Does mitigating ML’s disparate impact require disparate treatment?

Zachary C. Lipton¹, Alexandra Chouldechova¹, Julian McAuley²
Carnegie Mellon University¹
University of California, San Diego²
{zlipton, achould}@cmu.edu, jmcauley@cs.ucsd.edu
November 21, 2017

Abstract
Following related work in law and policy, two notions of prejudice have come to shape the study of fairness in algorithmic decision-making. Algorithms exhibit disparate treatment if they formally treat people differently according to a protected characteristic, like race, or if they intentionally discriminate (even if via proxy variables). Algorithms exhibit disparate impact if they affect subgroups differently. Disparate impact can arise unintentionally and absent disparate treatment. The natural way to reduce disparate impact would be to apply disparate treatment in favor of the disadvantaged group, i.e. to apply affirmative action. However, owing to the practice’s contested legal status, several papers have proposed trying to eliminate both forms of unfairness simultaneously, introducing a family of algorithms that we denote disparate learning processes (DLPs). These processes incorporate the protected characteristic as an input to the learning algorithm (e.g. via a regularizer) but produce a model that cannot directly access the protected characteristic as an input. In this paper, we make the following arguments: (i) DLPs can be functionally equivalent to disparate treatment, and thus should carry the same legal status; (ii) when the protected characteristic is redundantly encoded in the nonsensitive features, DLPs can exactly apply any disparate treatment protocol; (iii) when the characteristic is only partially encoded, DLPs may induce within-class discrimination. Finally, we argue the normative point that rather than masking efforts towards proportional representation, it is preferable to undertake them transparently.

1 Introduction
Effective decision-making relies on the ability of the decision maker to distinguish between options on the basis of available information. Selection processes, such as hiring and admissions, are typically driven by human assessments of applicants’ qualifications. This much is unavoidable, unless we opt to make trivial decisions and either select everyone, no one, or perform selection entirely at random. Yet some kinds of selection criteria violate ethical and legal principles. In many domains, the law explicitly prohibits adverse decisions made on the basis of an applicant’s irrelevant or protected characteristics. For example, in
the United States, in Title VII of the Civil Rights Act of 1964 [Civ64], the law forbids employment decisions that discriminate on the basis of the following protected characteristics: race, color, religion, sex, and national origin. The interpretation of this law has led to two widely-referenced notions of unfairness: disparate treatment and disparate impact.

Disparate treatment refers to intentional discrimination. This can include: (i) making decisions explicitly on the basis of a protected characteristic or (ii) making intentionally prejudiced decisions against members of a protected class via proxy variables. For example, in the 1900s, literacy tests were used to determine voting eligibility in order to disenfranchise racial minorities. Even absent disparate treatment, a facially neutral decision-making policy might exhibit disparate impact; i.e., unequal outcomes, for people in different classes with respect to some protected characteristic(s). This again may occur due to correlations between protected and unprotected characteristics.

In some cases, observed disparities are evidence of unfair treatment. For example, black defendants are sentenced to death more frequently than white defendants of the same crimes [For14]. This might owe in part to the racial biases of judges and juries; it also might owe to the correlation between race and wealth, and by extension, access to legal services. But disparate impact can also stem from more benign sources. For example, the over-representation of Asian students in prestigious US colleges appears not to stem from pro-Asian discrimination; on the contrary, investigative reports suggest that the over-representation actually arises despite admissions policies that set a higher bar for Asian applicants [HS17].

Disparate treatment and disparate impact are concepts rooted in United States labor law. In some texts, the terms have a more technical meaning (as defined above), and in others the meaning is tied up in the legal doctrine associated with a specific set of decisions, such as hiring. Throughout this paper, we will apply the technical definitions, following the convention in the existing Machine Learning (ML) literature that interprets these notions as widely applicable fairness criteria. Our discussion of disparate treatment and disparate impact in the context of algorithmic decision making may not at every stage adhere to prevailing legal doctrine. We also note that because current antidiscrimination laws were developed with human decision making in mind, there is considerable debate over their applicability to governing algorithmic decision-making systems [BS16, GW17, Kim17]. However, this topic exceeds the scope of the current paper.

1.1 Algorithmic Decision Making

Automated decision-making systems, based on ML models, are now trusted to make decisions of legal consequence, such as extending lines of credit, matching employees with employers, etc. These systems are typically built atop supervised ML models. In practice, the ML models do not take any actions themselves: they simply estimate the conditional probability of a label given some features. As a concrete example, a label might be a binary indicator \( \{0, 1\} \) of whether a loan defaults, and the features might be attributes related to a loan applicant’s financial history. Decisions are typically made by feeding the conditional probabilities into some decision rule.

In the simplest systems, the decision rule consists of a threshold applied
to the prediction. For example, if \( \hat{P}(\text{default}|x) > .2 \Rightarrow \text{reject loan} \). In many cases, while model outputs are generated programatically, decisions are made by humans. For example recidivism scores – predicted probabilities that an individual will be rearrested after being released – are taken into account as one among many criteria by judges when making decisions about bail, parole and sentencing.

A major concern is that a decision-making system based on an ML model might exhibit behavior that is unlawful with respect to protected characteristics. If the model explicitly incorporates a protected characteristic as input, this amounts to disparate treatment [Civ64, BS16, ZVGRG17]. An automated ML-based system can also exhibit disparate impact, which could arise via several possible mechanisms: (i) The choice of the target may be arbitrarily chosen among several that are all loose surrogates for the real quantity of interest (e.g. credit-worthiness). Any one particular choice might then benefit or disadvantage a given group; (ii) The targets themselves might reflect patterns of historical prejudice. For example, if members of the dominant group are more likely to keep their jobs during a layoff, then they may be more likely to repay their loans; (iii) If a group is grossly under-represented in the dataset, that could potentially lead to a model that is not as accurate when evaluating members of that group. In some of the literature on fairness in algorithmic decision-making, use of the protected characteristic is called direct discrimination, while points (i-iii) are sometime described as indirect discrimination [PRT08, KAS11].

Note that the behavior of an ML model depends on the dataset used for training, and thus these mechanisms could induce disparate impact absent a data scientist’s knowledge. As Barocas and Selbst note, “honest attempts to certify the absence of prejudice on the part of those involved in the data mining process may wrongly confer the imprimatur of impartiality on the resulting decisions” [BS16]. These issues have come to light starting with the ground-breaking paper [FN96], which recognized the ability of computer systems to disparately affect protected groups. Work specifically focused on discrimination owing to data-mining became more common following [PRT08]. Recently, several prominent cases of disparate impact garnered widespread media attention. For example, a report on ProPublica suggested that automated recidivism risk scores, used in sentencing and bail decisions, show racial bias [ALMK16]. Following the organization of workshops and conferences addressing related topics, research into algorithmic methods for mitigating disparate impact has accelerated.

1.2 An Overview of Disparate Learning Processes

To combat prejudice in ML-based decisions, there has been significant research interest in developing algorithms that constrain the level of disparate impact. The proposed approaches vary both in the working definition of fairness, and the algorithmic mechanisms used to ensure it. Most problem setups assume a binary classification setting in which the positive class is preferable. For example, the positive class might correspond to the designation “credit-worthy” or “employable”, which would lead a decision system to issue a loan, or recommend a job candidate, respectively. A common approach is then to cast the fairness problem as that of choosing a model that minimizes the disparate impact with minimal reduction in accuracy. A number of these papers [PRT08, KAS11, ZVGRG17, BL17] propose to remove disparate impact without resorting to dis-
parate treatment.

On the surface, this is a desirable goal. Disparate treatment and disparate impact may be seen as complementary definitions, both describing discriminatory mechanisms. In another sense, however, these definitions are in opposition. Given a decision-making system that exhibits disparate impact, the most obvious fix is to apply disparate treatment in favor of the disadvantaged group. Disparate treatment in the service of equality or diversity is sometimes described as affirmative action and reverse-discrimination [ZVGRG17]. In some cases, the courts have upheld the legality of applying disparate treatment to improve diversity, but there is also popular resentment of affirmative action and its future legality remains contested [BS16].

Many of the algorithmic proposals to reduce disparate impact while avoiding disparate treatment—which we denote disparate learning processes (DLPs)—operate according to the following principle: The protected characteristic may be used during training, but is not available to the model at prediction time. In the earliest such approach [PRT08], the protected characteristic was used to winnow the set of acceptable rules from an expert system. In other papers, the protected characteristic is incorporated into the learning objective (as a regularizer or constraint) or is used in pre-processing the training data [KC09, KCP10, ZVGRG17]. These approaches are grounded in the premise that DLPs are acceptable in cases where using a protected characteristic as a direct input to a model would constitute disparate treatment and thus be impermissible.

In this paper, we call this premise into question on the following grounds:

1. Disparate treatment in the service of improving diversity has been upheld as legal.
2. As we will show, the optimal way to trade off accuracy for proportional representation in the positive class is to apply disparate treatment directly.
3. When the protected characteristic is redundantly encoded in the other features, any disparate treatment can be equally implemented through a DLP.
4. When the protected characteristic is partially encoded in the other features, disparate treatment induces within-class discrimination applying the benefit of the affirmative action unevenly, and can even harm some members of the protected class.

The fact that DLPs and disparate treatment are functionally equivalent when the protected attributes are redundantly encoded in non-protected features should cast doubt on the legality of these algorithms in contexts where disparate treatment is prohibited. A legal opinion by Grimmelmann and Westreich supports this view [GW17]:

In our view, Title VII does not permit an employer to do indirectly what it could not do directly. An employer that explicitly selects applicants on the basis of [group membership] violates Title VII under a disparate treatment theory [...] regardless of whether it bears animus against particular [groups]. It is the selection “on the basis of” [group membership] that is the problem. An employer that uses home address to infer applicants’ [group membership] and then
selects applicants from particular groups does exactly the same, only in two steps rather than one. This too is a form of disparate treatment.

While to our knowledge none of the existing approaches have considered the task of maximizing disparate impact, in principle similar mechanisms could be used to do so. Applying a DLP to improve the fortunes of the dominant class might be judged by a court to constitute intentional discrimination and thus be a form of disparate treatment. Such a judgment might apply to the mechanism irrespective of the outcome it was used to effect. This would undermine the chief argument for DLPs, since the same ends can be achieved more effectively by proportional representation-promoting disparate treatment. It is worth noting that an algorithm that disproportionately fortunes the dominant class is more likely to raise red flags (under disparate impact) than one that effects representative outcomes. So the extent to which DLPs can mask intentional discrimination is limited to levels that do not subject the practice to scrutiny on the basis of significant disparate impact. Problematically, the law is often ambiguous, and the legal status of these algorithms has not yet, to our knowledge, been tested in the courts, leaving practitioners and researchers without clear guidelines.

One potential source of ambiguity lies in whether we consider the algorithms to be correcting for biases in the dataset, or if the disparate learning process is deemed to be an explicit diversity-promoting positive discrimination. Problematically, when a dataset with discriminatory labeling (such as historical hiring
decisions) arrives, it is seldom accompanied by meta-data precisely quantifying the prejudice. This leaves researchers to make extreme assumptions, e.g. that any correlation between the protected class and the label is attributable to discrimination. That assumption would entail numerous erroneous conclusions, such as that the high academic performance of Asian Americans is due to systemic pro-Asian discrimination, despite the abundant evidence to the contrary [HS17].

### 1.3 A Note on Organization

The rest of the paper is organized as follows: In Section 2, we give a more technical description of DLPs. In Section 3, we demonstrate some simple theoretical problems with disparate learning processes. In Section 4, we demonstrate the tendency of DLPs to perpetrate within-class discrimination, applying them both to a clean synthetic dataset and a real dataset of University admissions data. These sections aim to be objective and aim to make no value judgments. Then, in Section 6, we express the position that the policy and technical communities might benefit from accepting the efficacy and transparency of explicit pro-diversity disparate treatment. We conclude with a discussion of the challenges that remain such as the issues posed by data bias, and the difficulty in navigating the terrain between estimation and decision-making.

Throughout this paper we strive to maintain a distinction between the objective question of whether an algorithm discriminates along a protected characteristic, and the normative question of whether that discrimination should be permitted because it serves a socially desirable goal, such as ensuring proportional representation. In an attempt to distinguish between the descriptive and the normative, we use an organization scheme that separates the discussion of these different questions. Sections 2, 3, and 4 are descriptive while Section 6 is normative.

### 2 Disparate Learning Processes

To begin our formal description of the prior work, we’ll introduce some formal notation. A dataset \(X, Y\), consists of \(n\) examples, or data points \(\{x_i \in X, y_i \in Y\}\), each consisting of a feature vector \(x_i\) and a label \(y_i\). A supervised learning algorithm produces a model \(\hat{y} : X \to Y\), which given a feature vector \(x_i\), predicts the corresponding output \(y_i\). In this discussion, we’ll focus on binary classification, the setting in which the label \(y\) takes values from the set \(\{0, 1\}\). We’ll also focus on probabilistic classifiers, which produce estimates \(\hat{p}(x)\) of the conditional probability \(P(y = 1 \mid x)\) of the label given a feature vector \(x\). To make a prediction \(\hat{y}(x) \in Y\) given an estimated probability \(\hat{p}(x)\), a thresholding strategy is applied such that \(\hat{y}_i = 1\) if \(\hat{p} > t\). The optimal choice of the threshold \(t\) may depend on the performance metric being optimized. For instance, under 0–1 loss, the optimal decision rule thresholds \(P(y = 1 \mid x)\) at \(t = 0.5\). To optimize F1 score, the optimal threshold depends on the classifier’s confidence [LEN14]. Following prior work, we focus most of our analysis on the accuracy metric or 0–1 loss.

Note that a supervised learning algorithm is itself a function, mapping from datasets to models \(f : (X^n, Y^n) \to (X \to [0, 1])\). Additionally, some datasets
possess a sensitive attribute $Z$, making each example a three-tuple $x_i, y_i, z_i$. The protected characteristic may be real-valued, like age, or categorical, like race or gender. Following the related work, we focus on categorical characteristics, and look specifically at the binary case where the protected characteristic divides the population into groups $a$ and $b$. As shorthand, we will refer to the number of members of classes $a$ and $b$ as $n_a = \sum_{i}^n \mathbb{1}(z_i = a)$ and $n_b = \sum_{i}^n \mathbb{1}(z_i = b)$, respectively.

The papers which propose DLPs generally consider training a classifier on the protected characteristic, i.e. with feature vector $\tilde{x} = [x; z]$ to be impermissible, amounting to disparate treatment. However, even if the protected features $z_i$ are discarded, the model may still produce probabilities of belonging to the positive class $\hat{p}(x)$ that are correlated with $z$. Applying thresholds to make decisions, the ML-based decisions might correlate with the protected characteristic: $P(\hat{y} = 1 | z) \neq P(\hat{y} = 1)$. Empirically, we could estimate the proportions assigned to the positive class by evaluating the quantities $q_a = (\sum_{i: z_i = a} \mathbb{1}(\hat{y}(x_i) > .5))/n_a$ and $q_b = (\sum_{i: z_i = b} \mathbb{1}(\hat{y}(x_i) > .5))/n_b$.

When our goal in learning is simply to maximize accuracy, the estimation of $P(y = 1 | x)$ simplifies to the standard problem of binary classification. However, some papers propose trading off the accuracy of the classifier for reduction in disparate impact \cite{PRT08, KC09, ZVGRG17}. Different papers address different measures of disparate impact, but a few formulations are common. One approach addresses the Calders-Verwer (CV) gap, which quantifies disparate impact as $q_b - q_a$, the difference between the proportions assigned to the positive class in the disadvantaged group ($a$) and the advantaged group ($b$) \cite{KAS11}.

This definition is asymmetric in that it assumes a disadvantaged class, but could be made more generic by taking the absolute value of the difference. In \cite{ZVGRG17}, the proposed measure of disparate impact is $q_a/q_b$, following a book by Biddle on fair employment practices \cite{Bid06}, assuming that $a$ is the disadvantaged group. They propose constraining a model to satisfy a $p$-\% rule: $q_a/q_b > p/100$.

This generalizes the heuristic in \cite{Bid06} that disparate impact can be diagnosed when the ratio of proportions assigned to the positive class is less than .8 (a $p$-\% rule of 80). For simplicity, we focus on these two scores, but our results can be trivially extended to some (but possibly not all) measures of disparate impact. For example, \cite{BL17}, proposes another regularization-based scheme with two penalty terms. The first minimizes the difference in the false negative rate between the two groups, and the other minimizes the difference in the false positive rate. A similar analysis can be applied there.

Many papers propose to minimize their chosen measure of disparate impact to within some acceptable range with minimal reduction of accuracy. But even calculating these measures requires access to the sensitive feature. These papers state that incorporating the feature directly in determining the class assignments $\hat{y}_i$ constitutes disparate treatment. So they propose instead to incorporate the sensitive feature in the learning algorithm, but not the model. The formal function signature for a disparate learning process follows:

$$DLP : (X^n, Y^n, Z^n) \rightarrow (X \rightarrow Y).$$

Since $z$ is not a direct input of the resulting model, it is often asserted that such a model has a better legal standing than a model that uses $z$ directly.
We argue qualitatively that this claim is unreasonable, both with respect to the law \cite{GW17} and because under some circumstances the DLP is technically equivalent to disparate treatment.

Before addressing any qualitative arguments, we first demonstrate some theoretical properties of disparate treatment, demonstrating among other things that it optimally addresses the constrained optimization problem posed by several authors of DLP papers.

3 Theoretical Issues

In this section we introduce theoretical arguments in support of counterarguments (2-4) outlined in Section 1.2. We present a set of simple theoretical results that demonstrate the optimality of disparate treatment, and highlight properties of DLPs. Our optimality results are all derived in the population or “infinite data” setting where we assume knowledge of the true conditional probability function $p_{Y|X,Z}(x,z) \equiv \mathbb{P}(Y = 1 | X = x, Z = z)$. The main results can be summarized as follows.

1. Direct disparate treatment on the basis of $z$ is the optimal strategy for minimizing the expected $0 - 1$ loss subject to CV and $p$-% constraints.

2. When $X$ fully encodes $Z$, a sufficiently powerful DLP is equivalent to disparate treatment.

3. When $X$ only partially encodes $Z$, a DLP may be suboptimal and induce intra-group disparity.

3.1 Disparate treatment is optimal

Absent disparate impact constraints, the Bayes-optimal decision rule for minimizing expected $0 - 1$ loss (i.e., maximizing accuracy) is given by

$$d_{\text{uncon}}^*(x,z) = \begin{cases} 1 & p_{Y|X,Z}(x,z) \geq 0.5 \\ 0 & \text{otherwise} \end{cases}.$$  

In this section we show that the optimal decision rule in the CV and $p$-% constrained problems has a similar form. The optimal decision rule will again be based on thresholding $p_{Y|X,Z}(x,z)$, but at group-specific thresholds. These rules can be be thought of as operationalizing the following disparate treatment mechanism. Suppose that we start with the classifications of the unconstrained rule $d_{\text{uncon}}^*(x,z)$, and this allocated results in a CV gap of $q_b - q_a > \gamma$. To reduce the CV gap to $\gamma$ we have two mechanisms. We can flip predictions of cases in group $a$ from 0 to 1, and we can also flip predictions of cases in group $b$ from 1 to 0. The optimal strategy is to perform these flips on group $a$ cases that have the highest value of $p_{Y|X,Z}(x,z)$ and group $b$ cases that have the lowest value of $p_{Y|X,Z}(x,z)$.

The results in this section adapt the work of \cite{CDPF+17}, who establish optimal decision rules $d$ under different kinds of fairness constraints. In this work the authors characterize the optimal decision rule $d = d(x,z)$ that maximizes the immediate utility $u(d,c) = \mathbb{E}[Yd(X,Z) - cd(X,Z)]$ ($0 < c < 1$) under different
parity criteria. We begin with a lemma showing that expected classification accuracy has the functional form of an immediate utility function.

**Lemma 1.** Optimizing classification accuracy is equivalent to optimizing immediate utility with $c = 0.5$.

**Proof.** The expected accuracy of a binary decision rule $d(X)$ can be written as $E[Yd(X) + (1 - Y)(1 - d(X))]$. Expanding and rearranging this expression gives

$$E[Yd(X) + (1 - Y)(1 - d(X))] = E(2Yd(X) - d(X)) + E(Y) + 1$$

$$= 2u(d, 0.5) + E(Y) + 1$$

The only term in this expression that depends on $d$ is the immediate utility, $u$. Thus the decision rule that maximizes $u$ also maximizes accuracy.

For the next set of results, we follow [CDPF+17] and assume that $p_{Y|X,Z}(X, Z)$ viewed as a random variable has positive density on $[0, 1]$. This ensures that the optimal rules are unique and deterministic by disallowing point-masses of probability that would necessitate tie breaking among observations with equal probability. The first result that we state is a direct corollary of two results in [CDPF+17]. It considers the case where we desire exact parity; i.e., the constraint $q_a = q_b$.

**Corollary 2.** The optimal decision rules $d^*$ under various fairness constraints have the following form, and are unique up to a set of probability zero.

1. Among rules satisfying statistical parity (the 100% rule), the optimum is

$$d^*(x, z) = \begin{cases} 1 & p_{Y|X,Z}(x, z) \geq t_z \\ 0 & \text{otherwise} \end{cases}$$

where $t_z \in [0, 1]$ are constants that depend only on group membership $z$.

2. Among rules that have equal false positive rates across groups, the optimum is

$$d^*(x, z) = \begin{cases} 1 & p_{Y|X,Z}(x, z) \geq s_z \\ 0 & \text{otherwise} \end{cases}$$

where $s_z$ are constants that depend only on group membership $z$ (but are different from $t_z$).

3. (1) and (2) continue to hold even in the resource-constrained setting where the overall proportion of cases classified as positive is constrained.

**Proof.** (1) and (2) are direct corollaries of Lemma 1 combined with Theorem 3.2 and Prop 3.3 of [CDPF+17].

The next set of results establish optimality under general $p$-% and CV rules.

**Proposition 3.** Under the same assumptions as above, the optimum among rules that satisfy the CV constraint $0 \leq q_b - q_a < \gamma$ or the $p$-% rule also has the form

$$d^*(x, z) = \begin{cases} 1 & p_{Y|X,Z}(x, z) \geq t_z \\ 0 & \text{otherwise} \end{cases}$$
where \( t_z \in [0, 1] \) are constants that depend on the group membership \( z \), and on the choice of constraint parameter \( \gamma \) or \( p \). The thresholds \( t_z \) are different for the CV constraint and \( p \)-% rule.

**Proof.** Suppose that the optimal solution under the CV or \( p \)-% rule constraint classifies as positive \( q_a \) proportion of disadvantaged cases and \( q_b \) proportion of advantaged cases. As shown in [CDPF+17], we can rewrite the immediate utility as

\[
u(d, 0.5) = E\left[d(X, Z)(p_{Y|X,Z} - 0.5)\right].
\]

From this expression it is clear that the utility will be maximized precisely when

\[
d^*(X, Z) = 1 \quad \text{for the } q_z \text{ proportion of individuals in each group that have the highest values of } p_{Y|X,Z}.
\]

Since the optimal values of \( q_z \) may differ between the CV constrained solution and the \( p \)-% solution, the optimal thresholds may differ as well. The final result in this section shows that a decision rule that does not directly use \( z \) as an input variable or for determining the thresholds will have lower accuracy than the optimal rule that uses this information. That is, we show that DLPs are suboptimal for trading off between accuracy and disparate impact.

**Theorem 4.** Let \( d^*(X, Z) \) be the optimal decision rule under a the CV-\( \gamma \) or \( p \)-% constraint. Let \( d_{DLP}(X) \) be the optimal solution to a DLP. If \( d(X, Z) \) and \( d_{DLP}(X) \) satisfy CV or \( p \)-% constraints with the same \( q_a \) and \( q_b \), the DLP solution results in lower or equal accuracy. (Equal only if the solutions are the same.)

**Proof.** From Proposition 3 we know that the unique accuracy optimizing solution is given by

\[
d^*(x, z) = \begin{cases} 
1 & p_{Y|X,Z}(x, z) \geq t_z \\
0 & \text{otherwise}
\end{cases}
\]

where \( t_z \) is the 1 - \( q_z \) quantile of \( p_{Y|X,Z} \). The difference in immediate utility between the two decision rules can be expressed as follows.

\[
E[d^*(X, Z)(p_{Y|X,Z} - 0.5)] - E[d_{DLP}(X)(p_{Y|X,Z} - 0.5)]
= E[d^*(X, Z) - d_{DLP}(X)(p_{Y|X,Z} - 0.5)]
= E[p_{Y|X,Z} - 0.5 | d^* = 1, d_{DLP} = 0]P(d^* = 1, d_{DLP} = 0)
- E[p_{Y|X,Z} - 0.5 | d^* = 0, d_{DLP} = 1]P(d^* = 0, d_{DLP} = 1)
= (E[p_{Y|X,Z} - 0.5 | d^* = 1, d_{DLP} = 0]
- E[p_{Y|X,Z} - 0.5 | d^* = 0, d_{DLP} = 1])P(d^* = 1, d_{DLP} = 0)
\geq 0
\]

The final inequality follows from the observation that \( d^*(X, Z) = 1 \) for the highest values of \( p_{Y|X,Z} \), so \( p_{Y|X,Z} \) is stochastically greater on the event \( \{d^* = 1, d_{DLP} = 0\} \) than on \( \{d^* = 0, d_{DLP} = 1\} \). Note that equality holds only if \( P(d^* = 1, d_{DLP} = 0) = 0 \); that is, if the two rules are equivalent with probability 1. \( \square \)
All of the results in this section continue to hold under “do no harm” constraints where the proportion of cases in the disadvantaged group classified as positive is constrained to be no lower than the proportion under the unconstrained rule \( d_{\text{uncons}}(x, z) \) (or no lower than some fixed value \( q_{\text{min}} \)). This constraint imposes an upper bound on the optimal thresholds \( t_a \), but does not change the structure of the optimal rules.

### 3.2 Functional equivalence when protected characteristic is redundantly encoded

Consider the case where the protected characteristic \( z \) is redundantly encoded in the permissible data features \( x \). More precisely, suppose that there exists a known subcomputation \( g \) such that \( z_i = g(x_i) \forall i \). This allows for any function of the data \( f(x, z) \) to be represented as a function of \( x \) alone via \( \tilde{f}(x) = f(x, g(x)) \). While it remains the case that \( \tilde{f}(x) \) does not directly use \( z \) as an input variable, \( \tilde{f} \) should be no less suspect from a disparate impact perspective than the original function \( f \) that uses \( z \) directly. The main difference for the purpose of our discussion is that \( \tilde{f} \) is a valid possible result from a DLP whereas \( f \) is not. Yet it is clear that \( f \) and \( \tilde{f} \) are functionally equivalent in two senses: (i) as mathematical objects; and more importantly, (ii) as mechanisms for classifying cases.

### 3.3 Within-class differentiation when protected characteristic is partially redundantly encoded

When the protected characteristic is partially encoded in the other features, disparate treatment may induce within-class discrimination by applying the benefit of the affirmative action unevenly, and can even harm some members of the protected class. This claim asserts a possibility, so it is sufficient to produce one example to support the claim. In the following section, we establish the claim empirically using both synthetic data and real university admissions data. The ease of producing such examples might convince the reader that the highly varied effects of intervention with a DLP on members of the disadvantaged group raise serious questions about the usefulness of DLPs.

### 4 Empirical Analysis

This simple analysis that precedes makes plain several advantages to mitigating disparate impact by applying disparate treatment:

- **Optimality**: As demonstrated for CV score and for p-% rule, the disparate treatment intervention maximizes accuracy subject to a constraint on disparate impact.

- **Rational ordering**: Within each group, individuals with higher probability of belonging to the positive class are always assigned to the positive class ahead of those with lower probabilities.

- **Does no harm to the protected group**: The disparate treatment intervention can only benefit members of the disadvantaged class.
Figure 2: Left: probability of the sensitive variable versus (unconstrained) admission probability, on unseen test data. Points above 0.5 are individuals who are classified as ‘reject’ only after applying the fairness constraint; points below 0.5 are individuals who are classified as ‘admit’ only after applying the fairness constraint; the remaining ~4,000 individuals (whose labels were not altered by the fairness constraint) are shown as blue/yellow dots. Note that most students admitted due to the fairness approach are actually males who ‘look like’ females on the basis of their other features, whereas females who reflect male characteristics are more likely to be rejected. Center: detail view of the same plot. Right: summary statistics of the same plot.

But DLPs do not directly apply disparate treatment. Instead, they must recover a classifier that satisfies the disparate impact constraints, by relying upon the proxy features to minimize the disparate impact measure. In many of these papers, this is accomplished either by introducing constraints to a convex optimization problem, or by adding a regularization term and tuning the corresponding hyper-parameter. Because the CV score and $p$-% rule are non-convex in model parameters (scores only change when a point crosses the decision boundary), [KAS11, ZVGRG17] introduce convex surrogates aimed at reducing the correlation between the sensitive feature and the prediction.

All of these approaches assume that the proxy variables contain information about the sensitive attribute. Otherwise, the model could only achieve fairness by arriving at a trivial solution (e.g., assign everyone or no one to the positive class). So we must consider two scenarios: (i) the proxy variables $x$ fully redundantly encode $z$. In this case, an sufficiently powerful DLP will implicitly reconstruct $z$, because this gives the optimal solution to the impact-constrained objective. However, when $x$ doesn’t fully capture $z$, then the DLP may (i) be sub-optimal, (ii) violate rational ordering within groups, and (iii) harm members of the disadvantaged group.

4.1 Synthetic data example: work experience and hair length in hiring

To begin, we confirm these arguments empirically with a simple synthetic data experiment. To construct the data we sample $n_{all} = 2000$ total observations from the data generating process described below. 70% of the observations are
used for training, and the remaining 30% are reserved for model testing.

\[ z_i \sim \text{Bernoulli}(0.5) \]

\[ \text{hair\_length}_i \mid z_i = 1 \sim 35 \cdot \text{Beta}(2, 2) \]

\[ \text{hair\_length}_i \mid z_i = 0 \sim 35 \cdot \text{Beta}(2, 7) \]

\[ \text{work\_exp}_i \mid z_i \sim \text{Poisson}(25 + 6z_i) - \text{Normal}(20, \sigma = 0.2) \]

\[ y_i \mid \text{work\_exp} \sim 2 \cdot \text{Bernoulli}(p_i) - 1, \text{ where} \]

\[ p_i = 1 / (1 + \exp[-(-25.5 + 2.5\text{work\_exp})]) \]

This data generating process has the following key properties: (i) The historical hiring process was based solely on the number of years of work experience; (ii) Because women on average have fewer years of work experience than men (5 years vs. 11), men have been hired at a much higher rate than women; (iii) Women have longer hair than men, a fact that was irrelevant to historical hiring practice.

Figure 1 shows the test set results of applying a DLP to the available historical data to equalize hiring rates between men and women. We apply the DLP proposed by [ZVGRG17], using code available from the authors. While the DLP is successful in equalizing hiring rates (satisfying a 100%-rule), it does so through a problematic within-class discrimination mechanism. The DLP rule advantages individuals with very long hair length over those with short hair length and considerably longer work experience. We find that several women who would have been hired under historical practices owing to their 11+ years of work experience would not be hired under the DLP due to their short hair length (i.e., their male-like characteristics in the data). Similarly, several men who would not have been hired based on work experience alone are advantaged by the DLP on account of their longer hair length (i.e., their female-like characteristics in the data). The DLP mechanism violates rational ordering, and also has the effect of harming some of the most qualified individuals in the protected group. Group parity is achieved at the cost of significant individual unfairness.

Granted, factors such as hair length could not knowingly and defensibly be used as an input to a typical hiring algorithm. This example was constructed to illustrate a more general point. Since DLPs do not have direct access to the protected attribute, they must infer from the data cases that are most likely to be members of each subgroup. Using the protected attribute directly yields more reasonable policies: ones that hire the most qualified individuals in each group, rather than those that are most qualified among those that appear, from their unprotected characteristics, to be the most feminine.

### 4.2 Case Study: Gender Bias in CS Graduate Admissions

For our next example, we considered data from the Master’s admissions process of a large public university, considering a sample of ~9,000 students considered for admission over an 11-year period spanning 2006-2016. Half are withheld for testing. The available attributes include basic demographic information, such as country of origin, interest areas, and gender, as well as quantitative information

[https://github.com/mbilalzafar/fair-classification/](https://github.com/mbilalzafar/fair-classification/)
such as GRE scores. Finally, it includes a label in the form of a decision provided by an admissions committee.\(^2\)

Based on a superficial analysis, the data does not appear to exhibit gender bias (the admissions rates for male and female applicants are within 1% of each other). So, for the purposes of our experiments, we corrupt the data with synthetic discrimination. Of all women who were admitted, i.e., \(z_i = a, y_i = 1\), we flip 25% of those labels to 0: giving noisy labels \(\bar{y}_i = y_i \cdot \eta\), for \(\eta \sim Bernoulli(.25)\). This simulates a setting in which the training data exhibits a historical bias.

We then train three logistic regressors: (1) To predict the (prejudice-corrupted) labels from the non-sensitive features \(\{x_i, \bar{y}_i\}\); (2) The same model, applying the fairness constraint of [ZVGRG17]; and (3) A logistic regressor that predicts the sensitive feature from the non-sensitive features \(\{x_i, z_i\}\). The data contains limited information that can predict gender, though such predictions can be made better than random (AUC=0.59) due to different rates of gender imbalance across (e.g.) countries and interest areas.

Figure 2 (left) shapes our basic intuition for what is happening here: Considering the probability of admission for the unconstrained classifier (y-axis), students whose decisions are ‘flipped’ (after applying the fairness constraint) tend to be those close to the decision boundary. Furthermore, students predicted to be male (x-axis) tend to be flipped to the negative class (left half of plot) while students predicted to be female tend to be flipped to the positive class (right half of plot). This is shown in detail in Figure 2 (center and right). However, of the 19 students whose decisions are flipped to ‘admit,’ the majority (10) are males, each of whom has ‘female-like’ characteristics according to their other features. Demonstrated here with real-world data, the DLP both disrupts the within-group ordering, and violates the do no harm principle by disadvantaging some women who, but for the DLP, would have been admitted.

4.2.1 Comparison with Disparate Treatment

To demonstrate the better performance of disparate treatment, we implement a simple thresholding scheme. Assuming that our model gives us calibrated probabilities, and that this is all the information available to the decision maker, it’s easy to derive the optimal thresholds for maximizing accuracy under linear constraints on the proportions of predicted positives, like the CV-gap or \(p\)-% rule.

We now present a simple thresholding scheme for maximizing accuracy subject to a \(p\)-% rule. Recall that the \(p\)-% rule requires that \(q_a/q_b > p/100\). We can rewrite this as:

\[
\frac{p}{100}q_b - q_a < 0
\]

Like the CV-gap, the \(p\)-% rule imposes a linear constraint. We denote the quantity \(\frac{p}{100}q_b - q_a\) the \(p\)-gap. To maximize accuracy subject to satisfying the \(p\)-% rule, we construct a score, that quantifies reduction in \(p\)-gap per reduction in accuracy. Starting from the accuracy-maximizing predictions (thresholded at .5), we then flip those predictions which close the gap fastest:

\(^2\)These decisions do not precisely determine whether a student is made an offer, but rather represent an ‘above-the-bar’ assessment that is used to guide admissions decisions, and can be considered as a binary label.
1. Assign each example with \( \tilde{y}_i = 0, z_i = a \) or \( \tilde{y}_i = 1, z_i = b \), a score \( c_i \) equal to the reduction in the CV-gap divided by the reduction in accuracy:

(a) For each example in group \( a \) with initial \( \tilde{y}_i = 0 \),
\[
c_i = \frac{n}{n_a} \left( 1 - 2\hat{p}_i \right).
\]
(b) For each example in group \( b \) with initial \( \tilde{y}_i = 1 \),
\[
c_i = \frac{n_P}{100n_a (2\hat{p}_i - 1)}.
\]

2. Flip examples in descending order according to this score until the desired CV-score is reached.

These scores do not change after each iteration. So the greedy policy is optimal.

Overall, the fairness constraint of \( [ZVGRG17] \) achieves a \( p \)-% rule of 77.99%, compared to a \( p \)-% rule of 71.44% by naïve classification (on unseen test data). Both have similar accuracy: given that both positive labels and female applicants are a minority, assigning negative labels to males close to the boundary impacts accuracy very little. Both methods had accuracy of around 78% on this data. Critically though, by applying an optimal thresholding strategy, we were able to obtain the same accuracy as the method of \( [ZVGRG17] \), but with a higher \( p \)-% rule of 78.34%. Similarly, we could achieve a modest improvement in accuracy (<0.1%) while maintaining the same \( p \)-% rule as the method of \( [ZVGRG17] \).

5 Related Work

In this section we provide a brief overview of some of the other approaches that have been put forth for trading off between classification performance and disparate impact. One common approach consists of preprocessing or “massaging” the training data to reduce the dependence between the resulting model predictions and the sensitive attribute \( [KC09, KC12, FFM +15, AFF +16, JL17] \). These methods differ both in terms of what variables are affected by the data processing, and the degree of independence that is achieved. For instance, \( [KC09] \) propose flipping the negative labels of some observations in the disadvantaged class. \( [ZWS +13] \) proposes learning representations - in this case, cluster assignments - of each example such that each example maps to a cluster with some probability. They seek statistical parity in the percentage withing each group assigned to each cluster. \( [FFM +15] \) also investigates transformations of the features \( X \) into a new set of features that are constructed to be marginally independent from \( Z \). \( [JL17] \) demonstrate how to construct transformations to ensure that the derived features are jointly independent of \( Z \), and show that this produces distributional parity of the resulting fitted model.

A second widely adopted approach is to modify existing classification methods either through post-hoc corrections or in the training stage to constrain the level of disparate impact in the resulting model. \( [KAS11, CGGF16, CV10, KCPT10] \) consider modifications to methods such as SVM, logistic regression, Naive Bayes and decision trees. \( [ABDL17] \) show how disparate impact constraints can be framed as a cost-sensitive classification problem.
6 Discussion

Following our description of the problem, theoretical analysis, and empirical findings, we now offer a more normative take on these findings.

6.1 Coming to Terms with Disparate Treatment

At present, most legal scholarship and technical machine learning scholarship take place in a disjoint set of journals. These communities occasionally intersect when some paper, such as the widely influential California Law Review article by Barocas and Selbst reaches a cross-disciplinary audience [BS16]. However, the subsequent interdisciplinary technical work tends to be published in technical conferences, where the peer-reviewers may be ill-equipped to identify shortcomings in problem formulation.

For instance, the interpretations of anti-discrimination law that motivate DLPs appear not to consider that (i) present-day law might rule DLPs to be equivalent to disparate treatment if tested under the law (see e.g., arguments in [GW17]); and, (ii) disparate treatment may already be tolerated under the law in order to ensure more fair outcomes. This latter view is supported by Pauline Kim in her paper Data-driven discrimination at work [Kim17]:

A formalist reading of Title VII might appear to prohibit any use of variables capturing sensitive characteristics in a data model. Certainly, a simple model that relied on race or other protected characteristics as the basis for adverse decisions would run afoul of Title VII’s prohibitions. However, when dealing with a complex statistical model involving multiple variables, the appropriate treatment of these sensitive variables is more complicated. If the goal is to reduce biased outcomes, then a simple prohibition on using data about race or sex could be either wholly ineffective or actually counterproductive due to the existence of class proxies and the risk of omitted variable bias. Instead, avoiding classification bias may sometimes call for excluding sensitive demographic variables and at other times call for including them. Any response to biased data models must be sensitive to these nuances.

On the balance of these considerations, there are several compelling reasons for practitioners to promote equality more transparently through direct disparate treatment, rather than through hidden changes to the learning algorithm. As articulated earlier, a disparate treatment based approaches have three principal advantages over DLPs: they (i) optimally trade accuracy for representativeness; (ii) preserve rankings among members of each group (as compared to the unconstrained scores); and (iii) do no harm to members of the disadvantaged group.

In addition to these three properties, disparate treatment has another advantage. By setting class-dependent thresholds, it’s much easier to quantify how disparate treatment impacts individuals. Having an intuitive quantity to reason about might help policy-makers to decide what magnitude of disparate treatment best trades off group equality and individual fairness. For more indirect methods to satisfy disparate impact constraints, it might be hard to reason about the the intervention. As an example, it seems doubtful that policy-makers
would similarly intuitive meaning in the setting of a specific regularization co-
efficient.

Several key challenges still remain. The theoretical arguments in this paper
demonstrate that disparate treatment approaches are optimal in the setting
where we assume complete knowledge of the data generating distribution. It is
not always clear how best to realize these gains in practice, where imbalanced or
unrepresentative data sets can pose a significant obstacle to accurate estimation.
Furthermore, some of our results are tailored to the CV or the p-% rule notions
of fairness. As shown in [HPS+16, WGOS17] and [DIKL17], the situation can
much more complicated for other fairness criteria.

6.2 Separating Estimation and Decision-Making
In many algorithm-supported decision making contexts, it is desirable to obtain
not just a classification, but also an accurate probability estimate. These esti-
mates could then be incorporated into the decision-theoretic part of the pipeline
and appropriate measures could be taken at that stage to ensure the desired out-
come properties. By intervening at the modeling phase, DLPs risk distorting the
probabilities themselves. It is not clear what the probabilities that come out of
the resulting classifiers actually signify. In unconstrained learning approaches,
even if the label itself may reflect historical prejudice, one at least knows what
is being estimated. This leaves open the possibility of intervening at decision
time to promote more equitable outcomes.

While the distinction between building a model and making decisions is
stated clearly in the first modern work on fairness in discrimination-aware clas-
sification [PRT08], this distinction is frequently muddled in discussions of al-
gorithmic fairness. For example, [KC09] state that “a learned model may ex-
hibit unlawfully prejudiced behavior”. The conflation of modeling and decision-
making may lead to counterproductive corrections to the models that do not
adequately account for how the models are actually used. For example, it is
commonly assumed that decision makers desire to optimize accuracy and hence
that decisions will be made by thresholding probability estimates at .5. This
is often not the case. First, due to differences in the cost of false positives
and false negatives, accuracy is seldom a task-relevant metric. Furthermore,
decision-makers are often faced with a multi-objective problem that entails con-
siderations beyond what the algorithm is designed to predict.

6.3 Fairness beyond disparate impact
How best to quantify discrimination and unfairness remains an important open
question. The CV scores and p-% rules addressed in this paper offer one
set of definitions (often technically termed ‘disparate impact’), but there are
many other notions of fairness to which our results do not directly apply. For
example, equality of opportunity as introduced in [HPS+16]—requiring equality
of true positive rates across groups—has received considerable attention. Other
conditional notions of fairness and trade-offs between them have been studied
by [JKM+16, KMR16, Cho17, BHJ+17, RSZ17]. The work of [ZVR+17] departs
from parity-based definitions and proposes instead a preference-based notion of
fairness. [DIKL17] address the problem of how best to incorporate information
about protected attributes for several of these other fairness criteria.
Problematically, research into fair algorithms is often motivated by the case in which our ground-truth data is itself biased. It is not clear how to assess many of these other fairness criteria in the presence of biased data. Characterizing different forms of data bias and their impacts on fairness assessment remains an important outstanding challenge.

Even if we accept that the solution for many proportional representation problems will take the form of disparate treatment in favor of the disadvantaged class, a question remains of “how much?”. At what point is the disparate treatment simply correcting for biased labels? At what point does it more explicitly amount to affirmative action? Recent work on identifying proxy discrimination [DFK+17] and causal formulations of fairness [NS17, KRCP+17, KLRS17] offer approaches to framing such problems. To answer these questions, it would help to have a better understanding of by what mechanisms and to what degree the data has been influenced by prejudice. Perhaps data mining and machine learning have some role to play in asking these questions? The answers could guide decisions about where and how strongly to intervene.

References

[ABDL17] Alekh Agarwal, Alina Beygelzimer, Miroslav Dudík, and John Langford. A reductions approach to fair classification. 2017.

[AFF+16] Philip Adler, Casey Falk, Sorelle A Friedler, Gabriel Rybeck, Carlos Scheidegger, Brandon Smith, and Suresh Venkatasubramanian. Auditing black-box models by obscuring features. arXiv preprint arXiv:1602.07043, 2016.

[ALMK16] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias. 2016.

[BHJ+17] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. arXiv preprint arXiv:1703.09207, 2017.

[Bid06] Dan Biddle. Adverse impact and test validation: A practitioner’s guide to valid and defensible employment testing. Gower Publishing, Ltd., 2006.

[BL17] Yahav Bechavod and Katrina Ligett. Learning fair classifiers: A regularization-inspired approach. arXiv preprint arXiv:1707.00044, 2017.

[BS16] Solon Barocas and Andrew D Selbst. Big data’s disparate impact. California Law Review, 2016.

[CDPF+17] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. arXiv preprint arXiv:1701.08230, 2017.

[Cho17] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. Big Data, 2017.
Civil rights act of 1964, 1964. Accessed on September 11th, 2017.

Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery*, 21(2):277–292, 2010.

Anupam Datta, Matt Fredrikson, Gihyuk Ko, Piotr Mardziel, and Shayak Sen. Proxy non-discrimination in data-driven systems. *arXiv preprint arXiv:1707.08120*, 2017.

Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. Decoupled classifiers for fair and efficient machine learning. *arXiv preprint arXiv:1707.06613*, 2017.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 259–268. ACM, 2015.

Batya Friedman and Helen Nissenbaum. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, 1996.

Matt Ford. Racism and the execution chamber. 2014.

Gabriel Goh, Andrew Cotter, Maya Gupta, and Michael P Friedlander. Satisfying real-world goals with dataset constraints. In *Advances in Neural Information Processing Systems*, pages 2415–2423, 2016.

James Grimmelmann and Daniel Westreich. Incomprehensible discrimination. *California Law Review*, 2017.

Moritz Hardt, Eric Price, Nati Srebro, et al. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems*, 2016.

Anemona Hartocollis and Stepanie Saul. Affirmative action battle has a new focus: Asian americans. 2017.

Matthew Joseph, Michael Kearns, Jamie Morgenstern, Seth Neel, and Aaron Roth. Rawlsian fairness for machine learning. *arXiv preprint arXiv:1610.09559*, 2016.

James E Johndrow and Kristian Lum. An algorithm for removing sensitive information: application to race-independent recidivism prediction. *arXiv preprint arXiv:1703.04957*, 2017.

Toshihiro Kamishima, Shotaro Akaho, and Jun Sakuma. Fairness-aware learning through regularization approach. In *Data Mining Workshops (ICDMW), 2011 IEEE 11th International Conference on*, pages 643–650. IEEE, 2011.
[KC09] Faisal Kamiran and Toon Calders. Classifying without discriminat-
ing. In *Computer, Control and Communication, 2009. IC4 2009. 2nd International Conference on*, pages 1–6. IEEE, 2009.

[KC12] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowledge and Information Systems*, 33(1):1–33, 2012.

[KCP10] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Discrimi-
nation aware decision tree learning. In *Data Mining (ICDM), 2010 IEEE 10th International Conference on*, pages 869–874. IEEE, 2010.

[Kim17] Pauline T Kim. Data-driven discrimination at work. *William & Mary Law Review*, 58(3), 2017.

[KLRS17] Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *arXiv preprint arXiv:1703.06856*, 2017.

[KMR16] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. In-
herent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05587*, 2016.

[KRCP+17] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascand-
olo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. *arXiv preprint arXiv:1706.03744*, 2017.

[LEN14] Zachary C Lipton, Charles Elkan, and Balakrishnan Naryanaswamy. Optimal thresholding of classifiers to maxi-
imize f1 measure. In *ECML*, 2014.

[NS17] Razieh Nabi and Ilya Shpitser. Fair inference on outcomes. *arXiv preprint arXiv:1705.10378*, 2017.

[PRT08] Dino Pedreshi, Salvatore Ruggieri, and Franco Turini. Discrimination-aware data mining. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 560–568. ACM, 2008.

[RSZ17] Ya’acov Ritov, Yuekai Sun, and Ruofei Zhao. On conditional parti-
ness as a notion of non-discrimination in machine learning. *arXiv preprint arXiv:1706.08519*, 2017.

[WGOS17] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. *arXiv preprint arXiv:1702.06081*, 2017.

[ZVGRG17] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2017.
[ZVR\textsuperscript{+}17] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, Krishna P Gummadi, and Adrian Weller. From parity to preference-based notions of fairness in classification. \textit{arXiv preprint arXiv:1707.00010}, 2017.

[ZWS\textsuperscript{+}13] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In \textit{Proceedings of the 30th International Conference on Machine Learning}, 2013.