Bounding the Fairness of a Classifier Using Population-Level Statistics

SIVAN SABATO*, McMaster University, Canada, Ben-Gurion University of the Negev, Israel, and Canada CIFAR AI Chair, Vector Institute, Canada
ELAD YOM-TOV, Bar Ilan University, Israel

Background: Classifiers increasingly affect people's lives, necessitating their audit for fairness and accuracy on diverse populations. However, direct auditing is often not possible, due to a lack of access to the classifier or to suitable individual-level validation data.
Objectives: This work aims to assess the fairness and accuracy of black-box classifiers using only population-level statistics, without requiring access to the classifier or individual predictions. Specifically, it introduces a method to lower-bound the discrepancy of a classifier: a quantity that jointly captures inaccuracy and unfairness.
Methods: We define a novel measure of unfairness based on the equalized odds fairness criterion, quantifying the fraction of the population on which a classifier deviates from ideal fair behavior. Using this measure, we develop a computationally efficient procedure for calculating the tightest possible lower bound on the classifier's discrepancy, using only aggregated rates of positive predictions and true positives across protected sub-populations.
Results: Empirical evaluations confirm the tightness of the proposed lower bound in practical settings. The method is demonstrated on several use cases, including estimating the reliability of voting polls and assessing the fairness of patient identification from internet search data. The code and data are available at https://github.com/sivansabato/bfa2.
Conclusions: This work provides a practical and interpretable framework for auditing classifiers using population-level statistics. The proposed approach enables stakeholders to identify fairness and accuracy concerns in settings where traditional auditing is not feasible.

JAIR Associate Editor: Tongliang Liu

JAIR Reference Format:
Sivan Sabato and Elad Yom-Tov. 2026. Bounding the Fairness of a Classifier Using Population-Level Statistics. *Journal of Artificial Intelligence Research* 86, Article 47 (August 2026), 34 pages. DOI: 10.1613/jair.1.21172

1 Introduction

As machine learning models increasingly affect people's lives, it has become clear that it is not always sufficient to have accurate classifiers; in many cases, they should also be required to apply fair treatment to different sub-populations. The importance of auditing classifiers for fairness has been noted in many works (e.g., (Black, Elzayn, et al. 2022; Goyal et al. 2022; Roth et al. 2022)). However, in practice, in many cases it is difficult to audit classifiers for these properties, for instance due to lack of high-quality individual-level validation data, or because the classifier is considered proprietary by the developer in industry or government, and so direct access and use of the classifier is precluded. In these cases, we may only have access to aggregate statistics.

As a motivating example, suppose that a US-based health insurance company uses some unpublished method to decide whether a person should be classified as "at risk" for diabetes, for instance so as to offer diabetes screening.

*Corresponding Author.

Authors' Contact Information: Sivan Sabato, sabatos@mcmaster.ca, McMaster University, Hamilton, Ontario, Canada and Ben-Gurion University of the Negev, Beer Sheba, Israel and Canada CIFAR AI Chair, Vector Institute, Toronto, Ontario, Canada; Elad Yom-Tov, ORCID: 0000-0002-2380-4584, elad.yom-tov@biu.ac.il, Bar Ilan University, Ramat Gan, Israel.

DOI: 10.1613/jair.1.21172

Two desirable properties of such a classifier are accuracy and fairness. For this example, consider fairness with respect to the state of residence. The only details that are publicly available from the insurance company are aggregate statistics on the number of people identified as "at risk" in each state of residence. In addition, true proportions of diabetes in each state are provided by government agencies. Can this information be used to gain quantifiable insights on the quality of this proprietary classifier, in terms of its fairness and accuracy? Moreover, suppose that the company publishes some additional information about the accuracy of the classifier, or about its fairness with respect to the state of residence. Can this information be used to draw any conclusions about the other property?

A similar challenge arises when a classifier is designed or learned with little or no labeled training data, or when the available training data is not representative of the target distribution. When labeled data is insufficient, it is still possible to construct a classifier using various methodologies, such as unsupervised learning, transfer learning, and hand-crafting rules based on domain expertise. The challenge is to then estimate the quality of the classifier without a validation set. Here too, we are interested in both the accuracy of the classifier and its fairness with respect to some attribute of interest. For instance, suppose that the goal is to predict how individuals will vote in the upcoming elections based on their votes in the previous elections. Polling agencies use various techniques to predict how voters will vote, based on population demographics and other information. Can we draw any conclusions on the quality, that is, the accuracy and fairness, of the underlying prediction mechanisms, using only publicly available voting statistics? We show in this work how such conclusions can be derived.

To derive such conclusions, we exploit unavoidable trade offs between accuracy and fairness. It is well known that fairness and accuracy in classification do not always co-exist: in some cases, a more accurate classifier must be less fair, and vice versa (Goel et al. 2018; Kleinberg et al. 2017; Menon and Williamson 2018; Pleiss et al. 2017). We show that this property can be used to draw conclusions on classifier accuracy and fairness, based on partial aggregate information about the classifier.

We consider fairness in the sense of equalized odds (Hardt et al. 2016) (also termed *disparate mistreatment*; see, e.g., (Zafar, Valera, Gomez Rodriguez, et al. 2017)). This notion defines a classifier as fair if its false positive rate and its false negative rate, conditioned on the value of the fairness attribute, are the same for all attribute values. Our main interest is in cases where the protected attribute for which fairness is desired is also multi-valued, as in attributes such as race, level of education, or state of residence.

Clearly, if the confusion matrix conditioned on each of the attribute values is known, then it is easy to check whether a classifier is fair and also to calculate its error. However, the confusion matrix may not be known, as in the scenarios described above. We study the conclusions that can be drawn based on *label proportions*: this is a setting in which the only available information about the classifier is the proportions of predictions of each label in each sub-population. In addition, we assume access to the true label proportions in each sub-population and to the proportion out of the population of each sub-population. For instance, in the voting example, we have access to the fraction of votes for each party in the previous elections and in the predicted elections in each state, and the number of voters in each state.

We first discuss the case where the classifier is known to be fair. We show that in this case, a simple linear program can be applied to the known label proportions to obtain a lower bound on the error of the classifier. We then proceed to consider classifiers that are not necessarily completely fair. While complete fairness of a classifier is a desired property, it cannot always be guaranteed. It is thus of interest to be able to measure *how* unfair is a classifier: the least unfair, the better. Quantifying the unfairness of a classifier is an important and under-studied problem; Previous works (e.g., (Alghamdi et al. 2020; Calmon et al. 2017a; Donini et al. 2018; Jung et al. 2021) use various approaches, which usually involve a direct comparison of pairs of conditional probabilities. However, these comparisons are not directly interpretable, and the resulting quantification had significant disadvantages, as we discuss below.

In this work, we propose a principled approach for quantifying unfairness. We model the amount of unfairness as the fraction of the population treated differently from an underlying baseline treatment, and use this definition to derive a well-defined unfairness measure with a clear quantifiable meaning. We show that this measure satisfies properties that are natural to expect from an unfairness measure and are not always satisfied by previous approaches. We then show that given the confusion matrices of the classifier in each sub-population, the unfairness measure that we define can be formulated as the solution to a minimization problem over the unobserved baseline confusion matrix that can be solved accurately. We next study drawing conclusions on the unfairness and the error of a classifier given label proportion information, without access to the confusion matrices of the classifier in each sub-population. We define a measure called *discrepancy*, which balances unfairness and error. The minimal possible discrepancy for given label proportions captures the inherent accuracy-fairness trade off that can be inferred from these proportions. We provide an efficient algorithm that accurately solves this high-dimensional minimization problem, and provides the minimal possible discrepancy value given the known label proportions. This lower bound can be used to answer various questions, such as:

- Is it possible that the classifier is fair? If not, how unfair must it be based on the provided label proportions?
- Suppose that the classifier is known to be nearly fair. How accurate can it be?
- Suppose that the classifier is known to be quite accurate. How fair can it be?
- Suppose that there is a penalty for each person who gets a wrong prediction, and for each person who is treated unfairly; what is the smallest possible overall cost of the classifier?

We report experiments that demonstrate the success of the methods in various scenarios. Our experiments further demonstrate diverse applications of these tools for real-world scenarios, in applications such as election prediction, disease diagnosis, and analysis of access to health care. An implementation of the procedures proposed in this paper and the code for running the experiments are provided in https://github.com/sivansabato/bfa2.

Contributions. The main contributions of this paper are the following:

- A method for lower-bounding the error of a fair classifier using only label proportions;
- A new interpretable unfairness measure, derived based on a probabilistic model;
- An efficient procedure for calculating the value unfairness, given the conditional confusion matrices of the classifier;
- An efficient procedure for lower-bounding the classifier's discrepancy, which combines unfairness and error, using only label proportions;
- Experiments demonstrating the tightness of the lower bound in various real-world data sets;
- Experiments demonstrating the use of the approach to calculate Pareto-curves for classifiers without access to individual data, and to obtain revealing information about inequity in health-care;
- A theoretical and empirical study of the case of protected attributes that are mostly concentrated on two labels.

In this work we study binary classifiers, that is, classifiers that map individuals to one of two labels. A recent follow-up work (Sabato, Treister, et al. 2025) shows how to generalize and extend the measure and the methods to the multiclass setting, in which classifiers can map individuals to one of many possible labels. The discrepancy lower bound that we calculate is more informative when there are more sub-populations, since each sub-population adds aggregate statistics that can inform the lower bound. If all or almost all of the probability mass is concentrated on only two sub-populations, the bounds remain valid, but can sometimes be very loose. We demonstrate this phenomenon in our experiments.

An earlier short version of this work appeared as (Sabato and Yom-Tov 2020). The current work includes more in depth analyses and full proofs, including a new fast minimization algorithm for calculating a discrepancy lower bound (Section 7.3), a distributional derivation of the proposed unfairness measure (Section 5), and additional

experiments and discussions. In particular, experiments with decision trees were added (Section 8.1), and the case of protected attributes that are mostly concentrated on two values is discussed in depth, including a new analysis of the challenges of this case and new experiments demonstrating the possible effects on results (Section 8.3).

Paper structure. We discuss related work in Section 2. The setting and notation are defined in Section 3. In Section 4, we discuss lower-bounding the error using label proportions when the classifier is known to be fair. In Section 5, the new `unfairness` measure is proposed. In Section 6, we show how to calculate the `unfairness` of a classifier when the confusion matrices of each sub-population are known. In Section 7, algorithms for lower bounding the classifier discrepancy based on label proportions are provided. Experiments are reported in Section 8. We conclude with a discussion in Section 9.

2 Related Work

Fairness in classification has been a highly studied topic of research in recent years, due to its importance in legal, financial, and medical decisions (Barocas et al. 2017). A comparison of learning methods which incorporate fairness is given in (Ghosh et al. 2023). The importance of fairness has grown in parallel with the wide application of automated (and frequently, opaque) models in multiple areas affecting people. Various notions of fairness have been proposed (e.g., (Berk et al. 2018; Dwork et al. 2012; Grgic-Hlaca et al. 2016; Kusner et al. 2017; Verma and Rubin 2018)).

In this work, we focus on the notion of equalized odds (Hardt et al. 2016), which requires the false positive rates to be the same in each sub-population, and similarly for false negative rates. Methods for learning fair classifiers under the equalized-odds definition have been proposed in many works (e.g., (Feldman et al. 2015; Goh et al. 2016; Hardt et al. 2016; Nandy et al. 2022; Woodworth et al. 2017; Wu et al. 2019; Yu et al. 2024; Zafar, Valera, Gomez Rodriguez, et al. 2017; Zafar, Valera, Rogriguez, et al. 2017)). Learning methods that guarantee or approximate other definitions of fairness, such as equal opportunity and demographic parity, have also been widely studied in recent years (e.g., (Calmon et al. 2017b; Donini et al. 2018; Dwork et al. 2012; Goel et al. 2018; Johndrow and Lum 2019; Radovanović and Ivić 2023; Y. Wang et al. 2023; Zemel et al. 2013)).

Once a classifier is trained, estimating its fairness is a crucial task in process of making classifiers usable in the real-world (Bellamy et al. 2019). Fairness auditing has been studied in various contexts, including, for instance, assessment of algorithmic fairness of tax auditing (Black, Elzayn, et al. 2022), testing for discrimination in candidate rankings (Roth et al. 2022), and identifying bias in visual systems (Goyal et al. 2022).

Estimating fairness is easier when full access is given to both the model and to each attribute of individual-level datum. Removing the access to any of these three, or to a combination of them, raises increasing challenges to estimating fairness. Given access to the classifier and its individual predictions, Black et al. (Black, Yeom, et al. 2019) propose a method for fine-grained scrutiny of a classifier beyond group-fairness notions. McDuff et al. (McDuff et al. 2019) propose a simulation-based approach for interrogating the classifier. Kusner et al. (Kusner et al. 2017) define the property of "counterfactual fairness", which can be tested given access to the classifier or to individual classified examples.

In the missing-information regime, fairness auditing where the values of the protected attributes of individuals are not available for the auditor, whether at all or partially, has been widely studied (Ashurst and Weller 2023; Awasthi et al. 2021; Chen et al. 2019; Cornacchia et al. 2023; Fabris et al. 2023; Kallus et al. 2022; Kırnap et al. 2021). These works generally assume that information on individuals is available, and use this information to predict or impute the missing information about the protected attribute. Learning classifiers based on both individual covariates and population statistics has been studied under the name "ecological inference". Jackson et al. (Jackson et al. 2008, 2006) propose methods for regression based on both individual-level and aggregate-level statistics. Sun et al. (Sun et al. 2017) infer voting patterns using aggregate statistics. We are not aware of any works that

attempt to estimate properties of existing classifiers from aggregate statistics alone. In particular, we are not aware of such works in the context of fairness.

Requiring fair classifiers may degrade their accuracy. Menon et al. (Menon and Williamson 2018) provide a method for calculating the degradation in accuracy, as measured by the cost-sensitive risk, that results from requiring classifier fairness for a given population distribution. This degradation is defined as the difference in risk between the Bayes-optimal classifier and the Bayes-optimal fair classifier. Bell et al. (Bell et al. 2023) provide methods for finding a model that satisfies relaxed notions of fairness while maintaining high accuracy. Wang et al. (H. Wang et al. 2024) calculate a Pareto frontier for feasible fairness-constrained classifiers. In contrast, we quantify the fairness and accuracy of a given classifier, which is not necessarily optimal, based on limited aggregate information.

Approaches for quantifying unfairness for individual-fairness notions have been suggested by Speicher et al. (Speicher et al. 2018) and by Heidari et al. (Heidari et al. 2018). Speicher et al. (Speicher et al. 2018) propose a measure of unfairness based on a set of axioms that they suggest the measure should satisfy. Their approach is based on assigning a benefit to each individual as a function of its true label and its predicted label. Heidari et al. (Heidari et al. 2018) provide a family of measures for individual fairness that satisfy certain moral axioms and can be efficiently constrained. For group fairness, common approaches are to measure the unfairness using the absolute difference between relevant conditional distributions (e.g., (Donini et al. 2018; Jung et al. 2021)), or the ratio between them (e.g., (Alghamdi et al. 2020; Calmon et al. 2017a)).

Lamy et al. (Lamy et al. 2019) study fairness for binary protected attributes under a noise model of Scott et al. (Scott et al. 2013). In this model, it is assumed that the fairness attribute is noisy independently of the classifier's prediction, leading to a model in which the confusion matrix conditioned on the attribute value is a mixture of the two true confusion matrices conditioned on the two attribute values. They then show how this model can be combined with previously proposed fairness constraints to yield an adapted constrained optimization formulation for this restricted model.

3 Preliminaries

We consider a binary classification problem in which each individual in the population has a true label in $\mathcal{Y} \equiv \{0, 1\}$. We assume an attribute of interest (e.g, race, state of residence, age) that assigns a value for each individual. We denote the (finite) set of possible values of this attribute by $\mathcal{G}$. We do not impose a limit the number of possible attribute values. For concreteness, we henceforth call the possible values of the attribute *regions*, alluding to location-based attributes such as state of residence or country of origin, although our work is not limited to this type of attributes. A *sub-population* is a subset of the population that includes all the individuals in the same region.

The object of our study is an existing classifier, denote it C, which maps each individual from the population to a predicted label. The predicted label may or may not be equal to the true label of that individual. No assumptions are placed on the way the classifier was generated or on the classification mechanism. For a given classifier C, denote by $\mathcal{D}$ uniform distribution over the population of the triplets of true label, predicted label, and region. A random triplet drawn according to $\mathcal{D}$ is denoted by $(Y, \hat{Y}, G)$, where $Y \in \mathcal{Y}$ is the true label, $\hat{Y} \in \mathcal{Y}$ is the predicted label, and $G \in \mathcal{G}$ is the region of the individual. Denote the probability of an event E according to $\mathcal{D}$ by $\mathbb{P}[E]$, and the probability of an event E conditioned on the individual being from region $g \in \mathcal{G}$ by $\mathbb{P}_g[E] := \mathbb{P}[E \mid G = g]$. Throughout this work, the value of a conditional-probability expression in which the probability of the condition is zero is treated as zero. We define the following notation for properties of $\mathcal{D}$:

- The weight (prevalence) of each sub-population in the distribution is $w_g := \mathbb{P}[G = g]$; The vector of weights is $\mathbf{w} = (w_g)_{g \in \mathcal{G}}$.

Table 1. Notation summary

Notation	Meaning
$\mathbb{P}_g[E]$	$\mathbb{P}[E \mid G = g]$; Probability conditioned on sub-population from g
w_g	$\mathbb{P}[G = g]$; Prevalence of sub-population from region g
$\mathbf{w}$	$(w_g)_{g\in\mathcal{G}}$; Vector of prevalences
π_g^y	$\mathbb{P}_g[Y = y]$; Probability of label y in region g
$\boldsymbol{\pi}_g$	(π_g^0, π_g^1); Vector of label probabilities in region g
$\hat{p}_g^y$	$\mathbb{P}_g[\hat{Y} = y]$; Probability of predicted label y in region g by classifier C
$\hat{\mathbf{p}}_g$	$(\hat{p}_g^0, \hat{p}_g^1)$; Vector of predicted label probabilities in region g for C
α_g^y	$\mathbb{P}_g[\hat{Y} = 1 - y \mid Y = y]$; True and False Positive Rates in region g for C
$\mathbf{A}_g$	$\begin{pmatrix} 1-\alpha_g^0 & \alpha_g^0 \\ \alpha_g^1 & 1-\alpha_g^1 \end{pmatrix}$; The confusion matrix for region g for C
$\mathbf{A}_{\mathcal{G}}$	$\{\mathbf{A}_g\}_{g\in\mathcal{G}}$; The set of confusion matrices for all regions for C
$\mathcal{M}_{2\times 2}$	All 2-by-2 matrices with entries in $[0, 1]$ where each row sums to 1.

- The proportion of true label $y \in \mathcal{Y}$ in sub-population $g \in \mathcal{G}$ is $\pi_g^y := \mathbb{P}_g[Y = y]$; The vector of these proportions is $\boldsymbol{\pi}_g := (\pi_g^0, \pi_g^1)$.
- The proportion of predicted label $y \in \mathcal{Y}$ in sub-population $g \in \mathcal{G}$ is $\hat{p}_g^y := \mathbb{P}_g[\hat{Y} = y]$; The vector of these proportions is $\hat{\mathbf{p}}_g := (\hat{p}_g^0, \hat{p}_g^1)$.
- The false positive rate and the false negative rate for sub-population g are denote by α_g^0, α_g^1, respectively, defined as
$$\alpha_g^y := \mathbb{P}_g[\hat{Y} = 1 - y \mid Y = y].$$
- Denote the confusion matrix on a sub-population g by $\mathbf{A}_g$. This is a 2-by-2 matrix with entries that correspond to the true negative rate and the false positive rate on the first row, and the false negative rate and true positive rate on the second row, where each row sums to 1. Denote the set of all possible confusion matrices by $\mathcal{M}_{2\times 2}$. Denote the indexed set of all such confusion matrices by $\mathbf{A}_{\mathcal{G}} := \{\mathbf{A}_g\}_{g\in\mathcal{G}}$.

The notation provided above is summarized in Table 1. The population error of classifier C is given by:

$$\texttt{error}(\mathbf{A}_{\mathcal{G}}) := \mathbb{P}[\hat{Y} \neq Y] = \sum_{g\in\mathcal{G}} w_g \sum_{y\in\mathcal{Y}} \pi_g^y \alpha_g^y. \tag{1}$$

In the *label-proportions* setting, we assume that the confusion matrices of classifier C under study are unknown, and the only available information on the labels is the proportions of the true labels and the predicted labels in each region. Denote the available information in the label-proportions setting by

$$\texttt{Inputs} := (\mathbf{w}, \{(\boldsymbol{\pi}_g, \hat{\mathbf{p}}_g)\}_{g\in\mathcal{G}}). \tag{2}$$

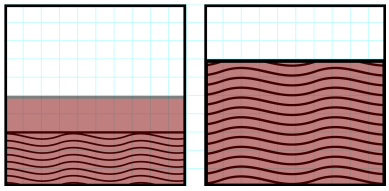
In the next section, we show how the error of a fair classifier C can be lower-bounded using only label proportions.

Label proportions for the population and a given classifier C:

	weight	true positive (π_g^1)	predicted positive ($\hat{p}_g^1$)
g = L(eft)	50%	30%	50%
g = R(ight)	50%	70%	70%

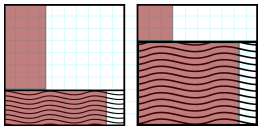
(I) Assume that classifier C has the minimal possible error.
The only solution given the label proportions above:

$$\mathcal{A}_L = \begin{pmatrix} \frac{5}{7} & \frac{2}{7} \\ 0 & 1 \end{pmatrix}, \quad \mathcal{A}_R = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$



This solution is consistent with a classifier that has `error` = 10%, but is unfair. (II) Assume that the classifier C is fair.
The only solution given the label proportions above:

$$\mathcal{A}_L = \mathcal{A}_R = \begin{pmatrix} 0.65 & 0.35 \\ 0.15 & 0.85 \end{pmatrix}$$



This solution is consistent with a classifier that is fair, but has `error` = 25%.

Legend: Predicted positives, True positives

Fig. 1. Two possible confusion matrices for a simple 2-region problem with a given classifier C, with accompanying illustrations representing the fractions of positives and predicted positives in each region. Grid lines are displayed for scale. Given the label proportions provided in the table above, each of the two displayed assumptions leads to a different solution. The first assumes that the error is minimal, while the other assumes that the fairness is maximal (that is, that the classifier C is completely fair under equalized odds). The details of the derivations are provided in Section 4.

4 Lower-Bounding the Error of a Fair Classifier Using Label Proportions

In this work, we consider fairness with respect to the fairness notion of *equalized odds* (Hardt et al. 2016). This notion requires that the false positive rates of the classifier in each sub-population are the same, and similarly for

the false negative rates. Formally, it was defined for the case $\mathcal{Y} = \{0, 1\}$ and $\mathcal{G} = \{0, 1\}$, as follows:

$$\forall y \in \mathcal{Y}, \quad \mathbb{P}[\hat{Y} = 1 \mid Y = y, G = 0] = \mathbb{P}[\hat{Y} = 1 \mid Y = y, G = 1]. \tag{3}$$

In our notation and with possibly more than two regions, this is equivalent to requiring that α_g^0 are the same for all $g \in \mathcal{G}$, and similarly for α_g^1.

We first consider the case where the studied classifier C is known to be fair with respect to the attribute G. This could be the case, for instance, if it is known that a fairness-preserving procedure (e.g., (Hardt et al. 2016; Woodworth et al. 2017)) was used to generate the classifier. More generally, the fact that the classifier is fair may be disclosed to the public even if the classifier itself or its predictions are not accessible. In the label-proportions setting, the confusion matrices of the classifier are not available. In this case, the provided label proportions information can be used to draw conclusions on the error of the classifier.

As a warm-up, we describe a simple example, which shows how the information that the given classifier C is fair along with label-proportions information can be used to draw conclusions on the error of the classifier. Consider a simple classification problem with two regions (see Figure 1). In this problem, $\mathcal{G} = \{L, R\}$ (for "Left" and "Right"). The weights of both regions is the same: $w_L = w_R = 50\%$. In the left region, 30% of the labels are positive ($\pi_L^1 = 0.3$) and the classifier C predicts a positive label for half of the individuals ($\hat{p}_L^1 = 0.5$). In the right region, 70% of the labels are positive ($\pi_R^1 = 0.7$), and this is also the fraction of labels predicted as positive by the classifier ($\hat{p}_R^1 = 0.7$).

If no additional information is provided, and we assume that the classifier C has the minimal possible error given these label proportions, then this is consistent with a classification error of $(50\% - 30\%)/2 = 10\%$ (see case (I) in Figure 1). However, this can only be the case if the classifier is unfair. In particular, this error is obtained only if $\frac{2}{7} = \alpha_L^{01} \neq \alpha_R^{01} = 0$. In contrast, if it is known that C is fair, then $\alpha_L^{yz} = \alpha_R^{yz}$ for all $y, z \in \{0, 1\}$. Under this constraint, only one configuration of the confusion matrices in each region is possible (see case (II) in Figure 1), with $\alpha_L^{01} = \alpha_R^{01} = 0.35$ and $\alpha_L^{10} = \alpha_R^{10} = 0.15$, giving a classification error of $\texttt{error} = 25\%$. Thus, in this example, by using the information that C is fair, we are able to infer its error exactly.

More generally, the information that the classifier is fair can be used to infer a lower bound on the error of the classifier, since fairness imposes constraints on the confusion matrices. Combined with constraints derived from the given label proportions, a linear optimization problem can be derived.

First, given `Inputs` for the classifier C, C must satisfy the following constraints on $\mathbf{A}_{\mathcal{G}}$:

$$\forall g \in \mathcal{G}, \quad \mathbf{A}_g \in \mathcal{M}_{2\times 2}, \text{ and } \mathbf{A}_g^T \boldsymbol{\pi}_g = \hat{\mathbf{p}}_g. \tag{4}$$

In addition, if C is known to be fair in the sense of equalized odds, then all matrices in $\mathbf{A}_{\mathcal{G}}$ are the same. The error of C is determined by $(\mathbf{A}_g)_{g\in\mathcal{G}}$ as well. Thus, a tight lower bound on the error of C can be obtained by solving the following optimization problem, over a matrix $\mathbf{A}$ that represents the confusion matrix of each of the sub-populations.

$$\begin{aligned} \text{Minimize}_{\mathbf{A}} \quad & \texttt{error}(\mathbf{A}) \equiv \sum_{g\in\mathcal{G}} w_g \sum_{y\in\mathcal{Y}} \pi_g^y \alpha_g^y \\ \text{s.t.} \quad & \mathbf{A} \in \mathcal{M}_{2\times 2}, \\ & \forall g \in \mathcal{G}, \quad \mathbf{A}^T \boldsymbol{\pi}_g = \hat{\mathbf{p}}_g. \end{aligned}$$

If this problem is infeasible, and the input values are accurate, then the classifier C cannot be fair. Another option is that C is nearly fair, but the input values are inaccurate, for instance due to estimation error, thus leading to infeasibility of the problem constraints. In both of these cases, C can be analyzed using a measure of *unfairness* of C. In the next section, we propose a new measure of unfairness for unfair classifiers. We further discuss how it can be used also when inputs are inaccurate.

5 Quantifying Unfairness

In the previous section, we considered a fair classifier, which means that the classifier satisfies the equalized odds requirements exactly. However, in many cases, exact fairness may not be guaranteed. For instance, the classifier C may have been constructed using heuristic methods that attempt to obtain fairness (e.g., (Goh et al. 2016; Zafar, Valera, Rogriguez, et al. 2017)). Nonetheless, being *close* to fairness as much as possible is a desired property. To draw conclusions about unfair classifiers, a meaningful measure of the unfairness of a classifier is required. In this section, we propose such an unfairness measure with respect to the equalized-odds fairness criterion. The guiding principle behind the derivation of this measure is that it is derived based on its intended meaning, which is rooted in a probabilistic interpretation.

We define the amount of unfairness as the *fraction of the population on which the classifier behaves differently from its baseline behavior*. This is analogous to the definition of `error`, which indicates the fraction of the population on which the classifier makes a mistake.

Formally, recall that $(Y, \hat{Y}, G) \sim \mathcal{D}$ for the given classifier C. For $y \in \mathcal{Y}$, denote by $\mathcal{D}_g^y$ the conditional distribution of $\hat{Y} \mid Y = y, G = g$. For $g \in \mathcal{G}$, we model $\mathcal{D}_g^y$ as a mixture of a baseline distribution $\mathcal{D}_{\text{b}}^y$ and local "nuisance" distributions $\mathcal{N}_g^y$. Denote by $\eta_g^y \in [0, 1]$ the probability, conditioned on $Y = y, G = g$, that $\hat{Y}$ is drawn from $\mathcal{N}_g^y$. Then we model $\mathcal{D}_g^y$ as the following mixture:

$$\mathcal{D}_g^y = (1 - \eta_g^y)\mathcal{D}_{\text{b}}^y + \eta_g^y \mathcal{N}_g^y. \tag{5}$$

It is easy to see that if the classifier C is fair under the equalized odds criterion, then $\mathcal{D}_g^y = \mathcal{D}_{\text{b}}^y$ for all $g \in \mathcal{G}, y \in \mathcal{Y}$, hence in this case Eq. (5) holds for all $g \in \mathcal{G}, y \in \mathcal{Y}$ for $\eta_g^y = 0$. In contrast, if the C is not completely fair under equalized odds, then any choice of $\mathcal{D}_{\text{b}}^y$ requires $\eta_g^y > 0$ for at least one $g \in \mathcal{G}$. Thus, an η_g^y-fraction of the individuals with $Y = y, G = g$ suffer a different prediction behavior than the baseline prediction behavior defined by $\mathcal{D}_{\text{b}}^y$. Thus, the value of the coefficients $\{\eta_g^y\}$ is related to the deviation of the classifier from the equalized-odds fairness criterion.

We define `unfairness` as the fraction of the population for which prediction behaves differently from the baseline prediction behavior. Since the decomposition to $\mathcal{D}_{\text{b}}^y$ and $\mathcal{N}_g^y$ is unobserved, we define this fraction with respect to the *best possible decomposition*. Formally, let $\boldsymbol{\eta}_g := (\eta_g^0, \eta_g^1)$. We define the `unfairness` measure for the classifier C described by $\mathcal{D}$ as follows:

$$\texttt{unfairness}(\mathcal{D}) := \min_{\{\boldsymbol{\eta}_g\}} \sum_{g \in \mathcal{G}} w_g \boldsymbol{\pi}_g^T \boldsymbol{\eta}_g, \tag{6}$$

where the minimum is over $\{\boldsymbol{\eta}_g\}$ such that Eq. (5) holds for some $\{\mathcal{D}_{\text{b}}^y, \{\mathcal{N}_g^y\}_{g \in \mathcal{G}}\}_{y \in \mathcal{Y}}$. Note that $\texttt{unfairness}(\mathcal{D}) = 0$ if and only if the equalized odds criterion is satisfied by the classifier.

This definition of unfairness has several advantages: First and foremost, it is fully interpretable, since it has a specific and well defined meaning. The value of `unfairness` is always in $[0, 1]$, where $\texttt{unfairness}(\mathcal{D}) = 0$ if and only if the classifier C is completely fair. The value of `unfairness` is meaningful even if the label distribution is extremely unbalanced: Suppose that the target label concerns the diagnosis of a rare disease. This label is highly unlikely, yet if the disease is twice as likely to be wrongly diagnosed in one sub-population than in another, this would manifest as a high `unfairness` value. Moreover, since `unfairness` and `error` are both defined as fractions of the population, they can be compared and combined meaningfully, without the need for rescaling or other manipulations. This contrasts with other fairness measures, which do not satisfy all of these properties. Common approaches for measuring unfairness suggest using the absolute difference between relevant conditional distributions (e.g., (Donini et al. 2018; Jung et al. 2021)), or the ratio between them (e.g., (Alghamdi et al. 2020; Calmon et al. 2017a)). However, neither of these approaches provide an interpretable measure that is always meaningful. The absolute-difference approach generates near-zero values when the labels are unbalanced, while

the ratio approach generates a possibly unbounded value, a concern raised also in (Calmon et al. 2017a). The issue of interpretability is even more pronounced when the protected attribute has many values, requiring these measures to take into account all pairs of differences or ratios.

The interpretability of unfairness as a fraction of the population allows a direct adaptation to possible inaccuracies in input quantities, for instance due to estimation error. Given the value of unfairness that results from the given input information, bounds on the true value can be derived by taking into account known accuracy bounds on the inputs. For instance, if the population weights $\mathbf{w}$ are known to be accurate up to an additive error of ϵ, then the true population unfairness is at most $|\mathcal{G}|\epsilon$-far from the value of unfairness calculated using the given $\mathbf{w}$. More generally, if the inputs are consistent with a distribution that is up to ϵ-far from the true population distribution, then the true unfairness is up to ϵ-far from the value of unfairness that results from using the input values.

6 Calculating Unfairness when Conditional Confusion Matrices are Known

We first consider a scenario in which we have access to the confusion matrices of the classifier C in each sub-population. For instance, this could be the case if we have a validation set, or if the confusion matrices are reported by the owner of C.

We derive a useful formula for unfairness for a known distribution $\mathcal{D}$. Note that $\mathcal{D}$ is fully determined by $\mathbf{w}$, the region weights, and $\mathbf{A}_{\mathcal{G}}$, the confusion matrices of each region. Below, we assume some fixed $\mathbf{w}$ and denote $\texttt{unfairness}(\mathbf{A}_{\mathcal{G}}) := \texttt{unfairness}(\mathcal{D})$.

Consider $\{\boldsymbol{\eta}_g\}$ values that minimize the RHS in the definition of unfairness in Eq. (6). Suppose first that the distributions $\{\mathcal{D}_{\text{b}}^y\}$ that obtain these values are known, and consider nuisance distributions $\{\mathcal{N}_g^y\}$ which are consistent with this choice. Let the random variable N be equal to 1 if the triplet $(\hat{Y}, Y = y, G = g)$ was drawn from $\mathcal{N}_g^y$ and equal to 0 if it was drawn from $\mathcal{D}_{\text{b}}^y$, so that $\eta_g^y = \mathbb{P}_g[N = 1 \mid Y = y]$. Denote the false positive rate and the false negative rate of $\{D_b^y\}$ by $\alpha_{\text{b}}^y := \mathbb{P}_g[\hat{Y} = 1 - y \mid Y = y, N = 0]$, and let $\mathbf{A}_{\text{b}}$ be the corresponding confusion matrix. Denote $\beta_g^y = \mathbb{P}_g[\hat{Y} = 1 - y \mid Y = y, N = 1]$, and note that the matrix $\mathbf{B}_g := (1 - \beta_g^0, \beta_g^0; \beta_g^1, 1 - \beta_g^1)$ is in $\mathcal{M}_{2\times 2}$.

the following relationship holds:

$$\begin{aligned}
\alpha_g^y &= \mathbb{P}_g[\hat{Y} \neq y \mid Y = y] \\
&= \quad \mathbb{P}_g[\hat{Y} \neq y \mid Y = y, N = 0] \cdot \mathbb{P}_g[N = 0 \mid Y = y] \\
&\quad + \mathbb{P}_g[\hat{Y} \neq y \mid Y = y, N = 1] \cdot \mathbb{P}_g[N = 1 \mid Y = y] \\
&= \alpha_{\text{b}}^y(1 - \eta_g^y) + \beta_g^y \cdot \eta_g^y.
\end{aligned}$$

Since $\beta_g^y \in [0, 1]$, we get

$$\alpha_{\text{b}}^y(1 - \eta_g^y) \leq \alpha_g^y \leq \alpha_{\text{b}}^y(1 - \eta_g^y) + \eta_g^y.$$

Thus, for $\alpha_{\text{b}}^y \in (0, 1)$, we have the following lower bound: $\eta_g^y \geq \max(1 - \frac{\alpha_g^y}{\alpha_{\text{b}}^y}, 1 - \frac{1-\alpha_g^y}{1-\alpha_{\text{b}}^y})$. In addition, if $\alpha_{\text{b}}^y = 0$, then $\eta_g^y \geq \alpha_g^y$, and if $\alpha_{\text{b}}^y = 1$, then $\eta_g \geq 1 - \alpha_g^y$. Moreover, η_g^y satisfies these lower bounds with equality. This can be seen by considering a deterministic nuisance distribution which draws y if $\alpha_g^y < \alpha_{\text{b}}^y$ and $1 - y$ otherwise. Thus, given α_{b}^y and α_g^y, the minimal value of η_g^y is $\eta(\alpha_{\text{b}}^y, \alpha_g^y)$, where $\eta : [0, 1]^2 \to [0, 1]$ is defined as follows (see Figure 2):

$$\eta(a, b) = \begin{cases} 1 - b/a & b < a, \\ 1 - (1 - b)/(1 - a) & b > a, \\ 0 & b = a. \end{cases}$$

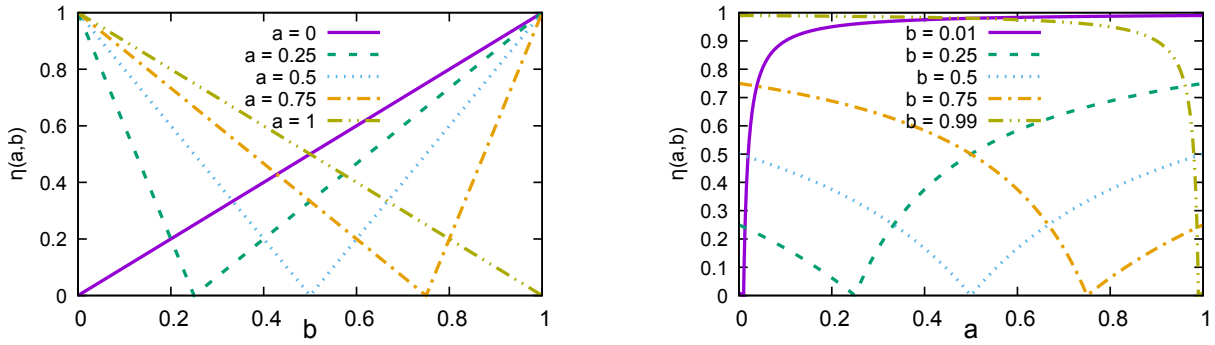


Fig. 2. Left: $b \mapsto \eta(a, b)$ for fixed a. Right: $a \mapsto \eta(a, b)$ for fixed b.

To find the value of $\texttt{unfairness}$ for the best possible decomposition, we minimize over α_{b}^0 and α_{b}^1, the false positive rate and false negative rate that determine the baseline distribution:

$$\texttt{unfairness}(\mathbf{A}_{\mathcal{G}}) = \sum_{y \in \mathcal{Y}} \min_{\alpha_{\mathrm{b}}^y \in [0,1]} \sum_{g \in \mathcal{G}} w_g \pi_g^y \eta(\alpha_{\mathrm{b}}^y, \alpha_g^y). \tag{7}$$

We now show that this function can be easily minimized exactly. Define the function $\psi^y(a) := \sum_{g \in \mathcal{G}} w_g \pi_g^y \eta(a, \alpha_g^y)$. First, suppose that for all $g \in \mathcal{G}, y \in \mathcal{Y}$, we have $\alpha_g^y \in (0, 1)$. For any fixed $b \in (0, 1)$, $a \mapsto \eta(a, b)$ is concave on the intervals $[0, b]$ and $[b, 1]$. Thus, ψ^y is concave on any closed interval which is a subset of $[0, \alpha_g^y]$ or of $[\alpha_g^y, 1]$ for all $g \in \mathcal{G}$. In each such interval, the minimizer of ψ^y is one of the end points of the interval. Therefore, ψ^y is minimized at an end point of a maximal interval which satisfies this property, that is, at a point in $A := \{\alpha_g^y\}_{g,y} \cup \{0, 1\}$.

Now, if for some g, y, we have $\alpha_g^y \in \{0, 1\}$, then $\eta(a, \alpha_g^y) = \mathbf{1}_{a \neq \alpha_g^y}$. Thus, this does not add other possible minimizers to A. Hence,

$$\texttt{unfairness}(\mathbf{A}_{\mathcal{G}}) = \sum_{y \in \mathcal{Y}} \min_{\alpha_{\mathrm{b}}^y \in A} \sum_{g \in \mathcal{G}} w_g \pi_g^y \eta(\alpha_{\mathrm{b}}^y, \alpha_g^y).$$

We conclude that given $\mathbf{A}_{\mathcal{G}}$, the value of $\texttt{unfairness}$ can be calculated exactly, in time linear in $|\mathcal{G}|$.

7 Lower-Bounding Classifier Discrepancy from Label Proportions

In the previous section, we discussed the calculation of $\texttt{unfairness}$ based on the confusion matrices of the classifier C in each sub-population. However, as discussed in the Introduction, these confusion matrices might be unavailable, especially if we do not have access to the classifier or to individual-level validation data. Nonetheless, we show in this section that it is possible to provide meaningful conclusions on the properties of C from label proportions only. Formally, we assume the label-proportions setting in which the only available information is the aggregate statistics in $\texttt{Inputs}$, defined in Eq. (2).

We define a combined measure for the accuracy and fairness of C, which we term the *discrepancy*, denoted $\texttt{disc}_\beta$, where $\beta \in [0, 1]$ is a trade-off parameter:

$$\texttt{disc}_\beta(\mathbf{A}_{\mathcal{G}}) := \quad \beta \cdot \texttt{unfairness}(\mathbf{A}_{\mathcal{G}}) + (1 - \beta) \cdot \texttt{error}(\mathbf{A}_{\mathcal{G}}). \tag{8}$$

Bounding the value of $\texttt{disc}_\beta$ with an appropriate value of β can provide answers to the questions about the classifier C which are listed in Section 1.

Each of the questions is equivalent to asking for a lower bound on $\texttt{disc}_\beta$, for some $\beta \in [0,1]$, or for a β that optimizes a constraint. For instance, if it is known that $\texttt{unfairness} \leq U$ for some known U, then finding the best possible accuracy of C is equivalent to minimizing $\texttt{disc}_\beta$ for the smallest β such that the minimizing solution satisfies $\texttt{unfairness} \leq U$. Moreover, one can assign a cost per individual who is treated unfairly and a (possibly different) cost per individual whose classification is incorrect. Setting β according to the ratio between these costs then makes $\texttt{disc}_\beta$ of a classifier equivalent to its cost, to facilitate auditing and classifier selection.

Given β, we are therefore interested in the smallest possible $\texttt{disc}_\beta$ value under the constraints imposed by $\texttt{Inputs}$. Define:

$$\texttt{mindisc}_\beta(\texttt{Inputs}) := \min\{\texttt{disc}_\beta(\mathbf{A}_\mathcal{G}) \mid \forall g \in \mathcal{G}, \mathbf{A}_g \in \mathcal{M}_{2\times 2}, \mathbf{A}_g^T \boldsymbol{\pi}_g = \hat{\mathbf{p}}_g\}. \tag{9}$$

Our goal is thus to calculate or bound $\texttt{mindisc}_\beta(\texttt{Inputs})$. Note that in the case of $\beta = 0$, $\texttt{disc}_0 = \texttt{error}$. Thus, it is easy to see that $\texttt{mindisc}_0(\texttt{Inputs}) = \sum_g w_g |\pi_g^1 - \hat{p}_g^1|$. We henceforth consider $\beta \in (0,1]$.

We show below that $\texttt{mindisc}_\beta(\texttt{Inputs})$ can be calculated exactly using an efficient procedure. Note that a lower bound on $\texttt{mindisc}_\beta$ implies a lower bound on $\texttt{disc}_\beta$, while an upper bound on $\texttt{mindisc}_\beta$ implies a limitation on the conclusions that can be drawn on $\texttt{disc}_\beta$ using the available information.

To derive the procedures, we first define a useful decomposition of $\texttt{disc}_\beta$. From Eq. (7) and Eq. (1), and setting $\boldsymbol{\alpha}_\text{b} := (\alpha_\text{b}^0, \alpha_\text{b}^1)$, we have

$$\texttt{disc}_\beta(\mathbf{A}_\mathcal{G}) = \beta \min_{\boldsymbol{\alpha}_\text{b} \in [0,1]^2} \sum_{y \in \mathcal{Y}} \sum_{g \in \mathcal{G}} w_g \pi_g^y \eta(\alpha_\text{b}^y, \alpha_g^y) + (1-\beta) \cdot \sum_{g \in \mathcal{G}} w_g \sum_{y \in \mathcal{Y}} \pi_g^y \alpha_g^y. \tag{10}$$

Denote $\tau_\beta(a,b) := \beta\eta(a,b) + (1-\beta)b$, Then

$$\texttt{disc}_\beta(\mathbf{A}_\mathcal{G}) = \min_{\boldsymbol{\alpha}_\text{b} \in [0,1]^2} \sum_{g \in \mathcal{G}} w_g \sum_{y \in \mathcal{Y}} \pi_g^y \tau(\alpha_\text{b}^y, \alpha_g^y). \tag{11}$$

We now turn to the calculation of $\texttt{mindisc}_\beta$. We show that the set of possible solutions for $\mathbf{A}_\mathcal{G}$ can be sufficiently restricted, so that it is computationally tractable to consider them all, and to globally minimize $\mathbf{A}_\text{b}$ with respect to each of these possible solutions. Thus, the value of $\texttt{mindisc}_\beta$ can be calculated accurately.

The constraint $\mathbf{A}_g^T \boldsymbol{\pi}_g = \hat{\mathbf{p}}_g$ is equivalent to requiring $\hat{p}_g^1 = \pi_g^0 \alpha_g^0 + \pi_g^1 (1 - \alpha_g^1)$. Define the set $\mathcal{G}^1 = \{g \in \mathcal{G} \mid \pi_g^1 > 0\}$. For $g \in \mathcal{G}^1$, $\alpha_g^1 = 1 - \hat{p}_g^1/\pi_g^1 + (\pi_g^0/\pi_g^1)\alpha_g^0$. We denote $r_g := 1 - \hat{p}_g^1/\pi_g^1$ and $q_g := \pi_g^0/\pi_g^1$, so that

$$\forall g \in \mathcal{G}^1, \quad \alpha_g^1 = r_g + q_g \alpha_g^0. \tag{12}$$

Note that for $g \notin \mathcal{G}^1$, $\pi_g^0 = 1$ and so $\alpha_g^0 = \hat{p}_g^1$ and $\alpha_g^1 = 0$. Define

$$D_{\beta,g}(\boldsymbol{\alpha}_\text{b}, \alpha_g^0) = \pi_g^0 \tau_\beta(\alpha_\text{b}^0, \alpha_g^0) + \pi_g^1 \tau_\beta(\alpha_\text{b}^1, r_g + q_g \alpha_g^0),$$

where we define the second term to be zero if $\pi_g^1 = 0$. Since $\alpha_g^y \in [0,1]$, it follows from Eq. (12) that the feasible domain of α_g^0 for $g \in \mathcal{G}^1$ is

$$\text{dom}_g := [\max\{-r_g/q_g, 0\}, \min\{(1-r_g)/q_g, 1\}].$$

If $g \notin \mathcal{G}^1$, define $\text{dom}_g := \{\hat{p}_g^1\}$. From Eq. (9), Eq. (11) and Eq. (12), we thus have

$$\texttt{mindisc}_\beta(\texttt{Inputs}) = \min_{\boldsymbol{\alpha}_\text{b} \in [0,1]^2} \left(\sum_{g \in \mathcal{G}} w_g \cdot \min_{\alpha_g^0 \in \text{dom}_g} D_{\beta,g}(\boldsymbol{\alpha}_\text{b}, \alpha_g^0) \right). \tag{13}$$

Below, we show that this minimization problem can be reduced to a small set of one-dimensional minimization problems. We further provide a procedure for minimizing these one-dimensional problems. Combining the two, this gives a procedure for calculating $\texttt{mindisc}_\beta$ accurately. For the sake of presentation, we first provide the procedure and a sketch of its derivation in Section 7.1, and then provide the full derivation, by first showing the

Algorithm 1 Calculating $\texttt{mindisc}_\beta$

Input: $\texttt{Inputs} \equiv (\mathbf{w}, \{(\boldsymbol{\pi}_g, \hat{\mathbf{p}}_g)\}_{g\in\mathcal{G}})$, $\beta \in [0,1]$, tolerance $\gamma > 0$

Output: $V \in [\texttt{mindisc}_\beta(\texttt{Inputs}), \texttt{mindisc}_\beta(\texttt{Inputs}) + \gamma]$, and (u, ϵ), the values of `unfairness` and `error` that obtain the discrepancy V.

For each $g \in \mathcal{G}^+$, $v_g \leftarrow \texttt{ArgMin1D}(g, \gamma, \beta, \texttt{Inputs})$. *#Get minimizer of H^g*
$V_0 \leftarrow \{0,1\} \cup \{-r_g/q_g, (1-r_g)/q_g\}_{g\in\mathcal{G}^1} \cup \{\hat{p}_g\}_{g:\pi_g^0=1}$.
$V_1 \leftarrow \{0,1\} \cup \{r_g, r_g + q_g\}_{g\in\mathcal{G}^1}$.
$P \leftarrow (V_0 \times V_1) \cup \{(v_g, r_g + q_g v_g) \mid g \in \mathcal{G}^+\}$.
$\hat{\boldsymbol{\alpha}}_\mathrm{b} \leftarrow \operatorname{argmin}_{\boldsymbol{\alpha}_\mathrm{b}\in P} F(\boldsymbol{\alpha}_\mathrm{b})$
$V \leftarrow F(\hat{\boldsymbol{\alpha}}_\mathrm{b})$
$\forall g \in \mathcal{G}$, $\hat{\alpha}_g^0 \leftarrow \operatorname*{argmin}_{\alpha_g^0 \in S_g(\hat{\boldsymbol{\alpha}}_\mathrm{b})} D_{\beta,g}(\hat{\boldsymbol{\alpha}}_\mathrm{b}, \alpha_g^0)$
$\forall g \in \mathcal{G}^1$, $\hat{\alpha}_g^1 \leftarrow r_g + q_g \hat{\alpha}_g^0$
$u \leftarrow \sum_{g\in\mathcal{G}} w_g \sum_y \pi_g^y \eta(\hat{\alpha}_\mathrm{b}^y, \hat{\alpha}_g^y)$.
$\epsilon \leftarrow \sum_{g\in\mathcal{G}} w_g \sum_y \pi_g^y \hat{\alpha}_g^y$.
Return V, (u, ϵ)

reduction to few one-dimensional problems in Section 7.2, and then providing the procedure for minimizing these in Section 7.3.

7.1 Calculating $\texttt{mindisc}_\beta$

The algorithm for calculating $\texttt{mindisc}_\beta$ up to a given tolerance is listed as Alg. 1. Alg. 1 gets as input the label proportions `Inputs`, the value of β that specifies the required minimization instance of Eq. (13), and a tolerance parameter $\gamma > 0$. It outputs a discrepancy $V \in \mathbb{R}$ which is the value of $\texttt{mindisc}_\beta(\texttt{Inputs})$ up to a tolerance of the input parameter γ. It further provides the values of `unfairness` and `error` that obtain the discrepancy V. We give here an overview of Alg. 1. For simplicity of presentation, we defer some of the definitions used by the algorithm to Section 7.2.

Denote $\mathcal{G}^+ := \{g \in \mathcal{G} \mid \forall y \in \mathcal{Y}, \pi_g^y \neq 1\}$. Alg. 1 first runs `ArgMin1D`, a one-dimensional minimization procedure, for each $g \in \mathcal{G}^+$. `ArgMin1D` is provided in Section 7.3 below. Running `ArgMin1D` gives a set of values $\{v_g\}_{g\in\mathcal{G}^+}$, which are then used along with two other sets V_0 and V_1 to construct a set P of possible assignments to $\boldsymbol{\alpha}_\mathrm{b}$. We show below that P necessarily includes an assignment which minimizes $\texttt{disc}_\beta$, subject to the provided label proportions in `Inputs`, up to a tolerance of γ. Alg. 1 then finds a minimizing assignment in P using a proxy function F, which is defined in Section 7.2 below. This function is easy to calculate and we show below that minimizing it is equivalent to minimizing the objective. The minimizing assignment is then used to calculate the value of the objective, as well as the value of the `unfairness` and `error` that obtain this objective. The next two subsections provide the full derivation of Alg. 1, proving that Alg. 1 outputs $\texttt{mindisc}_\beta$ up to a tolerance of γ, as required.

7.2 A Reduction to One-Dimensional Minimization Problems

We first show that given a fixed assignment for $\boldsymbol{\alpha}_\mathrm{b} \in [0,1]^2$, $\operatorname{argmin}_{\alpha_g^0 \in \mathrm{dom}_g} D_{\beta,g}(\boldsymbol{\alpha}_\mathrm{b}, \alpha_g^0)$ belongs to a small set of possible values.

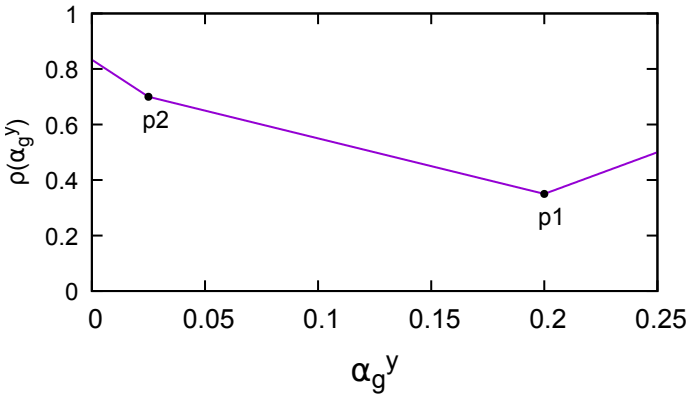


Fig. 3. $\rho(\alpha_g^y)$ and the inflection points p_1, p_2 for $\alpha_{\mathrm{b}}^0 = 0.2, \alpha_{\mathrm{b}}^1 = 0.6, \pi_g^1 = 0.1, \hat{p}_g^1 = 0.1$.

Lemma 1. *Fix some set of parameters* `Inputs`. *For* $g \in \mathcal{G}$, $\boldsymbol{\alpha}_b \in [0,1]^2$, *define*

$$S_g(\boldsymbol{\alpha}_b) := \{s_g^i(\boldsymbol{\alpha}_b)\}_{i\in[4]} \cap \mathrm{dom}_g,$$

where for $g \in \mathcal{G}^1$,

$$\begin{aligned}
s_g^1(\boldsymbol{\alpha}_b) \equiv s_g^1 &:= \max\{0, -r_g/q_g\},\\
s_g^2(\boldsymbol{\alpha}_b) \equiv s_g^2 &:= \min\{1, (1-r_g)/q_g\},\\
s_g^3(\boldsymbol{\alpha}_b) \equiv s_g^3(\alpha_b^0) &:= \alpha_b^0,\\
s_g^4(\boldsymbol{\alpha}_b) \equiv s_g^4(\alpha_b^1) &:= (\alpha_b^1 - r_g)/q_g.
\end{aligned}$$

For $g \in \mathcal{G} \setminus \mathcal{G}^1$, *define* $s_g^i(\boldsymbol{\alpha}_b) = \{\hat{p}_g^1\}$ *for all* $i \in [4]$. *Then*

$$\min_{\alpha_g^0 \in \mathrm{dom}_g} D_{\beta,g}(\boldsymbol{\alpha}_b, \alpha_g^0) = \min_{\alpha_g^0 \in S_g(\boldsymbol{\alpha}_b)} D_{\beta,g}(\boldsymbol{\alpha}_b, \alpha_g^0).$$

Proof. Assume that $\pi_g^0 \in (0,1)$, since the other case is trivial. Consider the function $b \mapsto \tau_\beta(a,b)$ for a fixed a. It is a convex combination of $\eta(a,b)$ and of b. Now, for a fixed $a \in [0,1]$, $b \mapsto \eta(a,b)$ is a convex piecewise-linear function: it is linear and decreasing over $[0,a]$ and linear and increasing over $[a,1]$, as can be verified from the definition of η (see Figure 2 (left) for illustration). Since by Eq. (12), α_g^1 is linear in α_g^0, it follows that $\rho(\alpha_g^0)$ is convex and piecewise-linear, with inflection points at $\alpha_g^0 = \alpha_{\mathrm{b}}^0 =: p_1$, and at $\alpha_g^1 = \alpha_{\mathrm{b}}^1$, that is at $\alpha_g^0 = (\alpha_{\mathrm{b}}^1 - r_g)/q_g =: p_2$ (see Figure 3 for illustration). It follows that the minimizer of $b \mapsto \tau_\beta(\alpha_g^0, b)$ must be one of p_1, p_2 or the end points of dom_g. These are exactly the values in $S_g(\boldsymbol{\alpha}_{\mathrm{b}})$ defined above. □

Define

$$F(\boldsymbol{\alpha}_{\mathrm{b}}) := \sum_{g\in\mathcal{G}} w_g \cdot \min_{\alpha_g^0 \in S_g(\boldsymbol{\alpha}_{\mathrm{b}})} D_{\beta,g}(\boldsymbol{\alpha}_{\mathrm{b}}, \alpha_g^0).$$

Note that $F(\boldsymbol{\alpha}_{\mathrm{b}})$ is easy to calculate, since the minimizations are over $S_g(\boldsymbol{\alpha}_{\mathrm{b}})$, which is a small finite set of values. From Lemma 1 and Eq. (13), it follows that

$$\texttt{mindisc}_\beta = \min_{\boldsymbol{\alpha}_{\mathrm{b}} \in [0,1]^2} F(\boldsymbol{\alpha}_{\mathrm{b}}). \tag{14}$$

While this two-dimensional minimization could be approached using grid-search, the grid resolution might be prohibitively small. This is especially evident when data is extremely unbalanced, as in the case of cancer occurrence statistics, on which we report experiments in Section 8. In such unbalanced cases, the necessary grid size can be prohibitive. For instance, cancer occurrence statistics can be of the order 10^{-6}. A two-dimensional grid search of this resolution would be of order 10^{12}, which could be impractical to perform. Thus, a further reduction of the search size is necessary. We now show that this two-dimensional problem can be reduced to a small set of one-dimensional minimization problems.

Theorem 2. *Let V_0 and V_1 be as defined in Alg. 1. Define the following set of vectors in $\mathbb{R}^2$:*

$$\mathit{Pairs} := (V_0 \times V_1) \cup \{(v, r_g + q_g v) \mid v \in \mathrm{dom}_g, g \in \mathcal{G}^+\}.$$

Then

$$\texttt{mindisc}_\beta = \min_{\boldsymbol{\alpha}_b \in \mathit{Pairs}} F(\boldsymbol{\alpha}_b).$$

The proof of Theorem 2 is given in Appendix A. In the proof, we show that given α_{b}^0, it is easy to solve $\min_{\alpha_{\mathrm{b}}^1 \in [0,1]} F((\alpha_{\mathrm{b}}^0, \alpha_{\mathrm{b}}^1))$, since there is a small finite number of possible minimizers. This is because the domain $[0, 1]$ can be partitioned into intervals that depend on α_{b}^0, such that $\alpha_{\mathrm{b}}^1 \to F(\boldsymbol{\alpha}_{\mathrm{b}})$ is concave in each of these intervals.

By Theorem 2, finding $\texttt{mindisc}_\beta$ can be done by minimizing $F(\boldsymbol{\alpha}_{\mathrm{b}})$ over the set Pairs, which includes $O(|\mathcal{G}|^2)$ discrete solutions $V_0 \times V_1$, and $|\mathcal{G}^+|$ one-dimensional solution sets of the form $(v, r_g + q_g v)$ for some $g \in \mathcal{G}^+$. Define, for $g \in \mathcal{G}^+$,

$$H^g(v) := F(v, r_g + q_g v), \text{ and } V_g := \min_{v \in \mathrm{dom}_g} H^g(v).$$

Then, we have

$$\texttt{mindisc}_\beta = \min\{\min_{\boldsymbol{\alpha}_{\mathrm{b}} \in V_0 \times V_1} F(\boldsymbol{\alpha}_{\mathrm{b}}), \min_{g \in \mathcal{G}^+} V_g\}.$$

Finding V_g is a one-dimensional minimization problem. Due to the irregularity of H^g and its possibly large absolute derivatives, especially in unbalanced cases, obtaining a reasonable solution accuracy using a naive grid search could be prohibitively expensive. In the next section, we provide a procedure, termed `ArgMin1D`, that searches over a smaller number of values when minimizing H^g, making the procedure practical even in extremely unbalanced cases. `ArgMin1D` gets as input the region identifier g, the solution value tolerance γ, the discrepancy coefficient β, and the set of inputs `Inputs`. It returns the minimizer of H^g.

7.3 Fast Minimization of H^g

Fix $g \in \mathcal{G}^+$. We give the procedure `ArgMin1D` that calculating $\mathrm{argmin}_{v \in \mathrm{dom}_g} H^g(v)$ up to a given tolerance. Let dom_g^o be the open interval which is the interior of the closed interval dom_g. Note that H^g is continuous on dom_g^o. To minimize H^g on dom_g^o, it suffices to find a constant $L > 0$ such that H^g is L-Lipschitz on dom_g^o, that is,

$$\forall a, b \in \mathrm{dom}_g^o, |H^g(a) - H^g(b)| \le L|a - b|.$$

Then, to find an optimal solution up to a tolerance of γ, one can perform a grid search with a resolution of L/γ. However, H can have very large derivatives in small intervals, leading to a prohibitively large value of L. We therefore partition dom_z^o into separate intervals, and derive a separate Lipschitz upper bound for each interval. We show that the size of intervals with a large Lipschitz constant is small; Thus, the resulting search size is significantly smaller than L/γ.

H^g is a linear combination of minimizations over applications of τ_β. For $x \in [0, 1], g \in \mathcal{G}$, denote $[x]_g = (x, r_g + q_g x)$. Where H^g is differentiable, its derivative can be upper bounded by

$$\frac{dH^g(v)}{dv} \leq \sum_{z\in\mathcal{G}} w_z \max_{x\in S_z([v]_g)} \sum_{y\in\{0,1\}} \pi_z^y \left|\frac{\partial \tau_\beta([v]_g(y),[x]_z(y))}{\partial v}\right|. \tag{15}$$

$\tau_\beta(a,b)$ is differentiable except at $a\in\{0,1\}$, $b\in\{0,1\}$ and $a=b$. Based on this observation, we partition the search interval $[0,1]$ into sub-intervals such that all the τ_β terms in the sum are differentiable in the interior of each of the sub-intervals.

Then the Lipschitz constant L_I in sub-interval I is upper bounded by

$$L_I \leq \sum_{z\in\mathcal{G}} w_z \sum_{y\in\{0,1\}} \pi_z^y \max_{x\in S_z([v]_g)} \max_{v\in I} \left|\frac{\partial \tau_\beta([v]_g(y),[x]_z(y))}{\partial v}\right|.$$

Note that for $x\in S_z([v]_g)$, x is a linear transformation of v. Therefore, all the maximizations above are of the form

$$\max_{v\in I} \left|\frac{\partial \tau_\beta(a+bv, c+dv)}{\partial v}\right|,$$

for some known real values a,b,c,d. The following lemma provides an upper bound for this expression, depending on the values of a,b,c,d.

Lemma 3. *Let $b>0, c\leq 1, a\leq 1, d\geq 0$. Define $\psi(x) := \tau_\beta(a+bx, c+dx)$, defined over x such that $a+bx, c+dx\in[0,1]$. Then ψ is differentiable in the interior of its domain, except perhaps for x such that $a+bx=c+dx$. Define*

$$D(a,b,c,d) := \begin{cases} b/c & (d=0)\wedge(c>0) \\ d^2/|bc-ad| & (d>0)\wedge(a/b>c/d) \\ |bc-ad|/(a+b(c-a)/(b-d))^2 & (d>0)\wedge(a/b<c/d)\wedge(b>d) \\ 0 & \textit{otherwise.} \end{cases}$$

Then for any x in the interior of the domain of ψ such that $a+bx\neq c+dx$,

$$\left|\frac{d\psi(x)}{dx}\right| \leq \begin{cases} D(a,b,c,d) & (d>b\wedge x<\frac{c-a}{b-d})\vee(d<b\wedge x>\frac{c-a}{b-d}) \\ D(1-a-b,b,1-c-d,d)) & (d>b\wedge x>\frac{c-a}{b-d})\vee(d<b\wedge x<\frac{c-a}{b-d}) \\ \frac{d}{|a-c|} & d=b, a\neq c \\ 0 & \textit{otherwise.} \end{cases}$$

The proof is provided in Appendix B. The procedure `ArgMin1D` calculates an upper bound on L_I for each sub-interval I using the upper bound provided in Lemma 3, and then searches with a grid resolution of L_I/γ in the sub-interval. In addition, it searches each of the end points of all sub-intervals. This provides a minimizer up to a tolerance of γ using a limited number of search points compared to a naive fixed-resolution search.

8 Experiments

In this section, we report experiments demonstrating the various algorithms proposed in this work, as well as the diversity of possible applications of bounding fairness from aggregate data. The procedures proposed in this paper and the code for running the experiments are provided in https://github.com/sivansabato/bfa2.

We report three sets of experiments. In the first set of experiments, reported in Section 8.1, we use data sets for which labeled data is available and can be used for validation. For various classifiers, we calculate the value of $\texttt{mindisc}_\beta$ using our proposed procedure, and compare it to the true discrepancy $\texttt{disc}_\beta$ of the classifier, which we calculate using the labeled data. This provides an evaluation of the tightness of $\texttt{mindisc}_\beta$ as a lower bound for $\texttt{disc}_\beta$. Since $\texttt{mindisc}_\beta$ is calculated using only the limited information of aggregate statistics, we do not expect

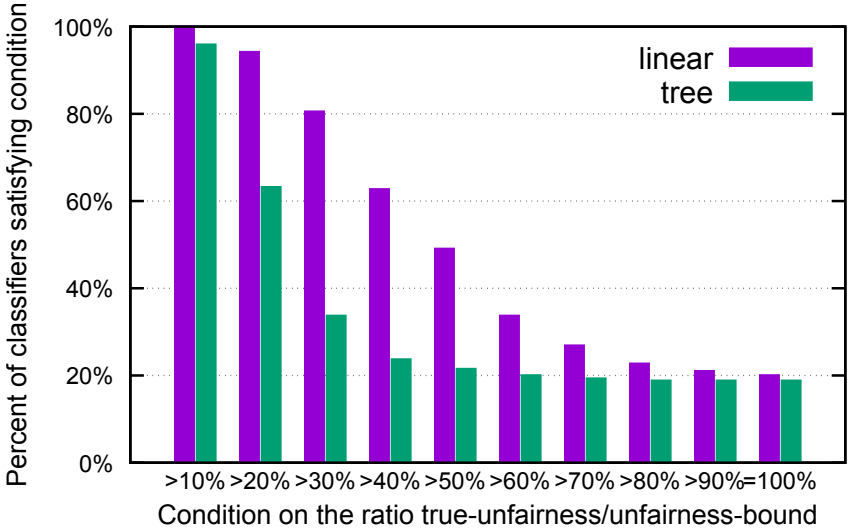


Fig. 4. US Census experiments. The figure shows the fraction of classifiers for which $\texttt{unfairness}/\texttt{mindisc}_1$ is larger than a given threshold. A ratio of 100% indicates that the lower bound is equal to the true unfairness. The median ratio for the linear method is 48% and 23% for the tree method. 97% of all classifiers obtained a ratio of at least 10%.

this lower-bound to always be tight. However, our experiments show that in many cases, it is quite close to the actual discrepancy.

In the second set of experiments, reported in Section 8.2, we demonstrate possible uses and outcomes of the calculation of $\texttt{mindisc}_\beta$ in various applications. We study several classifiers for which we only have aggregate statistics, calculate lower-bound Pareto curves of the trade-off between unfairness and error for each classifier, and discuss how these curves can help in decision making. In all the experiments below, we considered classification of US-based individuals, and the fairness attribute was the state of residence (or work) of the individual.

The experiments in Section 8.1 and Section 8.2 focus on classification problems in which there are many sub-populations with non-negligible probability mass. In these cases, the derived bounds are usually relatively tight. In contrast, in Section 8.3, we discuss and demonstrate phenomena that can occur when minimizing the discrepancy over distributions in which all or almost all of the probability mass is concentrated on only two sub-populations. While the bounds remain valid in this regime, we demonstrate that they can sometimes be very loose.

8.1 Comparing the Minimal Discrepancy to the Actual Discrepancy

For this comparison, we used the US Census (1990) data set (Dua and Graff 2019) to generate over 800 classifiers, on which we tested the tightness of the lower bound $\texttt{mindisc}_\beta$ by comparing it to the true discrepancy $\texttt{disc}_\beta$.

The US Census data set includes approximately 2.5 million records, with 124 attributes for each record. We studied fairness with respect to the "state of work" attribute, which specifies the state in which the individual worked. We used for the experiment the approximately 1.1 million records in which this attribute was present. We split this data into two halves at random, using one half as a training set to generate classifiers, and the other half as a test set to calculate the aggregate statistics of the classifier, as well as its true unfairness and error. For each attribute in the data set except for the state of work, and for each value v of the attribute, we generated a binary classification problem in which the examples are the records without this attribute and without the

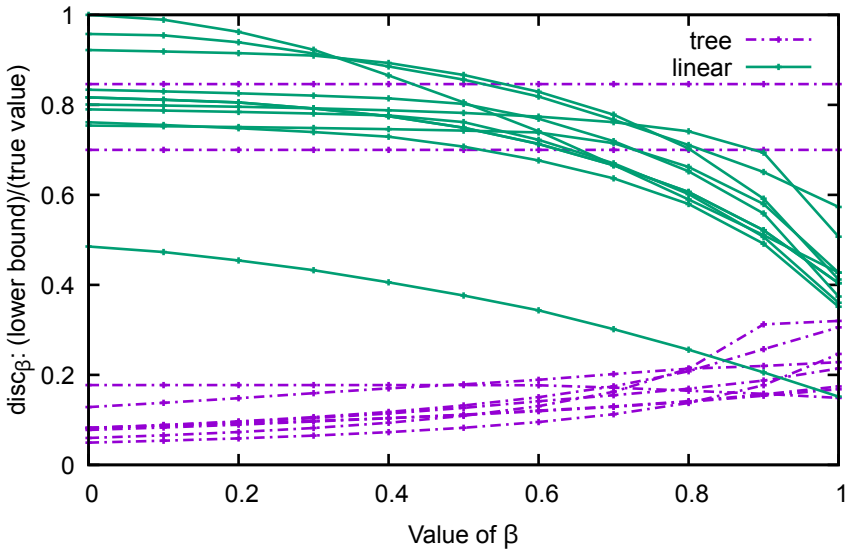


Fig. 5. US Census experiments. For 10 classifiers in each method, the ratio between $\texttt{disc}_\beta$ lower bound and the true $\texttt{disc}_\beta$, as a function of β. The figure shows that for linear classifiers $\texttt{mindisc}_\beta$ and $\texttt{disc}_\beta$ are usually more similar than they are for tree classifiers.

"state of work" attribute, and the label of a record was 1 if the attribute had value v. For attributes with more than 10 values, we binned the values of the attribute to 10 bins, and the label was 1 if the attribute had value at least v. If the resulting classification problem had more than 99% of the examples assigned to the same label, this classification problem was discarded. This process resulted in 410 classification problems. For each classification problem, we generated classifiers using two learning methods: a linear fit and a decision tree learning, using standard Matlab packages.

We ran Alg. 1 on Inputs, as calculated for each of these classifiers on the test set. The vector $\mathbf{w}$ that holds the fraction of the population in each state was also calculated based on the test set. First, we used Alg. 1 to calculate $\texttt{mindisc}_1$, which is a lower bound on $\texttt{disc}_1 \equiv \texttt{unfairness}$. We then calculated the ratio between the true unfairness (as calculated on the same test set) and the calculated $\texttt{mindisc}_1$. Figure 4 shows the fraction of the classifiers that achieved a ratio above a range of thresholds. For the linear method, the median ratio was 48%; it was 23% for the tree method. 97% of all classifiers obtained a ratio of at least 10%. We conjecture that the difference between the two learning methods might be due to the local nature of tree methods, which may make them inherently less fair (see a related discussion in (Valdivia et al. 2021)).

Next, we used Alg. 1 with a range of values of β, for 10 randomly selected classifiers. Figure 5 plots the ratio between the true $\texttt{disc}_\beta$ and our obtained lower bound for $\beta \in [0, 1]$ in each of the classifiers. Here too, it can be observed that for linear classifiers $\texttt{mindisc}_\beta$ and $\texttt{disc}_\beta$ are usually more similar than they are for tree classifiers.

8.2 Pareto Curves for Classifiers with Unknown Confusion Matrices

In the next set of experiments, we demonstrate possible uses and outcomes of the calculation of $\texttt{mindisc}_\beta$. We study several classifiers for which we only have aggregate statistics, and generate Pareto curves of the resulting lower bounds which provide the best possible trade-off between unfairness and error. We further demonstrate answering the types of questions in Section 7 using $\texttt{mindisc}_\beta$ and the resulting Pareto curves. The values in the

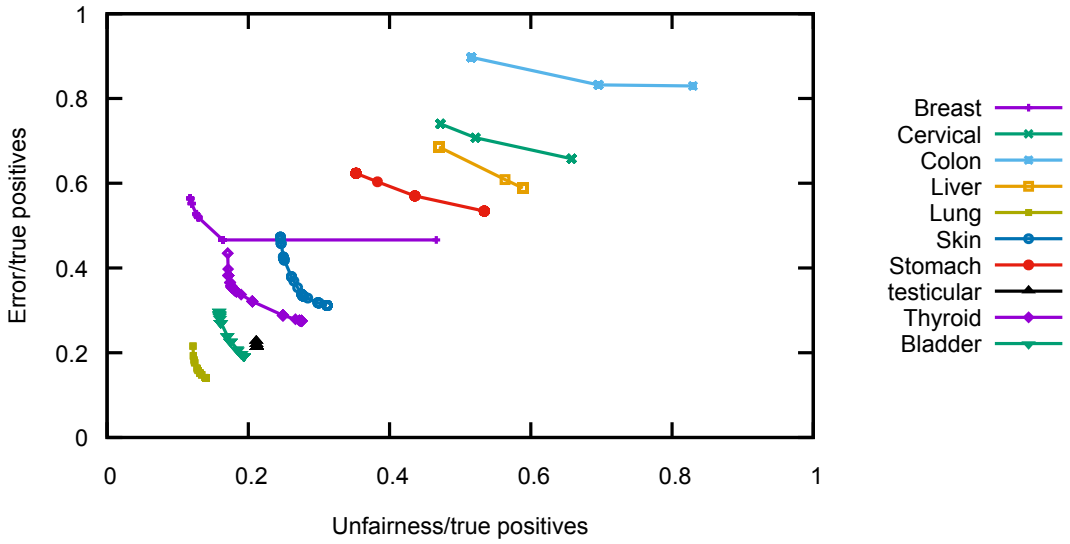


Fig. 6. Pareto curves for cancer by search queries. Each curve corresponds to a classifier for a specific cancer type, as listed in the plot legend. The Pareto curve of each classifier includes best-case pairs of $(\texttt{unfairness}, \texttt{error})$ that minimized $\texttt{disc}_\beta$ for each of the values of β that were tested for each classifier. To make the values of different cancer types comparable in the plot, all the values for a given classifier are normalized in the plot by the overall occurrence rate of the cancer type it predicts. Thus, values close to 1 indicate a poor classifier.

weight vector $\mathbf{w}$, which describe the fraction of the population in each state, were obtained from (US Census Bureau 2019).

In the first experiment, we consider a classification problem of identifying people diagnosed with a certain type of cancer from their queries in the Microsoft Bing search engine. The end goal was to identify possible participants for an anonymous patient study. We studied the 18 cancer types listed in (CDC and NCI. 2019), which reports their true-positive rate (incidence rate) in each state, giving the values in $\{\boldsymbol{\pi}_g\}_{g\in\mathcal{G}}$. For each cancer type, we constructed a classifier that predicted which anonymous US-based Bing users were likely ill with a specific cancer. The obtained classifier statistics are provided as part of the experiment code provided in https://github.com/sivansabato/bfa2. Note that we did not have individual validation data connecting users to a medically-validated diagnostic status. A user was predicted positive (having cancer) if they mentioned in their queries made between January 1st and June 30th, 2019 that they are suffering from a specific type of cancer. Following the methodology in (Hochberg et al. 2019; Shaklai et al. 2021; Yom-Tov et al. 2024), a user was predicted positive if they made at least one query that included the terms "I have [condition]" or "diagnosed with [condition]", where [condition] is one of the 18 specific cancer types. We excluded queries where the following phrases were included: "how do I know if I have [condition]", "do I have [condition]", "I think I have [condition]", "did I have [condition]" or the words "sister", "brother", "wife", "husband", "father", "mother", "dog", "cat".

For each classifier, we wish to discover a best-case accuracy-fairness trade-off, so as to identify classifiers that might be useful for a future, more detailed study. We note that unfairness may ensue using this classification method due to differences in health literacy between states. Out of the 18 cancer types, classifiers for 8 types had an overall positive prediction rate of more than twice or less than half of the true positive rate. We removed these classifiers from the experiment, since the final goal is classifiers that are both accurate and fair, and these classifiers were clearly very inaccurate. We note that when error is very large, the discrepancy is necessarily

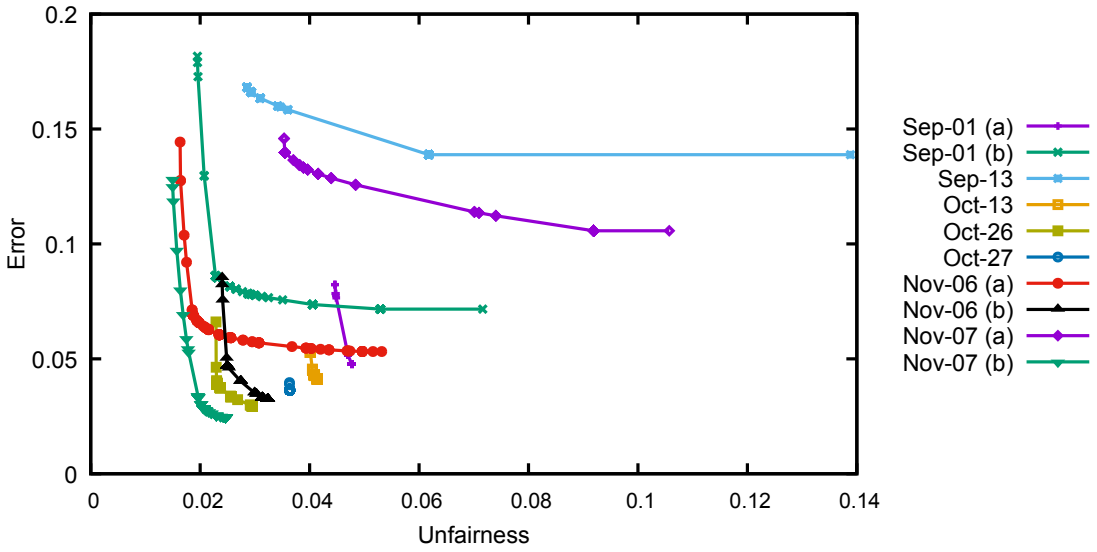


Fig. 7. Pareto curves for pre-election polls. Each curve corresponds to a classifier defined by a single poll, where the legend gives the date the poll was taken. The Pareto curve of each classifier includes best-case pairs of (unfairness, error) that minimized $\texttt{disc}_\beta$ for each of the values of β that were tested for each classifier.

large, and so lower-bounding it is uninformative. For the remaining 10 classifiers, we ran Alg. 1 for values of $\beta \in [0, 1]$.

We started with 10 evenly-spaced values for β, and added more values where the changes with respect to β were large. The pairs of (unfairness, error) that minimized $\texttt{disc}_\beta$ for each of the values of β for each classifier (cancer type) define a Pareto curve, representing the best-possible combinations for these classifiers. Figure 6 plots these curves. To make the values of different cancer types comparable in the plot, all the values for a given classifier are normalized in the plot by the overall occurrence rate of the cancer type it predicts. Thus, values close to 1 indicate a poor classifier. Recall that we do not have validation data these classifiers, thus we cannot compare to their true unfairness and error values. However, since $\texttt{mindisc}_\beta$ is a lower bound on $\texttt{disc}_\beta$, the true (unfairness, error) pair of each classifier is necessarily above its Pareto curve. We concluded from the results in Figure 6 that the classifier for lung cancer is the most promising for a future study, while many of the other classifiers are guaranteed to perform poorly on both measures.

The numerical values that generated the Pareto-curves in Figure 6 can further be used to answer many questions of the forms listed in Section 1. We demonstrate with a few examples below, based on the numerical results available at the linked github repository. All unfairness and error values below are normalized by the overall prevalence of true positives.

- "Is it possible that this classifier is fair? If not, how unfair must it be based on the provided label proportions?" This question can be answered by checking the leftmost point (minimal unfairness, $\beta = 1$) in the Pareto-curve. This is the minimal possible value of unfairness for this classifier, regardless of its true accuracy. For instance, the Breast Cancer classifier has a minimal possible unfairness value of 12%, and the Stomach Cancer classifier has a minimal possible unfairness value of 35%.
- "Suppose that the classifier is known to be nearly fair. How accurate can it be?". If it is known that the classifier's unfairness is smaller than some value, then the minimal possible error of the classifier is the

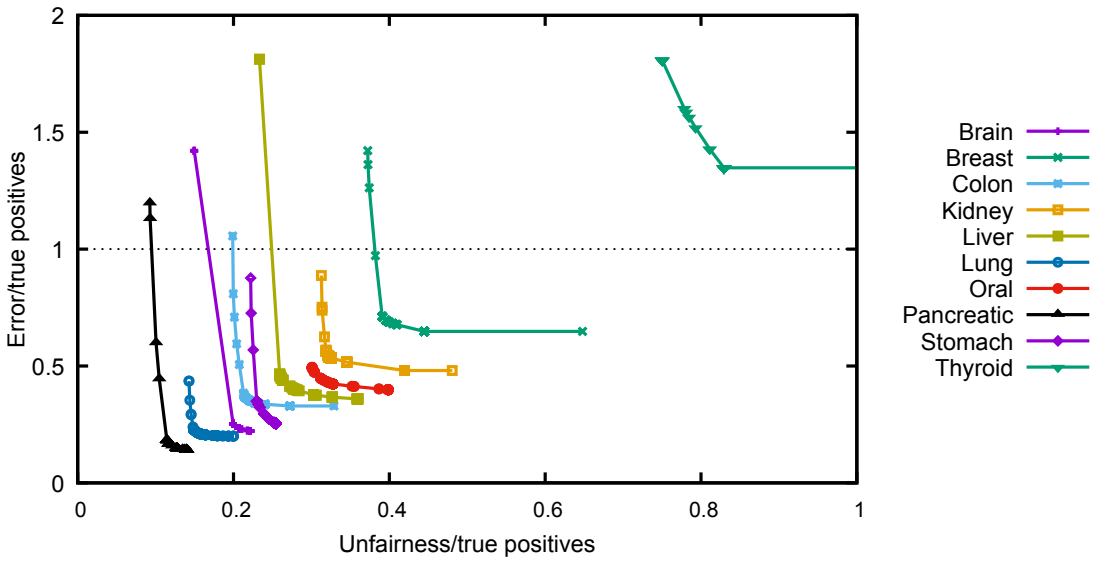


Fig. 8. Pareto curves for cancer diagnosis and mortality. Each curve corresponds to a classifier that would predict mortality of a diagnosed individual in each state with a probability equal to the overall mortality rate of the cancer type specified in the legend. The Pareto curve of each classifier includes best-case pairs of (`unfairness`, `error`) that minimized $\texttt{disc}_\beta$ for each of the values of β that were tested for each classifier. To make the values of different cancer types comparable in the plot, all the values for a given classifier are normalized in the plot by the overall occurrence rate of the cancer type it predicts. Thus, values close to 1 indicate a poor classifier.

smallest value of the `error` lower bound in the part of the Pareto curve that is to the left of this `unfairness` value. For instance, if the Thyroid Cancer classifier has `unfairness` $\leq$ %18, then its minimal possible `error` value is 35%, while for the Bladder Cancer classifier its minimal possible `error` value is 23%.

- "Suppose that the classifier is known to be quite accurate. How fair can it be?". If it is known that the classifier's `error` is smaller than some value, then the minimal possible unfairness of the classifier is the smallest value of the `unfairness` lower bound in the part of the Pareto curve that is below this `error` value. For instance, if the Thyroid Cancer classifier has `error` $\leq$ 35%, then its `unfairness` is at least 18%. For the Skin Cancer classifier, under the same upper bound on `error`, its `unfairness` is at least 28%.

For the last question, "Suppose that there is a penalty for each person who gets a wrong prediction, and for each person who is treated unfairly; what is the smallest possible overall cost of this classifier?", under any given trade-off between the cost of `unfairness` and the cost of `error`, the lowest possible cost of the classifier is $\texttt{mindisc}_\beta$, where β is set to the desired trade-off between the two quantities. For instance, setting $\beta = 3/4$ assigns `unfairness` to have triple the cost of `error`. Under this setting, the cost of the Breast Cancer classifier is $0.13 \cdot 0.75 + 0.53 \cdot 0.25 = 0.23$ and the cost of the Stomach Cancer classifier is $0.35 \cdot 0.75 + 0.62 \cdot 0.25 = 0.42$.

In the next experiment of this section, we study pre-election polls, and use them to provide the value of $\texttt{mindisc}_\beta$ for the classifiers that they might represent. Statistics on 10 pre-election polls of the 2016 US Presidential elections were obtained from (Five Thirty Eight 2016). The label was set to 1 if the individual voted for the Democratic candidate and to 0 otherwise. Each poll predicted a voting rate for the candidate in each state. The true positive rate was obtained from the actual results of the presidential elections in each state (Federal Elections Commission 2016). Treating each poll prediction as the aggregate statistics of an unknown classifier, we calculated the Pareto

curves representing the best-possible combinations of `unfairness` and `error` for these classifiers, using Alg. 1 and values of $\beta \in [0.01, 0.99]$, as in the experiment described above. The resulting Pareto curves, shown in Figure 7 by date of poll publication, can be used to compare the classifiers that could be underlying the different polls. For instance, although some of the more accurate polls are also the latest ones, this is not always the case. Moreover, some pairs of polls are incomparable, since they perform better in different parts of the curve. For instance, the Nov-06(b) poll has a lower error bound than the Nov-06(a) poll if its unfairness is larger than 0.05, but it cannot obtain unfairness smaller than 0.05, while the other poll can, hinting perhaps at a more biased polling methodology. This information can be used to further analyze the polling and prediction strategies employed in the various polls.

In the last experiment, we use the proposed method to explore the variation in cancer mortality rates across US states. Rates of cancer diagnosis and mortality in each US state, for 10 cancer types, were taken from the data published in (CDC and NCI. 2019). Note that in cases where data from some states was missing, we removed these states from the list of regions and renormalized the weights of the other states. The true occurrence rates $\{\boldsymbol{\pi}_g\}_{g\in\mathcal{G}}$ for each cancer type were set to the fraction of people in each state who died from the given cancer. We generated aggregate statistics for a classifier that would predict mortality of a diagnosed individual in each state with a probability equal to the *overall* mortality rate, which is the ratio between the overall mortality and the overall diagnosis rate. Thus, this classifier is based on the premise that in each state, the mortality rate is the same. Any deviation from this premise would mean that the classifier cannot be fully accurate or fair. In this context, a high `unfairness` value may be interpreted as a large fraction of individuals who have a non-typical manifestation of the disease or are treated differently by the health care system, while a high `error` may indicate a large fraction of the population whose mortality differs from the expected mortality, perhaps due to a difference in access to health services such as screening and care.

The Pareto curves that we calculated for these classifiers are shown in Figure 8; values for each cancer type are normalized by the occurrence rate of that cancer. It can be observed, for instance, that in several of the cancer types, allowing a small amount of `unfairness`, interpreted as modeling a fraction of the population as having non-typical variations of the disease, leads to a significantly smaller error, interpreted as a small difference in mortality rate between the states on the population with the typical variation of the disease. This use of our method sheds light on the diversity of possibilities of using such a tool in various ways for studying real-world phenomena.

We note that conclusions on unfairness can be combined with other available information to provide possible explanations for the unfairness. For instance, consider the value of the unfairness lower bound when minimizing the error. This is the rightmost point in the Pareto curve of the classifier. Assuming that the classifier is relatively accurate, this value can serve as a possible indication of its true unfairness. For the cancer diagnosis classifiers that were based on search queries (Figure 6), we compared this unfairness value to the average cost of care for each of the cancer types, as taken from (American Cancer Society 2019). As Figure 9 shows, there is a strong correlation ($R = 0.5$) between the cost of care and the unfairness value of the classifier. This may indicate that there are large differences in awareness and treatment between states when the cost of care is higher. As another example, we compared the unfairness of the cancer mortality classifiers (Figure 8), with the 5-year relative survival rate for each type of cancer, as taken from (American Cancer Society 2019). As Figure 10 shows, there is a very strong correlation ($R = 0.79$) between the survival rate and the unfairness of the mortality predictor. We speculate that this is because in cancers with higher survival rates, the influence of quality treatment on survival is significant. These two examples demonstrate how unfairness analysis may be used to inform exploratory studies of various social phenomena.

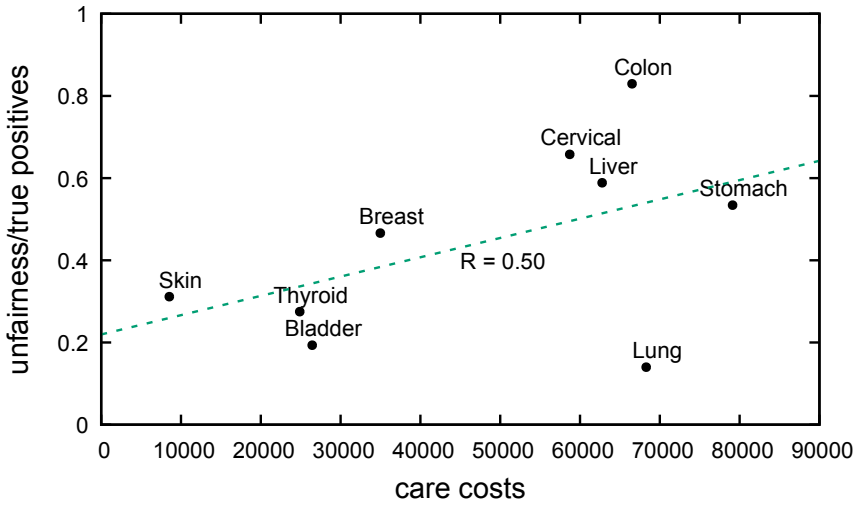


Fig. 9. The correlation between the (normalized) unfairness of the search-query cancer classifier and the cost of care. Each point in the plot corresponds to one type of cancer, as specified by the point's label. The x-axis shows the average annual per-patient cost of care for the cancer type in USD, based on (American Cancer Society 2019). The y-axis shows ratio between the unfairness value and the fraction of true-positives for the cancer type.

8.3 Minimizing the Discrepancy when there are Only Two Sub-Populations

In this section, we discuss and demonstrate two phenomena that can arise when minimizing the discrepancy over distributions in which all or almost all of the probability mass is concentrated on only two sub-populations. These phenomena can sometimes limit the usefulness of lower-bounding the discrepancy in these cases, since the resulting lower bound may be significantly lower than the true unfairness. This is a direct result of the small number of meaningful aggregate statistics that are available in this case. Nonetheless, any obtained lower bounds are still valid.

As discussed in Section 4, a fair classifier has a confusion matrix $\mathbf{A}$ which is the same for each of the sub-populations. A given set of `Inputs` statistics is consistent with a fair classifier if and only if there exists a matrix $\mathbf{A} \in \mathcal{M}_{2\times 2}$ such that $\mathbf{A}^T \boldsymbol{\pi}_g = \hat{\mathbf{p}}_g$ for each $g \in \mathcal{G}$. Equivalently, for each $g \in \mathcal{G}$, $\hat{p}_g^1 = \pi_g^0 \alpha_g^0 + \pi_g^1 (1 - \alpha_g^1)$. Thus, each sub-population induces one linear constraint. $\mathbf{A}$ has two degrees of freedoms. Therefore, if there are only two sub-populations, this problem is always feasible, unless the solution is outside of the legal range of $[0, 1]$. This means that in general, in problems in which almost all of the probability mass is concentrated on two sub-populations, it should be unsurprising if a classifier's minimal unfairness based on `Inputs` is found to be zero or close to zero.

A second phenomenon that can sometimes be observed in data sets with only two sub-populations, is the possible irrelevance of the value of β when minimizing $\texttt{mindisc}_\beta$. In cases where the same confusion matrix minimizes both the unfairness and the error, $\texttt{mindisc}_\beta$ would have the same solution regardless of β. This can be the case, for instance, if in all sub-populations, the frequency of positive predictions is larger than the frequency of true positives, or vice versa. Formally, consider a case of `Inputs` in which the following holds:

$$\forall g \in \mathcal{G}, \hat{p}_g^1 \geq \pi_g^1 \text{ or } \forall g \in \mathcal{G}, \hat{p}_g^1 \leq \pi_g^1. \tag{16}$$

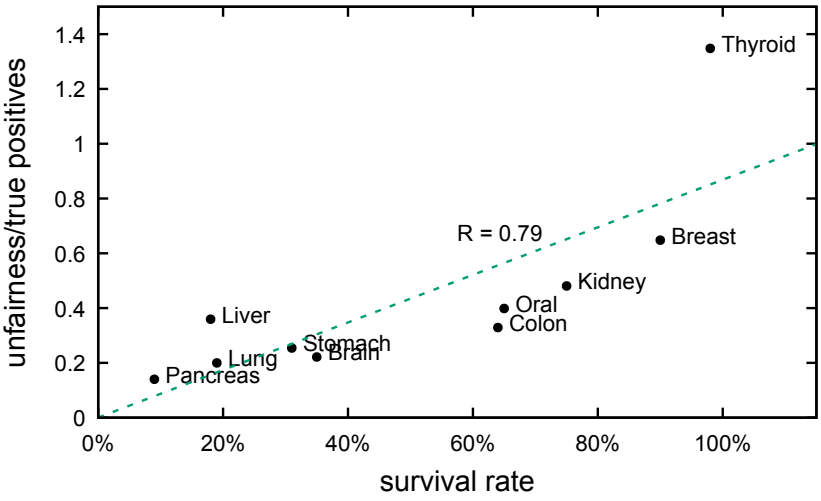


Fig. 10. The correlation between the (normalized) unfairness of the mortality prediction classifier and the survival rate of the cancer type. Each point in the plot corresponds to one type of cancer, as specified by the point's label. The x-axis shows 5-year relative survival rate the cancer type, based on (American Cancer Society 2019). The y-axis shows ratio between the unfairness value and the fraction of true-positives for the cancer type.

Note that Eq. (16) is much more likely when there is a small number of possible sub-populations. Consider a classifier that satisfies the former inequality of Eq. (16); the latter inequality is symmetric. In this case, the minimal error for the given `Inputs` is obtained by assuming that there are no false negatives in either of the sub-populations. More specifically, the error is minimized when assuming the following confusion matrix for each sub-population $g \in \mathcal{G}$:

$$\mathbf{A}_g = \begin{pmatrix} \frac{1-\pi_g^1}{\pi_g^0} & \frac{\hat{p}_g^1-\pi_g^1}{\pi_g^0} \\ 0 & 1 \end{pmatrix}.$$

Note that in this solution, the positive label does not induce any unfairness, since the second row is the same for all sub-populations. This may be the reason that when Eq. (16) holds, it is common to have no effect of β on the minimizing solution for $\texttt{mindisc}_\beta$.

We demonstrate the two phenomena described above with two data sets that have been widely used in the fairness literature. The first is the UCI Adult data set (Becker and Kohavi 1996). This data set is used in many algorithmic fairness works (e.g., (Calders et al. 2009; Kumar et al. 2023; Turgut 2023)). It includes data on over 48,000 individuals. For each individual, 14 demographic features are provided. The goal is to predict whether an individual's income is below or above $50,000. The protected attribute can be set to the sex or the race of the individual. The sex attribute has two possible values, matching the discussion on distributions with only two-sub-populations. The race attribute has 5 possible values. However, more than 96% of the examples have one of only two races, Black and White. Thus, this is also a relevant example for this discusssion.

The second data set is the Propublica COMPAS Recidivism data set (ProPublica 2016), also used in many fairness works (see, e.g., (Rudin et al. 2020; Zafar, Valera, Rogriguez, et al. 2017)). This data set includes demographic and historical data on over 5,000 defendants from Broward County, Florida, from 2013 and 2014. The label indicates

Table 2. Values of `Inputs` parameters for the four tested classifiers. The table uses the following shorthand: African-American, Cauc. for Caucasian, log-reg for logistic regression. The last column provides an experiment number, which is used in the tables below to refer to a specific experiment. It can be seen that in experiments I, II, III, V, $\forall g \in \mathcal{G}, \hat{p}_g^1 \leq \pi_g^1$. In contrast, this is not the case in experiments IV, VI. This distinction is relevant to the effect of β on the solution for $\texttt{mindisc}_\beta$, as we discuss in Section 8.3.

Data set	Classifier	Protected Attribute	Sub-population	w_g	π_g^1	$\hat{p}_g^1$	#Exp
Adult	vanilla k-NN	Sex	Female	33.30%	10.88%	8.45%	I
			Male	66.70%	29.98%	27.21%	
		Race	Amer-Indian-Eskimo	0.98%	11.95%	6.29%	II
			Asian-Pac-Islander	2.95%	27.71%	23.33%	
			Black	9.59%	11.47%	9.48%	
			Other	0.83%	18.52%	7.41%	
			White	85.66%	25.03%	22.47%	
	weighted k-NN	Sex	Female	33.30%	10.88%	11.20%	III
			Male	66.70%	29.98%	35.11%	
		Race	Amer-Indian-Eskimo	0.98%	11.95%	7.55%	IV
			Asian-Pac-Islander	2.95%	27.71%	27.92%	
			Black	9.59%	11.47%	13.45%	
			Other	0.83%	18.52%	14.81%	
			White	85.66%	25.03%	29.00%	
COMPAS	vanilla log-reg	Race	African-American	59.01%	54.57%	51.56%	V
			Caucasian	40.99%	37.96%	28.15%	
	weighted log-reg	Race	African-American	59.01%	54.57%	57.30%	VI
			Caucasian	40.99%	37.96%	32.91%	

whether these individual have re-offended within the next two years. The protected attribute in this data set is race, and only two races are represented in the data: African-American and Caucasian.

We generated two classifiers for each data set. For each data set, the first classifier was a simple classifier, which turned out to satisfy Eq. (16). To obtain also classifiers in which Eq. (16) does not hold, we generated a second classifier for each data set, in which the positive label was assigned a higher weight than its default weight based on the training data set. We report in Table 2 the values of `Inputs` for each valid combination of data set, classifier, and protected attribute.

The first classifier for the Adult data set was a vanilla k-Nearest-Neighbors classifier with 9 neighbors and a normalized Euclidean distance, as implemented by Matlab's `fitcknn` procedure. The first classifier for the COMPAS data set was a vanilla logistic regression classifier, as implemented by python's `sklearn` package. It can be seen in Table 2 that in both data sets, the vanilla classifier (#Exp I, II, V) satisfies Eq. (16). For the Adult data set, a weighted variant of the k-Nearest-Neighbor classifier was generated by initializing the classifier to use a different prior distribution of the labels, assigning 30% to positive labels and 70% to negative labels. For the COMPAS data set, a weighted logistic regression classifier was generated by setting the weight ratio of the positive labels to negative labels to 1.1. This has led to two classifiers that did not satisfy Eq. (16) (#Exp IV, VI) and one that did (#Exp III).

Table 3. Results of unfairness and error minimization for the experiments in Table 2. `error` is the value of $\texttt{mindisc}_0$ and `unfairness` is the value of $\texttt{mindisc}_1$. The last column indicates whether the value of β when calculating $\texttt{mindisc}_\beta$ has any effect on the resulting minimizing solution.

#Exp	`unfairness`	Min. unfairness	`error`	Min. error	Eq. (16) holds?	β affects solution?
I	5.10%	0.52%	16.18%	2.66%	Yes	No
II	4.79%	0.27%	16.18%	2.66%	Yes	No
III	7.49%	2.07%	17.33%	3.53%	Yes	Yes
IV	6.00%	0.52%	17.33%	3.67%	No	Yes
V	10.44%	3.34%	33.93%	5.80%	Yes	No
VI	10.91%	3.59%	33.98%	3.67%	No	Yes

Table 4. For those experiments in Table 3 in which β affects the solution, this table provides the values of the found solutions and the ranges of β in which each value was obtained.

#Exp	β range	solution unfairness	solution error
III	$0 - 0.78$	3.27%	3.53%
III	$0.79 - 1.0$	2.07%	7.66%
IV	$0.01 - 0.05$	3.67%	3.67%
IV	$0.06 - 0.1$	3.47%	3.68%
IV	$0.06 - 0.32$	2.16%	3.83%
IV	$0.33 - 1.0$	0.52%	4.57%
VI	$0 - 0.98$	3.67%	3.67%
VI	$0.99 - 1.00$	3.59%	6.72%

Table 3 reports the values of $\texttt{mindisc}_\beta$ that correspond to the minimal unfairness value and the minimal error value for each of the classifiers given `Inputs`, using Alg. 1. These values are compared to the true `unfairness` and `error`, reported in the same table. It can be seen that indeed, in some cases (#Exp I,II,IV) the minimal `unfairness` is close to zero although the true unfairness is not, demonstrating the first phenomenon discussed above. Nonetheless, note that in the other experiments, the ratio between the true unfairness and the lower bound is non-trivial, ranging between 3.1 and 3.6. Thus, in these cases minimizing the discrepancy is more informative.

The second phenomenon, in which the value of β does not affect the solution to $\texttt{mindisc}_\beta$, can also be observed in the last two columns of Table 3. It can be seen that β has no effect in 3 out of 4 experiments in which Eq. (16) holds, but does have an effect in all of the experiments in which Eq. (16) does not hold. For the experiments in which β had an effect on the solution, Table 4 reports the ranges of β that led to each of the identified solutions to $\texttt{mindisc}_\beta$. The small number of solutions across the range of possible β is again related to the small number of sub-populations in these experiments, which results in a small number of meaningful constraints on the minimization problem.

We conclude that in general, while calculating $\texttt{mindisc}_\beta$ is always possible, its value may be less informative in some distributions that are concentrated on two sub-populations.

9 Discussion

Machine learning models increasingly influence decisions in domains such as credit scoring, hiring, healthcare, and criminal justice. When these systems behave unfairly, whether by amplifying existing biases in the data, or by treating certain groups systematically worse, the resulting harms can be significant. These harms are often compounded by the opacity of machine learning models, making it difficult for affected individuals or external auditors to identify and contest unfair treatment.

Recent work on algorithmic fairness has developed a wide range of definitions and learning methods, with equalized odds emerging as a central group-fairness criterion and inspiring many approaches for training fair classifiers. Fairness auditing has also grown into a substantial research area, with methods that typically assume access to individual-level data, protected attributes, and often the classifier itself. When any of these components are missing, existing techniques generally rely on imputing protected attributes or leveraging individual covariates. Our work goes beyond these techniques by showing how fairness and performance can be quantified without assuming access to individual-level attributes or full model transparency, complementing research on the trade-offs between accuracy and fairness and extending the scope of fairness auditing to a more restrictive and practically relevant regime.

In this work, we showed that aggregate statistics on classifiers can be used to draw useful conclusions about the fairness and the accuracy of these classifiers. We defined the new `unfairness` measure, which facilitates a quantifiable study of classifiers that are not completely fair, and provided several procedures for bounding unfairness, error and the trade-offs between them. Our experiments demonstrate how such analyses can be useful in various areas of study. In addition, such analyses can help in classifier design, when individual labeled data is not available. The `unfairness` measure is constructed so as to be easily interpretable: as discussed in Section 5, its value indicates what fraction of the population is treated unfairly (that is, differently from the baseline). Therefore, practitioners can use this measure to identify the size of the unfairly-treated population when using a given classifier, informing reporting, compliance auditing and decision making such as classifier choice. Moreover, as discussed in Section 7, by assigning costs to treating any one individual unfairly and to mis-classifying any one individual, and combining them using a coefficient based on the application-specific cost ratio, the discrepancy of a classifier becomes equivalent to its cost. Therefore, the bounds on the discrepancy that can be obtained using our methods provide information about the cost of a classifier. This can be used for decision making, for instance to decide whether a given classifier should be used in cost-sensitive environments.

Our experiments in Section 8.2 show specific cases where harms imposed by classifiers can be detected through aggregate statistics. In our analysis of cancer detection data, we compared state-level statistics with the number of search engine users who were labeled as having different types of cancer via a simple algorithm. The results of our analysis allow us to test whether, for example, users in specific states are less likely to be identified as suffering from a specific cancer. If, for example, companies use this model to help guide users to treatment (e.g., (Yom-Tov 2020)), the implication is that users in some states might be discriminated against for various reasons. The proposed algorithm provides a direct measure of this unfairness, allowing officials to take steps to mitigate this unfairness.

Our analysis of election data shows how to identify where election polls fail to correctly represent the will of people. This is of societal importance, because when polls misrepresent people's voting patterns, they can distort the perceived legitimacy of electoral outcomes and undermine confidence in democratic institutions. Moreover, voting polls can serve to guide the priorities of candidates in campaigning. Therefore, misleading polls can affect the final results of elections, and if they are misleading in some locations more than others, then this is also an issue of equity. Identifying where and why such failures occur is therefore essential to ensuring that all segments of society are equally visible in the democratic process.

Thus, as classifiers become more abundant and influential in our daily lives, studying their properties from limited information becomes more and more important. This work shows that this is possible and can provide meaningful information. We expect that future work will expand on this vision and extend this approach to other performance metrics and other types of classification tasks.

Limitations. In this work, we studied unfairness with respect to the Equalized Odds fairness criterion. Other fairness measures, such as Equal Opportunity and Demographic Parity, require a separate treatment, since they would induce a different set of constraints on population statistics, which would require other minimization algorithms. Any use of the methods provided here should take into account the application and the appropriateness of Equalized Odds and its generalization to `unfairness`, with respect to the desired fairness properties of the specific application. The accuracy of the methods in this work depends on the accuracy of the probability estimates of the statistics used by the methods. If these are inaccurate, a commensurate inaccuracy is expected in the output of the methods. The definitions and methods in this work are designed for binary classifiers, in which each individual is classified into one of two categories. A follow-up work (Sabato, Treister, et al. 2025) generalizes the measure to multiclass classifiers and proposes relevant computational approaches.

Acknowledgments

An earlier short version of this work appeared as (Sabato and Yom-Tov 2020).

We acknowledge the support of the Natural Sciences and Engineering Research Council of Canada (NSERC), [funding reference number RGPIN-2024-05907]. Resources used in preparing this research were provided, in part, by the Province of Ontario, the Government of Canada through CIFAR, and companies sponsoring the Vector Institute; see https://vectorinstitute.ai/partnerships/current-partners/.

Declarations

The authors were previously employed by Microsoft, owner of Bing. No specific funding was provided for this work. The authors have no current relevant financial or non-financial interests to disclose.

The code and data to support the paper are in the repository cited in the Abstract.

References

W. Alghamdi, S. Asoodeh, H. Wang, F. P. Calmon, D. Wei, and K. N. Ramamurthy. 2020. "Model Projection: Theory and Applications to Fair Machine Learning." In: *2020 IEEE International Symposium on Information Theory (ISIT)*, 2711–2716. doi:10.1109/ISIT44484.2020.9173988.

American Cancer Society. 2019. *Cancer facts and statistics, table 8.* (2019). https://www.cancer.org/content/dam/cancer-org/research/cancer-facts-and-statistics/annual-cancer-facts-and-figures/2019/cancer-facts-and-figures-2019.pdf.

C. Ashurst and A. Weller. 2023. "Fairness without demographic data: A survey of approaches." In: *Proceedings of the 3rd ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–12.

P. Awasthi, A. Beutel, M. Kleindessner, J. Morgenstern, and X. Wang. 2021. "Evaluating fairness of machine learning models under uncertain and incomplete information." In: *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, 206–214.

S. Barocas, M. Hardt, and A. Narayanan. 2017. "Fairness in machine learning." *NIPS Tutorial.*

B. Becker and R. Kohavi. 1996. *Adult.* UCI Machine Learning Repository. DOI: https://doi.org/10.24432/C5XW20. (1996).

A. Bell, L. Bynum, N. Drushchak, T. Zakharchenko, L. Rosenblatt, and J. Stoyanovich. 2023. "The possibility of fairness: Revisiting the impossibility theorem in practice." In: *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 400–422.

R. K. Bellamy et al.. 2019. "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias." *IBM Journal of Research and Development*, 63, 4/5, 4–1.

R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth. 2018. "Fairness in criminal justice risk assessments: The state of the art." *Sociological Methods & Research*, 0049124118782533.

E. Black, H. Elzayn, A. Chouldechova, J. Goldin, and D. Ho. 2022. "Algorithmic Fairness and Vertical Equity: Income Fairness with IRS Tax Audit Models." In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '22). Association for Computing Machinery, Seoul, Republic of Korea, 1479–1503. ISBN: 9781450393522.

E. Black, S. Yeom, and M. Fredrikson. 2019. "FlipTest: Fairness Auditing via Optimal Transport." *CoRR*, abs/1906.09218. http://arxiv.org/abs/1906.09218 arXiv: 1906.09218.

T. Calders, F. Kamiran, and M. Pechenizkiy. 2009. "Building classifiers with independency constraints." In: *2009 IEEE international conference on data mining workshops*. IEEE, 13–18.

F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney. 2017a. "Optimized Pre-Processing for Discrimination Prevention." In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 3992–4001. http://papers.nips.cc/paper/6988-optimized-pre-processing-for-discrimination-prevention.pdf.

F. Calmon, D. Wei, B. Vinzamuri, K. Natesan Ramamurthy, and K. R. Varshney. 2017b. "Optimized Pre-Processing for Discrimination Prevention." In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 3992–4001.

CDC and NCI.. 2019. *United States Cancer Statistics: Data Visualizations*. (2019). https://gis.cdc.gov/Cancer/USCS/DataViz.html.

J. Chen, N. Kallus, X. Mao, G. Svacha, and M. Udell. 2019. "Fairness under unawareness: Assessing disparity when protected class is unobserved." In: *Proceedings of the conference on fairness, accountability, and transparency*, 339–348.

G. Cornacchia, V. W. Anelli, G. M. Biancofiore, F. Narducci, C. Pomo, A. Ragone, and E. Di Sciascio. 2023. "Auditing fairness under unawareness through counterfactual reasoning." *Information Processing & Management*, 60, 2, 103224.

M. Donini, L. Oneto, S. Ben-David, J. S. Shawe-Taylor, and M. Pontil. 2018. "Empirical Risk Minimization Under Fairness Constraints." In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett. Curran Associates, Inc., 2791–2801.

D. Dua and C. Graff. 2019. *UCI Machine Learning Repository*. (2019). http://archive.ics.uci.edu/ml.

C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel. 2012. "Fairness through awareness." In: *Proceedings of the 3rd innovations in theoretical computer science conference*. ACM, 214–226.

A. Fabris, A. Esuli, A. Moreo, and F. Sebastiani. 2023. "Measuring fairness under unawareness of sensitive attributes: A quantification-based approach." *Journal of Artificial Intelligence Research*, 76, 1117–1180.

Federal Elections Commission. 2016. *Federal Elections 2016*. (2016). https://transition.fec.gov/pubrec/fe2016/federalelections2016.pdf.

M. Feldman, S. A. Friedler, J. Moeller, C. Scheidegger, and S. Venkatasubramanian. 2015. "Certifying and removing disparate impact." In: *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 259–268.

Five Thirty Eight. 2016. *National Presidential Polls, November 8th, 2016*. (2016). https://projects.fivethirtyeight.com/2016-election-forecast/national-polls/.

A. Ghosh, P. Kvitca, and C. Wilson. 2023. "When fair classification meets noisy protected attributes." In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, 679–690.

N. Goel, M. Yaghini, and B. Faltings. 2018. "Non-discriminatory machine learning through convex fairness criteria." In: *Thirty-Second AAAI Conference on Artificial Intelligence*.

G. Goh, A. Cotter, M. Gupta, and M. P. Friedlander. 2016. "Satisfying Real-world Goals with Dataset Constraints." In: *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, 2415–2423.

P. Goyal, A. R. Soriano, C. Hazirbas, L. Sagun, and N. Usunier. 2022. "Fairness Indicators for Systematic Assessments of Visual Feature Extractors." In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '22). Association for Computing Machinery, Seoul, Republic of Korea, 70–88. ISBN: 9781450393522.

N. Grgic-Hlaca, M. B. Zafar, K. P. Gummadi, and A. Weller. 2016. "The case for process fairness in learning: Feature selection for fair decision making." In: *NIPS Symposium on Machine Learning and the Law*. Vol. 1, 2.

M. Hardt, E. Price, and N. Srebro. 2016. "Equality of opportunity in supervised learning." In: *Advances in neural information processing systems*, 3315–3323.

H. Heidari, C. Ferrari, K. Gummadi, and A. Krause. 2018. "Fairness Behind a Veil of Ignorance: A Welfare Analysis for Automated Decision Making." In: *Advances in Neural Information Processing Systems 31*. Ed. by S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, 1265–1276.

I. Hochberg, D. Daoud, N. Shehadeh, and E. Yom-Tov. 2019. "Can internet search engine queries be used to diagnose diabetes? Analysis of archival search data." *Acta Diabetologica*, 56, 1149–1154.

C. Jackson, N. Best, and S. Richardson. 2008. "Hierarchical related regression for combining aggregate and individual data in studies of socio-economic disease risk factors." *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 171, 1, 159–178.

C. Jackson, N. Best, and S. Richardson. 2006. "Improving ecological inference using individual-level data." *Statistics in medicine*, 25, 12, 2136–2159.

J. E. Johndrow and K. Lum. 2019. "An algorithm for removing sensitive information: application to race-independent recidivism prediction." *The Annals of Applied Statistics*, 13, 1, 189–220.

S. Jung, D. Lee, T. Park, and T. Moon. 2021. "Fair feature distillation for visual recognition." In: *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 12115–12124.

N. Kallus, X. Mao, and A. Zhou. 2022. "Assessing algorithmic fairness with unobserved protected class using data combination." *Management Science*, 68, 3, 1959–1981.

Ö. Kırnap, F. Diaz, A. Biega, M. Ekstrand, B. Carterette, and E. Yilmaz. 2021. "Estimation of fair ranking metrics with incomplete judgments." In: *Proceedings of the Web Conference 2021*, 1065–1075.

J. Kleinberg, S. Mullainathan, and M. Raghavan. 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.

D. Kumar, T. Grosz, N. Rekabsaz, E. Greif, and M. Schedl. 2023. "Fairness of recommender systems in the recruitment domain: an analysis from technical and legal perspectives." *Frontiers in big Data*, 6.

M. J. Kusner, J. Loftus, C. Russell, and R. Silva. 2017. "Counterfactual Fairness." In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, 4066–4076.

A. Lamy, Z. Zhong, A. K. Menon, and N. Verma. 2019. "Noise-tolerant fair classification." In: *Advances in Neural Information Processing Systems 32*, 294–306. http://papers.nips.cc/paper/8322-noise-tolerant-fair-classification.pdf.

D. McDuff, S. Ma, Y. Song, and A. Kapoor. 2019. "Characterizing Bias in Classifiers using Generative Models." In: *Advances in Neural Information Processing Systems 32*, 5404–5415.

A. K. Menon and R. C. Williamson. 2018. "The cost of fairness in binary classification." In: *Conference on Fairness, Accountability and Transparency*, 107–118.

P. Nandy, C. Diciccio, D. Venugopalan, H. Logan, K. Basu, and N. El Karoui. 2022. "Achieving Fairness via Post-Processing in Web-Scale Recommender Systems." In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, 715–725.

G. Pleiss, M. Raghavan, F. Wu, J. Kleinberg, and K. Q. Weinberger. 2017. "On fairness and calibration." In: *Advances in Neural Information Processing Systems*, 5680–5689.

ProPublica. 2016. *COMPAS Analysis GitHub Repository*. https://github.com/propublica/compas-analysis. (2016).

S. Radovanović and M. Ivić. 2023. "Enabling Equal Opportunity in Logistic Regression Algorithm." *Management: Journal of Sustainable Business and Management Solutions in Emerging Economies*, 28, 2, 55–66.

J. Roth, G. Saint-Jacques, and Y. Yu. 2022. "An Outcome Test of Discrimination for Ranked Lists." In: *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency* (FAccT '22). Association for Computing Machinery, Seoul, Republic of Korea, 350–356. ISBN: 9781450393522.

C. Rudin, C. Wang, and B. Coker. 2020. "The age of secrecy and unfairness in recidivism prediction." *Harvard Data Science Review*, 2, 1, 1.

S. Sabato, E. Treister, and E. Yom-Tov. July 2025. "Disparate Conditional Prediction in Multiclass Classifiers." In: *Proceedings of the 42nd International Conference on Machine Learning* (Proceedings of Machine Learning Research). Ed. by A. Singh, M. Fazel, D. Hsu, S. Lacoste-Julien, F. Berkenkamp, T. Maharaj, K. Wagstaff, and J. Zhu. Vol. 267. PMLR, (July 2025), 52508–52525.

S. Sabato and E. Yom-Tov. July 2020. "Bounding the fairness and accuracy of classifiers from population statistics." In: *Proceedings of the 37th International Conference on Machine Learning* (Proceedings of Machine Learning Research). Ed. by H. D. III and A. Singh. Vol. 119. PMLR, (July 2020), 8316–8325.

C. Scott, G. Blanchard, and G. Handy. June 2013. "Classification with Asymmetric Label Noise: Consistency and Maximal Denoising." In: *Proceedings of the 26th Annual Conference on Learning Theory* (Proceedings of Machine Learning Research). Ed. by S. Shalev-Shwartz and I. Steinwart. Vol. 30. PMLR, Princeton, NJ, USA, (June 2013), 489–511.

S. Shaklai, R. Gilad-Bachrach, E. Yom-Tov, and N. Stern. 2021. "Detecting impending stroke from cognitive traits evident in internet searches: analysis of archival data." *Journal of Medical Internet Research*, 23, 5, e27084.

T. Speicher, H. Heidari, N. Grgic-Hlaca, K. P. Gummadi, A. Singla, A. Weller, and M. B. Zafar. 2018. "A unified approach to quantifying algorithmic unfairness: Measuring individual &group unfairness via inequality indices." In: *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 2239–2248.

T. Sun, D. Sheldon, and B. O'Connor. Nov. 2017. "A Probabilistic Approach for Learning with Label Proportions Applied to the US Presidential Election." In: *2017 IEEE International Conference on Data Mining (ICDM)*. (Nov. 2017), 445–454. doi:10.1109/ICDM.2017.54.

O. Turgut. 2023. "Fair Classification with Ensembles." In: *2023 Innovations in Intelligent Systems and Applications Conference (ASYU)*. IEEE, 1–4.

US Census Bureau. 2019. *Annual Estimates of the Resident Population for the United States, Regions, States, and Puerto Rico: April 1, 2010 to July 1, 2019*. (2019). https://www2.census.gov/programs-surveys/popest/tables/2010-2019/state/totals/nst-est2019-01.xlsx.

A. Valdivia, J. Sánchez-Monedero, and J. Casillas. 2021. "How fair can we go in machine learning? Assessing the boundaries of accuracy and fairness." *International Journal of Intelligent Systems*, 36, 4, 1619–1643. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/int.22354. doi:https://doi.org/10.1002/int.22354.

S. Verma and J. Rubin. 2018. "Fairness definitions explained." In: *2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*. IEEE, 1–7.

H. Wang, L. He, R. Gao, and F. Calmon. 2024. "Aleatoric and epistemic discrimination: Fundamental limits of fairness interventions." *Advances in Neural Information Processing Systems*, 36.

Y. Wang, D. Sridhar, and D. Blei. 2023. "Adjusting Machine Learning Decisions for Equal Opportunity and Counterfactual Fairness." *Transactions on Machine Learning Research*.

B. Woodworth, S. Gunasekar, M. I. Ohannessian, and N. Srebro. July 2017. "Learning Non-Discriminatory Predictors." In: *Proceedings of the 2017 Conference on Learning Theory* (Proceedings of Machine Learning Research). Ed. by S. Kale and O. Shamir. Vol. 65. PMLR, Amsterdam, Netherlands, (July 2017), 1920–1953.

Y. Wu, L. Zhang, and X. Wu. 2019. "On Convexity and Bounds of Fairness-aware Classification." In: *The World Wide Web Conference*. ACM, 3356–3362.

E. Yom-Tov. 2020. "Screening for cancer using a learning internet advertising system." *ACM Transactions on Computing for Healthcare*, 1, 2, 1–13.

E. Yom-Tov, I. Navar, E. Fraenkel, and J. D. Berry. 2024. "Identifying amyotrophic lateral sclerosis through interactions with an internet search engine." *Muscle & Nerve*, 69, 1, 40–47.

Z. Yu, J. Chakraborty, and T. Menzies. 2024. "FairBalance: How to Achieve Equalized Odds With Data Pre-processing." *IEEE Transactions on Software Engineering*.

M. B. Zafar, I. Valera, M. Gomez Rodriguez, and K. P. Gummadi. 2017. "Fairness beyond disparate treatment & disparate impact: Learning classification without disparate mistreatment." In: *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, 1171–1180.

M. B. Zafar, I. Valera, M. G. Rogriguez, and K. P. Gummadi. 2017. "Fairness Constraints: Mechanisms for Fair Classification." In: *Artificial Intelligence and Statistics*, 962–970.

R. Zemel, Y. Wu, K. Swersky, T. Pitassi, and C. Dwork. June 2013. "Learning Fair Representations." In: *Proceedings of the 30th International Conference on Machine Learning* (Proceedings of Machine Learning Research). Ed. by S. Dasgupta and D. McAllester. Vol. 28(3). PMLR. Atlanta, Georgia, USA, (June 2013), 325–333.

A Proof of Theorem 2

In this section, we prove Theorem 2. We start with the observation that each of the functions $s_g^i(\boldsymbol{\alpha}_\mathrm{b})$ defined in Lemma 1 is continuous in $\boldsymbol{\alpha}_\mathrm{b}$. Thus, F can be redefined using continuous functions. Define

$$\forall g \in \mathcal{G}^+, i \in [4], \quad f_g^i(\boldsymbol{\alpha}_\mathrm{b}) := \begin{cases} \pi_g^0 \tau_\beta(\alpha_\mathrm{b}^0, s_g^i(\boldsymbol{\alpha}_\mathrm{b})) + \pi_g^1 \tau_\beta(\alpha_\mathrm{b}^1, r_g + q_g s_g^i(\boldsymbol{\alpha}_\mathrm{b})) & s_g^i(\boldsymbol{\alpha}_\mathrm{b}) \in \mathrm{dom}_g, \\ \infty & \text{otherwise.} \end{cases}$$

In addition, for $g \in \mathcal{G} \setminus \mathcal{G}^+$, define

$$f_g^i(\boldsymbol{\alpha}_\mathrm{b}) := \begin{cases} \sum_{y:\pi_g^y=1} \tau_\beta(\alpha_\mathrm{b}^y, 1 - \hat{p}_g^y) & i = 1 \\ \infty & \text{otherwise.} \end{cases}$$

Then, it easily follows that $F(\boldsymbol{\alpha}_\mathrm{b}) = \sum_{g \in \mathcal{G}} w_g \min_{i \in [4]} f_g^i(\boldsymbol{\alpha}_\mathrm{b})$.

The following lemma shows that each of the functions f_g^i is concave in α_b^1 on an appropriate segmentation of its domain. A technical detail is that in some cases, $f_g^i(\boldsymbol{\alpha}_\mathrm{b})$ might not be continuous in α_b^1 at $\alpha_\mathrm{b}^1 = 0$ or at $\alpha_\mathrm{b}^1 = 1$. This is because the function η is not continuous in its first argument in specific cases: For $b \in \{0, 1\}$, we have $1 = \lim_{a \to b} \eta(a, b) \neq \eta(b, b) = 0$. Thus, in some cases the concavity holds on a restricted domain, which does not include 0 or 1.

Lemma 4. *Fix $\alpha_b^0 \in [0, 1]$. For $g \in \mathcal{G}$, $i \in [4]$, consider the function $\alpha_b^1 \mapsto f_g^i(\alpha_b^0, \alpha_b^1)$ and the domain I_g^i on which it is finite. Define $\theta_g^i(\alpha_b^0)$ as follows:*

$$\theta_g^i(\alpha_b^0) = \begin{cases} \max\{r_g, 0\} & i = 1, g \in \mathcal{G}^+; \\ \min\{r_g + q_g, 1\} & i = 2, g \in \mathcal{G}^+; \\ r_g + q_g \alpha_b^0 & i \in \{3, 4\}, g \in \mathcal{G}^+, \alpha_b^0 \in \mathrm{dom}_g; \\ 1 - \hat{p}_g^1 & i = 1, g \in \mathcal{G}, \pi_g^1 = 1; \\ 0 & \textit{otherwise.} \end{cases}$$

Then $\alpha_b^1 \mapsto f_g^i(\alpha_b^0, \alpha_b^1)$ is concave on each of the intervals $(0, \theta_g^i(\alpha_b^0)] \cap I_g^i$ and $[\theta_g^i(\alpha_b^0), 1) \cap I_g^i$. Moreover, if $\theta_g^i(\alpha_b^0) \in (0, 1)$, then this holds also for the closed intervals.

Proof. For given $g \in \mathcal{G}, i \in [4]$, define $v^0 := s_g^i(\boldsymbol{\alpha}_\mathrm{b})$, and $v^1 := r_g + q_g v^0$. First, let $i \in [3]$, $g \in \mathcal{G}^+$, and assume $s_g^i(\alpha_\mathrm{b}^0) \in \mathrm{dom}_g$. In this case, $f_g^i(\boldsymbol{\alpha}_\mathrm{b}) = \pi_g^0 \tau_\beta(\alpha_\mathrm{b}^0, v^0) + \pi_g^1 \tau_\beta(\alpha_\mathrm{b}^1, v^1)$. Observe that s_g^1, s_g^2, s_g^3 are constant in α_b^1, thus v^0 does not depend on α_b^1, and the same holds for $\tau_\beta(\alpha_\mathrm{b}^0, v^0)$. Therefore, from the definition of τ_β, the function $\alpha_\mathrm{b}^1 \mapsto f_g^i(\boldsymbol{\alpha}_\mathrm{b})$ is an affine function of $\eta(\alpha_\mathrm{b}^1, v^1)$ (with non-negative coefficients). Note that v^1 does not depend on α_b^1. It is easy to verify that for any fixed $v^1 \in (0, 1)$, $\alpha_\mathrm{b}^1 \mapsto \eta(\alpha_\mathrm{b}^1, v^1)$ is concave on the interval $[0, v^1]$ and on the interval $[v^1, 1]$ (see illustration in Figure 2 (right)). Therefore, $\alpha_\mathrm{b}^1 \mapsto f_g^i(\boldsymbol{\alpha}_\mathrm{b})$ is also concave on the same intervals. Substituting v^1 with its definition, we get $\theta_g^i(\alpha_\mathrm{b}^0)$ as defined above. If $v^1 \in \{0, 1\}$ then the function is concave on $(0, v^1]$ and on $[v^1, 1)$.

Next, consider $i = 4$, $g \in \mathcal{G}^+$. Note that we have $I_g^4 = [r_g, q_g + r_g] \cap [0, 1]$, since $\alpha_\mathrm{b}^1 \in I_g^4$ iff $v^0 \equiv s_g^4(\alpha_\mathrm{b}^1) \in \mathrm{dom}_g$. In this case, we have $v^1 = \alpha_\mathrm{b}^1$, thus $\eta(\alpha_\mathrm{b}^1, v^1) = 0$ and so $\tau_\beta(\alpha_\mathrm{b}^1, v^1)$ is linear in α_b^1. In addition, $\tau_\beta(\alpha_\mathrm{b}^0, v^0)$ is an affine function of $\eta(\alpha_\mathrm{b}^0, v^0) = \eta(\alpha_\mathrm{b}^0, (\alpha_\mathrm{b}^1 - r_g)/q_g)$. This function is linear in α_b^1 on each of the intervals $[0, r_g + q_g \alpha_\mathrm{b}^0] \cap I_g^i$ and $[r_g + q_g \alpha_\mathrm{b}^0, 1] \cap I_g^i$. Thus, $\alpha_\mathrm{b}^1 \mapsto f_g^4(\boldsymbol{\alpha}_\mathrm{b})$ is linear on each of these intervals.

Lastly, consider $g \in \mathcal{G} \setminus \mathcal{G}^+$. In this case, only $i = 1$ is finite. If $\pi_g^0 = 1$, then $\alpha_\mathrm{b}^1 \mapsto f_g^1(\boldsymbol{\alpha}_\mathrm{b})$ is constant. If $\pi_g^1 = 1$, then f_g^1 is an affine function with positive coefficients of $\eta(\alpha_\mathrm{b}^1, 1 - \hat{p}_g^y)$. Similarly to the first case, this implies that $\alpha_\mathrm{b}^1 \mapsto f_g^i(\boldsymbol{\alpha}_\mathrm{b})$ is concave on $[0, 1 - \hat{p}_g^1]$ and on $[1 - \hat{p}_g^1, 1]$. □

From the concavity property proved in Lemma 4, we can conclude that it is easy to minimize $F(\boldsymbol{\alpha}_\mathrm{b})$ over α^1_b when α^0_b is fixed, since there is only a small finite number of solutions. This is proved in the following lemma.

Lemma 5. *Let $\alpha^0_b \in [0,1]$, and define*

$$\Theta(\alpha^0_b) := \{\theta^i_g(\alpha^0_b)\}_{g\in\mathcal{G},i\in[4]} \cup \{0,1\}.$$

Then,

$$\min_{\alpha^1_b\in[0,1]} F(\boldsymbol{\alpha}_b) = \min_{\alpha^1_b\in\Theta(\alpha^0_b)} F(\boldsymbol{\alpha}_b).$$

Proof. Fix α^0_b. Ordering the points in $\Theta(\alpha^0_\mathrm{b})$ in an ascending order and naming them $0 = p_1 < p_2 < \ldots < p_N = 1$, where $N = |\Theta(\alpha^0_\mathrm{b})|$, define the set of intervals $\mathcal{I} = \{[p_i, p_{i+1}] \setminus \{0,1\}\}_{i\in[N-1]}$. Observe that the intervals in $\mathcal{I}$ partition $(0,1)$ (with overlaps at end points). Moreover, for all $g \in \mathcal{G}^+, i \in [4]$, each of the intervals $I \in \mathcal{I}$ satisfies either $I \subseteq (0, \theta^i_g(\alpha^0_\mathrm{b})] \cap I^i_g$ or $I \subseteq [\theta^i_g(\alpha^0_\mathrm{b}), 1) \cap I^i_g$. By Lemma 4, it follows that for all $g \in \mathcal{G}^+, i \in [4]$, $\alpha^1_\mathrm{b} \mapsto \tilde{f}^i_g(\boldsymbol{\alpha}_\mathrm{b})$ is concave on I.

F is a conic combination of minima of concave functions on I, thus it is concave on I. Thus, on each $I \in \mathcal{I}$ which is a closed interval (that is, with endpoints other than 0 or 1), $F(\boldsymbol{\alpha}_\mathrm{b})$ is minimized by one of the endpoints of I. Now, consider I with an endpoint 0 or 1. For simplicity, assume $I = (0, p_2]$; the other case is symmetric. $F(\boldsymbol{\alpha}_\mathrm{b})$ is concave on I. Thus, it is either minimized on I at p_2 or satisfies $\lim_{\alpha_1\to 0} F(\boldsymbol{\alpha}_\mathrm{b}) \le \inf_{\alpha^1_\mathrm{b}\in I}(F(\boldsymbol{\alpha}_\mathrm{b}))$. In the latter case, there are two options: if $\alpha^1_\mathrm{b} \mapsto F(\boldsymbol{\alpha}_\mathrm{b})$ is continuous at 0, then $F(\alpha^0_\mathrm{b}, 0) = \lim_{\alpha_1\to 0} F(\boldsymbol{\alpha}_\mathrm{b})$, implying that the function is minimized on I by $\alpha^1_\mathrm{b} = 0$. If it is not continuous at 0, this can only happen if $r_g + q_g s^i_g(\boldsymbol{\alpha}_\mathrm{b}) = 0$, which leads to $\eta(\alpha^1_\mathrm{b}, r_g + q_g s^i_g(\boldsymbol{\alpha}_\mathrm{b})) = 0$, whereas $\lim_{\alpha^1_\mathrm{b}\to 0} \eta(\alpha^1_\mathrm{b}, r_g + q_g s^i_g(\boldsymbol{\alpha}_\mathrm{b})) = 1$. Thus, in this case, $F(\alpha^0_\mathrm{b}, 0) \le \lim_{\alpha_1\to 0} F(\boldsymbol{\alpha}_\mathrm{b})$. It follows that in this case as well, F is minimized on I by 0.

It follows that the minimizer of $\alpha^1_\mathrm{b} \mapsto F(\boldsymbol{\alpha}_\mathrm{b})$ must be one of the end points of $\Theta(\alpha^0_\mathrm{b})$, as claimed. □

Now, due to the complete symmetry of the problem definition between the two labels, Lemma 5 implies also the symmetric property, that for a fixed $\alpha^1_\mathrm{b} \in [0,1]$, $\min_{\alpha^0_\mathrm{b}\in[0,1]} F(\boldsymbol{\alpha}_\mathrm{b}) = \min_{\alpha^0_\mathrm{b}\in\tilde{\Theta}(\alpha^1_\mathrm{b})} F(\boldsymbol{\alpha}_\mathrm{b})$, where $\tilde{\Theta}$ is symmetric to Θ and is obtained by switching the roles of $y = 1$ and $y = 0$ in all definitions. By explicitly calculating the resulting values, we get that

$$\tilde{\Theta}(\alpha^1_\mathrm{b}) = V_0 \cup \{(\alpha^1_\mathrm{b} - r_g)/q_g\}_{g\in\mathcal{G}^+},$$
$$\Theta(\alpha^0_\mathrm{b}) = V_1 \cup \{r_g + q_g\alpha^0_\mathrm{b}\}_{g\in\mathcal{G}^+},$$

Where V_0, V_1 are defined as in Theorem 2. Using the above results, we can now prove Theorem 2.

of Theorem 2. By Eq. (14), $\mathtt{mindisc}_\beta = \min_{\boldsymbol{\alpha}_\mathrm{b}\in[0,1]^2} F(\boldsymbol{\alpha}_\mathrm{b})$. Thus, we have left to show that there exists a minimizer for $F(\boldsymbol{\alpha}_\mathrm{b})$ in the set Pairs defined in Theorem 2. From Lemma 5, it follows that there exists a minimizing solution $(a^0, a^1) \in \operatorname{argmin}_{\boldsymbol{\alpha}_\mathrm{b}\in[0,1]^2} F(\boldsymbol{\alpha}_\mathrm{b})$ such that $a^1 \in \Theta(a^0)$. The symmetric counterpart of Lemma 5 implies that there exists some $b^0 \in \tilde{\Theta}(a^1)$ such that $(b^0, a^1) \in \operatorname{argmin}_{\boldsymbol{\alpha}_\mathrm{b}\in[0,1]^2} F(\boldsymbol{\alpha}_\mathrm{b})$. Thus, at least one of the following options must hold:

- $(b^0, a^1) \in V_0 \times V_1$.
- $a^1 \notin V_1$; this implies that $a^1 = r_g + q_g a^0$ for some $g \in \mathcal{G}^+$.
- $b^0 \notin V_0$; this implies that $b^0 = (a^1 - r_g)/q_g$ for some $g \in \mathcal{G}^+$.

In the first case, there exists a minimizer in $V_0 \times V_1$. In the second and third case, there exists a minimizer in the set $\{(v, r_g + q_g v) \mid v \in [0,1], g \in \mathcal{G}^+\}$. Thus, in all cases there is a minimizer in the set Pairs, as required. □

B Proof of Lemma 3

of Lemma 3. Let $\phi[a,b,c,d](x) := 1-(c+dx)/(a+bx)$. Denote $a' = 1-a-b$, $c' = 1-c-d$, $x' = 1-x$. Then, it can be easily verified that

$$\eta(a+bx, c+dx) = \begin{cases} \phi[a,b,c,d](x) & a+bx > c+dx \\ \phi[a',b,c',d](x') & a'+bx' > c'+dx'. \end{cases}$$

Note that $a+bx > c+dx \Leftrightarrow a'+bx' < c'+dx'$. Therefore, it suffices to find an upper bound on $|d\phi[a,b,c,d](x)/dx| = \frac{|bc-ad|}{(a+bx)^2}$, for a,b,c,d such that $b>0$ and $d \geq 0$, on $x \in [0,1]$ such that $a+bx > c+dx$, and $a+bx, c+dx \in [0,1]$.

We consider the possible cases.

- Suppose $d = 0$
  - If $c > 0$ then $a+bx > c$, thus $|d\phi[a,b,c,d](x)/dx| = \frac{|bc-ad|}{(a+bx)^2} \leq \frac{|bc|}{c^2} = b/c$.
  - If $c = 0$, then $bc-ad = 0$, thus the derivative is zero.
  - If $c < 0$, then $c+dx < 0$, thus the domain is empty.
- Otherwise, $d > 0$. Let x_1 such that $a+bx_1 = 0$ and x_2 such that $c+dx_2 = 0$. Then $x_1 = -a/b$, $x_2 = -c/d$. Note that we must have $x \geq \max\{x_1, x_2\}$.
  - If $x_1 < x_2$ $(-a/b < -c/d)$ then $a+bx \geq a+bx_2 > a+bx_1 = 0$. Since $|a+bx_2| = |a-bc/d| = |bc-ad|/d$, we have $|d\phi[a,b,c,d](x)/dx| = \frac{|bc-ad|}{(a+bx)^2} \leq d^2/|bc-ad|$.
  - If $x_1 > x_2$ $(-a/b > -c/d)$, then $c+dx_1 > c+dx_2 = 0 = a+bx_1$. Thus, $c-a+(d-b)x_1 > 0$. Let x be any value in the feasible domain. Then $a+bx \geq c+dx$, hence $c-a+(d-b)x \leq 0$.
    - If $d-b = 0$ then this contradicts the property of x_1, thus there is no such x.
    - If $d-b > 0$, then $x < x_1$, thus $a+bx < 0$, again implying that x is not in the domain. Thus, in this case the domain is empty as well.
    - If $d-b < 0$, then $x \geq (c-a)/(b-d) > x_1$, thus $a+bx \geq a+b(c-a)/(b-d) > a+bx_1 \geq 0$. Therefore, $|d\phi[a,b,c,d](x)/dx| = \frac{|bc-ad|}{(a+bx)^2} \leq \frac{|bc-ad|}{a+b(c-a)/(b-d)}$.
  - If $x_1 = x_2$, Then $-c/d = -a/b$, which means that $ad = bc$, thus the derivative in this case is zero.

Combining the cases gives the upper bound for the derivative of ψ as given by $D(a,b,c,d)$, for x such that $a+bx > c+dx$. In the special case that $b = d$, this holds for all x if $a > c$ and for no x otherwise. In this special case, if $a > c$, then $D(a,b,c,d) = d^2/|bc-ad| = d/|a-c|$.

For the other case in the definition of η, the upper bound is obtained by $D(a',b,c',d)$ as described above, including the special case $b = d$ and $a' > c'$, that is $a < c$. □

Received 26 November 2025; accepted 25 June 2026