Increasing Fairness via Combination with Learning Guarantees[1]

Yijun BIAN

National University of Singapore

26 October 2023

[1] Yijun Bian et al. "Increasing Fairness via Combination with Learning Guarantees". In: *arXiv preprint arXiv:2301.10813* (2023). Under Review.

Overview

Examples of bias

AI detectors were more likely to flag writing by international students (i.e., non-native speakers) as AI-generated[2]

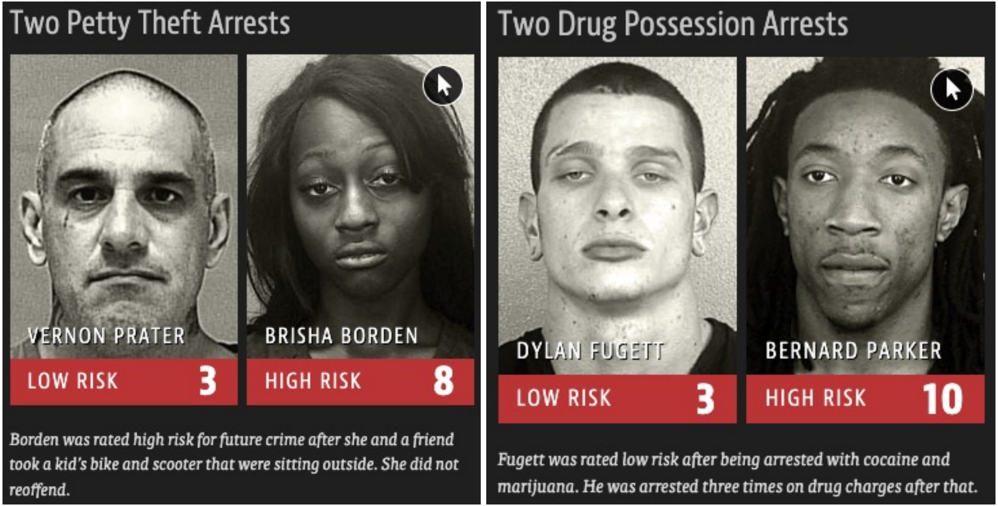
[2] Weixin Liang et al. "GPT detectors are biased against non-native English writers". In: *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*. 2023.

Examples of bias

When people of color have complex medical needs, they are less likely to be referred to programmes that provide more individualised care[2]

[2] Linda Nordling. "A fairer way forward for AI in health care". In: *Nature* 573.7775 (2019), S103–S103.

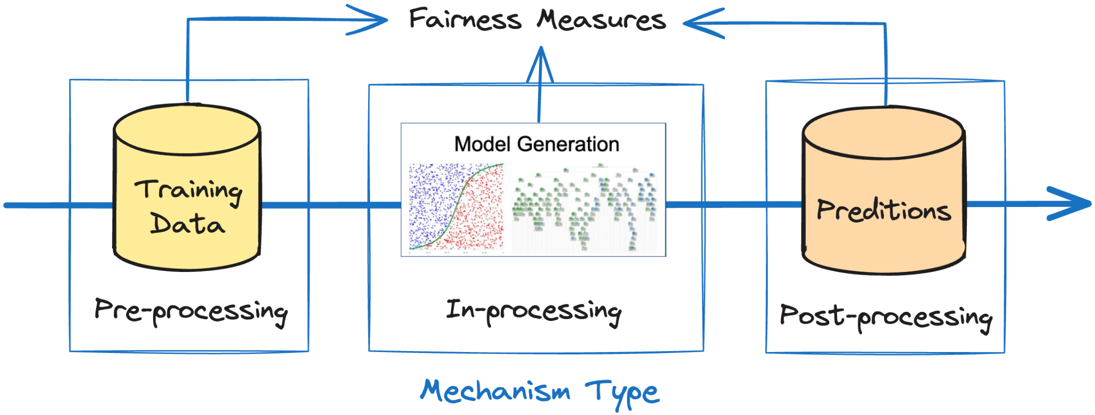
Examples of bias



Black defendants were mislabelled as high risk more often than white defendants[2]

[2]Lorenzo Belenguer. "AI bias: exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry". In: *AI and Ethics* 2.4 (2022), pp. 771–787.

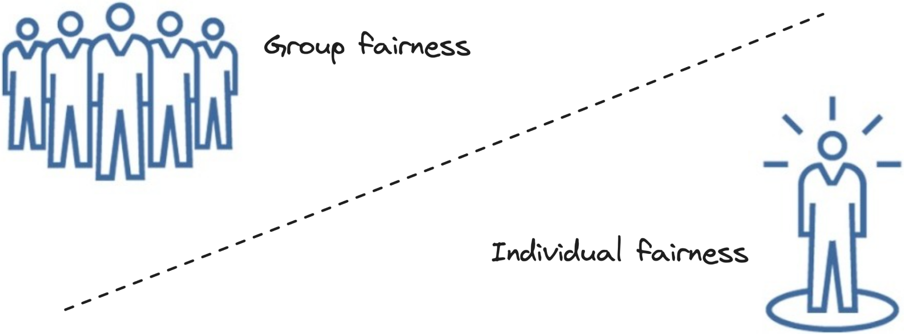
Mechanisms to enhance fairness[3]



Pre- and *post-processing mechanisms* normally function by manipulating input or output, while *inprocessing mechanisms* introduce fairness constraints into training procedures or algorithmic objectives

[3] Simon Caton and Christian Haas. "Fairness in machine learning: A survey". In: *ACM Comput Surv* (2020); Sorelle A Friedler et al. "A comparative study of fairness-enhancing interventions in machine learning". In: *FAT*. Atlanta, GA, USA: Association for Computing Machinery, 2019, pp. 329–338; Cynthia Dwork et al. "Decoupled classifiers for group-fair and efficient machine learning". In: *FAT*. vol. 81. PMLR, 2018, pp. 119–133.

Types of fairness measures



Group fairness[4] focuses on statistical/demographic equality among groups defined by sensitive attributes, while *individual fairness* follows a principle that "similar individuals should be evaluated or treated similarly."

[4]Michael Feldman et al. "Certifying and removing disparate impact". In: *SIGKDD*. Sydney, NSW, Australia: Association for Computing Machinery, 2015, pp. 259–268; Pratik Gajane and Mykola Pechenizkiy. "On formalizing fairness in prediction with machine learning". In: *FAT/ML*. 2018; Moritz Hardt, Eric Price, and Nathan Srebro. "Equality of opportunity in supervised learning". In: *NIPS*. vol. 29. Barcelona, Spain: Curran Associates Inc., 2016, pp. 3323–3331; Alexandra Chouldechova. "Fair prediction with disparate impact: A study of bias in recidivism prediction instruments". In: *Big Data* 5.2 (2017), pp. 153–163; Sahil Verma and Julia Rubin. "Fairness definitions explained". In: *FairWare*. IEEE. 2018, pp. 1–7.

Our target in this work

Research gap

- The hard compatibility among these measures means that unfair decisions may still exist even if one of them is satisfied[5]
- The possibility of theoretical guarantees of boosting fairness is rarely discussed in the existing fairness-aware ensemble-based methods[6]

[5] Solon Barocas, Moritz Hardt, and Arvind Narayanan. *Fairness and machine learning*. fairmlbook.org, 2019; Richard Berk et al. "Fairness in criminal justice risk assessments: The state of the art". In: *Sociol Methods Res* 50.1 (2021), pp. 3–44; Geoff Pleiss et al. "On fairness and calibration". In: *NIPS*. vol. 30. 2017; Hardt, Price, and Srebro, see n. 4.

[6] Vasileios Iosifidis and Eirini Ntoutsi. "AdaFair: Cumulative fairness adaptive boosting". In: *CIKM*. New York, NY, USA: ACM, 2019, pp. 781–790; Wenbin Zhang et al. "FARF: A fair and adaptive random forests classifier". In: *PAKDD*. Springer. 2021, pp. 245–256; André F Cruz et al. "FairGBM: Gradient Boosting with Fairness Constraints". In: *ICLR*. 2023.

Our target in this work

Research gap

- The hard compatibility among these measures means that unfair decisions may still exist even if one of them is satisfied[5]
- The possibility of theoretical guarantees of boosting fairness is rarely discussed in the existing fairness-aware ensemble-based methods[6]

***Questions* that we endeavour to answer**

1. *How to properly measure the discriminative level of a classifier from both individual and group fairness aspects?*
2. *Can fairness be boosted with some learning guarantee? Will* COMBINATION *help mitigate discrimination in multiple biassed individual classifiers?*

[5] Barocas, Hardt, and Narayanan, see n. 5; Berk et al., see n. 5; Pleiss et al., see n. 5; Hardt, Price, and Srebro, see n. 4.
[6] Iosifidis and Ntoutsi, see n. 6; Zhang et al., see n. 6; Cruz et al., see n. 6.

Overview

1. Background
2. **Methodology**
3. Discussions
4. Appendix

Research question recap

1. *How to properly measure the discriminative level of a classifier <u>from both individual and group fairness aspects</u>?*

Discriminative risk (DR) —from an individual aspect

Following the principle of individual fairness, the fairness quality of one hypothesis[7] $f(\cdot)$ could be evaluated by

$$
\begin{aligned}
\ell_{\text{fair}}(f, \boldsymbol{x}) &= \overbrace{\mathbb{I}}^{\text{the indicator function}}\big(f(\overbrace{\breve{\boldsymbol{x}}}^{\text{general attributes}}, \overbrace{\boldsymbol{a}}^{\text{sensitive attribute(s)}}) \neq f(\overbrace{\breve{\boldsymbol{x}}}^{\text{general attributes}}, \overbrace{\tilde{\boldsymbol{a}}}^{\text{sensitive attribute(s) that are slightly disturbed}})\big) \\
&= \mathbb{I}\big(\underbrace{f(\breve{\boldsymbol{x}}, \boldsymbol{a})}_{\text{model prediction on the raw instance}} \neq \underbrace{f(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}})}_{\text{model prediction when only sensitive attribute(s) are changed}}\big)\,,
\end{aligned} \tag{1}
$$

similarly to the 0/1 loss. Note that Eq. (1) is evaluated on only one instance with sensitive attributes $\boldsymbol{x}$.

[7]The hypothesis used in this equation could indicate an individual classifier or an ensemble classifier.

Discriminative risk (DR) —from a group aspect

To describe this characteristic of the hypothesis on multiple instances (aka. from a group level), then the empirical discriminative risk on one dataset S is expressed as

$$\hat{\mathcal{L}}_{\text{fair}}(f,S) = \frac{1}{n}\sum_{i=1}^{n} \underbrace{\ell_{\text{fair}}(f,\boldsymbol{x}_i)}_{\text{discriminative risk of } f(\cdot) \text{ on one instance}} \,, \tag{2}$$

and the true discriminative risk[8] of the hypothesis over a data distribution is

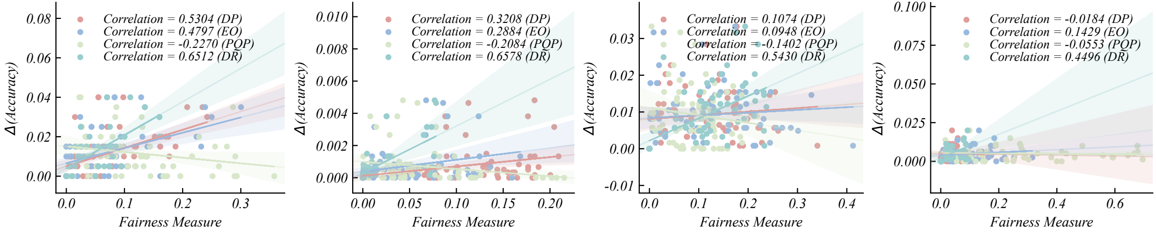
$$\mathcal{L}_{\text{fair}}(f) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\underbrace{\ell_{\text{fair}}(f,\boldsymbol{x})}_{\text{discriminative risk of } f(\cdot) \text{ on one instance}}] \,, \tag{3}$$

respectively.

[8]The instances from S 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 $\mathcal{D}$.

Empirical results of *DR* in comparison with group fairness measures[9,10]

Observation: DR captures better the characteristic of the changed treatment



(a) Credit **(b)** Income **(c)** PPR **(d)** PPVR

Figure 1: Comparison of the proposed DR with three group fairness measures, that is, DP, EO, and PQP. (a–d) Scatter diagrams with the degree of correlation on the credit, income, ppr, and ppvr datasets, respectively, where the x- and y-axes are different fairness measures and the variation of accuracy between the raw and disturbed data.

---

[9] They are demographic parity (DP) (Feldman et al., see n. 4; Gajane and Pechenizkiy, see n. 4), equality of opportunity (EO) (Hardt, Price, and Srebro, see n. 4), and predictive quality parity (PQP) (Chouldechova, see n. 4; Verma and Rubin, see n. 4).

[10] Five public datasets that we use include Ricci, Credit, Income, PPR, and PPVR, aka. Propublica-Recidivism and Propublica-Violent-Recidivism.

Research question recap

2. *Can fairness be boosted with some learning guarantee? Will* COMBINATION *help mitigate discrimination in multiple biassed individual classifiers?*

Oracle bounds of fairness

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,

$$\ell_{\text{fair}}(\mathbf{wv}_\rho, \boldsymbol{x}) \leqslant \mathbb{I}\Big(\mathbb{E}_\rho\big[\mathbb{I}(f(\breve{\boldsymbol{x}}, \boldsymbol{a}) \neq f(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}})) \big] \geqslant 0.5 \Big) \,. \tag{4}$$

discriminative risk of an ensemble $\mathbf{wv}_\rho(\cdot)$

that is, $\ell_{\text{fair}}(f, \boldsymbol{x})$
discriminative risk of an individual classifier $f(\cdot)$ on one instance $\boldsymbol{x}$

Meaning of $\mathbf{wv}_\rho(\cdot)$

Ensemble classifiers (via *weighted voting*)

- take a weighted combination of predictions by hypotheses, and
- predict a label that receives the largest number of votes

In other words, the ρ-weighted majority vote $\mathbf{wv}_\rho(\cdot)$ predicts

$$\mathbf{wv}_\rho(\boldsymbol{x}) = \operatorname*{argmax}_{y \in \mathcal{Y}} \mathbb{E}_\rho[\mathbb{I}[f(\boldsymbol{x}) = y)] \,,$$

where ρ corresponds to a potential ensemble over a hypothesis space.

Oracle bounds of fairness

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,

$$\underbrace{\ell_{\text{fair}}(\mathbf{wv}_\rho, x)}_{\text{discriminative risk of an ensemble } \mathbf{wv}_\rho(\cdot)} \leqslant \mathbb{I}\Big(\mathbb{E}_\rho\big[\, \underbrace{\mathbb{I}(f(\breve{x}, a) \neq f(\breve{x}, \tilde{a}))}_{\text{that is, } \ell_{\text{fair}}(f,x) \text{: discriminative risk of an individual classifier } f(\cdot) \text{ on one instance } x} \,\big] \geqslant 0.5 \Big)\,. \tag{4}$$

Theorem 1 (First-order oracle bound)

$$\underbrace{\mathcal{L}_{fair}(\mathbf{wv}_\rho)}_{\text{discriminative risk of an ensemble } \mathbf{wv}_\rho} \underbrace{\leqslant 2}_{\text{the worst case is controlled to a constant multiple}} \mathbb{E}_\rho[\, \underbrace{\mathcal{L}_{fair}(f)}_{\text{discriminative risk of an individual classifier } f} \,]\,. \tag{5}$$

Tandem discriminative risk

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,[11] by

hypothesis $f(\cdot)$ predicts differently for similar instances

hypothesis $f'(\cdot)$ also predicts differently for them

$$\boxed{\ell_{\text{fair}}(f,f',\boldsymbol{x})} = \mathbb{I}\big(f(\breve{\boldsymbol{x}}, \boldsymbol{a}) \neq f(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}}) \boxed{\wedge} f'(\breve{\boldsymbol{x}}, \boldsymbol{a}) \neq f'(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}}) \big) \,. \tag{6}$$

tandem discriminative risk

discriminative risks present in both of them

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

$$\ell_{\text{fair}}(f, \boxed{f}, \boldsymbol{x}) = \boxed{\ell_{\text{fair}}(f, \boldsymbol{x})} \,. \tag{7}$$

when $f'(\cdot)$ and $f(\cdot)$ are identical

discriminative risk of $f(\cdot)$

[11] Andrés R Masegosa et al. "Second order PAC-Bayesian bounds for the weighted majority vote". In: *NeurIPS*. vol. 33. Curran Associates, Inc., 2020, pp. 5263–5273.

Oracle bounds of fairness (cont.)

Then the expected tandem fairness quality is defined by $\mathcal{L}_{\text{fair}}(f,f') = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell_{\text{fair}}(f,f',x)]$.

Theorem 3 (Second-order oracle bound)

$$\overbrace{\mathcal{L}_{fair}(\mathbf{wv}_\rho)}^{\textit{discriminative risk of an ensemble } \mathbf{wv}_\rho} \underbrace{\leqslant 4}_{\textit{the worst case is controlled to a constant multiple}} \mathbb{E}_{\rho^2}[\overbrace{\mathcal{L}_{fair}(f,f')}^{\textit{tandem discriminative risk of two individuals } f \textit{ and } f'}] \,. \tag{8}$$

Lemma 2

In multi-class classification,

$$\mathbb{E}_{\rho^2}[\underbrace{\mathcal{L}_{\text{fair}}(f,f')}_{\text{the expected tandem discriminative risk}}] = \mathbb{E}_{\mathcal{D}}[\mathbb{E}_\rho[\overbrace{\ell_{\text{fair}}(f,x)}^{\text{discriminative risk of } f(\cdot)}]^2] \,. \tag{9}$$

♣

Empirical results of oracle bounds

Observation: The *discriminative risk (DR) of an ensemble* is indeed smaller than *the bounds presented in Theorems 1 and 3* in most cases, indicating that these inequalities are reliable

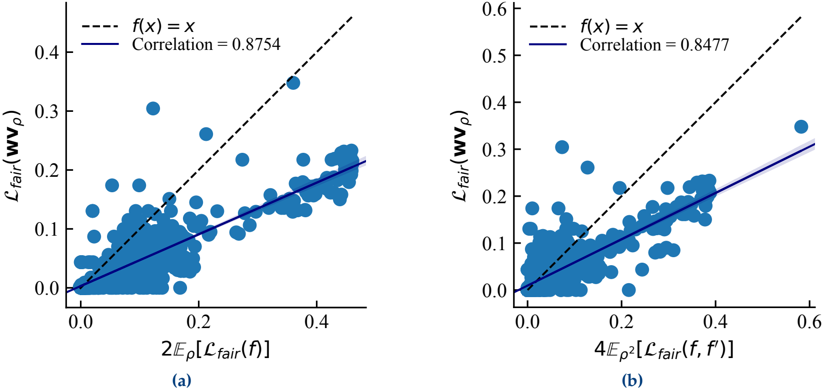


Figure 2: Correlation for oracle bounds. (a–b) Correlation between $\mathcal{L}_{\text{fair}}(\mathbf{wv}_\rho)$ and oracle bounds, where $\mathcal{L}_{\text{fair}}(\mathbf{wv}_\rho)$ is indicated on the vertical axis and the horizontal axes represent the right-hand sides of inequalities (5), and (8), respectively.

Overview

Summary[12]

RQ 2. *Can fairness be boosted with some learning guarantee? Will* COMBINATION *help mitigate discrimination in multiple biassed individual classifiers?*

Ensemble combination: fairness can be boosted without *being dependent on specific (hyper-)parameters*

$$\mathcal{L}_{\text{fair}}(\mathbf{wv}_\rho) \leqslant 2\,\mathbb{E}_\rho[\,\mathcal{L}_{\text{fair}}(f)\,] \qquad \text{cf. Theorem 1}$$

$$\mathcal{L}_{\text{fair}}(\mathbf{wv}_\rho) \leqslant 4\,\mathbb{E}_{\rho^2}[\,\mathcal{L}_{\text{fair}}(f,f')\,] \qquad \text{cf. Theorem 3}$$

[12] **P.S.** *Please refer to our paper for full methodology and empirical results*

Summary[12]

RQ 1. *How to properly measure the discriminative level of a classifier from both individual and group fairness aspects?*

Discriminative risk (DR) is proposed, that is,

$$\ell_{\text{fair}}(f, x) = \mathbb{I}(f(\breve{x}, a) \neq f(\breve{x}, \tilde{a}))\,.$$

DR is widely applicable, with two reasons enlarging its applicable fields/scenarios:

1. suitable for both binary and multi-class classification
2. allows one or multiple sensitive attributes, and each sensitive attribute allows binary and multiple values

[12] **P.S.** *Please refer to our paper for full methodology and empirical results*

Future work

Limitations

1. The computational results of DR may be affected somehow by a randomness factor
2. The degree of influence due to the number of values in sensitive attributes may vary, although its property remains

PROS

1. *Discriminative risk (DR)* is widely applicable
2. *Ensemble combination*: fairness can be boosted without *being dependent on specific (hyper-)parameters*

Thanks! Questions?

Overview

1. Background
2. Methodology
3. Discussions
4. Appendix

Sources of bias[13]

Biases from data

Data collected from

- biassed device measurements
- erroneous reports
- historically biassed human decisions
- or other reasons

[13]Verma and Rubin, see n. 4.

Sources of bias[13]

Biases from data

Biases from algorithms[a]

[a] e.g., caused by proxy attributes for sensitive attributes or tendentious algorithmic objectives

[13] Verma and Rubin, see n. 4.

Ensemble combination

The *weighted voting* prediction by an ensemble of m trained individual classifiers parameterised by a weight vector $\rho = [w_1, w_2, ..., w_m]^\mathsf{T} \in [0,1]^m$, such that $\sum_{j=1}^m w_j = 1$, wherein w_j is the weight of individual classifier $f_j(\cdot)$, is given by

$$\underset{\text{ensemble combination via weighted vote}}{\underbrace{\mathbf{wv}_\rho(\boldsymbol{x})}} = \operatorname*{argmax}_{y \in \mathcal{Y}} \sum_{j=1}^m \overset{\text{weight corresponding to } f_j(\cdot)}{\overbrace{w_j}}\, \mathbb{I}\big(\overset{\text{individual classifier}}{\overbrace{f_j(\boldsymbol{x})}} = \underset{\text{optional choice of labels}}{\underbrace{y}} \big) \,. \tag{10}$$

where a function $f \in \mathcal{F} : \mathcal{X} \mapsto \mathcal{F}$ denotes a hypothesis in a space of hypotheses $\mathcal{F}$. Note that ties are resolved arbitrarily.

Ensemble combination

The *weighted voting* prediction by an ensemble of m trained individual classifiers parameterised by a weight vector $\rho = [w_1, w_2, ..., w_m]^\mathsf{T} \in [0,1]^m$, such that $\sum_{j=1}^m w_j = 1$, wherein w_j is the weight of individual classifier $f_j(\cdot)$, is given by

$$\underbrace{\mathbf{wv}_\rho(\boldsymbol{x})}_{\text{ensemble combination via weighted vote}} = \operatorname*{argmax}_{y\in\mathcal{Y}} \sum_{j=1}^{m} \overbrace{w_j}^{\text{weight corresponding to } f_j(\cdot)} \mathbb{I}\big(\overbrace{f_j(\boldsymbol{x})}^{\text{individual classifier}} = \underbrace{y}_{\text{optional choice of labels}} \big) \,. \tag{10}$$

where a function $f \in \mathcal{F} : \mathcal{X} \mapsto \mathcal{F}$ denotes a hypothesis in a space of hypotheses $\mathcal{F}$. Note that ties are resolved arbitrarily. Ensemble classifiers predict by taking a weighted combination of predictions by hypotheses from $\mathcal{F}$, and the ρ-weighted majority vote $\mathbf{wv}_\rho(\cdot)$ predicts

$$\mathbf{wv}_\rho(\boldsymbol{x}) = \operatorname*{argmax}_{y\in\mathcal{Y}} \underbrace{\mathbb{E}_\rho}_{\text{potential } \rho \text{ corresponding to an ensemble over } [0,1]^m} [\mathbb{I}(f(\boldsymbol{x}) = y)] \,.$$

Oracle bounds regarding fairness for weighted vote

Theorem 4 (C-tandem oracle bound)

If $\mathbb{E}_\rho[\mathcal{L}_{fair}(f)] < 1/2$, *then*

the worst case is controlled, alternative bound based on Chebyshev-Cantelli inequality

$$\mathcal{L}_{fair}(\mathbf{wv}_\rho) \leqslant \frac{\mathbb{E}_{\rho^2}[\,\mathcal{L}_{fair}(f,f')\,] - \mathbb{E}_\rho[\,\mathcal{L}_{fair}(f)\,]^2}{\mathbb{E}_{\rho^2}[\,\mathcal{L}_{fair}(f,f')\,] - \mathbb{E}_\rho[\,\mathcal{L}_{fair}(f)\,] + \frac{1}{4}}\,. \tag{11}$$

discriminative risk of an ensemble $\mathbf{wv}_\rho$

tandem discriminative risk

discriminative risk

Oracle bounds regarding fairness for weighted vote

Theorem 4 (C-tandem oracle bound)

If $\mathbb{E}_\rho[\mathcal{L}_{fair}(f)] < 1/2$, *then*

the worst case is controlled, alternative bound based on Chebyshev-Cantelli inequality

$$\mathcal{L}_{fair}(\mathbf{wv}_\rho) \leqslant \frac{\mathbb{E}_{\rho^2}[\, \mathcal{L}_{fair}(f,f') \,] - \mathbb{E}_\rho[\, \mathcal{L}_{fair}(f) \,]^2}{\mathbb{E}_{\rho^2}[\, \mathcal{L}_{fair}(f,f') \,] - \mathbb{E}_\rho[\, \mathcal{L}_{fair}(f) \,] + \frac{1}{4}} \,. \tag{11}$$

discriminative risk of an ensemble $\mathbf{wv}_\rho$

tandem discriminative risk

discriminative risk

All oracle bounds are expectations that can only be estimated on finite samples instead of being calculated precisely. They could be transformed into empirical bounds via PAC-Bayesian 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 5 and 6.

PAC-Bayesian bounds for the weighted vote

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 a single hypothesis $f(\cdot)$,

the worst case is controlled with a specific bound

$$\mathcal{L}_{fair}(f) \leqslant \hat{\mathcal{L}}_{fair}(f,S) + \sqrt{\frac{1}{2n}\ln\frac{1}{\delta}}\,. \tag{12}$$

discriminative risk of a hypothesis

empirical discriminative risk of this hypothesis

Theorem 6

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 $\mathcal{F}$,

the worst case is controlled with a specific bound

$$\mathcal{L}_{fair}(\mathbf{wv}_\rho) \leqslant \hat{\mathcal{L}}_{fair}(\mathbf{wv}_\rho,S) + \sqrt{\frac{1}{2n}\log\frac{|\mathcal{F}|}{\delta}}\,. \tag{13}$$

discriminative risk of an ensemble

empirical discriminative risk of this ensemble

Our distinction

Despite the similar names of "first- and second-order oracle bounds" from our inspiration,[14] the essences of our bounds are distinct from theirs. To be specific, their work investigates the bounds for generalisation error and is not relevant to fairness issues, while ours focus on the theoretical support for bias mitigation. In other words, their bounds are based on the 0/1 loss

$$\overbrace{\ell_{\text{err}}(f,\boldsymbol{x})}^{\text{the loss of the classifier } f(\cdot)} = \mathbb{I}\big(\underbrace{f(\boldsymbol{x})}_{\text{model prediction on the raw data}} \overbrace{\neq y}^{\text{label of this instance, which means it makes mistakes on the instance}} \big)\,, \tag{14}$$

while ours are built on $\ell_{\text{fair}}(f,\boldsymbol{x})$ in Eq. (1), that is,

$$\underbrace{\ell_{\text{fair}}(f,\boldsymbol{x})}_{\text{the discriminative risk of } f(\cdot)} = \mathbb{I}\big(\overbrace{f(\breve{\boldsymbol{x}},\boldsymbol{a})}^{\text{model prediction on the raw data}} \neq \underbrace{f(\breve{\boldsymbol{x}},\tilde{\boldsymbol{a}})}_{\text{model prediction when only sensitive attribute(s) are changed}} \big)\,.$$

[14] Masegosa et al., see n. 11.