The Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)

Eliminating Latent Discrimination: Train Then Mask

Soheil Ghili[*]
School of Management
Yale University

Ehsan Kazemi[*†]
Yale Institute for Network Science
Yale University

Amin Karbasi
Yale Institute for Network Science
Yale University

Abstract

How can we control for latent discrimination in predictive models? How can we provably remove it? Such questions are at the heart of algorithmic fairness and its impacts on society. In this paper, we define a new operational fairness criteria, inspired by the well-understood notion of omitted variable-bias in statistics and econometrics. Our notion of fairness effectively controls for sensitive features and provides diagnostics for deviations from fair decision making. We then establish analytical and algorithmic results about the existence of a fair classifier in the context of supervised learning. Our results readily imply a simple, but rather counter-intuitive, strategy for eliminating latent discrimination. In order to prevent other features proxying for sensitive features, we need to include sensitive features in the training phase, but exclude them in the test/evaluation phase while controlling for their effects. We evaluate the performance of our algorithm on several real-world datasets and show how fairness for these datasets can be improved with a very small loss in accuracy.

1 Introduction

Nowadays, many sensitive decision-making tasks rely on automated statistical and machine learning algorithms. Examples include targeted advertising, credit scores and loans, college admissions, prediction of domestic violence, and even investment strategies for venture capital groups. There has been a growing concern about errors, unfairness, and transparency of such mechanisms from governments, civil organizations and research societies (White House 2016; Barocas and Selbst 2016; ProPublica 2018). That is, whether or not we can prevent discrimination against protected groups and attributes (e.g., race, gender, etc). Clearly, training a machine learning algorithm with the standard aim of loss function minimization (i.e., high accuracy, low prediction error, etc) may result in predictive behaviors that are unfair towards certain groups or individuals (Hardt et al. 2016; Liu et al. 2018; Zhang and Bareinboim 2018).

In many real-world applications, we are not allowed to use some sensitive features. For example, EU anti-discrimination law prohibits the use of protected attributes (directly or indirectly) for several decision-making tasks (Ellis and Watson 2012). A naive approach towards fairness is to discard sensitive attributes from training data. However, if other (seemingly) non-sensitive variables are correlated with the protected ones, the learning algorithm may use them to *proxy for protected features* in order to achieve a lower loss.[1] We call this a *latent* form of discrimination. Mitigating this kind of latent discrimination has received considerable attention in the machine learning community and interesting heuristic algorithms have been proposed (e.g., Zemel et al. (2013) and Kamiran and Calders (2009)). Since this type of discrimination is latent, most previous works fail to provide an operational description of this notion and usually resort to descriptive statements.

The first contribution of this paper is to propose a new operational definition of fairness, called *EL-fairness* (it stands for explicit and latent fairness), that controls for sensitive features. This definition rules out explicit discrimination in the conventional way by treating individuals with similar non-sensitive features similarly. It also provides a detection mechanism for observing latent discrimination of a classifier by comparing simple statistics within protected-feature groups to the ones provided by the optimum unconstrained classifier (trained on the full data set).

Proxying or *omitted variable bias* (OVB) occurs when a feature which is correlated with some other attributes is left out. In many models, for example, linear regression, it is well known that provided enough data, keeping the sensitive feature controls for OVB which enables us to separate its effect from other correlated attributes (Seber and Lee 2012). Building on our notion of EL-fairness and existing methods to remove/reduce the proxying effect or OVB (e.g., Žliobaitė and Custers (2016)), we develop a procedure for obtaining fair classifiers. In particular, we show that in order to eliminate latent discrimination one needs to consider the sensitive features in the training phase (in order to obtain reliable statistics to control for such features) and then mask them in evaluation/test phase. This way, we can ensure that correlated variables do not proxy the sensitive features and, more impor-

[*]The first two authors have equally contributed to this paper.
[†]Corresponding author.

[1]In Section 6 we observe that in many datasets: (i) the admission rate is hugely against the protected groups, and (ii) there are several features that are tightly correlated with the sensitive attribute.

tantly, decisions are not made based on protected attributes. Furthermore, such a train-then-mask approach achieves EL-fairness with almost no additional computational cost as the training phase is intact.

More specifically, in this paper we make two algorithmic contributions: (i) keep the sensitive feature during the training phase to control for OVB, and (ii) find an EL-fair classifier with the maximum accuracy by choosing the parameters of our algorithm properly. We use this idea to control for OVB in a general class of separable functions.[2]

As a final note, we should point out that our notion of fairness is robust against *double discrimination*. This is a peculiar situation (that happens surprisingly often) where a minority group outperforms the rest of the population despite discrimination. We show that our proposed procedure still removes the bias against the protected group in such scenarios, while group-fairness based notions do not.

The rest of this paper is organized as follows. In Section 2, we review the related literature. In Section 3, we define our notion of EL-fairness. In Section 4, we characterize the existence and properties of the optimal fair classifier and explain the train-then-mask algorithm. In Section 5, we discuss the relation of EL-fairness with double-unfairness and separability. In Section 6, we perform an exhaustive set of empirical studies to establish that our proposed approach reliably reduces latent discrimination with little loss in accuracy.

2 Related Work

This paper brings together pieces of literature from econometrics and machine learning. It is well known in both fields that if a variable is omitted from an analysis (on purpose or because it is unobservable), it might distort the results of the analysis in case it is correlated with variables that are not omitted (Greene 2003; Dwork et al. 2018; Kamiran and Calders 2009; Dwork et al. 2012; Hardt et al. 2016). In the econometrics literature, this concern arises mainly in the context of *causal inference* where the main objective is to estimate a *treatment effect*. There, if there is a factor that (i) impacts the outcome, (ii) is omitted from the analysis, and (iii) is correlated with the treatment, it can bias the estimated treatment effect, since (part of) the effect of the unobserved variable may be picked up by the estimation process as the effect of the treatment. This is generally called the *Omitted Variable Bias* (Greene 2003). The typical solution is to incorporate such variables as controls in the statistical model. This is an integral part of empirical and experimental research in multiple fields such as econometrics, marketing, and medicine (Sudhir 2001; Clarke 2005; Scheffler, Brown, and Rice 2007; Hendel and Nevo 2013; Ghili 2016).

However, the same strategy (i.e., controls) has not been explicitly used in the context of fairness in machine learning. This, partly stems from the fact that the objective functions are more complicated in the real world applications that machine learning algorithms aim to solve. On the one hand, it is not desirable if omitting a sensitive feature leads to latent discrimination, since correlated (and seemingly) non-sensitive features now can act as proxies for the sensitive one (Hardt et al. 2016; Pedreshi, Ruggieri, and Turini 2008). Unlike the causal inference literature, however, this problem may not be resolved by incorporating the sensitive feature in the analysis. This would eliminate latent discrimination but would come at the larger expense of explicit discrimination, i.e., the model might treat similar individuals of two different groups differently. The approaches suggested in the fairness literature to deal with this problem have either been mainly based on relabeling the data (Kamiran and Calders 2009) or based on mapping the data to a set of prototypes (Dwork et al. 2012). These approaches attempt to eliminate latent discrimination by directly or indirectly entering a notion of group fairness into the objective function of the optimization problem. That is, for instance, they try to achieve high admission accuracy but restrict the ratio of the number of admitted individuals form the unprotected group over admitted ones form the protected group.

More recently, Kilbertus et al. (2017) and Nabi and Shpitser (2018) framed the problem of fairness based on sensitive features in the language of causal reasoning in order to resolve the effect of proxy variables. Their focus is on the theoretical analysis of cases in which the full causal relationships among all (sensitive and nonsensitive) features are precisely known. In addition, Zhang and Bareinboim (2018) proposed a causal explanation formula to quantitatively evaluate fairness. We should point out that while the structure of causality could be learned by data generating models (e.g., in some special cases under certain linearity assumptions), our approach does not require such information.

Furthermore, several studies (such as the works of Hu and Chen (2018) and Liu et al. (2018)) consider the long-term effect of classification on different groups in the population. For another instance, Jabbari et al. (2017) investigated the long-term price of fairness in reinforcement learning. Similarly, Gillen et al. (2018) considered the fairness problem in online learning scenarios where the main objective is to minimize a game theoretic notion of regret. Also, fairness is studied in many other machine learning settings, including ranking (Celis, Straszak, and Vishnoi 2018), personalization and recommendation (Celis and Vishnoi 2017; Kamishima et al. 2018; Burke, Sonboli, and Ordonez-Gauger 2018), data summarization (Celis et al. 2018), targeted advertisement (Speicher et al. 2018), fair PCA (Samadi et al. 2018), empirical risk minimization (Donini et al. 2018; Hashimoto et al. 2018), privacy preserving (Ekstrand, Joshaghani, and Mehrpouyan 2018) and a welfare-based measure of fairness (Heidari et al. 2018). Finally, due to the massive size of today's datasets, practical algorithms with fairness criteria should be able to scale. To this end, Grgic-Hlaca et al. (2018) and Kazemi, Zadimoghaddam, and Karbasi (2018) have developed several scalable methods with the aim of preserving fairness in their predictions.

Our approach has important implications for other notions of fairness such as group fairness and individual fairness. It is well-known that there are inherent trade-offs among different notions of fairness and therefore satisfying multiple

[2]Note that linear and logistic regressions are two simple members of this general class which we define later.

fairness criteria simultaneously is not possible (Kleinberg, Mullainathan, and Raghavan 2017; Pleiss et al. 2017). For example, all methods that aim at solving the issue of proxying (including ours) do not satisfy the calibration property.

3 Setup and Problem Formulation

Let $\mathbf{X} \in \mathcal{X} \subset \mathbb{R}^{\ell+1}$ be a random variable with $\ell+1$ dimensions $\mathbf{X}_0$ through $\mathbf{X}_\ell$. That is, each sample draw $\mathbf{x}^i$ has $\ell+1$ real-valued components $\mathbf{x}_0^i$ through $\mathbf{x}_\ell^i$ where the dimensions are possibly correlated. Dimension $\mathbf{X}_0$ is binary and represents the status of the sensitive feature. For example, when the sensitive feature is gender, 1 represents female and 0 represents male. In this paper, we consider the binary classification problem, where we assume that there is a binary label y^i for each data point $\mathbf{x}^i$, i.e., the set of possible labels y^i is denoted by $\mathcal{Y} \in \{0,1\}$. We are given n training samples $z^1, \cdots, z^n$, where $z^i = (\mathbf{x}^i, y^i) \in \mathcal{X} \times \mathcal{Y}$.

Mathematically, a classifier is a function $h : \mathbb{R}^{\ell+1} \to [0,1]$ from a set of hypothesis (possible classifiers) $\mathcal{H}$, where each input sample $\mathbf{x} \in \mathbb{R}^{\ell+1}$ is mapped to a value in the interval $[0,1]$; a data point $\mathbf{x}$ is classified to 1 if $h(\mathbf{x}) > 1/2$, and to 0 otherwise. The ultimate goal of a classification task is to optimize some loss function $\mathcal{L}(y, h(\mathbf{x}))$ over all possible functions $h \in \mathcal{H}$, when applied to the training set. We denote by h^* the classifier that minimizes this loss function.[3] In other words, h^* is the most accurate classifier from the set $\mathcal{H}$ of functions, where all information–including sensitive feature $\mathbf{x}_0$–is used to achieve the highest accuracy.[4] Next, we turn to our fairness definition, articulating first the explicit dimension, then the latent one.

Definition 1 (Explicit Discrimination)**.** *Classifier h exhibits no explicit discrimination if for every pair $(\mathbf{x}^1, \mathbf{x}^2) \in \mathcal{X}^2$ such that $(\mathbf{x}_1^1, ..., \mathbf{x}_\ell^1) = (\mathbf{x}_1^2, ..., \mathbf{x}_\ell^2)$, regardless of $\mathbf{x}_0^1$ and $\mathbf{x}_0^2$ (i.e., the status of the sensitive features) we have $h(\mathbf{x}^1) = h(\mathbf{x}^2)$.*

Definition 1 captures the simple and conventional way of thinking about explicit discrimination: a fair classifier should treat two similar individuals (irrespective of their sensitive features) similarly. Latent discrimination is, however, less trivial to formally capture. Thus, we diagnose latent discrimination based on a subtle indirect implication that it has. We first give the formal definition and then discuss the diagnostic intuition behind it.

Definition 2 (Latent Discrimination)**.** *Classifier h exhibits no latent discrimination if for every pair $(\mathbf{x}^1, \mathbf{x}^2) \in \mathcal{X}^2$ such that $\mathbf{x}_0^1 = \mathbf{x}_0^2$ (i.e., pairs with similar sensitive features) we have*:

$$h^*(\mathbf{x}^1) = h^*(\mathbf{x}^2) \Rightarrow h(\mathbf{x}^1) = h(\mathbf{x}^2), \text{ and} \quad (1)$$

$$h^*(\mathbf{x}^1) > h^*(\mathbf{x}^2) \Rightarrow h(\mathbf{x}^1) \geq h(\mathbf{x}^2). \quad (2)$$

[3] We do not make any assumption regarding how the class $\mathcal{H}$ and/or loss function $\mathcal{L}$ should be chosen. Our approach guarantees that given a class and loss function, we can always design an EL-fair classifier.

[4] Through the whole paper, we define h^* to be the classifier from class $\mathcal{H}$ that minimizes the empirical loss. In many practical settings, we can find h^* in polynomial time.

In words, Definition 2 says that flipping the order of the classes of two individuals of the *same group* compared to h^* is a sign of latent discrimination. To see the intuition behind this definition, consider $\bar{h}$, representing the most accurate classifier that satisfies Definition 1 (i.e., it minimizes the loss function subject only to explicit non-discrimination). Here, by minimizing the loss function, we would ideally like to get as close as possible to h^*, but that is not generally possible given the constraint that the information about $\mathbf{x}_0$ may not be used. Thus, the minimizer would potentially treat the other ℓ features differently than h^* does in order to proxy for the missing $\mathbf{x}_0$ attribute. This proxying, however, inevitably changes how the classifier treats individuals within the same group, possibly by flipping the orders between some pairs. This is exactly what we call latent discrimination that we would like to control for. Definition 2 formalizes this idea in a very operational manner. Indeed, in Definition 2 we argue that the optimal unconstrained classifier provides a non-discriminatory ordering between individuals within each group. In other words, if $h^*(\mathbf{x}^1) > h^*(\mathbf{x}^2)$ for $\mathbf{x}_0^1 = \mathbf{x}_0^2$, we can conclude $\mathbf{x}^1$ is more qualified than $\mathbf{x}^2$. If a classifier h changes this ordering, then it could be a sign of latent discrimination. We are now equipped with the following definition for fairness.

Definition 3 (EL-fair)**.** *Classifier h is "EL-fair" if it exhibits neither explicit nor latent discrimination as described in Definitions 1 and 2.*

Note that h^* might not be EL-fair because it could suffer from explicit discrimination as it uses all features.

4 Characterization of Optimal Fair Classifier

With a formal definition of fairness in hand, we turn to the next natural step:

> What are the characteristics of an optimal classifier that satisfies EL-fairness condition?

While there is not a trivial answer to this question, in this section we show, however, that our notion of fairness lends itself into a practical algorithmic framework with the following properties. First, the computation of the optimal fair classifier is straightforward. In fact, it is not more complicated than computing the optimal unconstrained classifier h^*. Second, it provides an intuitive interpretation in line with the idea of *controlling for* different factors traditionally used in fields such as statistics and econometrics. Our first theoretical result establishes the existence of an EL-fair classifier. Then, in Theorem 2, we characterize the optimal classifier under fairness constraints of Definition 3. Finally, in Theorems 3 and 4, we outline the properties of a simple algorithm that computes the optimal EL-fair classifier.

Theorem 1. *An EL-fair classifier exists if the set $\mathcal{H}$ (set of all possible functions in our model) includes at least one constant function.*

Note that (almost) all practical models used in machine learning (e.g., logistic, linear, neural net, etc) allow for constant functions, therefore, they include an EL-fair classifier. We next turn to the characterization of the optimal fair classifier. But before that, we need to give a definition that (i) is

necessary for the statement of the theorem; and (ii) as we argue in Section 5, is conceptually crucial to the understanding of individual fairness.

Definition 4 (A separable classifier)**.** *Classifier h is "separable in the sensitive feature" if there are continuous functions $g : \mathbb{R}^2 \to \mathbb{R}$ and $K : \mathbb{R}^\ell \to \mathbb{R}$ such that: $\forall \mathbf{x} \in \mathcal{X}$ we have $h(\mathbf{x}) = g\left(\mathbf{x}_0, K(\mathbf{x}_1, \cdots, \mathbf{x}_\ell)\right)$.*

A wide range of classifiers satisfy this intuitive definition. For instance, any logistic model can be represented by choosing an appropriate linear function for K and choosing $g(z_1, z_2) \equiv \frac{e^{z_1+z_2}}{1+e^{z_1+z_2}}$. Later in the paper, we discuss the close ties between the notions of separability and individual fairness. For now, we state our main result.

Theorem 2. *Suppose the unconstrained optimal classifier h^* satisfies the definition of separability with a given g. Denote by h^*_{fair} the optimal classifier (in terms of accuracy) subject to EL-fairness criteria as described in Definition 3. There is a $\tau^* \in \mathbb{R}$ such that for all $\mathbf{x} = (\mathbf{x}_0, \mathbf{x}_1, ..., \mathbf{x}_\ell) \in \mathcal{X}$:*

- *if $h^*_{\text{fair}}(\mathbf{x}) > \frac{1}{2}$ then $h^*(0, \mathbf{x}_1, \cdots, \mathbf{x}_\ell) + \tau^* > \frac{1}{2}$.*
- *if $h^*_{\text{fair}}(\mathbf{x}) < \frac{1}{2}$ then $h^*(0, \mathbf{x}_1, \cdots, \mathbf{x}_\ell) + \tau^* < \frac{1}{2}$.*

Theorem 2 demonstrates that for a properly chosen τ^*,[5] there is an $h^*(0, \mathbf{x}_1, ..., \mathbf{x}_\ell) + \tau^*$ that *mimics* the optimal fair classifier h^*_{fair} by recommending all the decisions that h^*_{fair} would recommend. In Theorem 3 we prove that, under a mild assumption, such an $h^*(0, \mathbf{x}_1, ..., \mathbf{x}_\ell) + \tau^*$ classifier is also EL-fair.

Theorem 3. *If the function $h^* \in \mathcal{H}$ is separable in the sensitive feature $\mathbf{x}_0$, i.e., there is a function $g : \mathbb{R}^2 \to \mathbb{R}$ such that $h^*(\mathbf{x}) = g(\mathbf{x}_0, K(\mathbf{x}_1, ..., \mathbf{x}_\ell))$, and the function g is strictly monotone in its second argument, then all classifiers of the form $h^*_\tau \triangleq h^*(0, \mathbf{x}_1, \cdots, \mathbf{x}_\ell) + \tau$ are EL-fair.*

Thus, the only further step to find the optimal EL-fair classifier, in addition to computing h^*, is to search for τ^*. Theorem 4 shows that when the function g is monotone, then searching for τ^* is quite straightforward.

Theorem 4. *Assume that the function h^* is separable, i.e., there is a function $g : \mathbb{R}^2 \to \mathbb{R}$ such that $h^*(\mathbf{x}) = g(\mathbf{x}_0, K(\mathbf{x}_1, ..., \mathbf{x}_\ell))$ and the function g is strictly monotone in its second argument. Furthermore, assume τ^* is the value of τ such that it maximizes the classification accuracy of $h^*_\tau \triangleq h^*(0, \mathbf{x}_1, \cdots, \mathbf{x}_\ell) + \tau$. The function $h^*_{\tau^*}$ is the optimal EL-fair classifier.*

The above property makes the search for an optimal EL-fair classifier practical. That is, no matter how large the dataset is, as long as h^* can be computed, h^*_{fair} can be too. We call this approach the *train-then-mask* algorithm for eliminating latent discrimination. Algorithm 1 describes train-then-mask.

In spite of the fact that our formal definition of fairness is indirect, that is it turns to *within-group* variation to capture a concept that is essentially only meaningful between groups,

[5]Note that the use of threshold $1/2$ is only for the purpose of exposition. All our theoretical results will hold if the threshold is chosen adaptively.

Algorithm 1 The Train-Then-Mask Algorithm

1: Compute the optimal classifier $h^*(\mathbf{x}_0, \mathbf{x}_1, \cdots, \mathbf{x}_\ell)$ over all available features $\mathbf{x}_0, \mathbf{x}_1, \cdots,$ and $\mathbf{x}_\ell$.
2: Keep the sensitive feature $\mathbf{x}_0$ fixed (e.g., define $\mathbf{x}_0 = 0$) for all data points.
3: Find the value of τ^* such that it maximizes the accuracy of $h^*(0, \mathbf{x}_1, \cdots, \mathbf{x}_\ell) + \tau^*$ over the validation set.

Theorem 2 provides an intuitive characterization. Basically, to prevent other variables from proxying a sensitive feature, we must *control for* the sensitive feature when estimating the parameters that capture the importance of other nonsensitive variables. Crucially, the sensitive feature should not be left out of the model *before* training. In contrast, we do not want the sensitive feature to impact our prediction/evaluation when all else is equal (to ensure individual or explicit fairness). This is why the sensitive feature does eventually need to be excluded *after* training. Theorems 2 to 4 connect the less intuitive Definition 3 to this simple and established algorithmic procedure.

Generalization to a set of sensitive features: In many applications, there might be more than one sensitive feature (e.g., both gender and race might be present). It is straightforward to generalize our framework for such cases. All of the definitions, theorems, algorithms, and interpretations remain intact if instead of $\mathbf{x}_0 \in \mathbb{R}$ we assume $\mathbf{x}_0 \in \mathbb{R}^m$ for some $m \in \mathbb{N}$, where m is the number of sensitive features. Thus, our framework accommodates multiple sensitive features. More specifically, to apply our method we first train the model on all features. In the prediction step, we keep all the sensitive attributes fixed for all data points (e.g., if the sensitive features are age and gender we assume all people are young and female). The value of τ^* is then chosen in the way to maximize the accuracy on the validation set.

5 Discussion

In this section, we further discuss several important features of our proposed fairness notion and the algorithmic solution. In particular, (i) we overview the relationship with the important concept of group fairness, and (ii) we further elaborate on the significance of the separability property. We also argue that separability is a central notion in understanding the individual fairness property.

5.1 Relationship with Group Fairness

Unlike other suggested solutions to the problem of proxying, our approach does not incorporate some notion of group fairness to alleviate this issue. For example Kamiran and Calders (2009) suggested *massaging* the training set in order to exhibit group fairness, or Zemel et al. (2013) directly incorporated group fairness into the loss function. Although in Section 6 we show that our model performs well on the group fairness measure, it has not been directly incorporated into the objectives of our model. The reason we avoid mixing group fairness with the problem of proxying (which is essentially a matter of individual fairness) is the potential for

what we call *double unfairness*, a concept which we discuss below.

Double unfairness can happen when the protected group performs better than the unprotected group in spite of the discrimination. For instance, consider a dataset on college admissions with two groups $\mathcal{A}$ (the protected group) and $\mathcal{B}$ (the unprotected group): (i) A person from group $\mathcal{A}$, on average, has a lower chance of admission to the college compared to a person from group $\mathcal{B}$ with the same SAT score and extracurricular activities; (ii) Nevertheless, group $\mathcal{A}$ does better than group $\mathcal{B}$ on the SAT by a wide enough margin that on average the admission rate for $\mathcal{A}$ is higher than that for $\mathcal{B}$. The following synthesized dataset (see Table 1) illustrates an example for this potential scenario.

Table 1: Toy example: for the sensitive attribute, 1 represents the protected group $\mathcal{A}$ and 0 represents group $\mathcal{B}$.

ID	Admission	Sensitive	SAT	Extracurricular
1	1	1	1600	4
2	1	1	1500	6
3	1	1	1500	4
4	0	1	1400	6
5	1	0	1400	6
6	1	0	1300	5
7	0	0	1200	4
8	0	0	1200	4

It can be seen, from Table 1, that the admission process has been unfair to applicants from the protected group $\mathcal{A}$. Candidates 4 and 5 are identical with the sole exception that candidate 4 is from group $\mathcal{A}$ and 5 is from group $\mathcal{B}$. Candidate 4 has been denied but 5 has been admitted.[6] On the other hand, group $\mathcal{A}$ performs better than group $\mathcal{B}$ since they have an acceptance rate of $3/4$ while that of group $\mathcal{B}$ is only $1/2$. Thus, group $\mathcal{A}$ does on average 50% better. Intuitively, if we are to alleviate the discrimination against the protected group, we should expect a classifier that gives even a higher edge than 50% to them.

One can verify the danger of proxying in this toy example when the sensitive feature is omitted, by giving a higher (positive) weight to extracurricular activities compared to SAT score. This would happen because SAT has a higher correlation with the sensitive feature. Our approach does not allow for such weight adjustments since, by Theorem 2, it controls for the sensitive feature when training the rest of the weights. In doing so, train-then-mask gives a higher edge than the original 50% to group $\mathcal{A}$ in terms of admission ratio. This provides an advantage over approaches that tackle the problem of proxying by forcing a notion of group fairness. For instance, the methodology by Kamiran and Calders (2009) would first massage the data by relabeling applicant 3 to denied (or applicant 7 to admitted) and then train the classifier. The algorithm would do this to equate the admission rates between the two groups. Thus, this algorithm tries to get the acceptance rates of group $\mathcal{A}$ closer to that of $\mathcal{B}$. This, clearly, will only further discriminate candidates from the protected group. Similar concerns exist about other approaches that somehow employ a notion of group fairness to address proxying.

[6] A more precise way to detect unfairness against the protected group would be to run a model (such as linear regression) on the data and observe that the coefficient on the sensitive feature is negative. This means that, on aggregate, candidates from the protected group are treated worse than similar candidates from the other group.

5.2 Separability and Individual Fairness

At the heart of the sufficient conditions for Theorem 2 is the separability of function f between the sensitive features and all other features. What separability roughly says is that using a separable classifier, one can rank two individuals of the same group without knowing what their (common) sensitive feature is. In this section, we argue that our notion of fairness introduces separability as a central concept in the understanding of individual fairness; and sheds light on important future research directions.

Note that the separability of h^* is a sufficient (and not a necessary) condition for the existence of an optimal EL-fair classifier. In Theorem 5, we show that under a slightly stronger notion of fairness, if there is a fair classifier then the separability property is also necessary.

Definition 5 (Strictly EL-fair)**.** *Classifier h satisfies strong fairness criteria if it satisfies Definitions 1 and 2 but instead of Eq.* (2) *in Definition 2, it satisfies*: $h^*(\mathbf{x}^1) > h^*(\mathbf{x}^2) \Rightarrow h(\mathbf{x}^1) > h(\mathbf{x}^2)$.

Theorem 5. *Suppose there is a classifier that satisfies the strictly EL-fairness notion. Then, the function h^* is separable in the sense of definition 4, and the corresponding function $g : \mathbb{R}^2 \to \mathbb{R}$ of the separable representation is strictly monotone in its second argument.*

To see the intuition, consider an h^* that does not satisfy the separability and monotonicity properties: suppose the impact of a nonsensitive feature $\mathbf{x}_i$ on the outcome of h^*depends on $\mathbf{x}_0$, e.g., for the protected group (i.e., when $\mathbf{x}_0 = 1$) larger values of $\mathbf{x}_i$ results in a higher chance of positive classification (and vice versa). This means that the ordering implied under $\mathbf{x}_0 = 0$ is different from the ordering under $\mathbf{x}_1 = 1$. As a result, there is no classifier that satisfies the required corresponding orderings withing both groups.

The concept of separability provides a lens through which we can systematically think about some of the recent papers on fairness. For instance, Dwork et al. (2018) propose a decoupling technique which, although focused mainly on group fairness, is motivated precisely by the fact that the weight of a factor on the outcome might have different signs for different groups. It is important to note that we do not claim one should only use separable models (even if not appropriate in the context) to ensure EL-fairness. Indeed, we argue that under the non-separability assumption: (i) our method for detecting latent discrimination does not work, and (ii) by using currently existing methods several other problems arise (explained in other works such as (Dwork et al. 2018)).

We close this discussion by mentioning a few open questions. The first is to consider a novel methodology for measuring the *degree of non-separability* for general classifiers. Another important question is to detect latent discrimination

in non-separable environments and to design algorithms to ensure EL-fairness in these cases. Finally, we need a measure to identify the extent of proxying and a strategy that efficiently trades off accuracy with fairness.

6 Experiments

In this section, we compare the performance of the train-then-mask algorithm to a number of baselines on real-world scenarios. In our experiments, we compare train-then-mask (i) to the unconstrained optimum classifier (i.e., the one that tries to maximize the accuracy without any fairness constraints), (ii) to a model in which only the sensitive feature has been removed from training procedure (note that this algorithm might suffer from the latent discrimination), (iii) to the trivial majority classifier which always predict the most frequent label, (vi) to a data massaging algorithm introduced by Kamiran and Calders (2009), and (v) to the algorithm for maximizing a utility function subject to the fairness constraint introduced by Zemel et al. (2013). In our experiments we consider linear SVMs (separable) (Schölkopf and Smola 2002) and neural networks (non-separable) for the family of classifiers $\mathcal{H}$. To find the value of τ^* for our optimal fair classifier, we use a validation set; we take the value of τ^* such that it maximizes the accuracy over validation set and then we report the result of classification over the test set.

Datasets: We use the *Adult Income* and *German Credit* datasets from UCI Repository (Asuncion and Newman 2007; Blake and Merz 1998), and COMPAS Recidivism Risk dataset (ProPublica 2018). Adult Income dataset contains information about 13 different features of 48,842 individuals and the labels identifying whether the income of those individuals is over 50K a year. The German Credit dataset consists of 1,000 people described by a set of 20 attributes labeled as good or bad credit risks. The COMPAS dataset contains personal information (e.g., race, gender, age, and criminal history) of 3,537 African-American and 2,378 Caucasian individuals. The goal of the classification tasks in these datasets is to predict, respectively, the income status, credit risks and whether a convicted individual commit a crime again in the following two years.

Measures: We use the following measures to evaluate the performance of algorithms. *Accuracy* measures the quality of prediction of a classifier over the test set. It is defined by $\text{Acc.} = 1 - \frac{\sum_{i=1}^{n} |y^i - \hat{y}^i|}{n}$, where n is the number of samples in the test set, y^i and $\hat{y}^i$ are the real and predicted labels of a test sample $\mathbf{x}^i$. *Admittance* measures the ratio of samples assigned to the positive class in each group. It is defined by $\text{Admit}_1 = \frac{\sum_{i:\mathbf{x}_0^i=1} \hat{y}^i}{\sum_{i:\mathbf{x}_0^i=1} 1}$ and $\text{Admit}_0 = \frac{\sum_{i:\mathbf{x}_0^i=0} \hat{y}^i}{\sum_{i:\mathbf{x}_0^i=0} 1}$. *Group discrimination* measures the difference between the proportion of positive classifications within each one of the protected and unprotected groups, i.e., $\text{G}_{\text{Discr.}} = |\text{Admit}_1 - \text{Admit}_0|$. *Latent discrimination* is defined as the ratio of pairs that violates Definition 2 to the total number of pairs in each group. More precisely, we have

$$\text{L}_{\text{Discr.}} = \frac{\left|h^*(\mathbf{x}^i) > h^*(\mathbf{x}^j), h(\mathbf{x}^i) < h(\mathbf{x}^j) | i \neq j, \mathbf{x}_0^i = \mathbf{x}_0^j\right|}{\binom{\sum_{i:\mathbf{x}_0^i=0} 1}{2} + \binom{\sum_{i:\mathbf{x}_0^i=1} 1}{2}}.$$

Consistency measures a (rough) notion of individual fairness by assuming the prediction for data samples that are close to each other should be (almost) similar. More precisely, it provides a quantitative way to compare the classification prediction of a model for a given sample $\mathbf{x}^i$ to the set of its k-nearest neighbors (denoted by $kNN(\mathbf{x}^i)$), i.e., $\text{kNN-Pred}(\mathbf{x}^i) = \frac{1}{k} \sum_{j \in kNN(\mathbf{x}^i)} \hat{y}^j$.

We should mention that the admittance ratios in all these datasets are always lower for the protected groups. For example, in the Adult Income dataset, while the income status of 31% of the male population is positive, this value is 11% for females. In addition, in all these datasets there are several attributes that are highly correlated with the sensitive feature. For example, in German Credit dataset, the correlation of the sensitive feature, i.e., "age", with "Present employment since", "Housing" and "Telephone" features are 0.24, 0.28 and 0.21, respectively.

We first consider the linear SVM classifiers which are separable. As shown in Table 2, train-then-mask represents the best performance in terms of removing the latent discrimination (see $\text{L}_{\text{Discr.}}$). Indeed, both discrimination measures are lower under train-then-mask than it is under the unconstrained model or the model in which the sensitive feature has been omitted. This demonstrates that train-then-mask indeed helps with the issue of proxying. More precisely, we observe that *omitting the sensitive feature* has lower accuracy but also lower discrimination compared to the unconstrained classifier.

We also observe that train-then-mask performs very well in reducing the group discrimination at the expense of a very little decrease in the accuracy. To see this, let us compare train-then-mask to the data massaging technique (Kamiran and Calders 2009). Under the Adult Income dataset, train-then-mask achieves higher accuracy than data-massaging but yields also higher $\text{G}_{\text{Discr.}}$. Under the German Credit dataset, it does better on both the accuracy and group discrimination fronts. These results, combined with the intuitive interpretation of our algorithm, as well as its straightforward computation, suggests that train-then-mask as an algorithm can be easily employed to alleviate (explicit and latent) discrimination in various datasets. This observation further demonstrates that although Definition 3 does not seem directly related to discrimination between groups, it does capture a symptom of latent discrimination.

We should point out that in our applications (and a lot of practical ones) the sensitive feature does indeed increase the accuracy of the model. Note that, for example, in the Adult Income dataset the accuracy of admitting all individuals, i.e., the trivial baseline classifier, is 0.756; thus going from 0.826 to 0.824 is not "negligible". Our main claim is not that we do not lose much accuracy. We argue that train-then-mask, compared to other approaches that aim at resolving proxying, does well. It sometimes offers both a higher accuracy and a lower discrimination than other approaches (i.e., it dominates them) and it is never dominated in our experiments by any other approach.

To investigate the effect of our algorithm on discrimination for non-separable classifiers we consider a neural network with three hidden layers. In Table 2, we observe that our

Table 2: Comparison between the performance of different algorithms on Adult, German and COMPAS datasets. Data massage algorithm refers to the method presented by (Kamiran and Calders 2009). The results of algorithms with the best performance on Acc., $G_{Discr.}$ and $L_{Discr.}$ are represented in blue. The unconstrained model, as we expect, in almost all cases is the most accurate classifier. Train-then-mask shows the best performance in terms of removing the latent discrimination (i.e., $L_{Discr.}$), while its accuracy is also close to the unconstrained model and even better in one case. In addition, while the main goal of Train-then-mask is to remove the effect of the latent discrimination, it also performs better than the other algorithms in terms of reducing the group discrimination. Note that for the algorithm of (Zemel et al. 2013) and Majority the discrimination metrics do not provide any meaningful information, because their outputs for all instances are always constant. We left out the $L_{Discr.}$ for the optimal unconstrained classifier h^* since this metric measures the latent discrimination with respect to h^* itself.

	Adult dataset					German dataset					COMPAS dataset				
Algorithm	Acc.	Adm_1	Adm_0	$G_{Discr.}$	$L_{Discr.}$	Acc.	Adm_1	Adm_0	$G_{Discr.}$	$L_{Discr.}$	Acc.	Adm_1	Adm_0	$G_{Discr.}$	$L_{Discr.}$
Unconstrained model h^* (SVM)	**0.825**	0.078	0.248	0.170	-	**0.75**	0.60	0.86	0.26	-	**0.768**	0.27	0.62	0.35	-
Omit sensitive feature (SVM)	0.824	0.080	0.243	0.163	0.016	0.73	0.64	0.87	0.23	0.016	0.765	0.34	0.56	0.22	0.005
Train-then-mask (SVM)	0.823	0.096	0.188	0.092	**0.000**	0.74	0.61	0.80	**0.19**	**0.000**	0.766	0.44	0.64	**0.20**	**0.000**
Data massage (SVM)	0.807	0.183	0.236	**0.053**	0.109	0.73	0.63	0.83	0.20	0.114	0.747	0.35	0.58	0.23	0.025
(Zemel et al. 2013)	0.756	0.000	0.000	0.000	0.000	0.67	1.00	1.00	0.00	0.000	0.509	1.00	1.00	0.00	0.000
Majority	0.756	0.000	0.000	0.000	0.000	0.67	1.00	1.00	0.00	0.000	0.509	1.00	1.00	0.00	0.000
Unconstrained model h^* (NN)	**0.825**	0.093	0.266	0.173	-	0.72	0.62	0.85	0.23	-	**0.767**	0.25	0.60	0.35	-
Omit sensitive feature (NN)	0.824	0.092	0.259	0.167	0.083	0.69	0.67	0.84	0.17	0.325	0.741	0.42	0.67	0.25	0.095
Train-then-mask (NN)	0.823	0.091	0.192	0.101	**0.058**	**0.73**	0.60	0.76	**0.16**	**0.264**	0.740	0.31	0.53	**0.22**	**0.076**
Data massage (NN)	0.808	0.183	0.247	**0.064**	0.146	0.70	0.62	0.81	0.19	0.391	0.738	0.39	0.63	0.24	0.098

algorithm performs well in reducing the discrimination while maintaining the accuracy for neural network classifiers.[7] It is important to point out that in neural networks because the classifiers are not separable and it is possible to have higher levels of proxying, the latent discrimination (i.e., $L_{Discr.}$) is also increased in comparison to the SVM classifier. Note that even though our theory holds only for separable classifiers, we find that our notion of fairness is relevant in other practical scenarios where classifiers are not separable.

Effect of Threshold τ: In Theorem 3, we showed that under certain conditions for all value of τ, the function $h^*(0, \mathbf{x}_1, \cdots, \mathbf{x}_\ell) + \tau$ is fair based on Definition 3. Theorem 4 argues that the value of τ^* that maximizes the accuracy has the same classification outcome as the optimal fair classifier. However, if in an application one is more interested in lowering discrimination than in accuracy, she may choose values of τ other than τ^* in order to fine-tune the accuracy-discrimination trade-off according to the specifics of the application. Fig. 1 illustrates this point on the Adult Income, German Credit, and COMPAS Recidivism Risk datasets. Each sub-figure consists of points in the accuracy-discrimination space where each point comes from a specific choice of τ. The orange part of each curve (circles) is the *Pareto frontier*. That is, one cannot choose a τ that does better on both accuracy and discrimination fronts than an orange colored point. Whereas any blue points (triangles) correspond to choices of τ that are dominated by one other choice of τ on both fronts. The black point (square) corresponds to the value of τ^* with the maximum accuracy on the validation set. Note that accuracy and discrimination values reported in Table 2 are for this choice of τ^*.

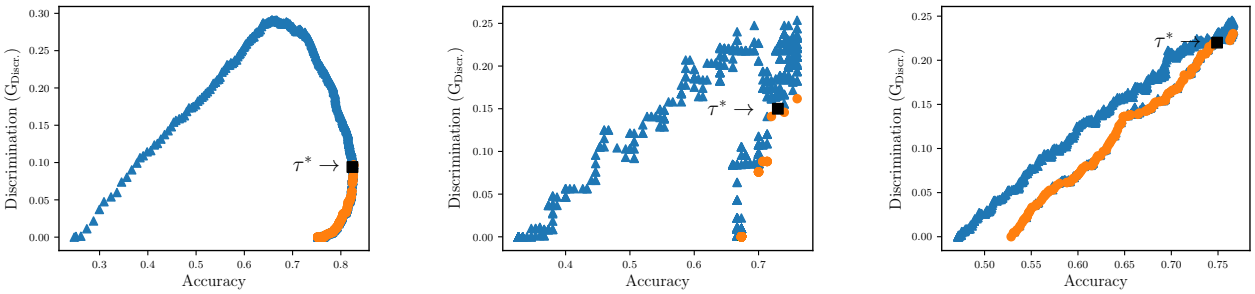
Consistency and Individual Fairness: In this part, we provide experimental evidence to support our claim regarding the individual fairness of our proposed classifier. For this reason, we compare the consistency in the output of our fair classifiers, i.e., for each sample $\mathbf{x}^i$, we compare the value of $\hat{y}^i$ to kNN-Pred($\mathbf{x}^i$). In Fig. 2, we observe that, not only for two data samples $\mathbf{x}^1$ and $\mathbf{x}^2$ such that $(\mathbf{x}_1^1, ..., \mathbf{x}_\ell^1) = (\mathbf{x}_1^2, ..., \mathbf{x}_\ell^2)$ the prediction is exactly the same (explicit individual fairness), for data samples which are close to each other (based on their Euclidean distances in the feature space) the predictions remain close.

Multiple Sensitive Features: In this part, we use an SVM classifier to evaluate the performance of our algorithm over multiple sensitive features. In the first experiment, we consider "sex" and "race" as two sensitive features in Adult Income dataset. In the prediction step, we assume all people are "female" and "black". The accuracy is 0.824 with $Adm_1 = 0.091$, $Adm_0 = 0.174$ and $G_{Discr.} = 0.083$ for "sex". For the second experiment, in the German Credit dataset, we consider "Personal status and sex" along with the "age" as sensitive attributes. Also, in the training step, we assume all individuals are "young" and "female and single". The accuracy is 0.71 with $Adm_1 = 0.78$, $Adm_0 = 0.86$ and $G_{Discr.} = 0.08$ for "age". For these experiments, when we naïvely omit the sensitive features $G_{Discr.}$ are, respectively, 0.170 and 0.196 for Adult Income and German Credit datasets.

7 Conclusion

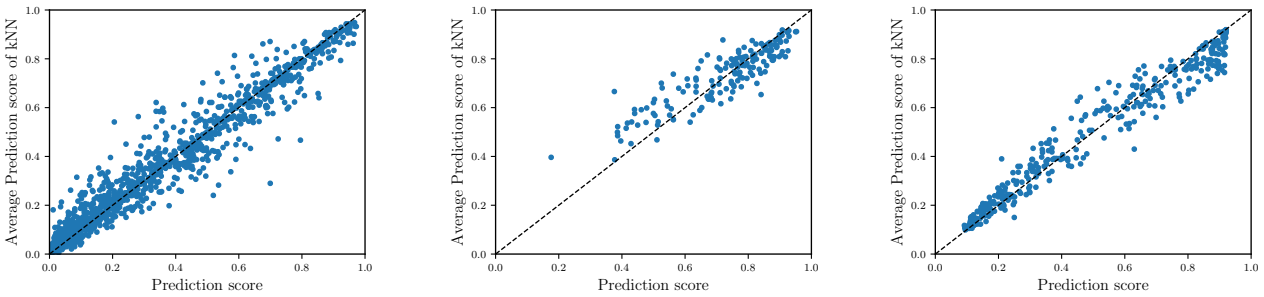
It is well established in the literature that simply omitting the sensitive feature from the model will not necessarily give a fair classifier. This is because often, nonsensitive features are correlated with the sensitive one and can act as proxies of that feature, bringing about latent discrimination. In spite of the consensus on the importance of latent discrimination and the attempts to eliminate it, no formal definition of it has been provided based on the notion of within-group fairness. Our main observation for providing an operational definition of latent discrimination relied on diagnosing this phenomenon by examining its symptoms. We argue that changing the

[7] Note that the algorithm of Zemel et al. (2013) does not depend on the choice of the family of classifiers $\mathcal{H}$.



(a) Adult Income dataset (b) German Credit dataset (c) COMPAS Recidivism dataset

Figure 1: The trade-off between accuracy and fairness for different values of τ. The orange points represent the Pareto frontier samples of τ, which means that one cannot do better than them in both fronts of accuracy and discrimination. The black point (square) corresponds to the value of τ^* with the maximum accuracy on the validation set.



(a) Adult Income dataset ($k = 10$) (b) German Credit dataset ($k = 5$) (c) COMPAS Recidivism dataset ($k = 5$)

Figure 2: The comparison between the prediction score of each data point and the average score of its k nearest neighbors in the test set. We set $\tau = 0$ in this experiment. Note that any other value of τ shifts both axes equally.

order of the values assigned to two samples within the same group compared to the optimal unconstrained classifier is a symptom of proxying, and we call a classifier free of latent discrimination if it does not exhibit any such disorders.

We demonstrated that our notion of fairness has multiple favorable features, making it suitable for analysis of individual fairness. First, we proved that the optimal fair classifier can be represented in a simple fashion. It enjoys an intuitive interpretation that the sensitive feature should be omitted after, rather than before training. This way, we control for the sensitive feature when estimating the weights on other features; but at the same time, we do not use the sensitive feature in the decision-making process. Based on this intuition, we then provided a simple two-step algorithm, called *train-then-mask*, for computing the optimal fair classifier. We showed that aside from simplicity and ease of computation, our notion of fairness had the advantage that it does not lead to *double discrimination*. That is when the group that is discriminated against is also the group that performs better overall, our method still removes the bias against that protected group. Finally, we should point out while h^* can be computed in many practical scenarios, but in the worst case, it is not possible to have h^*. Therefore, there is a gap between the surrogated accuracy and real accuracy which consequently results in a gap between surrogated and real discriminations.

Acknowledgements. The work of Amin Karbasi was supported by AFOSR Young Investigator Award (FA9550-18-1-0160).

References

Asuncion, A., and Newman, D. 2007. UCI Machine Learning Repository.

Barocas, S., and Selbst, A. D. 2016. Big data's disparate impact. *California Law Review* 104:671.

Blake, C. L., and Merz, C. J. 1998. Uci repository of machine learning databases. *Department of Information and Computer Science* 55.

Burke, R.; Sonboli, N.; and Ordonez-Gauger, A. 2018. Balanced Neighborhoods for Multi-sided Fairness in Recommendation. In *FAT*.

Celis, L. E., and Vishnoi, N. K. 2017. Fair Personalization. *CoRR* abs/1707.02260.

Celis, L. E.; Keswani, V.; Straszak, D.; Deshpande, A.; Kathuria, T.; and Vishnoi, N. K. 2018. Fair and Diverse DPP-Based Data Summarization. In *ICML*, 715–724.

Celis, L. E.; Straszak, D.; and Vishnoi, N. K. 2018. Ranking with Fairness Constraints. In *ICALP*.

Clarke, K. A. 2005. The phantom menace: Omitted variable bias in econometric research. *Conflict management and peace science* 22(4):341–352.

Donini, M.; Oneto, L.; Ben-David, S.; Shawe-Taylor, J.; and Pontil, M. 2018. Empirical risk minimization under fairness constraints. *CoRR* abs/1802.08626.

Dwork, C.; Hardt, M.; Pitassi, T.; Reingold, O.; and Zemel, R. 2012. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, 214–226. ACM.

Dwork, C.; Immorlica, N.; Kalai, A. T.; and Leiserson, M. D. M. 2018. Decoupled Classifiers for Group-Fair and Efficient Machine Learning. In *FAT*, 119–133.

Ekstrand, M. D.; Joshaghani, R.; and Mehrpouyan, H. 2018. Privacy for All: Ensuring Fair and Equitable Privacy Protections. In *FAT*.

Ellis, E., and Watson, P. 2012. *EU anti-discrimination law*. Oxford University Press.

Ghili, S. 2016. Network formation and bargaining in vertical markets: The case of narrow networks in health insurance. Available at SSRN: https://ssrn.com/abstract=2857305.

Gillen, S.; Jung, C.; Kearns, M. J.; and Roth, A. 2018. Online Learning with an Unknown Fairness Metric. *CoRR* abs/1802.06936.

Greene, W. H. 2003. *Econometric analysis*. Pearson Education India.

Grgic-Hlaca, N.; Zafar, M. B.; Gummadi, K. P.; and Weller, A. 2018. Beyond Distributive Fairness in Algorithmic Decision Making: Feature Selection for Procedurally Fair Learning. In *AAAI*.

Hardt, M.; Price, E.; Srebro, N.; et al. 2016. Equality of opportunity in supervised learning. In *NIPS*, 3315–3323.

Hashimoto, T. B.; Srivastava, M.; Namkoong, H.; and Liang, P. 2018. Fairness Without Demographics in Repeated Loss Minimization. In *ICML*, 1934–1943.

Heidari, H.; Ferrari, C.; Gummadi, K. P.; and Krause, A. 2018. Fairness Behind a Veil of Ignorance: A Welfare Analysis for Automated Decision Making. *CoRR* abs/1806.04959.

Hendel, I., and Nevo, A. 2013. Intertemporal price discrimination in storable goods markets. *American Economic Review* 103(7):2722–51.

Hu, L., and Chen, Y. 2018. A Short-term Intervention for Long-term Fairness in the Labor Market. In *Proceedings Conference on World Wide Web (WWW)*, 1389–1398. ACM.

Jabbari, S.; Joseph, M.; Kearns, M.; Morgenstern, J.; and Roth, A. 2017. Fairness in Reinforcement Learning. In *ICML*, 1617–1626.

Kamiran, F., and Calders, T. 2009. Classifying without discriminating. In *2nd International Conference on Computer, Control and Communication*, 1–6. IEEE.

Kamishima, T.; Akaho, S.; Asoh, H.; and Sakuma, J. 2018. Recommendation Independence. In *FAT*.

Kazemi, E.; Zadimoghaddam, M.; and Karbasi, A. 2018. Scalable Deletion-Robust Submodular Maximization: Data Summarization with Privacy and Fairness Constraints. In *ICML*, 2549–2558.

Kilbertus, N.; Carulla, M. R.; Parascandolo, G.; Hardt, M.; Janzing, D.; and Schölkopf, B. 2017. Avoiding discrimination through causal reasoning. In *NIPS*.

Kleinberg, J. M.; Mullainathan, S.; and Raghavan, M. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS)*, 43:1–43:23.

Liu, L. T.; Dean, S.; Rolf, E.; Simchowitz, M.; and Hardt, M. 2018. Delayed Impact of Fair Machine Learning. In *ICML*, 3156–3164.

Nabi, R., and Shpitser, I. 2018. Fair Inference on Outcomes. In *AAAI*.

Pedreshi, D.; Ruggieri, S.; and Turini, F. 2008. Discrimination-aware data mining. In *ACM SIGKDD*.

Pleiss, G.; Raghavan, M.; Wu, F.; Kleinberg, J. M.; and Weinberger, K. Q. 2017. On Fairness and Calibration. In *NIPS*.

ProPublica. 2018. COMPAS Recidivism Risk Score Data and Analysis.

Samadi, S.; Tantipongpipat, U.; Morgenstern, J.; Singh, M.; and Vempala, S. 2018. The Price of Fair PCA: One Extra Dimension. *CoRR* abs/1811.00103.

Scheffler, R. M.; Brown, T. T.; and Rice, J. K. 2007. The role of social capital in reducing non-specific psychological distress: The importance of controlling for omitted variable bias. *Social Science & Medicine* 65(4):842–854.

Schölkopf, B., and Smola, A. J. 2002. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press.

Seber, G. A., and Lee, A. J. 2012. *Linear regression analysis*, volume 329. John Wiley & Sons.

Speicher, T.; Ali, M.; Venkatadri, G.; Ribeiro, F. N.; Arvanitakis, G.; Benevenuto, F.; Gummadi, K. P.; Loiseau, P.; and Mislove, A. 2018. Potential for Discrimination in Online Targeted Advertising. In *FAT*.

Sudhir, K. 2001. Structural analysis of manufacturer pricing in the presence of a strategic retailer. *Marketing Science* 20(3):244–264.

White House. 2016. *Big data: A report on algorithmic systems, opportunity, and civil rights*. Executive Office of the President.

Zemel, R.; Wu, Y.; Swersky, K.; Pitassi, T.; and Dwork, C. 2013. Learning Fair Representations. In *ICML*, 325–333.

Zhang, J., and Bareinboim, E. 2018. Fairness in Decision-Making - The Causal Explanation Formula. In *AAAI*.

Žliobaitė, I., and Custers, B. 2016. Using sensitive personal data may be necessary for avoiding discrimination in data-driven decision models. *Artificial Intelligence and Law* 24(2):183–201.