. We wish to learn a metric that ignores the variance attributed to the sensitive attributes but remains sensitive to the variance attributed to the relevant attributes. This way, the metric declares any pair of samples that differ mainly in their sensitive attributes as comparable. One possible choice of Σ is the projection matrix onto the orthogonal complement of $\mathsf{ran}(A_*)$, where $\mathsf{ran}(A_*)$ is the column space of A_*. Indeed,

$$\begin{aligned} d_x(x_1, x_2) &= \langle \varphi_1 - \varphi_2, (I - P_{\mathsf{ran}(A_*)})(\varphi_1 - \varphi_2) \rangle \\ &\approx \langle B_*(v_1 - v_2), (I - P_{\mathsf{ran}(A_*)}) B_*(v_1 - v_2) \rangle, \end{aligned}$$

which ignores differences between φ_1 and φ_2 due to differences in the sensitive attributes. Although $\mathsf{ran}(A_*)$ is unknown, it is possible to estimate it from the learned representations and groups of comparable samples by factor analysis (see Algorithm 1). We remark that our target is $\mathsf{ran}(A_*)$, not A_* itself. This frees us from cumbersome identification restrictions common in the factor analysis literature.

Algorithm 1 estimating $\mathsf{ran}(A_*)$ by factor analysis

1: **Input:** $\{\varphi_i\}_{i=1}^n$, comparable groups $\mathcal{I}_1, \ldots, \mathcal{I}_G$
2: $\widehat{A}^T \in \arg\min_{W_g, A} \{\frac{1}{2} \sum_{g=1}^G \|H_g \Phi_{\mathcal{I}_g} - W_g A^T\|_F^2\}$, where $H_g \triangleq I_{|\mathcal{I}_g|} - \frac{1}{|\mathcal{I}_g|} 1_{|\mathcal{I}_g|} 1_{|\mathcal{I}_g|}^T$ is the centering matrix
3: $Q \leftarrow \mathtt{qr}(\widehat{A})$ {get orthonormal basis of $\mathsf{ran}(\widehat{A})$}
4: $\widehat{\Sigma} \leftarrow I_d - QQ^T$

Algorithm 1 is based on the observation that groups of comparable samples have similar relevant attributes; *i.e.*

$$\begin{aligned} H\Phi_{\mathcal{I}} &= HU_{\mathcal{I}}A_*^T + \cancelto{\approx 0}{HV_{\mathcal{I}}B_*^T} + HE_{\mathcal{I}} \\ &\approx HU_{\mathcal{I}}A_*^T + HE_{\mathcal{I}}, \end{aligned} \tag{2.4}$$

where $H \triangleq I_{|\mathcal{I}|} - \frac{1}{|\mathcal{I}|}1_{|\mathcal{I}|}1_{|\mathcal{I}|}^T$ is the centering matrix and $\Phi_{\mathcal{I}}$ (resp. $U_{\mathcal{I}}$, $V_{\mathcal{I}}$) is the matrix whose rows are the φ_i's (resp. u_i's, v_i's). This is the factor model that Algorithm 1 fits in Step 2 to obtain $\widehat{A}$ whose range is close to that of $\widehat{A}$. In Steps 3 and 4, the algorithm forms the projector onto the orthogonal complement of $\mathsf{ran}(\widehat{A})$.

2.2. EXPLORE: Embedded Xenial Pairs Logistic Regression

EXPLORE learns a fair metric from pair-wise comparisons. More concretely, the data comes from human feedback in the form of triplets $\{(x_{i_1}, x_{i_2}, y_i)\}_{i=1}^n$, where $y_i \in \{0, 1\}$ indicates whether the human considers x_{i_1} and x_{i_2} comparable ($y_i = 1$ indicates comparable). We posit (x_{i_1}, x_{i_2}, y_i) satisfies a binary response model.

$$\begin{aligned} y_i \mid x_{i_1}, x_{i_2} &\sim \mathsf{Ber}(2\sigma(-d_i)), \\ d_i &\triangleq \|\varphi_{i_1} - \varphi_{i_2}\|_{\Sigma_0}^2 \\ &= (\varphi_{i_1} - \varphi_{i_2})^T\Sigma_0(\varphi_{i_1} - \varphi_{i_2}) \\ &= \langle \underbrace{(\varphi_{i_1} - \varphi_{i_2})(\varphi_{i_1} - \varphi_{i_2})^T}_{D_i}, \Sigma_0\rangle \end{aligned} \tag{2.5}$$

where $\sigma(z) \triangleq \frac{1}{1+e^{-z}}$ is the logistic function, φ_{i_1} (resp. φ_{i_2}) is the learned representations of x_{i_1} (resp. x_{i_2}), and $\Sigma_0 \in \mathbf{S}_+^d$. The reason for multiplying by 2 is to make $P(y_i = 1|x_{i_1}, x_{i_2})$ close to 1 when φ_{i_1} is close to φ_{i_2} with respect to this scaled distance. This ensures that if we have two comparable samples, then the corresponding $y_i = 1$ with high probability. To estimate Σ_0 in EXPLORE from the humans' feedback, we seek the maximum of the log-likelihood

$$\begin{aligned} \ell_n(\Sigma) = \frac{1}{n}\sum_{i=1}^n y_i \log \frac{2\sigma(-\langle D_i, \Sigma\rangle)}{1 - 2\sigma(-\langle D_i, \Sigma\rangle)} \\ + \log(1 - 2\sigma(-\langle D_i, \Sigma\rangle)). \end{aligned} \tag{2.6}$$

on $\mathbf{S}_+^d$. As ℓ_n is concave (in Σ), we appeal to a stochastic gradient descent (SGD) algorithm to maximize ℓ_n. The update rule is

$$\Sigma_{t+1} = \texttt{ProjPSD}(\Sigma_t + \eta_t \partial\tilde{\ell}_n(\Sigma_t)),$$

where $\tilde{\ell}_n$ is the likelihood of the t-th minibatch, $\eta_t >$ is a step size parameter, and `ProjPSD` is the projection onto the PSD cone.

2.3. FACE vs EXPLORE

At first blush, the choice of which approach to use seems clear from the data. If the data consists of groups of comparable samples, then the factor analysis approach is appropriate. On the other hand, if data consists of pair-wise comparisons, then the logistic-regression approach is more appropriate. However, the type of data is usually part of the design, so the question is best rephrased as which type of data should the learner solicit. As we shall see, if the data is accurate and consistent, then FACE usually leads to good results. However, if the data is noisy, then EXPLORE is more robust.

The core issue here is a bias variance trade-off. Data in the form of a large group of comparable samples is more informative than pair-wise comparisons. As FACE is capable of fully utilizing this form of supervision, it leads to estimates with smaller variance. However, FACE is also more sensitive to heterogeneity within the groups of comparable samples as FACE is fully unbiased if all the variation in the group can be attributed to the sensitive attribute. If some of the variation is due to the discriminative attributes, then FACE leads to biased estimates. On the other hand, EXPLORE imposes no conditions on the homogeneity of the comparable and incomparable pairs in the training data. While EXPLORE cannot fully utilize comparable groups of size larger than two, it is also more robust to heterogeneity in the pairs of samples in the training data.

In the end, the key factor is whether it is possible for humans to provide homogeneous groups of comparable samples. In some applications, there are homogeneous groups of comparable samples. For example, in natural language processing, names are a group of words that ought to be treated similar in many ML tasks. For such applications, the factor analysis approach usually leads to better results. In other applications where there is less consensus on whether samples are comparable, the logistic regression approach usually leads to better results. As we shall see, our computational results validate our recommendations here.

2.4. Related work

Metric learning The literature on learning the fair metric is scarce. The most relevant paper is (Ilvento, 2019), which considers learning the fairness metric from consistent humans. On the other hand, there is a voluminous literature on metric learning in other applications (Bellet et al., 2013; Kulis, 2013; Suárez et al., 2018; Moutafis et al., 2017), including a variety of methods for metric learning from human feedback (Frome et al., 2007; Jamieson & Nowak, 2011; Tamuz et al., 2011; van der Maaten & Weinberger, 2012; Wilber et al., 2014; Zou et al., 2015; Jain et al., 2016). The approach described in subsection 2.1 was inspired by (Bolukbasi et al., 2016; Bower et al., 2018).

Learning individually fair representations There is a complementary strand of work on enforcing individual fairness by first learning a fair representation and then training an ML model on top of the fair representation (Zemel et al., 2013; Bower et al., 2018; Madras et al., 2018; Lahoti et al., 2019). Although it works well on some ML tasks, these methods lack theoretical guarantees that they train individually fair ML models.

Enforcing individual fairness We envision FACE and EXPLORE as the first stage in a pipeline for training individually fair ML models. The metrics from FACE and EXPLORE may be used in conjunction with methods that enforce individual fairness (Kim et al., 2018; Rothblum & Yona, 2018; Yurochkin et al., 2020; Yurochkin & Sun, 2020) or methods for individual fairness auditing (Xue et al., 2020). There are other methods that enforce individual fairness without access to a metric (Gillen et al., 2018; Jung et al., 2019). These methods depend on an oracle that detects violations of individual fairness, and can be viewed as combinations of a metric learning method and a method for enforcing individual fairness with a metric.

3. Theoretical properties of FACE

In this section, we investigate the theoretical properties of FACE. We defer proofs and theoretical properties of EXPLORE to the Appendix.

3.1. Learning from pairwise comparison

In this subsection, we establish theory of FACE when we learn the fair metric from comparable pairs. Given a pair $(\varphi_{i,1}, \varphi_{i,2})$ (the embedded version of $(x_{i,1}, x_{i,2})$), define for notational simplicity $z_i = \varphi_{i_1} - \varphi_{i_2}$. Here, we only consider those z_i's which come from a comparable pair, i.e., with corresponding $y_i = 1$. Under our assumption of factor model (see equation (2.3)) we have:

$$\begin{aligned} z_i &= \varphi_{i_1} - \varphi_{i_2} \\ &= A_*(u_{i_1} - u_{i_2}) + B_*(v_{i_1} - v_{i_2}) + (\epsilon_{i_1} - \epsilon_{i_2}) \\ &= A_*\mu_i + B_*\nu_i + w_i \end{aligned} \tag{3.1}$$

Here we assume that the sensitive attributes have more than one dimension which corresponds to the setting of *intersectional fairness* (*e.g.* we wish to mitigate gender and racial bias). We also assume μ_i's and ν_i's are isotropic, variance of w_i is $\sigma^2 I_d$ and μ_i, ν_i, w_i are all independent of each other. The scalings of μ_i and ν_i are taken care of by the matrices A_* and B_* respectively. Let Σ_Z be covariance matrix of z_i's. From model equation 3.1 and aforementioned assumptions:

$$\Sigma_Z = A_*A_*^T + B_*B_*^T + \sigma^2 I_d \tag{3.2}$$

We assume that we know the dimension of the sensitive direction beforehand which is denoted by k here. As ϕ_{i_1} is comparable to ϕ_{i_2}, we expect that variability along the protected attribute is dominant. Mathematically speaking, we assume $\lambda_{min}(A_*A_*^T) > \|B_*B_*^T + \sigma^2 I_d\|_{op}$. Here the fair metric we try to learn is:

$$d_x(x_1, x_2) = \langle(\varphi_1 - \varphi_2), \Sigma_0(\varphi_1 - \varphi_2)\rangle$$

where $\Sigma_0 = \left(I - P_{\mathsf{ran}(A_*)}\right)$. To estimate (and hence eliminate) the effect of the protected attribute, we compute the SVD of the sample covariance matrix $S_n = \frac{1}{n}\sum_{i=1}^n z_i z_i^\top$ of the z_i's and project out the eigen-space corresponding to the top k eigenvectors, denoted by $\hat{U}$. Our estimated distance metric will be:

$$\hat{d}_x(x_1, x_2) = \langle(\varphi_1 - \varphi_2), \widehat{\Sigma}(\varphi_1 - \varphi_2)\rangle\,,$$

where $\hat{\Sigma} = \left(I - \hat{U}\hat{U}^\top\right)$. The following theorem quantifies the statistical error of the estimator:

Theorem 3.1. *Suppose z_i's are centered sub-gaussian random vectors, i.e. $\|z_i\|_{\psi_2} < \infty$ where ψ_2 is the Orlicz-2 norm. Then we have with probability at-least $1 - 2e^{-ct^2}$:*

$$\|\hat{\Sigma} - \Sigma_0\|_{op} \leq b + \frac{\delta \vee \delta^2}{\tilde{\gamma} - (\delta \vee \delta^2)} \tag{3.3}$$

for all $t < (\sqrt{n}\tilde{\gamma} - C\sqrt{d}) \wedge (\sqrt{n\tilde{\gamma}} - C\sqrt{d})$, where:

1. $b = \left(\frac{\lambda_{min}(A_*A_*^T)}{\|B_*B_*^T + \sigma^2 I_d\|_{op}} - 1\right)^{-1}$
2. $\delta = \frac{C\sqrt{d} + t}{\sqrt{n}}$.
3. $\tilde{\gamma} = \lambda_{min}(A_*A_*^T) - \|B_*B_*^T\|_{op}$.

The constants C, c depend only on $\|x_i\|_{\psi_2}$, the Orlicz-2 norm of the x_i's.

The error bound on the right side of (3.3) consists of two terms. The first term b is the approximation error/bias in the estimate of the sensitive subspace due to heterogeneity in the similar pairs. Inspecting the form of b reveals that the bias depends on the relative sizes of the variation in the sensitive subspace and that in the relevant subspace: the larger the variation in the sensitive subspace relative to that in the relevant subspace, the smaller the bias. In the ideal scenario where there is no variation in the relevant subspace, Theorem 3.1 implies our estimator converges to the sensitive subspace. The second term is the estimation error, which vanishes at the usual $\frac{1}{\sqrt{n}}$-rate. In light of our assumptions on the sub-Gaussianity of the z_i's, this rate is unsurprising.

3.2. Learning from group-wise comparisons

In this subsection, we consider the complementary setting in which we have a single group of n comparable samples. We posit a factor model for the features:

$$\varphi_i = m + A_*\mu_i + B_*\nu_i + \epsilon_i \quad i = 1, 2, \ldots, n, \tag{3.4}$$

where $m \in \mathbf{R}^d$ is a mean term that represents the common effect of the relevant attributes in this group of comparable samples, $A_*\mu_i$ represents the variation in the features due to the sensitive attributes, and $B_*\nu_i$ represents any residual variation due to the relevant attributes (*e.g.* the relevant attributes are similar but not exactly identical). As before, we assume μ_i, ν_i's are isotropic, $\mathsf{Var}(\epsilon_i) = \sigma^2 I_d$ and the scale factors of μ_i's and ν_i's are taken care of by the matrices A_* and B_* respectively due to identifiability concerns. In other words, the magnitudes of $B_*\nu_i$'s are uniformly small. As the residual variation among the samples in this group due to the relevant factors are small, we assume that B_* is small compared to A_*, which can be quantified as before by assuming $\lambda_{\min}(A_*A_*^\top) > \|B_*B_*^\top + \sigma^2 I\|$. Hence to remove the effect of protected attributes, we estimate the column space of A_* from the sample and then project it out. From the above assumptions we can write the (centered) dispersion matrix of φ as:

$$\Sigma_\phi = A_*A_*^\top + B_*B_*^\top + \sigma^2 I, .$$

Note that the structure of Σ_z in the previous sub-section is same as Σ_φ as z is merely difference of two φ's. As before we assume we know dimension of the protected attributes which is denoted by k. Denote (with slight abuse of notation) by $\hat{U}$, the top k eigenvalues of $S_n = \frac{1}{n}\sum_{i=1}^n \varphi_i\varphi_i^\top$. Our final estimate of Σ_0 is $\widehat{\Sigma} = \left(I - \hat{U}\hat{U}^\top\right)$ and the corresponding estimated fair metric becomes:

$$d_x(x_1, x_2) = \langle(\varphi_1 - \varphi_2), \widehat{\Sigma}(\varphi_1 - \varphi_2)\rangle .$$

The following theorem provides a finite sample concentration bound on the estimation error:

Theorem 3.2. *Assume that φ_i have subgaussian tail, i.e $.\|\varphi_i\|_{\psi_2} < \infty$. Then with probability $\geq 1 - 2e^{-ct^2}$ we have:*

$$\|\widehat{\Sigma} - \Sigma_0\|_{op} \leq b + \tfrac{\delta \vee \delta^2}{\tilde{\gamma} - (\delta \vee \delta^2)} + \tfrac{t}{n}$$

for all $t < (\sqrt{n}\tilde{\gamma} - C\sqrt{d}) \wedge (\sqrt{n}\tilde{\gamma} - C\sqrt{d})$ where:

1. $b = \left(\frac{\lambda_{min}(A_*A_*^T)}{\|B_*B_*^T + \sigma^2 I_d\|_{op}} - 1\right)^{-1}$
2. $\delta = \frac{C\sqrt{d}+t}{\sqrt{n}}$.
3. $\tilde{\gamma} = \lambda_{min}(A_*A_*^T) - \|B_*B_*^T\|_{op}$.

The constants C, c only depend on the subgaussian norm constant of ϕ_i.

The error bound provided by Theorem 3.2 is similar to the error bound provided by Theorem 3.1 consists of two terms. The first term $\bar{B}$ is again the approximation error/bias in the estimate of the sensitive subspace due to heterogeneity in the group; it has the same form as the bias as in Theorem 3.1 and has a similar interpretation. The second term is the estimation error, which is also similar to the estimation error term in Theorem 3.1. The third term is the error incurred in estimating the mean of the φ_i's. It is a higher order term and does not affect the rate of convergence of the estimator.

3.3. Training individually fair ML models with FACE and SenSR

We envision FACE as the first stage in a pipeline for training fair ML models. In this section, we show that FACE in conjunction with SenSR (Yurochkin et al., 2020) trains individually fair ML models. To keep things concise, we adopt the notation of (Yurochkin et al., 2020). We start by stating our assumptions on the ML task.

1. We assume the embeded feature space of φ is bounded $R \triangleq \max\{\mathsf{diam}(\varphi), \mathsf{diam}_*(\varphi)\} < \infty$, where diam_* is the diameter of φ in the (unknown) exact fair metric
$$d_x^*(x_1, x_2) = \langle(\varphi_1 - \varphi_2), \Sigma_0(\varphi_1 - \varphi_2)\rangle^{1/2},$$
and diam is the diameter in the learned fair metric
$$d_x(x_1, x_2) = \langle(\varphi_1 - \varphi_2), \widehat{\Sigma}(\varphi_1 - \varphi_2)\rangle^{1/2}.$$
2. Define $\mathcal{L} = \{\ell(\cdot, \theta) \,:\, \theta \in \Theta\}$ as the loss class. We assume the functions in the loss class $\mathcal{L} = \{\ell(\cdot, \theta) : \theta \in \Theta\}$ are non-negative and bounded: $0 \leq \ell(z, \theta) \leq M$ for all $z \in \mathcal{Z}$ and $\theta \in \Theta$, and L-Lipschitz with respect to d_x:
3. the discrepancy in the fair metric is uniformly bounded: there is $\delta_c > 0$ such that
$$\sup_{(x_1,x_2)\in\mathcal{Z}} |d_x^2(x_1, x_2) - (d_x^*(x_1, x_2))^2| \leq \delta_c R^2.$$

The third assumption is satisfied with high probability as long as $\delta_c \geq (b + \frac{\delta \vee \delta^2}{\tilde{\gamma} - (\delta \vee \delta^2)})$.

Theorem 3.3. *Under the preceding assumptions, if we define $\delta^* \geq 0$ such that:*

$$\min_{\theta\in\Theta} \sup_{P: W_*(P, P_*) \leq \epsilon} \mathbb{E}_P\left[\ell(Z, \theta)\right] = \delta^* \tag{3.5}$$

and

$$\widehat{\boldsymbol{\theta}} \in \arg\min_{\theta\in\Theta} \sup_{P: W(P, P_n) \leq \epsilon} \mathbb{E}_P\left[\ell(Z, h)\right],$$

then the estimator $\hat{\theta}$ satisfies:

$$\sup_{P: W_*(P, P_*) \leq \epsilon} \mathbb{E}_P\left[\ell(Z, \widehat{\boldsymbol{\theta}})\right] - \mathbb{E}_{P_*}\left[\ell(Z, \widehat{\boldsymbol{\theta}})\right] \leq \delta^* + 2\delta_n, \tag{3.6}$$

where W and W_* are the learned and exact fair Wasserstein distances induced by the learned and exact fair metrics (see Section 2.1 in Yurochkin et al. (2020)) and

$$\delta_n \leq \tfrac{48\mathfrak{C}(\mathcal{L})}{\sqrt{n}} + \tfrac{48LR^2}{\sqrt{n\epsilon}} + \tfrac{L\delta_c R^2}{\sqrt{\epsilon}} + M\left(\tfrac{\log\frac{2}{t}}{2n}\right)^{\frac{1}{2}}.$$

where $\mathfrak{C}(\mathcal{L}) = \int_0^\infty \sqrt{\log\left(\mathcal{N}_\infty\left(\mathcal{L}, r\right)\right)}\, dr$, *with* $\mathcal{N}_\infty\left(\mathcal{L}, r\right)$ *being the covering number of the loss class* $\mathcal{L}$ *with respect to the uniform metric.*

Theorem 3.3 guarantees FACE in conjunction with SenSR trains an individually fair ML model in the sense that its fair gap (3.6) is small. Intuitively, a small fair gap means it is not possible for an auditor to affect the performance of the ML model by perturbing the training examples in certain "sensitive" ways.

The same conclusion can also be drawn using Theorem 3.2 with essentially similar line of arguments.

Remark 3.4. *The theory of EXPLORE is same in spirit with the theory of Face. In EXPLORE, we try to learn fair metric from comparable and incomparable pairs. As mentioned in the previous section, we solve MLE under the assumption of quadratic logit link to estimate* Σ_0. *Under the assumption that the parameter space and the space of embedded covariates* ($\varphi(x)$) *are boudned, we can establish the finite sample concentration bound of our estimator. It is also possible to combine our results with the results of Yurochkin et al. (2020) to obtain guarantees on the individual fairness of ML models trained with EXPLORE and SenSR (see Corollary B.9 in Supplement).*

4. Computational results

In this section, we investigate the performance of the learned metrics on two ML tasks: income classification and sentiment analysis.

4.1. Eliminating biased word embeddings associations

Many recent works have observed biases in word embeddings (Bolukbasi et al., 2016; Caliskan et al., 2017; Brunet et al., 2019; Dev & Phillips, 2019; Zhao et al., 2019). Bolukbasi et al. (2016) studied gender biases through the task of finding analogies and proposed a popular debiasing algorithm. Caliskan et al. (2017) proposed a more methodological way of analyzing various biases through a series of Word Embedding Association Tests (WEATs). *We show that replacing the metric on the word embedding space with a fair metric learned by FACE or EXPLORE eliminates most biases in word embeddings*.

Word embedding association test Word embedding association test (WEAT) was developed by (Caliskan et al., 2017) to evaluate semantic biases in word embeddings. The tests are inspired by implicit association tests (IAT) from the psychometrics literature (Greenwald et al., 1998). Let $\mathcal{X}, \mathcal{Y}$ be two sets of word embeddings of *target words* of equal size (*e.g.* African-American and European-American names respectively), and $\mathcal{A}, \mathcal{B}$ be two sets of *attribute words* (*e.g.* words with positive and negative sentiment respectively). For each word $x \in \mathcal{X}$, we measure its association with the attribute by

$$s(x, \mathcal{A}, \mathcal{B}) \triangleq \frac{1}{|\mathcal{A}|}\sum_{a\in\mathcal{A}} \frac{\langle x, a\rangle}{\|x\|\|a\|} - \frac{1}{|\mathcal{B}|}\sum_{b\in\mathcal{B}} \frac{\langle x, b\rangle}{\|x\|\|b\|} \quad (4.1)$$

If x tends to be associated with the attribute (*e.g.* it has positive or negative sentiment), then we expect $s(x, A, B)$ to be far from zero. To measure the association of $\mathcal{X}$ with the attribute, we average the associations of the words in $\mathcal{X}$:

$$s(\mathcal{X}, \mathcal{A}, \mathcal{B}) \triangleq \frac{1}{|\mathcal{X}|}\sum_{x\in\mathcal{X}} s(x, \mathcal{A}, \mathcal{B}).$$

Following (Caliskan et al., 2017), we consider the absolute difference between the associations of $\mathcal{X}$ and $\mathcal{Y}$ with the attribute as a test statistic:

$$s(\mathcal{X}, \mathcal{Y}, \mathcal{A}, \mathcal{B}) \triangleq |s(\mathcal{X}, \mathcal{A}, \mathcal{B}) - s(\mathcal{Y}, \mathcal{A}, \mathcal{B})|.$$

Under the null hypothesis, $\mathcal{X}$ and $\mathcal{Y}$ are equally associated with the attribute (*e.g.* names common among different races have similar sentiment). This suggests we calibrate the test by permutation. Let $\{(X_\sigma, Y_\sigma)\}_\sigma$ be the set of all partitions of $\mathcal{X} \cup \mathcal{Y}$ into two sets of equal size. Under the null hypothesis, $s(\mathcal{X}, \mathcal{Y}, \mathcal{A}, \mathcal{B})$ should be typical among the values of $\{s(\mathcal{X}_\sigma, \mathcal{Y}_\sigma, \mathcal{A}, \mathcal{B})\}$. We summarize the "atypicality" of $s(\mathcal{X}, \mathcal{Y}, \mathcal{A}, \mathcal{B})$ with a two-sided p-value[2]

$$\mathbf{P} = \frac{\sum_\sigma \mathbf{1}\{s(\mathcal{X}_\sigma, \mathcal{Y}_\sigma, \mathcal{A}, \mathcal{B}) > s(\mathcal{X}, \mathcal{Y}, \mathcal{A}, \mathcal{B})\}}{\mathbf{card}(\{(X_\sigma, Y_\sigma)\}_\sigma)}.$$

Following (Caliskan et al., 2017), we also report a standardized effect size

$$\mathbf{d} = \frac{s(\mathcal{X}, \mathcal{Y}, \mathcal{A}, \mathcal{B})}{\mathrm{SD}(\{s(x, \mathcal{A}, \mathcal{B})\}_{x\in\mathcal{X}\cup\mathcal{Y}})}$$

for a more fine-grained comparison of the methods.

Learning EXPLORE and FACE: To apply our fair metric learning approaches we should define a set of comparable samples for FACE and a collection of comparable and incomparable pairs for EXPLORE.

For the set of comparable samples for FACE we choose embeddings of a side dataset of 1200 popular baby names in New York City[3]. The motivation is two-fold: (i) from the perspective of individual fairness, it is reasonable to say that human names should be treated similarly in NLP tasks such as resume screening; (ii) multiple prior works have observed that names capture biases in word embeddings

[2] Caliskan et al. (2017) used one-sided p-value, however we believe that inverse association is also undesired and use a two-sided one

[3] available from https://catalog.data.gov/dataset/

Table 1. Word Embedding Association Test (WEAT) results. p-values that are significant/insignificant at the 0.05-level are shown in bold. See Table 3 in Supplement for the unabbreviated forms of the targets and attributes

Target	Attribute	Euclidean						EXPLORE		FACE-3		FACE-10		FACE-50	
		P	**d**	**P**	**d**	**P**	**d**	**P**	**d**	**P**	**d**	**P**	**d**	**P**	**d**
FLvINS	PLvUPL	**0.00**	1.58	**0.00**	1.55	**0.00**	1.55	**0.00**	1.41	**0.00**	1.59	**0.00**	1.56	**0.00**	1.27
INSTvWP	PLvUPL	**0.00**	1.46	**0.00**	1.45	**0.00**	1.46	**0.00**	1.44	**0.00**	1.48	**0.00**	1.58	**0.00**	1.49
MNTvPHS	TMPvPRM	**4e-5**	1.54	**4e-5**	1.54	**4e-5**	1.54	**4e-4**	1.31	**4e-5**	1.56	**0.00**	1.6	**0.00**	1.68
EAvAA	PLvUPL	0.00	1.36	0.00	1.36	0.00	1.38	1e-2	0.62	**5e-1**	0.17	7e-2	0.46	**2e-1**	0.33
EAvAA	PLvUPL	0.00	1.49	0.00	1.51	0.00	1.51	**2e-1**	0.49	**7e-1**	0.15	**6e-2**	0.67	**2e-1**	0.51
EAvAA	PLvUPL	8e-5	1.31	4e-5	1.41	4e-5	1.41	**1e-1**	0.55	**4e-1**	0.31	**4e-1**	0.3	**4e-1**	0.34
MNvFN	CARvFAM	0.00	1.69	2e-3	1.23	2e-3	1.23	**2e-1**	0.25	1e-3	1.24	6e-3	1.13	**8e-2**	0.53
MTHvART	MTvFT	8e-5	1.5	3e-2	0.84	1e-3	1.34	1e-3	1.34	1e-3	1.35	4e-3	1.18	6e-3	1.16
SCvART	MTvFT	9e-3	1.05	9e-3	1.08	4e-2	0.76	**6e-2**	0.65	4e-2	0.72	3e-2	0.84	**1e-1**	0.3
YNGvOLD	PLvUPL	1e-2	1.0	2e-4	1.5	1e-4	1.53	**7e-2**	0.6	**9e-2**	0.5	2e-3	1.27	4e-3	1.16

and used them to improve fairness in classification tasks (Romanov et al., 2019; Yurochkin et al., 2020). We consider three choices for the number of factors of FACE: 3, 10 and 50.

For EXPLORE we construct comparable pairs by sampling pairs of names from the same pool of popular baby names, however because there are too many unique pairs, we subsample a random 50k of them. To generate the incomparable pairs we consider random 50k pairs of positive and negative words sampled from the dataset proposed by Hu & Liu (2004) for the task of sentiment classification.

WEAT results First we clarify how the associations (4.1) are computed for different methods. The Euclidean approach is to use word embeddings and directly compute associations in the vanilla Euclidean space; the approaches of Bolukbasi et al. (2016) [4] and Dev & Phillips (2019) [5] is debias word embeddings before computing associations; associations with FACE and EXPLORE are computed in the Mahalanobis metric space parametrized by a corresponding Σ, i.e. the inner product $\langle x, y\rangle = x^T\Sigma y$ and norm $\|x\| = \sqrt{\langle x, \Sigma x\rangle}$. When computing **P**, if the number of partitions of target words $\mathbf{card}(\{(X_\sigma, Y_\sigma)\}_\sigma)$ is too big, we subsample 50k partitions.

We evaluate all of the WEATs considered in (Caliskan et al., 2017) with the *exact same* target and attribute word combinations. The results are presented in Table 1.

First we verify that all of the methods preserve the celebrated ability of word embeddings to represent semantic contexts — all WEATs in the upper part of the table correspond to meaningful associations such as Flowers vs Insects and Pleasant vs Unpleasant and all p-values are small corresponding to the significance of the associations.

On the contrary, WEATs in the lower part correspond to racist (European-American vs African-American names and Pleasant vs Unpleasant) and sexist (Male vs Female names and Career vs Family) associations. The presence of such associations may lead to biases in AI systems utilizing word embeddings. Here, larger p-value **P** and smaller effect size **d** are desired. We see that previously proposed debiasing methods (Bolukbasi et al., 2016; Dev & Phillips, 2019), although reducing the effect size mildly, are not strong enough to statistically reject the association hypothesis. Our fair metric learning approaches EXPLORE and FACE (with 50 factors) each successfully removes 5 out of 7 unfair associations, including ones not related to names. We note that there is one case, Math vs Arts and Male vs Female terms, where all of our approaches failed to remove the association. We think that, in addition to names, considering a group of comparable gender related terms for FACE and comparable gender related pairs for EXPLORE can help remove this association.

When comparing FACE to EXPLORE, while both performed equally well on the WEATs, we note that learning fair metric using human names appears more natural with FACE. We believe that *all* names are comparable and any major variation among their embeddings could permeate bias in all of the word embeddings. FACE is also easier to implement and utilize than EXPLORE, as it is simply a truncated SVD of the matrix of names embeddings.

4.2. Applying EXPLORE with SenSR

SenSR is a method for training fair ML system given a fair metric (Yurochkin et al., 2020). In this paper we apply SenSR along with the fair metric learned using EXPLORE on the adult dataset (Bache & Lichman, 2013). This data-

[4] https://github.com/tolga-b/debiaswe

[5] github.com/sunipa/Attenuating-Bias-in-Word-Vec

Table 2. Summary of **Adult** experiment over 10 restarts. Results for all prior methods are copied from Yurochkin et al. (2020)

	B-Acc,%	S-Con.	GR-Con.	$\text{Gap}_G^{\text{RMS}}$	$\text{Gap}_R^{\text{RMS}}$	$\text{Gap}_G^{\max}$	$\text{Gap}_R^{\max}$
SenSR+Explore (With gender)	79.4	**0.966**	0.987	**0.065**	**0.044**	**0.084**	**0.059**
SenSR+Explore (Without gender)	78.9	0.933	0.993	0.066	0.05	**0.084**	0.063
SenSR	78.9	.934	.984	.068	.055	.087	.067
Baseline	**82.9**	.848	.865	.179	.089	.216	.105
Project	82.7	.868	**1.00**	.145	.064	.192	.086
Adv. debiasing	81.5	.807	.841	.082	.070	.110	.078
CoCL	79.0	-	-	.163	.080	.201	.109

set consists of 14 attributes over 48842 individuals. The goal is to predict whether each individual has income more than 50k or not based on these attributes. For applying EXPLORE, we need comparable and incomparable pairs. We define two individuals to be comparable if they belong to the same income group (i.e. both of them has $>$ 50k or $<$ 50k annual salary) but with opposite gender, whereas two individuals are said to be incomparable if they belong to the different income group. Based on this labeling, we learn fair metric $\hat{\Sigma}$ via EXPLORE. Finally, following Yurochkin et al. (2020), we project out a "sensitive subspace" defined by the coefficients of a logistic regression predicting gender from $\hat{\Sigma}$ i.e.:

$$\hat{\Sigma} \longleftarrow (I - P_{gender})\hat{\Sigma}(I - P_{gender})\,.$$

where P_{gender} is the projection matrix on the span of this sensitive subspace. We then apply SenSR along with

$$d_x(x_1, x_2) = (x_1 - x_2)^\top \hat{\Sigma}(x_1 - x_2)\,.$$

Although most of the existing methods use protected attribute to learn a fair classifier, this is not ideal as in many scenarios protected attributes of the individuals are not known. So, it is advisable to learn fair metric without using the information of protected attributes. In this paper we learned our metrics in two different ways (with or without using protected attribute) for comparison purpose:

1. **SenSR + EXPLORE (with gender)** utilizes gender attribute in classification following prior approaches.
2. **SenSR + EXPLORE (without gender)** discards gender when doing classification.

In Yurochkin et al. (2020), the authors provided a comparative study of the individual fairness on Adult data. They considered balanced accuracy (B-Acc) instead of accuracy due to class imbalance. The other metrics they considered for performance evaluations are prediction consistency of the classifier with respect to marital status (S-Con., i.e. spouse consistency) and with respect to sensitive attributes like race and gender (GR-Con.). They also used RMS gaps and maximum gaps between true positive rates across genders ($\text{Gap}_G^{\text{RMS}}$ and $\text{Gap}_G^{\max}$) and races ($\text{Gap}_R^{\text{RMS}}$ and $\text{Gap}_R^{\max}$) for the assessment of group fairness (See Appendix for the detailed definition). Here we use their results and compare with our proposed methods. The results are summarized in Table 2. It is evident that SenSR + EXPLORE (both with gender and without gender) outperforms SenSR (propsoed in (Yurochkin et al., 2020)) in almost every aspect. Discarding gender in our approach prevents from violations of individual fairness when flipping the gender attribute as seen by improved gender and race consistency metric, however accuracy, spouse consistency and group fairness metrics are better when keeping the gender. Despite this we believe that it is better to avoid using gender in income classification as it is highly prone to introducing unnecessary biases.

5. Summary and discussion

We studied two methods of learning the fair metric in the definition of individual fairness and showed that both are effective in ignoring implicit biases in word embeddings. Our methods remove one of the main barriers to wider adoption of individual fairness in machine learning. We emphasize that our methods are probabilistic in nature and naturally robust to inconsistencies in the data. Together with tools for training individually fair ML models (Yurochkin et al., 2020), the methods presented here complete a pipeline for ensuring that ML models are free from algorithmic bias/unfairness.

Acknowledgements

This paper is based on work supported by the National Science Foundation (NSF) under grants no. 1712962, 1830247, and 1916271. Any opinions, findings, and conclusions or recommendations expressed in this paper are those of the authors and do not necessarily reflect the views of the NSF.

References

Angwin, J., Larson, J., Mattu, S., and Kirchner, L. Machine Bias. www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing, May 2016.

Bache, K. and Lichman, M. UCI machine learning repository. 2013.

Bellet, A., Habrard, A., and Sebban, M. A Survey on Metric Learning for Feature Vectors and Structured Data. *arXiv:1306.6709 [cs, stat]*, June 2013.

Bertrand, M. and Mullainathan, S. Are Emily and Greg More Employable Than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *American Economic Review*, 94(4):991–1013, September 2004. ISSN 0002-8282. doi: 10.1257/0002828042002561.

Bolukbasi, T., Chang, K.-W., Zou, J., Saligrama, V., and Kalai, A. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *arXiv:1607.06520 [cs, stat]*, July 2016.

Bower, A., Niss, L., Sun, Y., and Vargo, A. Debiasing representations by removing unwanted variation due to protected attributes. *arXiv:1807.00461 [cs]*, July 2018.

Brock, A., Donahue, J., and Simonyan, K. Large Scale GAN Training for High Fidelity Natural Image Synthesis. *arXiv:1809.11096 [cs, stat]*, September 2018.

Brunet, M.-E., Alkalay-Houlihan, C., Anderson, A., and Zemel, R. Understanding the origins of bias in word embeddings. In *International Conference on Machine Learning*, pp. 803–811, 2019.

Caliskan, A., Bryson, J. J., and Narayanan, A. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, April 2017. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.aal4230.

Chouldechova, A. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *arXiv:1703.00056 [cs, stat]*, February 2017.

Dastin, J. Amazon scraps secret AI recruiting tool that showed bias against women. *Reuters*, October 2018.

Davis, C. and Kahan, W. The Rotation of Eigenvectors by a Perturbation. III. *SIAM Journal on Numerical Analysis*, 7 (1):1–46, March 1970. ISSN 0036-1429. doi: 10.1137/0707001.

Dev, S. and Phillips, J. Attenuating bias in word vectors. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pp. 879–887, 2019.

Dwork, C., Hardt, M., Pitassi, T., Reingold, O., and Zemel, R. Fairness Through Awareness. *arXiv:1104.3913 [cs]*, April 2011.

Friedler, S. A., Scheidegger, C., and Venkatasubramanian, S. On the (im)possibility of fairness. *arXiv:1609.07236 [cs, stat]*, September 2016.

Frome, A., Singer, Y., Sha, F., and Malik, J. Learning Globally-Consistent Local Distance Functions for Shape-Based Image Retrieval and Classification. In *2007 IEEE 11th International Conference on Computer Vision*, pp. 1–8, October 2007. doi: 10.1109/ICCV.2007.4408839.

Garg, S., Perot, V., Limtiaco, N., Taly, A., Chi, E. H., and Beutel, A. Counterfactual Fairness in Text Classification through Robustness. *arXiv:1809.10610 [cs, stat]*, September 2018.

Garg, S., Perot, V., Limtiaco, N., Taly, A., Chi, E. H., and Beutel, A. Counterfactual fairness in text classification through robustness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 219–226. ACM, 2019.

Gillen, S., Jung, C., Kearns, M., and Roth, A. Online Learning with an Unknown Fairness Metric. *arXiv:1802.06936 [cs]*, February 2018.

Greenwald, A. G., McGhee, D. E., and Schwartz, J. L. K. Measuring Individual Differences in Implicit Cognition: The Implicit Association Test. *Journal of Personality and Soclal Psychology*, 74(6):17, 1998.

Hu, M. and Liu, B. Mining and Summarizing Customer Reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 10, Seattle, WA, August 2004.

Ilvento, C. Metric Learning for Individual Fairness. *arXiv:1906.00250 [cs, stat]*, June 2019.

Jain, L., Jamieson, K., and Nowak, R. Finite Sample Prediction and Recovery Bounds for Ordinal Embedding. *arXiv:1606.07081 [cs, stat]*, June 2016.

Jamieson, K. G. and Nowak, R. D. Low-dimensional embedding using adaptively selected ordinal data. In *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1077–1084, September 2011. doi: 10.1109/Allerton.2011.6120287.

Jung, C., Kearns, M., Neel, S., Roth, A., Stapleton, L., and Wu, Z. S. Eliciting and Enforcing Subjective Individual Fairness. *arXiv:1905.10660 [cs, stat]*, May 2019.

Kannan, H., Kurakin, A., and Goodfellow, I. Adversarial Logit Pairing. *arXiv:1803.06373 [cs, stat]*, March 2018.

Kim, M. P., Reingold, O., and Rothblum, G. N. Fairness Through Computationally-Bounded Awareness. *arXiv:1803.03239 [cs]*, March 2018.

Kleinberg, J., Mullainathan, S., and Raghavan, M. Inherent Trade-Offs in the Fair Determination of Risk Scores. *arXiv:1609.05807 [cs, stat]*, September 2016.

Kulis, B. Metric Learning: A Survey. *Foundations and Trends® in Machine Learning*, 5(4):287–364, 2013. ISSN 1935-8237, 1935-8245. doi: 10.1561/2200000019.

Lahoti, P., Gummadi, K. P., and Weikum, G. iFair: Learning Individually Fair Data Representations for Algorithmic Decision Making. *arXiv:1806.01059 [cs, stat]*, February 2019.

Madaan, N., Mehta, S., Agrawaal, T., Malhotra, V., Aggarwal, A., Gupta, Y., and Saxena, M. Analyze, Detect and Remove Gender Stereotyping from Bollywood Movies. In *Conference on Fairness, Accountability and Transparency*, pp. 92–105, January 2018.

Madras, D., Creager, E., Pitassi, T., and Zemel, R. Learning Adversarially Fair and Transferable Representations. *arXiv:1802.06309 [cs, stat]*, February 2018.

Massart, P., Nédélec, É., et al. Risk bounds for statistical learning. *The Annals of Statistics*, 34(5):2326–2366, 2006.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. Distributed Representations of Words and Phrases and their Compositionality. In Burges, C. J. C., Bottou, L., Welling, M., Ghahramani, Z., and Weinberger, K. Q. (eds.), *Advances in Neural Information Processing Systems 26*, pp. 3111–3119. Curran Associates, Inc., 2013.

Monteith, L. L. and Pettit, J. W. Implicit and explicit stigmatizing attitudes and stereotypes about depression. *Journal of Social and Clinical Psychology*, 30(5):484–505, 2011.

Moutafis, P., Leng, M., and Kakadiaris, I. A. An Overview and Empirical Comparison of Distance Metric Learning Methods. *IEEE Transactions on Cybernetics*, 47(3):612–625, March 2017. ISSN 2168-2267. doi: 10.1109/TCYB.2016.2521767.

Nosek, B. A., Banaji, M. R., and Greenwald, A. G. Harvesting implicit group attitudes and beliefs from a demonstration web site. *Group Dynamics: Theory, Research, and Practice*, 6(1):101, 2002a.

Nosek, B. A., Banaji, M. R., and Greenwald, A. G. Math= male, me= female, therefore math≠ me. *Journal of personality and social psychology*, 83(1):44, 2002b.

Radford, A., Metz, L., and Chintala, S. Unsupervised Representation Learning with Deep Convolutional Generative Adversarial Networks. *arXiv:1511.06434 [cs]*, November 2015.

Romanov, A., De-Arteaga, M., Wallach, H., Chayes, J., Borgs, C., Chouldechova, A., Geyik, S., Kenthapadi, K., Rumshisky, A., and Kalai, A. T. What's in a Name? Reducing Bias in Bios without Access to Protected Attributes. *arXiv:1904.05233 [cs, stat]*, April 2019.

Rothblum, G. N. and Yona, G. Probably Approximately Metric-Fair Learning. *arXiv:1803.03242 [cs]*, March 2018.

Suárez, J. L., García, S., and Herrera, F. A Tutorial on Distance Metric Learning: Mathematical Foundations, Algorithms and Software. *arXiv:1812.05944 [cs, stat]*, December 2018.

Tamuz, O., Liu, C., Belongie, S., Shamir, O., and Kalai, A. T. Adaptively Learning the Crowd Kernel. *arXiv:1105.1033 [cs]*, May 2011.

van der Maaten, L. and Weinberger, K. Stochastic triplet embedding. In *2012 IEEE International Workshop on Machine Learning for Signal Processing*, pp. 1–6, September 2012. doi: 10.1109/MLSP.2012.6349720.

van der Vaart, A. W. *Asymptotic Statistics*. Cambridge University Press, October 1998. doi: 10.1017/CBO9780511802256.

van der Vaart, A. W. and Wellner, J. A. Weak convergence. In *Weak convergence and empirical processes*, pp. 16–28. Springer, 1996.

Vershynin, R. Introduction to the non-asymptotic analysis of random matrices. *arXiv:1011.3027 [cs, math]*, November 2011.

Vigdor, N. Apple Card Investigated After Gender Discrimination Complaints. *The New York Times*, November 2019. ISSN 0362-4331.

Wang, H., Grgic-Hlaca, N., Lahoti, P., Gummadi, K. P., and Weller, A. An Empirical Study on Learning Fairness Metrics for COMPAS Data with Human Supervision. *arXiv:1910.10255 [cs]*, October 2019.

Wilber, M. J., Kwak, I. S., and Belongie, S. J. Cost-Effective HITs for Relative Similarity Comparisons. *arXiv:1404.3291 [cs]*, April 2014.

Xue, S., Yurochkin, M., and Sun, Y. Auditing ML Models for Individual Bias and Unfairness. In *International Conference on Artificial Intelligence and Statistics*, pp. 4552–4562, June 2020.

Yurochkin, M. and Sun, Y. SenSeI: Sensitive Set Invariance for Enforcing Individual Fairness. *arXiv preprint arXiv:2006.14168*, 2020.

Yurochkin, M., Bower, A., and Sun, Y. Training individually fair ML models with sensitive subspace robustness. In *International Conference on Learning Representations*, Addis Ababa, Ethiopia, 2020.

Zemel, R., Wu, Y., Swersky, K., Pitassi, T., and Dwork, C. Learning Fair Representations. In *International Conference on Machine Learning*, pp. 325–333, February 2013.

Zhao, J., Wang, T., Yatskar, M., Cotterell, R., Ordonez, V., and Chang, K.-W. Gender bias in contextualized word embeddings. *arXiv preprint arXiv:1904.03310*, 2019.

Zou, J. Y., Chaudhuri, K., and Kalai, A. T. Crowdsourcing Feature Discovery via Adaptively Chosen Comparisons. *arXiv:1504.00064 [cs, stat]*, March 2015.