On the Impossibility of Non-trivial Accuracy in Presence of Fairness Constraints

Carlos Pinzón,[1,3] **Catuscia Palamidessi**,[1,3] **Pablo Piantanida**[2,4] and **Frank Valencia**[2,3,5]
[1]Inria, Paris-Saclay, France
[2]CNRS, France
[3]Laboratoire d'informatique de l'École Polytechnique (LIX)
[4]Laboratoire des Signaux et Systèmes (L2S), Université Paris-Saclay, CentraleSupélec
[5]Pontificia Universidad Javeriana Cali, Colombia
carlos.pinzon@lix.polytechnique.fr

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. 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 (random guessing) classifier. In our 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 trade-off between accuracy and EO loss (opportunity difference), and provide a sufficient condition on the data source under which EO and non-trivial accuracy are compatible.

Introduction

During the last decade, the intersection between machine learning and social discrimination has gained considerable attention from the academia, the 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 (Pujol et al. 2020; Cummings et al. 2019; Kearns and Roth 2019; Agarwal 2020).

Fairness, has proven to be harder to conceptualize than privacy, for which differential privacy has become the defacto definition. Fairness is subjective and laws vary between countries. Even in academia, depending on the application, the words fairness and bias have different meanings (Crawford 2017). The current general consensus is that fairness can not be summarized into a unique universal definition; and for the most popular definitions, several trade-offs, implementation difficulties and impossibility theorems have been found (Kleinberg, Mullainathan, and Raghavan 2017; Chouldechova 2017). One such definition of fairness is equal-opportunity (Hardt, Price, and Srebro 2016).

To contrast equal-opportunity (EO) with accuracy, we borrow the notion of trivial accuracy from (Cummings et al. 2019). 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 (Cummings et al. 2019; Agarwal 2020), 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 (Cummings et al. 2019; Agarwal 2020) 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 (Hardt, Price, and Srebro 2016), 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. This principle is even applied in decisions that have significant social impact such as the Diversity Visa Program to qualify for a Green Card in the United States (State.gov 2021), and the Beijing lottery for getting a car license plate (Global Times 2018).

On the other hand, we consider probabilistic data sources. This is motivated by two different reasons:

1. It enables a more realistic and general representation of reality: one in which the information in the input may be insufficient to conclude definitely the yes-no decision, or in which real-life constraints force the decision to be different for identical inputs.
2. It provides a more general, yet simple, perspective for understanding the trade-off between fairness and accuracy. Also, it can take into account that in practice, input datasets are a noisy (thus probabilistic) approximation of reality.

Our contributions are the following.

1. We prove that for certain probabilistic distributions, no predictor can achieve EO and non-trivial accuracy simul-

taneously.

2. We provide a sufficient condition that guarantees compatibility between non-trivial accuracy and EO.
3. 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 sufficient and necessary conditions for non-trivially accurate predictors to exist.
4. We prove and depict several algebraic and geometric properties about the feasible region in the plane of opportunity-difference versus error.

Related Works

Our paper is strongly related to the following two works that consider a randomized learning algorithm guaranteeing (exact) equal-opportunity and also satisfying differential privacy: (Cummings et al. 2019) shows that, for certain distributions, these constraints imply trivial accuracy. (Agarwal 2020) proves the same claim for any arbitrary distribution and for non-exact equal-opportunity, i.e. bounded opportunity-difference. It also highlights that, although there appears to be an error in the proof of (Cummings et al. 2019), 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) equal-opportunity constraint, without any differential privacy assumption.

Another work about differential-privacy and fairness is (Pujol et al. 2020), but their notion of fairness is not equal-opportunity, and this allows them to provide examples in which privacy and fairness are not necessarily in a trade-off.

There are also several works that focus on the compatibility of fairness constraints. In (Kleinberg, Mullainathan, and Raghavan 2017), it is shown that several different fairness notions cannot hold simultaneously, except for exceptional cases. Similarly, in (Lipton, Chouldechova, and McAuley 2018), 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, (Corbett-Davies and Goel 2018) show issues inherent to anti-classification, classification parity and calibration separately. 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 on both classes.

Lastly, our geometric plots differ from those in the seminal paper of equal-opportunity (Hardt, Price, and Srebro 2016). Graphically, their analysis is carried out over ROC curves while we plot directly the two metrics of interest. In this sense, we provide a complementary geometric perspective for analyzing equal-opportunity and accuracy together.

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 sensible attributes or sensible features, are input features that represent race, gender, religion, nationality, age, or any other variable that could be used to discriminate against a group of people. A feature may be considered as 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 $q(x,a) \stackrel{\text{def}}{=} \mathbb{E}[\,Y \mid X=x, A=a\,]$ (conditional expectation, see the supplementary material for more details) and the random variable $Q \stackrel{\text{def}}{=} \mathbb{E}[\,Y \mid X, A\,] = q(X,A)$.

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.

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.

(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
$\text{acc}(\hat{Q})$	Accuracy of $\hat{Q}$: $\mathbb{P}[\hat{Y}=Y]$
$\text{oppDiff}(\hat{Q})$	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: Notation used in the paper.

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* (Hardt, Price, and Srebro 2016), we introduce a continuous metric called the *opportunity-difference*. The opportunity-difference of a predictor $\hat{Q} \in \mathcal{Q}$ is defined as

$$\text{oppDiff}(\hat{Q}) \stackrel{\text{def}}{=} (\,\mathbb{P}[\hat{Y}=1 \mid A=1, Y=1] - \mathbb{P}[\hat{Y}=1 \mid A=0, Y=1]\,)$$

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

$$\text{err}(\hat{Q}) \stackrel{\text{def}}{=} \mathbb{P}[\hat{Y} \neq Y]$$
$$\text{acc}(\hat{Q}) \stackrel{\text{def}}{=} 1 - \text{err}(\hat{Q})$$

Additionally, we consider a minimal reference level of accuracy that should be outperformed intuitively by any well-trained predictor. The *trivial accuracy* (Cummings et al. 2019) is defined as $\tau \stackrel{\text{def}}{=} \max\left\{\text{acc}(\hat{Q}) : \hat{Q} \in \text{Triv}\right\}$, where Triv is the 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.

[1] See for instance Chapter 3 of (Fukunaga 2013).

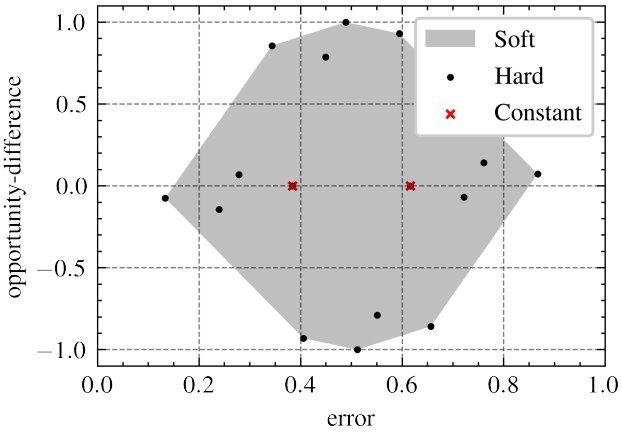

Figure 1: Region M for an arbitrary source distribution. The data source is not deterministic because the Bayes classifier does not satisfy equal-opportunity. The Python code for generating this figure is in the supplementary material.

The Error vs Opportunity-Difference Region

In this section, we remark several properties of 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).

M determines the tension between error and opportunity difference. Figure 1 shows an example of this region.

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 in the supplementary material. It uses the fact that affine transformations map polytopes into polytopes (See Chapter 3 of (Grünbaum 2013)). □

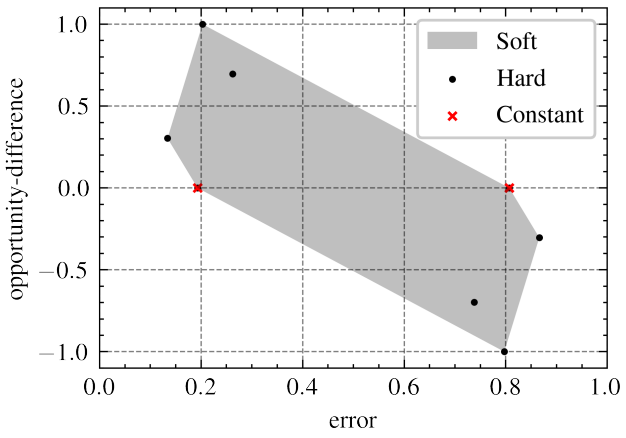

Figure 2: Randomly generated example using Algorithm 1 (Python code in supplementary material). The constant classifiers are vertices of the polygon, thus the constraints of equal-opportunity and non-trivial accuracy can not be satisfied simultaneously.

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

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. This figure illustrates the strong incompatibility that may occur (especially in highly probabilistic distributions). Namely, among the predictors satisfying equal-opportunity (those in the X-axis), the minimal error is achieved by a constant classifier.

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 $D \subseteq \mathbb{R}^k$ 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 $\mathbb{R}^n$.

[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]`.

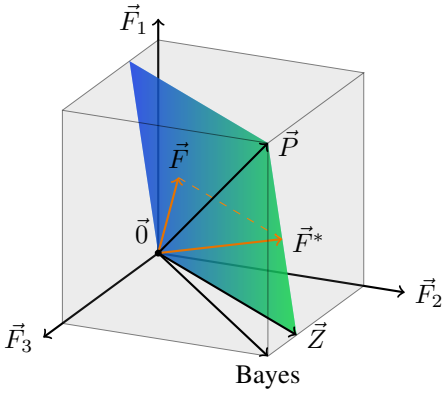

Figure 3: In vectorial form, the predictors that satisfy equal-opportunity form a plane inside the rectangular box of all predictors.

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 τ.*

Proof. The complete proof is contained in the supplementary material. Here we highlight only the sketch, the intuition and some relevant details.

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:

$$R_1 = \mathcal{X}_1 \times \{0\}$$
$$R_2 = \mathcal{X}_2 \times \{0\}$$
$$R_3 = \mathcal{X} \times \{1\}$$

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 predictor $\hat{Q} \in \mathcal{Q}$ satisfying EO. As shown in the supplementary material, the following constraints suffice:

C1. $\vec{P} \in (0,1)^3$ and, for probabilism, also $\vec{Q} \in (0,1)^3$.
C2. The accuracy of $\hat{Q} = \hat{1}$ is higher than that of $\hat{Q} = \hat{0}$ (for fixing an orientation).
C3. $\vec{Q}_1 < 1/2$ and $\vec{Q}_2, \vec{Q}_3 > 1/2$.
C4. $\vec{Q}_3 + \vec{Q}_1 \geq 1$, and
C5. $\vec{P}_1\vec{Q}_1 + \vec{P}_2\vec{Q}_2 < \vec{P}_3\vec{Q}_1$.

The last three constraints are not straightforward, but their main consequence can be explained graphically. Let us characterize each predictor $\hat{Q}$, with a vector $\vec{F}$ given by $\vec{F}_j \stackrel{\text{def}}{=} \mathbb{P}[\,\hat{Y}=1, (X,A) \in R_j\,]$, so that $\hat{Q} = \hat{1}$ corresponds to $\vec{F} = \vec{P}$. Figure 3 depicts the set of all predictors (box), and those that satisfy EO (plane). This plane can be characterized by the two vectors $\vec{Z}$ and $\vec{P}$.

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 $\vec{P}, \vec{Q}$

Constraint C3 simply fixes the location of the Bayes classifier at $(0, \vec{P}_2, \vec{P}_3)$. Constraints C4 and C5 force the gradient of the accuracy along the plane to be non-decreasing in the directions $\vec{Z}$ and $\vec{P} - \vec{Z}$, so that for each predictor $\vec{F}$ there is a pivot $\vec{F}^*$ with higher accuracy than $\vec{F}$ and lower accuracy than $\vec{P}$. As a consequence, when the constraints are satisfied, $\hat{1}$ has maximal accuracy in $\mathcal{Q}$.

In order to satisfy the constraints, we propose a randomized algorithm (Algorithm 1) that generates random vectors $\vec{P}$, $\vec{Q}$ satisfying the five constraints, regardless of the seed and the random sampling function, e.g. uniform. The proof is presented in the supplementary material. To corroborate, Figure 2 shows a particular output of the algorithm[3].

The proof concludes by showing that given $\vec{P}$ and $\vec{Q}$ (produced by the algorithm), it is always possible to split $\mathcal{X} \times \{0, 1\}$ into the three regions R_j (for $j \in \{1, 2, 3\}$) that satisfy $\mathbb{P}[\,(X, A) \in R_j\,] = \vec{P}_j$ and $\mathbb{P}[\,Y = 1 \mid (X, A) \in R_j\,] = \vec{Q}_j$. □

Finally, to conclude this section we present Example 1. It 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 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.

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

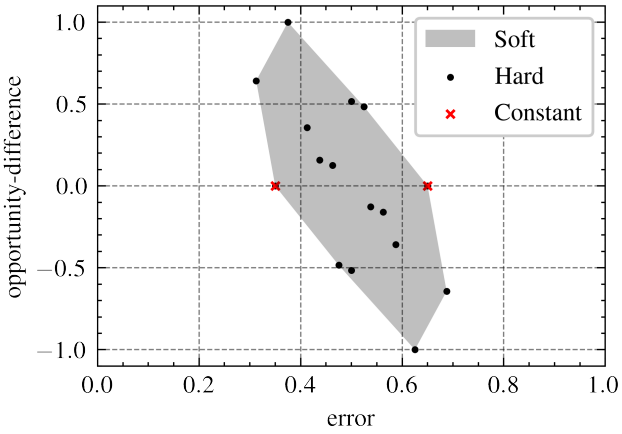

Figure 4: Example 1. One of the constant classifiers is Pareto-optimal.

Probabilistic vs 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.

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, if $\tau < 1$, then there is always a non-trivial predictor, for instance the Bayes classifier $Q_{1/2}$. Otherwise (degenerated case) all predictors are trivially accurate.

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

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 (Hardt, Price, and Srebro 2016) for defining equal-opportunity, because other notions of fairness, including statistical parity, are incompatible with accuracy.

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 secondary reference value $\tau^* \in [0,1]$. We let

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

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 is in the supplementary material. Intuitively, if $\mathbb{P}[\,Q \geq 1/2\,] = 1$, then predicting 1 for any input is optimal, and vice versa. □

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 satisfy 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 $\mathbb{E}[\,Q_{1/2}(X,A) \,|\, A{=}a, Y{=}1\,]$ can be reduced more simply into $\mathbb{E}[\,A \,|\, A{=}a, Y{=}1\,] = a$. Therefore, the Bayes classifier satisfies $\text{oppDiff}(Q_{1/2}) = 1 - 0 > 0$.

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

Sufficiency Condition

In this section, we provide a simple sufficient (but not necessary) condition (Theorem 9) that guarantees that equal-opportunity and non-triviality are compatible. It is not very restrictive and 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. It can also be used to verify whether a data source (X, A, Y) of a particular application is pathogenic for equal-opportunity or not.

Figure 5 summarizes the sufficiency condition in simple manner. The proof consists of showing that when the 4 events highlighted in Figure 5 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.

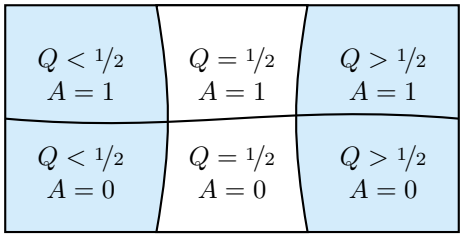

Figure 5: Sufficiency condition: If the 4 blue events have positive probability, then equal-opportunity and non-triviality are compatible.

Lemma 8. *Assume, for EO to be well defined, that* $\mathbb{P}[\,Y{=}1, A{=}a\,] > 0$ *for each* $a \in \{0,1\}$. *For any predictor* $\hat{Q}$, *we have*

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

Proof. Proved in the supplementary material. □

Theorem 9. *(Sufficiency condition, Figure 5)*

For any given data source (X, A, Y), *not-necessarily discrete, if*

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

for each $a \in \{0,1\}$, *then equal-opportunity and non-triviality are compatible.*

Proof. We begin by noticing that $\mathbb{P}[\,Q{>}1/2, A{=}a\,] > 0$ implies $\mathbb{P}[\,Y{=}1, A{=}a\,] > 0$ for each $a \in \{0,1\}$, thus equal-opportunity is well-defined.

The proof is divided into two cases depending on which constant classifier is optimal (either $\hat{0}$ or $\hat{1}$). The distinction is needed because equal-opportunity treats $Y = 1$ and $Y = 0$ differently.

Case 1. Assume $\text{err}(\hat{0}) \leq \text{err}(\hat{1})$.

We will show that there are constants $\hat{q}_0, \hat{q}_1 \in [0,1]$ such that the following predictor satisfies equal-opportunity and non-triviality.

$$\hat{Q} \stackrel{\text{def}}{=} \begin{cases} 0 & \text{if } Q \leq 1/2 \\ \hat{q}_0 & \text{if } Q > 1/2, A = 0 \\ \hat{q}_1 & \text{if } Q > 1/2, A = 1 \end{cases}$$

For $\hat{Q}$ to satisfy non-triviality, it suffices to have smaller error than $\hat{0}$. This holds whenever $\hat{q}_0$ or $\hat{q}_1$ are positive, because for each $a \in \{0,1\}$, the optimal classification over the region $\{Q > 1/2, A = a\}$ is 1, thus any positive value $\hat{q}_a$ improves the misclassification of $\hat{0}$. For this reason, we restrict $\hat{q}_0, \hat{q}_1 \in (0,1]$.

Regarding equal-opportunity, recall from Lemma 8 that

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

Let us call $\alpha_a \stackrel{\text{def}}{=} \mathbb{E}[\,\hat{Q}Q \,|\, A{=}a\,]$ to the numerator. Since $\hat{Q} = 0$ for $Q \leq 1/2$, then α_a may be computed as

$$\alpha_a = \hat{q}_a \mathbb{E}[\,Q \,|\, A{=}a, Q{>}1/2\,]\mathbb{P}[\,Q{>}1/2 \,|\, A{=}a\,]$$

and it is positive.

Hence, equal-opportunity may be stated as

$$\frac{\hat{q}_1}{\hat{q}_0} = \frac{\alpha_0 \mathbb{E}[\,Q \,|\, A{=}1\,]}{\alpha_1 \mathbb{E}[\,Q \,|\, A{=}0\,]}$$

The right hand side term is always well-defined, and it is a positive real number. Since $\hat{q}_0$ and $\hat{q}_1$ can be made arbitrarily small, there are always solutions to this equation in the range $\hat{q}_0, \hat{q}_1 \in (0,1]$.

Case 2. Assume $\text{err}(\hat{1}) < \text{err}(\hat{0})$.

Analogously, for $\hat{q}_0, \hat{q}_1 \in [0,1]$, let $\hat{Q}$ be given by

$$\hat{Q} \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } Q \geq 1/2 \\ \hat{q}_0 & \text{if } Q < 1/2, A = 0 \\ \hat{q}_1 & \text{if } Q < 1/2, A = 1 \end{cases}$$

If $\hat{q}_0, \hat{q}_1 < 1$, then $\hat{Q}$ satisfies non-triviality, because it has less error than $\hat{1}$. Hence, we restrict $\hat{q}_0 \hat{q}_1 < 1$.

Equal-opportunity, may be equivalently stated in terms of $\mathbb{P}[\,\hat{Y}{=}0 \,|\, Y{=}1, A{=}a\,]$ because it is the complement of $\mathbb{P}[\,\hat{Y}{=}1 \,|\, Y{=}1, A{=}a\,]$. Recall from Lemma 8 that

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

Let us call $\alpha_a \stackrel{\text{def}}{=} \mathbb{E}[\,(1{-}\hat{Q})Q \,|\, A{=}a\,]$ to the numerator. Since $(1 - \hat{Q}) = 0$ for $Q \geq 1/2$, then α_a may be computed as

$$\alpha_a = (1 - \hat{q}_a)\mathbb{E}[\,Q \,|\, A{=}a, Q{<}1/2\,]\mathbb{P}[\,Q{<}1/2 \,|\, A{=}a\,]$$

and it is non-negative.

Hence, equal-opportunity my be stated as

$$(1 - \hat{q}_1)\alpha_1 \mathbb{E}[\,Q \,|\, A{=}0\,] = (1 - \hat{q}_0)\alpha_0 \mathbb{E}[\,Q \,|\, A{=}1\,]$$

If $\alpha_1 \mathbb{E}[\,Q \,|\, A{=}0\,] = 0$, we let $\hat{q}_0 = 1$ and $\hat{q}_1 = 1/2$. If $\alpha_0 \mathbb{E}[\,Q \,|\, A{=}1\,] = 0$, we let $\hat{q}_1 = 1$ and $\hat{q}_0 = 1/2$. And if none of the two is zero, we use the same argument as in the first case: since $1 - \hat{q}_0$ and $1 - \hat{q}_1$ can be made arbitrarily small, there are always solutions to this equation in the range $\hat{q}_0, \hat{q}_1 \in [0,1)$. □

Conclusion and Future Work

Our 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 provided a sufficient condition on the data source under which EO and non-trivial accuracy are compatible.

Our method focuses on the fairness notion of equal-opportunity, which seeks for equal true positive rates. A symmetric analysis can be carried out for equal false positive rates using the same ideas. Since the notion of equal-odds seeks for both equal true positive rates and equal false positive rates, our methodology and results can be adapted to equal-odds. However, we believe that our results do not extend to statistical parity or to individual fairness notions.

An interesting question left for future work is whether the scenarios in which equal-opportunity and non-trivial accuracy are incompatible require the data source to be unfair on its own in some sense. If this is true, it would provide additional theoretical justification for equal-opportunity as a fairness notion. Nevertheless, any practical limitation of equal-opportunity that applies for the chosen application should always be prioritized, e.g. those shown for the COMPAS study case in (Corbett-Davies and Goel 2018).

Another line of research is the extension of our geometric and impossibility results to continuous distributions, possibly using existing theory from (Chzhen et al. 2019).

We also intend to characterize completely the conditions under which equal-opportunity and non-trivial accuracy are compatible.

Furthermore, we plan to study the trade-off and Pareto-optimality between accuracy and opportunity-difference as well as the accuracy gap between the Bayes classifier and the most accurate predictor that satisfies equal-opportunity.

Finally, we aim at bounding the opportunity-difference by taking into account the learning process and the statistical sampling.

References

Agarwal, S. 2020. *Trade-Offs between Fairness, Interpretability, and Privacy in Machine Learning*. Master's thesis, University of Waterloo.

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

Chzhen, E.; Denis, C.; Hebiri, M.; Oneto, L.; and Pontil, M. 2019. Leveraging labeled and unlabeled data for consistent fair binary classification. *Advances in Neural Information Processing Systems*, 32.

Corbett-Davies, S.; and Goel, S. 2018. The Measure and Mismeasure of Fairness: A Critical Review of Fair Machine Learning. *ArXiv*, abs/1808.00023.

Crawford, K. 2017. The trouble with bias. In *Conference on Neural Information Processing Systems (NIPS)*.

Cummings, R.; Gupta, V.; Kimpara, D.; and Morgenstern, J. 2019. On the compatibility of privacy and fairness. In *Adjunct Publication of the 27th Conference on User Modeling, Adaptation and Personalization*, 309–315.

Fukunaga, K. 2013. *Introduction to statistical pattern recognition*. San Diego, CA, US: Elsevier.

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

Grünbaum, B. 2013. *Convex polytopes*, volume 221. New York, NY, US: Springer Science & Business Media.

Hardt, M.; Price, E.; and Srebro, N. 2016. Equality of opportunity in supervised learning. *Advances in neural information processing systems*, 29.

Kearns, M.; and Roth, A. 2019. *The ethical algorithm: The science of socially aware algorithm design*. New York, NY, US: Oxford University Press.

Kleinberg, J.; Mullainathan, S.; and Raghavan, M. 2017. Inherent Trade-Offs in the Fair Determination of Risk Scores. In *8th Innovations in Theoretical Computer Science Conference (ITCS 2017)*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik.

Lipton, Z. C.; Chouldechova, A.; and McAuley, J. 2018. Does mitigating ML's impact disparity require treatment disparity? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, 8136–8146.

Pujol, D.; McKenna, R.; Kuppam, S.; Hay, M.; Machanavajjhala, A.; and Miklau, G. 2020. Fair decision making using privacy-protected data. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 189–199.

State.gov. 2021. Diversity Visa Program.
http://dvprogram.state.gov/.