Alternative Microfoundations for Strategic Classification

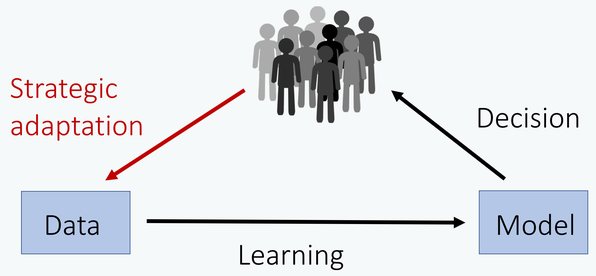
Meena Jagadeesan, Celestine Mendler-Dünner, Moritz Hardt

Berkeley UNIVERSITY OF CALIFORNIA

Contributions

- ✓ We show that standard microfoundations serve as a poor basis for studying agent behavior in binary classification.
- ✓ We explore alternative microfoundations for strategic classification, and we identify noisy response as a promising candidate model.

Classification in Strategic Environments



Decision rule induces users to strategically adapt their features.

Microfoundations for Strategic Adaptation

Microfoundations ≈ grounding theories of aggregate outcomes in microeconomic assumptions about individual behavior [5]

Benefit from ML perspective: Microfoundations endow strategic distribution shifts with structure.

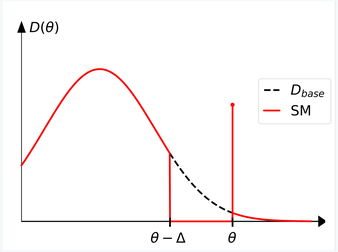
Standard microfoundations (SM): "*Agents maximize a utility function on the basis of perfectly accurate information*"

1. Cost $c: X \times X \to \mathbb{R}^{\geq 0}$ for changing features.
2. Utility of changing features to x' is $f_\theta(x') - c(x, x')$
3. Agents change features to: $\text{argmax}_{x \in X}[f_\theta(x') - c(x, x')]$

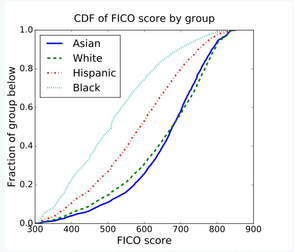
Degeneracies of Standard Microfoundations

#1

Proposition: any aggregate distribution resulting from standard microfoundations (SM) in response to a decision rule is either discontinuous or trivial.



Distribution predicted by SM in response to threshold classifier θ

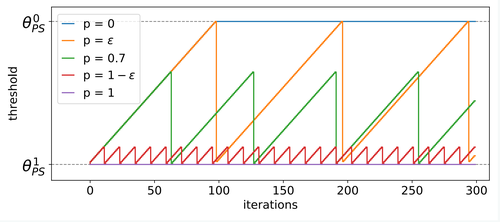


Observed distribution (FICO) in [2] is continuous

#2

Performative stability [4]: repeatedly retrain classifier weights to be optimal on the data distribution induced by the previous classifier.

Proposition: repeated retraining does not converge when *any* randomly chosen fraction $p \in (0,1)$ of agents are non-strategic.



Convergence of repeated retraining for SM

#3

Performative optimality [4]: anticipate distribution shifts based on microfoundation models (e.g. SM) to find the optimal point.

Proposition: standard microfoundations lead to extreme solutions that maximize negative externalities within a large class of alternative models for agent behavior.

Selecting Alternative Microfoundations

Idea: identify properties to navigate the space of alternative models

Property 1: Aggregate smoothness
- ➢ Requires that the aggregate distribution is smooth
- ➢ Guarantees the *robust existence* of fixed points of repeated retraining methods

Property 2: Expenditure constraint
- ➢ Constrains how much agents expend on gaming
- ➢ Helps ensure that agent responses are natural

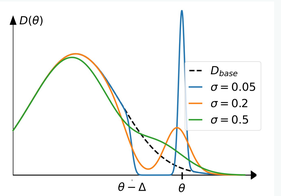
Candidate Model: Noisy Response

Captures imperfect agents using intuition from smoothed analysis (remaining agnostic to sources of imperfection).

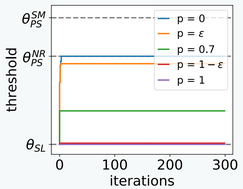
Idea: add random perturbations to perceptions

$$\text{argmax}_{x \in X}[f_{\theta+t}(x') - c(x, x')] \quad \text{(NR)}$$

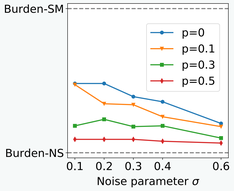
where t is distributed as a Gaussian across the population.



Distributions predicted by NR in response to threshold classifier θ



Convergence of repeated retraining for NR



Negative externalities of optimal points for NR

Selected References

[1] Hardt, Megiddo, Papadimitriou, Wooters. "*Strategic Classification*". 2016.
[2] Hardt, Price, Srebro. "*Equality of Opportunity in Supervised Learning*". 2016.
[3] Björkegren, Blumenstock, Knight. "*Manipulation-Proof Machine Learning.*" 2020.
[4] Perdomo, Zrnic, Mendler-Dünner, Hardt. "*Performative Prediction*". 2020.
[5] Robert E Lucas Jr. "*Econometric policy evaluation: A critique*." 1976.