

Fairness in representation: quantifying stereotyping as a representational harm*

Mohsen Abbasi[†] Sorelle A. Friedler[‡] Carlos Scheidegger[§]
Suresh Venkatasubramanian[¶]

Abstract

While harms of allocation have been increasingly studied as part of the subfield of algorithmic fairness, harms of representation have received considerably less attention. In this paper, we formalize two notions of stereotyping and show how they manifest in later allocative harms within the machine learning pipeline. We also propose mitigation strategies and demonstrate their effectiveness on synthetic datasets.

1 Introduction

In the rapidly growing area of fairness, accountability and transparency in machine learning, one of the fundamental questions is the problem of *discrimination*: are there disparities between social groups in the way decisions are made? This question has been typically studied as a *harm of allocation*: a problem in the way a learned model allocates decisions to entities. In a talk at NIPS 2017[13], Kate Crawford proposed studying instead harms of *representation*: the ways in which individuals might be *represented* differently in a feature space even before training a model. For example, she describes the representation of Black people as inherently more criminal as a harm whether or not hiring decisions are made based on that representation. Similarly, the work that has been done showing that word embeddings contain gender bias [4, 8, 33] identifies a harm of representation.

To study representational harm and how to minimize it, we must quantify it. Friedler et al. [18] provide a framework that explicitly calls out the distinction between the *construct space*, the desired representation of individuals, and the *observed space*, the measured attributes. They then propose measures of distortion between these spaces as a way to measure structural bias

in representation.

Harms of representation come in many different forms. Perhaps the most ubiquitous one is *stereotyping* – the tendency to assign characteristics to *all* members of a group based on stereotypical features shared by a few. In this paper, we focus on quantifying stereotyping as a form of representation distortion. Our goal is to apply the general framework of Friedler et al. in this specific context, in order to design measures that are specifically sensitive to stereotyping effects.

1.1 Our Contributions. Our main contributions in this paper are:

- A formal mechanism for stereotyping as a function from construct to observed space. This mechanism can be interpreted probabilistically or geometrically: in its former form it aligns with literature in psychology that explores how people form stereotypes.
- A demonstration of the effects of stereotyping in model building.
- A proposal for mitigating the effects of stereotyping – in effect an attempt to “invert” stereotyping as defined above – and experimental evidence demonstrating its effectiveness.

2 Literature review

In psychology, a stereotype is defined as an over-generalized belief about a particular group or class of people [10]. Stereotypes can be positive: *Asians are good at math*, or negative: *African-American names are more associated with criminal backgrounds* [35]; Stereotypes are not limited to just racial groups and they can change over time [25]. One of the central purposes served by applying stereotypes is to simplify our social world as they reduce the amount of information processing we have to do when faced with situations similar to our past experiences [20]. Though stereotypes can be seen as helping people respond to different social situations more promptly, they make us overlook individual

*This research was funded in part by the NSF under grants IIS-1633387, IIS-1633724, IIS-1513651 and IIS-1815238.

[†]University of Utah, <mailto:mohsena@cs.utah.edu>

[‡]Haverford College, <mailto:sorelle@cs.haverford.edu>

[§]University of Arizona, <mailto:cscheid@cs.arizona.edu>

[¶]University of Utah, <mailto:suresh@cs.utah.edu>

differences; this can lead to prejudice [34].

The approach we take in studying stereotypes is inspired by the literature on social cognition. This approach defines stereotypes as beliefs about the characteristics, attributes, and behaviors of members of certain groups and views stereotypic thinking as a mechanism which serves a variety of cognitive and motivational processes e.g. simplifying information processing to save cognitive resources [20]. Various models have been proposed and used to represent stereotyping. In the *prototype* model, people store and use an abstraction of stereotyped group’s typical features and judge the the said group members by their similarity to this prototype [9]. In the *exemplar* model, people use specific, real world individuals as representatives for the groups. As a group might have a number of exemplars; which one comes to mind when the stereotyping process is being activated depends on the context and situation in which the encounter with the stereotyped group member has occurred [32]. In *associative networks*, stereotypes are considered as linked attributes which are extensively interconnected [27]. In the *schemas* model, stereotypes are thought of as highly generalized beliefs about group members with no specific abstraction or exemplar for an attribute tied with such beliefs. Finally, in the *representativeness* model, stereotyping is defined as distorted perception of the relative frequency of a *type* in the stereotyped group compared to that of a base group [5]. This definition is based on the *representativeness heuristic* due to Kahnemann and Tversky [22, 23], a similarity heuristic that people rely on to judge the likelihood of uncertain events, instead of following the principles of probability theory [36].

Stereotyping also makes an appearance in the economic literature on statistical discrimination [2, 1]. Statistical discrimination describes the process where employers, unable to perfectly assess worker’s productivity at the time of hiring, use information such as sex and race as proxies for the expected productivity. This is estimated by the employer’s prior knowledge of the average productivity of the group the worker belongs to. In other words, the stereotyping mechanism described here is the replacement of individual scores by a single aggregated score over a group, where the score is perceived as relevant for the individual. This literature views stereotyping as a rational response to insufficient information about individuals, rather than as a choice of representation that might distort outcomes.

The literature on algorithmic fairness has for the most part focused on bias caused by skew in training data distributions and the training process itself, and has quantified the bias in terms of fairness measures that are typically outcome-based (see, e.g., [7, 19, 39, 24]

and surveys [29, 37]). Some of the research in this area has taken a representational, preprocessing approach to reversing training data skew [17, 26, 40, 6, 16] by changing the inputs so that a classifier finds fair outcomes. In the context of unsupervised learning, recent work on fair clustering[12, 31, 3] and PCA[30, 28] seeks to generate a modified representation of the input points so that the new representation (clustering or reduced dimension) satisfies a notion of “balance” with respect to groups. In all of these, the goal is to use representation to guide (fair) learning, rather than look at skew in the representation itself.

There have been a few works that look at bias that emerges from the representation process, most notably when looking at *learned* representations that come from word-vector embeddings[4, 8, 33, 14]. The goal of these methods is to show how biases in language are preserved after doing such an embedding.

3 Modeling Stereotypes

We now propose mechanisms by which stereotyping might occur. We first present a novel geometric approach to stereotyping (and a variant on it), and then review a *probabilistic* approach first proposed in [5]. Finally, we show that these different perspectives on stereotyping can be unified in a common algebraic framework. Each mechanism will have an associated stereotyping measure, with larger values indicating a greater degree of stereotyping.

3.1 Stereotyping via Exemplars. Stereotyping via exemplars refers to generalizing features attributed to a small subset of a group, called *stereotypic exemplars*, to all of its members. In the simplest version of stereotyping by exemplars, a single exemplar *pulls* points towards itself, so that in the observed representation, points from one group are perceived to be closer to the exemplar (and thus closer in feature space) than they actually are.

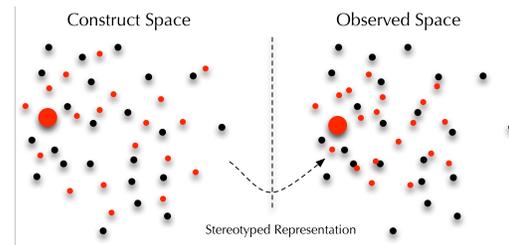


Figure 1: An example of stereotyping: points are drawn to the exemplar

This mechanism is illustrated in Figure 1. The left subfigure represents the original representation in

the construct space with two groups colored in red and black. Once an exemplar (the larger red point) is chosen, the observation process associates other point more closely with it – this is effectively achieved by having the red points mapped closer to the exemplar, as we see in the right subfigure.

Formally, we model this mechanism as follows. Let the minority (protected) group point set be denoted by P and the majority (unprotected) group point set be U . Fix an exemplar $c \in P$. Then each point $p \in P$ is shifted as follows:

$$(3.1) \quad p_\alpha = (1 - \alpha)p + \alpha c$$

The term α is the *measure of stereotyping*. Note that if $\alpha = 0$, no stereotyping happens, and if $\alpha = 1$, all points are collapsed to the exemplar. Points in U are not shifted at all.

Stereotyping Using Features. Stereotyping might happen only along some dimensions of the data, and not others. For example, we might stereotype all Asians as having higher aptitude for math based on an exemplar, but we might not borrow other attributes of the exemplar (say proficiency in sports, or crafting skills) and extend them also to all Asian people. Formally, this amounts to performing stereotyping in a subspace of relevant features. Assume that only k of d features are influenced by stereotyping. Then we can write the mechanism for shifting a point $p \in P$ as before as

$$(3.2) \quad p_\alpha = \begin{pmatrix} (1 - \alpha)\mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{d-k} \end{pmatrix} p + \begin{pmatrix} \alpha\mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} c$$

where we have reordered features so that the stereotyped ones are the first k .

3.2 Stereotyping via Representativeness. We now turn to a probabilistic mechanism for stereotyping first proposed by Bordalo et al[5].

For ease of exposition, imagine a data representation consisting of single feature that takes T distinct values that we call “types”. For example, the feature might be **age** and the types might be specific age ranges. Let G be a *group* of individuals. The *representativeness* of a type $t \in T$ for group G is defined as the ratio of conditional probability of individuals in G having value t to the corresponding probability for the complement \overline{G} :

$$R(t, G, \overline{G}) = \frac{\Pr[t|G]}{\Pr[t|\overline{G}]}$$

Intuitively, the representativeness measures how distinctive t is at distinguishing groups. The larger it is, the more likely it is that the presence of t predicts group membership.

Stereotyping now occurs by *amplifying* perceived probabilities based on representativeness. Formally, the distorted perceived probabilities are

$$(3.3) \quad \Pr^{st}(t|G) = \Pr(t|G) \times \frac{h_t(R(t, G, \overline{G}))}{\sum_{s \in T} \Pr[s|G] h_s(R(s, G, \overline{G}))}$$

Where $h_t : R_+^T \rightarrow R_+$ is weakly increasing in t and weakly decreasing in the other $T - 1$ types.

For concreteness (and as suggested by Bordalo et al) let us assume that $h(\cdot)$ takes the form $h_t(x) = x^\rho$. Here $\rho > 1$ is again the measure of stereotyping. The larger it is, the more representativeness influences the perceived probabilities.

3.3 A unified view of stereotyping. While the geometric and probabilistic mechanisms described above look quite different on the surface, they are actually examples of a more general linear framework for thinking about stereotyping. For a point $p \in P$, let $v(p) : P \rightarrow F$ be an invertible transformation of p into a feature space F . We will define a *generalized stereotyping transform* parametrized by matrices A, B as the transformation $p' = v^{-1}(Av(p) + B)$ or more conveniently $v(p') = Av(p) + B$. Note that setting v to the identity mapping already recovers the two geometric transformations via equations (3.1),(3.2).

Consider the probabilistic transform defined by (3.3). As suggested, let us set $h(x) = x^\rho$. Further, let us define $\lambda(t) = \Pr(t | G) / \Pr(t | \overline{G})$. We can now rewrite (3.3).

$$(3.4) \quad p'_{t|G} = \frac{p_{t|G} \lambda(t)^\rho}{\sum_s p_{s|G} \lambda(s)^\rho}$$

By assumption, probabilities in the majority group \overline{G} are not modified, which implies that $p'_{t|\overline{G}} = p_{t|\overline{G}}$. Dividing both sides of (3.4) by the left and right sides of this equality, we get

$$(3.5) \quad \lambda'(t) = \frac{\lambda(t)^{1+\rho}}{\sum_s p_{s|G} \lambda(s)^\rho}$$

which after taking logs yields

$$(3.6) \quad \ln \lambda'(t) = (1 + \rho) \ln \lambda(t) - \ln \sum_s p_{s|G} \lambda(s)^\rho$$

Setting $v(p) = \ln p$ and noting that $\ln \lambda(t) = \ln p_{t|G} - \ln p_{t|\overline{G}}$ where $p_{t|\overline{G}}$ is fixed, we can recover the same linear relationship as before.

While we can unify the different mechanisms of stereotyping mathematically, we observe that the actual processes by which these mechanisms might modify

data are different. Representativeness is a form of cognitive bias that would manifest itself in data collection: for example in predictive policing, officers that might stereotype potential criminals based on race are more likely to observe (and therefore collect data) on crimes perpetuated by minorities. Stereotyping via exemplars (and features) might manifest itself at the time of feature selection: by omitting key differentiating features in the representation, points might end up appearing closer than they actually are.

4 Harms of Stereotyping

While we believe that stereotypes are harms independent of their implications for the machine learning process, the use of stereotyped data representations as training data additionally leads to disparities in outcomes by group. In this section, we demonstrate this analytically for Naive Bayes classification, linear regression, and clustering; thus demonstrating these harms in classification, regression, and unsupervised settings. For Naive Bayes classification, we adopt the probabilistic interpretation of stereotyping, while the other two are studied under the geometric interpretation.

4.1 Naive Bayes Classifier. For an arbitrary input X , Bayes' rule states that the probability of holding a class label $C_k, 1 \leq k \leq m$ given X is: $\Pr[C_k|X] = \frac{\Pr[X|C_k]\Pr[C_k]}{\Pr[X]}$. The goal in classification is to predict the most likely label for a given input. Naive Bayes classification makes the assumption that conditioned on the class label, these features are independent of each other, yielding the following predictive rule. $\hat{y} = \arg \max_k \Pr[C_k] \prod_{i=1}^n \Pr[x_i|C_k]$ where $x_i, 1 \leq i \leq n$, is a single feature of X . In the presence of stereotyping such an assumption might have an undesired impact on the classification results. Consider a dataset with sensitive attribute A , e.g., being *Asian*, and an attributes D_p relevant to known stereotypes for A , e.g. being *good at math*. Let's assume D_p is *positively* correlated with being Asian. In this setting, if the data gathering process is susceptible to this type of stereotyping, the Naive Bayes formulation would turn into:

$$(4.7) \quad \hat{y} = \arg \max_k \Pr[C_k] \prod_{i=1}^{n-3} \Pr[x_i|C_k] \times \Pr[D_p|A, C_k] \Pr[A|C_k]$$

where $\Pr[D_p|C_k]$ is replaced by $\Pr[D_p|A, C_k]$. According to the probabilistic interpretation of stereotyping [5], the perceived probability $\Pr[D_p|A]$ for the target group could be different from its actual value: $\Pr^{st}[D_p|A, C_k] \geq \Pr[D_p|A, C_k]$ which can influence the

classification results. Indeed, the degree of difference is related directly to the representativeness as a function of the ratio $\Pr^{st}[D_p|A, C_k]/\Pr[D_p|A, C_k]$.

This transition – from $\Pr[D_p|C_k]$ to $\Pr[D_p|A, C_k]$ is key to understanding stereotyping in this context. In other words, the point of a predictive tool is that it uses “other” features other than the class label to determine the probabilities for desirable attributes like D_p . Alternatively, we can think of $\Pr[D_p|A, C_k]$ as reflecting a process by which the variable A affects the data used to compute the conditional probability, either by removing data that does not have A or by overweighting data that does.

4.2 Linear Regression. In linear regression, we are given points $X = x_1, \dots, x_n \in \mathbb{R}^d$ and corresponding labels $y_1, \dots, y_n \in \mathbb{R}$. The goal is to find a vector of parameters β such that $y = X\beta + \epsilon$. It is well known that the least-squares solution to this regression is given by $\beta = (X^T X)^{-1} X^T y$. and our goal is to understand how perturbing the input X (via stereotyping) will change the coefficients¹.

We will consider a very simple form of perturbation, where only a single coordinate s is perturbed while the rest (including the dependent variable y) stay fixed. Assume that the data matrix X (in which each point is a row) is organized as $\begin{pmatrix} Q \\ P \end{pmatrix}$ where Q is the set of all majority group points and P consists of all minority group points. Let c be the exemplar that points are pulled towards (in dimension s). We can then write the perturbation as

$$X' = X + \alpha(\Xi c^{*T} - I_r X I_s)$$

where Ξ is a $n \times 1$ vector with values of 1 for minority data points and zeros for majority ones and c^* is a vector of all zeros except in the s^{th} position, where its value is c_s . I_r is an $n \times n$ diagonal matrix with values of 1 for rows in X representing minority data points and zeros everywhere else; I_s is a $d \times d$ matrix where its only non-zero element, which is 1, is at row and column s .

For a general perturbation of the form $X' = X + \Delta$ the coefficients in β are updated to:

$$(4.8) \quad \beta' = [(X + \Delta)^T (X + \Delta)]^{-1} (X + \Delta)^T y$$

¹There is an extensive literature on the stability of linear regression to *random* perturbations of the input[11]; the behavior of the coefficients is well understood for example when the inputs are perturbed using Gaussian noise. In our setting, the perturbations are not random and are structured in a specific way, and thus the prior forms of analysis, and even more recent work that looks at other forms of structured perturbations[15] appear to not apply directly.

The key term in the right hand side is the inverse, which can be written in general form as $[X^\top X + (\Delta^\top \Delta + X^\top \Delta + \Delta^\top X)]^{-1}$. This is asking for the inverse of a perturbation of a given matrix, for which we can use the famous Woodbury formula[38]:

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)VA^{-1})$$

Let the number of rows in P (the number of minority points) be m and μ be the centroid of the points in P . We can then write

$$\begin{aligned} X'^\top X' = & \\ X^\top X + \alpha(m\mu c^{*\top} - P^\top P I_s) + \alpha(m c^* \mu^\top - I_s P^\top P) & \\ - m\alpha^2(c^* \mu^\top I_s + I_s \mu c^{*\top} + c^* c^{*\top}) + \alpha^2 I_s P^\top P I_s & \end{aligned}$$

Therefore, the update to $X^\top X$ consists of two of rank-1 updates: $X'^\top X' = X^\top X + wu^\top + uw^\top$ where:

$$w_i = \begin{cases} \alpha(m c \mu_s - \|P_s^\top\|^2) + \frac{m\alpha^2(c^2 - 2c\mu_s) + \alpha^2 \|P_s^\top\|^2}{2}, & \text{if } i = s \\ \alpha(m c \mu_i - P_i^\top \cdot P_s^\top), & \text{otherwise} \end{cases}$$

and u is a basis vector with a 1 in dimension s . Setting $U_{d \times 2} = \begin{pmatrix} w & u \end{pmatrix}$, $V_{2 \times d} = \begin{pmatrix} u \\ w \end{pmatrix}$, $C = I_2$ and embedding these matrices in the Woodbury formula yields

$$(4.9) \quad \beta' = \underbrace{(X^\top X)^{-1} X'^\top Y}_{p_1} - \underbrace{(X^\top X)^{-1} M (X^\top X)^{-1} X'^\top Y}_{p_2}$$

The amount of perturbation is controlled by α . If we split Equation 4.9 into two parts as illustrated, the values in part p_1 change linearly in α because of the linear dependence on X' . However, each element in M is a quadratic function of α . Therefore, the values in part p_2 , change quadratically when α is increased.

4.3 Clustering. Unlike in the previous two cases, we will not present a formal analysis of how clustering is affected by stereotyping, because clustering (and especially k -means) is a global objective where it is often difficult to predict how perturbations will affect the outcome. Rather, we will demonstrate empirically the effect of stereotyping on clustering in Section 6.

However, we argue qualitatively here for why clustering will be affected by stereotyping. The effect of moving points closer to an arbitrarily chosen exemplar, especially if this exemplar is not the mean of the set of minority points, has the effect of shifting and concentrating clusters to make them look more homogeneous with respect to group identity. But if the exemplar is an outlier, then the cost of the clustering will increase: for k -means this increase in cost is related both to the distance between the exemplar and the cluster mean and the number of points that are moved.

5 Mitigating the Effects of Stereotyping

Since it is impossible to access construct space, in order to mitigate the unwanted influences of stereotyping on machine learning pipeline, we need to make assumptions about data in the ideal world. One useful assumption to make is based on *We're All Equal* worldview [18], where the idea is that in the ideal world, different groups look essentially the same. Though such assumption might not hold true in every possible scenario (e.g. women are on average shorter than men), it implicitly appears in much of literature on statistical discrimination. This motivates us to adopt *WAE* as an appropriate axiom in our work. Let's consider two social groups in a hypothetical dataset: minority and majority. The minority group is the one being stereotyped in the mapping from construct to observed space. Based on the *WAE* worldview, we assume the two groups are generated by same distribution which we can estimate, by looking at the majority group in the observed space. Therefore, the goal is to recover the true representation of the minority group in the observed space, based on majority group.

5.1 Mitigation of exemplar-based stereotyping.

Recall that in landmark-based stereotyping, a landmark c is fixed first, as well as a measure of stereotyping α . Then each point in the minority group is pulled towards the landmark resulting in a new point p_α . Let the resulting mean of the modified points be μ_α . Using equation (3.1) and noting that the stereotyping process is linear, we can write

$$(5.10) \quad \mu_\alpha = (1 - \alpha)\mu_m + \alpha c$$

where μ_m is the mean of all the points in the minority group. Our goal is now to determine the values of α and c so that we can reconstruct the original points. We will invoke the WAE assumption by assuming that the mean μ_m is close to the mean μ_M of the majority group: specifically, that $\|\mu_m - \mu_M\| \leq \epsilon$.

Our goal is to determine candidates for c, α . (5.10) tells us that any feasible exemplar c must lie on the line between μ_α and μ_m . Since by assumption μ_m lies in a ball of radius ϵ around μ_M (denoted by $B_\epsilon(\mu_M)$), the feasible region for c is a cone with apex at μ_α such that the surface of its complementary cone is tangent to the ball around μ_M . Let $\|\mu_\alpha - \mu_M\| = d$, and fix a point c . Then c is a feasible exemplar if the angle θ made by $c - \mu_\alpha$ with the vector $\mu_\alpha - \mu_m$ is such that $\sin \theta \leq \epsilon/d$. We denote the set of such vectors by $C_{\mu_\alpha, \mu_M}(\epsilon)$ and will drop the subscripts for notational ease.

We have assumed that the exemplar is a point in P , so the set of feasible exemplars can be written as $P \cap C(\epsilon)$. For each such c the set of possible locations for μ_m

is a line segment resulting from the intersection of the line supporting the segment $\overline{\mu_\alpha c}$ and $B_\epsilon(\mu_M)$. For each candidate choice of μ_m on this segment we can compute α using (5.10). We will seek the smallest possible α – this represents the most conservative measure of stereotyping that is consistent with the observed data. It is easy to see that this minimum is achieved by the endpoint of the line segment closest to μ_α and can be computed easily.

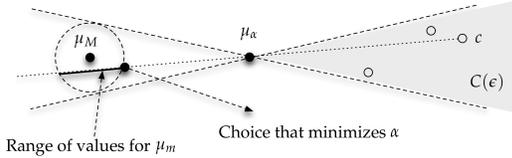


Figure 2: How we compute candidate c, α values

We now can associate a value of α with each $c \in P \cap C(\epsilon)$. We take the pair (c, α) such that α is minimized. This represents the most conservative reconstruction that is consistent with the WAE assumption and the observed data. We summarize this process as follows:

1. Compute the set $P \cap C(\epsilon)$ by verifying the angle test for each point of P .
2. For each candidate c in this set, compute its associated value of α .
3. Return the pair (c, α) that minimizes α .

5.1.1 Mitigating stereotyping with respect to features. The process above captures stereotyping with respect to exemplar points. We now consider the case of mitigating stereotypes that are limited to some features. As before, we can exploit the linearity of the perturbation (3.2) and the same idea as in the previous subsection to observe that the means μ_α, μ_m satisfy the same relationship.

$$\mu_\alpha = \begin{pmatrix} (1-\alpha)\mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{I}_{d-k} \end{pmatrix} \mu_m + \begin{pmatrix} \alpha\mathbf{I}_k & \mathbf{0} \\ \mathbf{0} & \mathbf{0} \end{pmatrix} c$$

This reduces to a k -dimensional version of the full stereotyping problem described above.

5.2 Mitigating Representativeness. We introduce the WAE modeling assumption: we assume that prior to stereotyping, the probabilities of types are very similar between majority and minority groups. Formally, we will express this condition as $1 - \epsilon \leq \lambda(t) \leq 1 + \epsilon$ where ϵ is a parameter that controls the degree to which the distributions are similar. Note that this assumption implies that the Kullback-Leibler divergence between the two conditional distributions is at most ϵ .

This follows from the fact that the KL-divergence can be written as $d_{KL}(\mathcal{D}_G, \mathcal{D}_{\bar{G}}) = \sum_t \Pr(t|G) \ln \lambda(t)$. Note that $\sum_s p_{s|G} \lambda(s)^\rho$ is a convex combination of the quantities $\{\lambda^\rho(s)\}$. It will be convenient to express this sum as $e^{\gamma\rho}$ where $e^\gamma \in [1 - \epsilon, 1 + \epsilon]$. Setting $\delta_t = \ln \lambda(t)$, we note that by the standard approximation of $\ln(1 + x)$, $|\delta_t| \leq \epsilon$.

We can now substitute these bounds on $\lambda(t)$ into (3.5), yielding $\ln \lambda'(t) = (1 + \rho)\delta_t - \gamma\rho = \delta_t + \rho(\delta_t - \gamma)$ where $|\delta_t|, |\gamma| \leq \epsilon$. Solving for ρ yields $\rho = \frac{\ln \lambda'(t) - \delta_t}{\delta_t - \gamma}$. There is one such equation for ρ for each value of t . The terms δ_t and γ are unknown (since they arise from the (unknown) distribution of types prior to stereotyping). However, we can use the bounds on these terms to provide bounds on the value of ρ .

The minimum value of ρ implied by any of the equations can be determined by setting δ_t to its largest value and γ to its smallest. This yields $\rho = \frac{\ln \lambda'(t) - \epsilon}{2\epsilon}$. Note that ρ can grow without bound, which implies that each equation yields a half-infinite range of possible values of ρ . Taking the intersection of all these ranges we conclude that ρ must lie in the range $[\frac{\ln \max \lambda'(t) - \epsilon}{2\epsilon}, \infty]$.

Each value of ρ in this range is a candidate measure of stereotyping and can be used to reconstruct the original probabilities $p(t|G)$. Specifically, it is easy to see that $\lambda(t)$ is proportional to $\lambda'(t)^{1/\rho}$, and since we can compute $p(t|\bar{G})$ directly we can obtain $p(t|G)$. Interestingly, if we now compute the KL-divergence between \mathcal{D}_G and $\mathcal{D}_{\bar{G}}$, it is a monotonically decreasing function of ρ . In other words, the smallest feasible value of ρ given above is consistent with the WAE assumption and is also the most conservative choice.

Summarizing, our procedure is:

1. Compute $\lambda'(t)$ for all types t .
2. Set $\rho = \frac{\ln \max \lambda'(t) - \epsilon}{2\epsilon}$.
3. Set $p_t = \lambda'(t)^{1/\rho}$.
4. Set $\tilde{p}(t|G) = \frac{p_t}{\sum_t p_t}$. Return $\{p(t|G)\}$.

6 Experiments

In this section, we provide an empirical assessment for harms of stereotyping to Naive Bayes classification, linear regression and clustering, using synthetic datasets. Also, for each one of these problems, we demonstrate the effectiveness of our mitigation methods.

6.1 Naive Bayes. In order to study the impact of stereotyping via representativeness we consider a synthetic dataset that allows us to manipulate the extent of stereotyping in the data and note its effect on a Naive Bayes classifier. The synthetic data contains three binary attributes and a class label: 1. the *sensitive*

attribute which is assigned randomly and specifies if an individual is Asian or not; 2. a randomly assigned attribute that has no correlation with the class label or sensitive attribute; 3. an attribute indicating if the individual is *good at math*, which contributes positively to receiving the desired classification; and 4. a class label indicating if the individual is *selected* for a job interview. 2000 instances are generated according to this procedure with a 50:50 training to test split ratio.

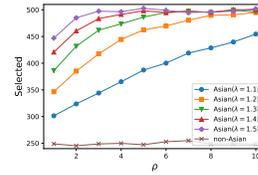
Stereotyping via representativeness can be thought of as amplifying disparities that already exist between target and base subgroups. The extent of these disparities is measured by $\lambda(t)$, and we assume that $\lambda(t) > 1$ for the type “good at math” with respect to the Asian subgroup. In order to simulate a stereotyped dataset, first we fix $\lambda(t)$ by assigning positive values to “good at math” for Asians, with a higher probability compared to non-Asians. Then, we gradually increase the positive correlation between being Asian and the attribute “good at math.” Specifically, the integer value of ρ is increased from 1 to 10 where

$$(6.11) \quad p'_{t|G} = \frac{p_{t|G} \lambda(t)^\rho}{\sum_s p_{s|G} \lambda(s)^\rho}$$

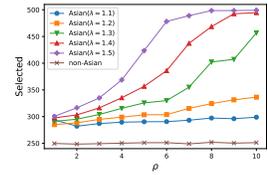
given that $\lambda(t) = \Pr(t | G) / \Pr(t | \bar{G})$ and G and \bar{G} denote Asian and non-Asian groups respectively.

The results are shown in Figure 3. As illustrated in Figure 3a, by boosting the representativeness of being “good at math,” the number of selected Asians increase while no difference is observed in the other group’s results. These results hold over different values of $\lambda(t)$. Figure 3b shows that the effects of stereotyping on the Asian subgroup’s results are significantly reduced by applying the representativeness mitigation solution. We should note there is a possibility for large values of $\lambda(t)$ and ρ , to saturate the probability of type t for the target group, e.g. $\lambda = 1.5, \rho = 10$ in Figure 3b. In such cases, since the stereotyped probabilities for the other types go down to zero, our mitigation method would not be able to retrieve the original probabilities.

6.2 Linear Regression. In this experiment, we study the effects of stereotyping under the geometric interpretation on linear regression. We again consider this on a synthetic dataset so that we can manipulate the extent of the stereotyping. The dataset contains four features: the first is a uniform randomly assigned binary *sensitive* attribute with values privileged and unprivileged; the second, third, and fourth features are numerical values between 0 and 1 assigned via a uniform distribution; and the dependent variable is a linear combination of the third and fourth attributes with noise



(a) Number of selected Asians for different values of λ and ρ



(b) Effects of mitigation on the number of selected Asians

Figure 3: Results of NB classification as the Asian subgroup’s representative type “good at math” is boosted.

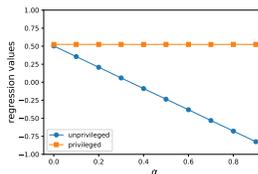
$-0.1 \leq \epsilon \leq 0.1$ added, i.e.:

$$(6.12) \quad y = -x_3 + 2x_4 + \epsilon.$$

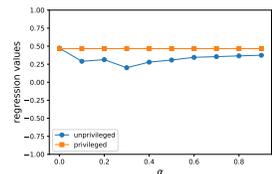
2000 instances are generated according to this procedure, and a 50:50 training to test split ratio is used.

We assume higher values for y are desired by individuals. In this experiment, we pick the individual with the lowest value for its dependent variable y as the exemplar, and modify values $x_2, x_3,$ and x_4 for individuals from the unprivileged group so that the distance between those individuals and the exemplar is decreased according to parameter $0 < \alpha < 1$ i.e. if $\alpha = 0$ the values don’t change and if $\alpha = 1$ the values for $x_2, x_3,$ and x_4 are the same as for the exemplar point.

The results of gradually increasing the value of α are shown in Figure 4. In the stereotyping process, since the values for the dependent variable are updated according to regression function 6.12, the regression coefficients stay the roughly same. But as a result of stereotyping, there will be a disparity in the regression values as shown in Figure 4a. Looking at Figure 4b, we observe that the disparities in regression values, which were caused by stereotyping, are reduced by applying the exemplar-based mitigation method.



(a) Linear decrease in \hat{Y} for unprivileged group as α increases



(b) The average regression values after applying mitigation

Figure 4: Changes in the regression values for the two groups as stereotyping gets more aggressive

6.3 Clustering. We now empirically study the effects of stereotyping under geometric interpretation on K -means clustering. We create a synthetic dataset with three features. The first feature is a uniform randomly assigned binary *sensitive* attribute with values “privileged” and “unprivileged”. We use two 2-dimensional normal distributions $\mathcal{N}(0, 0.3^2)$ and $\mathcal{N}(1, 0.3^2)$ to assign values to the second and third attributes; more specifically, in each social group, the second and third attributes for half the data are generated using one distribution and for the other half from the other. As for stereotyping, we pick an arbitrary exemplar within the unprivileged group and move the points representing the remaining individuals in this group towards it according to parameter $0 < \alpha < 1$, as illustrated in Figure 5.

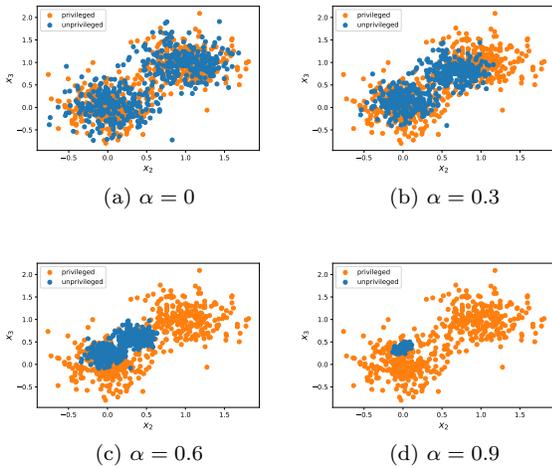


Figure 5: Stereotyping under geometrical interpretation for different values of α

In Figures 6a and 6b we compare the results of k -means clusterings on stereotyped data and its mitigated representations for different values of $0 \leq \alpha \leq 1$. We see a strong effect of the stereotyped representation for larger values of α and find that the mitigation strategy removes that effect. This holds under both the rand-index score [21] and the balance notion proposed by [12]. In addition, although the solution proposed by Chierichetti et al. [12] achieves a fair clustering, it would increase the cost of clustering in presence of stereotyping. This higher cost for fair clustering compared to k -means and the effectiveness of mitigation strategy are illustrated in Figures 6c and 6d respectively.

7 Discussion and Conclusion

In this paper, we formalized two notions of stereotyping and demonstrated how stereotyped representations

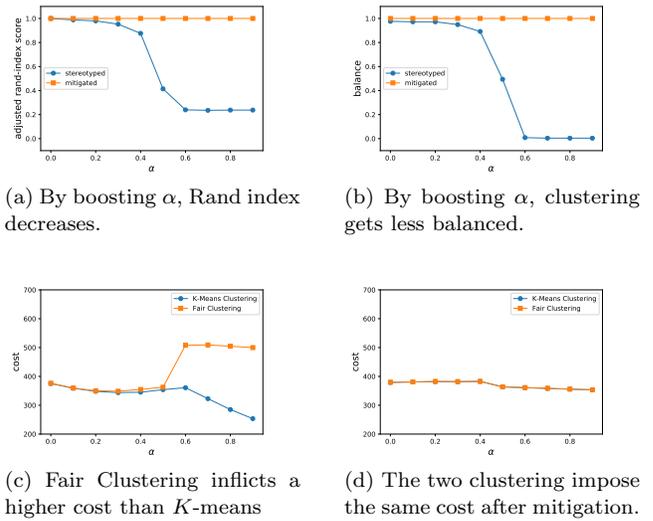


Figure 6: Clustering in the presence of stereotyping.

lead to skewed outcomes when part of a machine learning pipeline. We also presented mitigation strategies for these stereotype definitions and demonstrated via experiments on synthetic data that these strategies could largely remove the stereotype effects added to the data.

There are many aspects of stereotyped data that our approach does not include. One might assume that stereotyping (via exemplar or representativeness) might act more weakly on some individuals (e.g., celebrities) than others. Extensions to this framework that allow the stereotyping effect to act on a subset of the unprivileged group or that allow variations in α based, e.g., on the distance from the exemplar would be interesting to consider. We consider only a single exemplar, rather than many. Additionally, while there has been work validating the representativeness model of stereotyping in human subjects [5], the specific geometric model of stereotyping via exemplars that we consider here has not been similarly validated. There are also other cognitive models of stereotyping that we did not consider.

References

- [1] K. Arrow. The theory of discrimination. *Discrimination in labor markets*, 3(10):3–33, 1973.
- [2] G. S. Becker. The economics of discrimination, 1957.
- [3] S. K. Bera, D. Chakrabarty, and M. Negahbani. Fair algorithms for clustering. *arXiv preprint arXiv:1901.02393*, 2019.
- [4] T. Bolukbasi, K.-W. Chang, J. Zou, V. Saligrama, and A. Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Proc. 30th NIPS*, NIPS’16, pages 4356–4364, 2016.

- [5] P. Bordalo, K. Coffman, N. Gennaioli, and A. Shleifer. Stereotypes. *The Quarterly Journal of Economics*, 131(4):1753–1794, 2016.
- [6] M.-E. Brunet, C. Alkalay-Houlihan, A. Anderson, and R. Zemel. Understanding the Origins of Bias in Word Embeddings. *arXiv:1810.03611 [cs, stat]*, Oct. 2018. arXiv: 1810.03611.
- [7] T. Calders and S. Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining journal; special issue with selected papers from ECML/PKDD*, 2010.
- [8] A. Caliskan, J. J. Bryson, and A. Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, Apr. 2017.
- [9] N. Cantor and W. Mischel. Prototypes in person perception. In *Advances in experimental social psychology*, volume 12, pages 3–52. Elsevier, 1979.
- [10] M. Cardwell. *The Dictionary of Psychology*. Taylor & Francis, 1999.
- [11] S. Chatterjee and A. S. Hadi. *Sensitivity analysis in linear regression*, volume 327. John Wiley & Sons, 2009.
- [12] F. Chierichetti, R. Kumar, S. Lattanzi, and S. Vassilvitskii. Fair clustering through fairlets. In *Proc. NIPS*, pages 5029–5037, 2017.
- [13] K. Crawford. The trouble with bias. https://www.youtube.com/watch?v=fMym_BKWQzk, December 2017.
- [14] M. De-Arteaga, A. Romanov, H. Wallach, J. Chayes, C. Borgs, A. Chouldechova, S. Geyik, K. Kenthapadi, and A. T. Kalai. Bias in Bios: A Case Study of Semantic Representation Bias in a High-Stakes Setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19*, pages 120–128, New York, NY, USA, 2019. ACM.
- [15] J. A. Diaz-Garcia, G. González-Farías, and V. M. Alvarado-Castro. Exact distributions for sensitivity analysis in linear regression. *Applied Mathematical Sciences*, 1(22):1083–1100, 2007.
- [16] H. Edwards and A. Storkey. Censoring Representations with an Adversary. *arXiv:1511.05897 [cs, stat]*, Nov. 2015. arXiv: 1511.05897.
- [17] M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. Certifying and removing disparate impact. *Proc. of KDD*, pages 259–268, 2015.
- [18] S. A. Friedler, C. Scheidegger, and S. Venkatasubramanian. On the (im) possibility of fairness. *arXiv preprint arXiv:1609.07236*, 2016.
- [19] M. Hardt, E. Price, N. Srebro, et al. Equality of opportunity in supervised learning. In *Proc. NIPS*, pages 3315–3323, 2016.
- [20] J. L. Hilton and W. Von Hippel. Stereotypes. *Annual review of psychology*, 47(1):237–271, 1996.
- [21] L. Hubert and P. Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, Dec 1985.
- [22] D. Kahneman and A. Tversky. Subjective probability: A judgment of representativeness. *Cognitive psychology*, 3(3):430–454, 1972.
- [23] D. Kahneman and A. Tversky. On the psychology of prediction. *Psychological review*, 80(4):237, 1973.
- [24] T. Kamishima, S. Akaho, H. Asoh, and J. Sakuma. Fairness-aware classifier with prejudice remover regularizer. *Machine Learning and Knowledge Discovery in Databases*, pages 35–50, 2012.
- [25] D. Katz and K. W. Braly. Verbal stereotypes and racial prejudice. *Journal of abnormal and social psychology*, 28:280–290, 1933.
- [26] D. Madras, E. Creager, T. Pitassi, and R. Zemel. Learning adversarially fair and transferable representations. Technical report, arXiv preprint arXiv:1802.06309, 2018.
- [27] M. Manis, T. E. Nelson, and J. Shedler. Stereotypes and social judgment: Extremity, assimilation, and contrast. *Journal of Personality and Social Psychology*, 55(1):28, 1988.
- [28] M. Olfat and A. Aswani. Convex Formulations for Fair Principal Component Analysis. *arXiv:1802.03765 [cs, math, stat]*, Feb. 2018. arXiv: 1802.03765.
- [29] A. Romei and S. Ruggieri. A multidisciplinary survey on discrimination analysis. *The Knowledge Engineering Review*, pages 1–57, April 3 2013.
- [30] S. Samadi, U. Tantipongpipat, J. H. Morgenstern, M. Singh, and S. Vempala. The Price of Fair PCA: One Extra dimension. In *Proc. NeurIPS*, pages 10999–11010, 2018.
- [31] M. Schmidt, C. Schwiegelshohn, and C. Sohler. Fair coresets and streaming algorithms for fair k-means clustering. *arXiv preprint arXiv:1812.10854*, 2018.
- [32] E. R. Smith and M. A. Zarate. Exemplar-based model of social judgment. *Psychological review*, 99(1):3, 1992.
- [33] R. Speer. Conceptnet numberbatch 17.04: better, less-stereotyped word vectors. <http://bit.ly/speer-concept>, Apr 2017.
- [34] D. Statt. *The concise dictionary of psychology*. Routledge, 2002.
- [35] L. Sweeney. Discrimination in online ad delivery. *Queue*, 11(3):10, 2013.
- [36] A. Tversky and D. Kahneman. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157):1124–1131, 1974.
- [37] I. Žliobaitė. A survey on measuring indirect discrimination in machine learning. *arXiv preprint arXiv:1511.00148*, 2015.
- [38] Wikipedia contributors. Woodbury matrix identity. https://en.wikipedia.org/w/index.php?title=Woodbury_matrix_identity&oldid=856202501, 2018.
- [39] M. B. Zafar, I. Valera, M. G. Rodriguez, and K. P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. *Proc. of WWW*, pages 1171–1180, 2017.
- [40] R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. Learning fair representations. In *Proc. of Intl. Conf. on Machine Learning*, pages 325–333, 2013.