Universiteit van Amsterdam

Neural Information Processing Systems

Reproducibility Study: Equal Improvability

A New Fairness Notion Considering the Long-Term Impact

Berkay Chakar   Amina Izbassar   Mina Janicijevic   Jakub Tomaszewski

Motivation

- Most of existing fairness notions only consider immediate fairness, without taking into account the equal improvement possibility of the members of the different groups.

- In contrast, Equal Improvability (EI) is an effort-based fairness notion that concerns itself with long-term fairness.

- Real world applications could be found in areas where the group members can improve their features and be re-labelled, e.g. loan approval, college admissions.

Equal Improvability

Group 0's rejected samples are much closer to the boundary than Group 1 (less effort to cross the boundary)

Disparity in Improvability (unfairness)

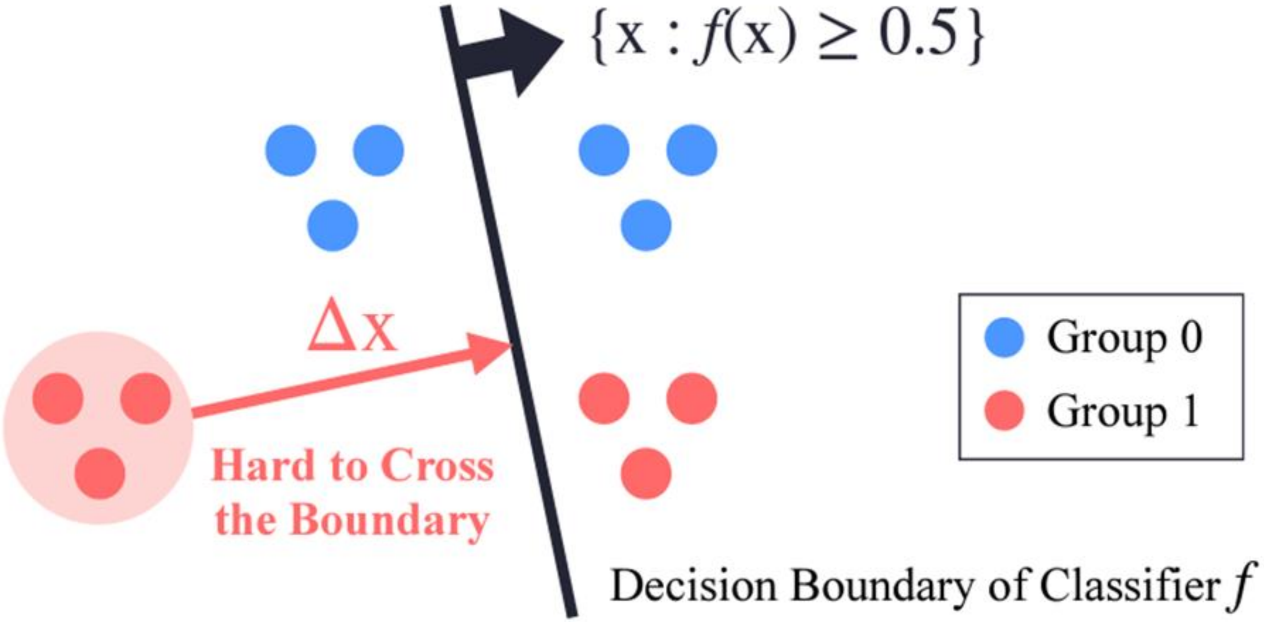
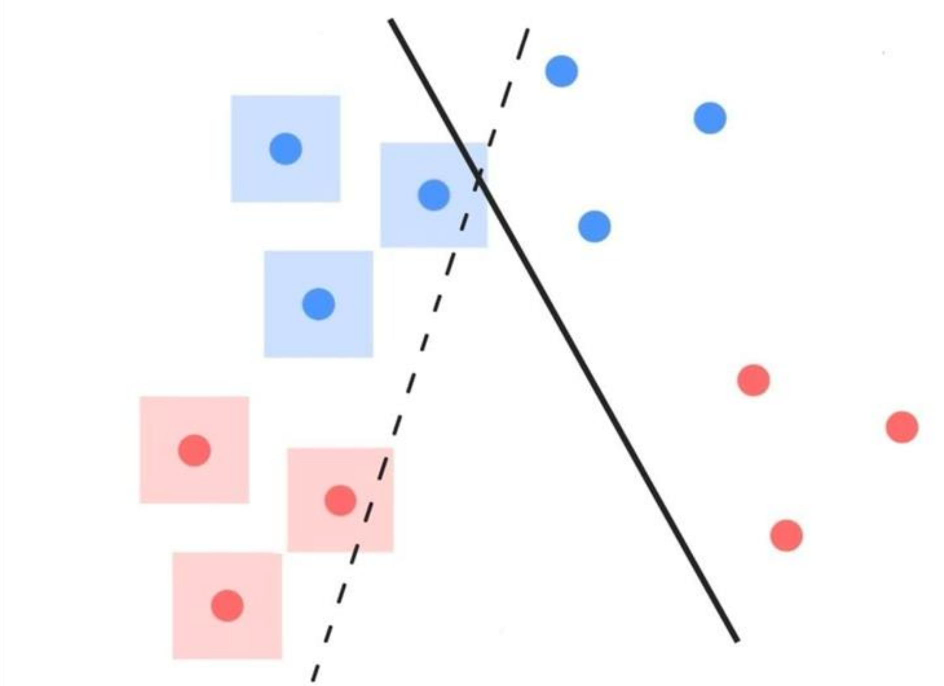


Image Source: Guldogan et al, 2023

Equal Improvability

Effort-based fairness notion that aims to balance the potential acceptance rates for rejected applicants across various groups, given a fixed amount of effort.



Classifier achieving EI Classifier **NOT** achieving EI

Equal Improvability Penalty

$$\min_{f \in \mathcal{F}} \left\{ (1-\lambda) \frac{1}{N} \sum_{i=1}^{N} \ell(y_i, f(x_i)) + \boxed{\lambda U_\delta} \right\}$$

- Covariance-based penalty: $\left(\mathrm{Cov}(\mathrm{z}, \max_{\|\Delta \mathbf{x}_\mathrm{I}\| < \delta} f(\mathbf{x} + \Delta \mathbf{x}) \mid f(\mathbf{x}) < 0.5)\right)^2$
- KDE-based penalty: $|KDE(EI\ Disparity)|$
- Loss-based penalty: $\sum_{z \in \mathcal{Z}} \left| \frac{1}{|I_{-,z}|} \sum_{i \in I_{-,z}} \ell\left(1, \max_{\|\Delta \mathbf{x}_{\mathrm{I}_i}\| \le \delta} f(\mathbf{x}_i + \Delta \mathbf{x}_i)\right) - \sum_{z \in \mathcal{Z}} \frac{I_{-,z}}{I_-} \tilde{L}_z \right|$

Claims of the original paper

Claim 1: A classifier obtained by each of the three proposed EI ensuring methods, has a significantly smaller EI disparity value than the ERM (Empirical Risk Minimization) approach and a comparable error.

Claim 2: Most existing methods have an adverse effect on long-term fairness, while EI continues to enhance it.

Claim 3: The introduced methods of achieving Equal Improvability prevent an over-parametrized classifier from overfitting the data.

Experimental setup - Datasets

Datasets	Samples	Classes	Sensitive attrs.
Synthetic	20,000	2	1
German Statlog Credit	1,000	2	1 & 2
ACS-Income-CA	195,665	2	1 & 2
$\boxed{\text{Default of Credit Card Clients}}$	30,000	2	1

New dataset!

Experimental setup - Models

- Logistic Regression
- Multilayer Perceptron Model

Reproducibility experiments

1

EI Disparity and Loss value check

(Claim 1)

2

Long-term (un)fairness check

(Claim 2)

3

Overfitting robustness check

(Claim 3)

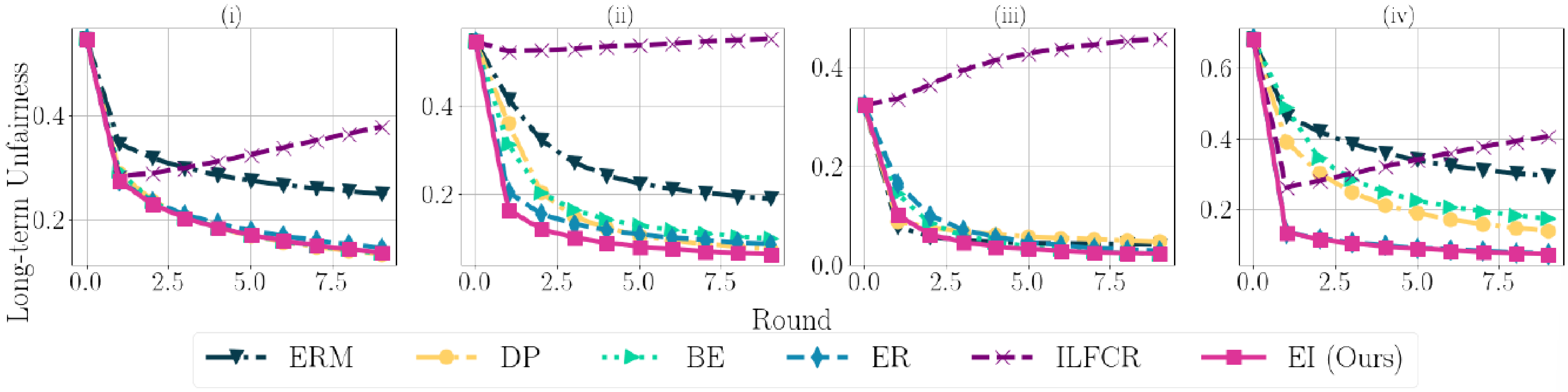
Reproducibility experiments

- Reported (left side) and Reproduced (right side) Error rate and EI Disparity values of ERM and three proposed methods for achieving EI with a Logistic Regression model.

Dataset	Metric	ERM	Covariance-Based	KDE-Based	Loss-Based
Synthetic	Error Rate	.221 | .222	.253 | .253	.250 | .253	.246 | .246
	EI Disp.	.117 | .118	.003 | .003	.003 | .003	.002 | .002
German Stat.	Error Rate	.220 | .262	.262 | .262	.243 | .249	.237 | .237
	EI Disp.	.041 | .021	.021 | .022	**.035** | **.226**	.015 | .016
ACSIncome-CA	Error Rate	.184 | .185	.200 | .200	.196 | .196	.193 | .195
	EI Disp.	.031 | .031	.008 | .008	.005 | .005	.006 | .006

Reproducibility Experiments

- The disparity between the sensitive group feature distributions reduces faster for the EI classifier than for the other metrics.
- This indicates that EI classifier is more favorable for achieving long-term fairness.



Reproducibility Experiments

- Error rate and EI disparities of ERM and the proposed EI-regularized methods on an overparameterized Multilayer perceptron (MLP) using a subset of German Statlog Credit dataset.

Metric	ERM	Covariance-Based	KDE-Based	Loss-Based
Train Error	$.218 \pm .004$	$.233 \pm .003$	$.226 \pm .009$	$.233 \pm .012$
Test Error	$.218 \pm .010$	$.218 \pm .010$	$.222 \pm .007$	$.231 \pm .007$
Train EI Disp.	$.024 \pm .017$	$.018 \pm .011$	$.018 \pm .012$	$.009 \pm .009$
Test EI Disp.	$.064 \pm .036$	$.050 \pm .024$	$.070 \pm .050$	$.062 \pm .015$

Reproducibility experiments

1 ✓

EI Disparity and Loss value check

(Claim 1)

2 ✓

Long-term (un)fairness check

(Claim 2)

3 ✓

Overfitting robustness check

(Claim 3)

Extended analysis

1

Evaluating EI Disparity and Error Rate values on a different dataset

(Claim 1)

2

Adding another sensitive feature

(Claim 1)

3

Long-term fairness with multiple sensitive features

(Claim 2)

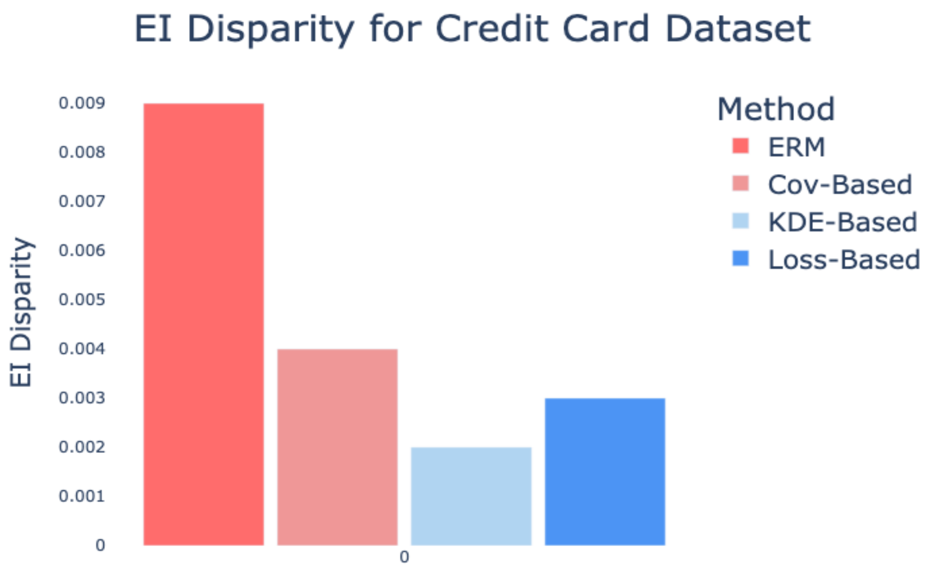
4

Overfitting robustness check with a more complex model

(Claim 3)

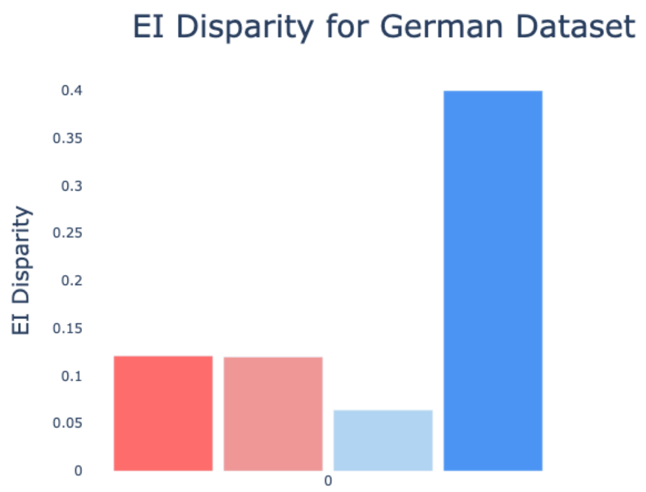
Result 1 – EI Disparity and Error Rate values on a different datasets

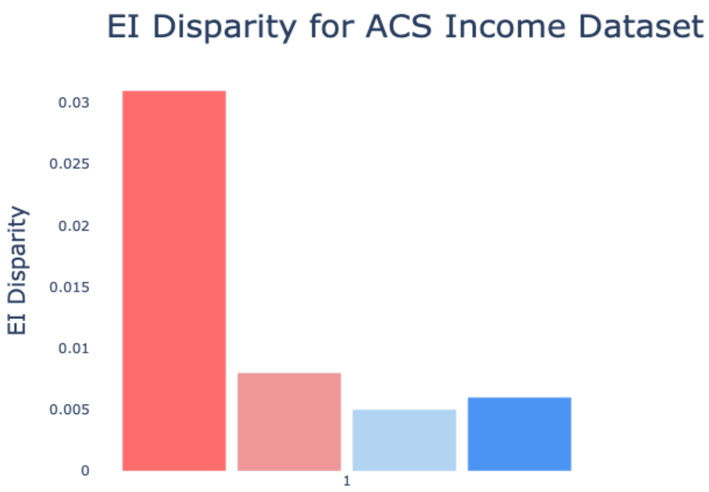
- EI-based classifiers still have a lower EI disparity without causing a significant increase in the error rate on the new dataset.

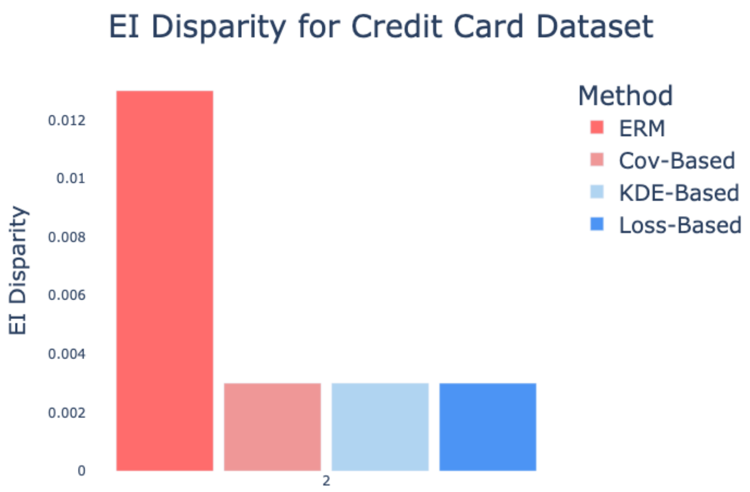


Result 2 – Adding another sensitive feature (Sex and Age)

- The measured EI Disparity for the original datasets using ERM and each of the 3 penalty terms optimized for 2 sensitive features.

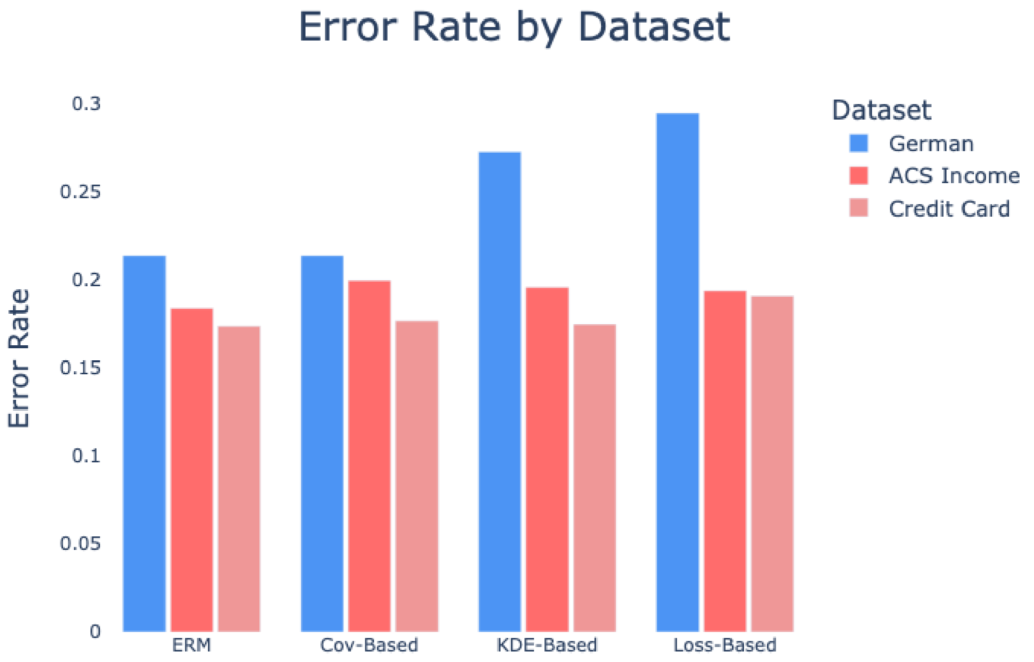






Result 2 – Adding another sensitive feature (Sex and Age)

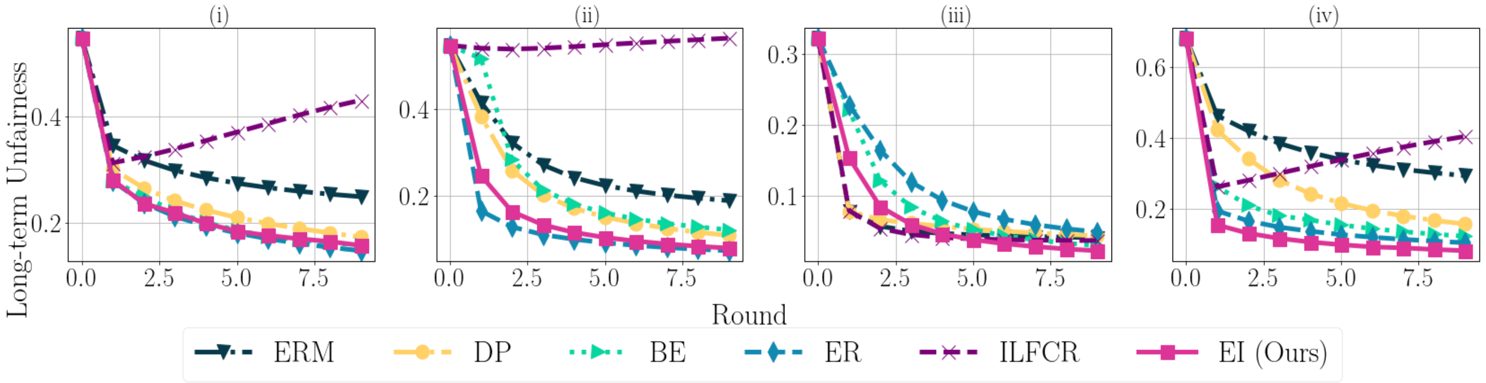
- The results indicate that EI Disparity of the proposed methods can still be low, without significantly increasing the Error Rate even with 2 sensitive features.



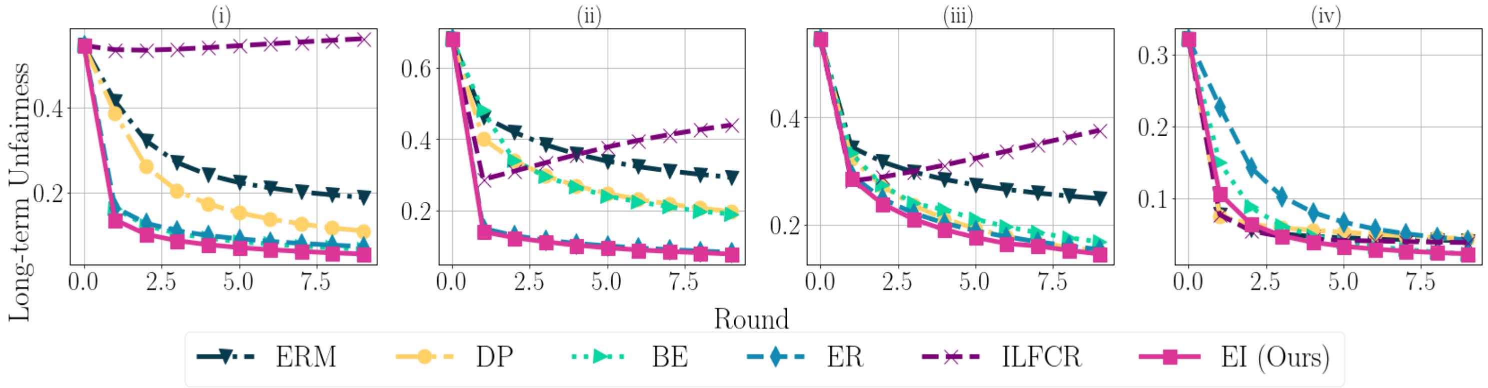
Result 3 – Long Term fairness with multiple sensitive features

The disparity between the feature distributions of different sensitive groups reduces faster for the EI classifier

Feature 1



Feature 2



Result 4 – Overfitting robustness

- Error rate and EI disparities of ERM and the proposed EI-regularized methods on an overparameterized Multilayer perceptron (MLP) using a subset of ACS-Income dataset.
- The error rate and EI disparity values for all methods are indicative of overfitting

Conclusion

- The reproducibility study proved the general claims of the original paper.
- Experiments on a different dataset also support the claims.
- Experiments with 2 sensitive features produced the results that were in line with the authors' claims except for the Loss-based method.
- Further experiments did not substantiate EI's robustness to overfitting.

Thank you for your attention!