Analyzing the Relationship Between Difference and Ratio-Based Fairness Metrics

Min-Hsuan Yeh
myeh@umass.edu
University of Massachusetts
Amherst, Massachusetts, USA

Blossom Metevier
bmetevier@umass.edu
University of Massachusetts
Amherst, Massachusetts, USA

Austin Hoag
austinthomashoag@gmail.com
Berkeley Existential Risk Initiative
Sacramento, California, USA

Philip S. Thomas
pthomas@cs.umass.edu
University of Massachusetts
Amherst, Massachusetts, USA

ABSTRACT

In research studying the fairness of machine learning algorithms and models, fairness often means that a metric is the same when computed for two different groups of people. For example, one might define fairness to mean that the false positive rate of a classifier is the same for people of different genders, ages, or races. However, it is usually not possible to make this metric identical for all groups. Instead, algorithms ensure that the metric is similar—for example, that the false positive rates are similar. Researchers usually measure this similarity or dissimilarity using either the *difference* or *ratio* between the metric values for different groups of people. Although these two approaches are known to be different, there has been little work analyzing their differences and respective benefits. In this paper we examine this relationship analytically and empirically, and conclude that unless there are application-specific reasons to prefer the difference approach, the ratio approach should be preferred.

CCS CONCEPTS

• **Computing methodologies** → **Machine learning**; • **General and reference** → *Metrics.*

KEYWORDS

Fair Machine Learning, Bias, Fairness Metrics, Classification

ACM Reference Format:
Min-Hsuan Yeh, Blossom Metevier, Austin Hoag, and Philip S. Thomas. 2024. Analyzing the Relationship Between Difference and Ratio-Based Fairness Metrics. In *The 2024 ACM Conference on Fairness, Accountability, and Transparency (FAccT '24), June 03–06, 2024, Rio de Janeiro, Brazil.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3630106.3658922

1 INTRODUCTION

In recent years, the application of *machine learning* (ML) models has become widespread across various domains, such as sentiment analysis [1], lie detection [24], and product recommendation [34], assisting individuals in complex decision-making tasks. However, the emergence of unfairness issues of ML models has raised significant concerns and ethical considerations. Existing literature highlights various risks associated with employing unfair ML models in real-life scenarios. For instance, Angwin et al. [3] examined an ML model used to predict whether a person will commit a violent crime in the future. Importantly, this model's predictions were considered by judges in eleven states during criminal sentencing. Their study revealed that, conditioned on individuals who did not commit a violent crime, the model was twice as likely to incorrectly predict that Black individuals would commit such a crime. To address such instances of unfairness, many ML researchers have shifted their focus from solely pursuing high performance to also ensuring fairness in prediction and detection.

A significant amount of research within the fair ML community focuses on the development of fair classification models (e.g., see the over 45 works surveyed by Mehrabi et al. [22]). These works predominantly address *group fairness*, aiming for similar outcomes across different demographic groups. This is often achieved by adhering to the fairness criteria outlined by Dwork et al. [14]. According to this criteria, an ML model is fair if it yields equal values for specific statistical metrics (e.g., accuracy, false positive rate, false negative rate) conditioned on *sensitive attributes* such as gender, race, or age. Therefore, many group fairness approaches calculate values of a particular metric for each group, and measure the degree of unfairness between these values.

For example, the ProPublica study [3] examined the *false positive rate* (FPR) across races, focusing on the probability of falsely predicting individuals, particularly from Black and White racial groups, to be at a high risk of criminal recidivism. In this example, let $\text{FPR}_{\text{Black}}$ and $\text{FPR}_{\text{White}}$ represent the average false positive rates for Black and White individuals, respectively. Ideally, these two values should be equal, which would indicate no racial bias in the model's predictions, i.e., that Black and White individuals have the same probability of being incorrectly labeled as high-risk for recidivism. In practice, these values may be close for models that are fair, but are almost never exactly the same. This leads to our primary research question: *When these values are different, how should the degree of unfairness be measured?*

FAccT '24, June 03–06, 2024, Rio de Janeiro, Brazil
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0450-5/24/06
https://doi.org/10.1145/3630106.3658922

There are two common ways of measuring the degree of unfairness between groups. First, unfairness can be quantified by measuring the absolute difference between the values for two groups ($|\text{FPR}_{\text{Black}} - \text{FPR}_{\text{White}}|$ in the above example), which is referred to as the *difference* method. An alternate approach for quantifying unfairness involves measuring the minimum ratio between the values ($\min\{\text{FPR}_{\text{Black}}/\text{FPR}_{\text{White}}, \text{FPR}_{\text{White}}/\text{FPR}_{\text{Black}}\}$ in the same example), known as the *ratio* approach. While it is clear that these two approaches are distinct, there has been a lack of comprehensive analysis in the literature regarding their differences and respective advantages. Furthermore, although choosing between these two methods can be application-dependent, papers proposing fair ML models often fail to provide explicit explanations for choosing one method over the other. In a discussion with five sets of authors of FAccT '21 and '22 papers that used one of these methods in training, we found that all of them adopted the difference method in their research simply because it had been employed in previous works, rather than due to its theoretical or empirical advantages.

Importantly, our study reveals that due to the fundamental differences between the difference and ratio approaches, satisfying a fairness constraint under one does *not* guarantee satisfaction under the other. This underscores a significant and potentially severe risk of not carefully selecting the appropriate approach to quantify fairness when evaluating or developing fair ML models. To illustrate this point, consider the Propublica study by Angwin et al. [3] discussed earlier. If a model has false positive rates of $\text{FPR}_{\text{Black}} = 0.02$ and $\text{FPR}_{\text{White}} = 0.01$, then under the difference method, the absolute difference between $\text{FPR}_{\text{Black}}$ and $\text{FPR}_{\text{White}}$ is 0.01. In the context of criminal recidivism prediction, this small difference (0.01) might suggest the model is fair. However, from the ratio-based perspective, these values are problematic, yielding a ratio of 0.5. This implies Black individuals are twice as likely to be falsely labeled high-risk compared to White individuals, highlighting a significant fairness concern.

To address this knowledge gap, we examine the differences between the ratio and difference approaches. We derive the theoretical relationship between the difference and ratio, then use a Monte Carlo method to show their empirical relationship over two datasets (LIAR and COMPAS) considering three common fairness definitions: predictive equality, equal opportunity, and overall accuracy equality. Additionally, we train models using loss functions that integrate both constraining approaches to show how the two metrics relate for optimized models (we call this relationship an "optimized relationship"). The results suggest that, in general, training with the ratio approach can prevent the optimized model from giving a misleading outcome. In addition, both methods can achieve the same optimized fairness value when choosing appropriate measures. The results also indicate that when reporting measures of fairness, using only the difference may be insufficient because each difference value can nap to a large range of the corresponding ratio values. However, constraining the ratio instead does induce a meaningful constraint on the difference as well. As a result, we encourage future research to adopt the ratio approach. Or, even better, to show the optimized relationship between the difference and ratio values of the proposed model.

2 NOTATION AND PROBLEM STATEMENT

In this paper, we study the relationship between two approaches for measuring fairness of binary classification models. Consider a dataset $\mathcal{D} = \{\mathbf{x}^{(i)}, y^{(i)}\}_{i=1}^{N}$ with N data points, where $\mathbf{x}^{(i)} \in \mathcal{X} \subset \mathbb{R}^K$ is an input vector and $y^{(i)} \in \mathcal{Y} = \{-1, +1\}$ is the corresponding label for the input $\mathbf{x}^{(i)}$. A common approach in group fairness is to divide the dataset $\mathcal{D}$ based on group affiliations associated with features. Accordingly, we separate $\mathcal{D}$ into two groups, $\mathcal{D}_0 = \{\mathbf{x}, y | \mathbf{x}_G = 0\}$ and $\mathcal{D}_1 = \{\mathbf{x}, y | \mathbf{x}_G = 1\}$, according to a binary sensitive attribute G, and evaluate the fairness of a model $f_\theta : \mathcal{X} \to \mathcal{Y}$ with respect to G.

We focus on statistical definitions of fairness, which use statistical metrics as measures of fairness [30]. The selection of the statistical metric depends on the specific fairness definition. For instance, *false positive rate* (FPR) is used to ensure predictive equality [11], while *false negative rate* (FNR) is used for equal opportunity [16].

For each statistical metric M, we define $M_i \in [0, 1]$ to be the metric value for people of type i, computed from data $\mathcal{D}_i$. For example, M_0 could correspond to the false positive rate of model f_θ on data $\mathcal{D}_0$, the accuracy of model f_θ on data $\mathcal{D}_0$, or any other metric of interest computed on $\mathcal{D}_0$. Notice that the same analysis holds if M_i is the metric value for the data generating distribution (conditioned on the sensitive attribute G), not just the empirical metric value for data $\mathcal{D}_i$.

The approaches we study are the difference and the ratio approach, which are two commonly used methods for measuring fairness. The difference approach evaluates fairness by measuring the absolute difference between M_1 and M_0. The fairness value measured by this approach is

$$\epsilon_d := |M_1 - M_0|. \tag{1}$$

The ratio approach, on the other hand, evaluates fairness by calculating the minimum ratio between M_1 and M_0. The fairness value measured by this approach is

$$\epsilon_r := \begin{cases} 0 & \text{if } M_1 = M_0 = 0, \\ 1 - \min\left\{\frac{M_1}{M_0}, \frac{M_0}{M_1}\right\} & \text{otherwise.} \end{cases} \tag{2}$$

Notice that instead of reporting the minimum ratio, we report *one minus* the minimum ratio. This ensures that fairness corresponds to lower values of both ϵ_r and ϵ_d.

On the dataset $\mathcal{D}$, the values of M_1 and M_0 vary with the model, i.e., they vary with the parameters θ. For each M_1 and M_0, we calculate the ϵ_d and ϵ_r values to form a (ϵ_d, ϵ_r) pair. By plotting all the (ϵ_d, ϵ_r) pairs on a 2D plane, we can observe the size and the shape of the region formed by the set of all possible (ϵ_d, ϵ_r) pairs, which represent the relationship between the outcome of the difference and the ratio approaches. In this paper, we investigate the implications of the size and the shape of the space of (ϵ_d, ϵ_r) pairs, by first considering the following questions:

(1) What is the size and shape of the set of all possible (ϵ_d, ϵ_r) pairs on a 2D plane?
(2) What makes the size and the shape of regions different in different settings?
(3) Can the relationship help us to select one approach over the other when training models and reporting results?

3 RELATED WORK

The majority of fair ML research addresses either the technical aspects of ML bias, or theories concerning its regulatory, societal, and moral implications. This work is concerned with technical approaches, which aim to promote fairness at various stages of the ML pipeline: pre-processing (data gathering and preparation), in-processing (selecting and training a model), or post-processing (adjusting model outputs).

Fairness definitions in ML can broadly be categorized into individual and group fairness. *Individual fairness*, first introduced by Dwork et al. [15], centers on the principle of treating similar individuals similarly, with the goal of ensuring fair treatment on an individual basis. On the other hand, *group (or statistical) fairness* focuses on achieving fair outcomes across different demographic groups, defined by sensitive attributes such as race, gender, or age [30]. It involves developing methods to detect and counteract biases in ML system decisions (and outcomes) to ensure more balanced treatment for these groups.

As stated previously, we consider statistical definitions of group fairness, which have been widely used to ensure fairness of classification models [30]. Our theoretical results apply to any methods that consider statistical definitions of fairness as defined in 2, including methods that mitigate bias at the stage of pre-, in-, and post-processing, and methods that go beyond these three categories, *e.g.*, intra-processing learning [25] and neural architecture search [13]. Our empirical evaluation focuses on in-processing methods, where a model is trained on prepared data, and its performance is optimized by directly modifying the learning process to produce fair outcomes. This is often achieved by integrating fairness definitions via constraints during optimization, e.g., [2, 27, 31].

To ensure such statistical definitions of group fairness, while a few works propose constrained optimization methods that directly treat a fairness definition as an equality constraint (to our knowledge, only the work of Baumann et al. [7]), most studies consider a *relaxation* of a fairness definition, i.e., introducing a positive amount of slack, $\tau > 0$, and forming an inequality constraint [4]. For example, Donini et al. [12] formulated equal opportunity [10] as the difference condition

$$|\mathrm{FNR}_a - \mathrm{FNR}_b| \leq \tau, \tag{3}$$

where FNR means the false negative rate; and Zafar et al. [33] formulated demographic parity [9] as the ratio condition

$$\min\{\mathrm{PR}_a/\mathrm{PR}_b, \mathrm{PR}_b/\mathrm{PR}_a\} \geq 1 - \tau, \tag{4}$$

where PR stands for positive rate. This ratio condition is called the disparate impact [5]. Beyond works that adhere to specific predefined fairness definitions, there more general algorithms that can satisfy multiple or arbitrary fairness definitions have also been proposed. For example, Thomas et al. [29] propose Seldonian algorithms, a class of algorithms that can constrain a model using a wide range of fairness metrics, including both the ratio or difference approach. While these algorithms offer the flexibility to choose from various fairness measures, there is little work studying the distinct advantages of each approach.

Studies introducing fair ML algorithms typically focus their evaluation on performance metrics, like accuracy, and the specific fairness approach used during the algorithm's training (difference or ratio) [2]. However, as we show in Section 4, a model that has a small difference (or ratio) value only means it can be considered fair under the difference (or ratio) condition. It can remain unfair if we measure the fairness in the opposite form (e.g., when two metric values are 0.01 and 0.02). Although there are studies focused on making the assessment of fairness more reliable, most of them targeted the issue of uncertainty and variation [6, 20, 21], or discussed the outcome from the aspect of social sciences [17, 19]. In this paper, we investigate the relationship between the difference and ratio-based relaxation in both the training stage and the assessment to improve the reliability of the fairness result.

Lastly, our work reaffirms the difference and ratio approach as two unique measurement scales of model fairness. As categorized by the work of Stevens [28], data measurement scales include nominal, ordinal, interval, and ratio types, each with its own analytical implications. Unlike ordinal scales, which only prioritize the order of values, the exact numerical differences and ratios of fairness metrics are crucial. Again, consider the example in Section 1, where the difference method resulted in a small numerical difference in the FPR rates between racial groups, and the ratio approach revealed a significant bias. This example highlights the need to carefully select the appropriate measurement scale (difference or ratio) in fairness evaluations, a point more explicitly addressed in the rest of this work.

4 ϵ_d–ϵ_r RELATIONSHIP

In order to analyze the disparities between the two training approaches, we begin by examining the relationship between the fairness measurements, ϵ_d and ϵ_r. Initially, we establish the theoretical relationship by deducing it from the formulae that define ϵ_d and ϵ_r. Additionally, we ascertain the empirical relationship by sampling various (ϵ_d, ϵ_r) pairs given a specific dataset and fairness definition. Moreover, we propose the optimized relationship, which we define as the relationship between ϵ_d and ϵ_r for models that were trained to optimize a trade-off between accuracy and one or both of the fairness measures (this differs from the values of ϵ_d and ϵ_r that occur for arbitrary models that are not designed to ensure fairness or accuracy). We demonstrate the optimized relationship by training models with two distinct loss functions, each incorporating the ϵ_d and ϵ_r terms, respectively.

4.1 Theoretical Relationship

Recall that ϵ_d is the difference in metric values and ϵ_r is the ratio of metric values for a model, and M_1 and M_0 are the metric values for two groups where M_1 and $M_0 \in [0, 1]$.

Theorem 1 (ϵ_d–ϵ_r relationship with unknown M_1 and M_0). *If there are two sensitive groups and the metric values of both groups are in* $[0, 1]$*, then it is always the case that*

$$\epsilon_r \in [\epsilon_d, 1]. \tag{5}$$

Proof. Since M_1 and $M_0 \in [0, 1]$, and $\epsilon_d = |M_1 - M_0|$, we know

$$\epsilon_d \leq \max\{M_1, M_0\} \leq 1 \Rightarrow 1 \geq \frac{\epsilon_d}{\max\{M_1, M_0\}} \geq \epsilon_d. \tag{6}$$

Then by Lemma 1, we have $1 \geq \epsilon_r \geq \epsilon_d$ and $\epsilon_r \in [\epsilon_d, 1]$. □

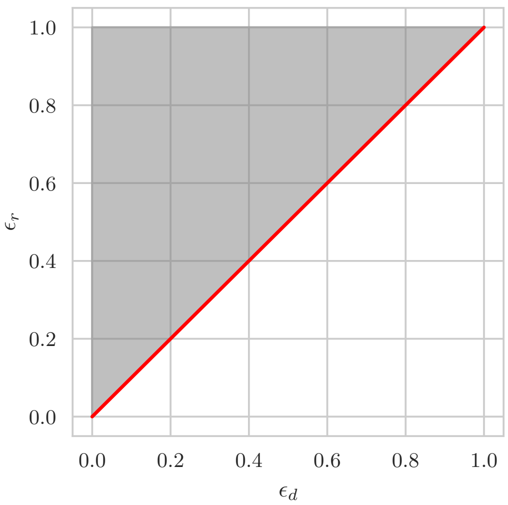


Figure 1: Theoretical region of the (ϵ_d, ϵ_r) pairs. The (ϵ_d, ϵ_r) pairs can only fall in the shaded region.

Lemma 1 (ϵ_d–ϵ_r relationship given M_1 and M_0). *Given M_1 and M_0,*

$$\epsilon_r = \frac{\epsilon_d}{\max\{M_1, M_0\}}. \tag{7}$$

Proof. Rewrite the equation for ϵ_r in terms of ϵ_d as

$$\epsilon_r = 1 - \min\left\{\frac{M_1}{M_0}, \frac{M_0}{M_1}\right\} = \frac{\max\{M_1, M_0\} - \min\{M_1, M_0\}}{\max\{M_1, M_0\}}$$
$$= \frac{|M_1, M_0|}{\max\{M_1, M_0\}} = \frac{\epsilon_d}{\max\{M_1, M_0\}}. \tag{8}$$

□

According to Theorem 1, the (ϵ_d, ϵ_r) pairs of a model f_θ must fall within the gray region shown in Figure 1. This region symmetrically demonstrates that for a certain ϵ_r, the possible range of the corresponding ϵ_d is $[0, \epsilon_r]$, i.e., the ratio approach upper-bounds the difference approach. This suggests that a classifier that is fair w.r.t. a tolerance T under the difference approach can be unfair w.r.t. any tolerance T' under the ratio approach. However, a classifier that is fair w.r.t. a tolerance T under the ratio approach will be fair w.r.t. any $T' \leq T$ under the difference approach. Additionally, Lemma 1 shows that the ratio approach can be viewed as a normalized difference approach, and the normalized term is $\max\{M_1, M_0\}$. This suggest that when considering fairness metrics that their value will decrease during optimization process (such as FPR or FNR), optimizing on the ϵ_r value would force the corresponding ϵ_d value to be smaller than directly optimizing on the ϵ_d value.

4.2 Empirical Relationship

The theoretical ϵ_d–ϵ_r relationship assumes M_0 and M_1 to be in $[0, 1]$. However, what if, in reality, M_0 and M_1 are only in a subset of $[0, 1]$? How will the feasible region of (ϵ_d, ϵ_r) pairs change? What factors affect the size (or shape) of the feasible region? Will the size or shape of the feasible region affect the decision to choose ratio or difference as the constraining approach? To answer these questions and understand more attributes of the ϵ_d–ϵ_r relationship, we apply a Monte Carlo method to sample logistic regression models from a parameter space, measure their ϵ_d and ϵ_r values under certain datasets and fairness definitions, and visualize the empirical ϵ_d–ϵ_r relationship for each setting in Figure 2.

4.2.1 Experiment Setting. We demonstrate the empirical relationship on two datasets:

(1) COMPAS [3]: COMPAS is a risk prediction dataset, which was widely used in many previous work (e.g., Mishler et al. [23], Sikdar et al. [26], Wang et al. [31]). Each datum of COMPAS describes a person, including their personal information, their criminal record, and a label indicating whether they committed crimes or violent crimes after 2 years. We take the race as the sensitive attribute and group data into two groups (African-American or Caucasian).
(2) LIAR [32]: LIAR is a text-based lie detection dataset. Each datum consists of a statement, information about the speaker, and a label indicating whether the statement was a lie. We take the U.S. political party affiliation as the sensitive attribute and group data into two groups (Democrat and Republican). We encode the statement as textual features, serving as an example to demonstrate the relationship between difference- and ratio-based approaches on natural language processing models.

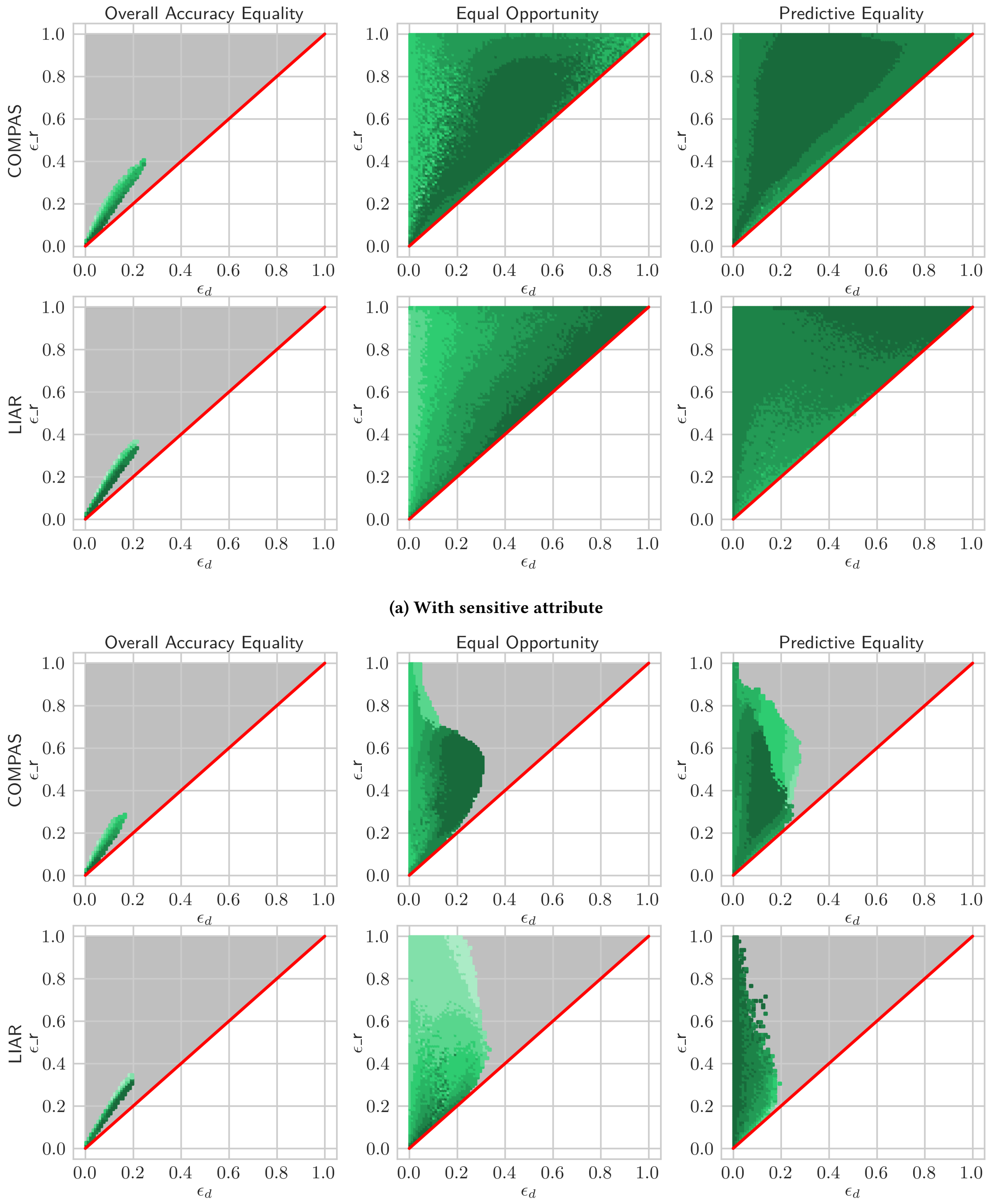
We choose three commonly used fairness definitions to examine the relationship between ϵ_d and ϵ_r:

(1) Predictive equality [11]: Requires that the FPR of a model is equal for each group.
(2) Equal opportunity [16]: Requires that the FNR of a model is equal for each group.
(3) Overall accuracy equality [8]: Requires that the accuracy of a model is equal for each group.

In addition, we design two kinds of input: 1) with and 2) without sensitive attributes, indicating whether the input feature includes the sensitive attributes (party affiliation for LIAR dataset and race for COMPAS dataset). For the setting of without sensitive attributes, we also remove the proxy features during data pre-processing.

We apply a Monte Carlo method for each setting. We randomly draw parameters of a logistic regression model from Uniform$(-10, 10)$ 1,500,000 times. We then measure their metric values conditioned on the sensitive attribute and plot the (ϵ_d, ϵ_r) pairs over the theoretical region. We shade the empirical area with different color values according to its accuracy. A region with higher (lower) accuracy is shaded with a darker (lighter) green.

4.2.2 Result. Figure 2 shows two factors that largely affect the shape and size of the feasible region of (ϵ_d, ϵ_r) pairs. The first one is the choice of fairness definitions. The empirical region of overall accuracy equality is approaching a straight line in all settings, while the regions of predictive equality and equal opportunity cover a large area of the theoretical region. This suggests that the statistics metric within the fairness definitions affects the empirical relationship between ϵ_d and ϵ_r. By analyzing the distribution of the metric value in the setting where the input features did not include sensitive attributes (see Figure 3), we found that the accuracy values of models in both datasets are in a range between 0.3 and



(a) With sensitive attribute

(b) Without sensitive attribute

Figure 2: The empirical region of the (ϵ_d, ϵ_r) pairs. We plot the (ϵ_d, ϵ_r) pairs of models with 1,500,000 distinct sample parameters. Each grid is assigned a green color if it contains (ϵ_d, ϵ_r) pairs. Grids with high-accuracy (low-accuracy) models are shaded with a darker (lighter) green. Figure 2a shows the result of models with the sensitive attribute included as a feature, and Figure 2b shows the result for models do not include the sensitive attribute. We present the outcomes of models on the COMPAS and LIAR datasets, considering three fairness definitions: overall accuracy equality, equal opportunity, and predictive equality.

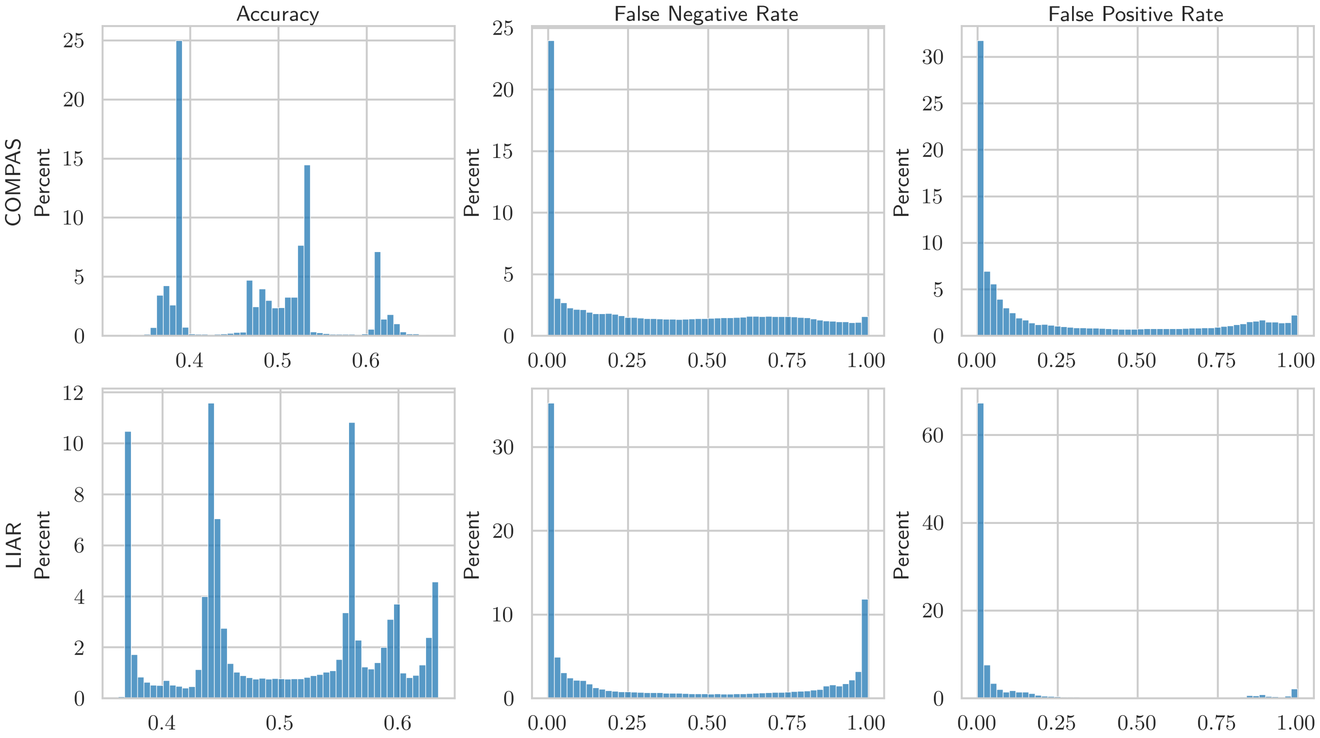


Figure 3: The distribution of metric value in different settings. We plot the distribution of metric values derived from 1,500,000 logistic regression models, each with unique sample parameters. The x-axis indicates the metric value and the y-axis represents the percentage of models for each metric value. Noted that the sensitive attributes were exclude from the input.

0.7. However, false negative and false positive rates range in $[0, 1]$. This phenomenon suggests that the feasible region of (ϵ_d, ϵ_r) pairs depends on the empirical range of metric value. In order to further study the ϵ_d–ϵ_r relationship of the settings where the metric values are bounded in a small range, we proposed Theorem 2.

Theorem 2 (ϵ_d–ϵ_r relationship with bounded M_1 and M_0). *If there are two sensitive groups and the metric values of both groups are in* $[a, b]$ *where* $0 \leq a \leq b \leq 1$, *then it is always the case that*

$$\epsilon_r \in \left[\frac{\epsilon_d}{b}, \frac{\epsilon_d}{a+\epsilon_d}\right]. \tag{9}$$

Proof. Let both $M_1 \in [a, b]$ and $M_0 \in [a, b]$, $0 \leq a \leq b \leq 1$. In this case, the value of ϵ_d is confined in $[0, b-a]$. Without loss of generality, we assume $M_1 \leq M_0$, then the smallest pair of (M_1, M_0) is $(a, a+\epsilon_d)$, and the largest pair is $(b-\epsilon_d, b)$. Thus, the value of ϵ_r corresponding to each ϵ_d would be

$$\epsilon_r \in \left[1-\frac{b-\epsilon_d}{b}, 1-\frac{a}{a+\epsilon_d}\right] \Rightarrow \epsilon_r \in \left[\frac{\epsilon_d}{b}, \frac{\epsilon_d}{a+\epsilon_d}\right]. \tag{10}$$

□

According to Theorem 2, once the range of the metric value is small, and the lower bound of the range is far from 0, the range of ϵ_r will be small for all ϵ_d. Figure 4 illustrates different ϵ_d–ϵ_r regions conditioned on the range of the metric values. The result

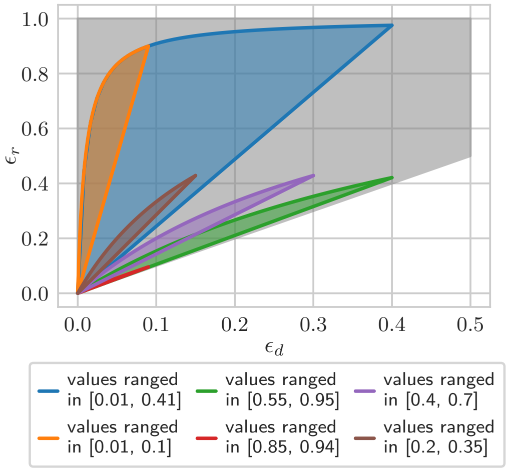


Figure 4: The $\epsilon_d - \epsilon_r$ region conditioned on different ranges of the metric value. The theoretical region is shaded with gray (here, we clip the region on $\epsilon_d = 0.5$), and the $\epsilon_d - \epsilon_r$ regions corresponding to different ranges of values are shaded with different colors.

shows that if the lower bound of the range is close to 0, e.g., 0.01, the range of ϵ_r corresponding to each ϵ_d will be large (see the blue and orange regions). On the other hand, if the lower bound of the metric value is far from 0, the range of ϵ_r corresponding to each ϵ_d will be largely decreased even if the range is large (see the green region). Moreover, once the range of metric becomes small, the range of ϵ_r will be close to 0 (see the red region), i.e., the ϵ_d–ϵ_r relationship is near linear. In this case, we proposed Theorem 3 to estimate the linear relationship between ϵ_d and ϵ_r.

Theorem 3. *Given ϵ_d and an error tolerance value $\tau > 0$, if there are two sensitive groups and the metric values of both groups are in $[a, b]$ where $1 - \frac{a}{b} \leq 2\tau$, we can estimate $\hat{\epsilon_r}$ as*

$$\hat{\epsilon_r} := \frac{\epsilon_d(b + a + \epsilon_d)}{2b(a + \epsilon_d)} \tag{11}$$

with a maximum error $|\hat{\epsilon_r} - \epsilon_{r*}| \leq \tau$ where ϵ_{r*} is the true value ϵ_r.

Proof. Let both $M_1 \in [a, b]$ and $M_0 \in [a, b]$, $0 \leq a \leq b \leq 1$. According to Theorem 2, given a ϵ_d, ϵ_r is ranged in $\left[\frac{\epsilon_d}{b}, \frac{\epsilon_d}{a+\epsilon_d}\right]$. When estimating $\hat{\epsilon_r}$ by the midpoint of $\frac{\epsilon_d}{b}$ and $\frac{\epsilon_d}{a+\epsilon_d}$, i.e.,

$$\hat{\epsilon_r} := \left(\frac{\epsilon_d}{a + \epsilon_d} + \frac{\epsilon_d}{b}\right)/2 = \frac{\epsilon_d(b + a + \epsilon_d)}{2b(a + \epsilon_d)}, \tag{12}$$

the estimated error would be $|\epsilon_{r*} - \frac{\epsilon_d(b+a+\epsilon_d)}{2b(a+\epsilon_d)}|$. In this case, the maximum error occurs when $\epsilon_{r*} \in \left\{\frac{\epsilon_d}{b}, \frac{\epsilon_d}{a+\epsilon_d}\right\}$ and $\frac{d}{d\epsilon_d}|\epsilon_{r*} - \frac{\epsilon_d(b+a+\epsilon_d)}{2b(a+\epsilon_d)}| = 0$.

Without loss of generality, assuming $\epsilon_{r*} = \frac{\epsilon_d}{a+\epsilon_d}$. Then, when the maximum error occurred, we have

$$\frac{d}{d\epsilon_d}\left(\frac{\epsilon_d}{a + \epsilon_d} - \frac{\epsilon_d(b + a + \epsilon_d)}{2b(a + \epsilon_d)}\right) = 0 \tag{13}$$

$$\Rightarrow \frac{d}{d\epsilon_d} \frac{\epsilon_d(b - a - \epsilon_d)}{2b(a + \epsilon_d)} = 0 \tag{14}$$

$$\Rightarrow \frac{ab - (\epsilon_d + a)^2}{2b(\epsilon_d + a)^2} = 0 \tag{15}$$

$$\Rightarrow \epsilon_d = -a \pm \sqrt{ab} \tag{16}$$

$$\Rightarrow \epsilon_d = -a + \sqrt{ab}.\ (-a - \sqrt{ab} \text{ is invalid since } \epsilon_d \geq 0) \tag{17}$$

Substituting (17) into (14), the maximum error between $\hat{\epsilon_r}$ and ϵ_{r*} would be

$$\frac{(-a + \sqrt{ab})(b - a - (-a + \sqrt{ab}))}{2b(a + (-a + \sqrt{ab}))} = \frac{b\sqrt{ab} - a\sqrt{ab}}{2b\sqrt{ab}} = \left(1 - \frac{a}{b}\right)/2. \tag{18}$$

Therefore, when $M_1 \in [a, b]$, $M_0 \in [a, b]$, and $1 - \frac{a}{b} \leq 2\tau$, the maximum error of estimating $\hat{\epsilon_r}$ is

$$\left(1 - \frac{a}{b}\right)/2 \leq \frac{2\tau}{2} = \tau. \tag{19}$$

□

According to Theorem 3, once the range of the metric value $[a, b]$ is small enough (i.e., $1 - \frac{a}{b} \leq 2\tau$ with a small τ), we can find an approximate transformation function g to map between ϵ_d and ϵ_r where

$$\epsilon_r \approx g(\epsilon_d) := \frac{\epsilon_d(b + a + \epsilon_d)}{2b(a + \epsilon_d)}. \tag{20}$$

In this case, there is no need to worry about the difference between using the difference or ratio approach, and we can choose the one that makes more sense in the applied scenario.

The second factor that affects the shape and size of the feasible region of (ϵ_d, ϵ_r) pairs is the input features. Figure 2 shows that excluding the sensitive attributes from the input features can largely restrict the region of the (ϵ_d, ϵ_r) pairs when fairness is defined as equal opportunity and predictive equality. However, it only ensures a small ϵ_d, and the possible value of ϵ_r still ranges from 0 to 1. To examine the reason for this phenomenon, we plot the (M_1, M_0) pairs of models in different settings (see Figure 5). For clear demonstration, we pair M_1 and M_0 in ascending order, i.e., the x-axis indicates the value of $\min\{M_1, M_0\}$ and the y-axis indicates the value of $\max\{M_1, M_0\}$. The result shows that when fairness is defined as equal opportunity and predictive equality, the (M_1, M_0) pairs of models that took sensitive attributes as input cover the whole upper triangle, while the (M_1, M_0) pairs of models that excluded sensitive attributes cover only a small portion of that triangle.

An intuitive reason for this discrepancy is that when sensitive attributes are included in the input, a model can exhibit complete unfairness by making decisions based solely on these attributes, making the FPR or FNR be one for a group and zero for another. For other cases where the model input does not include sensitive attributes, the above situation is harder to achieve. This suggests that for many cases, the value of the absolute difference between M_1 and M_0 is naturally constrained. This phenomenon points out a concern about using the difference approach as a fairness constraint because a constraint requiring ϵ_d to be small may be satisfied naturally in some settings, especially if the sensitive attributes do not exist in the input features. However, those models' ϵ_r value may be high because typical models (not trained to be fair with respect to the ratio) almost never satisfy a reasonable constraint on ϵ_r. In such cases, constraining on ϵ_d will be less meaningful or even misleading.

Furthermore, this result highlights an issue in reporting fairness. Studies usually reported fairness in the same form as the training constraint [2]. For example, if a model is constrained with the difference approach, reporting fairness means showing the ϵ_d value of the testing set. However, the result in Figure 2b shows that in some settings (e.g., defining fairness as equal opportunity), reporting only the ϵ_d value may over-simplify the result because each ϵ_d value can map to a large range of ϵ_r values. Instead, when using the difference method, we recommend reporting both the ϵ_d and ϵ_r values to provide a more holistic view of the (un-)fairness of the model.

4.3 Optimized Relationship

When visualizing the empirical relationship, we consider the entire parameter space instead of the optimal one. If we only consider models that trade-off accuracy and fairness, we obtain a further subset of the possible region. Here we are going to visualize this region of optimal logistic regression models for lie detection and risk prediction.

We consider two strategies for training a fair model. First, we consider approximately Pareto optimal models, i.e., we train the

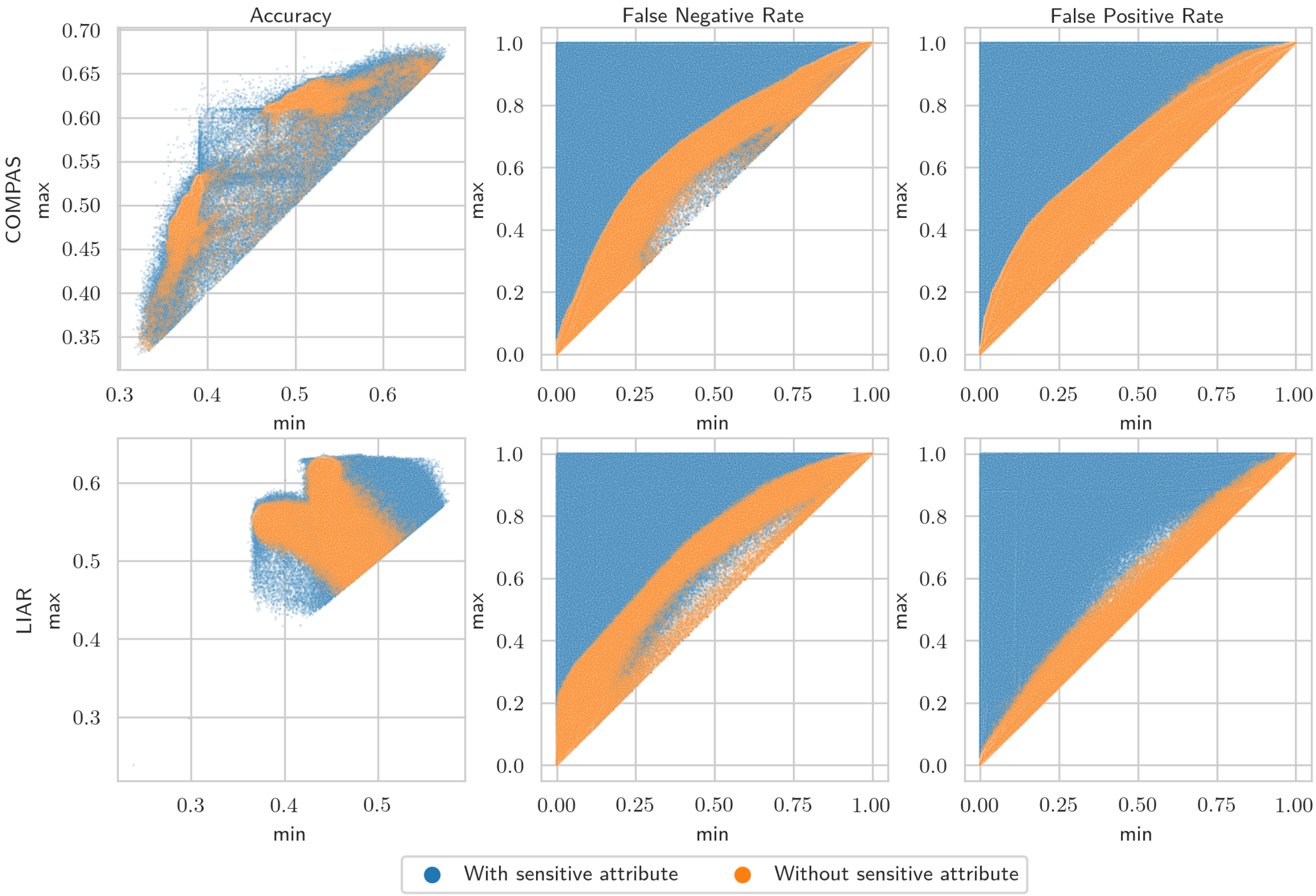


Figure 5: The (M_1, M_0) pairs of each setting. We pair M_1 and M_0 in an ascending order, *i.e.*, x-axis indicates the value of $\min\{M_1, M_0\}$ and y-axis indicates the value of $\max\{M_1, M_0\}$. The performance of model *with* sensitive attributes in the input are presented by blue dots, while the performance of model *without* sensitive attributes in the input are denoted by orange dots.

models with two different loss functions, $\mathcal{L}_{\epsilon_d}(\theta)$ and $\mathcal{L}_{\epsilon_r}(\theta)$, where

$$\mathcal{L}_{\epsilon_d}(\theta) = (1-\lambda)\mathcal{L}_{\text{BCE}}(\theta) + \lambda|P_M(f_\theta, \mathcal{D}_0) - P_M(f_\theta, \mathcal{D}_1)|, \quad (21)$$

and

$$\mathcal{L}_{\epsilon_r}(\theta) = (1-\lambda)\mathcal{L}_{\text{BCE}}(\theta) + \lambda\left(1 - \min\left\{\frac{P_M(f_\theta, \mathcal{D}_0)}{P_M(f_\theta, \mathcal{D}_1)}, \frac{P_M(f_\theta, \mathcal{D}_1)}{P_M(f_\theta, \mathcal{D}_0)}\right\}\right). \quad (22)$$

Here $\mathcal{L}_{\text{BCE}}(\theta)$ stands for the loss function of standard binary cross entropy, and P_M is a function for computing M_i given the model f_θ and data $\mathcal{D}_i$. We train 100 models with varying regularization terms $\lambda \in \{0.01, 0.02, ..., 1\}$. For each λ, we train 100 models initialized with different random seeds to reduce the effect of randomness.

For the second strategy, we train the logistic regression model with the Seldonian Toolkit [18], which allows us to constrain a model on arbitrary fairness metrics. The toolkit uses a Seldonian algorithm to provide high-confidence guarantees that the optimized model will not violate the fairness constraint. For this strategy, we propose two fairness constraints:

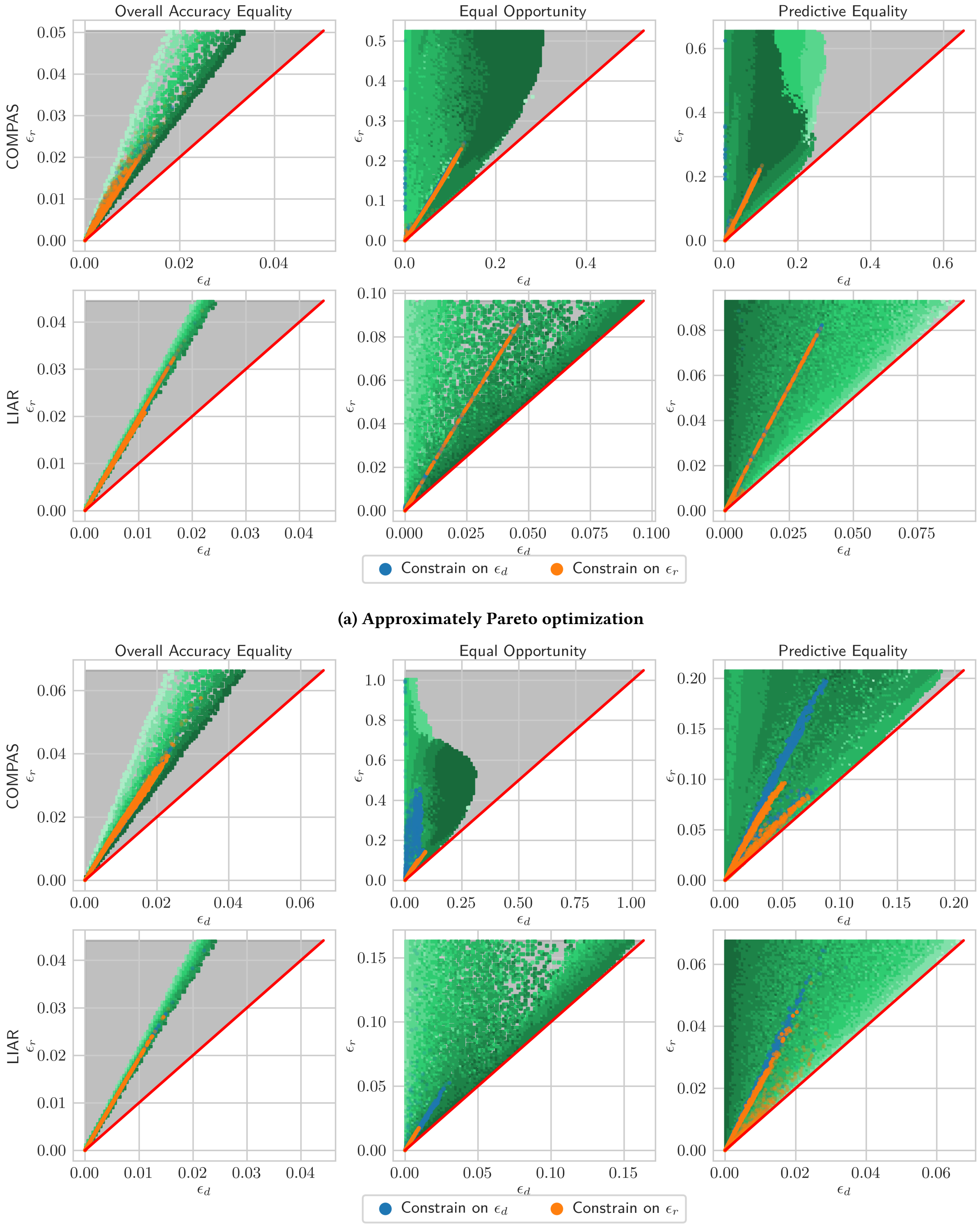
$$|P_M(f_\theta, \mathcal{D}_0) - P_M(f_\theta, \mathcal{D}_1)| \le \hat{\epsilon_d}, \quad (23)$$

and

$$\min\left\{\frac{P_M(f_\theta, \mathcal{D}_0)}{P_M(f_\theta, \mathcal{D}_1)}, \frac{P_M(f_\theta, \mathcal{D}_1)}{P_M(f_\theta, \mathcal{D}_0)}\right\} \ge 1 - \hat{\epsilon_r}, \quad (24)$$

where the $\hat{\epsilon_d}$ and $\hat{\epsilon_r}$ are the hyperparameters of the Seldonian algorithm. We consider six different $\hat{\epsilon_d}$ and $\hat{\epsilon_r}$: 0.2, 0.15, 0.1, 0.075, 0.05, and 0.025. For each setting, we run the Seldonian algorithm 1,000 times with different random initialized parameters.

Subsequently, we visualize the (ϵ_d, ϵ_r) pairs obtained from all trained models, as presented in Figure 6. The outcomes of models trained with the difference approach are represented by blue dots, while the results of models trained with the ratio approach are denoted by orange dots. The result shows that when training with approximately Pareto optimization, although $\mathcal{L}_{\epsilon_d}(\theta)$ and $\mathcal{L}_{\epsilon_r}(\theta)$ represent distinct loss functions, the blue dots and orange dots are nearly aligned along the same line. However, when training with a Seldonian algorithm, the region of (ϵ_d, ϵ_r) pairs looks different, especially in the setting of `[COMPAS, Equal Opportunity]` and `[COMPAS, Predictive Equality]`. For the setting of `[COMPAS, Equal Opportunity]`, the orange dots remain in a straight line while the blue dots cover a wide range of y-axis. For the setting of



(a) Approximately Pareto optimization

(b) Seldonian algorithm

Figure 6: The optimized region of the (ϵ_d, ϵ_r) pairs. We plot the optimized region over the theoretical and empirical region (Figure 2b) to show their relationship. The theoretical and empirical regions are cropped in order to clearly show the optimized region. The outcomes of models trained with the *difference* approach are represented by blue dots, while the results of models trained with the *ratio* approach are denoted by orange dots.

[COMPAS, Predictive Equality], the blue dots are nearly aligned along a line, while the orange dots vary in two lines underneath the blue dots. This result further confirms that optimizing with ϵ_r forces ϵ_d to be smaller.

Furthermore, this result suggests that with certain types of optimization strategies, constraining the model with difference or ratio are similar because of the ϵ_d–ϵ_r relationship shown in Lemma 1. For example, in the case of approximately Pareto optimization, we can rewrite the $\mathcal{L}_{\epsilon_r}(\theta)$ as

$$\mathcal{L}_{\epsilon_r}(\theta) = (1-\lambda)\mathcal{L}_{\text{BCE}}(\theta) + \lambda \frac{|P_M(f_\theta, \mathcal{D}_0) - P_M(f_\theta, \mathcal{D}_1)|}{\max\{P_M(f_\theta, \mathcal{D}_0), P_M(f_\theta, \mathcal{D}_1)\}}. \tag{25}$$

As a result, $\mathcal{L}_{\epsilon_r}(\theta)$ can be viewed as $\mathcal{L}_{\epsilon_d}(\theta)$ but with a different weight on the fairness term. However, this does not hold in the Seldonian algorithm, as it required to compute

$$\mathbb{E}\left[1 - \min\left\{\frac{P_M(f_\theta, \mathcal{D}_0)}{P_M(f_\theta, \mathcal{D}_1)}, \frac{P_M(f_\theta, \mathcal{D}_1)}{P_M(f_\theta, \mathcal{D}_0)}\right\}\right], \tag{26}$$

which is not equivalent to

$$\frac{\mathbb{E}[|P_M(f_\theta, \mathcal{D}_0) - P_M(f_\theta, \mathcal{D}_1)|]}{\mathbb{E}[\max\{P_M(f_\theta, \mathcal{D}_0), P_M(f_\theta, \mathcal{D}_1)\}]}. \tag{27}$$

Furthermore, the result suggests that when the relationship between the difference and ratio approaches in the training strategies is not clear, or the ratio approach can not be simplified as a normalized difference approach, it is better to consider the ratio approach. As in these cases, there is no guarantee that constraining with the difference approach and with the ratio approach will have the same outcome. Instead, there may be a case that constraining with the difference approach will result in models that have a small ϵ_d value but a large ϵ_r value.

5 CONCLUSION

In this study, we examine the relationship between ϵ_d and ϵ_r. We derive their theoretical relationship, showing that, theoretically, constraining models on the ratio is more restrictive than on the difference. Our empirical results suggests that when the optimized metric value is close to 0, it is better to constrain models with the *ratio* approach and to report the fairness with the ϵ_r value. This is because a classifier that is fair w.r.t. a tolerance T under the difference approach can be unfair w.r.t. any tolerance T' under the ratio approach. However, a classifier that is fair w.r.t. a tolerance T under the ratio approach will be fair w.r.t. any $T' \leq T$ under the difference approach Furthermore, the result of the optimized relationship indicates that under certain conditions, both the difference and ratio approaches can achieve the same optimized fairness value, and the relationship between ϵ_d and ϵ_r is linear. In such cases, we can use a transformation function to map between the difference and the ratio constraints. As no theoretical proof currently exists for why this special case arises, we assert that the ratio approach may be the safer option. When researchers are uncertain whether to adopt the difference or ratio approach, we expect this research to serve as a guide for researchers selecting the constraint and reporting fairness results.

There are still numerous unexplored questions in the relationship between these two approaches. For example, whether constraining models using one approach requires less data to achieve a fair result than the other, or which one is less sensitive to noise in the training data. Future work could address these questions, further clarifying the relationship between the difference and ratio approaches.

6 LIMITATIONS

Our study focuses on scenarios that only have two sensitive groups. However, for other scenarios that have three or more sensitive groups, there will be ϵ_d and ϵ_r values for each group combination. Therefore, the theoretical relationship we derived can only apply to each group combination independently. Another way for measuring ϵ_d and ϵ_r for all groups is to measure them with meta-metrics, such as max-min difference and max-min ratio, respectively [21]. In this case, the theoretical relationship should be further studied.

Another limitation is that our study focuses on which approach (ratio or difference) is more appropriate for arbitrary problems. However, there might be other problem-specific reasons for a study to select one approach over the other, e.g., legislation requiring fairness with respect to the difference, or convexity of the resulting fairness constraints.

ACKNOWLEDGMENTS

This work was supported by gifts from Meta Research and the Berkeley Existential Risk Initiative.

REFERENCES

[1] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. 2011. Sentiment Analysis of Twitter Data. In *Proceedings of the Workshop on Language in Social Media (LSM 2011)*. Association for Computational Linguistics, Portland, Oregon, 30–38. https://aclanthology.org/W11-0705

[2] Abdulaziz A. Almuzaini, Chidansh A. Bhatt, David M. Pennock, and Vivek K. Singh. 2022. ABCinML: Anticipatory Bias Correction in Machine Learning Applications. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 1552–1560. https://doi.org/10.1145/3531146.3533211

[3] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. 2016. Machine Bias: There's software used across the country to predict future criminals. And it's biased against blacks. *ProPublica* (May 2016). https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing

[4] Solon Barocas, Moritz Hardt, and Arvind Narayanan. 2019. *Fairness and Machine Learning: Limitations and Opportunities*. fairmlbook.org. http://www.fairmlbook.org.

[5] Solon Barocas and Andrew D. Selbst. 2016. Big Data's Disparate Impact. *California Law Review* 104, 3 (2016), 671–732. http://www.jstor.org/stable/24758720

[6] Ainhize Barrainkua, Paula Gordaliza, Jose A. Lozano, and Novi Quadrianto. 2023. Uncertainty in Fairness Assessment: Maintaining Stable Conclusions Despite Fluctuations. arXiv:2302.01079 (Feb 2023). http://arxiv.org/abs/2302.01079 arXiv:2302.01079 [cs].

[7] Joachim Baumann, Anikó Hannák, and Christoph Heitz. 2022. Enforcing Group Fairness in Algorithmic Decision Making: Utility Maximization Under Sufficiency. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 2315–2326. https://doi.org/10.1145/3531146.3534645

[8] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. 2021. Fairness in Criminal Justice Risk Assessments: The State of the Art. *Sociological Methods & Research* 50, 1 (Feb 2021), 3–44. https://doi.org/10.1177/0049124118782533

[9] Toon Calders and Sicco Verwer. 2010. Three naive Bayes approaches for discrimination-free classification. *Data Mining and Knowledge Discovery* 21, 2 (Sep 2010), 277–292. https://doi.org/10.1007/s10618-010-0190-x

[10] Alexandra Chouldechova. 2017. Fair Prediction with Disparate Impact: A Study of Bias in Recidivism Prediction Instruments. *Big Data* 5, 2 (Jun 2017), 153–163. https://doi.org/10.1089/big.2016.0047

[11] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD '17)*. Association for Computing Machinery, New York, NY, USA, 797–806. https://doi.org/10.1145/3097983.3098095

[12] Michele Donini, Luca Oneto, Shai Ben-David, John Shawe-Taylor, and Massimiliano Pontil. 2018. Empirical risk minimization under fairness constraints. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems* (Montréal, Canada) *(NIPS'18, Vol. 31)*, S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (Eds.). Curran Associates Inc., Red Hook, NY, USA, 2796–2806. https://proceedings.neurips.cc/paper_files/paper/2018/file/83cdcec08fbf90370fcf53bdd56604ff-Paper.pdf

[13] Samuel Dooley, Rhea Sukthanker, John Dickerson, Colin White, Frank Hutter, and Micah Goldblum. 2023. Rethinking Bias Mitigation: Fairer Architectures Make for Fairer Face Recognition. In *Advances in Neural Information Processing Systems*, A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (Eds.), Vol. 36. Curran Associates, Inc., 74366–74393. https://proceedings.neurips.cc/paper_files/paper/2023/file/eb3c42ddfa16d8421fdba13528107cc1-Paper-Conference.pdf

[14] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference (ITCS '12)*. Association for Computing Machinery, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[15] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference* (Cambridge, Massachusetts) *(ITCS '12)*. Association for Computing Machinery, New York, NY, USA, 214–226. https://doi.org/10.1145/2090236.2090255

[16] Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems (NIPS'16)*. Curran Associates Inc., Red Hook, NY, USA, 3323–3331.

[17] Sam Harper, Nicholas B. King, Stephen C. Meersman, Marsha E. Reichman, Nancy Breen, and John Lynch. 2010. Implicit value judgments in the measurement of health inequalities. *The Milbank Quarterly* 88, 1 (March 2010), 4–29. https://doi.org/10.1111/j.1468-0009.2010.00587.x

[18] Austin Hoag, James E. Kostas, Bruno Castro da Silva, Philip S. Thomas, and Yuriy Brun. 2023. Seldonian Toolkit: Building Software with Safe and Fair Machine Learning. In *Proceedings of the 45th International Conference on Software Engineering: Companion Proceedings* (Melbourne, Victoria, Australia) *(ICSE '23)*. IEEE Press, 107–111. https://doi.org/10.1109/ICSE-Companion58688.2023.00035

[19] Abigail Z. Jacobs and Hanna Wallach. 2021. Measurement and Fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (Virtual Event, Canada) *(FAccT '21)*. Association for Computing Machinery, New York, NY, USA, 375–385. https://doi.org/10.1145/3442188.3445901

[20] Disi Ji, Padhraic Smyth, and Mark Steyvers. 2020. Can I Trust My Fairness Metric? Assessing Fairness with Unlabeled Data and Bayesian Inference. In *Advances in Neural Information Processing Systems*, Vol. 33. Curran Associates, Inc., 18600–18612. https://proceedings.neurips.cc/paper/2020/hash/d83de59e10227072a9c034ce10029c39-Abstract.html

[21] Kristian Lum, Yunfeng Zhang, and Amanda Bower. 2022. De-Biasing "bias" Measurement. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (Seoul, Republic of Korea) *(FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 379–389. https://doi.org/10.1145/3531146.3533105

[22] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6, Article 115 (jul 2021), 35 pages. https://doi.org/10.1145/3457607

[23] Alan Mishler, Edward H. Kennedy, and Alexandra Chouldechova. 2021. Fairness in Risk Assessment Instruments: Post-Processing to Achieve Counterfactual Equalized Odds. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Virtual Event Canada, 386–400. https://doi.org/10.1145/3442188.3445902

[24] Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. 2015. Deception Detection using Real-life Trial Data. In *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction (ICMI '15)*. Association for Computing Machinery, New York, NY, USA, 59–66. https://doi.org/10.1145/2818346.2820758

[25] Yash Savani, Colin White, and Naveen Sundar Govindarajulu. 2020. Intra-Processing Methods for Debiasing Neural Networks. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 2798–2810. https://proceedings.neurips.cc/paper_files/paper/2020/file/1d8d70dddf147d2d92a634817f01b239-Paper.pdf

[26] Sandipan Sikdar, Florian Lemmerich, and Markus Strohmaier. 2022. GetFair: Generalized Fairness Tuning of Classification Models. In *2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*. Association for Computing Machinery, New York, NY, USA, 289–299. https://doi.org/10.1145/3531146.3533094

[27] Harvineet Singh, Rina Singh, Vishwali Mhasawade, and Rumi Chunara. 2021. Fairness Violations and Mitigation under Covariate Shift. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Virtual Event Canada, 3–13. https://doi.org/10.1145/3442188.3445865

[28] S. S. Stevens. 1946. On the Theory of Scales of Measurement. *Science* 103, 2684 (1946), 677–680. https://doi.org/10.1126/science.103.2684.677 arXiv:https://www.science.org/doi/pdf/10.1126/science.103.2684.677

[29] Philip S. Thomas, Bruno Castro da Silva, Andrew G. Barto, Stephen Giguere, Yuriy Brun, and Emma Brunskill. 2019. Preventing undesirable behavior of intelligent machines. *Science* 366, 6468 (Nov 2019), 999–1004. https://doi.org/10.1126/science.aag3311

[30] Sahil Verma and Julia Rubin. 2018. Fairness definitions explained. In *Proceedings of the International Workshop on Software Fairness*. ACM, Gothenburg Sweden, 1–7. https://doi.org/10.1145/3194770.3194776

[31] Jialu Wang, Yang Liu, and Caleb Levy. 2021. Fair Classification with Group-Dependent Label Noise. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*. ACM, Virtual Event Canada, 526–536. https://doi.org/10.1145/3442188.3445915

[32] William Yang Wang. 2017. "Liar, Liar Pants on Fire": A New Benchmark Dataset for Fake News Detection. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Vancouver, Canada, 422–426. https://doi.org/10.18653/v1/P17-2067

[33] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P. Gummadi. 2017. Fairness Constraints: Mechanisms for Fair Classification. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics (Proceedings of Machine Learning Research, Vol. 54)*, Aarti Singh and Jerry Zhu (Eds.). PMLR, 962–970. https://proceedings.mlr.press/v54/zafar17a.html

[34] Lei Zheng, Vahid Noroozi, and Philip S. Yu. 2017. Joint Deep Modeling of Users and Items Using Reviews for Recommendation. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM '17)*. Association for Computing Machinery, New York, NY, USA, 425–434. https://doi.org/10.1145/3018661.3018665