Removing Algorithmic Discrimination (With Minimal Individual Error)

El-Mahdi El-Mhamdi[1], Rachid Guerraoui[1], Lê Nguyen Hoang[1] and Alexandre Maurer[2]
[1]EPFL, Distributed Computing Laboratory
[2]UM6P, School of Computer & Communication Sciences

Abstract

We address for the first time the problem of correcting group discriminations within a score function, while minimizing the individual error. Each group is described by a probability density function on the set of profiles. We first solve the problem analytically in the case of two populations, with a uniform bonus-malus on the zones where each population is a majority. We then address the general case of n populations, where the entanglement of populations does not allow a similar analytical solution. We show that an approximate solution with an arbitrarily high level of precision can be computed with linear programming. Finally, we address the reverse problem where the error should not go beyond a certain value and we seek to minimize the discrimination.

Keywords: score function; discrimination; linear programming.

Corresponding author:

Alexandre Maurer

alexandre.maurer@um6p.ma

School of Computer & Communication Sciences
Mohammed VI Polytechnic University
Lot 660, Hay Moulay Rachid, Ben Guerir, 43150, Morocco

Competing interests statement: the authors have no competing interests to declare.

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

1 Introduction

As machine learning is being deployed, a growing number of cases of discriminatory behaviors is being highlighted. In 2016, a study by ProPublica[1] showed that some algorithmic assessment of recidivism risks was significantly racially biased against black criminals. Indeed, 44.9% of supposedly high-risk black criminals did not re-offend, as opposed to 23.5% of supposedly high-risk white criminals. Conversely, 28.0% of supposedly low-risk black criminals re-offended, as opposed to 47.7% of supposedly low-risk white criminals. Such concerns for algorithmic discrimination have fostered a lot of work.

A major difficulty posed by new machine learning techniques is that algorithms may have learned their biases from high-dimensional data, which ironically seems hard to handle without machine learning. Racial inequalities in facial recognition have for instance been showed in [1]. More disturbingly, it was discovered that the popular word2vec package [29] yields gender discriminative relations between word representations, e.g., $doctor - man + woman = nurse$. In other words, word2vec seems to infer from natural language processing that a man is to a woman what a doctor is to a nurse. Although this is only one example out of many, it illustrates the difficulty of mitigating algorithmic discrimination.

Many solutions have been proposed. Dwork et al. [11] introduced the concept of "fair affirmative action", to improve the treatment of specific groups while treating similar individuals similarly. Some approaches consist in pre-processing data used for machine learning [10] [41] [26] [27] or making it unbiased [12]. Some try to prevent discrimination during the learning phase [38] [34] [17], by using causal reasoning [21], or with graphical dependency models [16]. Other approaches try to achieve independence from specific sensitive attributes [40]. [28] considers the problem of learning fair classifiers, and [13] tries to achieve fairness in the context of adaptive boosting, support vector machines and logistic regression. Algorithmic discrimination was also considered in problems of subsampling [4], voting [3], personalization [6] or ranking [5].

More recently, [35] extended the results of [16] by showing the necessary and sufficient condition to remove discrimination while preserving an equal error rate between groups. Important negative results on fair classifiers were highlighted in [33], e.g., removing group discrimination is only compatible with a single error constraint (like equal false-negatives rates across groups). [22] proposes an approach to remove discrimination among several populations, while preserving the accuracy of populations for which the classifier is already accurate. [8] pointed out the problem of "self-fulfilling prophecies" (i.e. predictions affecting the outcome, and therefore future predictions), and [30] proposed a new counterfactual criterion better adapted to this kind of situations. [31] and [37] also studied how algorithmic predictions can interact with self-interested decision makers.

Most previous works focus on classifiers, i.e. binary decisions. Thus, even if a group fairness criterion is satisfied, from an individual point of view, one is either discriminated "totally" or "not at all". So far, the fairness of real-valued functions has been discussed in the contexts of linear regression [2], probabilistic models [14], collaborative filtering [39], dimensionality reduction [32] and statistical independence [18]. In [16], continuous scores are considered, but in the context of optimizing a fair binary classifier.[2]

In this paper, we consider a setting where a continuous score is attributed to several individuals. The typical example would be a ML model analyzing the CV of various job applicants, and attributing a score to each one. As many ML models are opaque and subject to biases, the recruiting company may consider that some criteria should not have an influence within some specific categories. For

[1] See `https://tinyurl.com/ml-bias-sentence`

[2] [16] considers binary predictors, and tries to reach statistical independence w.r.t. sensitive attributes. It first shows that this problem can be expressed as a linear programming problem. It then considers the particular case where the binary prediction results from the combination of a score function and an arbitrary threshold: this additional information on the nature of the predictor can be used to achieve a better precision. However, as stated in the paper, this specific version of the problem cannot be solved with linear programming (yet can be "efficiently optimized numerically using ternary search").

instance, among college-educated candidates with a similar technical background, the company may consider that race or gender should not have a large statistical influence. While it may be impossible to remove all biases in an opaque model, a reasonable objective could be that some categories (e.g. female college-educated candidates and male college-educated candidates) have the same *average score*. However, doing so may have a cost in terms of individual accuracy, that we may want to minimize.

We therefore address, for the first time, the problem of correcting group discrimination within a score function while minimizing the individual error. We consider a score function assigning a score to each individual. Our criterion of group fairness is the average score of a given population, that we may want to increase, decrease or equalize.[3] We propose a post-processing approach to modify the score function, where we tolerate an individual error ϵ. This error is defined as the maximum difference between the initial and modified score function. Our goal is to achieve group fairness while minimizing this individual error. In this setting, individuals are only "differentially discriminated", with an error at most ϵ. This is the first time this problem is considered in the literature.[4]

More specifically, we assume that we are given a score function f that computes a score $f(x)$ for each individual x. Here, the individual's profile $x \in S$ can be any sort of description of the individual. In simple settings, it may be a collection of real-valued features, i.e. $S = \mathbb{R}^d$, and the scoring function f may be interpretable. However, as machine learning improves, rawer data are being used to score individuals, e.g. they may be textual biographies of undetermined length. In such cases, the scoring function f is usually constructed via machine learning, and it often has to be regarded as some "black box". To remove group discrimination, rather than pre-processing raw data or modifying the learning phase, it may thus be simpler to perform some post-processing of the score function, i.e. deriving a non-discriminative score function h from the possibly discriminative function f.

An additional difficulty is that the individual's profile x may not clearly determine its sensitive features, e.g. gender or race. Nevertheless, evidently, even biographic texts may provide strong indications of the individual's likely sensitive features. A natural approach to analyze the dependency of the score function on sensitive features is to test its scoring on profiles that are representative of a certain gender or race. Interestingly, this approach can now be simulated using so-called generative models [15]. These models allow to draw representative samples of subpopulations of individuals.

Thus, we assume that any population i (women, men, black, white, ...) can be described by some generative model[5]. Formally, this corresponds to saying that the population i is represented by a probability density function p_i on S. Given p_i, we can determine the *average score* of population i (i.e., $\int_{x \in S} p_i(x) f(x) dx$), which can be well approximated by sampling the generative model associated to population i.

Contributions. For pedagogical reasons, we first study in this paper the simple case of two populations with a different average score. The goal here is to determine a new score function h where (a) the two populations have the same average score and (b) the *individual error* is minimized. We define the individual error as the maximal difference between f and h, i.e., $\max_{x \in S} |f(x) - h(x)|$ (also written $||f - h||_\infty$).[6]

[3] To our knowledge, this is the first time this criterion is considered. Similar criteria have been considered (see previous works on real-valued functions), but in a more probabilistic or statistical sense.

[4] Our claim is not that this is the first paper to consider fairness with real values (see references above), but the first to modify a score function under constraints.

[5] More precisely, many generative machine-learning models, such as Generative Adversarial Networks (GANs), precisely allow to infer a probability distribution of the features of a population (or a representative sample of this distribution) based on a finite set of example. Arguably, the probability distribution inferred by such algorithms is more relevant, as it avoids overfitting on a finite small set of given examples.

[6] Our motivation for using this norm (often called "max norm") are the following. First, it is very simple, and has (to our knowledge) not been considered before for this kind of problem. Second, in terms of social fairness, it sets a clear limit to the "worst-case treatment" of one particular individual. Among other existing norms, it is often possible to "sacrifice" a small group of individuals to reach the desired outcome (by giving them an extreme score).

We call the problem of determining the best function h the 2-ODR (2-Optimal Discrimination Removal) problem ("2" standing for "two populations"). We present an exact solution to the 2-ODR problem. Roughly speaking, we consider the subsets of S where $p_1(x) > p_2(x)$ and $p_2(x) > p_1(x)$, and apply a uniform bonus (or penalty) on these subsets. We show that our solution is indeed optimal for it minimizes the individual error.

Then we turn to the more general case of n populations, which is arguably the most relevant setting in practice. Indeed, it is for instance often considered important that a score function be both non-racist *and* non-sexist. Similarly, it may be relevant to compare the scores of several races, e.g. Black, White, Asian and Arabic. In fact, we may even demand greater granularity by also comparing black female and white female, in addition to already comparing black and white. We address this n-population setting by considering some desired average score y_i for each population i. This more general goal enables the modelers to describe more subtly what they consider desirable. We call this problem the Optimal Discrimination Removal (ODR) problem.

This problem is significantly more difficult with $n > 2$. In fact, we conjecture that it is computationally intractable for $n > 2$ and combinatorially large profile sets S. Indeed, intuitively, in the case $n = 2$, the general problem of removing discrimination could be fixed locally for each $x \in S$, by determining whether x is more likely to be of population 1 or 2. Unfortunately, this no longer seems to be the case when $n > 2$. To solve the ODR problem, it seems that a global solution h first needs to be derived. But this global solution seems to require at least $\Omega(|S|)$ computation steps in general.

Interestingly though, we show that an approximate solution (with an arbitrarily high level of precision) can be obtained with *linear programming* [9]. Linear programming problems are expressed in terms of a set of inequalities involving linear combinations of variables. These problems have been extensively studied, and a lot of algorithms have been proposed to solve them [19] [24] [36] [20]. Here, we show that this abundant literature of algorithms can also be leveraged to solve discrimination problems.

We proceed incrementally through 6 steps. We first show that the ODR problem is reducible to the simpler (to express) Optimal Bonus-Malus (OBM) problem, where each desired average score is 0. We then define an approximate version of OBM, which we denote AOBM. We consider an arbitrary partition $(S_1, \ldots, S_m)$ of S, as well as a set of functions Z which are "flat" on each subset S_j. The AOBM problem consists in approximating a solution to the OBM problem with a function $u \in Z$. The larger m, the more precise the solution. We show that the AOBM problem is equivalent to a linear programming problem with $2m + 1$ variables and $m + 2n$ inequalities[7]. We use the fact that the functions of Z can only take a finite number of values, to transform the continuous OBM problem into a discrete problem.

We finally also address the reverse problem, where the individual error is not allowed to be greater than ϵ. Here, the goal is to be *as close as possible* to the desired score of each population. We proceed in an analogous way through 6 steps.

The case of two populations is treated in Section 2, the general case in Section 3, and the reverse case in Section 4. We conclude in Section 5.

The max norm is a simple way to avoid this, as well as a clear and easily understandable guarantee.

[7]Excluding the inequalities requiring each variable to be positive (which are included in the canonical form of a linear programming problem).

2 The Case of Two Populations

Let S be a set of *profiles*[8]. Let f be a function from S to $\mathbb{R}$ associating a *score* to each profile. Let p_1 and p_2 be any two probability density functions[9] on S, representing two populations 1 and 2.

Let X be the set of functions g from S to $\mathbb{R}$ such that $\int_{x\in S} p_1(x)g(x)dx = \int_{x\in S} p_2(x)g(x)dx$ (i.e. population 1 and 2 have the same average score).

For any function g from S to $\mathbb{R}$, let $||g||_\infty = \max_{x\in S} |g(x)|$.

The 2-ODR (2-Optimal Discrimination Removal) problem consists in finding a function $h \in \arg\min_{g\in X} ||g - f||_\infty$, i.e., a function minimizing the individual error.

Solution. For $x \in S$, let $u(x) = 1$ if $p_1(x) > p_2(x)$ and -1 otherwise. Let $A = \int_{x\in S}(p_1(x) - p_2(x))u(x)dx$, $B = \int_{x\in S}(p_1(x) - p_2(x))f(x)dx$ and $k = -B/A$.

We define h by $h(x) = f(x) + ku(x)$.

Theorem 1. *Function h above solves the 2-ODR problem.*

Proof. By construction, $||h - f||_\infty = |k|$. If $k = 0$, h indeed minimizes $||h - f||_\infty$. We now suppose that $k \neq 0$.

The proof is by contradiction. Suppose the opposite of the claim: there exists a function $h' \in X$ such that $||h' - f||_\infty < |k|$. Then, $h'(x) = f(x) + v(x)$, with $|v(x)| < |k|$. Let $D = \int_{x\in S} p_1(x)h'(x)dx - \int_{x\in S} p_2(x)h'(x)dx$. Then, $D = \int_{x\in S}(p_1(x) - p_2(x))f(x)dx + \int_{x\in S}(p_1(x) - p_2(x))v(x)dx$.

By definition, $k = -B/A$, with $A = \int_{x\in S}(p_1(x)-p_2(x))u(x)dx$ and $B = \int_{x\in S}(p_1(x)-p_2(x))f(x)dx$. Thus, $B = -kA$, and $D = -k\int_{x\in S}(p_1(x)-p_2(x))u(x)dx + \int_{x\in S}(p_1(x)-p_2(x))v(x)dx = \int_{x\in S}(p_1(x)-p_2(x))(v(x) - ku(x))dx$.

Let S_1 (resp. S_2) be the subset of S such that, $\forall x \in S_1$ (resp. S_2), $p_1(x) > p_2(x)$ (resp. $p_2(x) \leq p_1(x)$). Then, $D = D_1 + D_2$, where $D_i = \int_{x\in S_i}(p_1(x) - p_2(x))(v(x) - ku(x))dx$.

We define function s as follows: $s(x) = 1$ if $x > 0$ and -1 otherwise.

If $p_1(x) > p_2(x)$ (resp. $p_2(x) \leq p_1(x)$), $u(x) = 1$ (resp. $u(x) = -1$). Then, as $|v(x)| < |k|$, we have $s(v(x) - ku(x)) = -s(k)$ (resp. $s(k)$). Thus, $s(D_1) = s(D_2) = -s(k)$, and $D = D_1 + D_2 \neq 0$.

Therefore, $\int_{x\in S} p_1(x)h'(x)dx \neq \int_{x\in S} p_2(x)h'(x)dx$, and $h' \notin X$: contradiction. Hence, our result. □

3 The Case of Many Populations

We now consider the case of n populations. This problem when $n \geq 2$ is significantly harder than the problem above due to the entanglement of several probability density functions. We show that an approximate solution of this problem can be obtained with linear programming. We proceed incrementally through 6 steps.

1. We define the general Optimal Discrimination Removal (ODR) problem, corresponding to the case $n \geq 2$.
2. We define a simpler (to express) problem, the Optimal Bonus-Malus (OBM) problem.
3. We show that solving OBM provides an immediate solution to ODR.

[8]Here, a profile can be any set of information describing an individual. The precise nature of these profiles has no importance here, since the function f is considered as a "black box".

[9]For the reasons to assume that such probability density functions are available: see the introduction.

4. We define an approximate version of the OBM problem (AOBM), where we restrict ourselves to functions which are "flat" on an arbitrarily large number of subsets of S.

5. We define a Linear Programming problem, that we simply call LP for convenience.

6. We show that LP also solves AOBM.

Note that this problem is not, strictly speaking, a generalization of Section 2.[10]

Step 1: The Optimal Discrimination Removal (ODR) Problem

Let $(p_1, p_2, \ldots, p_n)$ be n probability density functions on S, each one representing a population. Let $(y_1, y_2, \ldots, y_n)$ be n arbitrary values. Let Ω_0 be the set of functions g from S to $\mathbb{R}$ such that, $\forall i \in \{1, \ldots, n\}$, $\int_{x \in S} p_i(x) g(x) dx = y_i$ (i.e. the mean score of population i is y_i). If $\Omega_0 \neq \emptyset$, the ODR problem consists in finding a function $h \in \arg\min_{g \in \Omega_0} ||g - f||_\infty$.

Step 2: The Optimal Bonus-Malus (OBM) Problem

$\forall i \in \{1, \ldots, n\}$, let $b_i = y_i - \int_{x \in S} p_i(x) f(x) dx$. Let Ω be the set of functions g from S to $\mathbb{R}$ such that, $\forall i \in \{1, \ldots, n\}$, $\int_{x \in S} p_i(x) g(x) dx = b_i$. If $\Omega \neq \emptyset$, the OBM problem consists in finding a function $u \in \arg\min_{g \in \Omega} ||g||_\infty$.

Step 3: Reducing ODR to OBM

Theorem 2 below says that a solution to the OBM problem provides an immediate solution to the ODR problem.

Theorem 2. *If u solves the OBM problem, then $h = f + u$ solves the ODR problem.*

Proof. As u solves the OBM problem, we have the following: $\forall g \in \Omega$, $||u||_\infty \leq ||g||_\infty$.

Note that, if $g \in \Omega_0$, then $g - f \in \Omega$. Indeed, if $g \in \Omega_0$, then $\forall i \in \{1, \ldots, n\}$, $\int_{x \in S} p_i(x) g(x) dx = y_i$. Thus, $\forall i \in \{1, \ldots, n\}$, $\int_{x \in S} p_i(x)(g(x) - f(x)) dx = y_i - \int_{x \in S} p_i(x) f(x) dx = b_i$. Thus, $g - f \in \Omega$.

Therefore, $\forall g \in \Omega_0$, $||u||_\infty \leq ||g - f||_\infty$. As $u = h - f$, we have: $\forall g \in \Omega_0$, $||h - f||_\infty \leq ||g - f||_\infty$. Thus, $h \in \arg\min_{g \in \Omega_0} ||g - f||_\infty$. Thus, the result. □

Step 4: The Approximate OBM (AOBM) Problem

Since the OBM problem may be intractable, we restrict the space of solutions to a specific family of functions. We partition S into several subsets, and consider the set Z of functions with a constant value on each of these subsets, then restrict our minimization problem to Z. Note that, as the partitioning is arbitrary, a function of Z can approximate the real solution with an arbitrary precision (like a picture with an arbitrarily large number of pixels).

Let $(S_1, \ldots, S_m)$ be a partition of S: $S_1 \cup S_2 \cup \cdots \cup S_m = S$, and $\forall \{i, j\} \in \{1, \ldots, m\}$, $S_i \cap S_j = \emptyset$. Let Z be the set of functions z from S to $\mathbb{R}$ such that, $\forall i \in \{1, \ldots, m\}$, $\forall x \in S_i$ and $\forall x' \in S_i$, $z(x) = z(x')$ (i.e. z is "flat" on each subset S_i).

If $\Omega \cap Z \neq \emptyset$, the AOBM problem consists in finding a function $u \in \arg\min_{g \in \Omega \cap Z} ||g||_\infty$.

[10] The 2-ODR problem of Section 2 consists in making two average scores equal. The ODR problem of Section 3 consists in making n scores equal to n arbitrary values y_i. The reason for this difference is the following: the problem defined is Section 2 is more fit for a theoretical proof, and the problem defined in Section 3 is more fit for an approximate solution with linear programming. Besides, overall, we think that having arbitrary "goal values" y_i gives much more liberty to the user.

At this point, one may wonder which properties a partitioning $(S_1, \ldots, S_m)$ should ideally have. While there is no definitive answer to this question, let us give an example of such properties. The intuitive idea here is that, on any subset S_k, no relevant quantity (i.e., the function f and the probability density functions $(p_1, p_2, \ldots, p_n)$) should vary too much. If this property is satisfied, there is no reason to keep subdividing S_k: should we do so, the correction applied to each resulting part would be almost the same.

More formally, let δ_f and δ_p be two arbitrarily small positive constants. The goal is then to find a partitioning $(S_1, \ldots, S_m)$ such that, $\forall k \in \{1, \ldots, m\}$, the two following properties are satisfied:

1. $\max_{x \in S_k} f(x) - \min_{x \in S_k} f(x) \leq \delta_f$
2. $\forall i \in \{1, \ldots, n\}$, $\max_{x \in S_k} p_i(x) - \min_{x \in S_k} p_i(x) \leq \delta_p$

Step 5: The Linear Programming (LP) Problem

Let N and M be two integers. Let $(x_1, \ldots, x_N)$ be N variables. Let L and $(L_1, \ldots, L_M)$ be $M+1$ linear combinations of the variables $(x_1, \ldots, x_N)$. Let $(c_1, \ldots, c_M)$ be M constant terms.

A *linear programming* problem consists in finding values of $(x_1, \ldots, x_N)$ maximizing L while verifying the following inequalities:

- $\forall k \in \{1, \ldots, N\}$, $x_k \geq 0$
- $\forall k \in \{1, \ldots, M\}$, $L_k \leq c_k$

In the following, we define a specific linear programming problem, that we simply call LP problem for convenience.

$\forall i \in \{1, \ldots, n\}$ and $\forall j \in \{1, \ldots, m\}$, let $v(i,j) = \int_{x \in S_j} p_i(x)dx$.

Let $(\alpha_1, \ldots, \alpha_m)$, $(\beta_1, \ldots, \beta_m)$ and γ be $2m+1$ variables.

Consider the following inequalities:

1. $\gamma \geq 0$, and $\forall j \in \{1, \ldots, m\}$, $\alpha_j \geq 0$ and $\beta_j \geq 0$.
2. $\forall j \in \{1, \ldots, m\}$, $\alpha_j - \gamma \leq 0$ and $\beta_j - \gamma \leq 0$.
3. $\forall i \in \{1, \ldots, n\}$, $\Sigma_{j=1}^{j=m} \alpha_j v(i,j) - \Sigma_{j=1}^{j=m} \beta_j v(i,j) \leq b_i$
4. $\forall i \in \{1, \ldots, n\}$, $\Sigma_{j=1}^{j=m} \beta_j v(i,j) - \Sigma_{j=1}^{j=m} \alpha_j v(i,j) \leq -b_i$

The LP problem consists in finding values of $(\alpha_1, \ldots, \alpha_m)$, $(\beta_1, \ldots, \beta_m)$ and γ maximizing $-\gamma$ while satisfying the aforementioned inequalities.

Step 6: Reducing AOBM to LP

Let $(\alpha_1, \ldots, \alpha_m)$, $(\beta_1, \ldots, \beta_m)$ and γ be a solution to the LP problem. $\forall x \in S$, let $\lambda(x)$ be the integer j such that $x \in S_j$. Let u be the function from S to $\mathbb{R}$ such that, $\forall x \in S$, $u(x) = \alpha_{\lambda(x)} - \beta_{\lambda(x)}$.

Theorem 3 below says that u solves the AOBM problem. We first prove some lemmas.

Lemma 1. *$||u||_\infty \geq \max(\alpha^*, \beta^*)$, where $\alpha^* = \max_{j \in \{1,\ldots,m\}} \alpha_j$ and $\beta^* = \max_{j \in \{1,\ldots,m\}} \beta_j$.*

Proof. Suppose the opposite: $||u||_\infty < \max(\alpha^*, \beta^*)$. According to inequalities 2, $\gamma \geq \max(\alpha^*, \beta^*)$. Thus, $\gamma > ||u||_\infty$.

$\forall j \in \{1, \ldots, m\}$, we define $(\alpha'_1, \ldots, \alpha'_m)$ and $(\beta'_1, \ldots, \beta'_m)$ as follows:

- If $\alpha_j \geq \beta_j$, $\alpha'_j = \alpha_j - \beta_j$ and $\beta'_j = 0$.
- Otherwise, $\alpha'_j = 0$ and $\beta'_j = \beta_j - \alpha_j$.

Let $\gamma' = ||u||_\infty < \gamma$.

$\forall j \in \{1, \dots, m\}$, $\alpha'_j \leq \max(\alpha_j, \beta_j)$ and $\beta'_j \leq \max(\alpha_j, \beta_j)$. Thus, $\alpha'_j \leq \gamma'$ and $\beta'_j \leq \gamma'$.

We now show that $(\alpha'_1, \dots, \alpha'_m)$, $(\beta'_1, \dots, \beta'_m)$ and γ' satisfy the inequalities of the LP problem.

Inequalities 1 are satisfied by definition. $\forall j \in \{1, \dots, m\}$, $\alpha'_j \leq \gamma'$ and $\beta'_j \leq \gamma'$. Thus, $\alpha'_j - \gamma' \leq 0$ and $\beta'_j - \gamma' \leq 0$, and inequalities 2 are satisfied.

Inequalities 3 and 4 are equivalent to: $\forall i \in \{1, \dots, n\}$, $\Sigma_{j=1}^{j=m} \alpha_j v(i,j) - \Sigma_{j=1}^{j=m} \beta_j v(i,j) = b_i$. $\forall j \in \{1, \dots, m\}$:

- If $\alpha_j \geq \beta_j$, $\alpha'_j - \beta'_j = (\alpha_j - \beta_j) - 0 = \alpha_j - \beta_j$.
- Otherwise, $\alpha'_j - \beta'_j = 0 - (\beta_j - \alpha_j) = \alpha_j - \beta_j$.

Thus, $\forall j \in \{1, \dots, m\}$, $\alpha'_j - \beta'_j = \alpha_j - \beta_j$. Thus, $\Sigma_{j=1}^{j=m} \alpha'_j v(i,j) - \Sigma_{j=1}^{j=m} \beta'_j v(i,j) = \Sigma_{j=1}^{j=m} (\alpha'_j - \beta'_j) v(i,j) = \Sigma_{j=1}^{j=m} (\alpha_j - \beta_j) v(i,j) = \Sigma_{j=1}^{j=m} \alpha_j v(i,j) - \Sigma_{j=1}^{j=m} \beta_j v(i,j) = b_i$. Thus, inequalities 3 and 4 are satisfied.

Thus, there exists $(\alpha'_1, \dots, \alpha'_m)$, $(\beta'_1, \dots, \beta'_m)$ and γ' satisfying the inequalities of the LP problem with $-\gamma' > -\gamma$. Thus, $(\alpha_1, \dots, \alpha_m)$, $(\beta_1, \dots, \beta_m)$ and γ do not solve the LP problem: contradiction. Thus, the result. □

Lemma 2. $||u||_\infty = \gamma$.

Proof. $||u||_\infty = \max_{x \in S} |u(x)| = \max_{j \in \{1,\dots,m\}} |\alpha_j - \beta_j|$. $\forall j \in \{1, \dots, m\}$, $\alpha_j \leq \gamma$ and $\beta_j \leq \gamma$. Thus, $|\alpha_j - \beta_j| \leq \gamma$, and $||u||_\infty \leq \gamma$.

We now show that $\gamma \leq ||u||_\infty$. Suppose the opposite: $\gamma > ||u||_\infty$. As the LP problem consists in maximizing $-\gamma$ (and thus, minimizing γ), this implies that the inequalities of the LP problem are not compatible with $\gamma \leq ||u||_\infty$. Variable γ only appears in inequalities 1 and 2, and these inequalities impose to have $\gamma \geq 0$, $\gamma \geq \max_{j \in \{1,\dots,m\}} \alpha_j$ and $\gamma \geq \max_{j \in \{1,\dots,m\}} \beta_j$. Thus, $\gamma = \max(a^*, b^*)$, where $\alpha^* = \max_{j \in \{1,\dots,m\}} \alpha_j$ and $\beta^* = \max_{j \in \{1,\dots,m\}} \beta_j$. Thus, according to Lemma 1, $||u||_\infty \geq \gamma$.

Therefore, $||u||_\infty = \gamma$. □

Theorem 3. *Function u solves the AOBM problem.*

Proof. By definition, $u \in Z$.

Inequalities 3 and 4 of the LP problem are equivalent to: $\forall i \in \{1, \dots, n\}$, $\Sigma_{j=1}^{j=m} \alpha_j v(i,j) - \Sigma_{j=1}^{j=m} \beta_j v(i,j) = b_i$. Thus, $\forall i \in \{1, \dots, n\}$, $b_i = \Sigma_{j=1}^{j=m} (\alpha_j - \beta_j) v(i,j) = \Sigma_{j=1}^{j=m} (\alpha_j - \beta_j) \int_{x \in S_j} p_i(x) dx = \Sigma_{j=1}^{j=m} \int_{x \in S_j} u(x) p_i(x) dx = \int_{x \in S} p_i(x) u(x) dx$. Thus, $u \in \Omega$.

Therefore, $u \in \Omega \cap Z$. Now, suppose the opposite of the claim: $u \notin \arg\min_{g \in \Omega \cap Z} ||g||_\infty$. Let $w \in \arg\min_{g \in \Omega \cap Z} ||g||_\infty$. Thus, $||w||_\infty < ||u||_\infty$.

Let $(w_1, \dots, w_m)$ be such that, $\forall x \in S_j$, $w(x) = w_j$. Let $\gamma' = ||w||_\infty$. $\forall j \in \{1, \dots, m\}$, we define $(\alpha'_1, \dots, \alpha'_m)$ and $(\beta'_1, \dots, \beta'_m)$ as follows:

- If $w_j \geq 0$, $\alpha'_j = w_j$ and $\beta'_j = 0$.
- Otherwise, $\alpha'_j = 0$ and $\beta'_j = -w_j$.

Thus, inequalities 1 are satisfied.

As $\gamma' = ||w||_\infty$, $\forall j \in \{1, \ldots, m\}$, $\gamma' \geq |w_j| \geq \max(\alpha'_j, \beta'_j)$. Thus, inequalities 2 are satisfied.

As $w \in \Omega$, $\forall i \in \{1, \ldots, n\}$, $\int_{x \in S_j} p_i(x)w(x)dx = b_i$. Thus, $\forall i \in \{1, \ldots, n\}$, $b_i = \Sigma_{j=1}^{j=m} \int_{x \in S_j} p_i(x)w(x)dx$ $= \Sigma_{j=1}^{j=m} w_j \int_{x \in S_j} p_i(x)dx = \Sigma_{j=1}^{j=m} w_j v(i,j) = \Sigma_{j=1}^{j=m} \alpha'_j v(i,j) - \Sigma_{j=1}^{j=m} \beta'_j v(i,j) \leq b_i$. Thus, inequalities 3 and 4 are satisfied.

According to Lemma 2, $||u||_\infty = \gamma$. Thus, as $||w||_\infty < ||u||_\infty$, $\gamma' < \gamma$. Therefore, there exists $(\alpha'_1, \ldots, \alpha'_m)$, $(\beta'_1, \ldots, \beta'_m)$ and γ' satisfying the inequalities of the LP problem with $-\gamma' > -\gamma$. Thus, $(\alpha_1, \ldots, \alpha_m)$, $(\beta_1, \ldots, \beta_m)$ and γ do not solve the LP problem: contradiction. Thus, the result. □

4 The Reverse Case

In the previous section, we showed how to reach the desired scores for each population with a minimal individual error. However, even when minimized, the individual error may still be very high, and sometimes not acceptable.

In this section, we consider the reverse problem: assuming that we can accept an individual error which is at most ϵ, how can we reach a score which is *as close as possible* from the desired scores of each population? We call this problem the reverse ODR (R-ODR) problem.

We again proceed in 6 steps, following the same outline as the 6 steps of Section 3.

Step 1: The Reverse ODR (R-ODR) Problem

Let $\epsilon \geq 0$. Let Φ_0 be the set of functions g from S to $\mathbb{R}$ such that $||g - f||_\infty \leq \epsilon$ (i.e., functions for which the individual error remains acceptable).

Let g be a function from S to $\mathbb{R}$. $\forall i \in \{1, \ldots, n\}$, let $\mu_i(g) = |\int_{x \in S} p_i(x)g(x)dx - y_i|$ (i.e., the distance between the average score of population i and its desired average score y_i). Let $\mu(g) = \max_{i \in \{1,\ldots,n\}} \mu_i(g)$ (i.e., the upper bound of these distances).

The R-ODR problem consists in finding a function $h \in \arg\min_{g \in \Phi_0} \mu(g)$.

Note that in the previous case, we considered the set of functions for which each population has *exactly* the desired average score, and tried to minimize the individual error. Here, however, we cannot start with the assumption that $\epsilon = 0$, otherwise no change would be possible (by definition). Thus, we have to consider $\epsilon > 0$.

Step 2: The Reverse OBM (R-OBM) Problem

Let Φ be the set of functions g from S to $\mathbb{R}$ such that $||g||_\infty \leq \epsilon$.

$\forall i \in \{1, \ldots, n\}$, let $\Delta_i(g) = |\int_{x \in S} p_i(x)g(x)dx - b_i|$. Let $\Delta(g) = \max_{i \in \{1,\ldots,n\}} \Delta_i(g)$.

The R-OBM problem consists in finding a function $u \in \arg\min_{g \in \Phi} \Delta(g)$.

Step 3: Reducing R-ODR to R-OBM

In Theorem 4, we show that a solution to the R-OBM problem provides an immediate solution to the R-ODR problem.

Theorem 4. *If u solves the R-OBM problem, then $u + f$ solves the R-ODR problem.*

Proof. Let g be a function from S to $\mathbb{R}$. $\forall i \in \{1,\ldots,n\}$, $\Delta(u) = |\int_{x\in S} p_i(x)g(x)dx - b_i| = |\int_{x\in S} p_i(x)\ g(x)dx + \int_{x\in S} p_i(x)f(x)dx - \int_{x\in S} p_i(x)f(x)dx - b_i| = |\int_{x\in S} p_i(x)\ (g(x)+f(x))dx - y_i| = \mu_i(g+f)$. Thus, $\Delta(g) = \mu(g+f)$, and $\arg\min_{g\in\Phi}\Delta(g) = \arg\min_{g\in\Phi_0}\mu(g+f)$.

Therefore, if $u \in \arg\min_{g\in\Phi}\Delta(g)$, then $u+f \in \arg\min_{g\in\Phi_0}\mu(g)$. Thus, the result. □

Step 4: The Reverse AOBM (R-AOBM) Problem

The R-AOBM problem consists in finding a function $u \in \arg\min_{g\in\Phi\cap Z}\Delta(g)$.

Step 5: The Reverse LP (R-LP) Problem

Let $(\alpha_1,\ldots,\alpha_m)$, $(\beta_1,\ldots,\beta_m)$ and γ be $2m+1$ variables.

Consider the following inequalities:

1. $\gamma \geq 0$, and $\forall j \in \{1,\ldots,m\}$, $\alpha_j \geq 0$ and $\beta_j \geq 0$.
2. $\forall j \in \{1,\ldots,m\}$, $\alpha_j \leq \epsilon$ and $\beta_j \leq \epsilon$.
3. $\forall i \in \{1,\ldots,n\}$, $\Sigma_{j=1}^{j=m}\alpha_j v(i,j) - \Sigma_{j=1}^{j=m}\beta_j v(i,j) - b_i \leq \gamma$
4. $\forall i \in \{1,\ldots,n\}$, $\Sigma_{j=1}^{j=m}\beta_j v(i,j) - \Sigma_{j=1}^{j=m}\alpha_j v(i,j) + b_i \leq \gamma$

The R-LP problem consists in finding values of $(\alpha_1,\ldots,\alpha_m)$, $(\beta_1,\ldots,\beta_m)$ and γ maximizing $-\gamma$ while satisfying the aforementioned inequalities.

Step 6: Reducing R-AOBM to R-LP

Let $(\alpha_1,\ldots,\alpha_m)$, $(\beta_1,\ldots,\beta_m)$ and γ be a solution to the R-LP problem. Let u be the function from S to $\mathbb{R}$ such that, $\forall x \in S$, $u(x) = \alpha_{\lambda(x)} - \beta_{\lambda(x)}$.

In Theorem 5, we show that u solves the R-AOBM problem.

Lemma 3. $\Delta(u) \leq \gamma$.

Proof. $\forall i \in \{1,\ldots,n\}$, $\Delta_i(u) = |\int_{x\in S} p_i(x)u(x)dx - b_i| = |\Sigma_{j=1}^{j=m}\int_{x\in S_j} p_i(x)u(x)dx - b_i| = |\Sigma_{j=1}^{j=m}(\alpha_j - \beta_j)v(i,j)dx - b_i| = |\Sigma_{j=1}^{j=m}\alpha_j v(i,j) - \Sigma_{j=1}^{j=m}\beta_j v(i,j) - bi| \leq \gamma$, according to inequalities 3 and 4. Thus, $\Delta(u) = \max_{i\in\{1,\ldots,n\}}\Delta_i(u) \leq \gamma$. □

Lemma 4. $\Delta(u) \geq \gamma$.

Proof. Suppose the opposite: $\Delta(u) < \gamma$. Let $\gamma' = \Delta(u)$. $\forall i \in \{1,\ldots,n\}$, $\Delta_i(u) = |\int_{x\in S} p_i(x)u(x)dx - b_i| = |\Sigma_{j=1}^{j=m}\alpha_j v(i,j) - \Sigma_{j=1}^{j=m}\beta_j v(i,j) - b_i| \leq \Delta(u)$. Thus, as $\gamma' = \Delta(u)$, inequalities 3 and 4 are still satisfied if we replace γ by γ'. Thus, as $\gamma' < \gamma$, $(\alpha_1,\ldots,\alpha_m)$, $(\beta_1,\ldots,\beta_m)$ and γ do not solve the R-LP problem: contradiction. Thus, the result. □

Theorem 5. *The function u solves the R-AOBM problem.*

Proof. By definition, $u \in Z$. According to inequalities 2, $u \in \Phi$. Thus, $u \in \Phi \cap Z$. Now, suppose the opposite of the claim: $u \notin \arg\min_{g\in\Phi\cap Z}\Delta(g)$.

Let $w \in \arg\min_{g\in\Phi\cap Z}\Delta(g)$. Let $(w_1,\ldots,w_m)$ be such that, $\forall x \in S_j$, $w(x) = w_j$. Let $\gamma' = \Delta(w)$. $\forall j \in \{1,\ldots,m\}$, we define $(\alpha'_1,\ldots,\alpha'_m)$ and $(\beta'_1,\ldots,\beta'_m)$ as follows:

- If $w_j \geq 0$, $\alpha'_j = w_j$ and $\beta'_j = 0$.

- Otherwise, $\alpha'_j = 0$ and $\beta'_j = -w_j$.

By construction, inequalities 1 are satisfied.

As $w \in \Phi$, $\forall j \in \{1, \dots, m\}$, $|w_j| \leq \epsilon$. Thus, $\forall j \in \{1, \dots, m\}$, $\alpha'_j \leq |w_j| \leq \epsilon$ and $\beta'_j \leq |w_j| \leq \epsilon$. Therefore, inequalities 2 are satisfied.

As $\gamma = \Delta(w)$, $\forall i \in \{1, \dots, n\}$, $|\int_{x \in S} p_i(x) w(x) dx - b_i| = |\Sigma_{j=1}^{j=m} \alpha'_j v(i,j) - \Sigma_{j=1}^{j=m} \beta'_j v(i,j) - b_i| \leq \Delta(w) = \gamma'$. Thus, inequalities 3 and 4 are satisfied.

As $w \in \arg\min_{g \in \Phi \cap Z} \Delta(g)$ and $u \notin \arg\min_{g \in \Phi \cap Z} \Delta(g)$, we have $\Delta(w) < \Delta(u)$. We have $\Delta(w) = \gamma'$, and according to Lemma 3 and Lemma 4, $\Delta(u) = \gamma$. Thus, $\gamma' < \gamma$. Thus, $(\alpha_1, \dots, \alpha_m)$, $(\beta_1, \dots, \beta_m)$ and γ do not solve the R-LP problem: contradiction. Thus, the result. □

5 Conclusion and limitations

We consider the problem of removing algorithmic discrimination between several populations with a minimal individual error. We first describe an analytical solution to this problem in the case of two populations. We then show that the general case (with n populations) can be solved approximately with linear programming. We also consider the reverse problem where an upper bound on the error is fixed and we seek to minimize the discrimination.

A major challenge would be to either find an analytical solution to the general case with n populations or prove that it is indeed intractable. We conjecture the latter. Another interesting question would be to determine how to optimally choose the subsets $(S_1, \dots, S_m)$ used for the approximate solution.

Limitations. Finally, we make several clarifications on the scope of our approach, to avoid some misunderstandings.

1. We are "agnostic" w.r.t. the desirability of removing group discriminations. Such concerns are complex topics subject to many controversies, and are out of the scope of this paper. Our approach is the following: assuming that people want to remove group discriminations, we address the problem of doing this with a minimal error.

2. We do not pretend that any difference of score between two groups is problematic. These groups have to be carefully chosen by the user in order to be relevant. For instance, for a hiring process, comparing two groups of different ethnic origins but with the same education level may be more relevant than simply comparing two ethnic groups.[11]

3. The error considered is relative to the score function, not to the "ground truth" (which of course we do not have access to). The score function represents the best approximation we have of the ground truth.

4. The goal here is to adjust average scores with a minimal cost in terms of individual error. The decisions made with this score are out of the scope of this paper. However, we would like to underline the fact that this is not necessarily a binary decision. One could imagine (for instance) a set of job positions which "value" are proportional to the score. Replacing the score function with a binary function would be a different problem, but could be solved with a similar approach.

5. Finally, it is important to point out that solutions aiming to improve fairness may have unintended negative consequences. For instance, [25] shows that, in some temporal models, common fairness criteria may eventually cause more harm than good, due to feedback loops.

[11] This example is only here to illustrate, and is not meant to be a prescription.

The results of [23] suggest that some key notions of fairness are incompatible with each other, resulting in inherent trade-offs. In the case of recidivism, [7] shows that a recently-applied criterion of fairness may lead to considerable disparate impact when recidivism prevalence differs across groups.

Acknowledgements: *While doing this research work, the authors were paid by EPFL and UM6P.*

References

[1] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on Fairness, Accountability and Transparency*, pages 77–91, 2018.

[2] Toon Calders, Asim Karim, Faisal Kamiran, Wasif Ali, and Xiangliang Zhang. Controlling attribute effect in linear regression. In *2013 IEEE 13th International Conference on Data Mining, Dallas, TX, USA, December 7-10, 2013*, pages 71–80, 2013.

[3] L. Elisa Celis, Lingxiao Huang, and Nisheeth K. Vishnoi. Multiwinner voting with fairness constraints. In Jérôme Lang, editor, *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI 2018, July 13-19, 2018, Stockholm, Sweden*, pages 144–151. ijcai.org, 2018.

[4] L. Elisa Celis, Vijay Keswani, Damian Straszak, Amit Deshpande, Tarun Kathuria, and Nisheeth K. Vishnoi. Fair and diverse dpp-based data summarization. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 715–724. PMLR, 2018.

[5] L. Elisa Celis, Damian Straszak, and Nisheeth K. Vishnoi. Ranking with fairness constraints. In Ioannis Chatzigiannakis, Christos Kaklamanis, Dániel Marx, and Donald Sannella, editors, *45th International Colloquium on Automata, Languages, and Programming, ICALP 2018, July 9-13, 2018, Prague, Czech Republic*, volume 107 of *LIPIcs*, pages 28:1–28:15. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2018.

[6] L. Elisa Celis and Nisheeth K. Vishnoi. Fair personalization. *CoRR*, abs/1707.02260, 2017.

[7] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big Data*, 5(2):153–163, 2017.

[8] Amanda Coston, Alan Mishler, Edward H. Kennedy, and Alexandra Chouldechova. Counterfactual risk assessments, evaluation, and fairness. In Mireille Hildebrandt, Carlos Castillo, L. Elisa Celis, Salvatore Ruggieri, Linnet Taylor, and Gabriela Zanfir-Fortuna, editors, *FAT* '20: Conference on Fairness, Accountability, and Transparency, Barcelona, Spain, January 27-30, 2020*, pages 582–593. ACM, 2020.

[9] George Dantzig. *Linear programming and extensions*. Princeton university press, 2016.

[10] Flávio du Pin Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R. Varshney. Data pre-processing for discrimination prevention: Information-theoretic optimization and analysis. *IEEE J. Sel. Top. Signal Process.*, 12(5):1106–1119, 2018.

[11] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard S. Zemel. Fairness through awareness. In *Innovations in Theoretical Computer Science 2012, Cambridge, MA, USA, January 8-10, 2012*, pages 214–226, 2012.

[12] 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, Sydney, NSW, Australia, August 10-13, 2015*, pages 259–268, 2015.

[13] Benjamin Fish, Jeremy Kun, and Ádám Dániel Lelkes. A confidence-based approach for balancing fairness and accuracy. In *Proceedings of the 2016 SIAM International Conference on Data Mining, Miami, Florida, USA, May 5-7, 2016*, pages 144–152, 2016.

[14] Kazuto Fukuchi, Toshihiro Kamishima, and Jun Sakuma. Prediction with model-based neutrality. *IEICE Transactions*, 98-D(8):1503–1516, 2015.

[15] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. In *Advances in neural information processing systems*, pages 2672–2680, 2014.

[16] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. In *Advances in Neural Information Processing Systems 29: Annual Conference on Neural Information Processing Systems 2016, December 5-10, 2016, Barcelona, Spain*, pages 3315–3323, 2016.

[17] Shahin Jabbari, Matthew Joseph, Michael J. Kearns, Jamie Morgenstern, and Aaron Roth. Fairness in reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, pages 1617–1626, 2017.

[18] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Recommendation independence. In *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, pages 187–201, 2018.

[19] Narendra Karmarkar. A new polynomial-time algorithm for linear programming. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pages 302–311. ACM, 1984.

[20] Leonid G Khachiyan. Polynomial algorithms in linear programming. *USSR Computational Mathematics and Mathematical Physics*, 20(1):53–72, 1980.

[21] Niki Kilbertus, Mateo Rojas-Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 656–666, 2017.

[22] Michael P. Kim, Amirata Ghorbani, and James Y. Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In Vincent Conitzer, Gillian K. Hadfield, and Shannon Vallor, editors, *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES 2019, Honolulu, HI, USA, January 27-28, 2019*, pages 247–254. ACM, 2019.

[23] Jon M. Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In Christos H. Papadimitriou, editor, *8th Innovations in Theoretical Computer Science Conference, ITCS 2017, January 9-11, 2017, Berkeley, CA, USA*, volume 67 of *LIPIcs*, pages 43:1–43:23. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2017.

[24] Bernhard Korte and Jens Vygen. Linear programming algorithms. In *Combinatorial Optimization*, pages 73–99. Springer, 2012.

[25] Lydia T. Liu, Sarah Dean, Esther Rolf, Max Simchowitz, and Moritz Hardt. Delayed impact of fair machine learning. In Jennifer G. Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, volume 80 of *Proceedings of Machine Learning Research*, pages 3156–3164. PMLR, 2018.

[26] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S. Zemel. The variational fair autoencoder. In Yoshua Bengio and Yann LeCun, editors, *4th International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, May 2-4, 2016, Conference Track Proceedings*, 2016.

[27] Kristian Lum and James E. Johndrow. A statistical framework for fair predictive algorithms. *CoRR*, abs/1610.08077, 2016.

[28] Aditya Krishna Menon and Robert C. Williamson. The cost of fairness in binary classification. In *Conference on Fairness, Accountability and Transparency, FAT 2018, 23-24 February 2018, New York, NY, USA*, pages 107–118, 2018.

[29] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[30] Alan Mishler, Edward H. Kennedy, and Alexandra Chouldechova. Fairness in risk assessment instruments: Post-processing to achieve counterfactual equalized odds. In Madeleine Clare Elish, William Isaac, and Richard S. Zemel, editors, *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pages 386–400. ACM, 2021.

[31] Juan C. Perdomo, Tijana Zrnic, Celestine Mendler-Dünner, and Moritz Hardt. Performative prediction. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 7599–7609. PMLR, 2020.

[32] Adrián Pérez-Suay, Valero Laparra, Gonzalo Mateo-Garcia, Jordi Muñoz-Marí, Luis Gómez-Chova, and Gustau Camps-Valls. Fair kernel learning. In *Machine Learning and Knowledge Discovery in Databases - European Conference, ECML PKDD 2017, Skopje, Macedonia, September 18-22, 2017, Proceedings, Part I*, pages 339–355, 2017.

[33] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon M. Kleinberg, and Kilian Q. Weinberger. On fairness and calibration. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5680–5689, 2017.

[34] Novi Quadrianto and Viktoriia Sharmanska. Recycling privileged learning and distribution matching for fairness. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 677–688, 2017.

[35] Claire Lazar Reich and Suhas Vijaykumar. A possibility in algorithmic fairness: Can calibration and equal error rates be reconciled? In Katrina Ligett and Swati Gupta, editors, *2nd Symposium on Foundations of Responsible Computing, FORC 2021, June 9-11, 2021, Virtual Conference*, volume 192 of *LIPIcs*, pages 4:1–4:21. Schloss Dagstuhl - Leibniz-Zentrum für Informatik, 2021.

[36] James Renegar. A polynomial-time algorithm, based on newton's method, for linear programming. *Mathematical Programming*, 40(1-3):59–93, 1988.

[37] Yonadav Shavit, Benjamin L. Edelman, and Brian Axelrod. Causal strategic linear regression. In *Proceedings of the 37th International Conference on Machine Learning, ICML 2020, 13-18 July 2020, Virtual Event*, volume 119 of *Proceedings of Machine Learning Research*, pages 8676–8686. PMLR, 2020.

[38] Blake E. Woodworth, Suriya Gunasekar, Mesrob I. Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *Proceedings of the 30th Conference on Learning Theory, COLT 2017, Amsterdam, The Netherlands, 7-10 July 2017*, pages 1920–1953, 2017.

[39] Sirui Yao and Bert Huang. Beyond parity: Fairness objectives for collaborative filtering. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, 4-9 December 2017, Long Beach, CA, USA*, pages 2925–2934, 2017.

[40] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez-Rodriguez, and Krishna P. Gummadi. Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment. In *Proceedings of the 26th International Conference on World Wide Web, WWW 2017, Perth, Australia, April 3-7, 2017*, pages 1171–1180, 2017.

[41] Richard S. Zemel, Yu Wu, Kevin Swersky, Toniann Pitassi, and Cynthia Dwork. Learning fair representations. In *ICML 2013*, pages 325–333, 2013.