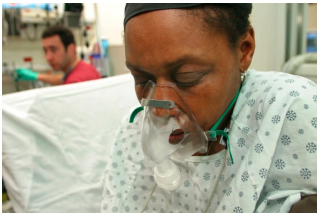
Study of FAIR ML *Models from a Theoretical Perspective*

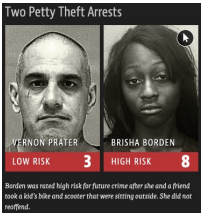
Yijun BIAN

Machine Learning Section
Department of Computer Science
University of Copenhagen

5 December 2024

Examples of bias [2,3,4]





Manuscripts under review[1]

1. **Yijun Bian*** and Kun Zhang. "Increasing fairness via combination with learning guarantees". In: *arXiv preprint arXiv:2301.10813v1* (2023).
2. **Yijun Bian**$^{\#\star}$ and Yujie Luo$^{\#}$. "Does machine bring in extra bias in learning? Approximating fairness in models promptly". In: *arXiv preprint arXiv:2405.09251* (2024).
3. **Yijun Bian**$^{\#\star}$, Yujie Luo$^{\#\star}$, and Ping Xu. "Approximating discrimination within models when faced with several non-binary sensitive attributes". In: *arXiv preprint arXiv:2408.06099* (2024).

[1] $^{\#}$Equal contribution; *corresponding author.

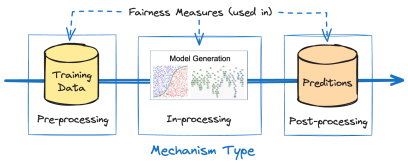
[2] AI detectors were more likely to flag writing by international students (i.e., non-native speakers) as AI-generated (Weixin Liang et al. "GPT detectors are biased against non-native English writers". In: *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*. 2023)

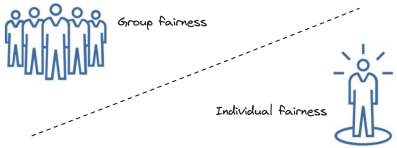
[3] When people of color have complex medical needs, they are less likely to be referred to programmes that provide more individualised care (Linda Nordling. "A fairer way forward for AI in health care". In: *Nature* 573.7775 [2019], S103–S103)

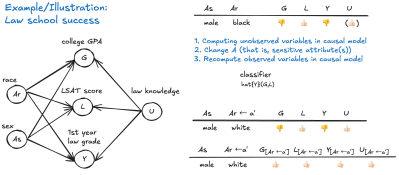
[4] Black defendants were mislabelled as high risk more often than white defendants (Lorenzo Belenguer. "AI bias: exploring discriminatory algorithmic decision-making models and the application of possible machine-centric solutions adapted from the pharmaceutical industry". In: *AI and Ethics* 2.4 [2022], 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







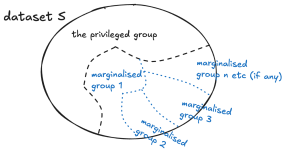
Our motivation

[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."

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[5]

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

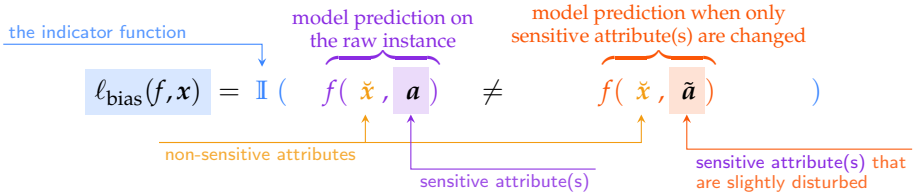
2. *How to evaluate the discrimination level within classifiers when facing more than one sensitive attribute (potentially with multiple values)?*

3. *How to make use of these proposed fairness measure/metric(s)?*

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

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



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

[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

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

($\ell_{\text{bias}}(f, \boldsymbol{x}_i)$: discriminative risk of $f(\cdot)$ on one instance)

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

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

($\ell_{\text{bias}}(f, \boldsymbol{x})$: 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*

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

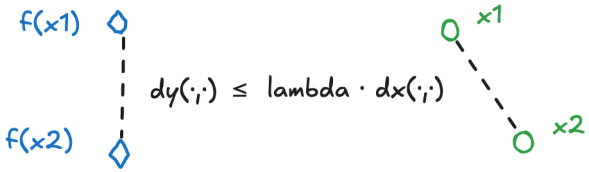
$$\hat{\mathcal{L}}_{\text{bias}}(f, S) = \frac{1}{n}\sum_{i=1}^{n} \ell_{\text{bias}}(f, \boldsymbol{x}_i)$$

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

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

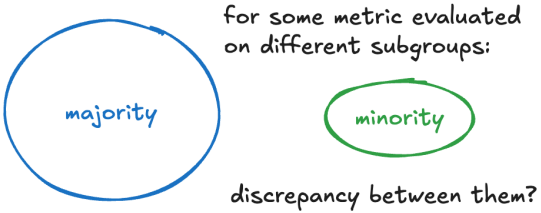
- Two distinctions from *individual fairness* measures
  1. relies on the choice of similarity/distance metric
  2. instance pairs in comparison coming from original data



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

Our distinction

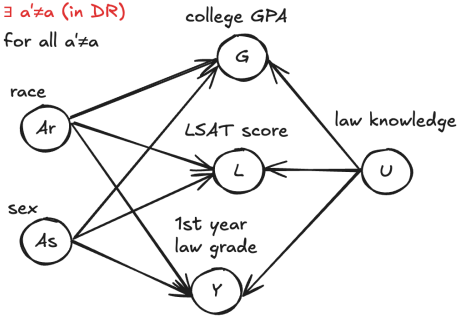
- Two distinctions from *individual fairness* measures
- Two distinctions from *group fairness* measures
  1. works for only one sensitive attribute (usually bi-valued)
  2. computing separately for each subgroup, then difference



- Five distinctions from *causal fairness*

Our distinction

- Two distinctions from *individual fairness* measures
- Two distinctions from *group fairness* measures
- Five distinctions from *causal fairness*
  1. works for only one sensitive attribute (although possibly multi-valued)
  2. based on causal models/graphs, not a quantitative measure
  3. non-sensitive attributes may vary with it in counterfactual fairness
  4. conditions for achieving them are stronger
  5. *DR* can be proved to be bounded vs. no such advantage



Our distinction

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

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

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

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

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

- Similarities that *DR* shares with the existing fairness measures
  - follows the same principle as *individual fairness* measures
  - is computed over a group of instances (like one dataset or a data distribution)
  - indicates the discrimination level from a statistical/demographic perspective

Validating *DR*, a fairness quality measure

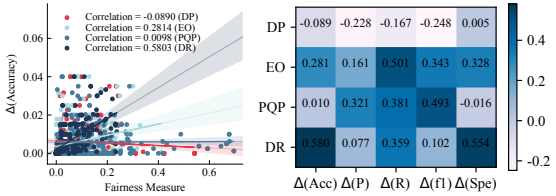


Figure 1: Comparison of the proposed discriminative risk (DR) with three group fairness measures, that is, DP, EO, and PQP. (a) Scatter diagrams with the degree of correlation, where the x- and y-axes are different fairness measures and the variation of accuracy between the raw and disturbed data. (b)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

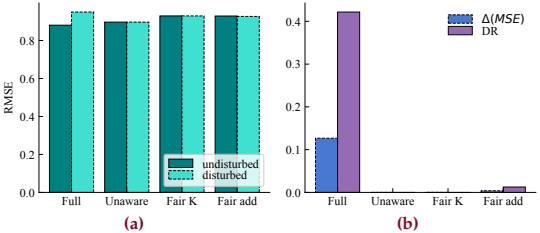


Figure 2: Example: law school success. (a) Test MSE of different models, where 'undisturbed' and 'disturbed' denote the results obtained from the original and disturbed data respectively. (b) The comparison between the change in MSE and *DR*, which suggests that DR$\approx$0 when the corresponding model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

Interim summary[8]

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,

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

Widely applicable, with two reasons enlarging its applicable fields/scenarios:

1. suitable for both binary and multi-class classification
2. allows one or more SAs, and each SA allows binary or multiple values

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 SAs may vary, although its property remains

[8] Bian and Zhang, see n. 5.

Research question recap[9]

2′. *How to evaluate the added discrimination introduced by ML models (on top of potential discrimination present in the raw data) properly?*

- *in the face of one sensitive attribute with binary values*
- *in the face of one sensitive attribute (with multiple values)*
- *in the face of several sensitive attributes (with multiple values)*

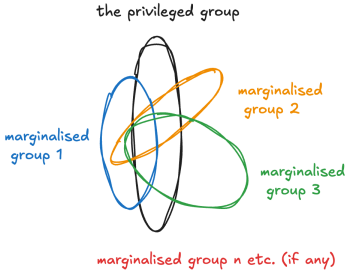
3′. *How to efficiently evaluate the added discrimination introduced in the learning process?*

- *A quick approximation of distances between sets for the Euclidean spaces*

[9] 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.

Harmonic Fairness via Manifolds (HFM)

If we view the *instances (with the same value of sensitive attributes)* as *data points on certain manifold(s)*, the manifold representing members from the marginalised/unprivileged group(s) is supposed to be as close as possible to that representing members from the privileged group.



To measure the fairness with respect to the sensitive attribute, we use a concept of 'distance of sets' introduced in mathematics, recorded as **D**, to *evaluate the discrepancy among groups divided by sensitive attributes*. Then a fairness measure[10] is built upon the concept of distances between sets.

[10] indicating difference from both individual- and group- apsects

Distance between sets —for one *bi-valued* SA

Given a specific distance metric $\mathbf{d}(\cdot,\cdot)$[11] on the feature space, the distance between two subsets is defined by

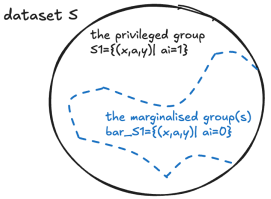
$$\mathbf{D}\big(\, S_1 \,,\, \bar{S}_1 \,\big) \triangleq \max\Big\{ \max_{(\boldsymbol{x},y)\in S_1} \overbrace{\min_{(\boldsymbol{x}',y')\in \bar{S}_1} \mathbf{d}\big((\ddot{\boldsymbol{x}},\ddot{y}),(\ddot{\boldsymbol{x}}',\ddot{y}')\big)}^{\text{to find the nearest data point in } \bar{S}_1}, \quad \max_{(\boldsymbol{x}',y')\in \bar{S}_1} \min_{(\boldsymbol{x},y)\in S_1} \mathbf{d}\big((\ddot{\boldsymbol{x}},\ddot{y}),(\ddot{\boldsymbol{x}}',\ddot{y}')\big)\Big\}, \tag{6}$$

S_1: the privileged group

$\bar{S}_1$: the marginalised/unprivileged group(s)

$\ddot{y}$: works for both the true label y and the prediction $\hat{y}$ of a trained classifier $f(\cdot)$

and is viewed as an approximation of the distance between the manifold of unprivileged groups and that of the privileged group.



Basic properties satisfied:

1. For any two data sets $S_0, S_1 \in \mathcal{X}\times\mathcal{Y}$, $\mathbf{D}(S_0,S_1)=0$ if and only if S_0 equals S_1; and
2. For any sets S_0, S_1, and S_2, we have the triangle inequality

$$\mathbf{D}(S_0,S_2) \leqslant \mathbf{D}(S_0,S_1) + \mathbf{D}(S_1,S_2)\,.$$

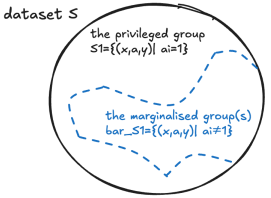
[11] Here we use the standard Euclidean metric. In fact, any two metrics $\mathbf{d}_1, \mathbf{d}_2$ derived from norms on the Euclidean space $\mathbb{R}^d$ are equivalent in the sense that there are positive constants c_1, c_2 such that $c_1\mathbf{d}_1(x,y) \leqslant \mathbf{d}_2(x,y) \leqslant c_2\mathbf{d}_1(x,y)$ for all $x, y \in \mathbb{R}^d$.

HFM[12] [Work 1 in this series]

For the distance between two subsets, and that of a trained classifier $f(\cdot)$, we have

$$\mathbf{D}(S_1, \bar{S}_1) \triangleq \max\Big\{ \max_{(x,y)\in S_1} \min_{(x',y')\in \bar{S}_1} \mathbf{d}\big((\tilde{x},y),(\tilde{x}',y')\big), \max_{(x',y')\in \bar{S}_1} \min_{(x,y)\in S_1} \mathbf{d}\big((\tilde{x},y),(\tilde{x}',y')\big)\Big\}, \tag{7a}$$

$$\mathbf{D}_f(S_1, \bar{S}_1) = \max\Big\{ \max_{(x,y)\in S_1} \min_{(x',y')\in \bar{S}_1} \mathbf{d}\big((\tilde{x},\hat{y}),(\tilde{x}',\hat{y}')\big), \max_{(x',y')\in \bar{S}_1} \min_{(x,y)\in S_1} \mathbf{d}\big((\tilde{x},\hat{y}),(\tilde{x}',\hat{y}')\big)\Big\}. \tag{7b}$$



We remark that $\mathbf{D}(S_1, \bar{S}_1)$ suggests the biases from the data and $\mathbf{D}_f(S_1, \bar{S}_1)$ suggests the biases from the algorithm. Then the following value could be used to indicate the fairness degree of this classifier, that is,

$$\mathbf{df}_{\text{prev}}(f) = \frac{\mathbf{D}_f(S_1, \bar{S}_1)}{\mathbf{D}(S_1, \bar{S}_1)} - 1\,. \tag{8}$$

[12] Bian and Luo, see n. 9.

Distance between sets —for one *multi-valued* SA

By extending Eq. (6), we introduce the following distance measures: (i) *maximal distance measure for one sensitive attribute*

$$\mathbf{D}_{\cdot,\boldsymbol{a}}(S,a_i) \triangleq \max_{1\leqslant j\leqslant n_{a_i}} \Big\{ \max_{(\boldsymbol{x},y)\in S_j} \overbrace{\min_{(\boldsymbol{x}',y')\in\bar{S}_j} \mathbf{d}\big((\ddot{\boldsymbol{x}},\ddot{y}),(\ddot{\boldsymbol{x}}',\ddot{y}')\big)}^{\text{to fine the nearest data point in } \bar{S}_j} \Big\}, \tag{9}$$

and (ii) *average distance measure for one sensitive attribute*

$$\mathbf{D}^{\text{avg}}_{\cdot,\boldsymbol{a}}(S,a_i) \triangleq \frac{1}{n}\sum_{j=1}^{n_{a_i}} \sum_{(\boldsymbol{x},y)\in S_j} \min_{(\boldsymbol{x}',y')\in\bar{S}_j} \mathbf{d}\big((\ddot{\boldsymbol{x}},\ddot{y}),(\ddot{\boldsymbol{x}}',\ddot{y}')\big), \tag{10}$$

where $\bar{S}_j = S\backslash S_j$, and $n_{a_i} = |\mathcal{A}_i| \geqslant 2$ is the number of optional values for the sensitive attribute $a_i \in \mathcal{A}_i$. Notice that $\mathbf{D}_{\cdot,\boldsymbol{a}}(S,a_i) = \mathbf{D}_{\cdot}(S_1,\bar{S}_1)$ when $\mathcal{A}_i = \{0,1\}$.

Distance between sets —for *several multi*-valued SAs

For the general case, we introduce the generalised distance measures: (i) *maximal distance measure for sensitive attributes*

$$\mathbf{D}_{\cdot,\boldsymbol{a}}(S) \triangleq \max_{1 \leqslant i \leqslant n_a} \mathbf{D}_{\cdot,\boldsymbol{a}}(S, a_i) \,, \tag{11}$$

and (ii) *average distance measure for sensitive attributes*

$$\mathbf{D}^{\text{avg}}_{\cdot,\boldsymbol{a}}(S) \triangleq \frac{1}{n_a} \sum_{i=1}^{n_a} \mathbf{D}^{\text{avg}}_{\cdot,\boldsymbol{a}}(S, a_i) \,. \tag{12}$$

HFM[13] [Work 2 in this series]

We remark that $\mathbf{D}_{\boldsymbol{a}}(S)$, $\mathbf{D}_{\boldsymbol{a}}^{\text{avg}}(S)$ reflect the biases from the data and that $\mathbf{D}_{f,\boldsymbol{a}}(S)$, $\mathbf{D}_{f,\boldsymbol{a}}^{\text{avg}}(S)$ reflect the extra biases from the learning algorithm. Then the following values could be used to reflect the fairness degree of this classifier, that is,

$$\mathbf{df}(f) = \log\left(\frac{\mathbf{D}_{f,\boldsymbol{a}}(S)}{\mathbf{D}_{\boldsymbol{a}}(S)}\right), \tag{13a}$$

$$\mathbf{df}^{\text{avg}}(f) = \log\left(\frac{\mathbf{D}_{f,\boldsymbol{a}}^{\text{avg}}(S)}{\mathbf{D}_{\boldsymbol{a}}^{\text{avg}}(S)}\right). \tag{13b}$$

We name the fairness degrees defined as above of one classifier by Eq. (13) as '*maximum harmonic fairness measure via manifolds (HFM)*' and '*average HFM*', respectively.

[13] Bian, Luo, and Xu, see n. 9.

Comparison between *HFM*[14] and baseline fairness measures

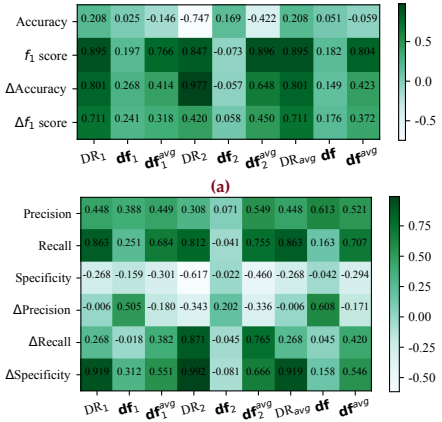


Figure 3: Correlation heatmap between normal evaluation metric and fairness measure, for all sensitive attributes within the dataset. Here we use $\mathrm{DR}_{avg} = \frac{1}{n_a}\sum_{i=1}^{n_a}\mathrm{DR}_i$ to reflect the bias level on the whole dataset.

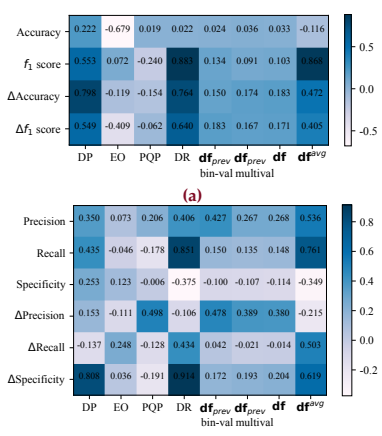


Figure 4: Correlation heatmap between normal evaluation metric and fairness, for one single sensitive attribute. Note that $\mathbf{df}_{\text{prev}} = \mathbf{D}_f(S_1, \tilde{S}_1)/\mathbf{D}(S_1, \tilde{S}_1) - 1$ represents our previous work, and ${}_{\mathbf{df}} = \log(\mathbf{D}_{f,a}(S, a_i)/\mathbf{D}_a(S, a_i))$ and ${}_{\mathbf{df}}{}^{\text{avg}} = \log(\mathbf{D}_{f,a}(S, a_i)/\mathbf{D}_a^{\text{avg}}(S, a_i))$ here represent *HFM* in this paper for each sensitive attribute.

[14] Bian, Luo, and Xu, see n. 9.

ExtendDist[15] [Work 2 in this series]

Algorithm 1 Approximation of extended distance between sets for several sensitive attributes with multiple values, aka. $\texttt{ExtendDist}\left(\{(\check{\boldsymbol{x}}_i, \boldsymbol{a}_i)\}_{i=1}^n, \{\ddot{y}_i\}_{i=1}^n; m_1, m_2\right)$,

Input: Dataset $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n = \{(\check{\boldsymbol{x}}_i, \boldsymbol{a}_i, y_i)\}_{i=1}^n$ where $\boldsymbol{a}_i = [a_{i,1}, a_{i,2}, ..., a_{i,n_a}]^\mathsf{T}$, prediction of S by the classifier $f(\cdot)$ that has been trained, that is, $\{\hat{y}_i\}_{i=1}^n$, and two hyper-parameters m_1 and m_2 as the designated numbers for repetition and comparison respectively

Output: Approximation of $\mathbf{D}_{\cdot,\boldsymbol{a}}(S)$ and $\mathbf{D}_{\cdot,\boldsymbol{a}}^{\text{avg}}(S)$

1: **for** j from 1 to n_a **do**

2: $d_{\max}^{(j)}, d_{\text{avg}}^{(j)} = \texttt{ApproxDist}\left(\{(\check{\boldsymbol{x}}_i, a_{i,j})\}_{i=1}^n, \{\ddot{y}_i\}_{i=1}^n; m_1, m_2\right)$

3: **return** $\max_{1 \leqslant j \leqslant n_a}\{d_{\max}^{(j)} \mid j \in [n_a]\}$ and $\frac{1}{n_a}\sum_{j=1}^{n_a} d_{\text{avg}}^{(j)}$

[15] Bian, Luo, and Xu, see n. 9.

ApproxDist[16] [Work 1&2 in this series]

Algorithm 2 (Simplified) Approximation of distance between sets, aka. $\texttt{ApproxDist}\left(\{(\check{\boldsymbol{x}}_i, a_i)\}_{i=1}^n, \{\ddot{y}_i\}_{i=1}^n; m_1, m_2\right)$

Input: Dataset $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n = \{(\check{\boldsymbol{x}}_i, \boldsymbol{a}_i, y_i)\}_{i=1}^n$, prediction of S by the classifier $f(\cdot)$ that has been trained, that is, $\{\hat{y}_i\}_{i=1}^n$, and two hyper-parameters m_1 and m_2 as the designated numbers for repetition and comparison respectively

Output: Approximation of distance $\mathbf{D}_{\cdot}(S_1, \tilde{S}_1)$ in Eq. (6)

1: **for** j from 1 to m_1 **do**
2: Take a random vector $\boldsymbol{w}$ from the space $\mathcal{W} = \{\boldsymbol{w} = [w_0, w_1, ..., w_{n_x}]^\mathsf{T} \mid \sum_{i=0}^{n_x} |w_i| = 1\} \subseteq [-1, 1]^{1+n_x}$
3: $d_{\max}^j = \texttt{AcceleDist}\left(\{(\check{\boldsymbol{x}}_i, \boldsymbol{a}_i)\}_{i=1}^n, \{\ddot{y}_i\}_{i=1}^n, \boldsymbol{w}; m_2\right)$
4: **return** $\min\{d_{\max}^j \mid j \in [m_1]\}$

Algorithm 2 Approximation of distance between sets (for one sensitive attribute with multiple values), aka. $\texttt{ApproxDist}\left(\{(\check{\boldsymbol{x}}_i, a_i)\}_{i=1}^n, \{\ddot{y}_i\}_{i=1}^n; m_1, m_2\right)$

Input: Dataset $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n = \{(\check{\boldsymbol{x}}_i, \boldsymbol{a}_i, y_i)\}_{i=1}^n$, prediction of S by the classifier $f(\cdot)$ that has been trained, that is, $\{\hat{y}_i\}_{i=1}^n$, and two hyper-parameters m_1 and m_2 as the designated numbers for repetition and comparison respectively

Output: Approximation of $\mathbf{D}_{\cdot,a}(S, a_i)$ and $\mathbf{D}_{\cdot,a}^{\text{avg}}(S, a_i)$

1: **for** j from 1 to m_1 **do**
2: Take two orthogonal vectors $\boldsymbol{w}_0$ and $\boldsymbol{w}_1$ where each $\boldsymbol{w}_k \in [-1, +1]^{1+n_x}$ $(k = \{0, 1\})$
3: **for** k from 0 to 1 **do**
4: $t_{\max}^k, t_{\text{avg}}^k = \texttt{AcceleDist}\left(\{(\check{\boldsymbol{x}}_i, a_i)\}_{i=1}^n, \{\ddot{y}_i\}_{i=1}^n, \boldsymbol{w}_k; m_2\right)$
5: $d_{\max}^j = \min\{t_{\max}^k \mid k \in \{0, 1\}\} = \min\{t_{\max}^0, t_{\max}^1\}$
6: $d_{\text{avg}}^j = \min\{t_{\text{avg}}^k \mid k \in \{0, 1\}\} = \min\{t_{\text{avg}}^0, t_{\text{avg}}^1\}$
7: **return** $\min\{d_{\max}^j \mid j \in [m_1]\}$ and $\frac{1}{n}\min\{d_{\text{avg}}^j \mid j \in [m_1]\}$

[16] Bian and Luo, see n. 9; Bian, Luo, and Xu, see n. 9.

Distance *approximation* for Euclidean spaces

We observe that *the distance between similar data points tends to be closer than others after projecting them onto a general one-dimensional linear subspace* (refer to[17]).

To estimate the distance between data points inside $\mathcal{X} \times \mathcal{Y}$,

$$g(\boldsymbol{x}, \ddot{y}; \boldsymbol{w}) = g(\check{\boldsymbol{x}}, \boldsymbol{a}, \ddot{y}; \boldsymbol{w}) = [\ddot{y}, x_1, ..., x_{n_x}]^\mathsf{T} \boldsymbol{w}\,, \tag{14}$$

where

- a random projection $g : \mathcal{X} \times \mathcal{Y} \mapsto \mathbb{R}$
- a non-zero random vector $\boldsymbol{w} = [w_0, w_1, ..., w_{n_x}]^\mathsf{T}$

That is to say, after sorting all the projected data points on $\mathbb{R}$, it is likely that *for one instance* $(\boldsymbol{x}, y)$ *in* S_j*, the desired instance* $\operatorname{argmin}_{(\boldsymbol{x}', y') \in \bar{S}_j} \mathbf{d}\big((\check{\boldsymbol{x}}, y), (\check{\boldsymbol{x}}', y')\big)$ *would be somewhere near it after the projection, and vice versa*. Thus, searching for it could be accelerated by checking several adjacent instances rather than traversing the whole dataset.

[17] Bian and Luo, see n. 9, Lemma 1.

AcceleDist[18] [Work 1&2 in this series]

Algorithm 3 Acceleration sub-procedure in approximation, aka. `AcceleDist` $(\{(\check{\boldsymbol{x}}_i, a_i)\}_{i=1}^n, \{\ddot{y}_i\}_{i=1}^n, \boldsymbol{w}; m_2)$

Input: Data points $\{(\check{\boldsymbol{x}}_i, a_i)\}_{i=1}^n$, its corresponding value $\{\ddot{y}_i\}_{i=1}^n$, where $\ddot{y}_i$ could be its true label y_i or prediction $\hat{y}_i$ by the classifier $f(\cdot)$, a random vector $\boldsymbol{w}$ for projection, and a hyper-parameter m_2 as the designated number for comparison

Output: Approximation of distance $\mathbf{D}_{\cdot,\cdot}(S_0, S_1)$ in Eq. (6)

Output: Approximation of $\mathbf{D}_{\cdot,a}(S, a_i)$ and $n\mathbf{D}_{\cdot,a}^{\text{avg}}(S, a_i)$

Project data points onto a one-dimensional space based on Eq. (14), in order to obtain $\{g(\boldsymbol{x}_i, \ddot{y}_i; \boldsymbol{w})\}_{i=1}^n$
Sort original data points based on $\{g(\boldsymbol{x}_i, \ddot{y}_i; \boldsymbol{w})\}_{i=1}^n$ as their corresponding values, in ascending order
for i from 1 to n **do**
    Set the anchor data point $(\boldsymbol{x}_i, \ddot{y}_i)$ in this round
    `// If` $a_i = j$ `(marked for clarity), in order to approximate` $\min_{(\boldsymbol{x}', \ddot{y}') \in S_j} \mathbf{d}((\check{\boldsymbol{x}}_i, \ddot{y}_i), (\check{\boldsymbol{x}}', \ddot{y}'))$
    Compute the distances $\mathbf{d}((\check{\boldsymbol{x}}_i, \ddot{y}_i), \cdot)$ for at most m_2 nearby data points that meets $a \neq a_i$ and $g(\check{\boldsymbol{x}}, \ddot{y}; \boldsymbol{w}) \leqslant g(\check{\boldsymbol{x}}_i, \ddot{y}_i; \boldsymbol{w})$
    Find the minimum among them, recorded as $d_{\min}^s$
    Compute the distances $\mathbf{d}((\check{\boldsymbol{x}}_i, \ddot{y}_i), \cdot)$ for at most m_2 nearby data points that meets $a \neq a_i$ and $g(\boldsymbol{x}, \ddot{y}; \boldsymbol{w}) \geqslant g(\boldsymbol{x}_i, \ddot{y}_i; \boldsymbol{w})$
    Find the minimum among them, recorded as $d_{\min}^r$
    $d_{\min}^{(i)} = \min\{d_{\min}^s, d_{\min}^r\}$
return $\max\{d_{\min}^{(i)} \mid i \in [n]\}$
return $\max\{d_{\min}^{(i)} \mid i \in [n]\}$ and $\sum_{i=1}^n d_{\min}^{(i)}$

[18] Bian and Luo, see n. 9; Bian, Luo, and Xu, see n. 9.

Validity of approximation[19] for distances in Euclidean spaces

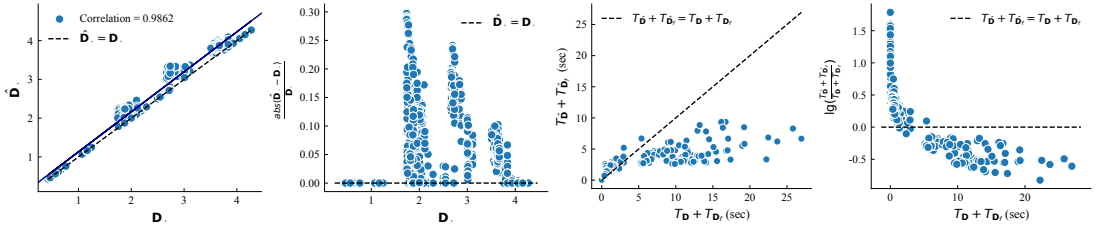


Figure 5: Comparison of approximation distances between sets with precise distances that are calculated directly by definition, evaluated on test data. (a) Scatter plot showing approximated values and precise values of distances between sets; (b) Relative difference comparison of *ApproxDist* with direct computation concerning distance values. (c–d) Comparison of time cost (second) between *ApproxDist* and direct computation based on Eq. (6).

[19] Bian and Luo, see n. 9.

Validity of *distance* approximation[20] in Euclidean spaces

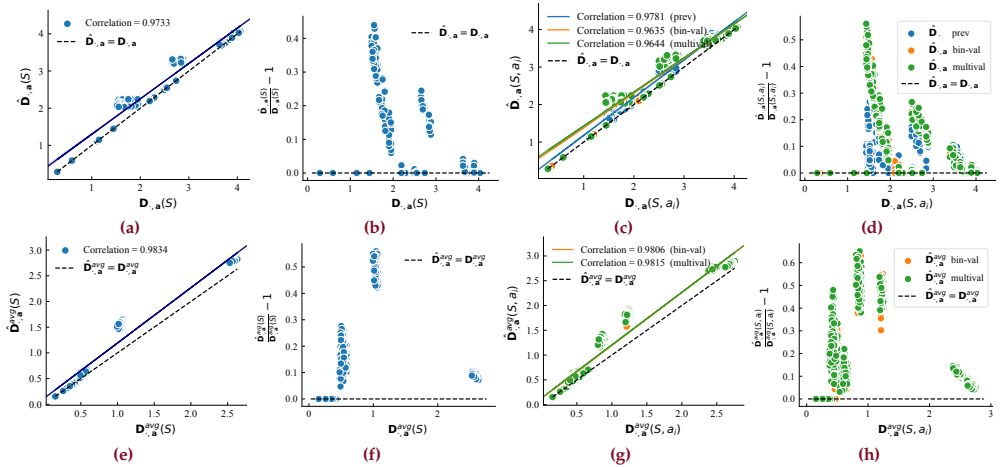


Figure 6: Comparison of approximation distances with precise distances that are calculated directly by definition, evaluated on test data. (a–b), (c–d), (e–f), and (g–h) Scatter plots for comparison between approximated and precise values of $\mathbf{D}_{\cdot,\boldsymbol{a}}(S)$, $\mathbf{D}_{\cdot,\boldsymbol{a}}(S, a_i)$, $\mathbf{D}^{\text{avg}}_{\cdot,\boldsymbol{a}}(S)$, and $\mathbf{D}^{\text{avg}}_{\cdot,\boldsymbol{a}}(S, a_i)$, respectively.

[20] Bian, Luo, and Xu, see n. 9.

Validity of *distance* approximation[20] in Euclidean spaces

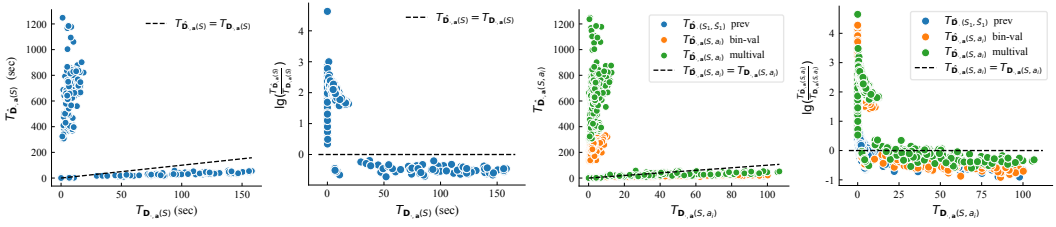


Figure 6: Comparison of approximation distances with precise distances that are calculated directly by definition, evaluated on test data. (i–j) Time cost comparison between *ExtendDist* and direct computation; (k–l) Time cost comparison between *ApproxDist* and direct computation. Note that 'prev' denotes approximation results obtained by the simplified Algorithm 2.

[20] Bian, Luo, and Xu, see n. 9.

Interim summary[21]

RQ 2. *How to efficiently measure the added discrimination introduced in learning by a classifier?*

	Work 1	**Work 2**	
Distance between sets	$\mathbf{D}(S_1, \tilde{S}_1)$	$\mathbf{D}_{\cdot,\boldsymbol{a}}(S, a_i)(S, a_i)$	$\mathbf{D}^{\text{avg}}_{\cdot,\boldsymbol{a}}(S, a_i)(S, a_i)$
		$\mathbf{D}_{\cdot,\boldsymbol{a}}(S)(S)$	$\mathbf{D}^{\text{avg}}_{\cdot,\boldsymbol{a}}(S)(S)$
HFM	$\mathbf{df}_{\text{prev}}(f)$	$\mathbf{df}(f)$	$\mathbf{df}^{\text{avg}}(f)$
Approximation for one SA	*AcceleDist*		*AcceleDist*
	ApproxDist		*ApproxDist*
Approximation for several SA			*ExtendDist*

[21] Bian and Luo, see n. 9; Bian, Luo, and Xu, see n. 9.

Why-and-how-to-use recap[22]

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

3.′ *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?*

[22] Bian and Zhang, see n. 5.

Cancellation-of-bias effect[23] 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

[23] Bian and Zhang, see n. 5.

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

weight corresponding to $f_j(\cdot)$ — individual classifier

$$\mathbf{wv}_\rho(\boldsymbol{x}) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \sum_{j=1}^{m} w_j \, \mathbb{I}(f_j(\boldsymbol{x}) = y) \, . \tag{15}$$

ensemble combination via weighted vote — optional choice of labels

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

$$\mathbf{wv}_\rho(\boldsymbol{x}) = \underset{y \in \mathcal{Y}}{\operatorname{argmax}} \sum_{j=1}^{m} w_j \mathbb{I}(f_j(\boldsymbol{x}) = y)\,. \tag{15}$$

where a function $f \in \mathcal{F}\colon \mathcal{X} \mapsto \mathcal{F}$ 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

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

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,

$$\ell_{\text{bias}}(\mathbf{wv}_\rho, \boldsymbol{x}) \leqslant \mathbb{I}\Big(\mathbb{E}_\rho\big[\ \mathbb{I}(f(\check{\boldsymbol{x}}, \boldsymbol{a}) \neq f(\check{\boldsymbol{x}}, \tilde{\boldsymbol{a}}))\ \big] \geqslant 0.5 \Big)\,. \tag{16}$$

$\ell_{\text{bias}}(\mathbf{wv}_\rho, \boldsymbol{x})$: discriminative risk of an ensemble $\mathbf{wv}_\rho(\cdot)$

$\mathbb{E}_\rho[\ \mathbb{I}(f(\check{\boldsymbol{x}}, \boldsymbol{a}) \neq f(\check{\boldsymbol{x}}, \tilde{\boldsymbol{a}}))\ ]$: that is, $\ell_{\text{bias}}(f, \boldsymbol{x})$, 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

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

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,

$$\ell_{\text{bias}}(\mathbf{wv}_\rho, \boldsymbol{x}) \leqslant \mathbb{I}\big(\, \mathbb{E}_\rho\big[\, \mathbb{I}(f(\check{\boldsymbol{x}}, \boldsymbol{a}) \neq f(\check{\boldsymbol{x}}, \tilde{\boldsymbol{a}}))\, \big] \geqslant 0.5\, \big)\,. \tag{16}$$

$\ell_{\text{bias}}(\mathbf{wv}_\rho, \boldsymbol{x})$: discriminative risk of an ensemble $\mathbf{wv}_\rho(\cdot)$

$\mathbb{E}_\rho[\, \mathbb{I}(f(\check{\boldsymbol{x}}, \boldsymbol{a}) \neq f(\check{\boldsymbol{x}}, \tilde{\boldsymbol{a}}))\,]$: 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)

$$\mathcal{L}_{bias}(\mathbf{wv}_\rho) \leqslant 2\, \mathbb{E}_\rho\big[\, \mathcal{L}_{bias}(f)\, \big]\,. \tag{17}$$

$\mathcal{L}_{bias}(\mathbf{wv}_\rho)$: *discriminative risk of an ensemble* $\mathbf{wv}_\rho$

$\mathcal{L}_{bias}(f)$: *discriminative risk of an individual classifier* f

$2\, \mathbb{E}_\rho[\, \mathcal{L}_{bias}(f)\,]$: *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,[24] by

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

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

$$\ell_{\text{bias}}(f, f', \boldsymbol{x}) = \mathbb{I}\big(f(\check{\boldsymbol{x}}, \boldsymbol{a}) \neq f(\check{\boldsymbol{x}}, \tilde{\boldsymbol{a}}) \wedge f'(\check{\boldsymbol{x}}, \boldsymbol{a}) \neq f'(\check{\boldsymbol{x}}, \tilde{\boldsymbol{a}}) \big) . \tag{18}$$

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

$$\ell_{\text{bias}}(f, f, \boldsymbol{x}) = \ell_{\text{bias}}(f, \boldsymbol{x}) . \tag{19}$$

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

discriminative risk of $f(\cdot)$

[24] 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'

$$\mathcal{L}_{bias}(\mathbf{wv}_\rho) \leqslant 4\,\mathbb{E}_{\rho^2}[\,\mathcal{L}_{bias}(f,f')] \,. \tag{20}$$

the worst case is controlled to a constant multiple

Lemma 2

In multi-class classification,

discriminative risk of $f(\cdot)$

$$\mathbb{E}_{\rho^2}[\,\mathcal{L}_{\text{bias}}(f,f')\,] = \mathbb{E}_{\mathcal{D}}[\mathbb{E}_{\rho}[\,\ell_{\text{bias}}(f,x)\,]^2] \,. \tag{21}$$

the expected tandem discriminative risk

Oracle bounds of fairness (cont.)

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

$$\mathcal{L}_{bias}(\mathbf{wv}_\rho) \leqslant \frac{\mathbb{E}_{\rho^2}[\,\mathcal{L}_{bias}(f,f')\,] - \mathbb{E}_\rho[\,\mathcal{L}_{bias}(f)\,]^2}{\mathbb{E}_{\rho^2}[\,\mathcal{L}_{bias}(f,f')\,] - \mathbb{E}_\rho[\,\mathcal{L}_{bias}(f)\,] + \frac{1}{4}}. \tag{22}$$

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 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 the *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 $f(\cdot)$,

$$\mathcal{L}_{bias}(f) \leqslant \hat{\mathcal{L}}_{bias}(f,S) + \sqrt{\frac{1}{2n}\ln\frac{1}{\delta}}\,. \tag{23}$$

the worst case is controlled with a specific bound

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 $\mathcal{F}$,

$$\mathcal{L}_{bias}(\mathbf{wv}_\rho) \leqslant \hat{\mathcal{L}}_{bias}(\mathbf{wv}_\rho,S) + \sqrt{\frac{1}{2n}\log\frac{|\mathcal{F}|}{\delta}}\,. \tag{24}$$

the worst case is controlled with a specific bound

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,[25] 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

$$\ell_{\text{err}}(f, \boldsymbol{x}) = \mathbb{I}(f(\boldsymbol{x}) \neq y), \tag{25}$$

model prediction on the raw data

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

model prediction on the raw data

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

the discriminative risk of $f(\cdot)$ — model prediction when only sensitive attribute(s) are changed

Besides, we have two more (PAC) generalisation bounds that they don't.

[25] Masegosa et al., see n. **??**.

Validating the oracle&PAC bounds

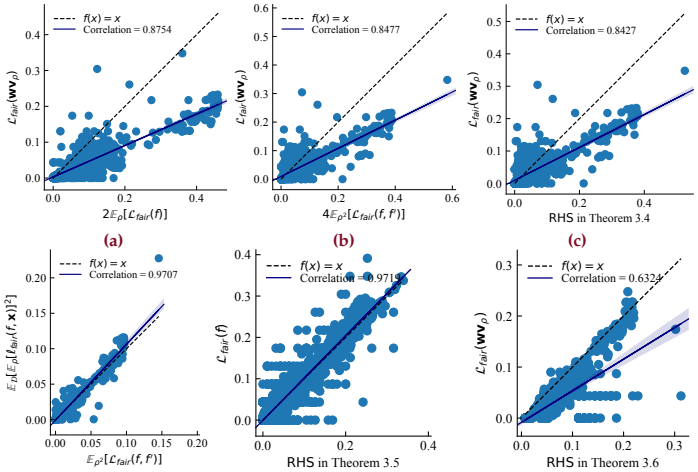


Figure 7: Correlation for oracle bounds and generalisation bounds. (a–c) Correlation between $\mathcal{L}_{\text{bias}}(\mathbf{wv}_\rho)$ and oracle bounds, where $\mathcal{L}_{\text{bias}}(\mathbf{wv}_\rho)$ is indicated on the vertical axis and the horizontal axes represent the right-hand sides of inequalities (17), (20), and (22), respectively. (d) The horizontal and vertical axes in (d) denote the right- and left-hand sides in (21), 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 (23) and (24) are indicated on the horizontal axes, respectively. Note that correlation here is calculated based on the results from all datasets.

Interim summary[26]

RQ 3. *Why and how can we make use of these proposed fairness measure/metric(s)?*

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

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

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

[26] Bian and Zhang, see n. 5.

Why-and-how-to-use recap[27]

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

3.′ *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?*

[27] Bian and Zhang, see n. 5.

Concept of *domination*[28]

We evaluate the accuracy quality of a hypothesis $f(\cdot)$ by the 0/1 loss $\ell_{\text{err}}(f, \boldsymbol{x})$, the empirical loss by $\hat{\mathcal{L}}_{\text{err}}(f, S) = \frac{1}{n}\sum_{i=1}^{n} \ell_{\text{err}}(f, \boldsymbol{x})$, and the expected loss by $\mathcal{L}_{\text{err}}(f) = \mathbb{E}_{\mathcal{D}}[\ell_{\text{err}}(f, \boldsymbol{x})]$, respectively.

objective

$G = (L_{err}, L_{bias})$

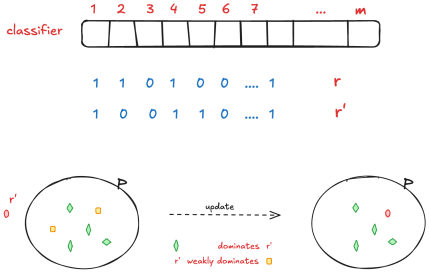
solutions $\rho \quad \leq \quad \leq \quad \pi$

Definition 7 (Domination)

Let $\mathcal{L}_{\text{err}}(\cdot)$ and $\mathcal{L}_{\text{bias}}(\cdot)$ be two sub-objectives to be minimised, and let $\mathcal{G} = (\mathcal{L}_{\text{err}}, \mathcal{L}_{\text{bias}})$ be the objective for a Pareto optimal solution. For two probability distributions ρ and π on $\mathcal{F}$ that are independent of S: 1) ρ weakly dominates π if $\mathcal{L}_{\text{err}}(\mathbf{wv}_\rho) \leqslant \mathcal{L}_{\text{err}}(\mathbf{wv}_\pi)$ and $\mathcal{L}_{\text{bias}}(\mathbf{wv}_\rho) \leqslant \mathcal{L}_{\text{bias}}(\mathbf{wv}_\pi)$, denoted as $\succeq_{\mathcal{G}}$; 2) ρ dominates π if $\rho \succeq_{\mathcal{G}} \pi$ and either $\mathcal{L}_{\text{err}}(\mathbf{wv}_\rho) < \mathcal{L}_{\text{err}}(\mathbf{wv}_\pi)$ or $\mathcal{L}_{\text{bias}}(\mathbf{wv}_\rho) < \mathcal{L}_{\text{bias}}(\mathbf{wv}_\pi)$, denoted as $\succ_{\mathcal{G}}$.

[28] A concept of domination (Chao Qian, Yang Yu, and Zhi-Hua Zhou. "Pareto ensemble pruning". In: *AAAI*. vol. 29. 1. 2015, pp. 2935–2941) is introduced to take fairness and accuracy into account simultaneously during pruning. Note that the domination relationship is used to achieve a Pareto optimal solution for this bi-objective minimisation case, where two sub-objectives are viewed as improving fairness and accuracy, respectively.

To construct fairer ensembles



A solution $\mathbf{wv}_\rho(\cdot)$ is *Pareto optimal* if there is no other solution in $\mathcal{F}$ that dominates $\mathbf{wv}_\rho(\cdot)$.

Algorithm 4 Pareto Optimal Ensemble Pruning via Improving Accuracy and Fairness Concurrently (*POAF*)

Input: training set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$, original ensemble $F = \{f_j(\cdot)\}_{j=1}^{m}$ via weighted vote, and threshold k as maximum size of the sub-ensemble after pruning
Output: Pruned sub-ensemble H ($H \subset F$ and $|H| \leqslant k$)

Randomly pick k individual members from F, indicated by $\mathbf{r}$
Initialise a candidate set for pruned sub-ensembles $\mathcal{P} = \{\mathbf{r}\}$
for $i = 1$ to k **do**
  Randomly choose $\mathbf{r}$ from $\mathcal{P}$ with equal probability
  Generate $\mathbf{r}'$ by flipping each bit of $\mathbf{r}$ with probability $1/m$
  if $\exists\, \mathbf{z} \in \mathcal{P}$ such that $\mathbf{z} \succ_{\mathcal{G}} \mathbf{r}'$ **then**
    continue
  $\mathcal{P} = (\mathcal{P} \setminus \{\mathbf{z} \in \mathcal{P} \mid \mathbf{r}' \succeq_{\mathcal{G}} \mathbf{z}\}) \bigcup \{\mathbf{r}'\}$
  Let $\mathcal{V} = \mathcal{N}_-(\mathbf{r}') \bigcup \mathcal{N}_+(\mathbf{r}')$
  Sort $\mathcal{V}$ by $\operatorname{argmin}_{\mathbf{v} \in \mathcal{V}} \hat{\mathcal{L}}(\mathbf{v}, S)$ in ascending order
  for $\mathbf{v} \in \mathcal{V}$ **do**
    if $\exists\, \mathbf{z} \in \mathcal{P}$ such that $\mathbf{z} \succ_{\mathcal{G}} \mathbf{v}$ **then**
      continue
    $\mathcal{P} = (\mathcal{P} \setminus \{\mathbf{z} \in \mathcal{P} \mid \mathbf{v} \succeq_{\mathcal{G}} \mathbf{z}\}) \bigcup \{\mathbf{v}\}$
$H = \operatorname{argmin}_{\mathbf{r} \in \mathcal{P}} \hat{\mathcal{L}}(\mathbf{r}, S)$

Comparison between *POAF* and fairness-aware ensemble-based methods[29]

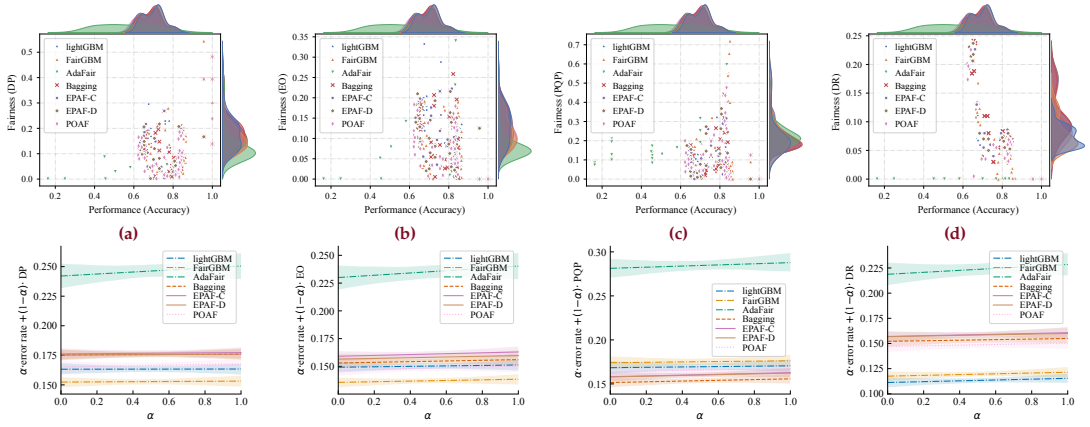


Figure 8: Comparison.. (a–d) Scatter plots showing fairness and accuracy of each algorithm, evaluated on the test data. (e–h) Plots of best test-set fairness-accuracy trade-offs per algorithm, where fairness is DP, EO, PQP, and *DR*, respectively. Lines show the mean value, and shades show 95% confidence intervals; The smaller the better.

[29] André F Cruz et al. "FairGBM: Gradient Boosting with Fairness Constraints". In: *ICLR*. 2023.

Comparison of *POAF* with *EPAF-C* and *EPAF-D* (& ensemble pruning methods)

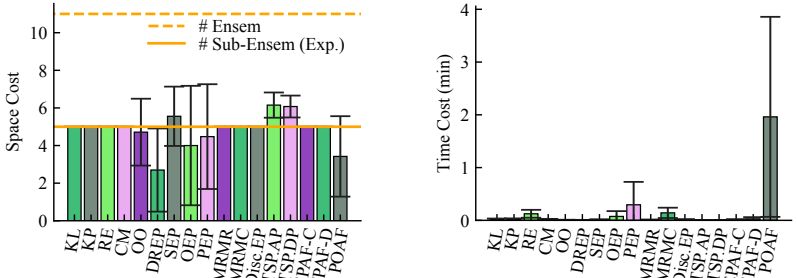


Figure 9: Comparison of the state-of-the-art pruning methods with *POAF*, where the dashed and solid lines indicate the size of the original ensemble and the expected size of the pruned sub-ensemble, respectively. (a–b) Comparison over the space and time cost, respectively.

Alternative: Bi-objective

An *alternative* way to compare two hypotheses by considering both fairness and accuracy is to define an objective function based on an *adaptive weighted sum* method, i.e.,

$$\mathcal{L}(f,f') = \lambda \frac{\mathcal{L}_{\text{err}}(f)+\mathcal{L}_{\text{err}}(f')}{2} + (1-\lambda)\mathcal{L}_{\text{bias}}(f,f')\,, \tag{26}$$

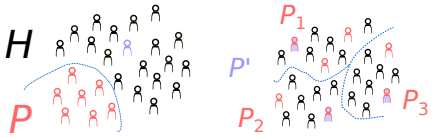
wherein $\lambda \in (0,1)$ represents a regularisation factor introduced to balance fairness and accuracy and to indicate their relative importance as well. Note that this is a symmetrical function, that is, $\mathcal{L}(f,f') = \mathcal{L}(f',f)$, and that in the degeneration case $\mathcal{L}(f,f) = \lambda\mathcal{L}_{\text{err}}(f) + (1-\lambda)\mathcal{L}_{\text{bias}}(f)$ when two hypotheses $f(\cdot)$ and $f'(\cdot)$ are identical. Then we can obtain the objective for the weighted vote as

$$\mathcal{L}(\mathbf{wv}_\rho) = \lambda\mathbb{E}_\rho[\mathcal{L}_{\text{err}}(f)] + (1-\lambda)\mathbb{E}_{\rho^2}[\mathcal{L}_{\text{bias}}(f,f')]\,, \tag{27}$$

which is expected to be minimised to improve fairness and accuracy concurrently.

Alternative: *EPAF*

Ensemble pruning via increasing accuracy and fairness concurrently (*EPAF*)



Algorithm 5 Centralised Version of Ensemble Pruning via Improving Accuracy and Fairness Concurrently (*EPAF-C*)

Input: training set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$, original ensemble $F = \{f_j(\cdot)\}_{j=1}^{m}$, and threshold k as maximum size after pruning

Output: Pruned sub-ensemble H ($H \subset F$ and $|H| \leqslant k$)

$H \leftarrow$ an arbitrary individual member $f_i \in F$
for $i = 2$ **to** k **do**
  $f^* \leftarrow \text{argmin}_{f_i \in F \setminus H} \sum_{f_j \in H} \hat{\mathcal{L}}(f_i, f_j, S)$
  Move f^* from F to H

Algorithm 6 Distributed Version of Ensemble Pruning via Improving Accuracy and Fairness Concurrently (*EPAF-D*)

Input: training set $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^{n}$, original ensemble $F = \{f_j(\cdot)\}_{j=1}^{m}$, threshold k as maximum size after pruning, and number of machines n_m

Output: Pruned sub-ensemble H ($H \subset F$ and $|H| \leqslant k$)

Partition F randomly into n_m groups as equally as possible, i.e., $F_1, ..., F_{n_m}$
for $i = 1$ **to** n_m **do**
  $H_i \leftarrow$ *EPAF-C*(F_i, k)
$H' \leftarrow$ *EPAF-C*($\bigcup_{i=1}^{n_m} H_i, k$)
$H \leftarrow \text{argmin}_{T \in \{H_1, ..., H_{n_m}, H'\}} \hat{\mathcal{L}}(T, S)$

Thanks! Questions?

Commonly used group fairness measures

There are three commonly-used group fairness measures, that is, ***demographic parity (DP),***[30] ***equality of opportunity (EO),***[31] ***and predictive quality parity (PQP)***[32].

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

$$\mathrm{DP}(f) = \left|\mathbb{P}_{\mathcal{D}}[f(\boldsymbol{x})=1 \,|\, \boldsymbol{a}=1] - \mathbb{P}_{\mathcal{D}}[f(\boldsymbol{x})=1 \,|\, \boldsymbol{a}=0]\right|, \tag{28a}$$

$$\mathrm{EO}(f) = \left|\mathbb{P}_{\mathcal{D}}[f(\boldsymbol{x})=1 \,|\, \boldsymbol{a}=1,\, y=1] - \mathbb{P}_{\mathcal{D}}[f(\boldsymbol{x})=1 \,|\, \boldsymbol{a}=0,\, y=1]\right|, \tag{28b}$$

$$\mathrm{PQP}(f) = \left|\mathbb{P}_{\mathcal{D}}[y=1 \,|\, \boldsymbol{a}=1, f(\boldsymbol{x})=1] - \mathbb{P}_{\mathcal{D}}[y=1 \,|\, \boldsymbol{a}=0, f(\boldsymbol{x})=1]\right|, \tag{28c}$$

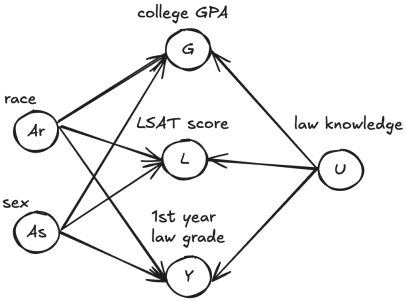
respectively, where $\boldsymbol{x} = (\check{\boldsymbol{x}}, \boldsymbol{a})$, y, and $f(\boldsymbol{x})$ are respectively features, the true label, and the prediction of this classifier for one instance. Note that $\boldsymbol{a} = 1$ and 0 respectively mean that the instance $\boldsymbol{x}$ belongs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

[30] 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.

[31] 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.

[32] 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



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[33] definition: A predictor $\hat{Y}$ is *counterfactually fair* if given observations $\mathcal{X}=\boldsymbol{x}$ and $A=a$ we have that, $\mathbb{P}(\hat{Y}_{A\leftarrow a}=y \mid \mathcal{X}=\boldsymbol{x}, A=a) = \mathbb{P}(\hat{Y}_{A\leftarrow a'}=y \mid \mathcal{X}=\boldsymbol{x}, A=a)$, for all y and $a' \neq a$.
Proxy discrimination[34] definition: A predictor $\hat{Y}$ exhibits no *individual proxy discrimination* if given observations $\mathcal{X}=\boldsymbol{x}$ and $A=a$ we have that, $\mathbb{P}(\hat{Y}=y \mid \mathrm{do}(A=a), \mathcal{X}=\boldsymbol{x}) = \mathbb{P}(\hat{Y}=y \mid \mathrm{do}(A=a'), \mathcal{X}=\boldsymbol{x})$, for all y and $a' \neq a$.

[33] Matt J Kusner et al. "Counterfactual fairness". In: *NIPS*. vol. 30. Curran Associates, Inc., 2017, pp. 4069–4079.
[34] Niki Kilbertus et al. "Avoiding discrimination through causal reasoning". In: *NIPS*. vol. 30. 2017.

Comparison between *HFM*[35] and baseline fairness measures

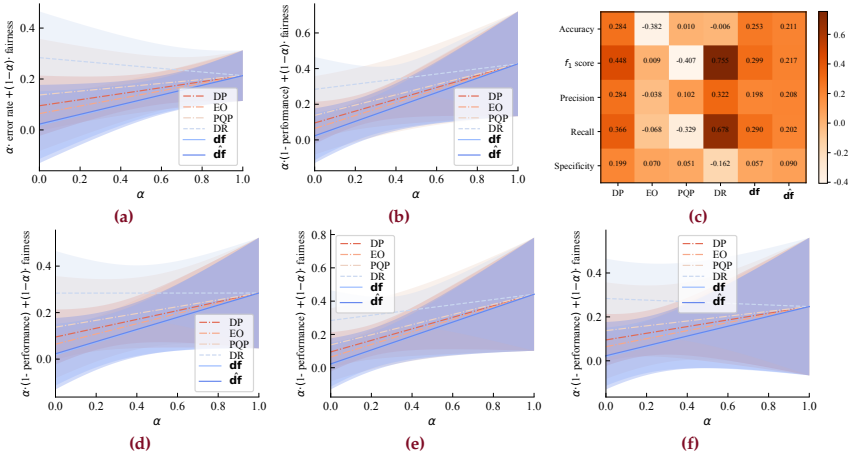


Figure 10: Comparison of baseline fairness measures and the proposed *HFM*, evaluated on test data. (a) Plot of trade-offs between fairness and error rates per fairness measure; (b) and (d–f) Plots of trade-offs between fairness and (1− performance) per fairness measures, where performance here are f_1 score, precision, recall/sensitivity, and specificity, respectively. (c) Correlation heatmap between evaluation metric and fairness.

[35] Bian and Luo, see n. 9; Cruz et al., see n. 28.

Comparison between *HFM*[36] and baseline fairness measures

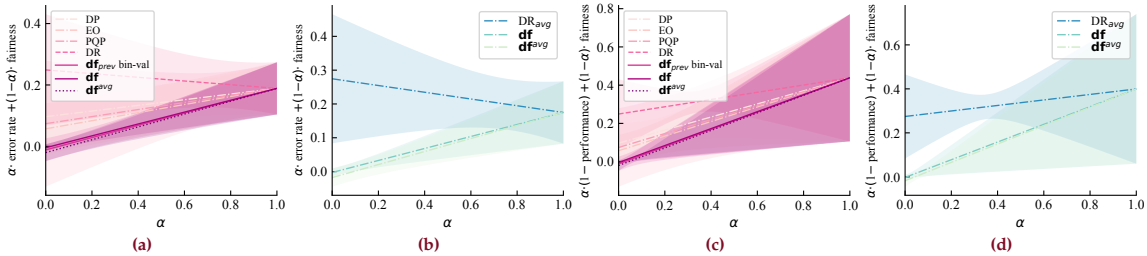


Figure 11: Plots of best test-set fairness-performance trade-offs per fairness metric (the smaller the better). (a) Plot of fairness-accuracy trade-off for one single sensitive attribute; (b) Plot of fairness-accuracy trade-off for all sensitive attributes; (c–d) Plots of fairness-f_1 score trade-off for one sensitive attribute and for all sensitive attributes, respectively. Note that the notations in (a) and (c) refer to those in Figure 4, and that in (b) and (d) refer to those in Figure 3.

[36] Bian, Luo, and Xu, see n. 9; Cruz et al., see n. 28.

Comparison of the state-of-the-art pruning method with *POAF*

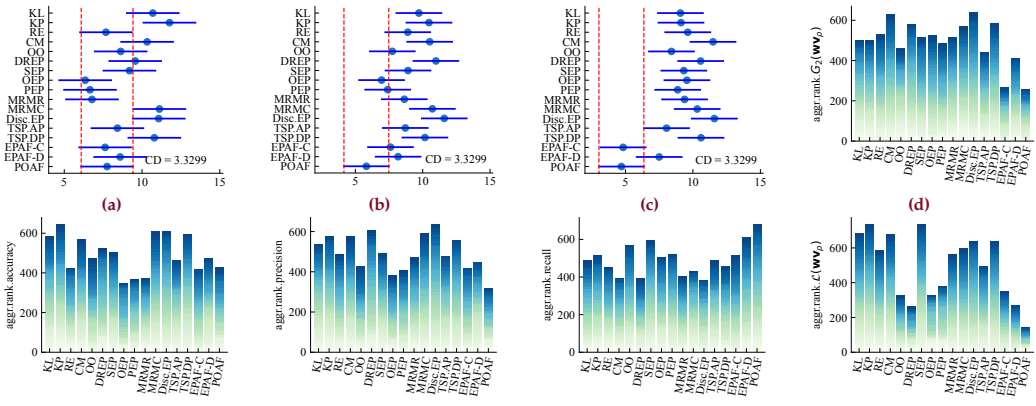


Figure 12: Comparison of the state-of-the-art pruning method with *POAF*, using bagging to conduct homogeneous ensembles. (a–c) Friedman test chart on the test accuracy, precision, and $\mathcal{L}_{\text{bias}}(\mathbf{wv}_\rho)$, respectively, of which each rejects the null hypothesis at the significance level of 5%; (d–h) The aggregated rank for each pruning method over the $\mathcal{L}_{\text{bias}}(\mathbf{wv}_\rho)$, test accuracy, precision, recall, and $\mathcal{L}(\mathbf{wv}_\rho)$, respectively.

Comparison between *EPAF-C* and *EPAF-D*

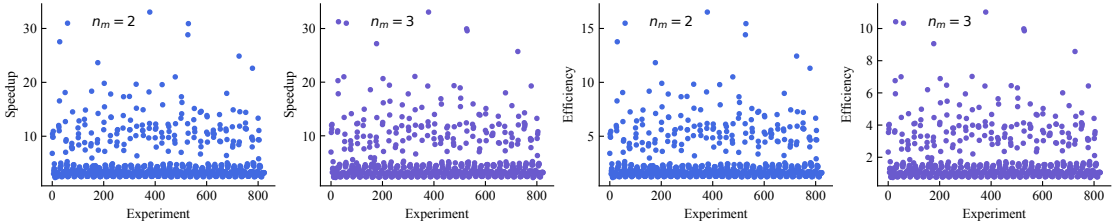


Figure 13: Comparison of speedup and efficiency between *EPAF-C* and *EPAF-D*. (a–b) Speedup with two and three machines, respectively; (c–d) Efficiency with two and three machines, respectively.