Fairness of Learned Classifiers under Performative Effects

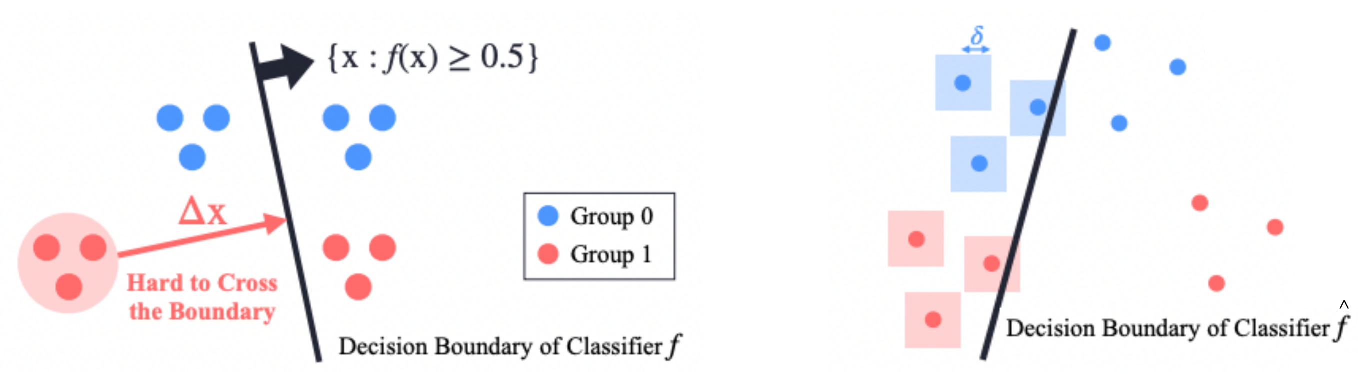
Sophia Günlük, António Góis, Simon Lacoste-Julien, Dhanya Sridhar

Introduction

Bail and Parole

Loan Applications

- Decision-making has been increasingly automized using machine learning, such as for banks giving out loans or judges determining bail or parole.
  - Trained using historical data: $(x_i, y_i) \overset{i.i.d.}{\sim} \mathscr{D}^\theta$
  - Goal: learn risk-scoring function $\hat{f}(x) \approx \mathbb{P}(Y = 0 \mid X = x)$ in order to identify positive predicted outcomes, using decision function $D(x) = 1\{\hat{f}(x) \leq \tau\}$
    - Objective: $\hat{f} = \arg\min_{f \in \mathcal{F}} \mathbb{E}_{(x,y)\sim\mathscr{D}^\theta}[\ell(f(x), y)]$
- Predictors naively trained can inherit bias from historic data, based on sensitive attributes such as the demographic attributes of an applicant.
  - Previous work to assess bias of learned classifier has led to static fairness metrics, such as demographic parity.
  - However, long-term effects of the classifier are important to consider; rejected applicants may adapt their features in order to get a better outcome if they reapply (which we call a *performative effect*)

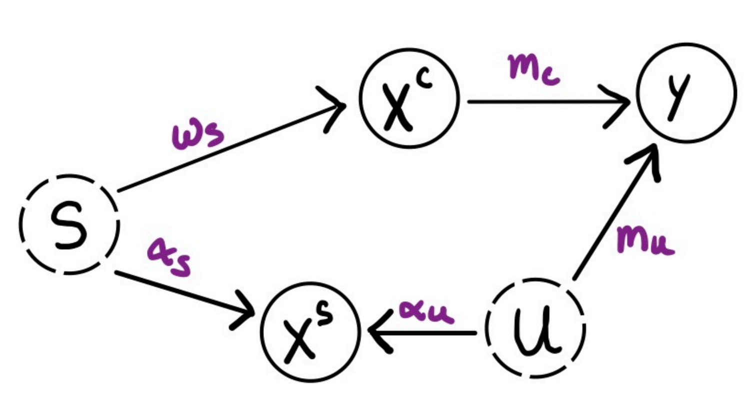


Guldogan, O., Zeng, Y. etc. Equal Improvability: A New Fairness Notion Considering the Long-term Impact. https://arxiv.org/abs/2210.06732

- Both classifiers have the same accuracy and achieves static fairness (demographic parity), but decision boundary f is harder for group 1 to cross than $\hat{f}$, and group 0 has an easier time than group 1 crossing both f and $\hat{f}$.
- ***Hypothesis***: In the performative setting, a learned classifier that uses non-causal and spurious features for prediction can lead to negative externalities, such as non-static unfairness.

Setup

- Variable Definition:
  - $X \in \mathbb{R}^{nxd}$: features observed by the decision-maker or classifier
  - $S, U \in \{0,1\}$: unobserved confounding characteristics/variables
  - $Y \in \mathbb{R}^n$: true outcome
  - $(\hat{f})\ f$: (classifier's prediction) outcome function
- Our Structural Causal Model (SCM):



Non-static Fairness Metrics

- Relevant definitions:

  δ: maximum effort, $(\mathbb{R}^{\geq 0})$

  μ: cost function, $(\mathcal{X} \to \mathbb{R}^{\geq 0})$

  $(\hat{f})\ f$: (estimated) probability function of Y = 1, $(\mathcal{X} \to [0,1])$

- Adaptation definition:

$$\Delta^{(f)} = \arg\max_{\Delta} \quad \delta * 1_{f(x+\Delta)\geq 0.5} - \mu(\Delta)$$

$$\hat{\Delta}^{(\hat{f})} = \arg\max_{\hat{\Delta}} \quad \delta * 1_{\hat{f}(x+\hat{\Delta})\geq 0.5} - \mu(\hat{\Delta})$$

- Our long term fairness metrics:

Improvability: $\mathbb{P}\left(f(x+\hat{\Delta}) \geq 0.5 \mid f(x) < 0.5,\ f(x+\Delta) \geq 0.5\right)$

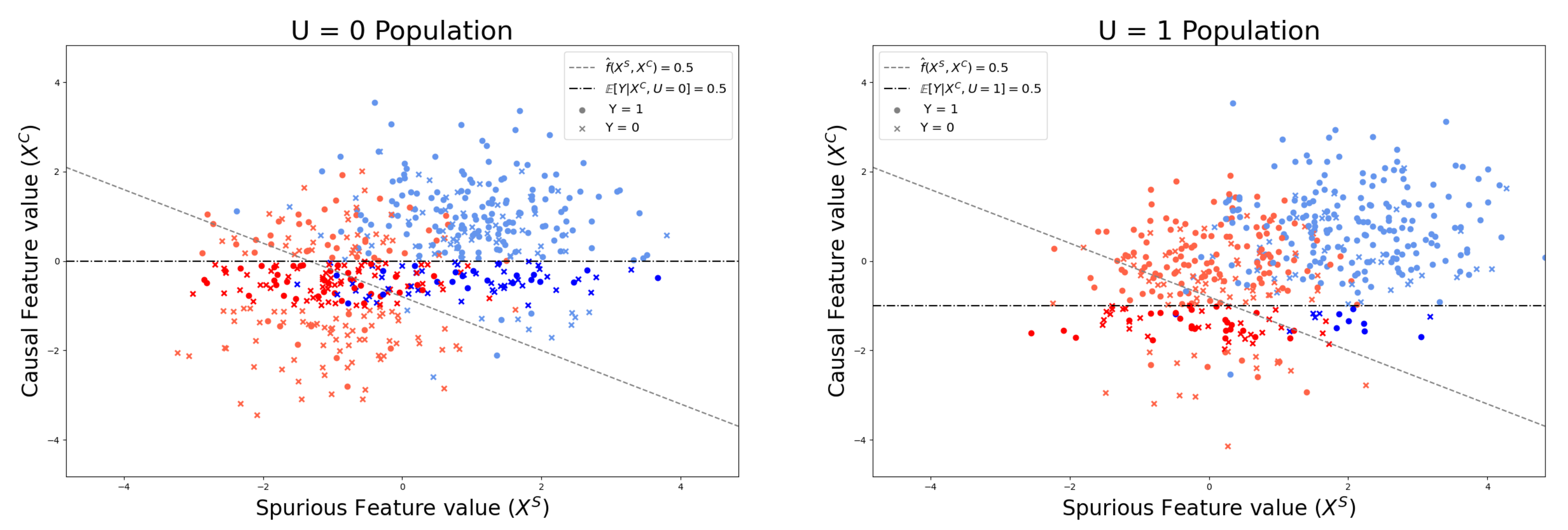
- *Out of those who could improve their real outcome with δ effort, what is the probability that they would improve their real outcome when adapting in response to the classifier?*

Gaming: $\mathbb{P}\left(\hat{f}(x+\hat{\Delta}) \geq 0.5 \mid f(x) < 0.5,\ f(x+\hat{\Delta}) < 0.5\right)$

- *Out of those who could not improve their real outcome with δ effort when adapting in response to the classifier, what is the probability that they would also improve their real outcome?*

Our Metrics Visualized

- For experiments: $\delta = 1$, and all DGM coefficients have value 1.
  - Each point represents a sampled individual, coordinate represents the features
  - $S = 0$: Red, $S = 1$: Blue. Darker colors represent individuals who could have improved their true label with the limited effort.



- Learned coefficients: $w = (w_s = 0.6,\ w_c = 1)$ and $b = 0.8$

	Minority	Majority
Improvability	0.277	0.031
Gaming	0.503	0.977

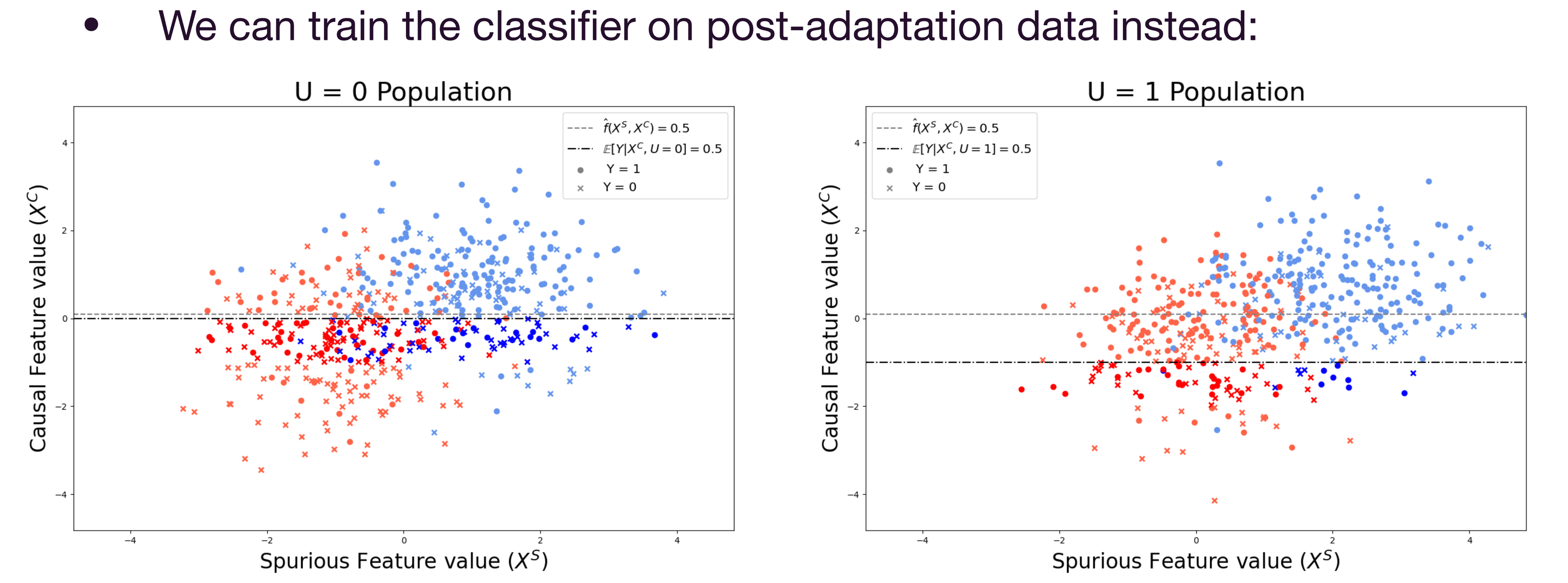
- Naively training by maximizing the accuracy:
  - The model is motivated to learn a classifier that uses non-causal feature X^S, due to the unobserved confounders in the true data generating process.
  - Gaming is very high for the majority group, and improvement is very low, it is less extreme for the minority group but still relatively poor metric values.

Ongoing Work

- Post-adaptation data: observations generated after individuals adapt their features in response to the classifier after one time-step.

$$X_{post}^{(\hat{f})} = x + \hat{\Delta}^{(\hat{f})} \qquad\qquad Y_{post}^{(\hat{f})} = f(X_{post}^{(\hat{f})})$$

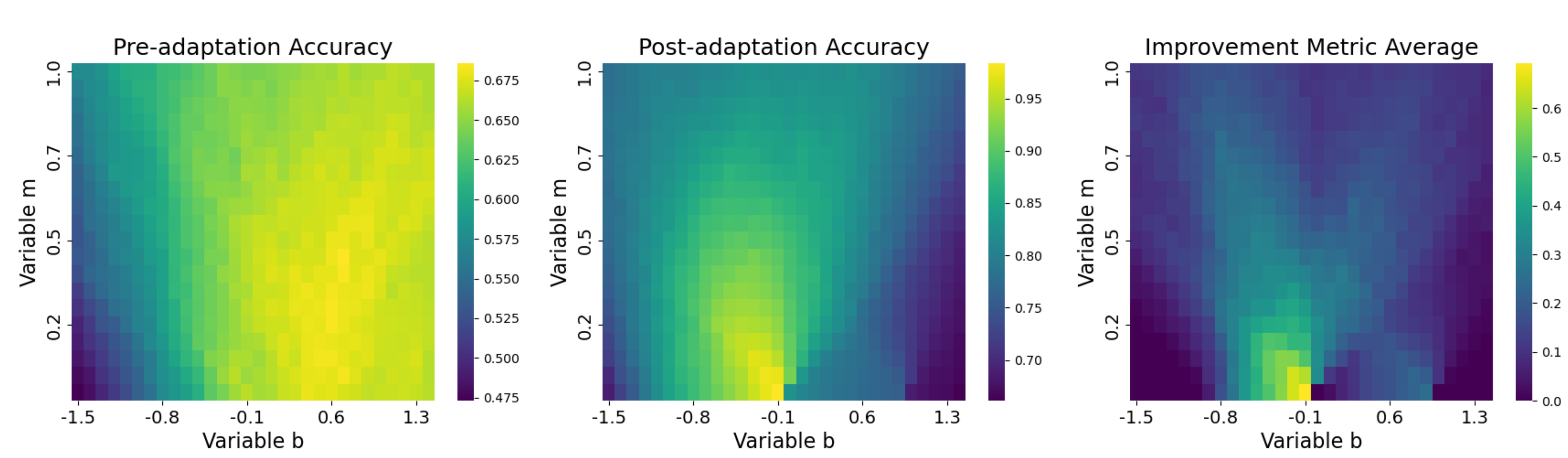
- We can train the classifier on post-adaptation data instead:



- Learned coefficients: $w = (w_s = 0,\ w_c = 1)$ and $b = -0.1$

	Minority	Majority
Improvability	0.526	0.723
Gaming	0.0	0.0

- The new classifier uses only the causal feature for prediction, and is more fair terms of improvability and gaming.
- Furthermore, it is the optimal classifier with respect to our improvability metric:



- Each point of the heat map represents the respective metric for a classifier with decision boundary: $m * X^S + X^C + b = 0$
- ***Conclusion***: a model trained with ERM on post-adaptation data finds a causal predictor, which is the optimal classifier with respect to improbability

Future Directions:

- Propose methods to approximately maximize post-adaptation metrics
- Propose post-adaptation goals (alternative to only accuracy) that bring more societal benefit or parity fairness
- Consider multiple time-steps