Finding and Curing Proxy Discrimination

This version, 28 March 2026

Dan O'Flaherty

Abstract: *We show how to find algorithms that are personally non-discriminatory, proxy non-discriminatory, and optimally accurate subject to the constraints of personal and proxy non-discrimination. Proxy non-discrimination is not ubiquitous, even when decision-makers use variables that are highly correlated with race. When it occurs, it can be cured, although at some cost to accuracy.*

Acknowledgments: *Jinxin Tian and Theo Burden have been tremendous research assistants. I think Jeff Fagan for New York City police search data and Will Dobbie for pointing me toward the New York State pretrial detention data.*

People who want to discriminate on the basis of race (or any other protected category) can use non-racial words to do so. For instance, "grandfather clauses" in Jim Crow voting laws effectively barred most Blacks from registering by mandating excruciatingly hard standards for residents whose male ancestors were not permitted to vote before the Civil War. If removing racial words is not sufficient to avoid discriminating, how can people who don't want to discriminate avoid doing so, especially when they use complex, high-powered algorithms to help them make decisions?

This paper proposes a new and very general answer to this question. While we talk about racial discrimination, almost everything also applies to discrimination on the basis of gender, age, disability, or membership in any other protected class. We use race in order to simplify exposition. Pope and Sydnor (2011), Ghili et al (2019), and Yang and Dobbie (2020) have also answered this question, but only for special cases.

Of course, anyone can avoid discriminating by acting blindly or letting dice make decisions for them, but this paper assumes that the decision-maker has some (non-racial) objective function they are trying to maximize (admitting the smartest possible class, for instance). More precisely, then, this paper shows how to maximize this objective function subject to the constraint that the decision be fully non-discriminatory (I will explain "fully" soon).

The question is most acute when attributes a non-discriminatory decision-maker values are also negatively correlated with protected class membership—for instance, when Black or women candidates for admission have on average lower SAT scores than white or men candidates. A decision-maker who denied admission to Blacks or women would definitely be discriminatory, and so presumably would be a decision-maker who denied admission to people with dark skin (precisely defined) or to people who wear dresses (at least once every few months). What is the

morally relevant difference between denying admission to these two classes and denying admission to people with low SAT scores? In my presumed setting, dark skin, dress-wearing, and low SAT scores all predict relatively poor academic performance and none of these attributes predicts race or gender perfectly.

My approach is to define two types of non-discrimination, personal and proxy, and say that a selection rule is *fully* non-discriminatory if satisfies both types. Roughly speaking, personal non-discrimination means that no person is treated worse because of their race; the decision-maker is blind to a person's race. Proxy non-discrimination means that no group is treated worse because of its racial composition; the decision-maker is blind to the group's racial composition.[1] Obviously personal non-discrimination requires that the decision-maker select (non-racially defined) groups and within each group, every member is treated the same way. A selection rule is both optimal and fully non-discriminatory if it would be optimal if all groups had the same racial composition. That is, it is non-discriminatory if it can be justified by appealing to some world where all groups had the same racial composition. (In colloquial economics terms, an optimal fully non-discriminatory selection rule is optimal if is optimal in an imagined where all selection effects have disappeared, or where no attribute or function of a set of attributes is correlated with race.)

The major difference between our approach and that of Pope and Sydnor (2011) and Yang and Dobbie (2020) is that we do not think about covariates. Picture a standard data set, one with $\mathcal{O}$ observations and $\mathcal{V}$ covariates. Conventionally, the names of the covariates appear across the top and names of the observations run down the left side. Covariates, conventionally, mean column

[1] In a society where race is a continuous variable not a binary one (Brazill perhaps) both types of non-discrimination would be defined by almost the same language.

vectors of $\mathcal{O}$ dimensions. In contrast, our primitive objects are "descriptions": row vectors of $\mathcal{V}$ dimensions. Changing the focus from columns to rows turns out to be quite helpful.

Note that our goal is non-discrimination, not fairness as that term has come to be used. The literature on fairness seeks algorithms that produce outcomes that appear balanced in various ways; in contrast, the non-discrimination literature is not directly concerned with outcomes. For discussions and summaries of the fairness literature, see, for instance, Chohlas-Wood et al (2023), Card and Smith (2020), and Corbett-Davies et al (2024).

We also work with a much stronger definition of nondiscrimination than many papers that develop "nondiscriminatory algorithms." For instance, Arnold et al (2025) show how bias can be removed from an algorithm by processing observed characteristics or post-processing calculated outputs but using either procedure violates what we call "personal nondiscrimination"—the requirement that two people who differ only in protected characteristics be treated the same.

The plan of the paper is the following. We begin by explaining our set up and definitions, especially what we mean by various types of nondiscrimination. We also show that proxy discrimination is related to Simpson's paradox. The next section concentrates on how to make optimal fully nondiscriminatory selections in a very wide class of problems. A short literature review follows. Understanding how this work fits in with the existing literature requires some understanding of what this work is. After that is a section about examples. We discuss one toy example from the literature and two examples with actual data that are related to well-known papers. (Since these papers used restricted data, we use related but publicly available data.) In only one of these three examples do we find proxy discrimination, and in that case we show how to eliminate proxy discrimination at minimum cost to accuracy. The cost in accuracy of curing proxy discrimination turns out to be small in this example. The final part of this section discusses why,

contrary to much of the literature, proxy discrimination does not seem to be common in carefully developed algorithms.

Set up, notation, and definitions

A decision-maker is trying to select some candidates out of a possibly larger number of candidates to be treated in a particular way. For instance, a police officer may be trying to select cars or pedestrians to stop, a university admissions office may be trying to select students for its incoming class, a customs officer may be trying to select travelers to be searched, a marketer may be trying to select email addresses to receive an advertisement, a voter may be trying to select a candidate to vote for, and so on. In deciding which candidate to select, the decision-maker can distinguish certain non-racial characteristics of candidates and can also distinguish a protected characteristic of each candidate. To be concrete, we call the protected characteristic race and restrict race to be binary --Black or white.

Let x denote a generic element of a set X, which is the finest partition of non-racial information[2] that the decision-maker can perceive. Assume w.l.o.g. that the measure of each x is the same, which we normalize to one. The measure of the set X is an integer n. The model has three data, which the decision-maker knows: a payoff function, a racial distribution function, and a cost. The *payoff function* is

$$p: \{b, w\} \times X \to (0,1)^2$$

The *racial distribution function* is

$$\beta: X \to (0,1)^n$$

[2] It is not at all clear that race can actually be distinguished from non-racial characteristics, but, following most of the economics literature, I will assume that the distinction is possible. See Rose (2023) for a discussion.

For instance, for a police officer, the payoff p(r,x) is the probability that a car with non-racial description x and driver of race r will be carrying discoverable contraband; and $\beta(x)$ is the proportion of the cars with characteristic x that have Black drivers. Estimating these parameters is in general a difficult task to which many scholars have devoted great attention; but in this paper we will assume that these parameters are common knowledge.

Finally, we assume that the cost of selecting a candidate is $c > 0$, the same for each candidate selected. Candidates not selected cost nothing, and so the decision-maker trades off the benefits from selecting good candidates against the costs that any selection incurs. We also assume that decision-makers are risk neutral.

Notice that in many cases x will be a multi-dimensional object—the most complete description available to the decision-maker of a candidate's non-racial characteristics. Hence we will call the elements of x "descriptions", for short. For most of this paper we will not be concerned with variables like height, education, color of vehicle that are often called "characteristics" and that are usually the main subjects of concern in discussion of non-discriminatory algorithms. Our results are more comfortable in the setting of trees or forests, not regressions. But in the appendix we show how to translate our methods into a world of regressions.

A *selection rule*

$$S\colon \{b,w\} \times X \times \mathbb{R}^+ \to \{0,1\}$$

tells decision-maker which candidates to select:

$$S(r,x|p,\beta,c) = 1 \; iff \; the \; decision$$
$$- maker \; selects \; all \; candidates \; of \; race \; r \; and \; description \; x$$

Thus, absent considerations of discrimination, the decision-maker's problem is to choose a selection rule $S(r, x|p, \beta, c)$ to maximize expected net benefits:

$$\max_S \sum_x [\beta(x)p(b,x)S(b,x|p,\beta) + \big(1-\beta(x)\big)p(w,x)S(w,x|p,\beta) - c]$$

We call this the cost-constrained problem.

An alternative perspective is that the decision-maker is not constrained explicitly by cost but instead is bound by a quota q on the expected number of candidates they may select.

$$\max_S \sum_x [\beta(x)p(b,x)S(b,x|p,\beta) + \big(1-\beta(x)\big)p(w,x)S(w,x|p,\beta)]$$

subject to

$$\sum_x [\beta(x)S(b,x|p,\beta) + \big(1-\beta(x)\big)S(w,x|p,\beta)] \leq q$$

We call this the quota-constrained problem. In this paper, we will generally be concerned with cost-constrained problems rather than quota-constrained problems because for most of results, cost-constrained problems are simpler to discuss. We will refer to quota-constrained problems when discussing related literature, however, because that literature often works with quota-constrained problems. Of course, every quota problem is equivalent to some cost problem and vice versa, and so this restriction is not serious.

Discrimination. We consider two kinds of discrimination. *Personal discrimination* occurs if some candidates with the same non-racial descriptions, but different races are

treated differently. *Proxy discrimination* occurs if candidates with some description x are selected solely because of the proportion of Black candidates among those with that description. Personal discrimination singles out individual candidates for adverse (or beneficial) treatment because of their race. Proxy discrimination singles out groups with particular non-racial descriptions for adverse (or beneficial) treatment because of their racial composition.

Specifically, a selection rule S(.) is personally non-discriminatory iff for all x, p, c, and β

$$S(b, x|p, \beta. c) = S(w, x|p, \beta, c)$$

A selection rule S(.) is proxy nondiscriminatory iff for all x, p, and β there is some scalar $a \in [0,1]$ such that

$$S(r, x|p, \beta, c) = 1\ , r = b, w, if\ and\ only\ if\ \varphi(a, x) = ap(b, x) + (1 - a)p(w, x) \geq c$$

The definition of personal non-discrimination is standard; it is the popular definition that has been used almost since the Civil Rights Act of 1866. It requires that the decision-maker act as if they were race-blind; alternatively, it requires that the decision-maker be able to defend their decision ex post without saying anything about race.

This definition of proxy discrimination is not so familiar. It can be motivated by the following argument. Consider a world in which the racial composition of every characteristic was the same:

$\beta(x) = a \; for \; all \; x.$

In such a world, no description could be a proxy for race, no candidate's description could be informative about their race, and no racial information could be informative about descriptions. No covariate would be correlated with race, to use the more standard terminology. The optimal selection rule would be to select applicants of description x if and only if $\varphi(a, x) = ap(b, x) + (1 - a)p(w, x) \geq c$.

Proxy non-discrimination requires that the decision-maker act as if they were in some world where non-racial characteristics are uninformative about race. (Notice that proxy non-discrimination requires that a selection rule be justified by appealing to some value of weights $a \in [0,1]$, not any particular value. The constraint is that the same value must be used to evaluate all descriptions; different weights cannot be used for different descriptions.)

The connection between personal and proxy discrimination is clear. Personal non-discrimination requires that the decision-maker act as if they cannot recognize race directly. Proxy non-discrimination requires that the decision-maker act as if they cannot recognize race indirectly through correlations with characteristics. Full non-discrimination requires both.

The idea of proxy discrimination is closely related to the better-known concept of disparate impact. Consider the high school degree requirements for internal promotion in the Duke Power generating plant in 1960s North Carolina, the subject of the 1971 Supreme Court decision *Griggs v Duke Power* that established the disparate impact document.

Duke Power required a high school diploma for promotion out of the poorly paid "labor" job category.

Let x_1 denote the description of having a high school diploma-- and x_2 denote not having one. So $\beta(x_1) < \beta(x_2)$. Assume that having a high school diploma makes no difference to one's ability to do one of the jobs outside fhe labor category; and race also makes no difference. Hence,

$p(w, x_1) = p(w, x_2)$=$p(b, x_1)$=$p(b, x_2)$.

Notice that it would not violate personal nondiscrimination for a decision-maker to favor the white-dominated description, denying promotions to people without high school diplomas. A decision-maker who acted in such a fashion could justify their actions by appealing solely to descriptions without invoking race. This decision might be sub-optimal, but it would not be personally discriminatory.

But this decision would be proxy discriminatory. For any a and c, it is impossible that $\varphi(a, x_1) > c > \varphi(a, x_2)$ and so the decision cannot be justified without appeal to the racial composition of descriptions. (A similar analysis can be applied to the grandfather clause example at the beginning of this paper.)

Our definition of proxy discrimination is also implicit in Yang and Dobbie's discussion. For the general case, they outline two procedures that they claim will always eliminate proxy discrimination and still allow decision-makers to achieve moderately good results. One method is to score every description x by $\varphi(0, x)$ and to select descriptions with high values of this score: that is to set $a = 0$. The other method they recommend is to

$a = \bar{\beta}$, where $\bar{\beta}$ is the average proportion Black over all descriptions. In the examples section below we discuss these methods in more detail. We differ from Yang and Dobbie in that we allow any value of a, not just 0 and $\bar{\beta}$, and we search for the value that maximizes expected payoff.

Domination

Proxy non-discrimination (especially in the quota-constrained problem) is closely related to the concept we call "domination." Description x *dominates* a description y whenever $p(b,x) \geq p(b,y),\ p(w,x) \geq p(w,y)$ and the converse. Description x *strictly dominates* description y whenever x dominates y and at least one of the inequalities is strict. A description x *dominates the cost scalar* whenever $p(b,x) \geq c \ and \ p(w,x) \geq c$. Description x *strictly dominates the cost scalar* whenever x dominates the cost scalar and at least one of the inequalities is strict. The cost scalar *dominates a description* x whenever $p(b,x) \leq c \ and \ p(w,x) \leq c$. The cost scalar *strictly dominates a description x* whenever the cost scalar dominates the description and at least one of the inequalities is strict.

Then we say that a decision rule is *dominance respectful* iff for all r, x, p, c, β, and y

(1) Whenever x dominates y, $S(r,x|p,\beta,c) \geq S(r,y|p,\beta,c)$ and

(2) Whenever x strictly dominates the coat scalar, $S(r,x|p,\beta,c) = 1$.

(3) Whenever the cost scalar strictly dominates x, $S(r,x|p,\beta,c) = 0$.

It is easy to see that all proxy non-discriminatory selection rules are dominance respectful, but the converse does not hold.

For instance, the unconstrained optimal selection rule is

$S(r, x|p, \beta, c) = 1 \; if \; and \; only \; if \; p(r, x) \geq c$ for r= b, w.

This selection rule is dominance respectful but not proxy non-discriminatory. On the other hand, the optimal selection rule constrained only by personal non-discrimination may not be dominance respectful and hence not proxy non-discriminatory.

Example A below illustrates.

Example A

c = .15

Descriptions	p(b, x)	p(w, x)	$\beta(x)$	$\varphi(\beta(x), x)$
x = 1	.20	.10	.10	.11
x = 2	.19	.08	.90	.179

Where $\varphi(\beta(x), x)$ is calculated from the preceding three columns.

The optimal policy unconstrained by either non-discrimination constraint is to select only Black candidates and to select all Black candidates:

$$S(b, x|p, \beta, c) = 1 \; for \; all \; x$$

$$S(w, x|p, \beta, c) = 0 \; for \; all \; x.$$

The expected payoff is 0.195 – 0.15=0.045 (since both descriptions are equally likely).

The optimal policy constrained only by personal non-discrimination is, for r= b, w

$$S(r, 1|p, \beta, c) = 0$$

$$S(r, 2|p, \beta, c) = 1$$

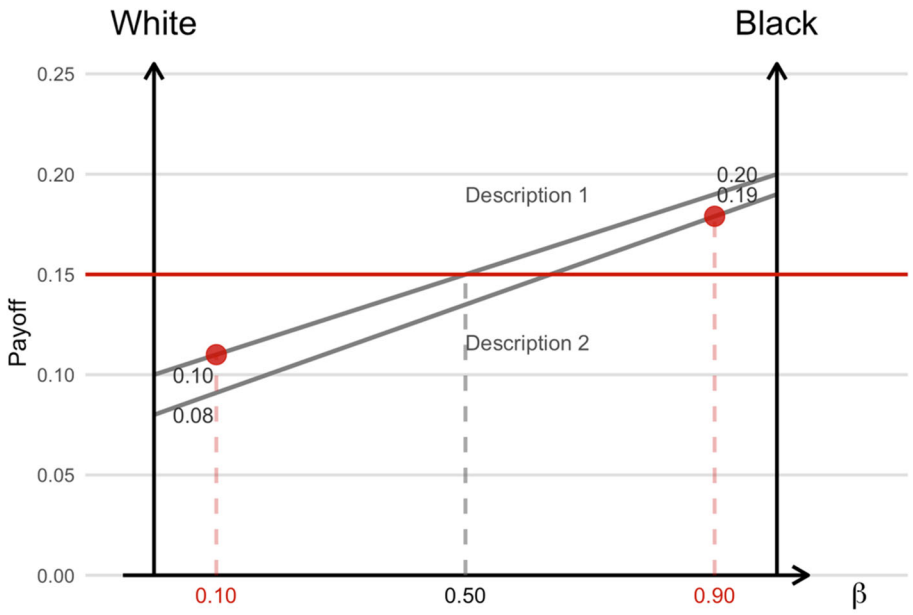
Only description 2 is selected, and the expected payoff is 0.179-0.150=0.029, less than the unconstrained expected payoff. From the first two columns, however, description 2 is dominated by description 1. Description 2 is being used as a proxy for being Black; choosing it violates proxy non-discrimination.

Because all proxy non-discriminatory policies respect dominance, only three selection rules are proxy non-discriminatory: either both descriptions are selected, or only description 1 is selected, or no description is selected. If only description 1 is selected, the expected net payoff is .11 - .150 < 0. Expected net payoff if both descriptions are satisfied is also negative, since .11 - .150 + .179 - .150 < 0. So the optimal policy that satisfies both personal and proxy non-discrimination in this example is to select no candidate. Any other dominance-respectful policy will result in negative expected net payoff. The no-proxy-discrimination constraint is binding and costly in this example.

We can illustrate this example in figure 1 with a diagram we call a bridge graph. We will see many bridge graphs in this paper. On the horizontal axis are values of ß ranging from 0 to 1. Payoffs are on the horizontal axis. The vertical axis at $\beta = 0$ corresponds to white payoffs and the vertical axis at $\beta = 1$ corresponds to Black payoffs. The line segment

from .10 on the white axis to .20 on the Black corresponds to expected payoffs under description 1 as a function of β. Similarly the line segment from .08 on the white axis to .19 on the Black axis corresponds to description 2. Finally, the line segment between .15 and .15 represents cost (which is constant with respect to β).

Figure 1: Bridge Graph Illustration



The line for description 1 is always higher than the line for description 2: this illustrates dominance.

Finally, we add actual values of β on the horizontal axis for the two descriptions, and directly above each of them we place red dots. For each description x the height of the red dot is the expected payoff from selecting that description, $\varphi(\beta(x), x)$. We see that the

red dot for description 2 is above the red dot for description 1 and above the cost line. That was why selecting only description 2 was the optimal selection rule constrained only by personal nondiscrimination. But since the line for description 1 is always above the line for description 2, proxy nondiscrimination forbids selecting description 2 alone.

The bridge graph shows that for $a < .5$, both lines are below .15: both descriptions give negative expected net payoff. The optimal selection rule we have calculated (nothing is selected) is proxy non-discriminatory. It can be justified without appealing to any differences in racial composition.

Simpson's paradox. Example A suggests that there is a close connection between proxy discrimination and Simpson's paradox. That is correct in a sense to be made precise shortly: Simpson's paradox implies susceptibility to proxy discrimination, but not the converse.

Following Sprenger and Weinberger (2021) we define Simpson's paradox the following way[3]: A combination (p, β) of payoff function and racial composition is an *example of Simpson's paradox* if there exist descriptions $x_1, x_2 \in X$ such that x_1 dominates x_2 and $\varphi(\beta(x_1), x_1) < \varphi(\beta(x_2), x_2)$.

For completeness, I define proxy discrimination in the obvious way: A problem (p, β, c) is an example of *proxy discrimination* if there is a selection rule that is optimal subject to the constraint of personal nondiscrimination that is not proxy nondiscriminatory.

These definitions imply:

[3] Specifically, this is the association reversal variety of Simpson's paradox.

Proposition. If (p, β) is an example of Simpson's paradox, then there exists a c such that (p, β, c) is an example of proxy discrimination.

Proof. Suppose (p, β) is an example of Simpson's paradox, x_1 dominates x_2 and $\varphi(\beta(x_1), x_1) < c < \varphi(\beta(x_2), x_2)$. Then under any selection rule that is optimal subject only to personal nondiscrimination, x_2 is selected and x_1 is not selected. However, because x_1 dominates x_2, this optimal personally nondiscriminatory selection rule is not proxy nondiscriminatory.

The converse of this proposition does not hold. There are examples of proxy nondiscrimination that are not examples of Simpson's paradox; for instance, figure 4 in the next section.

Finding optimal fully nondiscriminatory policies

These propositions about dominating descriptions greatly simplify the task of finding optimal non-discriminatory algorithms in many cases.

In any example, let Δ denote the set of descriptions that strictly dominate the cost scalar and let ∇ denote the set of descriptions that the cost scalar strictly dominates. The following proposition is obvious:

Domination proposition: If S(.) is an optimal non-discriminatory policy, then for all $x \epsilon \Delta, and\ r \epsilon \{b, w\}, S(r, x|p, \beta, c) = 1$, and for all $x \epsilon \nabla, and\ r \epsilon \{b, w\}, S(r, x|p, \beta, c) = 0$.

Proof: If x dominates the cost scalar, then $\varphi(\beta(x), x) - c > 0$, no matter what $\beta(x)$ is, and so selecting x will always improve the decision-maker's payoff. Since $\varphi(a, x) - c > 0$, no matter what

a is, this selection will always occur in a proxy non-discriminatory selection rule. A similar argument holds when the cost scalar dominates x, *mutatis mutandis*.

We call a problem *simple* iff every description either strictly dominates or is strictly dominated by the cost scalar; that is, iff $\Delta \cup \nabla = X$.

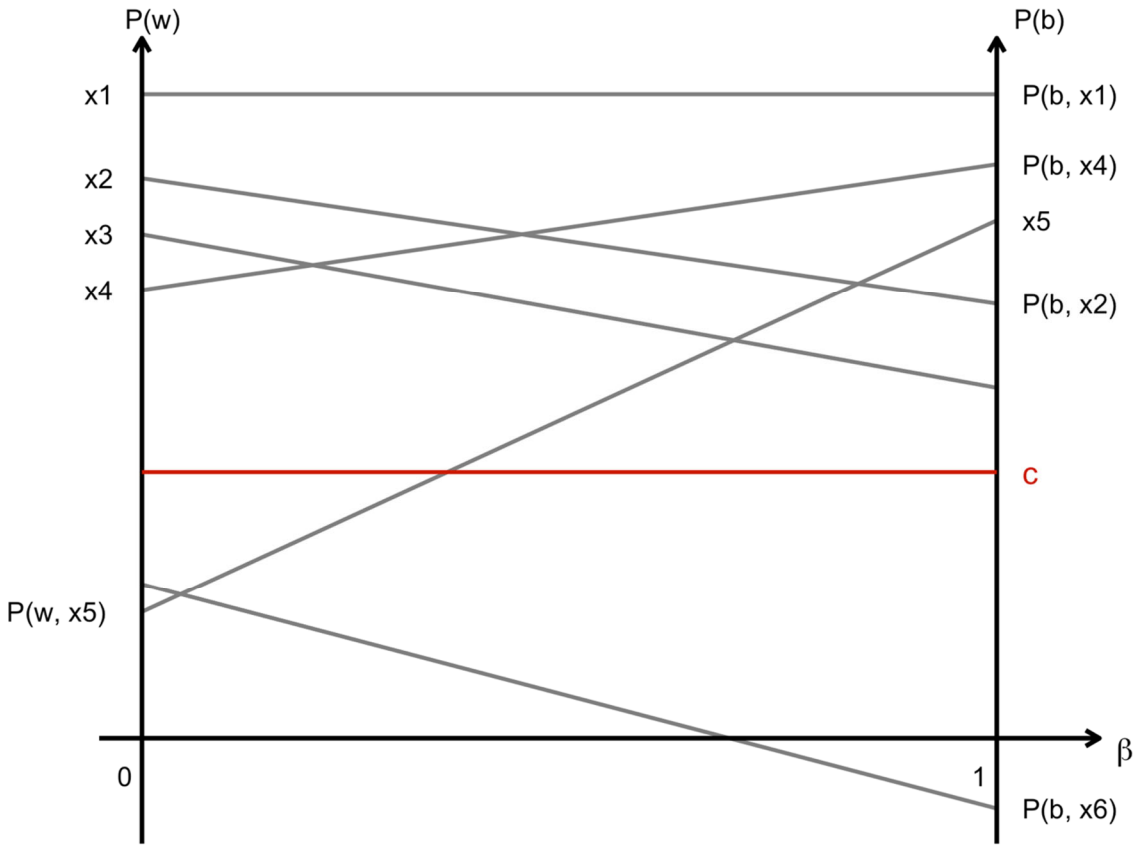
Corollary to the domination proposition: If a problem is simple, the domination proposition describes the optimal non-discriminatory policy completely. Any value $a \in (0,1)$ will justify this decision rule. This selection rule is the same as the optimal personally non-discriminatory selection rule.

Thus we can describe the optimal non-discriminatory policy easily in simple problems. This class of simple problems may not be small—for instance, a problem is simple if every $p(b,x)$ is a monotone tranaform of $p(w,x)$. The papers of Pope and Sydnor and of Ghili et al., for instance, are about a variety of simple problems.

The domination proposition also helps in solving problems that are not simple. It resolves selection for descriptions in $\Delta \cup \nabla$, and so only remainng issues are the selections in the complement of $\Delta \cup \nabla$ in X. Specifically, a description x is ambiguous iff min $(p(b,x), p(w,x)) < c < \max\{p(b,x), p(w,x))$. For instance, both descriptions in example A are ambiguous. Let M denote the set of ambiguous descriptions.

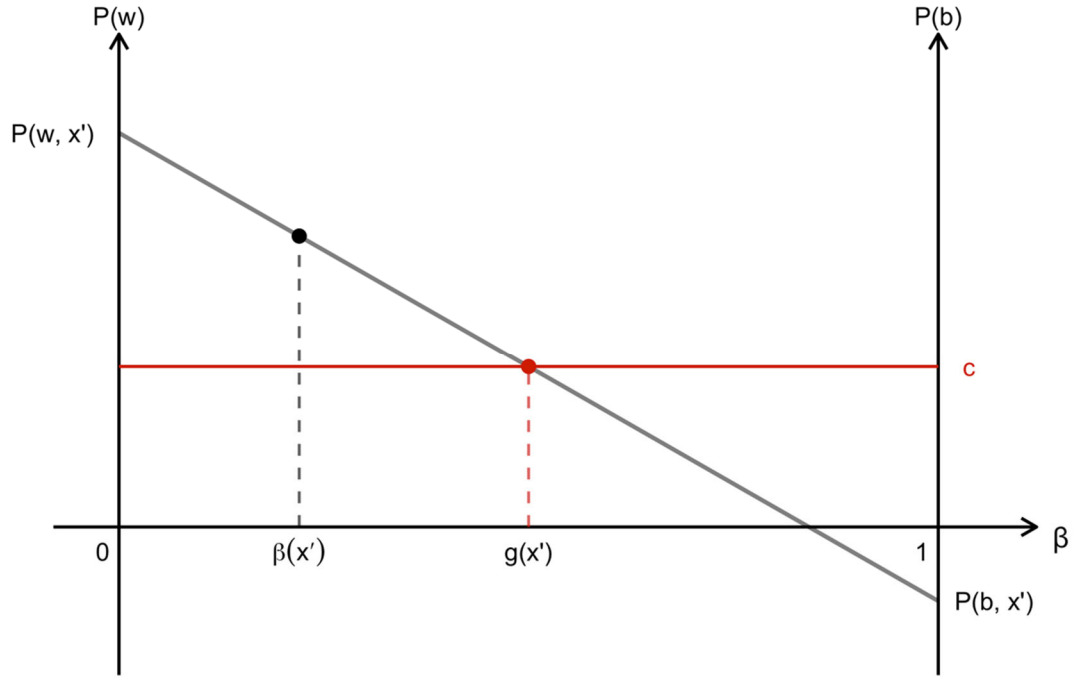
Figure 2 is a bridge graph with many descriptions to illustrate these three kinds of descriptions. Descriptions that dominate the cost scalar are at the top, ambiguous descriptions are in the middle, and descriptions that the cost scalar dominates are at the bottom.

Figure 2: Bridge Graph Illustration with Different Descriptions



To start thinking about ambiguous descriptions, consider an example where only one description x' is ambiguous, and assume that p(w,x')> p(b, x') so that the line segment for x' slopes down as illustrated in figure 3. The value of β that makes $\varphi(\beta, x') = c$ is called the cross-over value g(x') and we assume that g(x') > $\beta(x')$, as shown in figure 3. Thus $\varphi(\beta(x'), x') - c > 0$ (as again shown on in figure 3) and so in the optimal policy constrained only by personal non-discrimination x' is selected. Thus the optimal non-discriminatory policy selects x' also because x' would be selected with a < g(x). This shows how we find the optimal non-discriminatory policy when only one description is ambiguous. In general, when only one description is ambiguous, the optimal non-discriminatory policy coincides with the optimal policy subject only to the personal non-discrimination constraint.

Figure 3: Bridge Graph Illustration with One Ambiguous Description



This happy property does not carry over to cases with more than one ambiguous description as we saw in example A. Figure 4 illustrates another way that proxy nondiscrimination can be a stronger condition than personal nondiscrimination. Here we keep x' but add a new ambiguous description x" with p(w,x") < p(b, x") so that the line segment for x" slopes up. Assume g(x') < g(x") < $\beta(x")$, as shown in figure 3 and so $\varphi\left(\beta\left(x^{"}\right), x"\right) - c > 0$. In the optimal policy constrained only by personal nondiscrimination, both x' and x" are selected. But because g(x') < g(x"), both cannot be selected in a policy that is proxy non-discriminatory: for a < g(x') only x' is selected; for g(x') <a <g(x"), neither is selected; and for a > g(x"), only x" is selected.[4] In figure 4, $\varphi\left(\beta\left(x^{"}\right), x"\right) > \varphi(\beta(x'), x')$ and so a > g(x") and only x" is selected.

[4] Intuitively, what's wrong with selecting both x' and x"? Consider a college admissions problem. Description x' is applicants from public high schools near the center of large Northeastern cities who do well on an exam about basketball knowledge and x" is applicants from private high schools in the Upper Midwest who do well on an exam about hockey knowledge. The college is looking for smart students. Admitting both groups of students (evaluating the Midwesterners by using the white proportion and Northeasterners on the

Figure 4: Bridge Graph Illustration with Two Ambiguous Descriptions

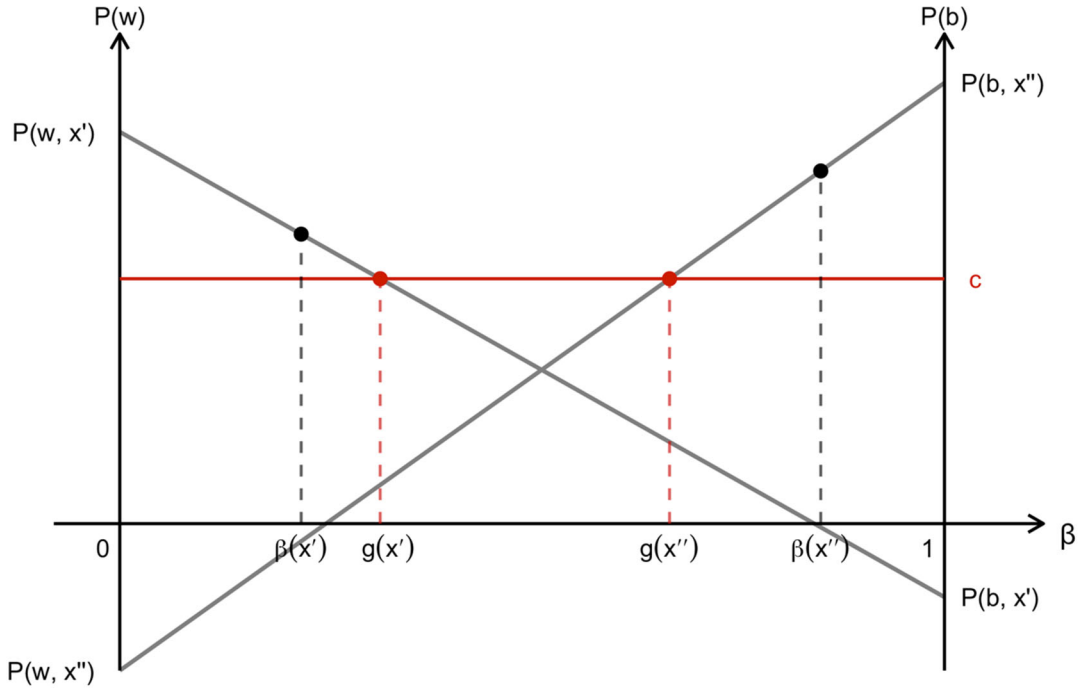
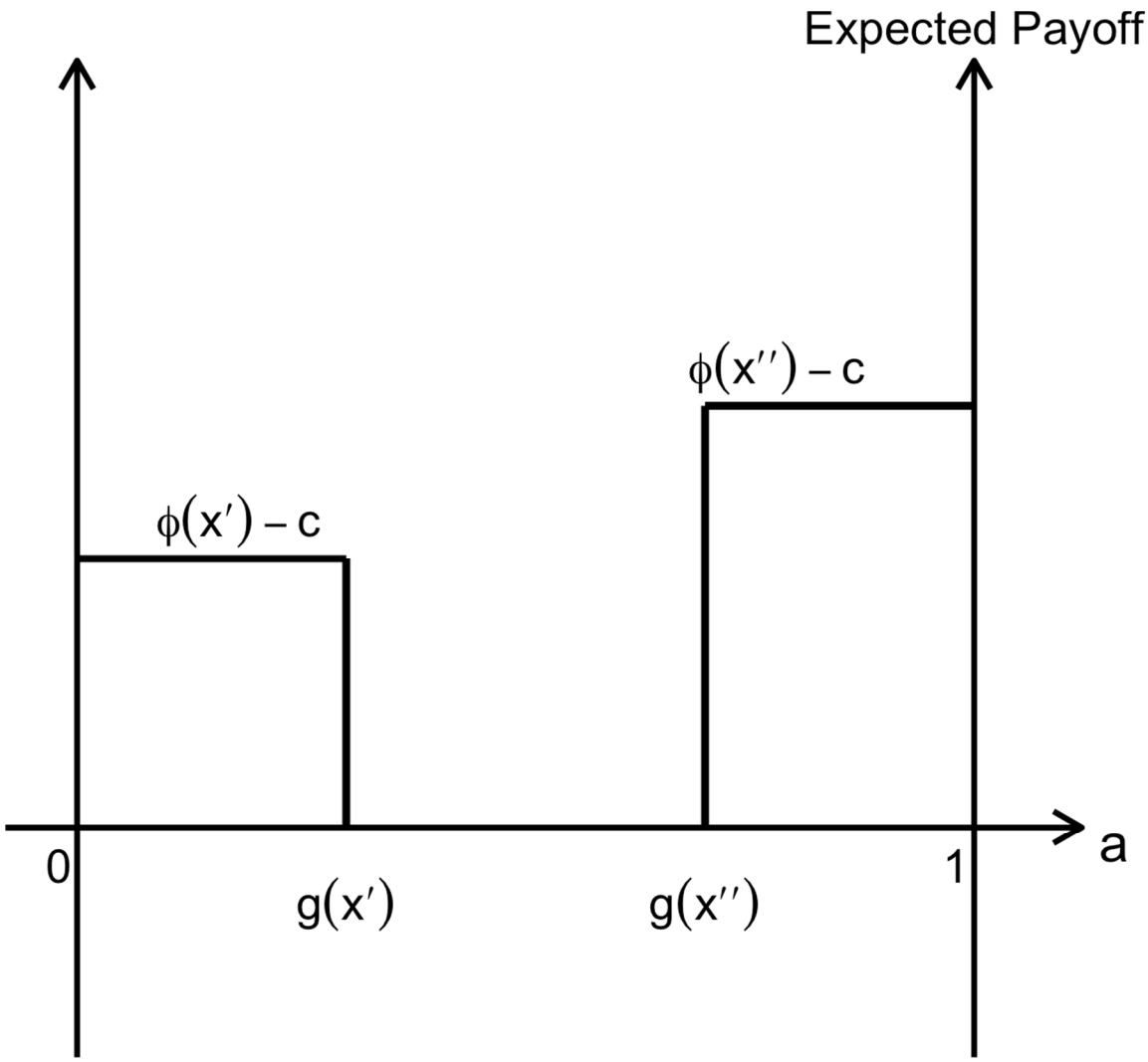


Figure 5 is an illustration of the calculation described in the last paragraph leading to the choice of a and selection. On the horizontal axis are the possible values from 0 to 1. On the vertical axis are the contributions among ambiguous descriptions to expected total payoff from using a as the constant proportion Black at which to evaluate expected payoff. The region of the horizontal axis that maximizes the value on the vertical axis determines the optimal non-discriminatory selection.

Black proportion) would maximize expected academic ability but would run counter to the spirit of personal non-discrimination. For instance, white applicants from Midwestern schools with great basketball knowledge would be denied admission while Black applicants from Northeastern schools with identical basketball knowledge would be admitted.

Figure 5: Illustration of the Choice of a and Selection



Notice that in both this example of proxy discrimination and in example A, the cure for proxy discrimination was effectively to consolidate the two descriptions involved into one description by requiring that either both be selected or neither be selected (the other alternative, selecting a description with negative net payoff, was clearly suboptimal). The decision-maker was required to act as if they could not distinguish between the two descriptions.

For future reference, the algebraic expression for the cross-over g(x) is

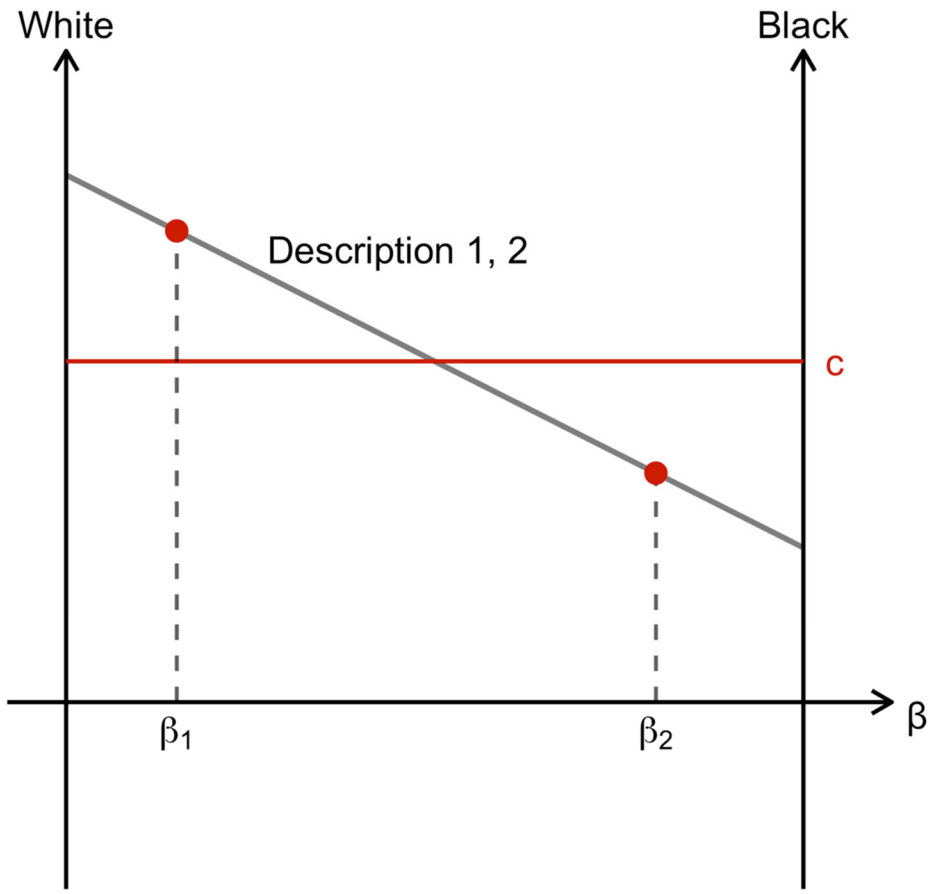
$$g(x) = \frac{c - p(w,x)}{p(b,x) - p(w,x)}.$$

Sanity check

With these tools in hand, we can show how our definition is consonant with the canonical intuitive example of proxy discrimination. Consider a problem with two descriptions. Description 1 is the same as description 2 as far as expected net payoffs from both Blacks and whites are concerned. For reasons unobserved by the decision-maker, the expected payoffs to Blacks are lower than the expected payoffs to whites. Description 2 differs only in some payoff-irrelevant detail. However, that payoff-irrelevant detail makes Blacks much more likely to satisfy description 2, although some Blacks and some whites satisfy each of the descriptions.

The bridge graph in figure 6 illustrates. Aside from the cost line, only one description line appears because the lines for the two descriptions are identical; racial composition is the only thing that sets the two descriptions apart. As drawn, in the optimum subject only to personal nondiscrimination, description 1 (the predominantly white description) is selected and description 2 (the predominantly Black description) is not. But if a selection rule is proxy non-discriminatory, then for every $a \in [0,1]$, $\varphi(a,1) = \varphi(a,2)$ and so $S(r,1|p,\beta,c) = S(r,2|p,\beta.c)\ for\ r = B,W$. So the optimal selection rule subject only to personal nondiscrimination is not proxy nondiscriminatory. And to satisfy proxy nondiscrimination, the decision-maker must consolidate the two descriptions and act as if they cannot distinguish between them.

Figure 6: Bridge Graph Illustration with Overlapped Descriptions



Our definition of proxy discrimination is consistent with intuition in the canonical case.

The general case

The general case is solved by generalizing the process that led to the diagram in figure 5.

Let

$$W = \{x | p(w, x) > p(b, x)\} \cap M$$

Let B denote the complement of W in M.

Fix a and consider the members of W that will be selected. Remember x' in figure 3. If the decision-maker acts as if the proportion Black in each description is a, then they will select all the members for W for which a < g(x), and only those members. Let $F^W(a)$ denote the subset of F that is selected when a is the assumed proportion Black:

$$F^W(a) = \{x | g(x) \geq a\} \cap W.$$

By similar reasoning,

$$F^B(a) = \{x | g(x) \leq a\} \cap B.$$

Hence the set of descriptions that are selected when a is used is

$$F(a) = F^W(a) \cup F^B(a).$$

While comparing a and g(x) determines whether an ambiguous description is selected when a is used, $\beta(x)$, the actual proportion Black, determines the contribution that that selection makes to net expected payoff $\varphi(\beta(x), x) - c$. Since $\beta(x)$ will in general be different from a, that net expected payoff may be negative as well as positive.

The aggregate net expected payoff in M from using a is called Q(a)[5]:

$$Q(a) = \sum_{x \, \epsilon F(a)} [\, \varphi(\beta(x), x) - c].$$

Figure 5 is an example of the graph of Q(a) for the simple example. Let A denote the set of values of a that maximize Q(a). In the example of Figure 5, the set A was a

[5] If we were to drop the assumption that all descriptions have equal population, the summands in this expression would have to be weighted by the population of the descriptions.

continuum: {a|a>g(x"}. For ambiguous descriptions, then, the optimal non-discriminatory policy selects descriptions that are in F(a) for some a that maximizes Q(.).

Summarizing then, the optimal non-discriminatory policy for r= b, w is

$$S(r,x) = 1 \; if \; x \in \Delta \; or \; if \; x \in F(a) for \; some \; a \in A; = 0 \; otherwise.$$

If we use this process to obtain the optimal non-discriminatory policy for a particular example and it coincides with the optimal policy subject only to personal nondiscrimination, then we will say that there is no proxy discrimination in this example. In practice when the number of ambiguous descriptions is small, we can start with the optimal personally nondiscriminatory policy, compare subsets of ambiguous descriptions to see whether the optimal selection rule for this subset is proxy nondiscriminatory or not.

Literature review

Ghili et al (2019), Pope and Sydnor (2011), and Yang and Dobbie (2020) are three papers closely related to this paper, although they all work in different frameworks and pay more attention than this paper does to estimating the parameters of their models, rather than assuming they were known. All three of these papers deal with quota problems rather than the cost problems that we concentrate on.

In our notation, Ghili et al adopt a different definition than we do of proxy non-discrimination—which they call "latent fairness"—but restrict the set of problems for which they can find optimal algorithms. Latent fairness is violated if for any $(x, x') \in X^2$ and either r=b or r=w, p(r,x) > p(r,x') and S(r,x) < S(r,x'). The selection rule must array descriptions in the same order as

either race arrays expected payoff. Ghili et al then restrict attention to parameter sets that satisfy a condition they call "separability": for any pair (x,x') of descriptions, either x weakly dominates x' or x' dominates x. They then show that if separability holds, there is a selection rule that satisfies both personal non-discrimination and latent fairness, and it is the obvious one for a quota problem: array descriptions in dominance order from the description that dominates all others down to the description that is dominated by all others, and select descriptions from the top until the quota is filled.

The major problem with extending the approach of Ghili et al to more general problems is that if p is not separable, any non-trivial search rule that satisfies personal non-discrimination is latent discriminatory. We define a search rule $S(r,x|p,\beta)\ as\ non-trivial\ if\ S(r,x|p,\beta)=1$ for at least one (r, x), and $S(r,x|p,\beta)=0$ for at least one (r, x).

To see this, suppose there are two descriptions x=1 and x=2, and p(b,1)>p(b,2) and p(w,1)<p(w,2). Separability does not hold. Personal non-discrimination requires S(b, 1)=S(w,1)=s(1) and S(b,2)=S(w,2)= s(2) where s(1) and s(2) are either 0 or 1. If s(1)=s(2), the stopping rule is trivial. So either s(1)>s(2) or s(1)<s(2). If s(1)<s(2), then latent fairness is violated for whites. If s(1)<s(2), latent fairness is violated for Blacks. All non-trivial stopping rules are discriminatory.

Pope and Sydnor (2011) work in a regression framework, and impose a condition we can call "strong separability": there is a constant d such that for all x, p(b,x) = p(w,x) + d. Strong separability implies separability, but the converse does not hold. Given strong separability, they find algorithms that are both person and proxy non-discriminatory.

Yang and Dobbie are more general in the sense that they consider the case where separability fails. For this case, they describe two algorithms and argue these algorithms are both person and proxy non-discriminatory. They do not argue that they are optimal.

Examples

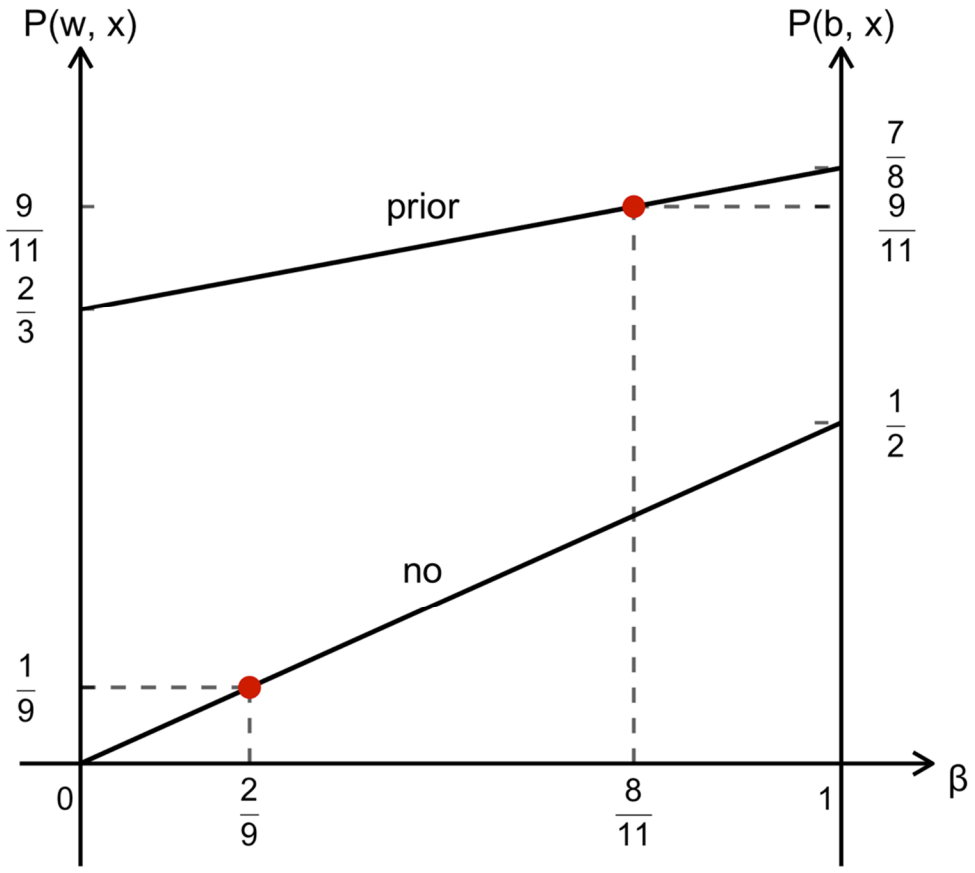
The Yang-Dobbie Toy Example

Yang and Dobbie (2021) present a toy example to illustrate their ideas. The question is which defendants should be detained before their trials, and the goal is to minimize failures to appear (FTA). To put this problem into our setting, there is a cost *c* for detaining a defendant (the financial cost to the government and social cost to the defendant and those who rely on the defendant). Defendants are either Black (B) or white (W). There are two descriptions: a record of prior convictions (x = prior) or no record of prior convictions (x = no). The data of their problem are:

Descriptions	p(B, x)	p(W, x)	ß(x)
x = prior	7/8	2/3	8/11
x = no	1/2	0	2/9

In addition, the proportion of Black defendants in the total population $\bar{\beta}$ is one-half. We illustrate this problem with the bridge graph as figure 7. This graph does not include a cost line because we are going to experiment with different costs.

Figure 7: Bridge Graph Illustration for the Yang-Dobbie Example



With this graph, the problem is easy: "prior" dominates "no" and $\varphi(\beta(prior), prior) > \varphi(\beta(no), no)$ so that no corrections are needed to preserve the dominance ranking. The optimal personally non-discriminatory selection rule is proxy non-discriminatory. Specifically, the optimal selected set under both personal and proxy non-discrimination is

$$the\ null\ set\ if\ c > \left(\frac{9}{11}\right)$$

$\{prior\}\ if\ \left(\frac{9}{11}\right) \geq c > \left(\frac{1}{9}\right)$

$\{prior, no\}\ if\ c \leq \left(\frac{1}{9}\right)$.

To prove rigorously that the optimal selection subject only to personal nondiscrimination is also proxy nondiscriminatory, consider each set of values of *c* in turn:

If $c > \left(\frac{9}{11}\right)$, then for any $a < \frac{8}{11}, \varphi(a, prior) < \frac{9}{11} < c\ and\ \varphi(a, no) < \frac{1}{2} < c$ and so proxy non-discrimination is satisfied.

If $\left(\frac{9}{11}\right) \geq c > \left(\frac{1}{9}\right)$, then for $a\ such\ that\ 2c > a \geq \frac{27}{5}\left(c - \frac{2}{3}\right), we\ have\ \varphi(a, prior) > c > \varphi(a, no)$ and so proxy non-discrimination is satisfied.

If $c \leq \left(\frac{1}{9}\right)$, then for any $a > \frac{2}{9}, \varphi(a, prior) > \frac{2}{3} > c\ and\ \varphi(a, no) = \frac{1}{2}a > \frac{1}{2} \cdot \frac{2}{9} = \frac{1}{9} \geq c$ and proxy nondiscrimination is satisfied.

In our set up, there are no biases or hidden proxies to worry about in this problem. For this problem, personal nondiscrimination plus optimality implies proxy nondiscrimination.

Yang and Dobbie seem to believe that something is discriminatory about standard algorithms dealing with this problem and propose two algorithms to eliminate proxy discrimination. They do not claim either is optimal. Both algorithms in fact guarantee proxy nondiscrimination, but in this problem are often suboptimal.

They call their first proposed algorithm minorities-as-whites. In our terminology, the algorithm selects the descriptions that can be justified by *a = 0:* description *x* is selected if and only

if $\varphi(0,x) \geq c$. This algorithm clearly guarantees proxy nondiscrimination but in the problem is often suboptimal. In the example the selected set is

$$the\ null\ set\ if\ c > \frac{1}{2}$$

$$\{prior\} if\ 0 \leq c \leq \frac{1}{2}$$

Relative to the nondiscriminatory optimum, this algorithm selects the null set instead of {prior} for $\frac{9}{11} \geq c > \frac{1}{1}$ and {prior} instead of {prior, no} for $c \geq \frac{1}{9}$.

They call their second proposal the color-blinding inputs algorithm. This algorithm is similar to the minorities-as-whites algorithm in that it uses a designated constant value of *a* to evaluate all descriptions but instead of using $a = 0$ like minorities as whites, it uses $a = \bar{\beta}$ (recall that $\bar{\beta}$ is the mean proportion Black across all descriptions).

To understand how this works, it's best to modify the example slightly so it meets the conditions that Pope and Sydnor (2011) implicitly imposed when they first proposed this algorithm. In our notation, that condition is

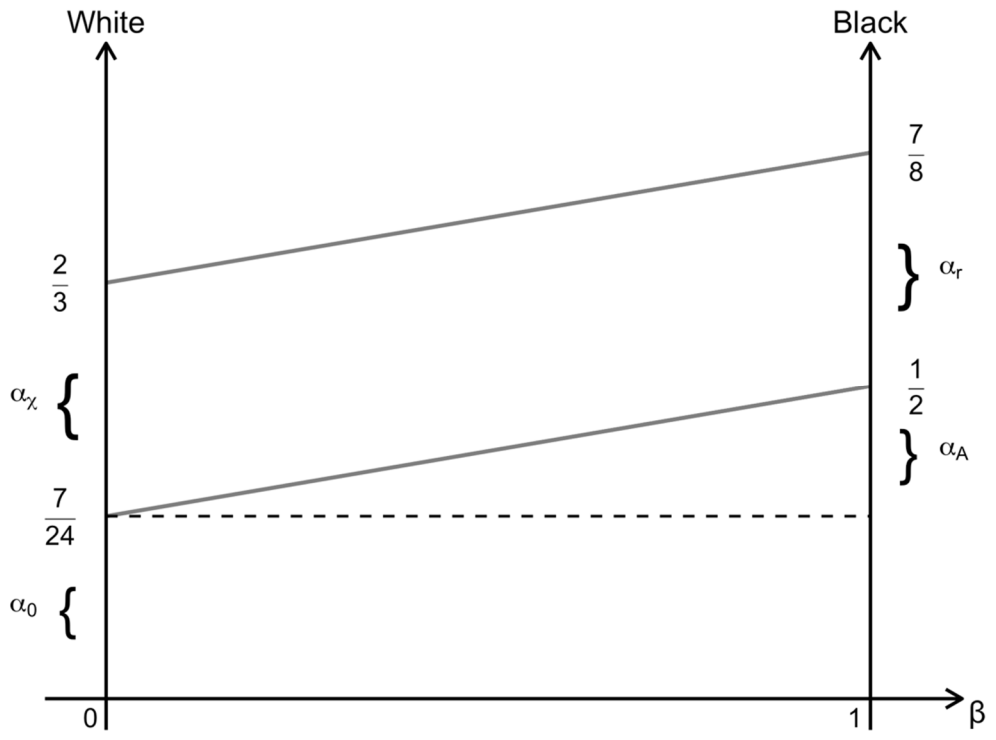
$$p(W, prior) - p(W, no) = p(B. prior) - p(B, no).$$

The lines of two descriptions are parallel as in figure 8 and so form a parallelogram. To accomplish this end, we amend $p(no, W) = 7/24$ instead of 0 and keep all other parameters of the problem. The data of the amended problem are now:

Descriptions	p(B, x)	p(W, x)	ß(x)
x = prior	7/8	2/3	8/11
x = no	1/2	7/24	2/9

The associated bridge diagram is figure 8.

Figure 8



Using this figure, Yang and Dobbie would fit the regression

$$p(\chi, r) = \alpha_0 + \alpha_r r + \alpha_\chi \chi$$

$$where\ r = 1\ if\ race = Black\ and\ r = 0\ if\ race = white\ and$$

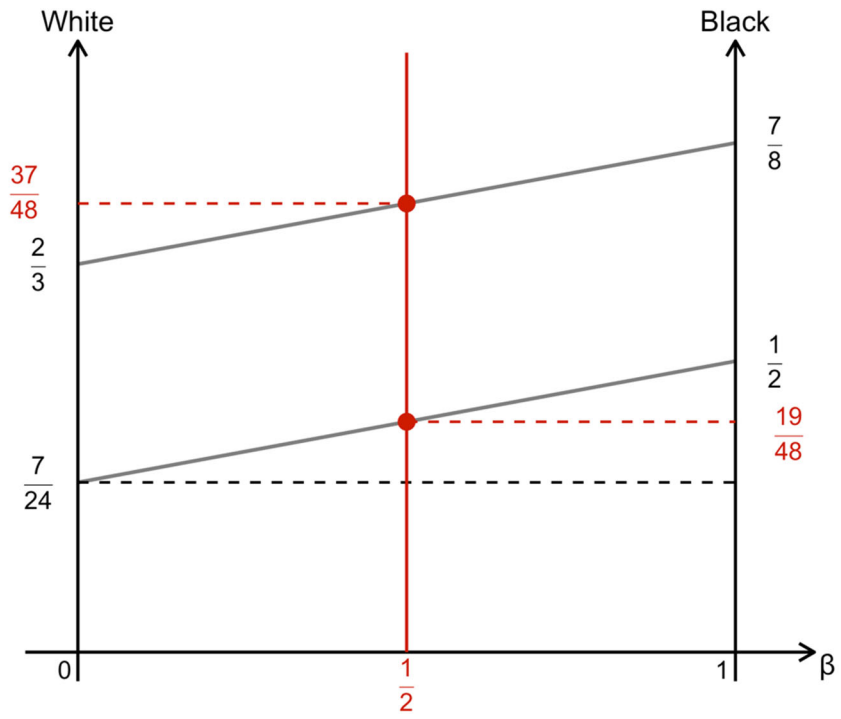
$$\chi = 1 \; if \; x = prior \; and \; race = Black; \;\; and \; \chi = 0 \; otherwise.$$

To fit the expression perfectly, you can just read the coefficients off figure 8:

$$\alpha_0 = \frac{7}{24}, \alpha_\chi = \frac{3}{8}, \alpha_r = 5/24.$$

They evaluate the resulting expression at $a = \bar{\beta}$= ½, as shown in figure 9.

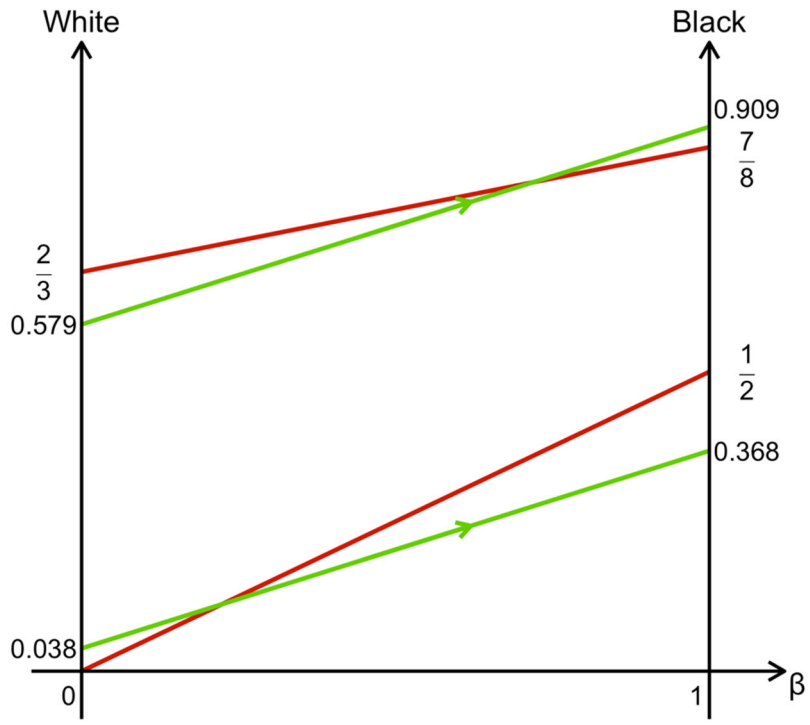
Figure 9



In the original Yang-Dobbie problem where the parallel condition is not met, what appears in the modified problem as a parallelogram is a quadrilateral in the original problem. As a result, the regression cannot be fit exactly. In the original problem, fitting the regression results in a parallelogram where the slope is between the slopes of the two lines in the quadrilateral of the original problem. This is shown in figure 10. With the new approximating parallelogram, proxy

nondiscrimination is guaranteed approximately. It could be guaranteed exactly if the regression exercise were dropped and the original quadrilateral were used instead of the approximating parallelogram (but they may not be easy when there is a continuum of descriptions instead of two).

Figure 10



Thus, we differ from Yang and Dobbie (2020) in that they recommend using one of two preset values of *a* to evaluate descriptions while we recommend searching for an optimal value of *a*. They (almost) always produce proxy nondiscriminatory results, and they do not claim that the results they get are either all the proxy nondiscriminatory selection rules or the optimal ones. So there is no direct conflict between our results and theirs. Perhaps we are being too permissive in allowing selection rules to be justified by any $a \in [0,1]$ but we have seen no reason to set limits. Accuracy is valuable (by definition).

New York City Search Decisions

Goel et al. (2016) analyzed 760,000 pedestrian stops that the New York Police Department (NYPD) made between 2008 and 2012 where the stated objective was to recover a weapon that was being carried illicitly (known as "CPW stops" for "criminal possession of a weapon"). This was part of a larger NYPD program called Stop Question and Frisk (SQF). As part of this study, Goel et al developed an algorithm for deciding which stopped pedestrians should be searched for weapons. This algorithm was simple, used information that the police already had, and was personally non-discriminatory. They showed that this algorithm was considerably more efficient than NYPD's actual methods: it could recover more weapons with fewer searches and concentrate less heavily on Blacks and Hispanics. We wanted to investigate whether this algorithm was proxy non-discriminatory.

The data that Goel et al used, however, were not fully public. Instead, we used a publicly available data set from the New York City public data depository for a longer period, 2008-2016. Professor Jeffrey Fagan kindly provided this data set. Following Goel et al (2016) we isolated stops where a search was conducted. In each of these stops, our dependent variable was whether a weapon was recovered. "Weapon" means a pistol, rifle, assault weapon, knife or other cutting instrument, or a similar object. We use the same independent variables that Goel et al used in their preferred algorithm. We also have information on the race of the pedestrians who were searched (although the NYPD uses a racial taxonomy different from the federal one.)

Both published papers that we are imitating as examples use results from people who were actually chosen to infer results about people who may not have been chosen. Arnold et al (2022)

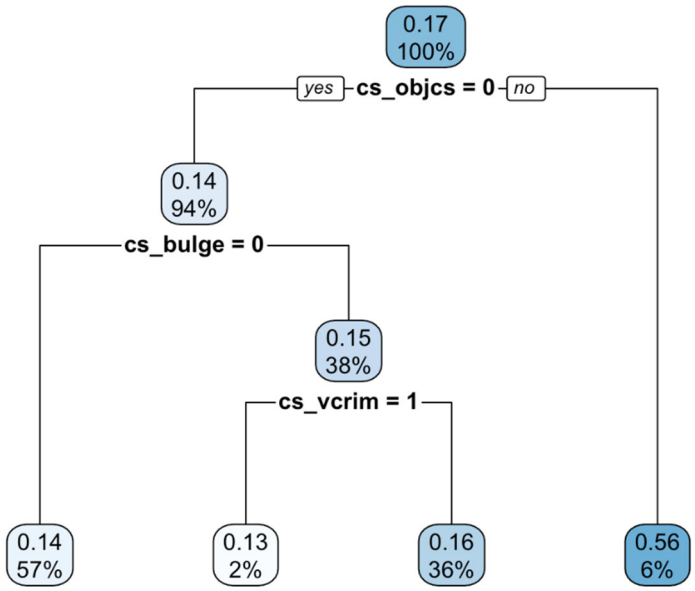
show how this problem can be addressed in some settings, but we will continue the simpler approach in the original papers.

Using rpart, we then constructed and pruned a classification tree for the full data set, which is shown as figure 11.[6] (In all the classification trees in this paper, we choose the complexity parameter that minimizes cross-validation error, following Diachkov 2023). We do this once excluding races as a set of independent variables and once including races. The pruned trees are exactly the same, whether or not race is included. The optimal unconstrained algorithm is personally non-discriminatory. (Using race would be over-fitting.)

To look for proxy discrimination, we examined the racial composition of the four leaves in figure 11. This is shown in table y (leaves are numbered from left to right in figure 11).

[6] We realize that classification trees are not the most sophisticated or powerful predictive tool. Their great pedagogic advantage is that their output is usually a small number of descriptions.

Figure 11: Classification Tree with NYC Search Data



Notes: cs_objcs = 0 means "primary stop circumstances are not suspicious objects";

Cs_bulge = 0 means "primary stop circumstances are not suspicious bulge";

Cs_vcrim = 1 means "primary stop circumstances are actions of violent crime".

We constructed bridge diagrams for the most important pairs of races and found no proxy discrimination. It is easy to see why. First, leaf 4 dominates the other three leaves, and is far above all three. The lowest hit rate (proportion of searches that recovered a weapon) was 0.53. For the other three leaves, the highest hit rate for any race is 0.34. So we can ignore leaf 4 in any search for proxy discrimination. Second, in the other three leaves, racial composition is quite similar. The table below shows the proportion of each leaf that each of the four largest races/ethnicities represents. Within each race, there is very little variation between leaves.

	Leaf 1	Leaf 2	Leaf 3
Black	.551	.567	.546
Black Hispanic	.066	.066	.064
White Hispanic	.256	.267	.262
White	.064	.050	.064

Knowing a person's leaf provides almost no information about their race/ethnicity. No matter what value of c we choose, we will not be able to find proxy discrimination in this example.

New York State pre-trial releases

Pointing to high correlations between race and key variables used in decision-making about whether arraignment judges should release arrested defendants before trial, Yang and Dobbie (2020) argued that regression-based algorithms using these variables would be proxy discriminatory. Having seen in the search data that optimal algorithms in arbitrary data sets might not be discriminatory, I asked Yang and Dobbie for their data. Unfortunately, it was also confidential. Professor Dobbie, however, suggested using New York State pre-trial release data, which is available at

https://ww2.nycourts.gov/pretrial-release-data-3313

From this website, we took data for 2022-2024 for the entire state for city, district, and superior courts and used the same selection criteria as Yang and Dobbie did: males, non-Hispanic Blacks and non-Hispanic whites with complete data who had at least one court appearance and who were released either on bail, with non-monetary conditions, or on their own recognizance. We have 205,919 Black defendants and 85,422 white defendants, the majority of whom were released on their own recognizance. Thus $\bar{\beta} = .707$.

Our independent variable is what we call "return failure." We consider a defendant to be a return failure if either they were rearrested or failed to appear for their scheduled trial (specifically, a warrant was issued between arraignment and disposition). Yang and Dobbie considered only the latter outcome, but we added rearrests to make it more likely that pruned trees would open up.

For independent variables, we followed Yang and Dobbie with the following exceptions: we could not find one variable in the state data (prior failures to appear), one variable we had to modify (geography), and four new variables we added. For geography, Yang and Dobbie used county income, which took five different values because New York City includes five counties. New York State has 62 counties that are divided into 13 judicial districts. We simplified the districts to 10, composed of the following counties, and included a dummy variable for each:

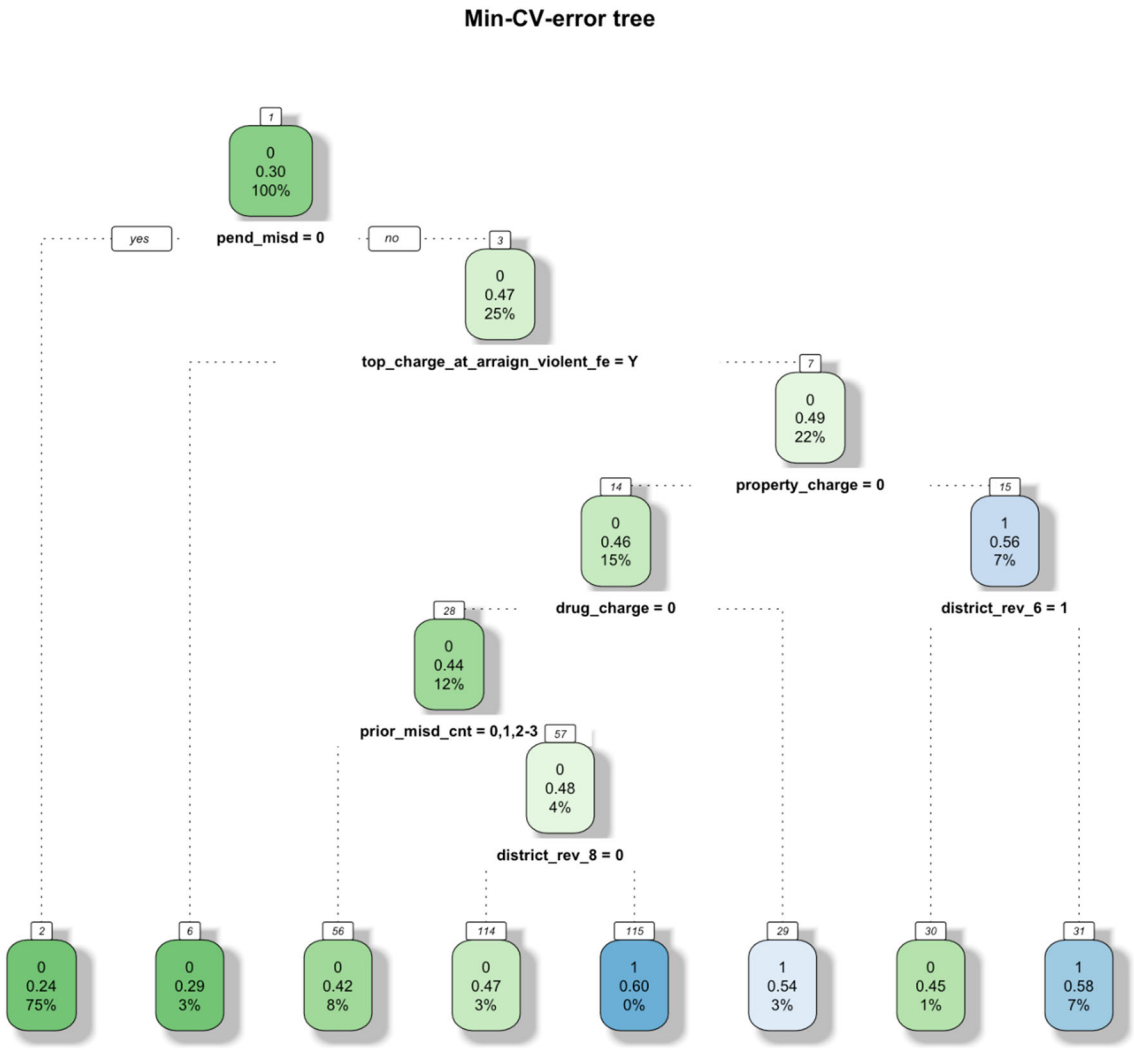
1. New York County (Manhattan)
2. Kings County (Brooklyn)
3. Monroe and surrounding counties (Rochester area) (7^{th} judicial district)
4. Erie and surrounding counties (Buffalo area) (8^{th} judicial district)
5. Dutchess, Orange, Putnam, Rockland, and Westchester (9^{th} judicial district)
6. Nassau and Suffolk
7. Queens

8. Bronx
9. Richmond (Staten Island)
10. Everything else

The new independent variables are whether at the time of arrest, the defendant was on probation or parole ("supervision"), had pending violent felony arrests ("pend vfo"), had pending nonviolent felony arrests ("pend nonvfo"), or had pending misdemeanor arrests ("pend misd").

Following the same practice that we used for NYC searches, we fit classification trees with and without using the race variables. Once again there was no difference: the optimal unconstrained tree was the same as the optimal tree constrained to be personally non-discriminatory. The tree is figure 12.

Figure 12: Classification Tree with NYS Pretrial Release Data



Notes: Pend_misd = 0 means "having one or more pending arrests for fingerprintable misdemeanor offenses";

Top_charge_at_arraign_violent_fe = Y means "Yes for the top arraign violent felony indicator";

Property_charge = 0 means "Arraignment charge category is not property or larceny";

Drug_charge = 0 means "Arraignment charge category is not drug";

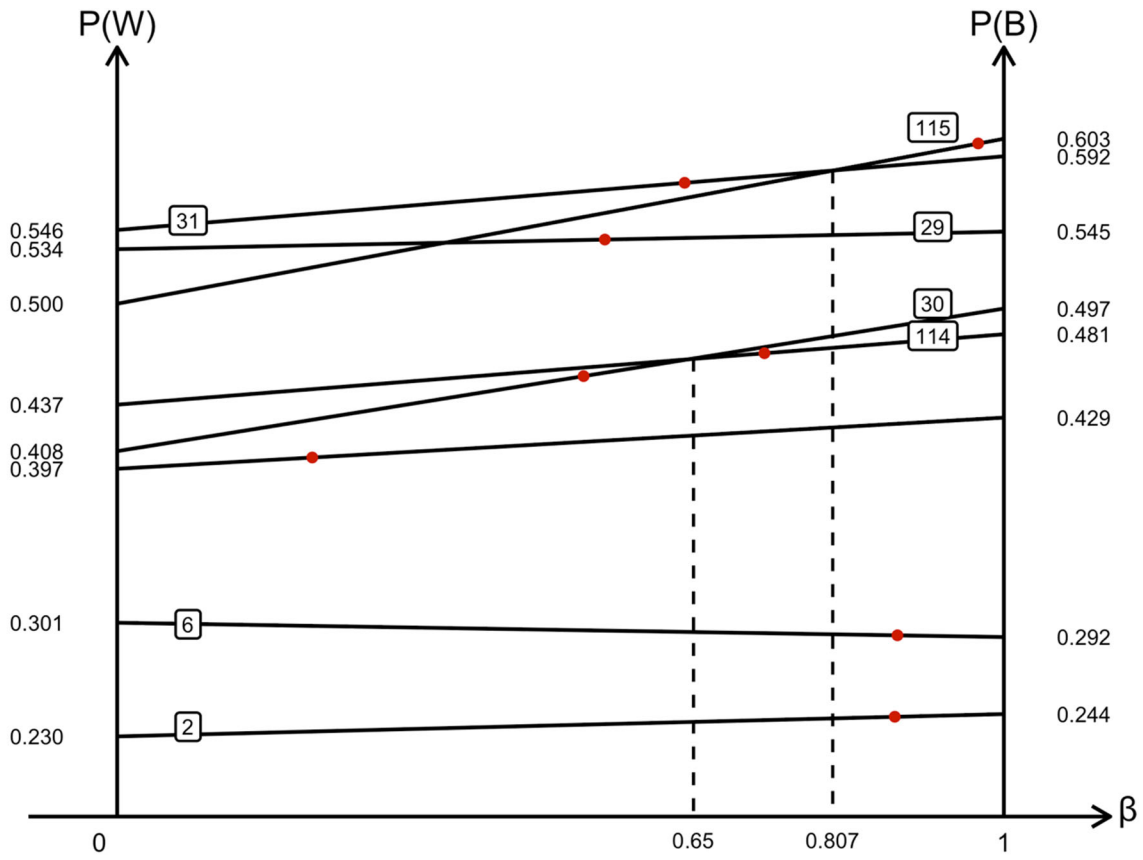
District_rev_6 = 1 means "in Nassau and Suffolk";

Prior_misd_cnt =0,1, 2-3 means "the count of convictions for misdemeanor offenses is 0, 1, 2-3";

District_rev_8 = 0 means "not in Bronx".

Next we used this tree to develop a bridge diagram (without costs so that we could experiment with different levels of cost). This is figure 13.

Figure 13: Bridge Graph for NYS Pretrial Detention Without Cost Line

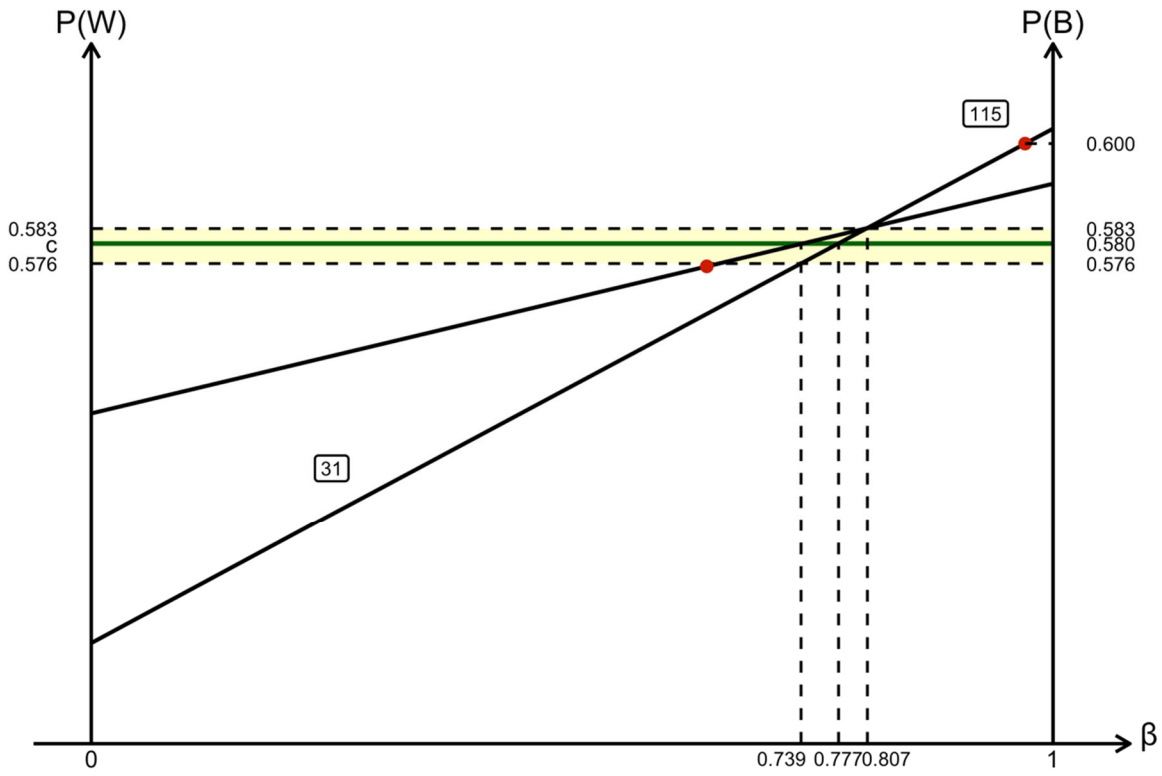


No pair of descriptions in this figure looks like figure 1 with parallel sloping lines and widely different racial compositions. The description lines are pretty widely dispersed vertically, and thus it is possible to analyze them a few at a time since for any cost only at most a few descriptions will be ambiguous. We concentrate on description lines that cross and find proxy discrimination for certain values of in two cases—one involving leaves 115 and 31 and one involving leaves 114 and

30. The lines for leaves 115 and 29 also cross, but no proxy discrimination involving these two leaves occurs.

Consider leaves 115 and 31 first. Proxy discrimination requires that both be ambiguous, and hence proxy discrimination can occur only for costs that satisfy $.592 \geq c \geq .546$ Since $\varphi(\beta(11), 11) = .600 > \varphi(\beta(15), 15) = .576$, leaf 31 will never be selected in the optimal personally nondiscriminatory policy when leaf 115 is not, and so proxy discrimination could happen in the optimal personally nondiscriminatory policy only if leaf 31 were selected and leaf 115 were not selected. Hence the range of relevant cost is reduced to $.592 \geq c \geq .576$. The two lines cross where $\beta = .807$ and expected payoff is .583. This is illustrated in figure 13a, which is the relevant portion of figure 13 magnified, plus a cost line.

Figure 13a: Bridge Graph for Leaves 115 and 31



Proxy discrimination in this case occurs for $.583 \geq c \geq .576$. (In figure 13a, this range is shaded.) To see this for a particular value of cost, set $c = .580$, as drawn In figure 13a. Let a vary from 0 to 1, and check which descriptions are selected and how expected aggregate payoff Q varies.

The cost line intersects the leaf 31 line at $a = .739$. For values of a less than this, the cost line is above both the leaf 31 line and the leaf 115 line, and so neither leaf would be selected. So Q(a)= 0 in this range.

The cost line intersects the leaf 115 line at $a = .777$. For $.739 \leq a \leq .777$, the cost line is below the leaf 31 line and above the leaf 115 line. Leaf 31 is selected and leaf 115 is not selected (the opposite of what happens in the optimal personally nondiscriminatory rule). The expected per person net payoff from leaf 31 is .576 - .580 = - .004. This must be weighted by the population of leaf 31, which is 9506, and so Q(a) = 9506 (-.004) = - 38.024 for a in this range.

Finally, for $a \geq .777$ both the leaf 115 line and the leaf 31 line are above the cost line and so both leaves are selected. The expected per person payoff from leaf 115 is .600 -.580 =.020. The population of leaf 115 is 487 and so the expected payoff from including leaf 31 is 9.74. Since leaf 31 is also included in this range, Q(a) = -38.024 + 9.74 = -28.284 in this range.

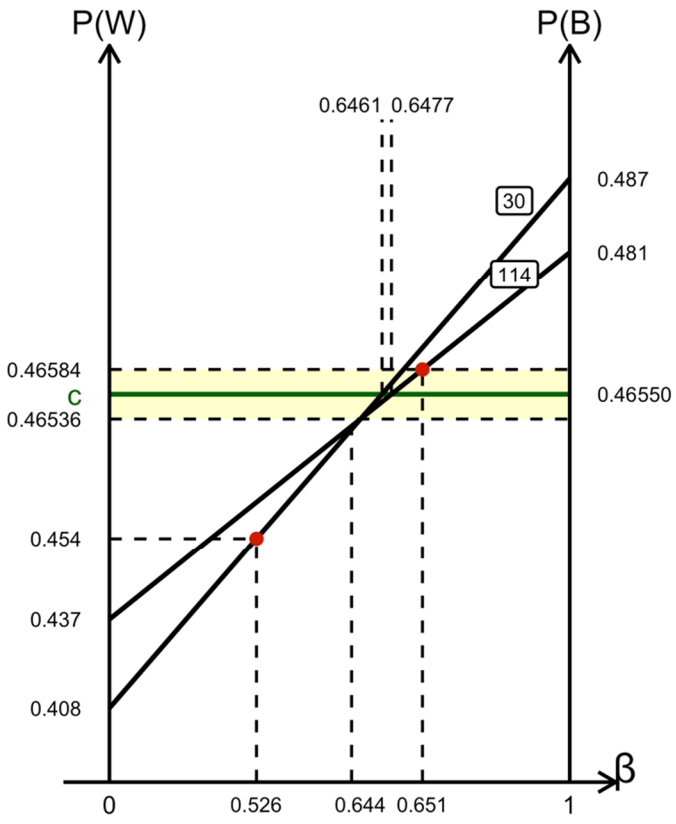
Thus the optimal proxy nondiscriminatory policy is to select neither description, which is different from the optimal personally nondiscriminatory policy, which is to select only leaf 115. Satisfying proxy nondiscrimination is not costless; in the absence of this constraint expected aggregate payoff would have been 9.74 (the expected payoff from leaf 115) instead of zero.

Once again, the cure for proxy discrimination is equivalent to consolidating the two leaves and requiring that either both be selected or neither. But consolidating the leaves is much more

complicated than the traditional idea of ignoring some variable. In both leaves the detainees have other pending charges for misdemeanors and the top charge is not a violent crime. But within that class, leaf 115 consists of people charged with drug offenses in the Bronx and leaf 31 consists of people charged with property offenses but not in Nassau and Suffolk.

Next consider leaves 114 and 30. Since $\varphi(\beta(10), 10) = .46564 > \varphi(\beta(14), 14) = .454$ leaf 15 will never be selected in the optimal personally nondiscriminatory policy when leaf 115 is not. . The two lines cross where $\beta = .644$ and expected payoff is .46536. This is illustrated in figure 13b, which is again the relevant portion of figure 13 magnified, plus a cost line.

Figure 13b: Bridge Graph for Leaves 114 and 30



Proxy discrimination in this case occurs for $.46564 \geq c \geq .46536$. (In figure 13b, this range is shaded.) To see this for a particular value of cost, set $c = .46550$, as drawn In figure 13b. With this cost, leaf 114 is selected and leaf 30 is not in the optimal personally nondiscriminatory rule.

Let a vary from 0 to 1, and check which descriptions are selected and how expected aggregate payoff Q varies.

The cost line intersects the leaf 30 line at $a = .6461$. For values of a less than this, the cost line is above both the leaf 30 line and the leaf 114 line, and so neither leaf would be selected. So Q(a)= 0 in this range.

The cost line intersects the leaf 115 line at $a = .6477$. For $.6461 \leq a \leq .6477$, the cost line is below the leaf 30 line and above the leaf 114 line. Leaf 30 is selected and leaf 114 is not selected (the opposite of what happens in the optimal personally nondiscriminatory rule). The population of leaf 30 is 1394 and the expected payoff is -16.0. So Q(a) = -16.0 for a in this range.

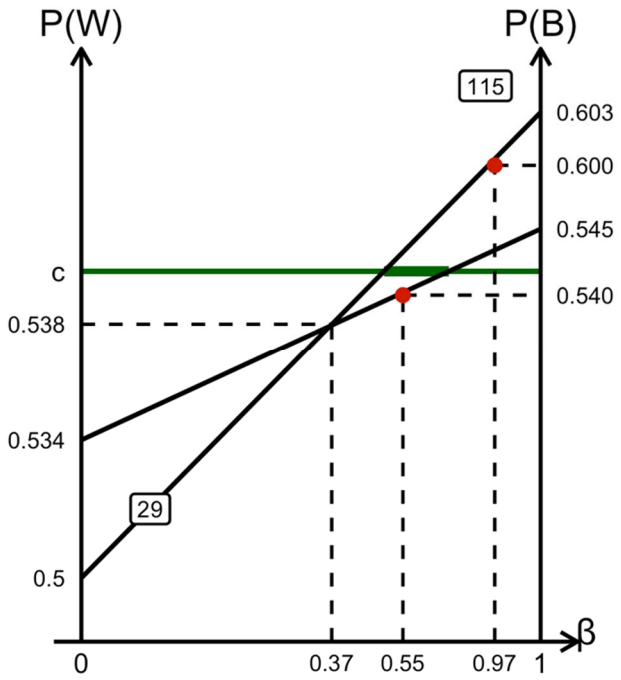
Finally, for $a \geq .6477$ both the leaf 114 line and the leaf 30 line are above the cost line and so both leaves are selected. The expected per person payoff from leaf 114 is .46564 -.46550 =.00014. The population of leaf 114 is 4863 and so the expected payoff from including leaf 114 is 0.68. Since leaf 30 is also included in this range, Q(a) = -16.0 + 0.68 = -15.32 in this range.

So the optimal proxy nondiscriminatory rule is to select no descriptions. This differs from the optimal personally nondiscriminatory algorithm, which is to select only leaf 114. But since the expected payoff from leaf 114 is only 0.68, the net cost of enforcing proxy nondiscrimination is small.

Leaves 114 and 30 are also quite dissimilar in description. Both leaves contain detainees with pending misdemeanor charges, but leaf 30 contains people charged with property offenses in Nassau and Suffolk, while leaf 114 contains people outside the Bronx with more than three prior misdemeanor convictions who are not charged with violent, property or drug offenses.

Finally, to show that not all crossings imply proxy nondiscrimination, consider leaves 115 and 29, which is magnified in figure 13c. For $.600 \geq c \geq .540$ in this case, the optimal personally nondiscriminatory rule selects leaf 115 and not leaf 29. Consider any cost in this range. When a is between the value of β at the intersection of this cost line and the leaf 115 line, and either 1 or the value of β at the intersection of this cost line and the leaf 29 line, leaf 115 is selected and leaf 29 is not. This segment is highlighted in figure 13c. This is the highest value of Q(a), and so the optimal proxy nondiscriminatory rule is the same as the optimal personally nondiscriminatory rule.

Figure 13c: Bridge Graph for Leaves 115 and 29

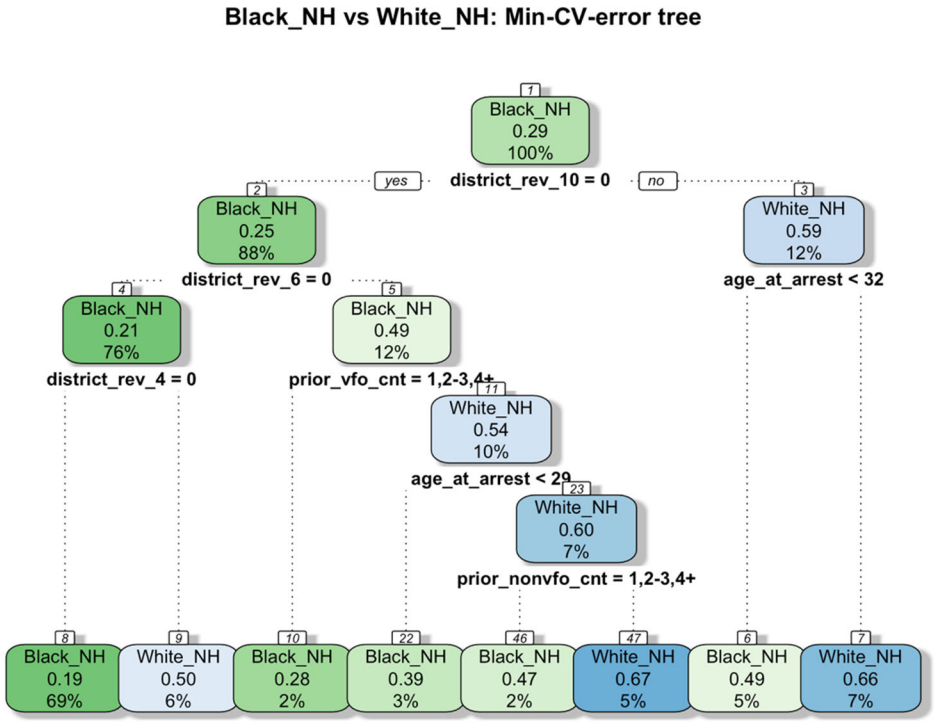


We also experimented with algorithms that were not optimal. At the extremes, suppose a malevolent racist were to design a facially neutral algorithm that was maximally anti-Black. They would first fit a classification tree using the non-racial variables to predict whether observations were of Black defendants or white defendants; that is, they would find propensity scores. Then they would lie about what this tree was—they would say that it predicted the probability of return failure instead of the probability of being Black. Given any cost, the leaves that would be selected to be detained would be the leaves with the greatest proportion of Blacks.

Perhaps this is the worst-case scenario that proponents of proxy-exile are worried about. We call this the diabolical tree.

Figure 14 shows the diabolical tree, and figure 15 is the associated bridge graph. In the most extreme case, our malevolent racist would lie and tell the decision-maker that the tree predicted the probability that a defendants would successfully return without detention (not that it was the probability of being white), and that defendants in the leaves with low values (really the leaves with the highest proportion Black) should be detained.

Figure 14: Diabolical Tree



Notes: district_rev_10 = 0 means "not in everything else in definition of districts";

District_rev_6 = 0 means "not in Nassau and Suffolk";

Age_at_arrest means "arrest age";

District_rev_4 = 0 means "not in Erie and surrounding counties (Buffalo area) (8th judicial district)";

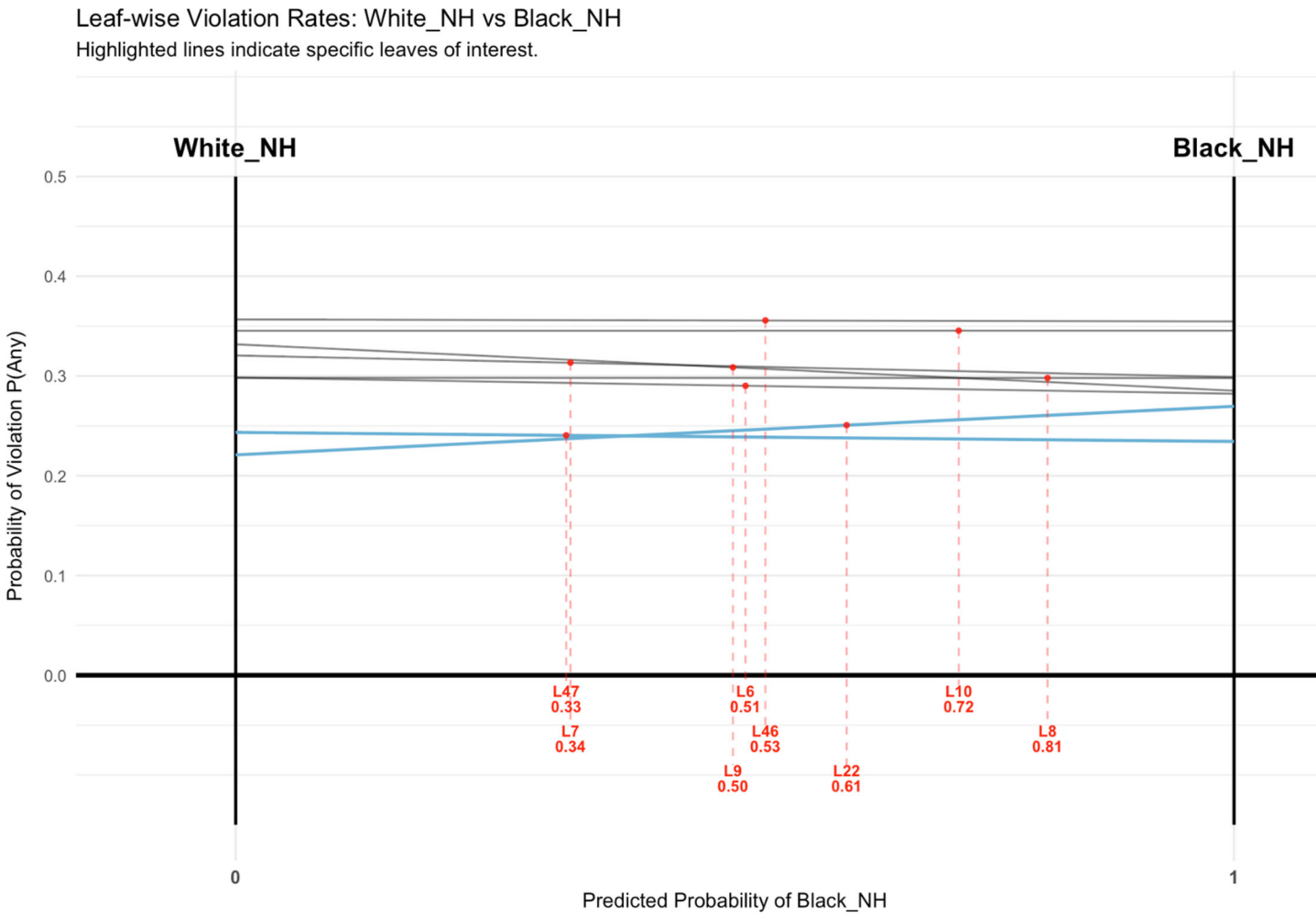
Prior_vfo_cnt = 1,2-3,4+ means "count of prior convictions for violent felonies is 1,2-3,4+"

Prior_nonvfo_cnt = 1,2-3,4+ means "count of prior convictions for non-violent felony offenses is 1,2-3,4+"

Since the designer of the diabolical tree was trying to proxy for race as best they could, an obvious question is whether using the tree is likely to cause an extraordinary amount of proxy discrimination. The answer is no. Only for the case of $c \in [.240, .241]$ does proxy discrimination occur. It involves leaves 22 and 47. In this range, the optimal policy without a proxy non-discrimination constraint is to select both leaves, but proxy non-discrimination requires that only

one be selected (the picture looks like figure 4). Proxy discrimination is no more common in the diabolical tree than it is in the optimal tree (figure 12).

Figure 15: Bridge Graph for the Diabolical Tree

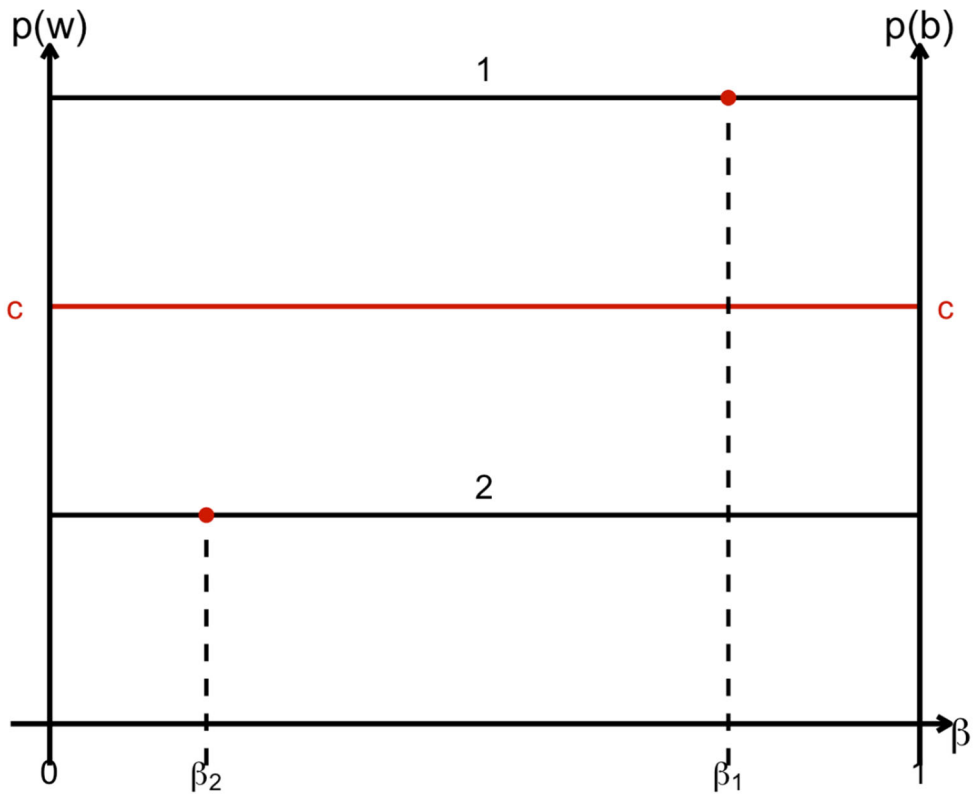


Since it seems that a tree rife with proxies should cause proxy discrimination, one reaction to the result in this case is that we are using a poor definition of proxy discrimination. I don't think so. The literature does not support such a conclusion. Because the lines cross, neither Pope and Sydnor nor Ghili et al would be able to say whether or not the bridge graph in figure 15 depicted proxy discrimination.

Yang and Dobbie would find no proxy discrimination with the diabolical tree for the color-blinding-inputs algorithm but would find it in two cases for minorities-as-whites. I will consider only quota problems, since that is what they wrote about. Consider minorities-as-whites first. Then the question is when the ordering of x by $\varphi(\beta(x), x)$ differs from the ordering of x by $p(W, x) = \varphi(0, x)$. Different ordering occurs for leaves 22 and 47 and for leaves 8 and 6. If the quota is reached between either of these pairs, proxy discrimination according to minorities-as-whites will occur. Under color-blinding-inputs, the question is when the ordering of x by $\varphi\big(x, \beta(x)\big)$ differs from the ordering of x by $\varphi\big(\bar{\beta}, x\big) = \varphi(.707, x)$. This never occurs in the diabolical tree.

Figure 16 shows a bridge graph where the existing literature speaks with one voice. In this graph, descriptions are highly predictive of race; that is, either description is a strong proxy for race. But Pope and Sydnor, Ghili et al, and both algorithms of Yang and Dobbie agree that description 1 should be chosen. This is the description that is optimal subject only to personal non-discrimination and so the proxy non-discrimination constraint is not binding. There is no proxy discrimination in this problem. It takes more than proxies to create proxy discrimination.

Figure 16



Why proxy discrimination is not ubiquitous

Much of the existing literature on algorithmic discrimination (surveyed in Yang and Dobbie 2020) seems to imply that proxy discrimination is both inevitable and incurable when explanatory variables are correlated with race. "No explanatory variables correlated with race" translates approximately in our set up to "$\beta(x)$ is the same for all x," and so we know the converse of this proposition is true: "If no explanatory variables are correlated with race, proxy discrimination cannot occur."

But in the first (theoretical) part of this paper we showed that if proxy discrimination did occur it could be cured; and in the second (empirical) part we showed that even in well-known data sets believed to be susceptible to proxy discrimination, it was present but not ubiquitous.

Why was that? There are three reasons.

First, optimality depends only on ambiguous descriptions—or in words more common with economists, only the margin determines the optimum. Descriptions that are not ambiguous are always ignored, even if they look like they would cause proxy discrimination. In fact, proxy discrimination can occur only if at least two descriptions are ambiguous, and it depends on the relationship between at least two ambiguous descriptions.

Second, proxy discrimination requires that at least two ambiguous descriptions differ substantially in racial composition, even though their payoffs are similar enough that they are both ambiguous. (Differences in racial composition among non-ambiguous descriptions are of course irrelevant.)

Finally, proxy discrimination requires that within the same ambiguous description, payoff probabilities differ substantially between races. This makes for steep lines in our bridge graphs and implies that differences in racial composition cause large differences in expected payoff. Using a proxy for race can be profitable only if racial composition makes a big difference to expected payoff.

In the classification tree format we have been using, this steepness requirement explains the relationship between personal nondiscrimination and proxy nondiscrimination. In the search for accuracy, a node is likely to split on race only if splitting on race is permitted and the inter-racial difference in expected payoff is great. So, when the personal nondiscrimination constraint is not binding—that is, when within-node interracial payoff differences were not great enough to merit a split, the proxy nondiscrimination constraint is also likely not to be binding.

These three requirements, moreover, are only necessary for proxy discrimination to occur. Even together, they are not sufficient. The association between proxy discrimination and Simpson's paradox is worth remembering. There are many examples of Simpson's paradox, but it is not ubiquitous. The same is true for proxy discrimination.

Conclusion

The good news from this paper is that decision makers need not quiver in fear of using strong automated optimality-seeking algorithms and variables correlated with race to make important decisions. They may encounter proxy discrimination but it's not likely to be everywhere, and if they encounter it, they can stomp it out.

But this good news is not cause for euphoria. This paper started with the assumption that the data of any problem (p, β, c) were known. The statistical and philosophical difficulties of learning those parameters in any non-trivial problem remain huge, although many scholars are trying to learn them. Machine learning also introduces many issues of its own; see for instance Kleinbard et al (2018). Artificial intelligence programs that learn from imitating existing practices are not likely to be much help. Moreover, many important decisions, from whom to choose as a life partner or animal companion to how to respond to a stranger at your door in the middle of the night, are also unlikely ever to be mechanized. And not every decision-maker is likely to be willing to sacrifice private gain for racial non-discrimination.

Racial justice requires solving many large and difficult problems. Proxy discrimination may not be prominent among them.

References

David Arnold, Will S. Dobbie, and Peter Hull, 2022, Measuring racial discrimination in bail decisions, American Economic Review 112(9): 2992-3038.

---, 2025, Building non-discriminatory algorithms in selected data, NBER working paper 32403.

Dallas Card and Noah A. Smith, 2020, On consequentialism and fairness, Frontiers of Artificial Intelligence, Volume 3 - 2020 | https://doi.org/10.3389/frai.2020.00034.

Corbett-Davies, S., Gaebler, J., Nilforoshan, H., Shroff, R. & Goel, S. The measure and mismeasure of fairness. *J. Mach. Learn. Res.* (in the press).

Chohlas-Wood, A., Coots, M., Goel, S. *et al.* Designing equitable algorithms. *Nat Comput Sci* **3**, 601–610 (2023). https://doi.org/10.1038/s43588-023-00485-4

Diachkov, Dima, 2023, Hyperparameter tuning and pruning: More about decision trees in R with rpart," R for Applied Economics Guide, part #31. Accessed at https://medium.com/data-and-beyond/hyperparameter-tuning-and-pruning-more-about-k-means-clustering-in-r-with-rpart-9a405685a09b

Ghili, Soheil, Ehsan Kazemi, and Amin Karbasi, 2019, Eliminating latent discrimination: Train then mask, Proceedings of the AAAI Conference on Artificial Intelligence 33(01): 3672-3680.

Goel, Sharad, Justin M. Rao, and Ravi Shroff. 2016, Precinct or prejudice? Understanding racial disparities in New York City's stop-and-frisk policy, Journal of Applied Statistics 10(1): 365-394.

Kleinberg, Jon, Jens Ludwig, Sendhil Mullainathan, and Cass Sunstein, 2018, Discrimination in the age of algorithms, Journal of Legal Studies 10: 113-174.

Brendan O'Flaherty, Rajiv Sethi, and Morgan Williams, The nature, detection, and avoidance of harmful discrimination in criminal justice, 2024, Journal of Policy Analysis and Management 43: 289-320.

Pople, Devin, and Justin Sydnor, 2011, Implementing anti-discrimination policies in statistical profiling models, American Economic Journal: Economic Policy 3 (3): 206-231.

Rose, Evan, 2023, A constructivist perspective on empirical discrimination research, Journal of Economic Literature 61(3): 906-933,

Sprenger, Jan and Naftali Weinberger, 2021, Simpson's Paradox, The Stanford Encyclopedia of Philosophy (summer 2021 edition), Edward N. Zalta, ed. Accessed at <https://plato.stanford.edu/archives/sum2021/entries/paradox-simpson/>.

Yang, Crystal and Will Dobbie, 2020, Equal protection under algorithms: A new statistical and legal framework, Michigan Law Review 119 (2): 291-396.

Appendix: Regressions

Standard regression results are not sufficient for finding fully non-discriminatory selection rules because they do not contain information about the racial composition of descriptions ($\beta(x)$).). However, racial composition information is often available in the source data that the regression results summarize, and so optimal non-discriminatory selection rules can sometimes be constructed from regressions. Ideally, if covariates are continuous variables, racial composition should be continuous too, which is rare in observational data. So even when standard racial composition variables are available in the source data, some adjustments may have to be made, like discretizing the continuous variables to match the information on racial composition. It may often be easier to rely on trees, forests or other methods *ab initio*, rather than adding racial composition to an existing regression.

Nevetheless, in this section we will show how a regression with racial composition information appended can form the basis for an optimal non-discriminatory policy. The steps we outlined in section x should be followed: finding descriptions that dominate the cost scalar, finding the set of ambiguous descriptions, finding the sets $F^b(a)$ and $F^w(a)$, finding the function $Q(a)$, and maximizing that function. In this section, we illustrate with a fictitious example.

Example B

In this example, a description is a pair of continuous variables $x = (z, y)$. Think of this perhaps as a college admissions process and z as a measure of cognitive achievement like SAT or GRE and y as a measure of social and behavioral skills like grit or maturity. The college seeks to

maximize the number of successful students it graduates, subject to a long-term budget constraint. Each admitted student costs the college an amount c to educate, where $c \in (1,2)$.

Regressions have found that the probability of becoming a successful graduate conditional on admission is a linear function of z and y, but the function differs between races:

$$p(b,z,y) = z + 2y$$

$$p(w,z,y) = 2z + y$$

In the population, (z,y) is distributed bivariate uniformly over the following rectangle: $\frac{c-1}{2} \leq z \leq \frac{c}{2}, 0 \leq y \leq 1$. (This triangle is designed to keep probabilities non-negative.)

Unconstrained optimal selection policy

This information, without any information about racial composition, is sufficient to find the unconstrained optimal selection policy. This is

$$S(w,(z,y)|p,\beta,c) = \{(z,y)|2z + y \geq c\}$$

$$S(b,(z,y)|p,\beta,c) = \{(z,y)|z + 2y \geq c\}$$

We show these selection sets in figure A-1. The area where the sets overlap is the set of descriptions that dominate the cost scalar. The set M of ambiguous descriptions consists of two subsets: B in the northwest corner where $p(b,z,y) > p(w,z,y)$ and W in the southeast corner where $p(w,z,y) > p(b,z,y)$.

The marginal rate of substitution between y and z for whites (the slope of the line in figure A-1 for whites) is (-2). The marginal rate of substitution for Blacks is (-1/2).

Optimal policy only constrained by personal non-discrimination

To find optimal partially and fully non-discriminatory policies, we must specify racial composition. For this example, we choose a simple form, given the parameters β_1 and β_2:

$$\beta(z,y) = \beta_1 \in [0,1] \; if \; z \leq {}^{c}/_{3}$$

$$\beta(z,y) = \beta_2 < \beta_1 \; if \; z > {}^{c}/_{3}$$

We also constrain β_2 to be positive.

The optimal selection policy constrained only by personal non-discrimination is

$$S(r,(z.y)|p,\beta,c) = \left\{(z,y)\middle| y \geq \frac{c}{1+\beta_1} - \frac{2-\beta_1}{1+\beta_1}z\right\} \; if \; z \leq {}^{c}/_{3}$$

$$= \left\{(z,y)\middle| y \geq \frac{c}{1+\beta_2} - \frac{2-\beta_1}{1+\beta_2}z\right\} \; if \; z > {}^{c}/_{3}$$

Figure A-2 illustrates.

The marginal rate of substitution is $\frac{2-\beta_1}{1+\beta_1}$ for $z \leq {}^{c}/_{3}$ and $\frac{2-\beta_1}{1+\beta_2}$ for $z > {}^{c}/_{3}$. It is easy to see that

$$2 > \frac{2-\beta_1}{1+\beta_2} > \frac{2-\beta_1}{1+\beta_1} > {}^{1}/_{2}.$$

The boundary of the area selected under personal non-discrimination is always below the envelope of the areas selected without constraints, and the area not selected in this case is concave to the origin. The boundary always runs through the set M.

The optimal fully non-discriminatory selection policy

To find the optimal fully non-discriminatory policy, start by considering an arbitrary description (z, y) in B, the region where $y \geq z$. From the discussion of the general case in section x, the description (z, y) will be selected if and only if g((z,yz)) <a. This simplifies to selecting (z, y) if and only if

(1) $y > \frac{c}{1+a} - \frac{2-a}{1+a} z = y(a, z)$.

If a=0, this condition coincides with the upper boundary of B, and so no descriptions in B are selected. If a =(1/2), the condition coincides with the lower boundary of B, and so all descriptions in B are selected. All descriptions in B are also selected if a > (1/2).

Within W, the same algebraic condition holds: a description is selected if and only if (1). But a=0 coincides with the upper boundary of W and a=(1/2) coincides with the lower boundary of W. When a=(1/2), no descriptions in W are selected; and when a=0, all descriptions on W are selected.

The boundary line (1) between selected and unselected descriptions is a line that pivots around the point (c/3. c/3). When the weight a goes up, the selected region in B expands and the selected region in W contracts. The opposite happens when the weight a goes down.

This describes the regions $F^r(a)$, r=b,w, as shown in figure A-3.

Let $\beta_i = \beta_1 or\ \beta_2$ as appropriate. Note that

(2) $\varphi\big(\beta_i, z, y(a, z)\big) - c = \frac{(c-3z)(\beta_i - a)}{1+a}$.

Following section x, the weight a must be chosen to maximize

$$Q(a) = \iint_{F^b(a)} [\varphi(\beta_i, z, y) - c]dy\, dz + \iint_{F^w(a)} [\varphi(\beta_i, z, y) - c]dy\, dz$$

To think about maximizing Q(a), think about the integrand for fixed arbitrary z. Let U(z) denote the upper boundary of M at z:

$$U(z) = \frac{c - z}{2}\ if\ y \geq z,$$

$$= c - 2z\ if\ y < z.$$

Then define

$$q(a, z) = \int_{y(a,z)}^{U(x)} \varphi(\beta_i, z.y) - c]dy$$

To maximize Q(a), we first look for first-order conditions. Note

$$\frac{d}{da} q(a, z) = -\frac{d}{da} y(a, z) * [\ \varphi\big(\beta_i, z, y(a, z)\big) - c]$$

From (1)

$$\frac{d}{da} y(a, z) = \frac{3z - c}{(1 = a)^2}.$$

Combining this with (2) we obtain

$$\frac{d}{da} q(a, z) = \frac{(3z - c)^2(\beta_i - a)}{(1 + a)^3}$$

Hence

$$\frac{d}{da}Q(a) = \int_{\frac{c-1}{2}}^{\frac{c}{3}} \frac{(3z-c)^2(\beta_1 - a)}{(1+a)^3}dz + \int_{c/3}^{c/2} \frac{(3z-c)^2(\beta_2 - a)}{(1+a)^3}$$

To simplify

(3)

$$\frac{d}{da}Q(a) = \frac{1}{(1+a)^3}\{(\beta_1 - a)\int_{\frac{c-1}{2}}^{\frac{c}{3}}(3z-c)^2\,dz + (\beta_2 - a)\int_{c/3}^{c/2}(3z-c)^2 dz\}$$

Let

$$\omega_1 = \frac{\int_{\frac{c-1}{2}}^{\frac{c}{3}}(3z-c)^2\,dz}{\int_{\frac{c-1}{2}}^{\frac{c}{3}}(3z-c)^2\,dz + \int_{\frac{c}{3}}^{\frac{c}{2}}(3z-c)^2 dz\}}$$

And $\omega_2 = 1 - \omega_1$.

Since $(3z-c)^2 \geq 0$ with strict inequality for at least one z in each interval, $\omega_i \in [0,1]\ for\ i = 1,2$.

Setting $\frac{d}{da}Q(a) = 0$, and solving for a we obtain:

$$a^* = \omega_1\beta_1 + \omega_2\beta_2$$

The optimal weight on Black candidates is a weighted average of the empirical proportions Black in the two populations. The weights depend on the relative aggregate distance from the pivot point (c/3,c/3).

The second-order condition for a^* to be a maximum is easily satisfied. In (3) the numerator is decreasing in a and the denominator is increasing in a. The second derivative of Q(a) with respect to a is therefore decreasing everywhere and a^* is a unique maximum.

Figure A-3 illustrates how the optimal fully non-discriminatory policy differs from the optimal partially non-discriminatory policy. In this example, the selected set under full non-discrimination is a proper subset of personal non-discrimination. This holds for both races. This result, however, is not robust: if $\beta_1 < \beta_2$, the opposite would hold. This graph also shows that the optimal fully non-discriminatory policy is different from the optimal non-discriminatory policy with selection constained to be determined by a single regression.

Figure A-1

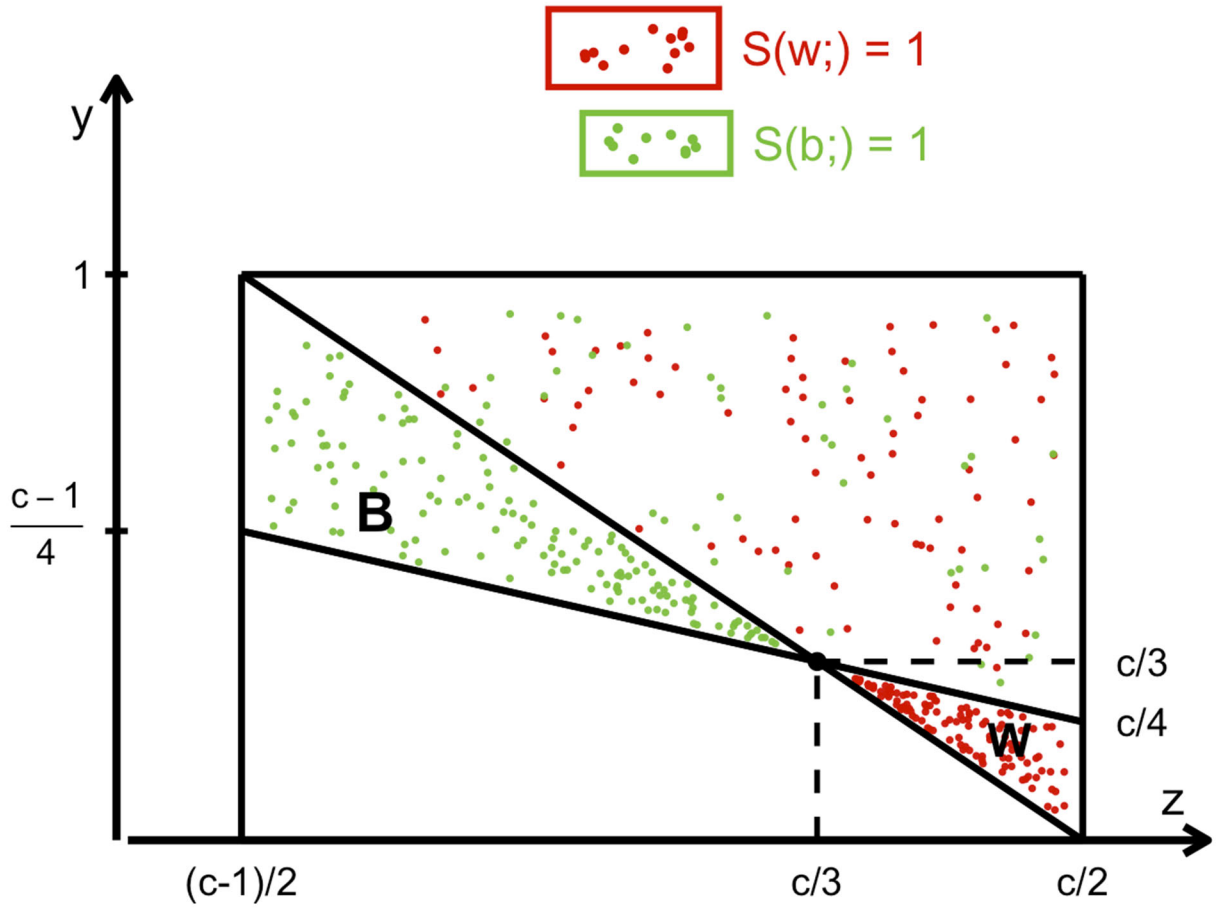


Figure A-2

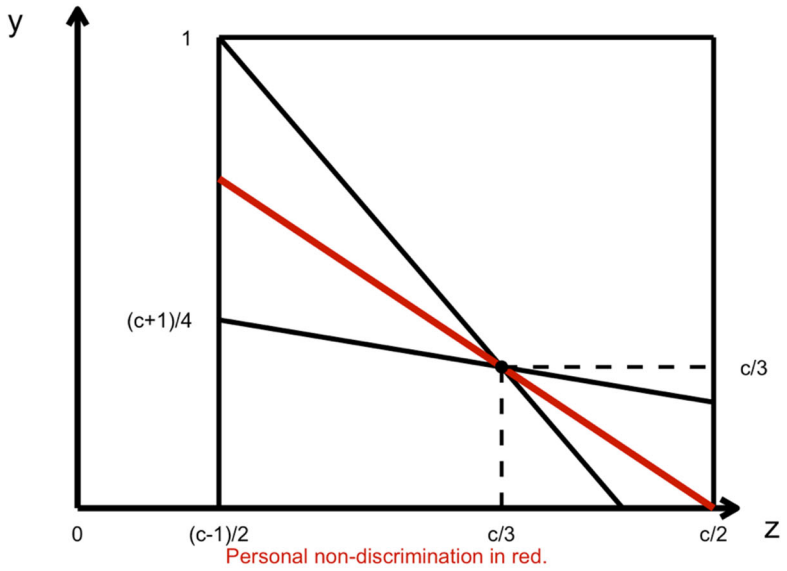
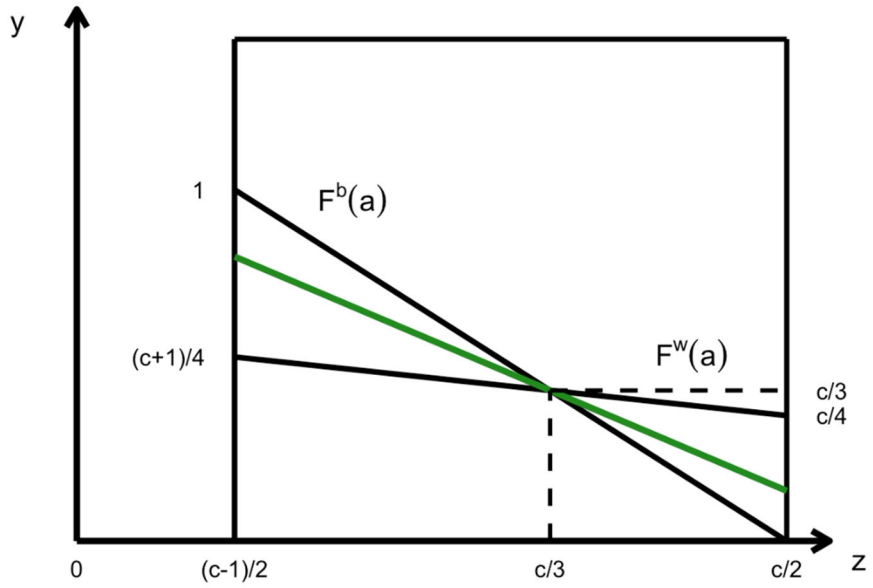
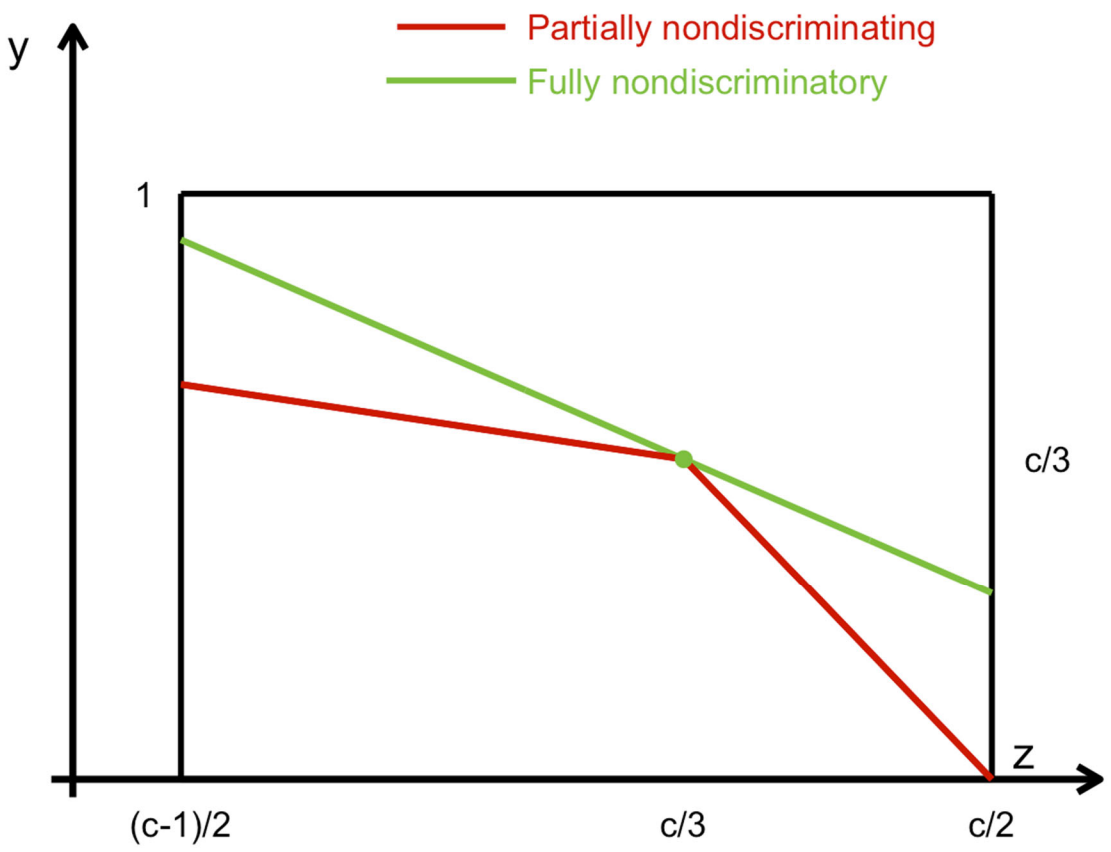


Figure A-3



-

Figure A-4


Partially nondiscriminating
Fully nondiscriminatory
y
1
c/3
z
(c-1)/2
c/3
c/2