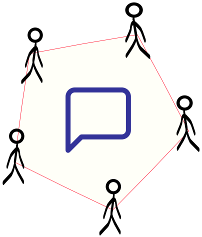
Increasing *Fairness* via Combination with Learning Guarantees

Yijun Bian

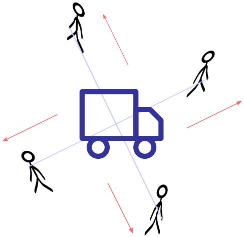
Machine Learning Section
Department of Computer Science
University of Copenhagen

4 November 2024

Ensemble diversity *(my research in the past)*



brainstorm
benefit by mutual discussion



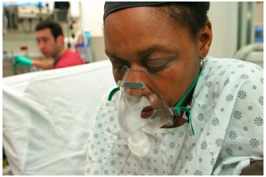
drift apart
where to go next? no idea

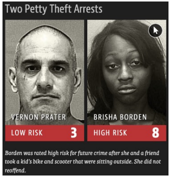
[1]Examples

1. *Panel interview*
2. *Group work*
3. *Co-supervisors*
4. *DIKU faculty board*

[1] Yijun Bian and Huanhuan Chen. “When does diversity help generalization in classification ensembles?” In: *IEEE Trans Cybern* 52.9 (2022), pp. 9059–9075. DOI: 10.1109/TCYB.2021.3053165; Yijun Bian et al. “Ensemble pruning based on objection maximization with a general distributed framework”. In: *IEEE Trans Neural Netw Learn Syst* 31.9 (2020), pp. 3766–3774. DOI: 10.1109/TNNLS.2019.2945116.

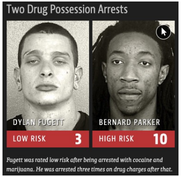
Fairness *(my current research)*





Two Petty Theft Arrests

Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.



Two Drug Possession Arrests

Fugett was rated low risk after being arrested with cocaine and marijuana. He was arrested three times on drug charges after that.

[2]**Examples** of discrimination/bias [3,4,5]

[2] Yijun Bian and Yujie Luo. "Does Machine Bring in Extra Bias in Learning? Approximating Fairness in Models Promptly". In: *arXiv preprint arXiv:2405.09251* (2024); Yijun Bian, Yujie Luo, and Ping Xu. "Approximating Discrimination Within Models When Faced With Several Non-Binary Sensitive Attributes". In: *arXiv preprint arXiv:2408.06099* (2024). Under Review.

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

[4] When people of color have complex medical needs, they are less likely to be referred to programmes that provide more individualised care (Linda Nordling. "A fairer way forward for AI in health care". In: *Nature* 573.7775 [2019], S103–S103)

[5] Black defendants were mislabelled as high risk more often than white defendants (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)

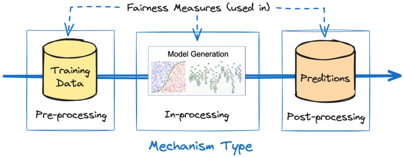
Challenging

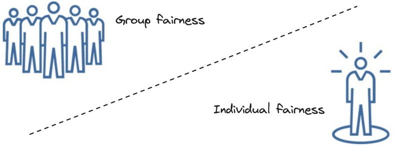
- Fairness definitions and measures/metrics [a,b]
- Incompatibility among fairness measures
- Multi-attribute fairness protection
- The trade-off between fairness and accuracy
- Fairness estimation based on a finite sample
- Insufficient data

Our motivation

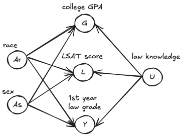
[a] *Pre-* and *post-processing mechanisms* normally function by manipulating input or output, while *inprocessing mechanisms* introduce fairness constraints into training procedures or algorithmic objectives

[b] *Group fairness* 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."





Example/Illustration:
Law school success



As	Ar	G	L	Y	U
male	black	👎	👍	👎	(👍)

1. Computing unobserved variables in causal model
2. Change A (that is, sensitive attribute(s))
3. Recompute observed variables in causal model

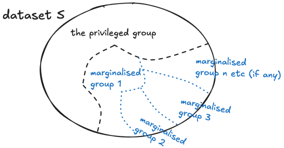
classifier
hat{Y}(G,L)

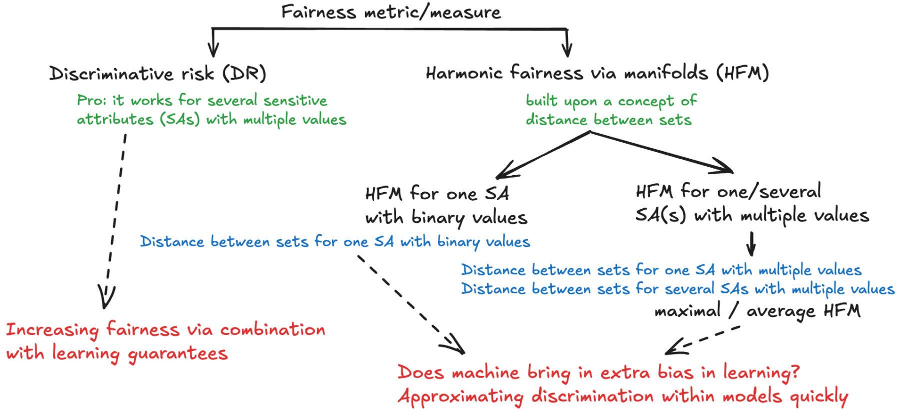
As	Ar ← a'	G	L	Y	U
male	white	👎	👍	👎	👍

As	Ar ← a'	$G_{[Ar \leftarrow a']}$	$L_{[Ar \leftarrow a']}$	$Y_{[Ar \leftarrow a']}$	$U_{[Ar \leftarrow a']}$
male	white	👍	👍	👍	👍

My research

My research in this direction, up to the present





Fairness metric/measure

Discriminative risk (DR)

Pro: it works for several sensitive attributes (SAs) with multiple values

Harmonic fairness via manifolds (HFM)

built upon a concept of distance between sets

HFM for one SA with binary values

HFM for one/several SA(s) with multiple values

Distance between sets for one SA with binary values

Distance between sets for one SA with multiple values
Distance between sets for several SAs with multiple values
maximal / average HFM

Increasing fairness via combination with learning guarantees

Does machine bring in extra bias in learning?
Approximating discrimination within models quickly

Research question recap

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?*

3. *How to utilise the proposed metric to obtain better ensemble classifiers?*

Discriminative risk (DR) —from an individual aspect

Following the principle of individual fairness, with an instance denoted by $\boldsymbol{x} = (\breve{\boldsymbol{x}}, \boldsymbol{a})$, the fairness quality of one hypothesis[6] $f(\cdot)$ could be evaluated by

sensitive attribute(s)

sensitive attribute(s) that are slightly disturbed

non-sensitive attributes

$$\ell_{\text{bias}}(f,\boldsymbol{x}) = \mathbb{I}\Big(\underbrace{f(\breve{\boldsymbol{x}},\boldsymbol{a})}_{\text{model prediction on the raw instance}} \neq \underbrace{f(\breve{\boldsymbol{x}},\tilde{\boldsymbol{a}})}_{\text{model prediction when only sensitive attribute(s) are changed}}\Big) \tag{1}$$

the indicator function

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

[6] 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

$$\hat{\mathcal{L}}_{\text{bias}}(f,S) = \frac{1}{n}\sum_{i=1}^{n} \ell_{\text{bias}}(f,\boldsymbol{x}_i) \,, \tag{2}$$

discriminative risk of $f(\cdot)$ on one instance

and the true discriminative risk[7] of the hypothesis over a data distribution is

$$\mathcal{L}_{\text{bias}}(f) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\, \ell_{\text{bias}}(f,\boldsymbol{x}) \,] \,, \tag{3}$$

discriminative risk of $f(\cdot)$ on one instance

respectively.

Note that the empirical DR on S is an unbiassed estimation of the true DR.

[7] 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 $\mathcal{D}$.

A property of *DR*

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

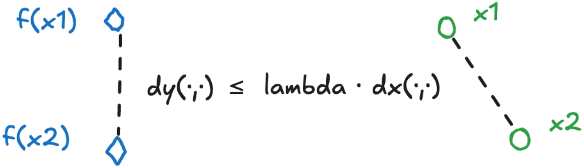
$$\hat{\mathcal{L}}_{\text{bias}}(f,S) = \frac{1}{n}\sum_{i=1}^{n} \ell_{\text{bias}}(f,\boldsymbol{x}_i)$$

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

1. For one random variable X representing instances, $\ell_{\text{bias}}(f,\boldsymbol{x})$ could be viewed as a new random variable obtained by using a few fixed operations on X, recorded as Y.
2. 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.
3. 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 $\mathsf{X}\sim\mathcal{D}$.
4. Therefore, it could be easily seen that the former is an unbiased estimation of the latter.

Our distinction

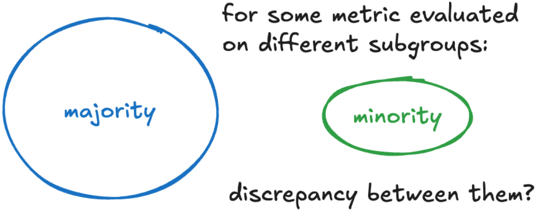
- Two distinctions from *individual fairness*
  1. the choice of similarity/distance metric
  2. instance pairs to be compared



- Two distinctions from *group fairness*
- Five distinctions from *causal fairness*

Our distinction

- Two distinctions from *individual fairness*
- Two distinctions from *group fairness*
  1. a way of computing the difference
  2. one sensitive attribute (usually with binary values)



for some metric evaluated on different subgroups:

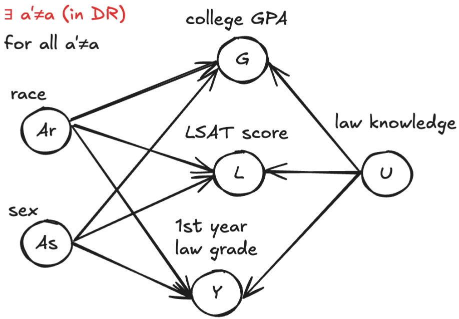
discrepancy between them?

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

- Five distinctions from *causal fairness*

Our distinction

- Two distinctions from *individual fairness*
- Two distinctions from *group fairness*
- Five distinctions from *causal fairness*
  1. both counterfactual fairness and proxy discrimination work for only one sensitive attribute (although possibly including multiple values)
  2. causal models/graphs vs. quantitative measure (easy to calculate)
  3. non-sensitive attributes may be changed as well
  4. conditions for achieving them are stronger
  5. *DR* can be proved to be bounded



Our distinction

- Two distinctions from *individual fairness*
- Two distinctions from *group fairness*
- Five distinctions from *causal fairness*

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

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

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

$$\begin{aligned}\mathcal{L}'_{\text{bias}}(f) = \ & |\mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}|\boldsymbol{a}=1}[\ell_{\text{bias}}(f,\boldsymbol{x})] \\ & - \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}|\boldsymbol{a}=0}[\ell_{\text{bias}}(f,\boldsymbol{x})]|\end{aligned}$$

- Similarities that *DR* shares with the existing fairness measures

Cancellation-of-bias effect[8] in ensemble combination

- Inspired by existing work for error rates and oracle bounds
- First- and second-order oracle bounds concerning fairness
- Similarly to the *cancellation-of-errors* effect in ensemble combination

The DR of an ensemble can be bounded by a constant times the DR of the individual classifiers

[8] Yijun Bian and Kun Zhang. "Increasing Fairness via Combination with Learning Guarantees". In: *arXiv preprint arXiv:2301.10813v1* (2023).

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

$$\underbrace{\mathbf{wv}_\rho(\boldsymbol{x})}_{\text{ensemble combination via weighted vote}} = \operatorname*{argmax}_{y \in \mathcal{Y}} \sum_{j=1}^{m} \overbrace{w_j}^{\text{weight corresponding to } f_j(\cdot)} \mathbb{I}(\overbrace{f_j(\boldsymbol{x})}^{\text{individual classifier}} = \underbrace{y}_{\text{optional choice of labels}}) \,. \tag{6}$$

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

$$\mathbf{wv}_\rho(\boldsymbol{x}) = \operatorname*{argmax}_{y\in\mathcal{Y}} \sum_{j=1}^{m} w_j \mathbb{I}(f_j(\boldsymbol{x}) = y)\,. \tag{6}$$

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

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

potential ρ corresponding to an ensemble over $[0,1]^m$

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,

$$\underbrace{\ell_{\text{bias}}(\mathbf{wv}_\rho, \boldsymbol{x})}_{\text{discriminative risk of an ensemble } \mathbf{wv}_\rho(\cdot)} \leqslant \mathbb{I}\big(\mathbb{E}_\rho[\, \underbrace{\mathbb{I}(f(\breve{\boldsymbol{x}}, a) \neq f(\breve{\boldsymbol{x}}, \tilde{a}))}_{\text{that is, } \ell_{\text{bias}}(f, \boldsymbol{x}) \text{ discriminative risk of an individual classifier } f(\cdot) \text{ on one instance } \boldsymbol{x}} \,] \geqslant 0.5 \big)\,. \tag{7}$$

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,

$$\ell_{\text{bias}}(\mathbf{wv}_\rho, \boldsymbol{x}) \leqslant \mathbb{I}\big(\mathbb{E}_\rho[\, \mathbb{I}(f(\breve{\boldsymbol{x}}, \boldsymbol{a}) \neq f(\breve{\boldsymbol{x}}, \tilde{\boldsymbol{a}}))\,] \geqslant 0.5 \big)\,. \tag{7}$$

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

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

Theorem 1 (First-order oracle bound)

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

discriminative risk of an individual classifier f

$$\mathcal{L}_{bias}(\mathbf{wv}_\rho) \leqslant 2\, \mathbb{E}_\rho[\, \mathcal{L}_{bias}(f)\,]\,. \tag{8}$$

the worst case is controlled to a constant multiple

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,[9] by

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

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

$$\ell_{\text{bias}}(f,f',\boldsymbol{x}) = \mathbb{I}\big(f(\breve{\boldsymbol{x}},a) \neq f(\breve{\boldsymbol{x}},\tilde{a}) \wedge f'(\breve{\boldsymbol{x}},a) \neq f'(\breve{\boldsymbol{x}},\tilde{a}) \big)\,. \tag{9}$$

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

$$\ell_{\text{bias}}(f, f, \boldsymbol{x}) = \ell_{\text{bias}}(f, \boldsymbol{x})\,. \tag{10}$$

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

discriminative risk of $f(\cdot)$

[9] 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 $\mathcal{L}_{\text{bias}}(f,f') = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell_{\text{bias}}(f,f',x)]$.

Theorem 3 (Second-order oracle bound)

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

tandem discriminative risk of two individuals f and f'

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

the worst case is controlled to a constant multiple

Lemma 2

In multi-class classification,

discriminative risk of $f(\cdot)$

$$\mathbb{E}_{\rho^2}[\,\mathcal{L}_{\text{bias}}(f,f')\,] = \mathbb{E}_{\mathcal{D}}[\mathbb{E}_{\rho}[\,\ell_{\text{bias}}(f,x)\,]^2]\,. \tag{12}$$

♣

the expected tandem discriminative risk

Oracle bounds regarding fairness for weighted vote

Theorem 4 (C-tandem oracle bound)

If $\mathbb{E}_\rho[\mathcal{L}_{bias}(f)] < {}^1\!/_2$, *then*

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

$$\mathcal{L}_{bias}(\mathbf{wv}_\rho) \leqslant \frac{\mathbb{E}_{\rho^2}[\,\mathcal{L}_{bias}(f,f')\,] - \mathbb{E}_\rho[\,\mathcal{L}_{bias}(f)\,]^2}{\mathbb{E}_{\rho^2}[\,\mathcal{L}_{bias}(f,f')\,] - \mathbb{E}_\rho[\,\mathcal{L}_{bias}(f)\,] + \frac{1}{4}}\,. \tag{13}$$

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

discriminative risk

tandem 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 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 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

$$\underbrace{\mathcal{L}_{bias}(f)}_{\text{discriminative risk of a hypothesis}} \leqslant \underbrace{\hat{\mathcal{L}}_{bias}(f,S)}_{\text{empirical discriminative risk of this hypothesis}} + \sqrt{\frac{1}{2n}\ln\frac{1}{\delta}}\,. \tag{14}$$

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

$$\underbrace{\mathcal{L}_{bias}(\mathbf{wv}_\rho)}_{\text{discriminative risk of an ensemble}} \leqslant \underbrace{\hat{\mathcal{L}}_{bias}(\mathbf{wv}_\rho, S)}_{\text{empirical discriminative risk of this ensemble}} + \sqrt{\frac{1}{2n}\log\frac{|\mathcal{F}|}{\delta}}\,. \tag{15}$$

Our distinction

Despite the similar names of "first- and second-order oracle bounds" from our inspiration,[10] 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

the loss of the classifier $f(\cdot)$

label of this instance, which means it makes mistakes on the instance

$$\ell_{\text{err}}(f,x) = \mathbb{I}(f(x) \neq y)\,, \tag{16}$$

model prediction on the raw data

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

model prediction on the raw data

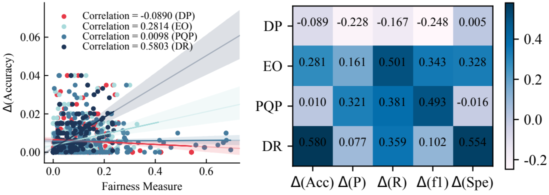
$$\ell_{\text{bias}}(f,x) = \mathbb{I}(f(\breve{x},a) \neq f(\breve{x},\tilde{a}))\,.$$

the discriminative risk of $f(\cdot)$

model prediction when only sensitive attribute(s) are changed

[10] Masegosa et al., see n. 9.

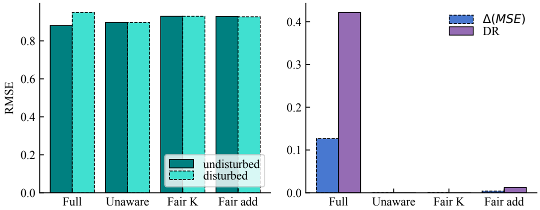
Validating the proposed fairness quality measure



	Δ(Acc)	Δ(P)	Δ(R)	Δ(f1)	Δ(Spe)
DP	-0.089	-0.228	-0.167	-0.248	0.005
EO	0.281	0.161	0.501	0.343	0.328
PQP	0.010	0.321	0.381	0.493	-0.016
DR	0.580	0.077	0.359	0.102	0.554

(a) **(b)**

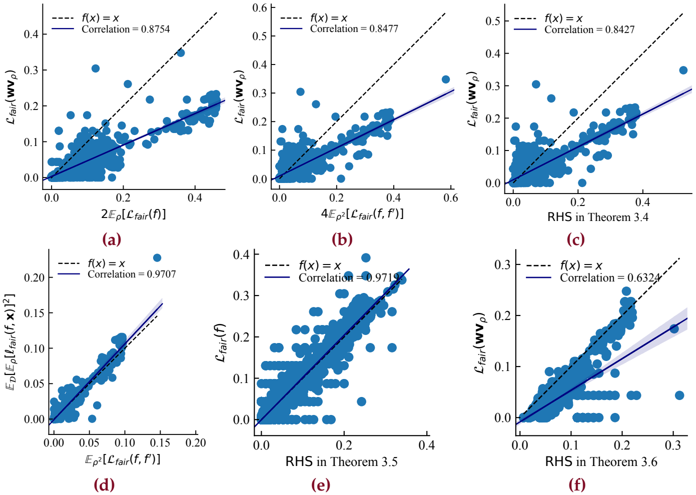
Figure 1: Comparison of the proposed discriminative risk (DR) with three group fairness measures, that is, DP, EO, and PQP. (a) Scatter diagrams with the degree of correlation, where the *x*- and *y*-axes are different fairness measures and the variation of accuracy between the raw and disturbed data.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



(a) **(b)**

Figure 2: Example: law school success. (a) Test MSE of different models, where ‘undisturbed‘ and ‘disturbed’ denote the results obtained from the original and disturbed data respectively. (b) The comparison between the change in MSE and *DR*, which suggests that DR $\approx$ 0 when the corresponding model satisfies or nearly satisfies counterfactual fairness.

Validating the oracle&PAC bounds



(a) (b) (c)

(d) (e) (f)

Figure 3: Correlation for oracle bounds and generalisation bounds. (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 (8), (11), and (13), respectively. (d) The horizontal and vertical axes in (d) denote the right- and left-hand sides in (12), respectively. (e–f) 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 (14) and (15) are indicated on the horizontal axes, respectively. Note that correlation here is calculated based on the results from all datasets.

Summary up to now[11,12]

***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*

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

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

[11] Bian and Zhang, see n. 8.

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

Summary up to now[11,12]

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,

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

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

[11] Bian and Zhang, see n. 8.

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

Summary up to now[11,12]

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,

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

DR is widely applicable

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

[11] Bian and Zhang, see n. 8.

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

Research question recap: Application

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?*

3. *How to utilise the proposed metric to obtain better ensemble classifiers?*

Commonly used group fairness measures

There are three commonly-used group fairness measures, that is, ***demographic parity (DP),***[13] ***equality of opportunity (EO),***[14] ***and predictive quality parity (PQP)***[15].

These three commonly used group fairness measures of one classifier $f(\cdot)$ are evaluated as

$$\begin{aligned}
\mathrm{DP}(f) &= \left|\mathbb{P}_{\mathcal{D}}[f(\boldsymbol{x})=1 \,|\, \boldsymbol{a}=1] - \mathbb{P}_{\mathcal{D}}[f(\boldsymbol{x})=1 \,|\, \boldsymbol{a}=0]\right| , && \text{(17a)}\\
\mathrm{EO}(f) &= \left|\mathbb{P}_{\mathcal{D}}[f(\boldsymbol{x})=1 \,|\, \boldsymbol{a}=1, y=1] - \mathbb{P}_{\mathcal{D}}[f(\boldsymbol{x})=1 \,|\, \boldsymbol{a}=0, y=1]\right|, && \text{(17b)}\\
\mathrm{PQP}(f) &= \left|\mathbb{P}_{\mathcal{D}}[y=1 \,|\, \boldsymbol{a}=1, f(\boldsymbol{x})=1] - \mathbb{P}_{\mathcal{D}}[y=1 \,|\, \boldsymbol{a}=0, f(\boldsymbol{x})=1]\right|, && \text{(17c)}
\end{aligned}$$

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

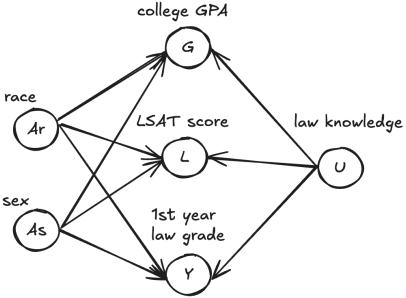
[13] Michael Feldman et al. "Certifying and removing disparate impact". In: *SIGKDD*. Sydney, NSW, Australia: Association for Computing Machinery, 2015, pp. 259–268; Pratik Gajane and Mykola Pechenizkiy. "On formalizing fairness in prediction with machine learning". In: *FAT/ML*. 2018.

[14] 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.

[15] 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.

Fairness based on causality

Example/Illustration:
Law school success



As	Ar	G	L	Y	U
male	black	👎	👍	👎	(👍)

1. Computing unobserved variables in causal model
2. Change A (that is, sensitive attribute(s))
3. Recompute observed variables in causal model

classifier
hat{Y}(G,L)

As	Ar ← a'	G	L	Y	U
male	white	👎	👍	👎	👍

As	Ar ←a'	$G_{[Ar \leftarrow a']}$	$L_{[Ar \leftarrow a']}$	$Y_{[Ar \leftarrow a']}$	$U_{[Ar \leftarrow a']}$
male	white	👍	👍	👍	👍

Counterfactual fairness[16] definition: A predictor $\hat{Y}$ is *counterfactually fair* if given observations $\mathcal{X}=x$ and $A=a$ we have that, $\mathbb{P}(\hat{Y}_{A\leftarrow a}=y \mid \mathcal{X}=x, A=a) = \mathbb{P}(\hat{Y}_{A\leftarrow a'}=y \mid \mathcal{X}=x, A=a)$, for all y and $a' \neq a$.

Proxy discrimination[17] definition: A predictor $\hat{Y}$ exhibits no *individual proxy discrimination* if given observations $\mathcal{X}=x$ and $A=a$ we have that, $\mathbb{P}(\hat{Y}=y \mid \mathrm{do}(A=a), \mathcal{X}=x)= \mathbb{P}(\hat{Y}=y \mid \mathrm{do}(A=a'), \mathcal{X}=x)$, for all y and $a' \neq a$.

[16] Matt J Kusner et al. "Counterfactual fairness". In: *NIPS*. vol. 30. Curran Associates, Inc., 2017, pp. 4069–4079.

[17] Niki Kilbertus et al. "Avoiding discrimination through causal reasoning". In: *NIPS*. vol. 30. 2017.