Improving Fairness with Ensemble Combination: Margin-Dependent Bounds [1]

Yijun Bian

Department of Computer Science
University of Copenhagen

26 June 2026

Funded by
the European Union

[1] ACM FAccT '26, Montreal, QC, Canada

Example [2]

Resume A

Emily Walsh
White-sounding name

Education: BA, 4 years experience
Skills: administrative, computer
Honours, volunteer work

$\leftrightarrow$

Resume B — identical except

Lakisha Washington
African-American-sounding name

Education: BA, 4 years experience
Skills: administrative, computer
Honours, volunteer work

Motivation

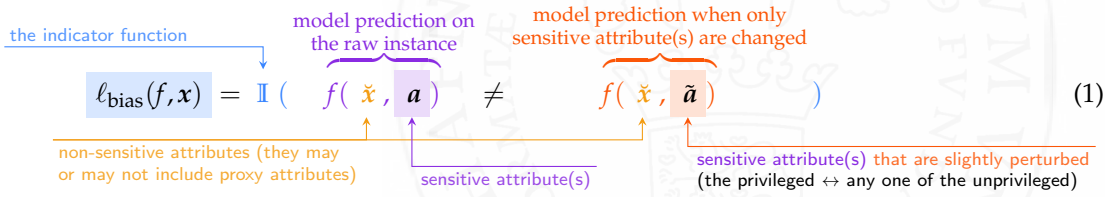
If perturbing a protected attribute changes the prediction
$\Rightarrow$ **discriminative risk exists**

[2] *White-associated names received 50% more callbacks for interviews* (Marianne Bertrand and Sendhil Mullainathan. "Are Emily and Greg more employable than Lakisha and Jamal? A field experiment on labor market discrimination". In: *Am Econ Rev* 94.4 [2004], pp. 991–1013)

Discriminative risk (DR)

—from an individual-level aspect

Following the principle of individual fairness *(the treatment/evaluation on one instance should not change solely due to minor changes in its protected/sensitive attributes)*, with an instance denoted by $\boldsymbol{x} = (\check{\boldsymbol{x}}, \boldsymbol{a})$, the fairness quality of one hypothesis[3] $f(\cdot)$ could be evaluated by



similarly to the 0/1 loss, where $\boldsymbol{a} = [a_1, ..., a_{n_a}]^\mathsf{T}$, $a_i \in \mathcal{A}_i$, $n_a \geqslant 1$, $|\mathcal{A}_i| \geqslant 2$, and $\tilde{\boldsymbol{a}}$ is a perturbed[4] $\boldsymbol{a}$. Note that Eq. (1) is evaluated on only one instance with sensitive attributes (SAs) $\boldsymbol{x}$.

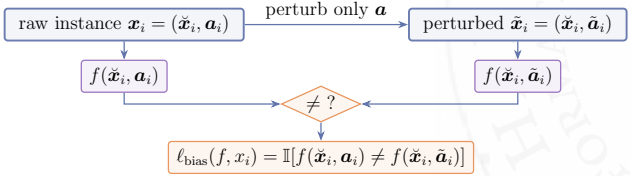
[3] The hypothesis used in this equation could indicate an individual classifier or an ensemble classifier.

[4] A member belonging to the marginalised group(s) will be perturbed into the privileged group, while a member of the privileged group is perturbed randomly into one of the marginalised groups.

Discriminative risk (DR)

—from an individual-level aspect
—from a group-level aspect

To describe this characteristic of the hypothesis on multiple instances, then the empirical *discriminative risk (DR)* on one dataset S is expressed as $\hat{\boldsymbol{L}}_{\text{bias}}(f,S) = \frac{1}{n}\sum_{i=1}^{n} \ell_{\text{bias}}(f,\boldsymbol{x}_i)$, and the true DR^3 of the hypothesis over a data distribution is $\boldsymbol{L}_{\text{bias}}(f) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\ \ell_{\text{bias}}(f,\boldsymbol{x})\ ]$, respectively.



$S = \{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}$

$\underbrace{0\ 1\ 0\ 0\ 1\ 0\ 1}$

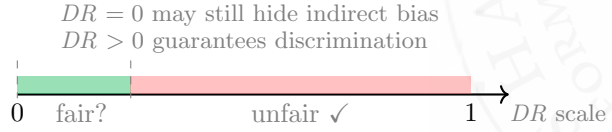
$\hat{\boldsymbol{L}}_{\text{bias}}(f,S) = \frac{1}{n}\sum_{i=1}^{n} \ell_{\text{bias}}(f,\boldsymbol{x}_i)$

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

Discriminative risk (DR) [4]

—from an individual-level aspect
—from a group-level aspect

To describe this characteristic of the hypothesis on multiple instances, then the empirical *DR* on one dataset S is expressed as $\hat{\boldsymbol{L}}_{\text{bias}}(f,S) = \frac{1}{n}\sum_{i=1}^{n} \ell_{\text{bias}}(f,\boldsymbol{x}_i)$, and the true *DR*[3] of the hypothesis over a data distribution is $\boldsymbol{L}_{\text{bias}}(f) = \mathbb{E}_{(\boldsymbol{x},y)\sim\mathcal{D}}[\ \ell_{\text{bias}}(f,\boldsymbol{x})\ ]$, respectively.



Key caveat

$DR = 0$ is *necessary but not sufficient* for fairness.

When DR > 0, discrimination is *guaranteed* to exist.

[3] 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 $\mathcal{D}$.

[4] $\hat{\boldsymbol{L}}_{\text{bias}}(f,S)$ is an *unbiased estimator* of $\boldsymbol{L}_{\text{bias}}(f)$

How *DR* relates to existing fairness measures

	Group fairness[5]	**Individual fairness**[6]	**Causal fairness**[7]	***DR***
Subgroup partition	required	—	—	*not required*
Similarity metric	—	required	—	*not required*
Causal graph	—	—	required	*not required*
Multi-valued SA	✗	✗	✓	✓
Original data only	Yes	Yes	No	No

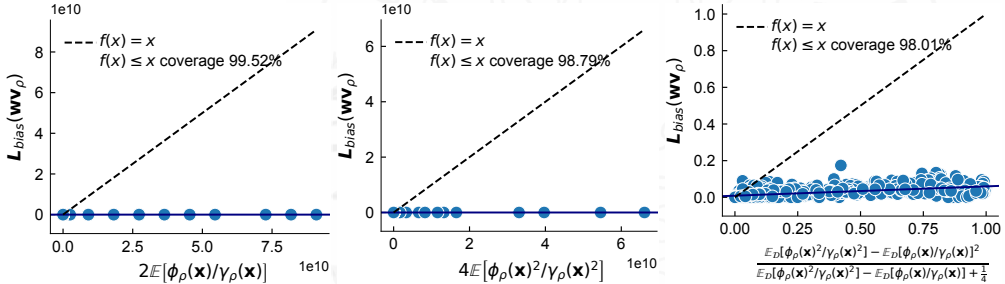
DR follows the individual-fairness principle, aggregates like group fairness, and probes sensitivity like CFF — without requiring a specific metric, a subgroup partition, or a causal model.

[5] e.g. DP, EO, EOpp, and PP. *(Empirically, DR has 0.58 Pearson correlation with accuracy change under perturbation, versus near-zero for DP and PP.)*
[6] e.g. Lipschitz condition
[7] e.g. CFF, proxy discrimination. *(DR reflects counterfactual fairness (CFF) in a quantifiable way, without requiring a causal graph.)*

Oracle bounds: can ensemble combination reduce *DR*?

1st order: $L_{\text{bias}}(\mathbf{wv}_\rho) \leqslant 2\,\mathbb{E}_{\mathcal{D}}[\phi_\rho(x)/\gamma_\rho(x)]$

2nd order: $L_{\text{bias}}(\mathbf{wv}_\rho) \leqslant 4\,\mathbb{E}_{\mathcal{D}}[\phi_\rho(x)^2/\gamma_\rho(x)^2]$



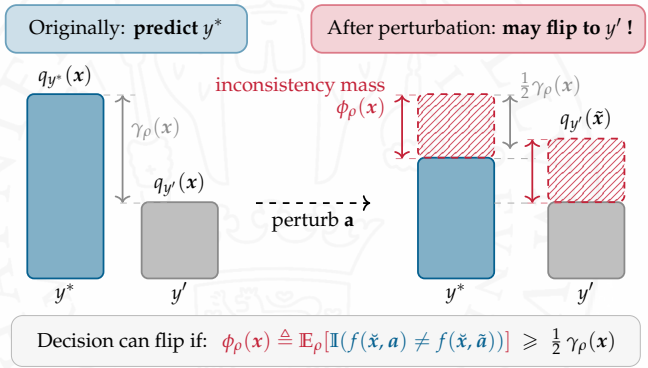
Bounds hold >98%, 2nd tighter, C-tandem tighter still
Points below the identity line = bound holds

What the bounds tell us

Ensembles of m classifier with weights ρ
Top-1 voting margin

$$\gamma_\rho(\mathbf{x}) \triangleq q_{y^*}(\mathbf{x}) - \max_{y \neq y^*} q_y(\mathbf{x}) \ \in [0,1]$$

If the ensemble exhibits discrimination risk, it means that the leading gap in the prediction class has been reversed, in other words, at least half of the votes for the prediction class will be altered. That is to say, the total weight of the inconsistency between the original prediction and the perturbed prediction has to hold



DR of the ensemble $\leqslant 2\times$ (individual *DR* / voting margin)
Larger margin $\rightarrow$ tighter bound $\rightarrow$ fairer ensemble
A *cancellation-of-biases* effect exists — but it depends on margins, unlike cancellation of errors

Summary [8]

What we proposed

- *DR* captures both individual and group fairness via SA perturbation (no metric, no causal graph, no subgroup partition needed), and applies naturally to multi-valued & multiple SAs
- *Oracle bounds* indicate that a cancellation-of-biases effect exists in combination, margin-dependent
- *POAF*: ensemble pruning via Pareto optimality, improving fairness with acceptable accuracy loss

What this means

- Fairness *can* be improved via ensembling with theoretical guarantees, not just empirical hope
- Larger voting margin $\Rightarrow$ tighter fairness bounds: boosting performance helps fairness too
- *POAF* saves effort on repetitive hyperparameter tuning

Limitations & future work

- *DR* involves randomness from SA perturbation
- *POAF* has higher time cost than simpler baselines
- Acceleration of *POAF* and stability of *DR* worth exploring

[8] Code: `https://github.com/eustomaqua/FairML`