Gradual (In)Compatibility of Fairness Criteria

Corinna Hertweck*[1,2] and Tim Räz*[3]

[1]Zurich University of Applied Sciences, [2]University of Zurich, [3]University of Bern

* Equal contribution

1. BACKGROUND

- *Impossibility theorems*: important fairness criteria are incompatible: independence, separation and sufficiency
- What if stakeholders have conflicting views about what fairness criteria are fitting?
- **To what extent are fairness criteria (in)compatible when only partial fulfillment is required?**

2. PARTIAL FULFILLMENT OF FAIRNESS CRITERIA

Use Information theory to define partial fulfillment:

Independence: $R \perp A$: $I(R;A) = 0$
Independence gap (IND) of degree d if $I(R;A) \leq d$

Separation: $R \perp A \mid Y$: $I(R;A|Y) = 0$
Separation gap (SEP) of degree d if $I(R;A|Y) \leq d$

Sufficiency: $Y \perp A \mid R$: $I(Y;A|R) = 0$
Sufficiency gap (SUF) of degree d if $I(Y;A|R) \leq d$

3. EXPERIMENTAL DESIGN

- Train logistic regression with the loss function L
- Evaluate trained models using normalized criteria and 5-fold cross validation

$$L = l_{fit} + \lambda \cdot l_2 + \mu \cdot l_{fair}$$

l_{fit} : cross-entropy

$\lambda \cdot l_2$: L2 regularization

$\mu \cdot l_{fair}$: regularizes {IND, SEP, SUF, balance, negative-accuracy}

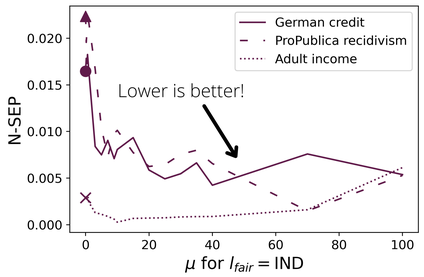
We can improve the fairness criteria independence and separation at the same time despite the impossibility theorems.

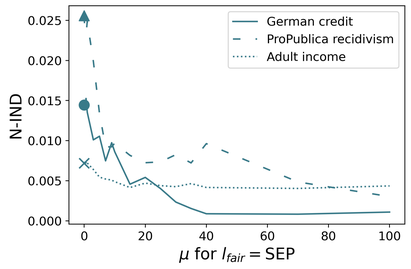
4. MAIN RESULT

Indirect regularization between independence and separation

$$\underbrace{I(R;A|Y)}_{\text{separation}} = -\underbrace{I(Y;R)}_{\text{accuracy}} + \underbrace{I(Y;R|A)}_{\text{balance}} + \underbrace{I(A;R)}_{\text{independence}}$$

Independence is part of the decomposition of separation, so we hope for indirect regularization effects between independence and separation.



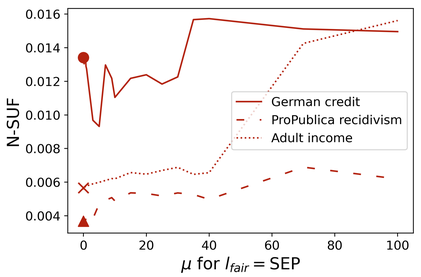


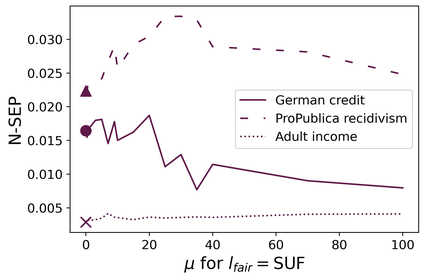
Regularizing independence gap (IND) **improves** normalized separation gap (N-SEP)
Regularizing separation gap (SEP) **improves** normalized independence gap (N-IND)

No indirect regularization between separation and sufficiency

$$\underbrace{I(R;A|Y)}_{\text{separation}} = -\underbrace{I(Y;R)}_{\text{accuracy}} + \underbrace{I(Y;R|A)}_{\text{balance}} + \underbrace{I(A;R)}_{\text{independence}} \qquad \underbrace{I(Y;A|R)}_{\text{sufficiency}} = -\underbrace{I(Y;R)}_{\text{accuracy}} + \underbrace{I(Y;R|A)}_{\text{balance}} + \underbrace{I(A;Y)}_{\text{legacy}}$$

Accuracy and balance are part of the decomposition of both separation and sufficiency, so we hope for indirect regularization effects between separation and sufficiency.





Regularizing separation gap (SEP) **doesn't improve** normalized sufficiency gap (N-SUF)
Regularizing sufficiency gap (SUF) **doesn't improve** normalized separation gap (N-IND)