On the incompatibility of accuracy and equal opportunity

Carlos Pinzón, Catuscia Palamidessi, Pablo Piantanida, Frank Valencia

On the incompatibility of accuracy and equal opportunity

Carlos Pinzón[1,4,†] Catuscia Palamidessi[1,4]
Pablo Piantanida[2,3] Frank Valencia[2,4,5]

[1] Inria, France
[2] CNRS, France
[3] International Laboratory on Learning Systems (ILLS), Mc Gill - ETS - MILA, Université Paris-Saclay
[4] LIX, École Polytechnique, Institut Polytechnique de Paris
[5] Pontificia Universidad Javeriana Cali

[†] Corresponding author: carlos.pinzon@lix.polytechnique.fr

Keywords: equal opportunity, fairness, accuracy, trade-off, impossibility.

Abstract

One of the main concerns about fairness in machine learning (ML) is that, in order to achieve it, one may have to trade off some accuracy. To overcome this issue, Hardt et al.[13] proposed the notion of equality of opportunity (EO), which is compatible with maximal accuracy when the target label is deterministic with respect to the input features.

In the probabilistic case, however, the issue is more complicated: It has been shown that under differential privacy constraints, there are data sources for which EO can only be achieved at the total detriment of accuracy, in the sense that a classifier that satisfies EO cannot be more accurate than a trivial (i.e., constant) classifier. In this paper, we strengthen this result by removing the privacy constraint. Namely, we show that for certain data sources, the most accurate classifier that satisfies EO is a trivial classifier. Furthermore, we study the admissible trade-offs between accuracy and EO loss (opportunity difference) and characterize the conditions on the data source under which EO and non-trivial accuracy are compatible.

1 Introduction

During the last decade, the intersection between machine learning and social discrimination has gained considerable attention from academia, industry, and the public in general. A similar trend occurred before between machine learning and privacy, and even the three fields have been studied together recently [24, 6, 15, 1].

Fairness has proven to be harder to conceptualize than privacy, for which differential privacy has become the de-facto definition. Fairness is subjective and laws vary between countries. Even in academia, depending on the application, the words fairness and bias have different meanings [5]. The current general consensus is that fairness cannot be summarized into a unique universal definition; and for the most popular definitions, several trade-offs, implementation difficulties, and impossibility theorems have been found [16, 3]. One such definition of fairness is equal opportunity [13], which is one of the most common group notions of fairness along with disparate impact, demographic parity, and equalized odds [21]. Equal opportunity is restricted to binary classification tasks with binary sensitive attributes.

To contrast equal opportunity (EO) with accuracy, we borrow the notion of trivial accuracy from [6]. A *non-trivial* classifier is one that has higher accuracy than any constant classifier. Since constant classifiers are independent of the input, trivial accuracy determines a very low-performance level that any correctly trained classifier should overcome. Yet, as shown in related works [6, 1], under the simultaneous constraints of differential privacy and equal opportunity, it is impossible to have non-trivially accurate classifiers.

In this paper, we strengthen the result of [6, 1] by showing that, even without the assumption of differential privacy, there are distributions for which equal opportunity implies trivial accuracy. In particular, this is possible when the data source is probabilistic, i.e., the correct label for a given input is not necessarily unique.

Probability plays two different roles in this paper. On the one hand, we allow classifiers to be probabilistic, i.e. we allow the classification to be influenced by controlled randomness for some inputs. This is needed because satisfying equal opportunity typically requires a probabilistic predictor [13], but also because it has a practical justification. Namely, in some cases, randomness is the only fair way to distribute an indivisible limited resource. For instance, a parent with one candy and two children might throw a coin to decide whom to give it to. This principle is even applied in decisions that have a significant social impact such as the Diversity Visa Program to qualify for a Green Card in the United States [26], and the Beijing lottery for getting a car license plate [10].

On the other hand, we consider probabilistic data sources. This provides

a more general framework for studying the trade-off between fairness and accuracy, as there are situations in which reality is more accurately represented by a probabilistic model. For instance, the information carried by the input may be insufficient to conclude definitely the yes-no decision, or there may be constraints that force the decision to be different for identical inputs.

The analysis of this paper is mostly theoretical and is limited to the notion of equal opportunity, hence to distributions with binary targets and binary sensitive attributes. On the other hand, the results are very general. For instance, whenever we state a property about all predictors, it includes all probabilistic classifiers without exception. Hence our results hold for classifiers that do not use the sensitive attribute for prediction, as well as for those that use it to compensate existing biases, or take into account proxy variables, or use multiple threshold mechanisms, or are based on causality, or do not use machine learning at all.

The contributions of this paper can be summarized as follows.

1. We prove that for certain probabilistic distributions, no predictor can achieve EO and non-trivial accuracy simultaneously.
2. We explain how to modify existing results that assume deterministic data sources to the probabilistic case:
   (a) We prove that for certain distributions, the Bayes classifier does not satisfy EO. As a consequence, in these cases, EO can only be achieved by trading-off some accuracy.
   (b) We give necessary and sufficient conditions for non-trivially accurate predictors to exist.
3. We prove and depict several algebraic and geometric properties about the feasibility region, i.e. the region containing all predictors in the plane of opportunity difference versus error.
4. We determine necessary and sufficient conditions under which non-trivial accuracy and EO are compatible.
5. We develop an algorithm that computes the Pareto-optimal boundary of the accuracy-fairness trade-off, and more generally, the feasibility region.
6. We illustrate how the incompatibility between EO and non-trivial accuracy may arise in practice.
7. We discuss the distortion effect that arises when we use the above algorithm on empirical distributions from sampled data.

For reproducibility, we published a repository [23] with Python code for generating the figures and algorithms mentioned in this paper, including Algorithms 1 and 2.

The rest of the paper is organized as follows. Section 2 discusses related work. Section 3 recalls the preliminary notions that are used in the rest

of the document. Section 4 introduces the plane of error versus opportunity difference and shows several geometric properties taking place in this plane. Section 5 presents the impossibility result: for certain probabilistic distributions, no predictor can achieve EO and non-trivial accuracy simultaneously. Section 6 compares deterministic sources against probabilistic ones, and shows how to modify existing results that hold in the former case to guarantee them in the latter. Section 7 presents algorithms for computing the Pareto-optimal frontier and all the vertices of the feasibility region in the plane of error versus opportunity difference. Section 8 states the necessary and sufficient conditions under which there exist predictors achieving EO and non-trivial accuracy simultaneously. Section 9 shows an example of the impossibility result arising in (a variant of) a real-life dataset. Section 10 discusses the distortion on the Pareto-optimal frontier when we compute and evaluate it using empirical distribution from sampled data. Section 11 draws the conclusion. Finally, Section 12 presents some auxiliary lemmas and their proofs.

A preliminary and partial version of this paper appeared in the proceedings of AAAI 2022 [22]. In this document, we extend the AAAI-2022 version by studying the Pareto-optimal boundary (Section 7), the necessary and sufficient conditions (Theorem 10) that characterize the impossibility between equal opportunity and non-trivial accuracy, and a practical example based on the Adult dataset (Figure 9).

2 Related Work

This paper contributes to the technical literature about equal opportunity (EO) [13], one of the most common group fairness notions [21]. For an overview of when EO is appropriate and how EO relates to other fairness notions, the reader is referred to the survey papers [18, 21, 2, 19] and the moral framework in [14].

Our paper is strongly related to the following two papers that consider a randomized learning algorithm guaranteeing (exact) EO and also satisfying differential privacy: [6] shows that, for certain distributions, these constraints imply trivial accuracy. [1] proves the same claim for any arbitrary distribution and for non-exact EO, i.e. bounded opportunity difference. It also highlights that, although there appears to be an error in the proof of [6], the statement is still correct. In contrast, in this paper, we prove the existence of particular distributions in which trivial accuracy is implied directly from the (exact) EO constraint, without any differential privacy assumption.

There are also several works that focus on the incompatibility of fairness constraints. In [16], it is shown that several different fairness notions cannot hold simultaneously, except for exceptional cases. Similarly, in [17], it is shown that the two main legal notions of discrimination are

in conflict for some scenarios. In particular, when impact parity and treatment parity are imposed, the learned model seems to decide based on irrelevant attributes. These works reveal contradictions when different notions of fairness are imposed together.

In contrast, [4] show issues inherent to anti-classification, classification parity, and calibration, separately, without inducing them simultaneously with another fairness notion. Regarding equal opportunity in the COMPAS case, they show that forcing equal and low false positive rates obliges the system to decide almost randomly (trivially) for black defendants. Our work presents theoretical scenarios in which this problem is even more extreme and the system becomes trivial for both classes. As shown in our sufficiency and necessary conditions, the extreme scenarios are characterized based on six population statistics. In this sense, our paper is also related to [25], which computes bounds on fairness and accuracy based on population statistics.

Lastly, in comparison to the seminal paper on equal opportunity [13], this paper uses a different geometric approach. Graphically, their analysis is carried out using ROC curves of fixed predictors. In contrast, we plot directly the error and the difference in opportunity of the two sensitive groups. In Section 5, Figure 9, we depict side by side the two perspectives. In this sense, we provide a complementary geometric perspective for analyzing equal opportunity and accuracy together.

3 Preliminaries

The notation described in this section is summarized in Table 1.

We consider the problem of binary classification with a binary protected feature. *Protected features*, also called sensitive attributes or sensitive features, are input features that represent race, gender, religion, nationality, age, or any other variable that could potentially be used to discriminate against a group of people. A feature may be considered a protected feature in some contexts and not in others, depending on whether the classification task should ideally consider that feature or not. For our purposes, we assume the simple and fundamental case in which there is a single protected attribute that can only take two values, e.g. man or woman, or, religious or non-religious.

Data Source

We consider an observable underlying statistical model consisting of three random variables over a probability space $(\Omega, \mathcal{E}, \mathbb{P})$: the *protected feature* $A : \Omega \to \{0, 1\}$, the *non-protected feature vector* $X : \Omega \to \mathbb{R}^d$ for some

positive integer d, and the *target label* $Y : \Omega \to \{0,1\}$. We refer to this statistical model as the *data source*.

The distribution of (X, A) is denoted by the measure π that computes for each ((X, A)-measurable) event $E \subseteq \mathbb{R}^d \times \{0,1\}$, the probability $\pi(E) \stackrel{\text{def}}{=} \mathbb{P}[\,(X, A) \in E\,]$. To reduce the verbosity of the discrete case, we denote the probability mass function as $\pi(x, a) \stackrel{\text{def}}{=} \pi(\{(x, a)\})$, i.e. $\pi(x, a) = \mathbb{P}[\,X = x, A = a\,]$.

The expectation of Y conditioned on (X, A) is denoted both as the function $q(x, a) \stackrel{\text{def}}{=} \mathbb{E}[\,Y \mid X = x, A = a\,]$ and the random variable $Q \stackrel{\text{def}}{=} \mathbb{E}[\,Y \mid X, A\,] = q(X, A)$. Importantly, the notation $\mathbb{E}[\,Y \mid X = x, A = a\,]$ for defining $q(x, a)$ is not an expectation conditioned on the possibly null event $(X = x, A = a)$. Instead, it is syntactic sugar for the conditional expectation function. Formally speaking, the function q is not necessarily unique in the way it is defined. It is defined almost everywhere uniquely, so that for any alternative conditional expectation function q', we have $q(X, A) = q'(X, A)$ almost surely. Throughout the paper, we prioritize studying the discrete case to avoid this extreme level of formalism without losing rigor.

The random variable Q plays the role of a soft target label because, since $q(x, a) = \mathbb{P}[\,Y = 1 \mid X = x, A = a\,]$, then Y can be modeled as a Bernoulli random variable with success probability Q.

The distribution of (X, A, Y) is completely characterized by the pair (π, q), hence we refer to this pair as the distribution of the data source. And we distinguish two cases: the data source is *probabilistic* in general, but if $Q \in \{0,1\}$ (with probability 1), then it is said to be *deterministic*. This distinction is crucial, because several statements hold exclusively in one of the two cases.

(X, A, Y)	Data source
X	Non-protected feature vector in $\mathbb{R}^d$
A	Protected feature in $\{0,1\}$
Y	Target label in $\{0,1\}$
Q, q	Soft target label $Q \stackrel{\text{def}}{=} \mathbb{E}[\,Y \mid X, A\,]$
π	Distribution of (X, A)
(π, q)	Distribution of (X, A, Y)
$\hat{Q}, \hat{q}$	Predictor $\hat{Q} = \hat{q}(X, A) = \mathbb{E}[\,\hat{Y} \mid X, A\,]$
$\hat{Y}$	Predicted label in $\{0,1\}$
$\mathcal{Q}$	Set of all predictors
$\mathrm{acc}(\hat{Q})$	Accuracy of $\hat{Q}$: $\mathbb{P}[\,\hat{Y} = Y\,]$
$\mathrm{oppDiff}(\hat{Q})$	Opportunity difference of $\hat{Q}$: $\mathbb{E}[\,\hat{Q} \mid Y = 1, A = 1\,] - \mathbb{E}[\,\hat{Q} \mid Y = 1, A = 0\,]$

Table 1: The notation used in the paper.

Classifiers and Predictors

Analogously to the data source, we model the estimation $\hat{Y}$ as a Bernoulli random variable with success probability $\hat{Q} = \hat{q}(X, A)$ for some ((X, A)-measurable) function $\hat{q}$. We refer to $\hat{Y}$ as a (hard) *classifier*, and to $\hat{Q}$ or $\hat{q}$ as a (soft) *predictor*. Notice that $\hat{Y}$ is deterministic when $\hat{Q} \in \{0, 1\}$ (with probability 1), in which case, $\hat{Y} = \hat{Q}$ (w.p. 1). Hence all deterministic classifiers are also predictors.

The set of all soft predictors is denoted as $\mathcal{Q}$. We highlight the following predictors in $\mathcal{Q}$:

1. the two constant classifiers, $\hat{0}$ and $\hat{1}$, given by $\hat{0}(x, a) \stackrel{\text{def}}{=} 0$ and $\hat{1}(x, a) \stackrel{\text{def}}{=} 1$,
2. for each $\hat{Q} \in \mathcal{Q}$, the $1/2$-threshold classifier given by $\hat{Q}_{1/2} \stackrel{\text{def}}{=} \mathbf{1}_{\hat{Q}>1/2}$,
3. the data source soft target Q, and
4. the Bayes classifier $Q_{1/2} = \mathbf{1}_{Q>1/2}$.

It is well known[1] that the Bayes classifier $Q_{1/2}$ has minimal error among all predictors in $\mathcal{Q}$, regardless of whether the data source is deterministic or not.

Evaluation Metrics

To refer to *equal opportunity* [13], we introduce a continuous metric called the *opportunity difference*. The opportunity difference of a predictor $\hat{Q} \in \mathcal{Q}$ is defined as

$$\begin{aligned}\text{oppDiff}(\hat{Q}) \stackrel{\text{def}}{=} (\, &\mathbb{P}[\,\hat{Y}{=}1 \,|\, A{=}1, Y{=}1\,] \\ &- \mathbb{P}[\,\hat{Y}{=}1 \,|\, A{=}0, Y{=}1\,],\,)\end{aligned}$$

and a predictor $\hat{Q} \in \mathcal{Q}$ is said to satisfy equal opportunity whenever $\text{oppDiff}(\hat{Q}) = 0$.

The *error* and the *accuracy* of a predictor $\hat{Q} \in \mathcal{Q}$ are defined as

$$\begin{aligned}\text{err}(\hat{Q}) &\stackrel{\text{def}}{=} \mathbb{P}[\,\hat{Y} \neq Y\,], \\ \text{acc}(\hat{Q}) &\stackrel{\text{def}}{=} 1 - \text{err}(\hat{Q}).\end{aligned}$$

Additionally, we consider a minimal reference level of accuracy that should be outperformed intuitively by any well-trained predictor. The *trivial accuracy* [6] is defined as $\tau \stackrel{\text{def}}{=} \max\left\{\text{acc}(\hat{Q}) : \hat{Q} \in \text{Triv}\right\}$, where Triv is the

[1] See for instance Chapter 3 of [9].

set of (trivial) predictors whose output does not depend on X and A at all, and as a consequence is in independent of Y as well. In other words, Triv consists of all constant soft predictors $\text{Triv} \stackrel{\text{def}}{=} \{((x,a) \mapsto c) : c \in [0,1]\}$. According to the Neyman-Pearson Lemma, the most accurate trivial predictor is always hard, i.e. must be either $\hat{0}$ or $\hat{1}$. Thus τ is well-defined and can be computed as

$$\tau = \max\{\mathbb{P}[Y=0], \mathbb{P}[Y=1]\}.$$

A predictor $\hat{Q} \in \mathcal{Q}$ is said to be *trivially accurate* if $\text{acc}(\hat{Q}) \leq \tau$, and *non-trivially accurate*, or *non-trivial* otherwise. Notice that for a degenerated data source in which the decision Y is independent of X and A, all predictors are forcibly trivially accurate.

4 The Error versus Opportunity-Difference Region

In this section, we analyze the region $M \subseteq [0,1] \times [-1,+1]$ given by

$$M \stackrel{\text{def}}{=} \{(\text{err}(\hat{Q}), \text{oppDiff}(\hat{Q})) : \hat{Q} \in \mathcal{Q}\},$$

which represents the feasible combinations of the evaluation metrics (error and opportunity difference) that can be obtained for a given source distribution (π, q). This region determines the tension between error and opportunity difference. Figure 1 shows an example of the region M.

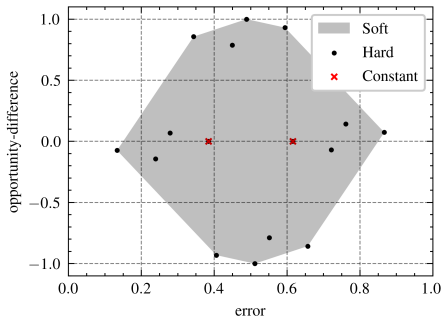


Figure 1: Region M for an arbitrary source distribution.

The results presented in this section assume that the data source is discrete and its range is finite. We will use the following vectorial notation to represent both the distribution (π, q) and any arbitrary predictor $\hat{Q} \in \mathcal{Q}$.

Definition 1. Suppose (X, A) can only take a finite number of outcomes $\{(x_i, a_i)\}_{i=1}^{n}$ (each with positive probability) for some integer $n > 0$. In

order to represent π, q and any $\hat{Q} \in \mathcal{Q}$ respectively, let $\vec{P}, \vec{Q}, \vec{F} \in \mathbb{R}^n$ be the vectors given by

$$\begin{aligned}
\vec{P}_i &\stackrel{\text{def}}{=} \mathbb{P}[\, X{=}x_i, A{=}a_i \,], \\
\vec{Q}_i &\stackrel{\text{def}}{=} \mathbb{P}[\, Y{=}1 \,|\, X{=}x_i, A{=}a_i \,], \\
\vec{F}_i &\stackrel{\text{def}}{=} \mathbb{P}[\, \hat{Y}{=}1, X{=}x_i, A{=}a_i \,].
\end{aligned}$$

For notation purposes, let also $\vec{Q}^{(0)}, \vec{Q}^{(1)} \in \mathbb{R}^n$ be given by $\vec{Q}_i^{(a)} \stackrel{\text{def}}{=} \vec{Q}_i \cdot \mathbf{1}_{a_i=a}$, and, following the definition of $\text{err}(\hat{Q})$ and $\text{oppDiff}(\hat{Q})$, let

$$\begin{aligned}
\text{err}(\vec{F}) &\stackrel{\text{def}}{=} \langle \vec{P}, \vec{Q} \rangle + \langle \vec{F}, 1{-}2\vec{Q} \rangle, \\
\text{oppDiff}(\vec{F}) &\stackrel{\text{def}}{=} \frac{\langle \vec{F}, \vec{Q}^{(1)} \rangle}{\langle \vec{P}, \vec{Q}^{(1)} \rangle} - \frac{\langle \vec{F}, \vec{Q}^{(0)} \rangle}{\langle \vec{P}, \vec{Q}^{(0)} \rangle}.
\end{aligned}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product:

$$\langle u, v \rangle \stackrel{\text{def}}{=} u_1 v_1 + \cdots + u_n v_n.$$

(End)

Regarding Definition 1, we highlight three important remarks:

1. $\vec{Q} \in [0,1]^n$, $\vec{P} \in (0,1]^n$, $|\ \vec{P}\ |_1 = 1$ and $\vec{F}$ lies in the rectangular n-dimensional box given by

   $$0 \preceq \vec{F} \preceq \vec{P},$$

   where $\preceq$ denotes the componentwise order in $\mathbb{R}^n$, i.e. $0 \leq \vec{F}_i \leq \vec{P}_i$ for each $i \in \{1, ..., n\}$. Moreover, from the definition of $\vec{P}$ and $\vec{F}$, the vertices of this rectangular box correspond precisely with the deterministic predictors.

2. The vectorial definitions of error and opportunity difference correspond to those of the non-vectorial case. Moreover, their gradients are constant.

3. There is a one-to-one correspondence between the predictors $\hat{q} \in \mathcal{Q}$ and the vectors $\vec{F}$ that satisfy $0 \preceq \vec{F} \preceq \vec{P}$. Indeed, each predictor is uniquely given by its pointwise values $\hat{q}(x_i, a_i) = \frac{\vec{F}_i}{\vec{P}_i}$ and each vector by its pointwise coordinates $\vec{F}_i = \vec{P}_i \hat{q}(x_i, a_i)$. Therefore

   $$M = \{(\text{err}(\vec{F}), \text{oppDiff}(\vec{F})) : 0 \preceq \vec{F} \preceq \vec{P}\}.$$

We now make use of results from a different research area in mathematics (geometry) to conclude the main properties of the region M.

Theorem 1. *Assuming a discrete data source with finitely many possible outcomes, the region M of feasible combinations of error versus opportunity difference satisfies the following claims.*

1. *M is a convex polygon.*
2. *The vertices of the polygon M correspond to some deterministic predictors.*
3. *M is symmetric with respect to the point $(1/2, 0)$.*

Proof. The proof is based on the fact that affine transformations map polytopes into polytopes (See Chapter 3 of [12]).

Assume the notation of Definition 1.

Part 1. In geometrical terms, M is the result of applying an affine transformation, i.e. a linear transformation and a translation, to the n-dimensional *polytope* given by $0 \preceq \vec{F} \preceq \vec{P}$.

Affine transformations are known to map polytopes into polytopes (See Chapter 3 of [12]), therefore M must be a 2-dimensional polytope, i.e. the region M is a convex polygon. In theory, this region may also be a 1-dimensional segment, but this can only occur in the extreme case that $Q = 1/2$ (with probability 1).

Part 2. The vertices of a polytope, also called extremal points, are the points in the polytope that are not in the segment between any two other points in the polytope. It is known from geometry theory that affine mappings preserve collinearity, i.e. they map segments into segments, thus they map non-vertices into non-vertices. As a consequence, the vertices of the polygon M correspond to some vertices of the polytope $0 \preceq \vec{F} \preceq \vec{P}$, that is, to some deterministic classifiers.

Part 3. Notice (Lemma 15) that

$$\begin{aligned} \mathrm{err}(\vec{P} - \vec{F}) &= 1 - \mathrm{err}(\vec{F}), \\ \mathrm{oppDiff}(\vec{P} - \vec{F}) &= -\mathrm{oppDiff}(\vec{F}). \end{aligned}$$

This implies that for each point $(\mathrm{err}(\vec{F}), \mathrm{oppDiff}(\vec{F})) \in M$, there is another one, namely $(\mathrm{err}(\vec{P} - \vec{F}), \mathrm{oppDiff}(\vec{P} - \vec{F})) \in M$ that is symmetrical to the former w.r.t the point $(1/2, 0)$. Geometrically, this means that the polygon M is symmetric with respect to the point $(1/2, 0)$. □

The reader is invited to visualize the properties of M mentioned in Theorem 1 in Figure 1, which depicts the region M for a particular instance [2] of $\vec{P}$ and $\vec{Q}$.

[2] Namely `P=[0.267 0.344 0.141 0.248]`, `Q=[0.893 0.896 0.126 0.207]` and `A=[0 1 0 1]`.

5 Strong Impossibility Result

Contrasting with Figure 1 in the previous section, Figure 2 shows a data source for which the constant classifiers are vertices of the polygon. Figure 2 was generated using the theory developed in this section and it illustrates the strong incompatibility that may occur in certain distributions. Namely, among the predictors satisfying equal opportunity (those in the X-axis), the minimal error is achieved by a constant classifier.

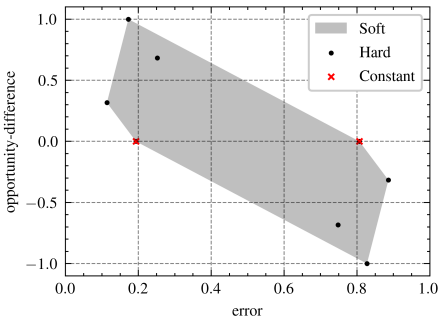


Figure 2: In this distribution, the constant classifiers are vertices of the polygon, thus the constraints of equal opportunity and non-trivial accuracy can not be satisfied simultaneously.

In other words, there are data sources for which no predictor can achieve equal opportunity and non-trivial accuracy simultaneously. This is Theorem 3.

Since Theorem 3 is our strongest result, we also show how to generalize it to non-finite domains. For this purpose, and focusing on formality, we state in Definition 2 very precisely, for which kind of domains it applies.

Definition 2. The *essential range* of a random variable $S : \Omega \to \mathbb{R}^k$ is the set

$$\{\vec{s} \in \mathbb{R}^k : (\forall \epsilon > 0)\ \mathbb{P}[\,|S - \vec{s}| < \epsilon\,] > 0\}.$$

We call a set $\mathcal{D} \subseteq \mathbb{R}^k$ an *essential domain* if it is the essential range of any random variable.

Definition 2 excludes pathological domains such as non-measurable sets, the Cantor set, or the irrationals. But it allows for isolated points, convex and closed sets, finite unions of them, and countable unions of them as long as the resulting set is closed. This includes typical domains, such as products of closed intervals $\prod_{i=1}^{n}[l_i, r_i]$, or the whole space $\mathbb{R}^n$.

Theorem 3. *For any essential domain $\mathcal{X} \subseteq \mathbb{R}^d$ with $|\mathcal{X}| \geq 2$, there exists a data source (X, A, Y) whose essential range is $\mathcal{X} \times \{0,1\}^2$ and such that the accuracy $acc(\hat{Q})$ of any predictor $\hat{Q} \in \mathcal{Q}$ that satisfies equal opportunity is at most the trivial accuracy $\tau \in [0, 1)$.*

Proof. The proof is divided into four parts. We will (i) reduce the problem into an algebraic one; (ii) find the linear constraints that solve the

algebraic problem when satisfied; (iii) provide an algorithm that generates vectors that satisfy the linear constraints; and finally, (iv) convert the vectorial solution back into a distribution (π, q) for the given domain.

Part 1. Reduction to an algebraic problem.

Partition the non-protected input space $\mathcal{X}$ into two non-empty sets $\mathcal{X}_1, \mathcal{X}_2$, and the input space $\mathcal{X} \times \{0,1\}$ into three regions R_j:

$$\begin{aligned} R_1 &= \mathcal{X}_1 \times \{0\}, \\ R_2 &= \mathcal{X}_2 \times \{0\}, \\ R_3 &= \mathcal{X} \times \{1\}. \end{aligned}$$

For any distribution (π, q) for which these 3 regions have positive probabilities, denote $\vec{P}_j \stackrel{\text{def}}{=} \mathbb{P}[\,(X,A) \in R_j\,] > 0$ and $\vec{Q}_j \stackrel{\text{def}}{=} \mathbb{P}[\,Y=1 \,|\, (X,A) \in R_j\,]$ for $j \in \{1,2,3\}$. We search for constraints over $\vec{P}$ and $\vec{Q}$ that are feasible and cause $\text{acc}(\hat{Q}) \leq \tau$ for any fair predictor $\hat{Q} \in \mathcal{Q}$ satisfying EO. The first such constraint is

C1. $\vec{P}, \vec{Q} \in (0,1)^3$.

That is, we require $\vec{P}_j$ to be positive, and Y to have at least some degree of randomness in each region.

Given a reference predictor $\hat{Q}$, let $\vec{F} \in [0,1]^3$ be the vector given by $\vec{F}_j \stackrel{\text{def}}{=} \mathbb{P}[\,\hat{Y}=1, (X,A) \in R_j\,]$. Lemma 14 shows that the accuracy and the opportunity difference of any predictor $\hat{Q}$ can be computed from $\vec{P}$, $\vec{Q}$ and $\vec{F}$ as

$$\begin{aligned} \text{acc}(\hat{Q}) &= \langle \vec{F}, 2\vec{Q} - 1 \rangle + C_{\vec{Q}}, \\ \text{oppDiff}(\hat{Q}) &= \frac{\vec{F}_3}{\vec{P}_3} - \frac{\vec{F}_1 \vec{Q}_1 + \vec{F}_2 \vec{Q}_2}{\vec{P}_1 \vec{Q}_1 + \vec{P}_2 \vec{Q}_2}, \end{aligned}$$

where $C_{\vec{Q}} \stackrel{\text{def}}{=} 1 - \langle \vec{P}, \vec{Q} \rangle$ is a constant and the operator $\langle \cdot, \cdot \rangle$ denotes the inner product explained in Definition 1. Since we are interested in relative accuracies with respect to the trivial predictors, the constant $C_{\vec{Q}}$ is mostly irrelevant. For this reason, we let $L(\vec{F}) \in [-1,1]$ denote the non-constant component of the accuracy $L(\vec{F}) \stackrel{\text{def}}{=} \langle \vec{F}, 2\vec{Q} - 1 \rangle$.

Both accuracy and opportunity difference are completely determined for any predictor by the vectors $\vec{P}$, $\vec{Q}$ and $\vec{F}$ as shown above. Moreover, both quantities are linear with respect to $\vec{F}$.

Regarding equal opportunity, the constraint $\text{oppDiff}(\hat{Q}) = 0$ forms a plane in $\mathbb{R}^3$, depicted in Figure 3. This plane passes through the origin, is determined by $\vec{P}$ and $\vec{Q}$, and contains all vectors $\vec{F}$ (restricted to $0 \leq$

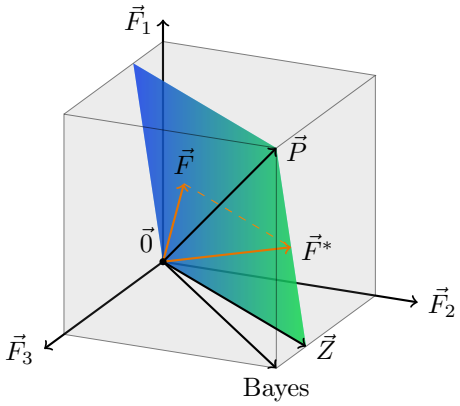


Figure 3: In vectorial form, the predictors that satisfy equal opportunity form a plane inside the rectangular box of all predictors.

$\vec{F}_j \leq \vec{P}_j$) that satisfy

$$\vec{F}_3(\vec{P}_1\vec{Q}_1 + \vec{P}_2\vec{Q}_2) - \vec{P}_3(\vec{F}_1\vec{Q}_1 + \vec{F}_2\vec{Q}_2) = 0,$$

or equivalently, all vectors F that are normal to the vector $(-\vec{P}_3\vec{Q}_1, -\vec{P}_3\vec{Q}_2, \vec{P}_1\vec{Q}_1 + \vec{P}_2\vec{Q}_2)$.

Regarding accuracy, the two constant predictors correspond to $\vec{F} = \vec{0}$ and $\vec{F} = \vec{P}$, thus $\tau = C_Q + \max\{L(\vec{0}), L(\vec{P})\}$. Importantly, both of them lie on the equal opportunity plane.

The problem is now reduced to finding vectors $\vec{P}$ and $\vec{Q}$ such that all vectors $\vec{F}$ in the equal opportunity plane satisfy $L(\vec{F}) \leq \max\{L(\vec{0}), L(\vec{P})\}$.

Part 2. Constraints for the algebraic solution.

To fix an orientation, let us impose these constraints:

C2. Among the constant predictors, the accuracy of $\vec{F} = \vec{P}$ is higher than that of $\vec{F} = \vec{0}$. This is $L(\vec{P}) > 0 = L(\vec{0})$.

C3. The Bayes classifier is located at $(0, \vec{P}_2, \vec{P}_3)$ as in Figure 3. Algebraically this means $\vec{Q}_1 < 1/2$ and $\vec{Q}_2, \vec{Q}_3 > 1/2$.

In order to derive the constraints that make the scalar field L maximal at $\vec{P}$ over the plane, consider the vector $\vec{Z}$ that lies on the plane and has minimal $\vec{Z}_1$ and maximal $\vec{Z}_2$, i.e.

$$\vec{Z} \stackrel{\text{def}}{=} (0, \vec{P}_2, \vec{P}_3 \frac{\vec{P}_2\vec{Q}_2}{\vec{P}_1\vec{Q}_1 + \vec{P}_2\vec{Q}_2}).$$

Since the gradient of L is given by $2\vec{Q} - 1$ and has signs $(-, +, +)$, then for any vector $\vec{F}$ in the plane, there is $\vec{F}^*$ in the segment between $\vec{P}$ and

$\vec{Z}$ such that $\vec{F}_1 = \vec{F}_1^*$ and $L(\vec{F}^*) \geq L(\vec{F})$ (refer to Figure 3). This implies that the L attains its maximal value on the segment between $\vec{P}$ and $\vec{Z}$. Hence, for L to be maximal at $\vec{P}$, it would suffice to have $L(\vec{P}) > L(\vec{Z})$. As shown in Lemma 20, this can be achieved by imposing, in addition,

C4. $\vec{Q}_3 + \vec{Q}_1 \geq 1$, and
C5. $\vec{P}_1\vec{Q}_1 + \vec{P}_2\vec{Q}_2 < \vec{P}_3\vec{Q}_1$.

Part 3. Solution to the constraints.

Algorithm 1 is a randomized algorithm that generates random vectors. We will prove that the output vectors $\vec{P}$ and $\vec{Q}$ satisfy the constraints of the previous parts of this proof, regardless of the seed and the random sampling function, e.g. uniform. For corroboration and illustration, the distribution in Figure 2 presented early was generated using this algorithm[3].

Algorithm 1 Random generator for Theorem 3.

procedure VectorGenerator(seed)
  Initialize random sampler with the seed
  $\vec{Q}_1 \leftarrow$ random in $(0, 1/2)$
  $\vec{Q}_2 \leftarrow$ random in $(1/2, 1)$
  $\vec{Q}_3 \leftarrow$ random in $(1 - \vec{Q}_1, 1)$
  $\vec{P}_3 \leftarrow$ random in $(1/2, 1)$
  $a \leftarrow \max\{(1 - \vec{P}_3)\vec{Q}_1, 1/2 - \vec{P}_3\vec{Q}_3\}$
  $b \leftarrow \min\{(1 - \vec{P}_3)\vec{Q}_2, \vec{P}_3\vec{Q}_1\}$
  $c \leftarrow$ random in (a, b)
  $\vec{P}_2 \leftarrow {}^{(c - \vec{Q}_1(1-\vec{P}_3))}/_{\vec{Q}_2 - \vec{Q}_1}$
  $\vec{P}_1 \leftarrow 1 - \vec{P}_3 - \vec{P}_2$
  return $\vec{P}, \vec{Q}$

Two immediate observations about Algorithm 1 are that the construction of $\vec{Q}$ implies that constraints C3 and C4 are satisfied, and the construction of $\vec{P}$ implies $\vec{P}_1 + \vec{P}_2 + \vec{P}_3 = 1$. To prove the correctness of the algorithm, it remains to prove that (i) $a < b$ (otherwise the algorithm would not be well-defined), that (ii) $\vec{P}_2 \in (0, 1)$ for constraint C1, and also that (iii) constraints C2 and C5 are satisfied. For better readability, the algebraic proof of these claims is moved to Lemma 21.

Part 4. Construction of the distribution.

Generate a pair of vectors $\vec{P}$ and $\vec{Q}$ using the algorithm of the previous part (Part 3). The first goal is to partition $\mathcal{X}$ into $\mathcal{X}_1$ and $\mathcal{X}_2$ to generate the regions R_1, R_2 and R_3. The second goal is to define π in such a way

[3]The algorithm's output was `P=[0.131 0.096 0.772]` and `Q=[0.274 0.858 0.891]`. Also, `A=[0 0 1]` from the partition $\{R_1, R_2, R_3\}$.

that $\mathbb{P}[\,(X,A)\in R_j\,] = \vec{P}_j$ for each $j \in \{1,2,3\}$. The third and last goal is to define q so that $\mathbb{E}[\,Q \mid (X,A) \in R_j\,] = \vec{Q}_j$ for each j. This can be done immediately by letting $q(x,a) \stackrel{\text{def}}{=} \vec{Q}_j$ for all $(x,a) \in R_j$. Thus only the first two goals remain.

For the first goal, since $|\mathcal{X}| \geq 2$, we may create a simple Voronoi clustering diagram by choosing two different arbitrary points $s_1, s_2 \in \mathcal{X}$, and letting $\mathcal{X}_1 \stackrel{\text{def}}{=} \{s \in \mathcal{X} : |s - s_1| \leq |s - s_2|\}$ and $\mathcal{X}_2 \stackrel{\text{def}}{=} \mathcal{X} \setminus \mathcal{X}_1$.

For the second goal, since $\mathcal{X}$ is an essential domain, there exists a random variable S whose essential range is $\mathcal{X}$. Notice that $\mathbb{P}[\,S \in \mathcal{X}_j\,] \geq \mathbb{P}[\,|S - s_j| < |s_1 - s_2|/2\,] > 0$ for each $j \in \{1,2\}$. For each ((X,A)-measurable) event E, let $E_a \stackrel{\text{def}}{=} \{x : (x,a) \in E\}$, and define $\pi(E)$ as

$$\begin{aligned}
\mathbb{P}[\,(X,A)\in E\,] &\stackrel{\text{def}}{=} \sum_{a=0,1} \mathbb{P}[\,X \in E_a, A = a\,], \\
\mathbb{P}[\,X \in E_0, A = 0\,] &\stackrel{\text{def}}{=} \sum_{j=1,2} \mathbb{P}[\,S \in E_0 \,|\, S \in \mathcal{X}_j\,]\vec{P}_j, \\
\mathbb{P}[\,X \in E_1, A = 1\,] &\stackrel{\text{def}}{=} \mathbb{P}[\,S \in E_1\,]\vec{P}_3.
\end{aligned}$$

This forces $\mathbb{P}[\,(X,A)\in R_j\,] = \vec{P}_j$ for each $j \in \{1,2,3\}$ as desired. □

Finally, to conclude this section we present Example 1, which shows that there are many other scenarios, not necessarily those of Theorem 3, in which EO and non-trivial accuracy are incompatible.

Example 1. Consider a data source (X,A,Y) over $\{0,1\}^3$ whose distribution is given by

x	a	$\pi(x,a)$	$q(x,a)$
0	0	3/8	9/20
0	1	2/8	15/20
1	0	1/8	15/20
1	1	2/8	16/20

Then, (i) there are predictors satisfying equal opportunity, (ii) there are predictors with non-trivial accuracy, but (iii) there are no predictors satisfying both. *(End)*

Indeed, Figure 4 depicts the region M for Example 1. On the one hand, the set of non-trivially accurate predictors corresponds to the area with an error strictly smaller than the left constant classifier. On the other hand, the set of equal opportunity predictors is (for this particular example) the closed segment between the two constant classifiers. As claimed in Example 1 (and depicted in Figure 4), these two sets are non-empty and do not intersect each other.

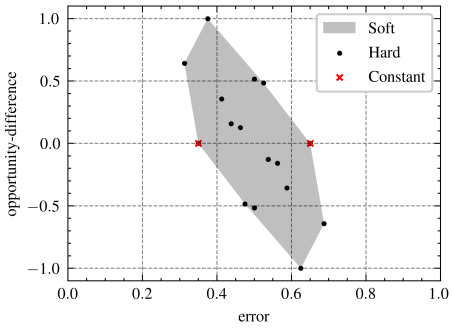


Figure 4: Example 1. One of the constant classifiers is Pareto-optimal.

6 Probabilistic versus Deterministic Sources

In this section, we compare the tension between error and opportunity difference when the data source is deterministic and probabilistic. The motivation for studying the probabilistic case is presented in the introduction. Particularly, we show that some known properties that apply for the discrete case may fail to hold for the probabilistic one, and under what conditions this happens.

6.1 Deterministic Sources

Under the assumption that the data source is deterministic, there are some important existing results showing the compatibility between equal opportunity and high accuracy:

Fact 4. Assuming a deterministic data source, the Neyman Pearson lemma [8] implies that if $\tau < 1$, then there is always a non-trivial predictor, for instance, the Bayes classifier $Q_{1/2}$. Otherwise (degenerated case with $\tau = 1$) all predictors are trivially accurate.

Fact 5. Assuming a deterministic data source, the Bayes classifier $Q_{1/2}$ satisfies equal opportunity necessarily [13].

As a consequence, EO and maximal accuracy (thus also non-trivial accuracy) are always compatible provided $\tau < 1$, because the Bayes classifier satisfies both. This is a celebrated fact and it was part of the motivations of [13] for defining equal opportunity, because other notions of fairness, including statistical parity, are incompatible with accuracy.

6.2 Probabilistic Sources

If we allow the data source to be probabilistic, the results of the deterministic case change. In particular, Fact 4 is generalized by Proposition 6 and Fact 5 is affected by Proposition 7 and Example 1.

Analogous to τ for deterministic sources, we define a second reference value $\tau^* \in [0, 1]$. We let

$$\tau^* \stackrel{\text{def}}{=} \max\left\{\mathbb{P}[\,Q \geq 1/2\,],\, \mathbb{P}[\,Q \leq 1/2\,]\right\},$$

highlighting that (i) $Q = q(X, A)$ is a random variable varying in $[0, 1]$, (ii) τ and τ^* are equal when the data source is deterministic, and (iii) the condition $\tau = 1$ implies $\tau^* = 1$, but not necessarily the opposite.

As shown in Proposition 6, the equation $\tau^* = 1$ characterizes the necessary and sufficient conditions on the data source for non-trivially accurate predictors to exist.

Particularly, in the deterministic case, we have $\tau^* = \tau$, and Proposition 6 resembles Fact 4.

Proposition 6. (Characterization of the impossibility of non-trivial accuracy)

For any arbitrary source distribution (π, q), non-trivial predictors exist if and only if $\tau^* < 1$.

Proof. The proof intuition is that if $\mathbb{P}[\,Q \geq 1/2\,] = 1$, then predicting 1 for any input is optimal, and vice versa.

We will prove that all predictors are trivially accurate if and only if $\tau^* = 1$.

($\Leftarrow$) Suppose $\tau^* = 1$, i.e. $\mathbb{P}[\,Q \leq 1/2\,] = 1$ or $\mathbb{P}[\,Q \geq 1/2\,] = 1$.

In the former case, the Bayes classifier $Q_{1/2}$ is the constant predictor $(x, a) \mapsto 0$, thus $\mathrm{acc}(Q_{1/2}) \leq \tau$ necessarily. In the latter case, the alternative Bayes classifier $Q^*_{1/2}$ (defined in Lemma 18) is the constant predictor $(x, a) \mapsto 1$, thus $\mathrm{acc}(Q^*_{1/2}) \leq \tau$. According to Lemma 18, $\mathrm{acc}(Q_{1/2}) = \mathrm{acc}(Q^*_{1/2})$, thus we may conclude $\mathrm{acc}(Q_{1/2}) \leq \tau$ as well.

It follows that $\mathrm{acc}(\hat{Q}) \leq \mathrm{acc}(Q_{1/2}) \leq \tau$ for all $\hat{Q} \in \mathcal{Q}$ because $Q_{1/2}$ has maximal accuracy in $\mathcal{Q}$.

($\Rightarrow$) Suppose $\tau^* < 1$. We will suppose that the Bayes classifier $Q_{1/2}$ is not trivially accurate and find a contradiction.

Suppose $\mathrm{acc}(Q_{1/2}) - \tau = 0$. According to Lemmas 16 and 19 we may rewrite this as $\mathbb{E}[\,|Q - 1/2| - |\mathbb{E}[\,Y\,] - 1/2|\,] = 0$. Using the reverse triangle inequality, we conclude $\mathbb{E}[\,|Q - \mathbb{E}[\,Y\,]|\,] = 0$, thus $Q = \mathbb{E}[\,Y\,]$ is constant.

If $\mathbb{E}[\,Y\,] \leq 1/2$, then $\mathbb{P}[\,Q \leq 1/2\,] = 1$. If $\mathbb{E}[\,Y\,] \geq 1/2$, then $\mathbb{P}[\,Q \geq 1/2\,] = 1$. In any case, we have $\tau^* = 1$ which contradicts the initial supposition.

□

Finally, in Proposition 7 and its proof, we show a simple family of probabilistic examples for which equal opportunity and optimal accuracy (obtained by the Bayes classifier) are not compatible. This issue does not merely arise from the fact that the Bayes classifier is hard while the data distribution is soft. Adding randomness to the classifier does not solve the issue. To justify this, and also for completeness, we considered the soft predictor Q and showed that it also fails to satisfy equal opportunity.

Proposition 7. There are data sources for which neither the Bayes classifier $Q_{1/2}$ nor the predictor Q satisfies equal opportunity.

Proof. Fix any data source with $\mathbb{P}[\,A{=}a, Y{=}1\,] > 0$ for each $a \in \{0,1\}$, pick an arbitrary ((X,A)-measurable) function $c : \mathbb{R}^d \to (0, 1/2)$ and let

$$q\,(x,a) \stackrel{\text{def}}{=} \begin{cases} 1/2 - c(x) & \text{if } a = 0 \\ 1/2 + c(x) & \text{if } a = 1 \end{cases}$$

for each $(x,a) \in \mathbb{R}^d \times \{0,1\}$.

Since we know that $Q_{1/2}(x,a) = a$, then the term $\mathbb{E}[\,Q_{1/2}(X,A) \mid A = a, Y = 1\,]$ can be reduced more simply into $\mathbb{E}[\,A \mid A = a, Y = 1\,] = a$. Therefore, the Bayes classifier satisfies $\text{oppDiff}(Q_{1/2}) = 1 - 0 > 0$.

Regarding Q, we have $\mathbb{E}[\,Q \mid A = 1, Y = 1\,] = 1/2 + \mathbb{E}[\,c(X) \mid A = 1, Y = 1\,]$ and $\mathbb{E}[\,Q \mid A = 0, Y = 1\,] = 1/2 + \mathbb{E}[\,c(X) \mid A = 0, Y = 1\,]$. Notice from the range of c, that $\mathbb{E}[\,Q \mid A = 1, Y = 1\,] \in (1/2, 1)$ and $\mathbb{E}[\,Q \mid A = 0, Y = 1\,] \in (0, 1/2)$. Hence $\text{oppDiff}(Q) > 0$.

Therefore, neither $Q_{1/2}$ nor Q satisfy equal opportunity.

□

As a remark, notice that the data sources proposed in the proof of Proposition 7, contrast the extreme case $Y = A$ because they allow some mutual information between X and Y after A is known, as one would expect in a real-life distribution. Nevertheless, there is an evident inherent demographic disparity in these distributions, and this can be the reason why equal opportunity hinders optimal accuracy for these examples.

7 Algorithms for the Pareto Frontier

In this section, we provide an algorithm for computing and depicting the Pareto frontier that optimizes the trade-off between error and the absolute value of opportunity difference (0 being EO). We consider (and aim at minimizing) the absolute value because we regard the difference in opportunity as bias, independently of the sign.

Three algorithms are explained and compared: the brute force, the one we propose, and the double-threshold method based on [13]. The methods are restricted to finite alphabets for the non-protected attributes, i.e. $\mathcal{X} = \{x_1, ..., x_n\}$, so the inputs (x, a) can only take a total of $|\mathcal{X} \times \{0, 1\}| = 2n$ values. For the convenience of the reader, we summarized them in Table 2.

Methodology	Complexity	Principle for finding the convex hull
Brute-force	$O(n\ 2^{2n})$	All corners correspond to deterministic classifiers.
Proposed	$O(n \log n)$	Algorithm 2. The n partial derivatives of error and opportunity difference are constant.
Double-threshold	$O(n^3 \log n)$	Algorithm 3. All corners correspond to single-threshold classifiers in V.

Table 2: Comparison of methods for finding the Pareto frontier and the feasibility region.

7.1 Brute-Force Algorithm

We begin by describing the brute-force algorithm for reference. The brute-force algorithm will compute not only the points that determine the Pareto frontier but all the vertices of the feasibility region M.

Recall that the set of all predictors forms a $2n$-dimensional polytope that is mapped into the region M when error and opportunity difference are measured. We know that each vertex of the region M corresponds to a deterministic classifier, or equivalently, to one of the 2^{2n} vertices of the polytope.

Therefore, it suffices to compute the error and opportunity difference for the 2^{2n} vertices of the polytope (first part), and then compute their convex hull (second part).

Assuming that each classifier is represented with an array of length $2n$, then the runtime complexity for computing the first part is $O(2n\ 2^{2n})$. For the second part, we may use Graham's scan algorithm [11] to find

the vertices of the convex hull. Since there are 2^{2n} points and Graham's scan has complexity $O(N \log N)$ where N is the number points, then the complexity is $O(2^{2n} \log 2^{2n}) = O(2n\ 2^{2n})$. Hence the complexity for the whole algorithm (adding up the first and second parts) is $O(4n\ 2^{2n}) = O(n\ 2^{2n})$.

7.2 Proposed Method

The proposed method (Algorithm 2) also computes all the vertices of the feasibility region M, but unlike the brute-force algorithm, it exploits greedily a property that appears to be local (depending on a chosen predictor), but in reality, is global (same for all predictors) in M.

For each predictor in the $2n$-dimensional polytope, let us consider its *taxicab neighbors*, i.e. the set of points that differ with it in at most one coordinate. Since the measurement function from the polytope into M is linear, these neighbors form a *star* in M around the given predictor (Figure 5).

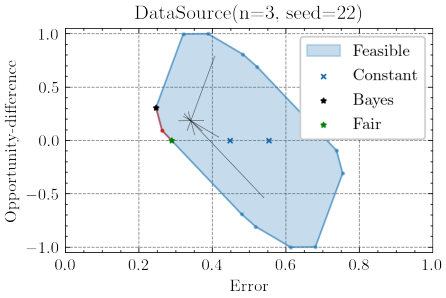


Figure 5: Feasibility region for a particular data source showing the Pareto frontier in red and the *taxicab neighbors*' star around an arbitrary predictor.

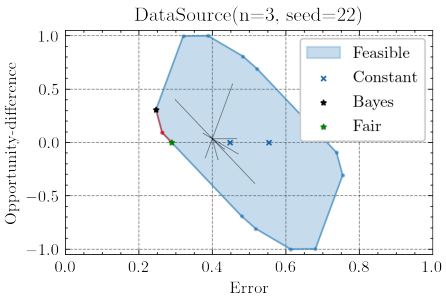


Figure 6: Same scenario as in Figure 5, but showing a star around a different arbitrary central predictor. The segments of the two stars differ exclusively in offset, not in slope or length. We exploit this fact in Algorithm 2.

Algorithm 2 Fast computation of the feasibility region vertices.

1: Letting $\alpha_a \stackrel{\text{def}}{=} \mathbb{P}[\,Y{=}1, A{=}a\,] > 0$ and $n \stackrel{\text{def}}{=} |\mathcal{X}|$,
2: **procedure** CONVEX HULL(α_0, α_1, Q)
3: $R \leftarrow [\,]$ ▷ Empty list of rays
4: **for** each (x, a) **do** ▷ $2n$ in total
5: sign $\leftarrow -1 + 2 \cdot \mathbf{1}_{a=1}$
6: $y \leftarrow 1$; $\theta \leftarrow \arctan2(\,1{-}2q(x,a),\ \text{sign}{\cdot}q(x,a)/\alpha_a\,)$
7: push tuple (θ, x, a, y) into R
8: $y \leftarrow 0$; $\theta \leftarrow (\,\theta + \pi \mod (-\pi, \pi]\,)$
9: push tuple (θ, x, a, y) into R
10: sort(R) ▷ by angle in $(-\pi, \pi]$
11: $V \leftarrow []$ (empty list of classifiers)
12: $\hat{Q} \leftarrow$ Bayes classifier $Q_{1/2}$
13: **for** each ray (θ, x, a, y) in R **do** ▷ $4n$ in total
14: update $\hat{q}(x, a) \leftarrow y$
15: push a copy of $\hat{Q}$ into V
16: **return** V ▷ classifiers that are vertices of M

The star consists of at most $4n$ rays ($2n$ segments crossing the middle) that represent the $2n$ degrees of freedom in the polytope. It reveals the possible combinations of error and opportunity difference that we can obtain from a given predictor by modifying a single component, i.e. the decision for a particular (x, a). In particular, when the central predictor is a vertex of the region M, then two of the rays of the star will land on the two neighboring vertices of the polygon.

The crucial fact exploited by Algorithm 2 is that the inclination and length of the segments of the star are the same regardless of the chosen central predictor. The only variation is the offset (compare Figures 5 and 6). As a consequence, the $2n$ segments that form the star can be visited in convenient order such that, starting from a vertex of the polygon M, all the visited predictors are vertices (or lie collinearly between two consecutive vertices) of the polygon.

More precisely, Algorithm 2 sorts the rays by angle, starts at the Bayes classifier, and then visits each ray, updating the current classifier according to the ray direction in the polytope. Each angle is computed in Line 6 using the gradients of error and opportunity difference as the x and y arguments respectively (derived from their definitions and Lemma 13). Both gradients were divided by a factor of $\pi(x, a)$ because the arctan2 function is indifferent to linear scales, and this allows the whole Algorithm to become independent of the distribution $\pi(\cdot, \cdot)$, except only for two population values, α_0 and α_1, defined as $\alpha_a \stackrel{\text{def}}{=} \mathbb{P}[\,Y{=}1, A{=}a\,]$.

The runtime complexity of Algorithm 2 is $O(n \log n)$ because of the sort instruction. All other instructions can be computed in linear time. Compared with the complexity of the brute-force algorithm, the proposed method enables the computation and visualization of the feasibility region M or the Pareto boundary for data sources with large (but finite)

n. Indeed, Figure 7 shows an example with $n = 1000$. Since the method computes all the vertices exactly, the visualization may be zoomed in at any level of detail.

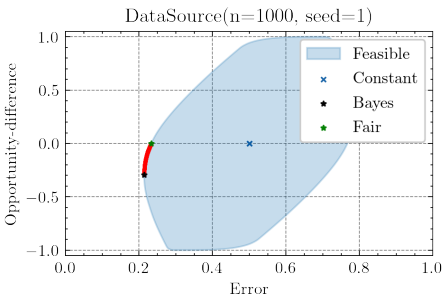


Figure 7: Pareto boundary (red) for a more elaborated example with $n = 1000$ in which the $O(n\ 2^{2n})$ brute-force algorithm is inconceivable. The feasibility region is guaranteed to be convex, and although its perimeter looks like a curve, it is a high-resolution piecewise linear path. Also, unlike Figures 5 and 6, the favored class is $a = 0$ and because of this, the Bayes classifier and the Pareto curve lie in the bottom half.

7.3 Double-Threshold Method

The following fact was shown by [13]. It allows parametrizing all the Pareto classifiers in a simple manner.

Fact 8. (Six parameters predictors) Any Pareto-optimal predictor $\hat{Q}$ can be written in terms of six parameters $l_0, l_1, r_0, r_1, p_0, p_1 \in [0, 1]$ ($l_a < r_a$, standing for left and right thresholds) as

$$\hat{q}(x, a) \stackrel{\text{def}}{=} \begin{cases} 0 & \text{if } q(x, a) \in [0, l_a) \\ p_a & \text{if } q(x, a) \in [l_a, r_a] \\ 1 & \text{if } q(x, a) \in (r_a, 1]. \end{cases}$$

This holds both discrete (as we assume) and non-discrete X.

Following Fact 8, a straightforward algorithm to approximate the Pareto-boundary consists of iterating over a large number of combinations of parameters, e.g. over a six-dimensional grid. This will produce a list of predictors of which we can filter only those that are Pareto optimal (optimal with respect to all other predictors in the list). The filtered predictors will form an approximation of the Pareto boundary.

As shown in Fact 9, if we concentrate on finding only the vertices of the Pareto-boundary and not all the points between them, the search space for the parameters can be reduced dramatically.

Fact 9. (Double threshold classifiers) For any vertex of the piecewise linear Pareto-boundary, there is a corresponding predictor $\hat{Q}$ (with that error and opportunity difference combination) that can be written in terms of two parameters $t_0, t_1 \in [0, 1]$ as either $\hat{q}(x, a) \stackrel{\text{def}}{=} \mathbf{1}_{q(x,a)>t_a}$, or $\hat{q}(x, a) \stackrel{\text{def}}{=}$

$\mathbf{1}_{q(x,a)\geq t_a}$, or a combination of the two, e.g.

$$\hat{q}(x,a) \stackrel{\text{def}}{=} \begin{cases} \mathbf{1}_{Q(x,0)>t_0} & \text{if } a=0 \\ \mathbf{1}_{Q(x,1)\geq t_1} & \text{if } a=1. \end{cases}$$

Proof. Let $l_0, l_1, r_0, r_1, p_0, p_1$ be the six parameters that define $\hat{Q}$ according to Fact 8. Since $\hat{Q}$ is a vertex on the Pareto-boundary it is also a vertex of the region M, and we know from Theorem 1 that the vertices of M correspond to deterministic predictors. Therefore, $\hat{Q}$ can only take values $0, 1$, which implies $p_0, p_1 \in \{0,1\}$. This restriction makes one of the two thresholds l_a or r_a irrelevant for each $a \in \{0,1\}$ and the predictor can be rewritten for each $a \in \{0,1\}$ as either $\hat{q}(x,a) = \mathbf{1}_{q(x,a)\geq l_a}$ or $\hat{q}(x,a) = \mathbf{1}_{q(x,a)>r_a}$. □

For our particular case of interest in which the variable for non-protected attributes X is discrete, $q(x,a)$ can only take a finite number of values $r_1, ..., r_m \in [0,1]$ with $r_i < r_{i+1}$. This makes the classifiers $\mathbf{1}_{q(x,a)>r_i}$ and $\mathbf{1}_{q(x,a)\geq r_{i+1}}$ equivalent. Therefore, we may unify all the possible cases of Fact 9 without loss of generality using only strict inequalities:

$$\hat{q}(x,a) \stackrel{\text{def}}{=} \mathbf{1}_{q(x,a)>t_a},$$

for two thresholds $t_0, t_1 \in \{q(x,a) \mid x \in \mathcal{X},\ a \in \{0,1\}\} \cup \{-1\}$. The special value -1 is added to contain the particular case $\mathbf{1}_{q(x,a)\geq 0}$ for which no strict threshold rule would exist. This is implemented in the 'candidates' procedure in Algorithm 3.

Algorithm 3 Computation of the feasibility region vertices

Letting $\alpha_a \stackrel{\text{def}}{=} \mathbb{P}[Y=1, A=a] > 0$ and $n \stackrel{\text{def}}{=} |\mathcal{X}|$,
procedure PARETO VERTICES(α_0, α_1, Q, P)
    $V \leftarrow$ candidates(Q)
    $W \leftarrow [\ (\text{err}(\hat{Q}), \text{oppDiff}(\hat{Q})) \mid \hat{Q} \in V\ ]$ ▷ needs Q, P
    $I \leftarrow$ indices of convex hull of W, sorted clockwise
    $i \leftarrow$ index in I with minimal x-coordinate ▷ V_i is Bayes
    $j \leftarrow$ first (or last) index in I with opposite y-sign to i
        ▷ (first or last depends on the y-sign of W_i)
    $I_{\text{Pareto}} \leftarrow$ indices in I between i and j
    return $[\ V_i \mid i \in I_{\text{Pareto}}\ ]$ ▷ Pareto vertices
procedure CANDIDATES(Q)
    $T_0 \leftarrow \{Q(x,0) \mid \text{for each } x\} \cup \{-1\}$ ▷ $|T_0| \leq n+1$
    $T_1 \leftarrow \{Q(x,1) \mid \text{for each } x\} \cup \{-1\}$ ▷ $|T_1| \leq n+1$
    $V \leftarrow [\,]$ ▷ Empty list of threshold classifiers
    for each $t_0 \in T_0$ **do**
        for each $t_1 \in T_1$ **do**
            push $\mathbf{1}_{q(x,a)>t_a}$ into V
    return V ▷ $|V| \leq (n+1)^2$

Algorithm 3, i.e. the 'Pareto vertices' procedure, computes the error and opportunity difference for each threshold classifiers of interest (each classi-

fier in V) and computes the convex hull to then filter the Pareto boundary. Since $|V| \leq (n+1)^2$, Algorithm 3 is polynomial. The exact complexity depends on the implementation of the computation $(\mathrm{err}(\hat{Q}), \mathrm{oppDiff}(\hat{Q}))$ for a fixed $\hat{Q} \in V$. Normally, this would take $O(n)$ by literally implementing their definition formulas for $|\mathcal{X}| = n$, hence the complexity of Algorithm 3 is $O(n^3 \log n)$.

8 Necessary and Sufficient Conditions

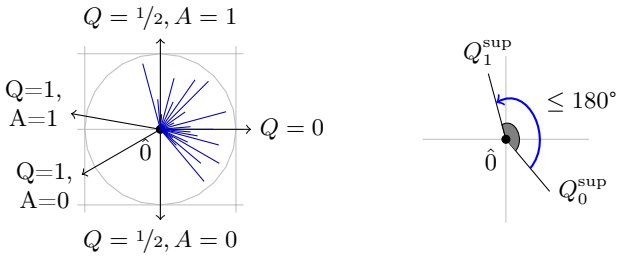
In this section, we provide a necessary and sufficient condition (Theorem 10), as well as a simple sufficient (but not necessary) condition (Corollary 11) that guarantees that equal opportunity and non-triviality are compatible. Finally, we discuss when and how a dataset may present this pathological incompatibility.

Theorem 10 (Necessary and sufficient condition for compatibility)**.** *Let (X, A, Y) be an arbitrary data source. Let $Q_a \stackrel{\text{def}}{=} \mathbb{E}[\, Q \mid A = a \,] = \mathbb{E}[\, Y \mid A = a \,]$ be the output average for each group. Let $Q_a^{\sup} \stackrel{\text{def}}{=} \sup\{q \in [0,1] \mid \exists S\ \mathbb{E}[\, Q \mid X \in S \wedge A = a \,] \geq q\}$ and analogously $Q_a^{\inf} \stackrel{\text{def}}{=} \inf\{q \in [0,1] \mid \exists S\ \mathbb{E}[\, Q \mid X \in S \wedge A = a \,] \leq q\}$.*

Then equal opportunity and non-triviality are compatible if and only if

$$
\begin{aligned}
0 \leq Q_1 Q_0^{\sup}(1 - 2Q_1^{\sup}) \quad &\leq Q_0 Q_1^{\sup}(2Q_0^{\sup} - 1) \quad \text{, or} \\
0 \leq Q_0 Q_1^{\sup}(1 - 2Q_0^{\sup}) \quad &\leq Q_1 Q_0^{\sup}(2Q_1^{\sup} - 1) \quad \text{, or} \\
0 \leq Q_1 Q_0^{\inf}(2Q_1^{\inf} - 1) \quad &\leq Q_0 Q_1^{\inf}(1 - 2Q_0^{\inf}) \quad \text{, or} \\
0 \leq Q_0 Q_1^{\inf}(2Q_0^{\inf} - 1) \quad &\leq Q_1 Q_0^{\inf}(1 - 2Q_1^{\inf}).
\end{aligned}
$$

Proof. Recall the star of rays around each classifier explained in Section 7.2, and consider the rays around the constant classifier $\hat{0}$ in the plane of error vs. opportunity difference. For each (x, a) in the domain, consider the predictor that maps everything to zero except (x, a) to one. The change in opportunity difference with respect to $\hat{0}$ is $\Delta y = \pi(x,a)\frac{q(x,a)}{Q_a}(2a-1)$, and the change in error is $\Delta x = \pi(x,a)(1-2q(x,a))$. Hence the angle of this ray is given by $\arctan_2(\Delta y, \Delta x) = \arctan_2(q(x,a)(2a-1),\, Q_a(1-2q(x,a)))$. In order to have an impossibility between EO and non-trivial accuracy, the constant classifier $\hat{0}$ must have either minimal error among the classifiers satisfying EO, or maximal error, in which case $\hat{1}$ is minimal. Geometrically, this means that $\hat{0}$ must be part of the convex hull, which holds if and only if all the angles of the rays departing from $\hat{0}$ lie in an interval of length at most $\pi = 180°$.



All the rays for $a = 1$ satisfy $\Delta y \geq 0$ and their angles lie between those of $Q_1^{\inf}$ counter-clockwise to $Q_1^{\sup}$. Similarly, all the rays for $a = 0$ satisfy $\Delta y \leq 0$ and their angles lie between those of $Q_0^{\inf}$ clockwise to $Q_0^{\sup}$. Therefore, checking that all rays lie in an interval of at most π is equivalent to checking that the counter-clockwise angle from $Q_0^{\sup}$ to $Q_1^{\sup}$ is at most π, or the clockwise angle from $Q_0^{\inf}$ to $Q_1^{\inf}$ is at most π. By replacing the values of Δy and Δx, and considering separately the cases $Q_0^{\sup} \leq 1/2$, $Q_1^{\sup} \leq 1/2$, $Q_0^{\inf} \geq 1/2$, and $Q_1^{\inf} \geq 1/2$, the four inequalities of the theorem statement are obtained. □

From Theorem 10 we can derive a simpler condition for EO and non-trivial accuracy to be compatible. It is only sufficient (i.e., not necessary), but it is easier to check and can be used to verify that a data source (X, A, Y) of a particular application is not pathological for equal opportunity. It is valid for discrete, continuous, and mixed data sources. Therefore, it may be used as a minimal assumption for any research work on equal opportunity dealing with probabilistic data sources.

Figure 8 summarizes the sufficiency condition in simple manner. The proof consists of showing that when the 4 events highlighted in Figure 8 have positive probabilities, then it is possible to use one of them to improve the performance of the best constant classifier and another one to compensate for equal opportunity.

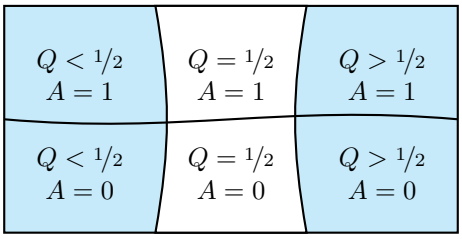


Figure 8: Sufficiency condition: If the 4 blue events have positive probability, then equal opportunity and non-triviality are compatible.

Corollary 11 (Sufficient condition)**.** *For any given data source* (X, A, Y)*, not-necessarily discrete, if for each* $a \in \{0, 1\}$*,*

$$\mathbb{P}[\, Q > 1/2, A = a \,], \mathbb{P}[\, Q < 1/2, A = a \,] > 0,$$

then equal opportunity and non-triviality are compatible. See Figure 8

Proof. According to Theorem 10, equal opportunity and non-triviality are compatible if and only if none of its four inequalities hold. If $\mathbb{P}[\,Q{>}1/2, A{=}a\,], \mathbb{P}[\,Q{<}1/2, A{=}a\,] > 0$ for each $a \in \{0,1\}$, then $Q_1^{\sup} > 1/2$, $Q_0^{\sup} > 1/2$, $Q_1^{\inf} < 1/2$ and $Q_1^{\inf} < 1/2$, which respectively violate the four inequalities of Theorem 10. Therefore compatibility is guaranteed. □

An alternative proof of Corollary 11 that does not use Theorem 10 can be found in the preliminary version of this paper [22] that was published in the proceedings of AAAI 2022.

Corollary 11 reveals an important property of the pathological distributions in which EO and non-triviality are incompatible, namely, that they must be already very biased in favor of either $A = 0$ or $A = 1$ and they are highly probabilistic, meaning that the decision Y depends largely on external information, e.g. noise. For instance, if $\mathbb{P}[\,Q{>}1/2, A{=}0\,] = 0$, then for all individuals in the class $A = 0$ the decision that minimizes error is $\hat{Y} = 0$, regardless of their value of X; and the only explanation for individuals with $A = 0$ and $Y = 1$ is external information not contained in X.

9 An example based on a real-life dataset

In this section, we show how the incompatibility may occur in practice with a variant of a real-life dataset. A consequence of Corollary 11 is that real world datasets should not incur an incompatibility between EO and non-trivial accuracy if sufficient information about the output is captured in the input features. However, the pathology may still arise when this property is violated. To illustrate this phenomenon, we consider a variant of the Adult dataset [7], where we eliminate some features (thus making it more probabilistic) and artificially reduce the rate of acceptance of the whole population to put the disadvantaged class in a more critical position.

Figure 9 shows the Adult dataset after applying the following process: (1) restricting the dataset to the 6 most relevant columns, (2) binarizing the columns using the mean as a threshold, and (3) randomly decreasing the probability of acceptance by 30% for both genders. The purpose of these operations was to illustrate the incompatibility, nevertheless, they are not so arbitrary. Indeed, the first two operations correspond to a simplification of the data, e.g. to perform a simple manual analysis, and the third was applied without direct use of the sensitive attribute (sex), meaning that no additional gender-specific bias was needed to derive the pathology. In other words, had the acceptance rate been lower for both classes, a simplification of the dataset into 6 binary columns would have sufficed to trigger the incompatibility.

More precisely, Figure 9 shows the feasibility region in the plane of error vs opportunity difference (the geometric perspective introduced in this paper) as well as the associated ROC curve for a classifier (the geometric perspective used in [13]). The left plot shows the constant classifier at the extreme left, on the convex hull of the feasibility region. The plot at the right is the ROC of a standard scikit-learn[20] random forest classifier of 100 decision trees, using a train-test split of 70%-30%. The parallel lines correspond to constant levels of accuracy and based on the slope and the direction of the gradient, it corroborates that accuracy is maximal at the left-bottom extreme point, which corresponds to $\hat{0}$ with 0 false positives and 0 true positives. The code for processing the dataset and generating the plots is available at [23].

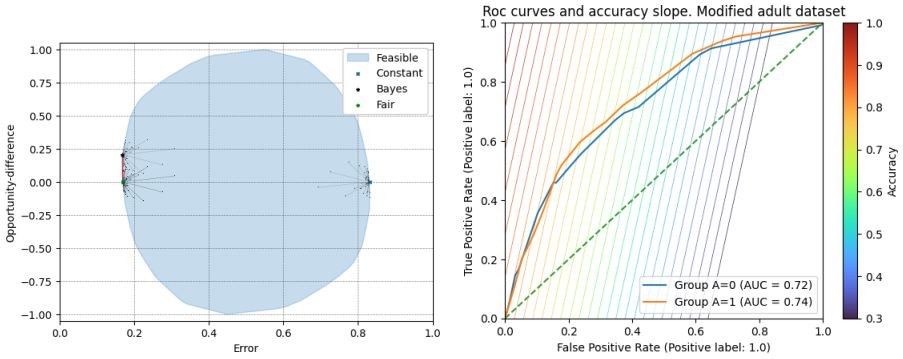


Figure 9: Adult dataset after simplification and reduction in acceptance rate. EO and non-trivial accuracy become incompatible.

10 Distortion effect of empirical distributions

In many situations, we do not have at our disposal a perfect description of the true data distribution, but only a dataset sampled from the distribution. This is the case, for instance, in machine learning, where the training and the testing are done on the basis of sets of samples. In this section, we discuss how using an empirical distribution from samples may distort the estimation and the evaluation of opportunity difference and accuracy. This distortion with respect to the true values is a consequence of the fact that an empirical distribution is only an approximation of the true one.

Figure 10 shows this mismatch from two points of view on an artificial dataset with $N = 100$ categories for X and $n = 1000$ samples. The dataset was generated by taking samples from a distribution consisting of a fixed categorical distribution for the $2N$ joint categories of X, A, and a binomial

distribution for $Y|X, A$ whose parameter depends on the conditioning pair X, A.

Figure 10 shows that when the empirical distribution of the dataset is used instead of the true distribution of the data source, the resulting (empirical) Pareto-optimal boundary obtained may mismatch the actual Pareto-optimal boundary, meaning that some classifiers that are empirically deemed as optimal are not optimal, and vice versa.

More precisely, in the left plot of Figure 10 the axes represent the measurements of the true error and opportunity difference, and the blue region shows the true convex hull. The orange line represents the empirical Pareto-optimal boundary, computed by applying the algorithms of Section 7 on the empirical distribution. As we can see, this boundary does not delimit a convex hull anymore, and it is at some distance from the true Pareto-optimal boundary. In particular, the empirical Fair (max accuracy subject to EO) and empirical Bayes predictors are not at the boundary of the true feasibility region, thus they are sub-optimal. Interestingly, the empirical Bayes classifier has less accuracy than the empirical Fair.

Conversely, the right plot of Figure 10 depicts the empirical apparent truth that a practitioner would observe in practice. Here, the axes are empirical (apparent) measurements of error and opportunity difference, and the orange area represents the empirical feasible region. The blue line represents the empirical evaluation of the true Pareto-optimal boundary. As we can see, in the empirical view the actual Bayes classifier and the fairest predictor appear to be sub-optimal.

Note that the classifiers that form the vertices of the orange convex hull in the right plot are exactly the orange points in the left plot and, vice versa, the classifiers that form the vertices of the blue convex hull in the left plot are exactly the blue points in the left plot.

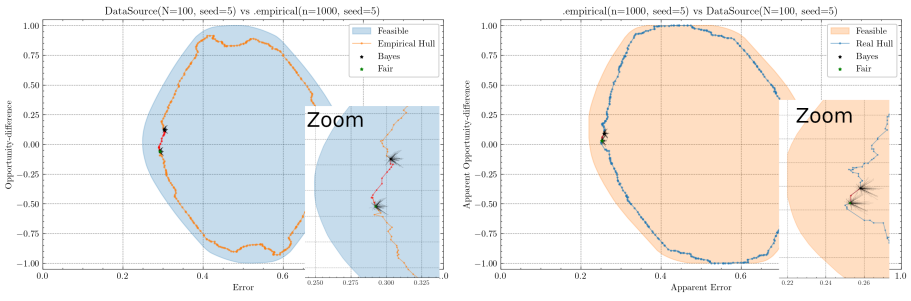


Figure 10: True distribution vs empirical dataset. Left: the empirical Pareto boundary mismatches the actual optimal boundary of error and opportunity difference. Right: using the dataset for measuring (apparent) error and opportunity difference makes the optimal predictors in the Pareto boundary to appear sub-optimal.

The unavailability of the true distribution, which causes a mismatch between the estimated error and opportunity difference and their true values, can have other unexpected consequences. For instance, the left plot in Figure 11 shows an example in which the best empirical fair classifier has less error and more opportunity difference than the empirical Bayes classifier. That is, for that particular data source and sampled dataset, training a model towards maximal accuracy results in more fairness than training taking fairness into account; conversely, training under the fairness constraint results in higher accuracy than training in an unconstrained manner towards maximal accuracy.

This distorting effect has a random nature from the sampling process and is reduced as the number of samples increases, making the empirical measurements closer to their real counterparts. The right plot in Figure 11 shows the result of computing the Pareto-optimal boundary on 100 different datasets sampled independently from the same data source distribution of $N = 100$ categories for X and $n = 2500$ samples. The plot shows that, on average, the positions of the empirical Bayes classifier and the empirical Fair classifier match the expected idea of the former having less error and more opportunity difference and vice versa. The empirical Fair classifier has indeed on average an opportunity difference close to zero, suggesting that even though there is no formal guarantee of achieving (true) equal opportunity using the Algorithms in this paper on (empirical) datasets, one does expect that, with an adequate number of samples, the empirical optimal classifiers will be close to the true optimal ones.

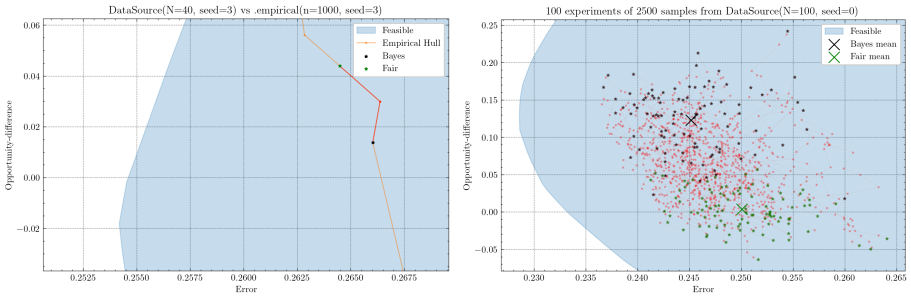


Figure 11: Left: a scenario in which the best empirical fair classifier has less error and more opportunity difference than the empirical Bayes classifier. Right: empirical Pareto boundaries for 100 randomly sampled datasets.

11 Conclusion

This work extends existing results about equal opportunity and accuracy from a deterministic data source to a probabilistic one. The main result, Theorem 3, states that for certain probabilistic data sources, no predictor

can achieve equal opportunity and non-trivial accuracy simultaneously. We also characterized in Theorem 10 the conditions on the data source under which EO and non-trivial accuracy are compatible and provided a simple sufficient condition that ensures compatibility (Corollary 11).

The methods used in this paper rely mostly on geometric properties of the feasibility region in the plane of error vs opportunity difference, thus they are tuned for the fairness notion of equal opportunity, which seeks equal true positive rates TPR. A symmetric analysis can be carried out for equal false positive rates using the same ideas. Since the notion of equal odds seeks both equal true positive rates and equal false positive rates, our methodology and results can be extended to equal odds. In particular, the impossibility theorem holds also for equal odds. However, the geometric methodology that we used was tuned for opportunity difference, they are therefore not directly useful for analyzing statistical parity or individual fairness notions.

12 Lemmas

Lemma 12. *For every* $\hat{Q} \in \mathcal{Q}$,

$$\mathit{err}(\hat{Q}) = \mathbb{E}[\,|\hat{Q} - Y|\,].$$

Proof. Notice that $\mathbb{P}[\,\hat{Y} \neq Y \,|\, Y = 1\,] = \mathbb{E}[1 - \hat{Q} \mid Y = 1\,]$ and $\mathbb{P}[\,\hat{Y} \neq Y \,|\, Y = 0\,] = \mathbb{E}[\,\hat{Q} \mid Y = 0\,]$. In both cases, we may write $\mathbb{P}[\,\hat{Y} \neq Y \,|\, Y = y\,] = \mathbb{E}[\,|Y - \hat{Q}| \mid Y = y\,]$.

Hence, marginalizing over Y we conclude $\mathbb{P}[\,\hat{Y} \neq Y\,] = \mathbb{E}[\,|Y - \hat{Q}|\,]$.

□

Lemma 13. *Assume* $\mathbb{P}[\,Y = 1, A = a\,] > 0$ *for each* $a \in \{0, 1\}$. *For any predictor* $\hat{Q}$, *we have*

$$\mathbb{P}[\,\hat{Y} = 1 \,|\, Y = 1, A = a\,] = \frac{\mathbb{E}[\,\hat{Q}Q \mid A = a\,]}{\mathbb{E}[\,Q \mid A = a\,]},$$

hence also

$$\mathit{oppDiff}(\hat{Q}) = \frac{\mathbb{E}[\,\hat{Q}Q \mid A = 1\,]}{\mathbb{E}[\,Q \mid A = 1\,]} - \frac{\mathbb{E}[\,\hat{Q}Q \mid A = 0\,]}{\mathbb{E}[\,Q \mid A = 0\,]}.$$

As an additional consequence, by considering the symmetric predictor $1 - \hat{Q}$, *it is also true that*

$$\mathbb{P}[\,\hat{Y} = 0 \,|\, Y = 1, A = a\,] = \frac{\mathbb{E}[\,(1 - \hat{Q})Q \mid A = a\,]}{\mathbb{E}[\,Q \mid A = a\,]}.$$

Proof. Indeed, by applying repetitively the Bayes rule, we get

$$\begin{aligned}\mathbb{P}[\,\hat{Y}{=}1\,|\,Y{=}1, A{=}a\,] &= \frac{\mathbb{P}[\,\hat{Y}{=}1, Y{=}1, A{=}a\,]}{\mathbb{P}[\,Y{=}1, A{=}a\,]} \\ &= \frac{\mathbb{P}[\,A{=}a\,]\mathbb{E}[\,\hat{Q}Q \mid A = a\,]}{\mathbb{P}[\,Y{=}1, A{=}a\,]} \\ &= \frac{\mathbb{E}[\,\hat{Q}Q \mid A = a\,]}{\mathbb{P}[\,Y{=}1\,|\,A{=}a\,]} \\ &= \frac{\mathbb{E}[\,\hat{Q}Q \mid A = a\,]}{\mathbb{E}[\,Q \mid A = a\,]}.\end{aligned}$$

The second equality holds because $(Y, \hat{Y}) \perp A \mid (Q, \hat{Q})$ and $Y \perp \hat{Y} \mid (Q, \hat{Q})$. □

Lemma 14. *(Vectorial metrics) Using the notation of Definition 1, we have*

$$\begin{aligned}\mathit{err}(\hat{Q}) &= \mathit{err}(\vec{F}), \\ \mathit{oppDiff}(\hat{Q}) &= \mathit{oppDiff}(\vec{F}).\end{aligned}$$

Proof. For the error, we marginalize over (X, A). Notice

$$\begin{aligned}\mathbb{P}[\,Y{\neq}\hat{Y}|\,X{=}x_i, A{=}a_i\,] =& + (1 - q(x_i, a_i))\hat{q}_(x_i, a_i) \\ &+ q(x_i, a_i)(1 - \hat{q}_(x_i, a_i)) \\ =&(1 - \vec{Q}_i)\frac{\vec{F}_i}{\vec{P}_i} + \vec{Q}_i\frac{\vec{P}_i - \vec{F}_i}{\vec{P}_i} \\ =&\frac{\vec{Q}_i\vec{P}_i + \vec{F}_i(1 - 2\vec{Q}_i)}{\vec{P}_i}.\end{aligned}$$

Thus $\mathbb{P}[\,Y{\neq}\hat{Y}, X{=}x_i, A{=}a_i\,] = \vec{F}_i\vec{P}_i + \vec{F}_i(1 - 2\vec{Q}_i)$, and

$$\begin{aligned}\mathrm{err}(f) &= \mathbb{P}[\,\hat{Y}{\neq}Y\,] \\ &= \sum_{i=1}^{n}\mathbb{P}[\,\hat{Y}{\neq}Y, X{=}x_i, A{=}a_i\,] \\ &= \langle\vec{P}, \vec{Q}\rangle + \langle\vec{F}, 1{-}2Q\rangle \\ &= \mathrm{err}(\vec{F}).\end{aligned}$$

For opportunity difference, we also marginalize over (X, A). Notice that

$$\begin{aligned}\mathbb{P}[\,\hat{Y}{=}1, Y{=}1, X{=}x_i, A{=}a_i\,] =&\vec{P}_i\,\mathbb{E}[\,\hat{Q}Q \mid X = x_i, A = a_i\,] \\ =&\vec{P}_i\frac{\vec{F}_i}{\vec{P}_i}\vec{Q}_i = \vec{F}_i\vec{Q}_i,\end{aligned}$$

hence $\mathbb{P}[\,\hat{Y}{=}1, Y{=}1, X{=}x_i, A{=}a\,] = \vec{F}_i\vec{Q}_i^{(a)}$. In addition, $\mathbb{P}[\,Y{=}1, X{=}x_i, A{=}a\,] =$

$\vec{P}_i\vec{Q}_i^{(a)}$ and

$$\begin{aligned}\mathbb{P}[\,\hat{Y}{=}1\,|\,Y{=}1, A{=}a\,] &= \frac{\sum_{i=1}^{n}\mathbb{P}[\,\hat{Y}{=}1, Y{=}1, X{=}x_i, A{=}a\,]}{\sum_{i=1}^{n}\mathbb{P}[\,Y{=}1, X{=}x_i, A{=}a\,]}\\ &= \frac{\langle\vec{F},\vec{Q}^{(a)}\rangle}{\langle\vec{P},\vec{Q}^{(a)}\rangle}.\end{aligned}$$

Therefore

$$\begin{aligned}\text{oppDiff}(\hat{Q}) &= +\,\mathbb{P}[\,\hat{Y}{=}1\,|\,Y{=}1, A{=}1\,]\\ &\quad -\,\mathbb{P}[\,\hat{Y}{=}1\,|\,Y{=}1, A{=}0\,]\\ &= \frac{\langle\vec{F},\vec{Q}^{(1)}\rangle}{\langle\vec{P},\vec{Q}^{(1)}\rangle} - \frac{\langle\vec{F},\vec{Q}^{(0)}\rangle}{\langle\vec{P},\vec{Q}^{(0)}\rangle}\\ &= \text{oppDiff}(\vec{F}).\end{aligned}$$

□

Lemma 15. *(Metrics symmetry) Using the notation of Definition 1, we have*

$$\begin{aligned}\mathit{err}(\vec{P}-\vec{F}) &= 1-\mathit{err}(\vec{F}),\\ \mathit{oppDiff}(\vec{P}-\vec{F}) &= -\mathit{oppDiff}(\vec{F}).\end{aligned}$$

Proof. According to Lemma 14, opportunity difference is a linear transformation. Since linear transformations preserve scalar multiplication and vector addition, it follows that $\text{oppDiff}(\vec{P}-\vec{F}) = \text{oppDiff}(\vec{P}) - \text{oppDiff}(\vec{F})$. Moreover, since $\text{oppDiff}(\vec{P}) = 1-1 = 0$, then $\text{oppDiff}(\vec{P}-\vec{F}) = -\text{oppDiff}(\vec{F})$.

According to the same lemma, the error is an affine transformation with offset $\langle\vec{P},\vec{Q}\rangle$. Hence

$$\begin{aligned}\text{err}(\vec{P}-\vec{F}) &= \text{err}(\vec{P}) - \text{err}(\vec{F}) + \langle\vec{P},\vec{Q}\rangle\\ &= 2\langle\vec{P},\vec{Q}\rangle - \langle\vec{P}, 1{-}2\vec{Q}\rangle - \text{err}(\vec{F})\\ &= \langle\vec{P}, 1\rangle - \text{err}(\vec{F})\\ &= 1 - \text{err}(\vec{F}).\end{aligned}$$

because $\sum_{i=1}^{n}\vec{P}_i = 1$. □

Lemma 16. *(Bayes accuracy)*

$$\mathit{acc}(Q_{1/2}) = 1/2 + \mathbb{E}[\,|Q-1/2|\,].$$

Proof. Out of Lemma 12, we know $\text{err}(Q_{1/2}) = 1-\mathbb{E}[\,\epsilon\,]$ where $\epsilon \stackrel{\text{def}}{=} |Q_{1/2}-Y|$. Let us condition on $Q < 1/2$ and $Q \geq 1/2$ separately (whenever these events have possible probabilities).

For $Q < 1/2$, we have $\mathbb{E}[\,\epsilon \mid Q < 1/2\,] = \mathbb{E}[\,Y \mid Q < 1/2\,] = \mathbb{E}[\,Q \mid Q < 1/2\,]$ and $Q = 1/2-(1/2-Q)$. For $Q \geq 1/2$, we have $\mathbb{E}[\,\epsilon \mid Q \geq 1/2\,] = \mathbb{E}[\,1-Y \mid Q \geq 1/2\,] = \mathbb{E}[\,1-Q \mid Q \geq 1/2\,]$ and $1-Q = 1/2-(Q-1/2)$.

These cases partition Ω and in both cases we have $\mathbb{E}[\,\epsilon\,] = 1/2 - \mathbb{E}[\,1/2 - Q\,]$. It follows that $\mathrm{err}(Q_{1/2}) = 1/2 - \mathbb{E}[\,|Q - 1/2|\,]$.

□

Lemma 17. *(Uniform case) Let $\hat{Q} \in \mathcal{Q}$ and $\epsilon \stackrel{\text{def}}{=} |\hat{Q} - Y|$ be the random variable of the error of $\hat{Q}$ (according to Lemma 12). If $\mathbb{P}[\,Q{=}1/2\,] > 0$, then*

$$\mathbb{E}[\,\epsilon \mid Q = 1/2\,] = 1/2.$$

Proof. Define $r \stackrel{\text{def}}{=} \mathbb{E}[\,\hat{Q}\,]$. Let us condition on $Y = 0$ and $Y = 1$ separately. For $Y = 0$, we have $\mathbb{E}[\,\epsilon \mid Q = 1/2, Y = 0\,] = r$, and for $Y = 1$, we have $\mathbb{E}[\,\epsilon \mid Q = 1/2, Y = 1\,] = 1 - r$.

Since $\mathbb{P}[\,Y{=}y | Q{=}1/2\,] = 1/2$, we can compute the marginal as

$$\mathbb{E}[\,\epsilon \mid Q = 1/2\,] = (1/2)(r{+}1{-}r) = 1/2$$

□

Lemma 18. *(Alternative Bayes) The* alternative Bayes classifier $Q_{1/2}$ *given by $\mathbf{1}_{q(x,a) \geq 1/2}$ ($\geq$ instead of $>$) has also maximal accuracy.*

Proof. We will prove that $\mathrm{err}(Q_{1/2}) = \mathrm{err}(Q^*_{1/2})$. Following Lemma 12, let $\epsilon \stackrel{\text{def}}{=} |Q_{1/2} - Y|$ and $\epsilon^* \stackrel{\text{def}}{=} |Q^*_{1/2} - Y|$.

Conditioned to $Q \neq 1/2$ we have $Q_{1/2} = Q^*_{1/2}$ from their definitions, and thus also $\mathbb{E}[\,\epsilon - \epsilon^* \mid Q \neq 1/2\,] = 0$. It suffices to check the complement event $Q = 1/2$. Suppose $\mathbb{P}[\,Q{=}1/2\,] > 0$. Conditioned to $Q = 1/2$, Lemma 17 implies that $\mathbb{E}[\,\epsilon - \epsilon^* \mid Q = 1/2\,] = 1/2 - 1/2 = 0$.

Hence $\mathbb{E}[\,\epsilon\,] = \mathbb{E}[\,\epsilon^*\,]$, i.e. $\mathrm{err}(q_{1/2}) = \mathrm{err}(q^*_{1/2})$.

□

Lemma 19. *(Trivial error as an expectation)*

$$\tau = 1/2 + |\mathbb{E}[\,Y\,] - 1/2|.$$

Proof. The constant 0 predictor ($\hat{0}$) has error $\mathbb{E}[\,Y\,]$, while the constant 1 predictor ($\hat{1}$) has error $1 - \mathbb{E}[\,Y\,]$. We can rewrite these quantities respectively as $1/2 - (1/2 - \mathbb{E}[\,Y\,])$ and $1/2 + (1/2 - \mathbb{E}[\,Y\,])$, whose maximum is $\tau = 1/2 + |1/2 - \mathbb{E}[\,Y\,]|$.

□

Lemma 20. *Let $\vec{P}, \vec{Q} \in (0,1)^3$, with $\vec{Q}_1 < 1/2$ and $\vec{Q}_2 - \vec{Q}_1 > 0$ (as in Theorem 3).*

If $\vec{P}$ and $\vec{Q}$ satisfy also $\vec{Q}_3 + \vec{Q}_1 \geq 1$ and $\vec{P}_1\vec{Q}_1 + \vec{P}_2\vec{Q}_2 < \vec{P}_3\vec{Q}_1$, then $\langle \vec{P}, 2\vec{Q}-1\rangle > \langle \vec{Z}, 2\vec{Q}-1\rangle$.

Proof. We have the following equivalences and at the end an implication.

$$\begin{aligned}
&\langle 2\vec{Q}-1, \vec{P}\rangle - \langle 2\vec{Q}-1, \vec{Z}\rangle > 0\\
\equiv&\langle 2\vec{Q}-1, \vec{P}-\vec{Z}\rangle > 0\\
\equiv&(2\vec{Q}_1-1)\vec{P}_1+(2\vec{Q}_3-1)\vec{P}_3\frac{\vec{P}_1\vec{Q}_1}{\vec{P}_1\vec{Q}_1+\vec{P}_2\vec{Q}_2}>0\\
\equiv&(2\vec{Q}_1-1)(\vec{P}_1\vec{Q}_1+\vec{P}_2\vec{Q}_2)+(2\vec{Q}_3-1)\vec{P}_3\vec{Q}_1>0\\
\equiv&(1-2\vec{Q}_1)(\vec{P}_1\vec{Q}_1+\vec{P}_2\vec{Q}_2)<(2\vec{Q}_3-1)\vec{P}_3\vec{Q}_1\\
\Leftarrow&(1-2\vec{Q}_1\leq 2\vec{Q}_3-1)\wedge(\vec{P}_1\vec{Q}_1+\vec{P}_2\vec{Q}_2<\vec{P}_3\vec{Q}_1)
\end{aligned}$$

It is given that $\vec{Q}_3 + \vec{Q}_1 \geq 1$ and $\vec{P}_1\vec{Q}_1 + \vec{P}_2\vec{Q}_2 < \vec{P}_3\vec{Q}_1$, which are equivalent to the last two inequalities. Thus, they imply that $\langle 2\vec{Q}-1, \vec{P}\rangle - \langle 2\vec{Q}-1, \vec{Z}\rangle > 0$.

□

Lemma 21. *(Complementary part of Theorem 3) Algorithm 1 is correct.*

Proof. We will prove $a < b$, $\vec{P}_2 \in (0, 1)$ and the fulfillment of constraints C2 and C5.

Part 1. Proof that $a < b$.

Recall $a = \max\{(1 - \vec{P}_3)\vec{Q}_1, 1/2 - \vec{P}_3\vec{Q}_3\}$ and $b = \min\{(1 - \vec{P}_3)\vec{Q}_2, \vec{P}_3\vec{Q}_1\}$.

1. Since $\vec{Q}_1 < 1/2 < \vec{Q}_2$ and $\vec{P}_3 \in (0, 1)$, then $(1 - \vec{P}_3)\vec{Q}_1 < (1 - \vec{P}_3)\vec{Q}_2$.
2. Since $\vec{P}_3 \in (1/2, 1)$, then $(1 - \vec{P}_3)\vec{Q}_1 < \vec{P}_3\vec{Q}_1$.
3. Since $\vec{P}_3 \in (0, 1)$ and $\vec{Q}_3 \in (1/2, 1)$, then $\vec{P}_3(\vec{Q}_2 - \vec{Q}_3) < 1 \cdot (\vec{Q}_2 - 1/2)$, or equivalently, $1/2 - \vec{P}_3\vec{Q}_3 < (1 - \vec{P}_3)\vec{Q}_2$.
4. Since $\vec{P}_3 > \frac{1}{2(\vec{Q}_1+\vec{Q}_3)}$ then $1/2 - \vec{P}_3\vec{Q}_3 < \vec{P}_3\vec{Q}_1$.

Since the inequalities hold for all available choices for a and b, then, in general, $a < b$ holds.

Part 2. Proof that $\vec{P}_2 \in (0, 1)$.
We know $c > \vec{Q}_1(1 - \vec{P}_3)$ and $c < \vec{Q}_2(1 - \vec{P}_3)$. These inequalities imply that $c - \vec{Q}_1(1 - \vec{P}_3) \in (0, \vec{Q}_2 - \vec{Q}_1)$, hence also that $\vec{P}_2 \in (0, 1)$.

Part 3. Constraint C2 is satisfied.

Since $\vec{P}_1 + \vec{P}_2 = 1 - \vec{P}_3$ and $\vec{Q}_2 > \vec{Q}_1$, then the term $\vec{P}_1\vec{Q}_1 + \vec{P}_2\vec{Q}_2$ is minimal when $\vec{P}_1 = 1 - \vec{P}_3$ and $\vec{P}_2 = 0$. Thus,

$$\begin{aligned}\vec{P}_1\vec{Q}_1 + \vec{P}_2\vec{Q}_2 + \vec{P}_3\vec{Q}_3 &\geq (1 - \vec{P}_3)\vec{Q}_1 + \vec{P}_3\vec{Q}_3 \\ &= \vec{Q}_1 + \vec{P}_3(\vec{Q}_3 - \vec{Q}_1) \\ &> \vec{Q}_1 + \frac{\vec{Q}_3 - \vec{Q}_1}{2} \\ &= \frac{\vec{Q}_3 + \vec{Q}_1}{2} \geq 1/2.\end{aligned}$$

Part 4. Constraint C5 is satisfied.
Since $b \leq \vec{P}_3\vec{Q}_1$, then $\vec{P}_2(\vec{Q}_2 - \vec{Q}_1) < \vec{P}_3\vec{Q}_1 - \vec{Q}_1(1 - \vec{P}_3)$. From this inequality, we may derive constraint C5 as follows.

$$\begin{aligned}\vec{P}_2(\vec{Q}_2 - \vec{Q}_1) &< \vec{P}_3\vec{Q}_1 - \vec{Q}_1(1 - \vec{P}_3) \\ \vec{P}_2\vec{Q}_2 &< (2\vec{P}_3 - 1 + \vec{P}_2)\vec{Q}_1 \\ \vec{P}_2\vec{Q}_2 &< \vec{P}_3\vec{Q}_1 - \vec{P}_1\vec{Q}_1 \\ \vec{P}_1\vec{Q}_1 + \vec{P}_2\vec{Q}_2 &< \vec{P}_3\vec{Q}_1.\end{aligned}$$

□

Declarations

- Funding: This work was supported by the European Research Council (ERC) project HYPATIA under the European Union's Horizon 2020 research and innovation programme. Grant agreement n. 835294.

- Conflicts of interest: Not applicable.
- Ethics approval: Not applicable.
- Consent to participate: Not applicable.
- Consent for publication: Not applicable.
- Availability of data and material: The document is self-contained.
- Code availability: We published a repository[23] containing Python implementation of Algorithms 1 and 2, as well as code for generating some of the figures presented.
- Authors' contributions: The 4 authors met regularly to discuss the correctness, presentation, and relevance of the results. Pinzón and Palamidessi proved the majority of the statements. Piantanida and Valencia reviewed the document repeatedly.

References

[1] Sushant Agarwal. Trade-offs between fairness, interpretability, and privacy in machine learning. Master's thesis, University of Waterloo, 2020.

[2] Simon Caton and Christian Haas. Fairness in machine learning: A survey. *arXiv preprint arXiv:2010.04053*, 2020.

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

[4] Sam Corbett-Davies and Sharad Goel. The measure and mismeasure of fairness: A critical review of fair machine learning. *ArXiv*, abs/1808.00023, 2018.

[5] Kate Crawford. The trouble with bias. In *Conference on Neural Information Processing Systems (NIPS)*, 2017.

[6] Rachel Cummings, Varun Gupta, Dhamma Kimpara, and Jamie Morgenstern. On the compatibility of privacy and fairness. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, pages 309–315, 2019.

[7] Dheeru Dua and Casey Graff. UCI machine learning repository, 2017.

[8] Keinosuke Fukunaga. Statistical pattern recognition. In *Handbook of pattern recognition and computer vision*, pages 33–60. World Scientific, 1993.

[9] Keinosuke Fukunaga. *Introduction to statistical pattern recognition*. Elsevier, San Diego, CA, US, 2013.

[10] Global Times. Beijing to release new license plate lottery policy. *https://www.globaltimes.cn/content/1190224.shtml*, 2018.

[11] Ronald L. Graham. An efficient algorithm for determining the convex hull of a finite planar set. *Info. Pro. Lett.*, 1:132–133, 1972.

[12] Branko Grünbaum. *Convex polytopes*, volume 221. Springer Science & Business Media, New York, NY, US, 2013.

[13] Moritz Hardt, Eric Price, and Nati Srebro. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29, 2016.

[14] Hoda Heidari, Michele Loi, Krishna P Gummadi, and Andreas Krause. A moral framework for understanding fair ml through economic models of equality of opportunity. In *Proceedings of the conference on fairness, accountability, and transparency*, pages 181–190, 2019.

[15] Michael Kearns and Aaron Roth. *The ethical algorithm: The science of socially aware algorithm design*. Oxford University Press, New York, NY, US, 2019.

[16] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

[17] Zachary C Lipton, Alexandra Chouldechova, and Julian McAuley. Does mitigating ml's impact disparity require treatment disparity? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 8136–8146, 2018.

[18] Karima Makhlouf, Sami Zhioua, and Catuscia Palamidessi. On the applicability of machine learning fairness notions. *ACM SIGKDD Explorations Newsletter*, 23(1):14–23, 2021.

[19] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[20] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[21] Dana Pessach and Erez Shmueli. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44, 2022.

[22] Carlos Pinzón, Catuscia Palamidessi, Pablo Piantanida, and Frank Valencia. On the impossibility of non-trivial accuracy in presence of fairness constraints. In *AAAI-22 Conference proceedings*. Association for the Advancement of Artificial Intelligence, AAAI, 2022.

[23] Carlos Pinzón. Github repository (impossibility-fairness-non-trivial-accuracy), 2022. `https://github.com/caph1993/impossibility-fairness-non-trivial-accuracy`, 2022 (accessed December 12 2022).

[24] David Pujol, Ryan McKenna, Satya Kuppam, Michael Hay, Ashwin Machanavajjhala, and Gerome Miklau. Fair decision making using privacy-protected data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pages 189–199, 2020.

[25] Sivan Sabato and Elad Yom-Tov. Bounding the fairness and accuracy of classifiers from population statistics. In *International Conference on Machine Learning*, pages 8316–8325. PMLR, 2020.

[26] State.gov. Diversity visa program. *http://dvprogram.state.gov/*, 2021.