Increasing Fairness via Combination with Learning Guarantees

Yijun Bian
Department of Computer Science
University of Copenhagen
yibi@di.ku.dk

Kun Zhang
School of Computer Science and Information Engineering
Hefei University of Technology
zhang1028kun@gmail.com

Abstract

The concern about hidden discrimination in machine learning (ML) models is growing, as their widespread real-world applications increasingly impact human lives. Various techniques, including commonly used group fairness measures and several fairness-aware ensemble-based methods, have been developed to enhance fairness. However, existing fairness measures typically focus on only one aspect—either group or individual fairness, and the hard compatibility among them indicates a possibility of remaining biases even when one of them is satisfied. Moreover, existing mechanisms to boost fairness usually present empirical results to show validity, yet few of them discuss whether fairness can be boosted with certain theoretical guarantees. To address these issues, we propose a fairness quality measure named '*discriminative risk (DR)*' to reflect both individual and group fairness aspects. Furthermore, we investigate its properties and establish the first- and second-order oracle bounds concerning fairness to show that fairness can be boosted via ensemble combination with theoretical learning guarantees. The analysis is suitable for both binary and multi-class classification. Comprehensive experiments are conducted to evaluate the effectiveness of the proposed methods.

1 Introduction

Machine learning (ML) is increasingly applied in sensitive decision-making domains such as recruitment, jurisdiction, and credit evaluation. As ML becomes more pervasive in real-world scenarios nowadays, concerns about the fairness and reliability of ML models have emerged and grown. Discriminative models may perpetuate or even exacerbate improper human prejudices, negatively impacting on both model performance and societal equity. Unfairness in ML models identified in the literature primarily stems from two causes: data biases and algorithmic biases [44]. Data biases mainly arise from inaccurate device measurements, erroneous reports, or historically biassed human decisions, misleading ML models to replicate them. Missing data can also distort the distribution of the dataset from the target population and introduce further biases. Algorithmic biases occur even with purely clean data. These biases may arise from proxy attributes for sensitive variables or from tendentious objectives of the learning algorithms themselves. For example, minimising aggregated prediction errors may inadvertently favour the privileged group over unprivileged minorities.

Numerous mechanisms have been proposed to mitigate biases and enhance fairness in ML models, typically categorised as pre-processing, inprocessing, and post-processing mechanisms. Pre- and post-processing mechanisms normally function by manipulating input or output, while inprocessing mechanisms incorporate fairness constraints into training procedures or algorithmic objectives. Determining the superior approach is challenging as results vary based on applied fairness measures, datasets, and even the training-test split handling [20, 16]. Various fairness measures have been developed to facilitate the design of fair ML models, such as group and individual fairness measures. Group fairness emphasises statistical/demographic equality among groups defined by sensitive

Mathematics of Modern Machine Learning Workshop at NeurIPS 2024.

attributes (SAs), including but not limited to demographic parity (DP) [17, 21], equality of opportunity (EO) [25], and predictive quality parity (PQP) [11, 44]. In contrast, individual fairness follows the principle that 'similar individuals should be evaluated or treated similarly,' with similarity measured by specific distances between individuals [27, 15]. However, these measures often conflict, meaning that unfair outcomes may persist even when one criterion is met, such as the incompatibility of group fairness measures themselves and that between individual and group fairness measures [4, 6, 41, 25].

To address these challenges, we propose a novel fairness quality measure named '*discriminative risk (DR)*', which reflects the bias degree of learners from both individual and group fairness perspectives. We further investigate its properties and suggest that the fairness quality of learners can benefit from combination with theoretical learning guarantees, inspired by a cancellation-of-errors effect of the ensemble combination, where combining multiple weak learners yields a more powerful learner. In essence, we explore the possibility of a cancellation-of-biases effect in combination, seeking to answer the question: *Will combination help mitigate discrimination in multiple biassed individual classifiers?* Our proposed bounds regarding fairness indicate a positive answer, demonstrating that the fairness quality of an ensemble is superior to that of one single individual classifier.

Our contributions in this work are three-fold: (1) We propose a novel fairness quality measure that assesses the bias level of classifiers from both individual and group fairness sides, along with its properties. (2) We establish first- and second-order oracle bounds concerning fairness, theoretically demonstrating the cancellation-of-biases effect in ensemble combination, and present two PAC bounds to further support this effect. (3) We conduct comprehensive experiments to validate the effectiveness of the proposed fairness measure and its corresponding bounds.

2 Methodology

In this section, we formally study the fairness properties of ensemble methods with a majority vote.

We denote a dataset by $S=\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, where instances are independent identically distributed (i.i.d.) drawn from an input/feature-output/label space $\mathcal{X}\times\mathcal{Y}$ according to an unknown distribution $\mathcal{D}$. The label space $\mathcal{Y}=\{1, 2, ..., n_c\}(n_c \geqslant 2)$ is finite and could represent binary or multi-class classification, and the feature space $\mathcal{X}$ is arbitrary. An instance including SAs $\boldsymbol{a}$ is represented as $\boldsymbol{x} \triangleq (\breve{\boldsymbol{x}}, \boldsymbol{a})$, and $\tilde{\boldsymbol{a}}$ indicates perturbed $\boldsymbol{a}$. A hypothesis in the space of hypotheses $\mathcal{F}$ is a function $f \in \mathcal{F} : \mathcal{X} \mapsto \mathcal{Y}$. The *weighted voting* prediction by an ensemble of m individual classifiers, parameterised by a weight vector $\rho = [w_1, ..., w_m]^\mathsf{T} \in [0, 1]^m$, is originally given by $\mathbf{wv}_\rho(\boldsymbol{x}) = \operatorname{argmax}_{y\in\mathcal{Y}} \sum_{j=1}^m w_j \mathbb{I}(f_j(\boldsymbol{x}) = y)$, such that $\sum_{j=1}^m w_j = 1$, where w_j is the weight of individual classifier $f_j(\cdot)$. Note that ties are resolved arbitrarily, and both parametric and non-parametric models can serve as individual classifiers here. The ensemble classifier can be reformulated to ρ-weighted majority vote $\mathbf{wv}_\rho(\boldsymbol{x}) = \operatorname{argmax}_{y\in\mathcal{Y}} \mathbb{E}_{f\sim\rho}[\mathbb{I}(f(\boldsymbol{x})=y)]$. Note that $\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\cdot]$ and $\mathbb{E}_{f\sim\rho}[\cdot]$ can be respectively abbreviated as $\mathbb{E}_{\mathcal{D}}[\cdot]$ and $\mathbb{E}_{\rho}[\cdot]$ for brevity when the context is unambiguous.

2.1 Fairness quality from both individual and group fairness aspects

To discuss the properties of fairness, it is essential to have a proper measure that accurately reflects the prediction quality of hypotheses. Lots of fairness measures have been proposed (see Section A.2), including three commonly used ones (i.e., DP, EO, and PQP).[1] However, the hard compatibility among them means that each one of them focuses only on one specific aspect of fairness (either group or individual fairness). Therefore, to capture the discriminative degree of hypotheses from both individual and group sides, we propose a new fairness quality measure named *discriminative risk (DR)*.

Presume that the considered dataset S consists of instances containing SAs, where the features of an instance $\boldsymbol{x}=(\breve{\boldsymbol{x}}, \boldsymbol{a})$ include the SAs $\boldsymbol{a}$ and non-sensitive attributes $\breve{\boldsymbol{x}}$. Note that $\boldsymbol{a}=[a_1, ..., a_{n_a}]^\mathsf{T}$ allows multiple attributes, with n_a as the number of SAs, and for each attribute, $a_i \in \mathbb{Z}_+ (1 \leqslant i \leqslant n_a)$ allows binary and multiple values. Following the principle of individual fairness, the treat-

[1] In this paper, these three group fairness measures of one hypothesis $f(\cdot)$ are evaluated respectively as

$$\mathrm{DP}(f)=|\mathbb{P}_{\mathcal{D}}[f(\boldsymbol{x})=1|\,\boldsymbol{a}=1]-\mathbb{P}_{\mathcal{D}}[f(\boldsymbol{x})=1|\,\boldsymbol{a}=0]|\,, \tag{1a}$$

$$\mathrm{EO}(f)=|\mathbb{P}_{\mathcal{D}}[f(\boldsymbol{x})=1|\,\boldsymbol{a}=1, y=1]-\mathbb{P}_{\mathcal{D}}[f(\boldsymbol{x})=1|\,\boldsymbol{a}=0, y=1]|, \tag{1b}$$

$$\mathrm{PQP}(f)=|\mathbb{P}_{\mathcal{D}}[y=1|\,\boldsymbol{a}=1, f(\boldsymbol{x})=1]-\mathbb{P}_{\mathcal{D}}[y=1|\,\boldsymbol{a}=0, f(\boldsymbol{x})=1]|, \tag{1c}$$

where $\boldsymbol{x} = (\breve{\boldsymbol{x}}, \boldsymbol{a})$, y is the true label, and $f(\boldsymbol{x})$ is the prediction. Note that $\boldsymbol{a} = 1$ and $\boldsymbol{a} = 0$ mean that the instance $\boldsymbol{x}$ belongs to the privileged group and marginalised groups, respectively.

ment/evaluation of one instance should not change solely due to minor changes in its SAs. This indicates the existence of underlying discriminative risks if one hypothesis/classifier makes different predictions for an instance based solely on changes in its SAs. Consequently, the fairness quality of one hypothesis $f(\cdot)$ can be evaluated by

$$\ell_{\text{bias}}(f, \boldsymbol{x}) = \mathbb{I}(f(\breve{\boldsymbol{x}}, \boldsymbol{a}) \neq f(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}}))\,, \tag{2}$$

similarly to the $0/1$ loss. Note that (2) is evaluated on only one instance. To describe this characteristic of the hypothesis on multiple instances (aka. from a group level), then the empirical discriminative risk on S and the true discriminative risk of the hypothesis are expressed as

$$\hat{\mathcal{L}}_{\text{bias}}(f, S) = \tfrac{1}{n}\sum_{i=1}^{n} \ell_{\text{bias}}(f, \boldsymbol{x}_i)\,, \tag{3}$$

and

$$\mathcal{L}_{\text{bias}}(f) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\ell_{\text{bias}}(f, \boldsymbol{x})]\,, \tag{4}$$

respectively. Therefore, by $\ell_{\text{bias}}(f, \boldsymbol{x})$, we are able to measure the discriminative risk of one hypothesis on one instance, guided by the principle of individual fairness that 'there will exist discriminative risk if similar instances are not treated similarly'.[2] Meanwhile, by $\hat{\mathcal{L}}_{\text{bias}}(f, S)$ and $\mathcal{L}_{\text{bias}}(f)$, we are able to weigh the discriminative risk of the hypothesis over the whole group including all subgroups divided by SAs, in the sense, consistent with the idea of group fairness as well. Overall, the proposed fairness quality could provide benefits to properly measure the discriminative risk of one hypothesis from both individual and group fairness aspects.

Ensemble classifiers predict by taking a weighted combination of predictions by hypotheses from $\mathcal{F}$, and the fairness quality of the ensemble will be evaluated corresponding to the total weight assigned to individual classifiers, that is, $\ell_{\text{bias}}(\mathbf{wv}_\rho, \boldsymbol{x}) = \mathbb{I}(\mathbf{wv}_\rho(\breve{\boldsymbol{x}}, \boldsymbol{a}) \neq \mathbf{wv}_\rho(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}}))$. Then we can obtain the empirical discriminative risk on S and the true discriminative risk of the ensemble analogously.

2.2 The distinction of *DR* compared with existing fairness measures

The design of *DR* shares some similarities with the ideas of individual fairness and group fairness, therefore it is supposed to be able to reflect the bias level of one hypothesis from both individual and group aspects. Besides, from some point of view, the proposed *DR* in (3) and (4) could be viewed as a simplified version of causal fairness (such as counterfactual fairness[3] (CFF) [35] and proxy discrimination[4] (PD) [33]), because both of them consider the situation where SAs are altered. However, there are still many distinctions between our proposed *DR* and these existing fairness and their measures.

Two distinctions from individual fairness The intuition behind individual fairness [15, 14] is that we want similar predicted outcomes for similar individuals based on specific similarity metrics. It is easily notable that the choice of the similarity metric (that is, $\mathbf{d}_{\mathcal{X}}$ and $\mathbf{d}_{\mathcal{Y}}$ in (5)) plays an essential role in comparison and may mislead results if it is chosen less properly. Besides, all these formulations make comparisons between existing individual pairs, while *DR* finds another way

[2] Apart from various choices of the metric, there are different formulations in the literature to mathematically capture the aforementioned intuition, such as Lipschitz mapping-based, probability Lipschitzness, and the $(\epsilon{-}\delta)$ language-based formulations. In other words, for a mapping $h\colon \mathcal{X}\mapsto\mathcal{Y}$, we say it satisfies individual fairness if for all possible $(\breve{\boldsymbol{x}}, \boldsymbol{a}), (\breve{\boldsymbol{x}}', \boldsymbol{a}') \in \mathcal{X}$, it holds any one of the following: (1) it is λ-Lipschitz w.r.t. appropriate metrics on the domain $\mathcal{X}$ and the codomain $\mathcal{Y}$, that is,

$$\mathbf{d}_{\mathcal{Y}}\big(h(\breve{\boldsymbol{x}}, \boldsymbol{a}), h(\breve{\boldsymbol{x}}', \boldsymbol{a}')\big) \leqslant \lambda \cdot \mathbf{d}_{\mathcal{X}}\big((\breve{\boldsymbol{x}}, \boldsymbol{a}), (\breve{\boldsymbol{x}}', \boldsymbol{a}')\big)\,; \tag{5}$$

(2) it is probability Lipschitz w.r.t. appropriate metrics on the domain and the codomain, that is,

$$\mathbb{P}_{\mathcal{D}}\left[\mathbf{d}_{\mathcal{Y}}(h(\breve{\boldsymbol{x}}, \boldsymbol{a}), h(\breve{\boldsymbol{x}}', \boldsymbol{a}'))/\mathbf{d}_{\mathcal{X}}((\breve{\boldsymbol{x}}, \boldsymbol{a}), (\breve{\boldsymbol{x}}', \boldsymbol{a}')) \geqslant \epsilon\right] \leqslant \delta\,; \tag{6}$$

or (3) it holds

$$\mathbf{d}_{\mathcal{X}}\big((\breve{\boldsymbol{x}}, \boldsymbol{a}), (\breve{\boldsymbol{x}}', \boldsymbol{a}')\big) \leqslant \epsilon \Rightarrow \mathbf{d}_{\mathcal{Y}}\big(h(\breve{\boldsymbol{x}}, \boldsymbol{a}), h(\breve{\boldsymbol{x}}', \boldsymbol{a}')\big) \leqslant \delta\,, \tag{7}$$

where we consider $\epsilon \geqslant 0$ and $\delta \geqslant 0$.

[3] Given a predictive problem where A, X and Y denote the protected attributes, remaining attributes, and output of interest respectively, Kusner et al. [35] assume that a causal model (U, V, F) is given, where $V \equiv A \cup X$. The *counterfactual fairness* is a postulated criterion for predictors of Y, that is, a predictor $\hat{Y}$ is counterfactual fair if under any context $X = x$ and $A = a$,

$$\mathbb{P}(\hat{Y}_{A\leftarrow a}(U) = y \mid X = x, A = a) = \mathbb{P}(\hat{Y}_{A\leftarrow a'}(U) = y \mid X = x, A = a)\,, \tag{8}$$

for any y and for any value a' attainable by A.

[4] Kilbertus et al. [33] generally consider causal graphs involving a protected attribute A, a set of proxy variables P, features X, a predictor R, and sometimes an observed outcome Y. A causal graph is a directed, acyclic graph whose nodes represent random variables. A variable V in a causal graph exhibits *potential proxy discrimination*, if there exists a direct path from A to V that is blocked by a proxy variable and V itself is not a proxy; An intervention on P is denoted by $\mathrm{do}(P = p)$, and a predictor R exhibits no *proxy discrimination* based on a proxy P if for all p, p',

$$\mathbb{P}(R \mid \mathrm{do}(P = p)) = \mathbb{P}(R \mid \mathrm{do}(P = p'))\,. \tag{9}$$

to compare individuals by slightly perturbing one individual's SA(s), in order to ensure the two individuals for comparison are similar enough. In other words, the distance between the original individual/instance and its slightly perturbed version is zero(i.e., $\mathbf{d}((\breve{\boldsymbol{x}}, \boldsymbol{a}), (\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}})) = 0$)—because the only possible difference that exists between $(\breve{\boldsymbol{x}}, \boldsymbol{a})$ and $(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}})$ is within SAs, and it is not restricted by or heavily relys on some carefully selected distance metric. *DR* does not compare the original instance pairs within the dataset, which is also different from the existing individual fairness computation. That is to say, all $(\breve{\boldsymbol{x}}, \boldsymbol{a})$ and $(\breve{\boldsymbol{x}}', \boldsymbol{a}')$ in (5) come from the original dataset S, while for an original instance $(\breve{\boldsymbol{x}}, \boldsymbol{a})$, its slightly perturbed version $(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}})$ in (2) does not.

Two distinctions from group fairness Group fairness focuses on statistical/demographic equality among groups defined by SAs, therefore, it is usually calculated in a way of computing the difference of one specific criterion or metric between two different subgroups that are divided by SAs. However, unlike three commonly-used group fairness measures in (1), the proposed *DR* in (3) and (4) does not need to split different subgroups and calculate them separately. *DR* can be gotten for a dataset as a whole including the privileged group and unprivileged groups. The reason is that in (2), if one hypothesis/classifier makes different predictions for one instance with only changes in SAs, there will exist underlying discriminative risks, regardless of which group this instance/member belongs to. Note that *DR* can be computed in the same way as (1),

$$\mathcal{L}'_{\text{bias}}(f) = \left|\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}|\boldsymbol{a}=1}[\ell_{\text{bias}}(f, \boldsymbol{x})] - \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}|\boldsymbol{a}=0}[\ell_{\text{bias}}(f, \boldsymbol{x})]\right|, \tag{10}$$

if the practitioner wants, except that they do not have to. Moreover, these three group fairness measures only work for the scenario of one SA (usually with binary values). Thus a bonus advantage of the proposed *DR* is that it is suitable for not only binary but also multi-class classification scenarios, as well as for the scenarios of several SAs and multiple values, enlarging its applicable fields.

Five distinctions from causal fairness Despite the similarity shared between *DR* and causal fairness (that is, CFF [35] and PD [33]), there are also differences. First, *DR* is a quantitative measure to evaluate the discriminative risk, while causal fairness is validated based on causal models/graphs (U, V, F)—not being computed in concrete values. As one measure without causal analysis, *DR* is easy to calculate and avoids the risk of misplacing the relationship among factors, yet is still somehow able to achieve similar observations guided by the same idea of evaluating what if SAs of instances are altered. Second, both CFF and PD only work for one SA, as shown in (8) and (9), while *DR* is more applicable as it can deal with scenarios of multiple SAs. Third, note that non-sensitive attributes may be changed as well in CFF when SAs are perturbed due to the existence of proxy attributes, which means $\breve{\boldsymbol{x}}$ might not remain the same. In contrast, *DR* and PD do not consider the subsequent change of proxy attributes after perturbing the SAs. Fourth, both CFF and PD have a quite strong condition in (8) and (9) by 'it is achieved only if the equality holds for any value $\tilde{\boldsymbol{a}}$ different from $\boldsymbol{a}'$, while *DR* has a weaker one by 'it is broken if the equality does not hold anymore when there exists one $\tilde{\boldsymbol{a}}$ different from $\boldsymbol{a}$, without checking all possible $\tilde{\boldsymbol{a}}$ that does not equal $\boldsymbol{a}'$. We implement it by using a probability of altering SAs within *DR* when computing it in practice. In other words, $\tilde{\boldsymbol{a}}$ is perturbed with a probability of $p \in [0, 1]$ from $\boldsymbol{a}$, which means it is with a teeny tiny little possibility that $\tilde{\boldsymbol{a}} = \boldsymbol{a}$ for certain instance (not all of them in one dataset), while usually $\tilde{\boldsymbol{a}} \neq \boldsymbol{a}$ in CFF and PD. Lastly, *DR* can be proved to be bounded, presented in Section 2.3, which is an advantage that CFF and PD do not have.

2.3 Properties of *DR* and bounds regarding fairness for weighted vote

In this subsection, we firstly discuss properties of *DR*, where no restrictions apply on the type of classifiers. In other words, it works for both individual/member classifiers and ensemble classifiers via weighted vote. We argue that *the empirical* DR *on S is an unbiassed estimation of the true* DR*, that is, $\hat{\mathcal{L}}_{bias}(f, S)$ in* (3) *is an unbiassed estimation of $\mathcal{L}_{bias}(f)$ in* (4). The reason is that for one random variable X representing instances, $\ell_{\text{bias}}(f, \boldsymbol{x})$ in (2) could be viewed as a new random variable obtained by using a few fixed operations on X, recorded as Y. Then for n random variables (i.e., $\mathsf{X}_1, \mathsf{X}_2, ..., \mathsf{X}_n$ representing instances) that are independent and identically distributed (iid.), by operating them in the same way, we can get random variables $\mathsf{Y}_1, \mathsf{Y}_2, ..., \mathsf{Y}_n$ that are iid. as well. Then we can rewrite $\hat{\mathcal{L}}_{\text{bias}}(f, S)$ as $\frac{1}{n}\sum_{i=1}^{n}\mathsf{Y}_i$ and $\mathcal{L}_{\text{bias}}(f)$ as $\mathbb{E}_{\mathsf{Y}\sim\mathcal{D}'}[\mathsf{Y}]$, where $\mathcal{D}'$ denotes the space after operating $\mathsf{X}\sim\mathcal{D}$. Therefore, it could be easily seen that the former is an unbiased estimation of the latter.

Then we provide some bounds concerning fairness for the weighted vote, inspired by the work of Masegosa *et al.* [37]. Following the notations described in Section 2.1, if the weighted vote makes a discriminative decision, then at least a ρ-weighted half of the classifiers have made a discriminative

decision and, therefore, $\ell_{\text{bias}}(\mathbf{wv}_\rho, \boldsymbol{x}) \leqslant \mathbb{I}(\mathbb{E}_\rho[\mathbb{I}(f(\breve{\boldsymbol{x}}, \boldsymbol{a}) \neq f(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}}))] \geqslant 0.5)$. This observation leads to our proposed first- and second-order oracle bounds for the fairness quality of weighted vote. Note that there are no more assumptions except the aforementioned notations in the following theorems. Also, note that a *prominent difference* between our work and the work of Masegosa *et al.* [37] is that they investigate the expected risk or accuracy rather than fairness quality.

2.3.1 Oracle bounds regarding fairness for weighted vote

Theorem 1 (First-order oracle bound)**.**

$$\mathcal{L}_{bias}(\mathbf{wv}_\rho) \leqslant 2\mathbb{E}_\rho[\mathcal{L}_{bias}(f)]\,. \tag{11}$$

To investigate the bound deeper, we introduce here the tandem fairness quality of two hypotheses $f(\cdot)$ and $f'(\cdot)$ on one instance $(\boldsymbol{x}, y)$, adopting the idea of the tandem loss [37], by

$$\ell_{\text{bias}}(f, f', \boldsymbol{x}) = \mathbb{I}\big(\big(f(\breve{\boldsymbol{x}}, \boldsymbol{a}) \neq f(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}})\big) \wedge \big(f'(\breve{\boldsymbol{x}}, \boldsymbol{a}) \neq f'(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}})\big)\big)\,. \tag{12}$$

The tandem fairness quality counts a discriminative decision on the instance $(\boldsymbol{x}, y)$ if and only if both $f(\cdot)$ and $f'(\cdot)$ give a discriminative prediction on it. Note that in the degeneration case $\ell_{\text{bias}}(f, f, \boldsymbol{x}) = \ell_{\text{bias}}(f, \boldsymbol{x})$. Then the expected tandem fairness quality is defined by $\mathcal{L}_{\text{bias}}(f, f') = \mathbb{E}_{\mathcal{D}}[\ell_{\text{bias}}(f, f', \boldsymbol{x})]$. Lemma 1 relates the expectation of the second moment of the standard fairness quality to the expected tandem fairness quality. Note that $\mathbb{E}_{f\sim\rho, f'\sim\rho}[\mathcal{L}_{\text{bias}}(f, f')]$ for the product distribution $\rho \times \rho$ over $\mathcal{F} \times \mathcal{F}$ can be abbreviated as $\mathbb{E}_{\rho^2}[\mathcal{L}_{\text{bias}}(f, f')]$ for brevity.

Lemma 1. *In multi-class classification,*

$$\mathbb{E}_{\mathcal{D}}[\mathbb{E}_\rho[\ell_{bias}(f, \boldsymbol{x})]^2] = \mathbb{E}_{\rho^2}[\mathcal{L}_{bias}(f, f')]\,. \tag{13}$$

Theorem 2 (Second-order oracle bound)**.** *In multi-class classification*

$$\mathcal{L}_{bias}(\mathbf{wv}_\rho) \leqslant 4\mathbb{E}_{\rho^2}[\mathcal{L}_{bias}(f, f')]\,. \tag{14}$$

Furthermore, we can also obtain an alternative second-order bound based on Chebyshev-Cantelli inequality, presented in Theorem 3.

Theorem 3 (C-tandem oracle bound)**.** *If* $\mathbb{E}_\rho[\mathcal{L}_{bias}(f)] < 1/2$*, then*

$$\mathcal{L}_{bias}(\mathbf{wv}_\rho) \leqslant \frac{\mathbb{E}_{\rho^2}[\mathcal{L}_{bias}(f, f')] - \mathbb{E}_\rho[\mathcal{L}_{bias}(f)]^2}{\mathbb{E}_{\rho^2}[\mathcal{L}_{bias}(f, f')] - \mathbb{E}_\rho[\mathcal{L}_{bias}(f)] + \frac{1}{4}}\,. \tag{15}$$

Up to now, we have gotten the first- and second-order oracle bounds regarding the fairness quality for the weighted vote, which will help us further investigate fairness. Furthermore, it is worth noting that, despite the similar names of 'first- and second-order oracle bounds' from our inspiration [37], the essences of our bounds are distinct from theirs. Specifically, their work investigates the bounds for generalisation error and is not relevant to fairness issues at all, while ours focus on the theoretical support for bias mitigation. In other words, their bounds are based on the $0/1$ loss $\ell_{\text{err}}(f, \boldsymbol{x}) = \mathbb{I}(f(\boldsymbol{x}) \neq y)$, while ours are built upon $\ell_{\text{bias}}(f, \boldsymbol{x})$ in (2).

2.3.2 PAC bounds for the weighted vote

All oracle bounds described above are expectations that can only be estimated on finite samples instead of being calculated precisely. They could be transformed into empirical bounds via PAC analysis as well to ease the difficulty of giving a theoretical guarantee of the performance on any unseen data, which we discuss in this subsection. Based on Hoeffding's inequality, we can deduct generalisation bounds presented in Theorems 4 and 5.

Theorem 4. *For any* $\delta \in (0, 1)$*, with probability at least* $(1-\delta)$ *over a random draw of* S *with a size of* n*, for a single hypothesis* $f(\cdot)$*,*

$$\mathcal{L}_{bias}(f) \leqslant \hat{\mathcal{L}}_{bias}(f, S) + \sqrt{\tfrac{1}{2n} \ln \tfrac{1}{\delta}}\,. \tag{16}$$

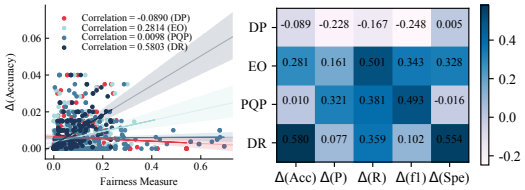
Theorem 5. *For any* $\delta \in (0, 1)$*, with probability at least* $(1 - \delta)$ *over a random draw of* S *with a size of* n*, for all distributions* ρ *on* $\mathcal{F}$*,*

$$\mathcal{L}_{bias}(\mathbf{wv}_\rho) \leqslant \hat{\mathcal{L}}_{bias}(\mathbf{wv}_\rho) + \sqrt{\tfrac{1}{2n} \log \tfrac{|\mathcal{F}|}{\delta}}\,. \tag{17}$$

Note that all proofs in this subsection are provided in Appendix B.

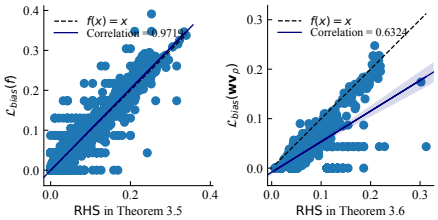
3 Empirical results

In this section, we present our experiments to evaluate the effectiveness of the proposed *discriminative risk (DR)* and its corresponding bounds. These experiments are conducted to explore the following research questions: **RQ1**. Compared with the baseline fairness measures, does *DR* capture the bias level of classifiers effectively, and can it capture discrimination from both individual and group fairness aspects? **RQ2**. Are the oracle bounds and generalisation bounds in Section 2.3 valid? Note that more details (including the experimental setup) are elaborated on in Appendix C to save space.



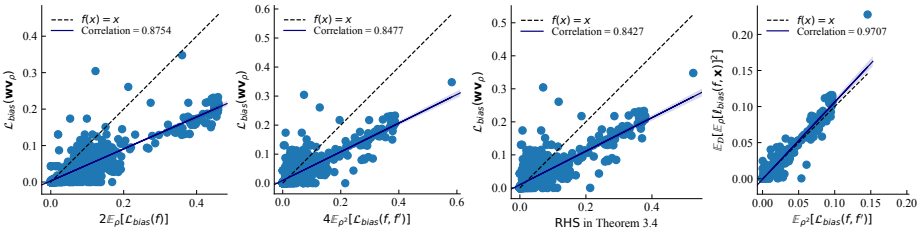
(a) (b)

Figure 1: Comparison of the proposed *discriminative risk (DR)* with three group fairness measures, that is, DP, EO, and PQP. (a) Scatter diagrams with the degree of correlation, where the x- and y-axes are different fairness measures and the variation of accuracy between the raw and perturbed data. (b) Correlation among multiple criteria. Note that correlation here is calculated based on the results from all datasets.



(a) (b)

Figure 2: Correlation for generalisation bounds in section 2.3.2. (a–b) Correlation between $\mathcal{L}_{\text{bias}}(\cdot)$ and generalisation bounds, where $\mathcal{L}_{\text{bias}}(\cdot)$ is indicated on the vertical axis and the right-hand sides of inequalities (16) and (17) are indicated on the horizontal axes, respectively. Note that correlation here is calculated based on the results from all datasets.



(a) (b) (c) (d)

Figure 3: Correlation for oracle bounds in section 2.3.1. (a–c) Correlation between $\mathcal{L}_{\text{bias}}(\mathbf{wv}_\rho)$ and oracle bounds, where $\mathcal{L}_{\text{bias}}(\mathbf{wv}_\rho)$ is indicated on the vertical axis and the horizontal axes represent the right-hand sides of inequalities (11), (14), and (15), respectively. (d) The horizontal and vertical axes in (d) denote the right- and left-hand sides in (13), respectively. Note that correlation here is calculated based on the results from all datasets.

Validating the proposed fairness quality measure We evaluate here the validity of the proposed fairness quality measure (namely *discriminative risk, DR*) in section 2.1, compared with three group fairness measures. Classifiers are conducted using several ensemble methods including bagging, AdaBoost, lightGBM, FairGBM, and AdaFair. The empirical results are reported in Fig. 1. As we can see from Fig. 1(a), compared with three group fairness measures, *DR* has the highest value of correlation (namely the Pearson correlation coefficient) between itself and the variation of accuracy. It means that *DR* captures better the characteristic of changed treatment than three other group fairness measures when SAs are perturbed, as the drop in accuracy indicates the existence of underlying discrimination hidden in models. Fig. 1(b) reports the correlation between the variation of other criteria (such as precision, recall/sensitivity, f_1 score, and specificity) including accuracy and multiple fairness measures. It shows that *DR* also has a high correlation with the variation of specificity.

Validating the oracle bounds and PAC-Bayesian bounds Experiments here are conducted to verify the proposed oracle bounds in section 2.3.1 and generalisation bounds in section 2.3.2, of which the validity is evaluated on scatter diagrams with the degree of correlation, namely the Pearson correlation coefficient, reported in Figures 3 and 2. As we may see from Fig. 3(a), it shows a high level of correlation between $\mathcal{L}_{\text{bias}}(\mathbf{wv}_\rho)$ and $2\mathbb{E}_\rho[\mathcal{L}_{\text{bias}}(f)]$ and that $\mathcal{L}_{\text{bias}}(\mathbf{wv}_\rho)$ is indeed smaller than $2\mathbb{E}_\rho[\mathcal{L}_{\text{bias}}(f)]$ in most cases, indicating the inequality (11) is faithful. Similar results are presented in Figures 3(b) and 3(c) as well for inequalities (14) and (15), respectively. Note that the correlation between $\mathbb{E}_{\mathcal{D}}[\mathbb{E}_\rho[\ell_{\text{bias}}(f,\boldsymbol{x})]^2]$ and $\mathbb{E}_{\rho^2}[\mathcal{L}_{\text{bias}}(f,f')]$ in (13) is close to one, indicating that (13) is faithful. As for Fig. 3(b), it shows a relatively high level of correlation between $\mathcal{L}_{\text{bias}}(\mathbf{wv}_\rho)$ and the generalisation bound in inequality (17) and that $\mathcal{L}_{\text{bias}}(\mathbf{wv}_\rho)$ is indeed smaller than the generalisation

bound in most of the cases, indicating the inequality (17) for an ensemble is reliable. Similar results for one single individual classifier are presented in Fig. 3(a) to demonstrate that inequality (16) is reliable.

4 Conclusion

We have presented a novel analysis of the expected fairness quality via weighted vote and demonstrated the existence of a cancellation-of-biases effect in ensemble combination, confirmed by extensive empirical results. Our work shows that combination could boost fairness with theoretical learning guarantees, which is helpful to save some fruitless efforts on (hyper-)parameter tuning.

Acknowledgments and Disclosure of Funding

This research is funded by the European Union (MSCA, FairML, project no. 101106768).

Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Executive Agency (REA). Neither the European Union nor the granting authority can be held responsible for them.

The authors would also like to personally thank Prof. Yevgeny Seldin at the University of Copenhagen and Prof. Shai Ben-David at the Univerity of Waterloo, for their discussions with the authors.

References

[1] A Agarwal, Alina Beygelzimer, Miroslav Dudík, John Langford, and Hanna Wallach. A reductions approach to fair classification. In *ICML*, volume 80, pages 60–69. PMLR, 2018.

[2] Alekh Agarwal, M Dudík, and Zhiwei S Wu. Fair regression: Quantitative definitions and reduction-based algorithms. In *ICML*, volume 97, pages 120–129. PMLR, 2019.

[3] Arturs Backurs, Piotr Indyk, Krzysztof Onak, Baruch Schieber, Ali Vakilian, and Tal Wagner. Scalable fair clustering. In *ICML*, volume 97, pages 405–413. PMLR, 2019.

[4] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning*. fairmlbook.org, 2019.

[5] Yahav Bechavod and Katrina Ligett. Penalizing unfairness in binary classification. *arXiv preprint arXiv:1707.00044*, 2017.

[6] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociol Methods Res*, 50(1):3–44, 2021.

[7] Leo Breiman. Bagging predictors. *Mach Learn*, 24(2):123–140, 1996.

[8] Toon Calders and Sicco Verwer. Three naive bayes approaches for discrimination-free classification. *Data Min Knowl Discov*, 21(2):277–292, 2010.

[9] Flavio Calmon, Dennis Wei, Bhanukiran Vinzamuri, Karthikeyan Natesan Ramamurthy, and Kush R Varshney. Optimized pre-processing for discrimination prevention. In *NIPS*, volume 30. Curran Associates, Inc., 2017.

[10] Flavio Chierichetti, Ravi Kumar, Silvio Lattanzi, and Sergei Vassilvitskii. Fair clustering through fairlets. In *NIPS*, volume 30. Curran Associates, Inc., 2017.

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

[12] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *SIGKDD*, pages 797–806. ACM, 2017.

[13] André F Cruz, Catarina Belém, João Bravo, Pedro Saleiro, and Pedro Bizarro. Fairgbm: Gradient boosting with fairness constraints. In *ICLR*, 2023.

[14] Cynthia Dwork and Christina Ilvento. Fairness under composition. In *ITCS*. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[15] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *ITCS*, pages 214–226. ACM, 2012.

[16] Cynthia Dwork, Nicole Immorlica, Adam Tauman Kalai, and Max Leiserson. Decoupled classifiers for group-fair and efficient machine learning. In *FAT*, volume 81, pages 119–133. PMLR, 2018.

[17] Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *SIGKDD*, pages 259–268, 2015.

[18] Yoav Freund. Boosting a weak learning algorithm by majority. *Inf Comput*, 121(2):256–285, 1995.

[19] Yoav Freund, Robert E Schapire, et al. Experiments with a new boosting algorithm. In *ICML*, volume 96, pages 148–156. Citeseer, 1996.

[20] Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *FAT*, pages 329–338. ACM, 2019.

[21] Pratik Gajane and Mykola Pechenizkiy. On formalizing fairness in prediction with machine learning. In *FAT/ML*, 2018.

[22] Nina Grgić-Hlača, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. Beyond distributive fairness in algorithmic decision making: Feature selection for procedurally fair learning. In *AAAI*, volume 32, 2018.

[23] Benjamin Guedj. A primer on pac-bayesian learning. *arXiv preprint arXiv:1901.05353*, 2019.

[24] Benjamin Guedj and John Shawe-Taylor. A primer on pac-bayesian learning. Technical report, 2019.

[25] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. In *NIPS*, volume 29, pages 3323–3331. Curran Associates Inc., 2016.

[26] Vasileios Iosifidis and Eirini Ntoutsi. Adafair: Cumulative fairness adaptive boosting. In *CIKM*, pages 781–790, New York, NY, USA, 2019. ACM.

[27] Matthew Joseph, Michael Kearns, Jamie H Morgenstern, and Aaron Roth. Fairness in learning: Classic and contextual bandits. In *NIPS*, volume 29. Curran Associates, Inc., 2016.

[28] Faisal Kamiran and Toon Calders. Data preprocessing techniques for classification without discrimination. *Knowl Inf Syst*, 33(1):1–33, 2012.

[29] Faisal Kamiran, Toon Calders, and Mykola Pechenizkiy. Discrimination aware decision tree learning. In *ICDM*, pages 869–874. IEEE, 2010.

[30] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Enhancement of the neutrality in recommendation. In *RecSys Workshop on Decisions*, pages 8–14, 2012.

[31] Toshihiro Kamishima, Shotaro Akaho, Hideki Asoh, and Jun Sakuma. Fairness-aware classifier with prejudice remover regularizer. In *ECML-PKDD*, pages 35–50. Springer Berlin Heidelberg, 2012.

[32] Guolin Ke, Qi Meng, Thomas Finley, Taifeng Wang, Wei Chen, Weidong Ma, Qiwei Ye, and Tie-Yan Liu. Lightgbm: A highly efficient gradient boosting decision tree. In *NIPS*, volume 30, pages 3146–3154, 2017.

[33] Niki Kilbertus, Mateo Rojas Carulla, Giambattista Parascandolo, Moritz Hardt, Dominik Janzing, and Bernhard Schölkopf. Avoiding discrimination through causal reasoning. In *NIPS*, volume 30, 2017.

[34] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. In *ITCS*, volume 67, page 43. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2017.

[35] Matt J Kusner, Joshua Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. In *NIPS*, volume 30, pages 4069–4079. Curran Associates, Inc., 2017.

[36] Christos Louizos, Kevin Swersky, Yujia Li, Max Welling, and Richard S Zemel. The variational fair autoencoder. In *ICLR*, 2016.

[37] Andrés R Masegosa, Stephan Sloth Lorenzen, Christian Igel, and Yevgeny Seldin. Second order pac-bayesian bounds for the weighted majority vote. In *NeurIPS*, volume 33, pages 5263–5273. Curran Associates, Inc., 2020.

[38] Aditya K Menon and Robert C Williamson. The cost of fairness in binary classification. In *FAT*, volume 81, pages 107–118. PMLR, 2018.

[39] Hamed Nilforoshan, Johann D Gaebler, Ravi Shroff, and Sharad Goel. Causal conceptions of fairness and their consequences. In *ICML*, volume 162, pages 16848–16887. PMLR, 2022.

[40] Dana Pessach and Erez Shmueli. Improving fairness of artificial intelligence algorithms in privileged-group selection bias data settings. *Expert Syst Appl*, 185:115667, 2021.

[41] Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. On fairness and calibration. In *NIPS*, volume 30, 2017.

[42] Novi Quadrianto and Viktoriia Sharmanska. Recycling privileged learning and distribution matching for fairness. In *NIPS*, volume 30. Curran Associates, Inc., 2017.

[43] Samira Samadi, Uthaipon Tantipongpipat, Jamie H Morgenstern, Mohit Singh, and Santosh Vempala. The price of fair pca: One extra dimension. In *NeurIPS*, volume 31. Curran Associates, Inc., 2018.

[44] Sahil Verma and Julia Rubin. Fairness definitions explained. In *FairWare*, pages 1–7, 2018.

[45] Michael Wick, Swetasudha Panda, and Jean-Baptiste Tristan. Unlocking fairness: a trade-off revisited. In *NeurIPS*, volume 32, pages 8783–8792. Curran Associates, Inc., 2019.

[46] Blake Woodworth, Suriya Gunasekar, Mesrob I Ohannessian, and Nathan Srebro. Learning non-discriminatory predictors. In *COLT*, volume 65, pages 1920–1953. PMLR, 2017.

[47] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rodriguez, and Krishna P Gummadi. Fairness beyond disparate treatment and disparate impact: Learning classification without disparate mistreatment. In *WWW*, pages 1171–1180, 2017.

[48] Muhammad Bilal Zafar, Isabel Valera, Manuel Gomez Rogriguez, and Krishna P Gummadi. Fairness constraints: Mechanisms for fair classification. In *AISTATS*, volume 54, pages 962–970. PMLR, 2017.

[49] Rich Zemel, Yu Wu, Kevin Swersky, Toni Pitassi, and Cynthia Dwork. Learning fair representations. In *ICML*, pages 325–333. PMLR, 2013.

[50] Wenbin Zhang, Albert Bifet, Xiangliang Zhang, Jeremy C Weiss, and Wolfgang Nejdl. Farf: A fair and adaptive random forests classifier. In *PAKDD*, pages 245–256. Springer, 2021.

[51] Indrė Žliobaitė. Measuring discrimination in algorithmic decision making. *Data Min Knowl Discov*, 31(4): 1060–1089, 2017.

A Related work

In this section, we first introduce existing techniques to mitigate bias issues in ML models and then summarise relevant fairness-aware ensemble-based methods in turn.

A.1 Mechanisms to enhance fairness

Three typical mechanisms are employed to mitigate biases and enhance fairness in ML models: pre-processing, inprocessing, and post-processing mechanisms. Pre-processing mechanisms usually manipulate features or labels of instances before they are fed into the algorithm, aiming to assimilate the distributions of unprivileged groups with those of the privileged group, thus making it harder for the algorithm to distinguish between them [3, 43, 10, 9, 36, 17, 28]. While advantageous for their flexibility across classification tasks, pre-processing mechanisms often suffer from high uncertainty in the final accuracy. Post-processing mechanisms adjust output scores or decouple predictions for each group [38, 16, 12, 25]. Though also task-agnostic, they typically yield inferior results due to their late application in the learning process, and may compromise individual fairness through differentiated treatment. Inprocessing mechanisms favour directly incorporating fairness constraints during the training by incorporating penalty/regularisation terms [47, 48, 46, 42, 31, 29], while some adjusting these constraints in minimax or multi-objective optimisation settings [2, 1, 49]. These mechanisms enforce explicit trade-offs between accuracy and fairness in objectives, but are closely tied to the specific ML algorithm used. Choosing the optimal mechanism is context-dependent, varying with fairness measures, datasets, and even the training-test split handling [20, 16].

A.2 Types of fairness measures

Various fairness measures have been proposed to facilitate the design of fair ML models, which can be generally divided into distributive and procedural fairness measures. Procedural fairness concerns the fairness of decision-making processes, encompassing feature-apriori fairness, feature-accuracy fairness, and feature-disparity fairness [22]. These measures depend on features and user perceptions of fairness but may still introduce hidden biases within the data. Distributive fairness pertains to the fairness of decision-making outcomes (predictions), including unconscious/unawareness fairness, group fairness, individual fairness, and counterfactual fairness. As the simplest, unawareness fairness means making predictions without explicitly using any protected SAs, though it does not prevent biases linked to associations between unprotected and protected attributes [15]. Group fairness focuses on statistical/demographic equality among groups defined by SAs, such as demographic parity, equality of opportunity, and predictive quality parity [6, 51, 34, 12, 25, 17, 30, 8]. In contrast, individual fairness operates on the principle that 'similar individuals should be evaluated or treated

similarly,' where similarity is measured by some certain distance between individuals while the specified distance also matters a lot [27, 15]. Besides, counterfactual fairness aims to explain the sources of discrimination and qualitative equity through causal interference tools [39, 35].

However, the group fairness measures are often hardly compatible with each other, for example, the occurrence between equalised odds and demographic parity, or that between equalised calibration and equalised odds [6, 41, 25]. Individual and group fairness such as demographic parity are also incompatible except in the case of trivial degenerate solutions. Moreover, three fairness criteria—independence, separation, and sufficiency—are demonstrated not to be satisfied concurrently unless in degenerate cases [4]. Furthermore, significant attention has been paid to compromising accuracy in the pursuit of higher levels of fairness [20, 38, 12, 5]. It is widely accepted that introducing fairness constraints into an optimisation problem likely results in reduced accuracy compared to optimising solely for accuracy. However, some researchers have recently proposed a few unique scenarios where fairness and accuracy can be simultaneously improved [40, 45].

A.3 Fairness-aware ensemble-based methods

One of the primary challenges in designing fair learning algorithms is the potential trade-off between fairness and accuracy. Recent studies have explored various fairness-enhancing techniques and their impact on ML models, including a few methods employing typical boosting mechanisms in ensemble learning, such as AdaFair [26], FARF [50], and FairGBM [13]. For instance, FARF and AdaFair combined different fairness-related criteria into the training phase, while FairGBM transformed non-differentiable fairness constraints into a proxy inequality constraint to facilitate gradient-based optimisation. Despite these advancements, there is a paucity of research addressing the theoretical guarantees of these methods in enhancing fairness, with most studies relying on empirical results to demonstrate practical effectiveness.

B Proofs for the proposed bounds in Section 2.3

In this section, we provide the corresponding proofs for Sections 2.3.1 and 2.3.2 in turn.

B.1 Proof of oracle bounds in section 2.3.1

Proof of theorem 1. We have

$$\mathcal{L}_{\text{bias}}(\mathbf{wv}_\rho) = \mathbb{E}_{\mathcal{D}}[\ell_{\text{bias}}(\mathbf{wv}_\rho, \boldsymbol{x})] \leqslant \mathbb{P}_{\mathcal{D}}(\mathbb{E}_\rho[\mathbb{I}(f(\breve{\boldsymbol{x}}, \boldsymbol{a}) \neq f(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}}))] \geqslant 0.5)\,.$$

By applying Markov's inequality to random variable $Z = \mathbb{E}_\rho[\mathbb{I}(f(\breve{\boldsymbol{x}}, \boldsymbol{a}) \neq f(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}}))]$ we get

$$\mathcal{L}_{\text{bias}}(\mathbf{wv}_\rho) \leqslant \mathbb{P}_{\mathcal{D}}(\mathbb{E}_\rho[\mathbb{I}(f(\breve{\boldsymbol{x}}, \boldsymbol{a}) \neq f(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}}))] \geqslant 0.5) \leqslant 2\mathbb{E}_{\mathcal{D}}[\mathbb{E}_\rho[\mathbb{I}(f(\breve{\boldsymbol{x}}, \boldsymbol{a}) \neq f(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}}))]] = 2\mathbb{E}_\rho[\mathcal{L}_{\text{bias}}(f)]\,.$$

□

Proof of lemma 1. We have

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}[\mathbb{E}_\rho[\ell_{\text{bias}}(f, \boldsymbol{x})]^2] &= \mathbb{E}_{\mathcal{D}}[\mathbb{E}_\rho[\mathbb{I}(f(\breve{\boldsymbol{x}}, \boldsymbol{a}) \neq f(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}}))]\mathbb{E}_\rho[\mathbb{I}(f'(\breve{\boldsymbol{x}}, \boldsymbol{a}) \neq f'(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}}))]] \\
&= \mathbb{E}_{\mathcal{D}}[\mathbb{E}_{\rho^2}[\mathbb{I}(f(\breve{\boldsymbol{x}}, \boldsymbol{a}) \neq f(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}}))\mathbb{I}(f'(\breve{\boldsymbol{x}}, \boldsymbol{a}) \neq f'(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}}))]] \\
&= \mathbb{E}_{\mathcal{D}}[\mathbb{E}_{\rho^2}[\mathbb{I}((f(\breve{\boldsymbol{x}}, \boldsymbol{a}) \neq f(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}})) \wedge (f'(\breve{\boldsymbol{x}}, \boldsymbol{a}) \neq f'(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}})))]] \\
&= \mathbb{E}_{\rho^2}[\mathbb{E}_{\mathcal{D}}[\mathbb{I}((f(\breve{\boldsymbol{x}}, \boldsymbol{a}) \neq f(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}})) \wedge (f'(\breve{\boldsymbol{x}}, \boldsymbol{a}) \neq f'(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}})))]] \\
&= \mathbb{E}_{\rho^2}[\mathbb{E}_{\mathcal{D}}[\ell_{\text{bias}}(f, f', \boldsymbol{x})]] \\
&= \mathbb{E}_{\rho^2}[\mathcal{L}_{\text{bias}}(f, f')]\,.
\end{aligned}$$

□

Proof of theorem 2. By applying second-order Markov's inequality to $Z = \mathbb{E}_\rho[\mathbb{I}(f(\breve{\boldsymbol{x}}, \boldsymbol{a}) \neq f(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}}))]$ and lemma 1, we get

$$\mathcal{L}_{\text{bias}}(\mathbf{wv}_\rho) \leqslant \mathbb{P}_{\mathcal{D}}(\mathbb{E}_\rho[\mathbb{I}(f(\breve{\boldsymbol{x}}, \boldsymbol{a}) \neq f(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}}))] \geqslant 0.5) \leqslant 4\mathbb{E}_{\mathcal{D}}[\mathbb{E}_\rho[\mathbb{I}(f(\breve{\boldsymbol{x}}, \boldsymbol{a}) \neq f(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}}))]^2] = 4\mathbb{E}_{\rho^2}[\mathcal{L}_{\text{bias}}(f, f')]\,.$$

□

Proof of theorem 3. By applying the Chebyshev-Cantelli inequality to $\mathbb{E}_\rho[\ell_{\text{bias}}(f, \boldsymbol{x})]$, we can obtain

$$\begin{aligned}
\mathcal{L}_{\text{bias}}(\mathbf{wv}_\rho) &\leqslant \mathbb{P}_{\mathcal{D}}(\mathbb{E}_\rho[\ell_{\text{bias}}(f, \boldsymbol{x})] \geqslant \tfrac{1}{2}) \\
&= \mathbb{P}_{\mathcal{D}}(\mathbb{E}_\rho[\ell_{\text{bias}}(f, \boldsymbol{x})] - \mathbb{E}_\rho[\mathcal{L}_{\text{bias}}(f)] \geqslant \tfrac{1}{2} - \mathbb{E}_\rho[\mathcal{L}_{\text{bias}}(f)])
\end{aligned}$$

$$\leqslant \frac{\mathbb{E}_{\rho^2}[\mathcal{L}_{\text{bias}}(f,f')] - \mathbb{E}_{\rho}[\mathcal{L}_{\text{bias}}(f)]^2}{(\frac{1}{2}-\mathbb{E}_{\rho}[\mathcal{L}_{\text{bias}}(f)])^2+\mathbb{E}_{\rho^2}[\mathcal{L}_{\text{bias}}(f,f')]-\mathbb{E}_{\rho}[\mathcal{L}_{\text{bias}}(f)]^2}$$
$$= \frac{\mathbb{E}_{\rho^2}[\mathcal{L}_{\text{bias}}(f,f')] - \mathbb{E}_{\rho}[\mathcal{L}_{\text{bias}}(f)]^2}{\mathbb{E}_{\rho^2}[\mathcal{L}_{\text{bias}}(f,f')] - \mathbb{E}_{\rho}[\mathcal{L}_{\text{bias}}(f)] + \frac{1}{4}} .$$

□

B.2 Proof of generalisation bounds in section 2.3.2

Proof of theorem 4. According to Hoeffding's inequality, for any $\varepsilon > 0$, we have

$$\mathbb{P}(\mathcal{L}_{\text{bias}}(f) - \hat{\mathcal{L}}_{\text{bias}}(f,S) \geqslant \varepsilon) \leqslant e^{-2n\varepsilon^2} . \tag{24}$$

Let $\delta \triangleq e^{-2n\varepsilon^2} \in (0,1)$, we can obtain $\varepsilon = \sqrt{1/(2n)\ln(1/\delta)}$. Then with probability at least $(1-\delta)$ we have (16). □

Proof of theorem 5. We have known that for one single hypothesis $f(\cdot)$, and for any $\varepsilon > 0$, it holds (24). Then for a finite hypotheses set F such that $|F| = m$, we will have

$$\begin{aligned}\mathbb{P}\big(\exists\, \rho \in [0,1]^m : \mathcal{L}_{\text{bias}}(\mathbf{wv}_\rho) - \hat{\mathcal{L}}_{\text{bias}}(\mathbf{wv}_\rho, S) \geqslant \varepsilon\big) &\leqslant \textstyle\sum_{j=1}^m \mathbb{P}\big(\mathcal{L}_{\text{bias}}(f_j) - \hat{\mathcal{L}}_{\text{bias}}(f_j, S) \geqslant \varepsilon\big) \\ &\leqslant \textstyle\sum_{j=1}^m e^{-2n\varepsilon^2} \\ &= me^{-2n\varepsilon^2} .\end{aligned}$$

Let $\delta \triangleq me^{-2n\varepsilon^2}$, then we obtain $\varepsilon = \sqrt{1/(2n)\,\ln(m/\delta)}$. Therefore with probability at least $(1-\delta)$ we have (17). □

Note that lemma 2 is deducted based on McAllester Bound [23, 24].

Lemma 2. *For any probability distribution π on $\mathcal{F}$ that is independent of S and any $\delta \in (0,1)$, with probability at least $(1-\delta)$ over a random draw of S, for all distributions ρ on $\mathcal{F}$,*

$$\mathcal{L}_{bias}(\mathbf{wv}_\rho) \leqslant \hat{\mathcal{L}}_{bias}(\mathbf{wv}_\rho) + \sqrt{\tfrac{1}{2n}\Big(\mathrm{KL}(\rho\|\pi) + \ln\tfrac{2\sqrt{n}}{\delta}\Big)} , \tag{26}$$

where $\mathrm{KL}(\rho\|\pi)$ is the Kullback-Leibler divergence between distributions ρ and π.

C Supplemental empirical results

In this section, we elaborate on our experiments to evaluate the effectiveness of the proposed *discriminative risk (DR)* and its corresponding bounds.

C.1 Experimental setups

In this subsection, we present the experimental setting, including datasets, evaluation metrics, baseline fairness measures, baseline fairness-aware ensemble-based methods, baseline ensemble pruning methods, and implementation details.

Datasets Five public datasets[5] that we use include Ricci, Credit, Income, PPR, and PPVR, with more details about the dataset statistics provided in Table 1.

Evaluation metrics, and baseline fairness measures As data imbalance usually exists within unfair datasets, we consider multiple criteria to evaluate the prediction performance from different perspectives, including accuracy, precision, recall (aka. sensitivity), f_1 score, and specificity. Moreover, to measure the discrimination degree within classifiers as well as to evaluate the validity of *DR* in capturing the discriminative risk of classifiers, we consider three commonly-used group fairness measures (that is, demographic parity (DP) [17, 21], equality of opportunity (EO) [25], and predictive quality parity (PQP) [11, 44]) and compare *DR* with these three baseline fairness measures.

[5]Ricci `https://rdrr.io/cran/Stat2Data/man/Ricci.html`, Credit `https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)`, Income `https://archive.ics.uci.edu/ml/datasets/adult`, Propublica-Recidivism (PPR) and Propublica-Violent-Recidivism (PPVR) datasets `https://github.com/propublica/compas-analysis/`

Table 1: Dataset statistics. Note that the columns named "#inst" and "#sen att" represent the number of instances and the number of sensitive attributes (SAs) in the dataset, respectively. The joint SA of one instance represents both two of the SAs of it belong to the corresponding privileged group.

Dataset	**#inst**	**#feature**		**#sen att**	**#privileged group for sensitive attribute**		
		raw	binarized		1st sen att	2nd sen att	joint
ricci	118	5	6	1	68 in race	—	—
credit	1000	21	59	2	690 in sex	851 in age	625
income	30162	14	99	2	25933 in race	20380 in sex	18038
ppr	6167	11	402	2	4994 in sex	2100 in race	1620
ppvr	4010	11	328	2	3173 in sex	1452 in race	1119

Baseline fairness-aware ensemble-based methods To evaluate the validity of *DR* in Section 2.1, ensemble classifiers in Section 3 are conducted using several ensemble methods including bagging [7], AdaBoost [19, 18], lightGBM [32], as well as two fairness-aware ensemble-based methods (that is, FairGBM [13] and AdaFair [26]). Besides, to verify the proposed oracle bounds and generalisation bounds in Section 2.3, bagging, AdaBoostM1, and SAMME are used in Section 3 to constitute an ensemble classifier on various kinds of individual classifiers including decision trees (DT), naive Bayesian (NB) classifiers, k-nearest neighbours (KNN) classifiers, Logistic Regression (LR), support vector machines (SVM), linear SVMs (linSVM), and multilayer perceptrons (MLP).

Implementation details We use standard five-fold cross-validation in these experiments, that is to say, in each iteration, the entire dataset is split into two parts, with 80% as the training set and 20% as the test set. Note that GPU is not necessarily required, and CPU would be sufficient to run our experiments. All experiments were run on our lab servers using a Docker image named "continuumio/anaconda3".[6]

C.2 Limitations and discussion

Discrimination mitigation techniques are meaningful given the wide applications of ML models nowadays, therefore, bringing such a technique with learning guarantees matters as it could provide theoretical foundations to boost fairness without potentially vain and repetitive practical attempts. In this view, our work throws away a brick in order to get a gem, showing that fairness can indeed be boosted with learning guarantees instead of being dependent on specific (hyper-)parameters. The proposed *discriminative risk (DR)* measure and the proposed oracle bounds are suitable for both binary and multi-class classification, enlarging the applicable fields, which is advantageous. However, there is also a limitation in *DR*. For instance, its computation is relevant to perturbed SAs, which means a randomness factor exists and may affect somehow computational results. Therefore, the effect of randomness on discriminative risk is worth exploring in the future.

[6]https://hub.docker.com/r/continuumio/anaconda3