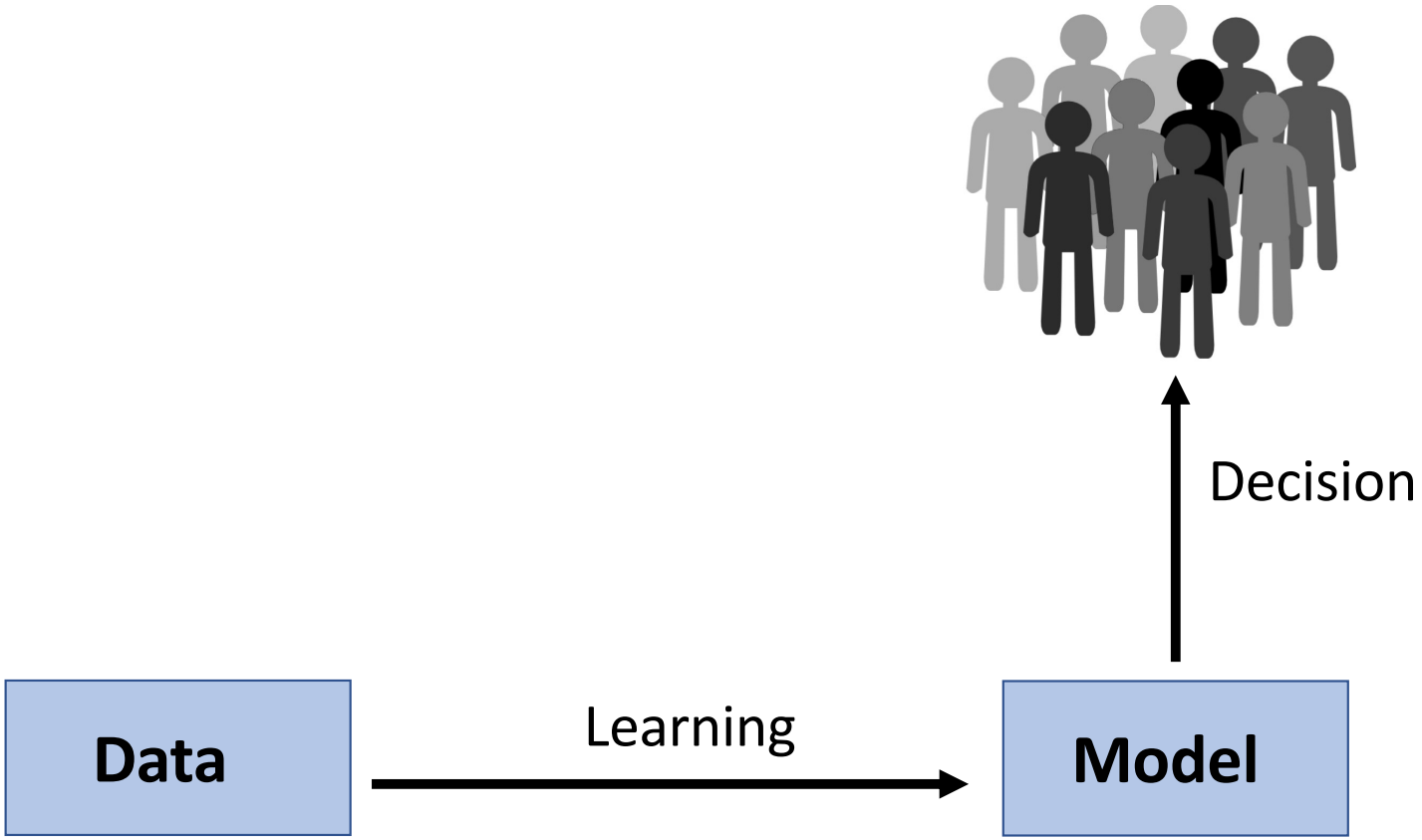
Alternative Microfoundations for Strategic Classification

Meena Jagadeesan, Celestine Mendler-Dünner, and Moritz Hardt

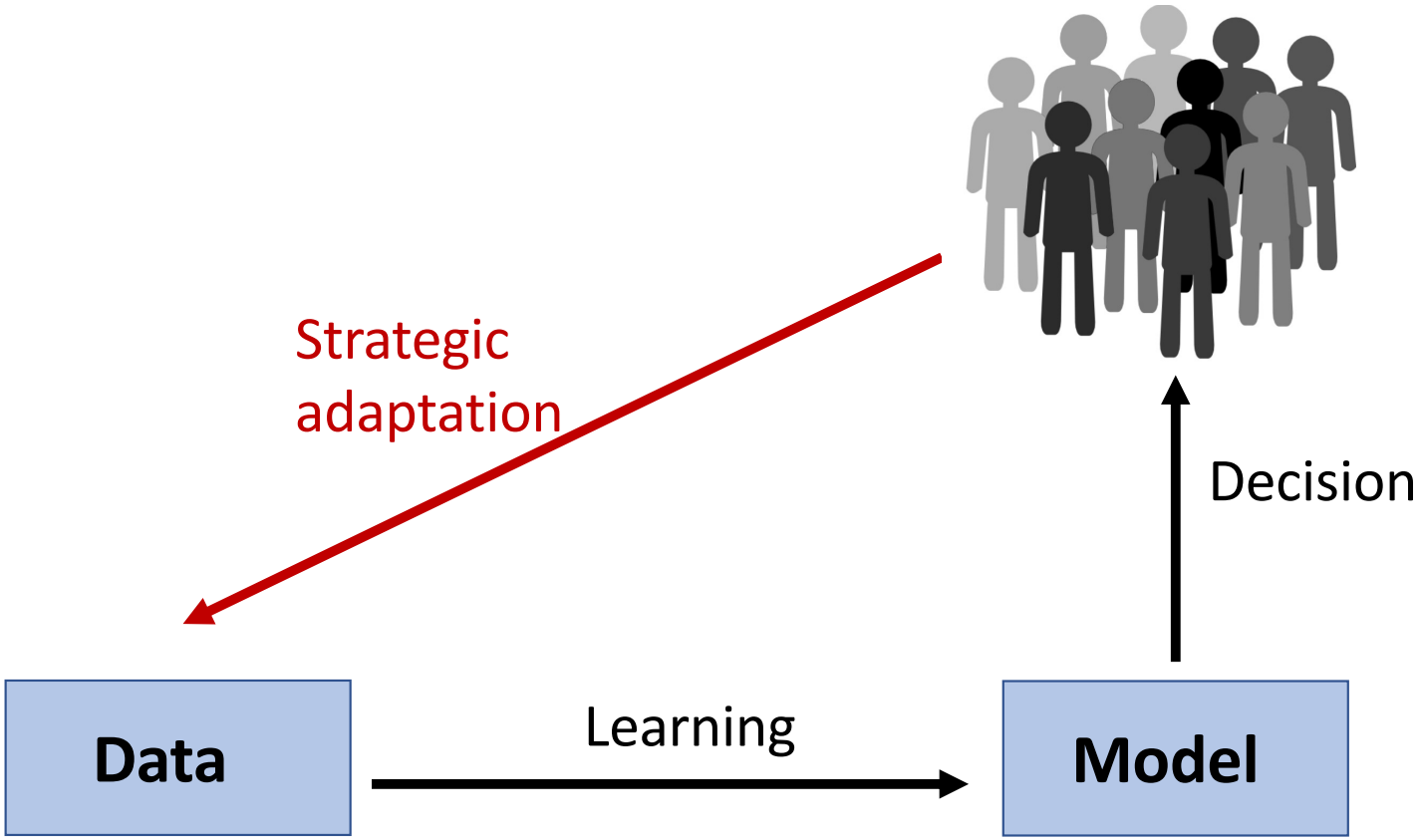
UC Berkeley

ICML 2021

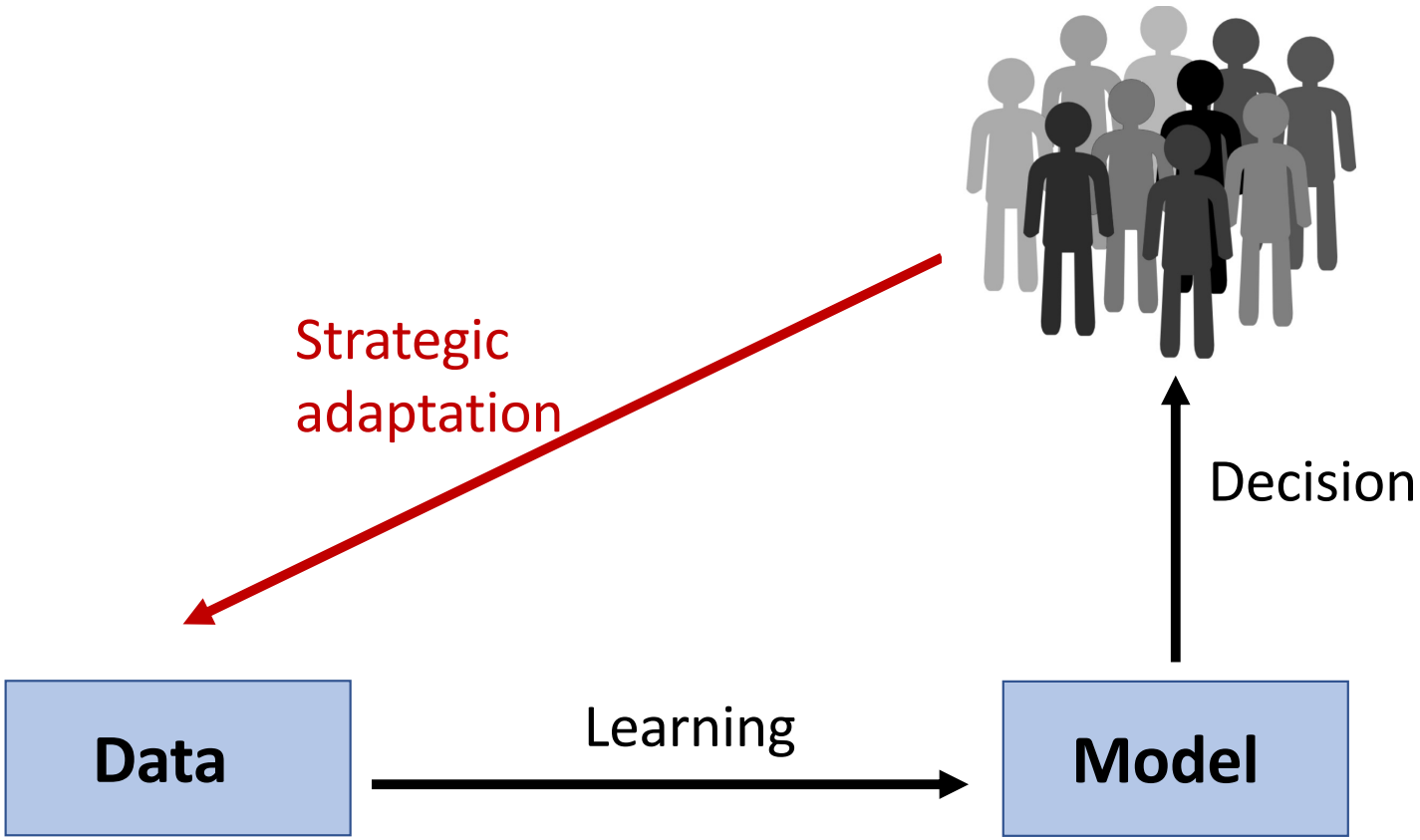
Algorithmic Decision-Making in Strategic Environments



Algorithmic Decision-Making in Strategic Environments

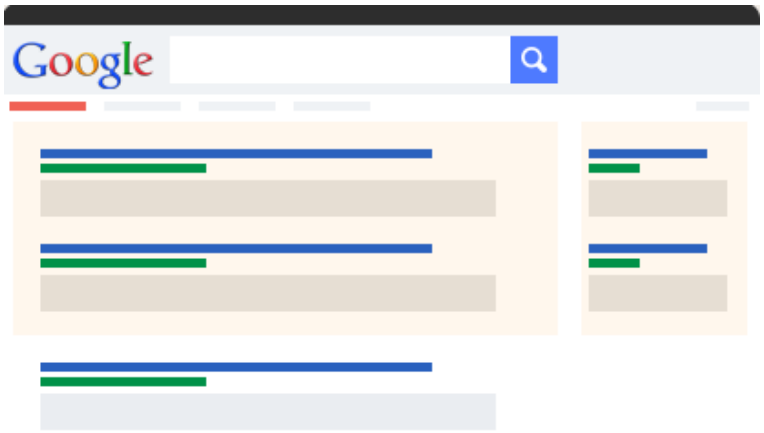


Algorithmic Decision-Making in Strategic Environments

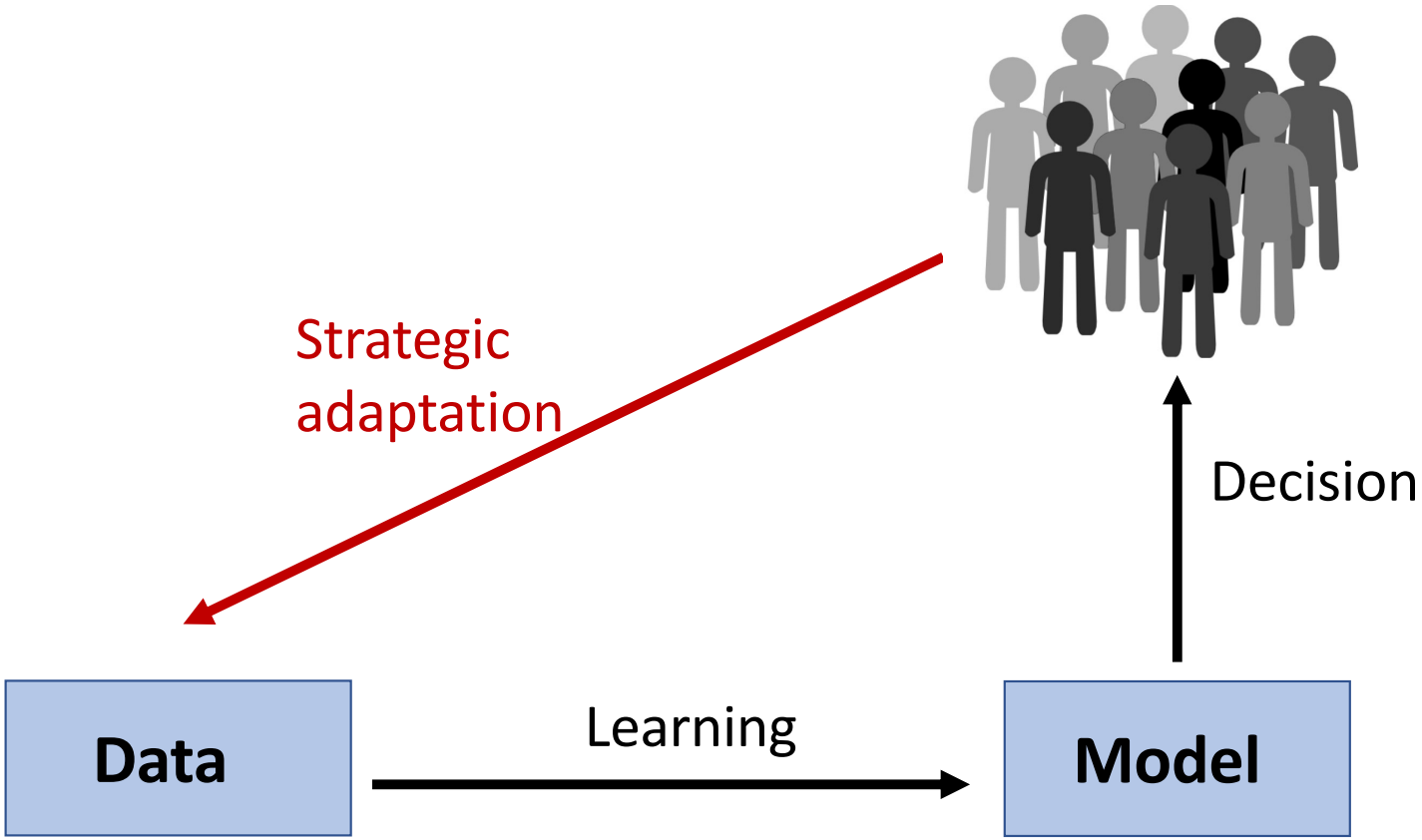


We show standard approaches to anticipate strategic adaptation combine poorly with binary classification.

Decision-Making Tasks with Strategic Adaptation



Algorithmic Decision-Making in Strategic Environments



The decision rule can trigger changes in the observed data distribution.

Using Microfoundations to Anticipate Distribution Shifts

Microfoundations $\approx$ grounding theories of aggregate outcomes in *microeconomic assumptions* about individual behavior.

Using Microfoundations to Anticipate Distribution Shifts

Microfoundations $\approx$ grounding theories of aggregate outcomes in *microeconomic assumptions* about individual behavior.

"Standard microfoundations": *agents maximizes a utility function on the basis of perfectly accurate information*

Using Microfoundations to Anticipate Distribution Shifts

Microfoundations $\approx$ grounding theories of aggregate outcomes in *microeconomic assumptions* about individual behavior.

"Standard microfoundations": *agents maximizes a utility function on the basis of perfectly accurate information*

Standard microfoundations (SM) are followed in *strategic classification*.

- Agents have cost $c : X \times X \to \mathbb{R}^{\geq 0}$ of changing features.
- Agents change features to: $\arg\max_{x' \in X} \ [f_\theta(x') - c(x, x')]$.

Using Microfoundations to Anticipate Distribution Shifts

Microfoundations $\approx$ grounding theories of aggregate outcomes in *microeconomic assumptions* about individual behavior.

"Standard microfoundations": *agents maximizes a utility function on the basis of perfectly accurate information*

Standard microfoundations (SM) are followed in *strategic classification*.

- Agents have cost $c : X \times X \to \mathbb{R}^{\geq 0}$ of changing features.
- Agents change features to: $\arg\max_{x' \in X} \ [f_\theta(x') - c(x, x')]$.

Our Contribution

Standard microfoundations are a poor basis for studying strategic behavior in binary classification. We propose alternative microfoundations models.

Result I: SM Cannot Capture Observed Distributions

Proposition (Informal)

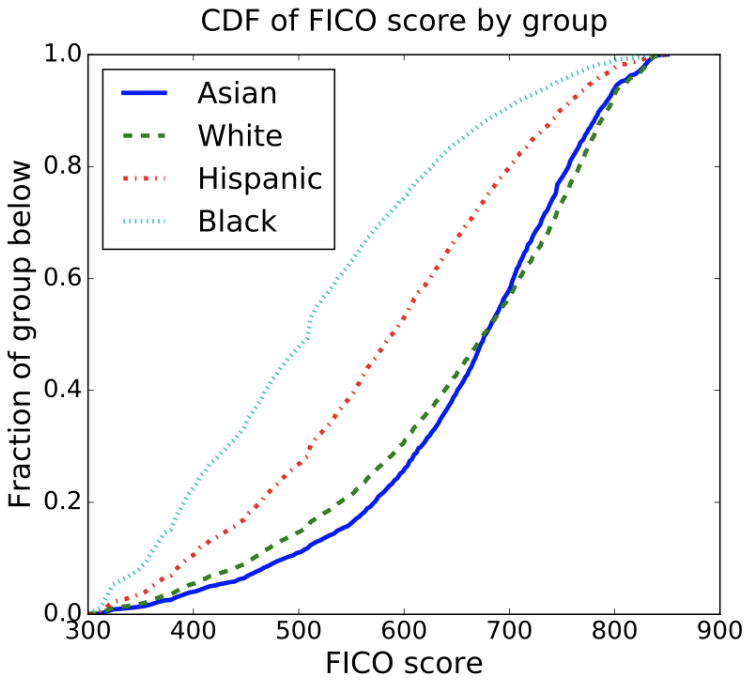
Any distribution induced by an aggregrate of agents following standard microfoundations is necessarily discontinuous.

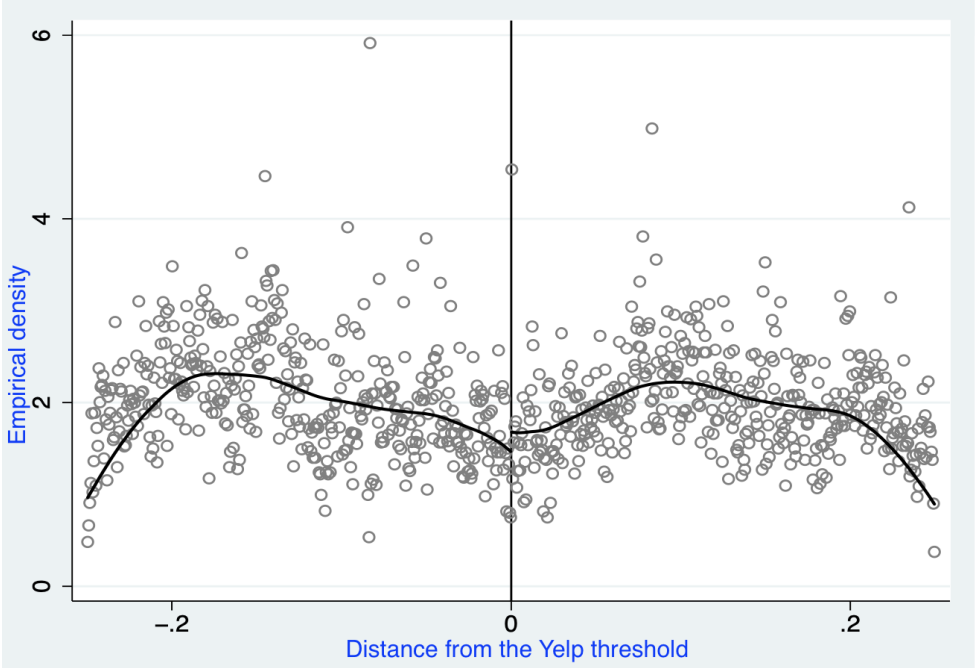
Result I: SM Cannot Capture Observed Distributions

Proposition (Informal)

Any distribution induced by an aggregate of agents following standard microfoundations is necessarily discontinuous.

But observed distributions often do not exhibit significant discontinuities:





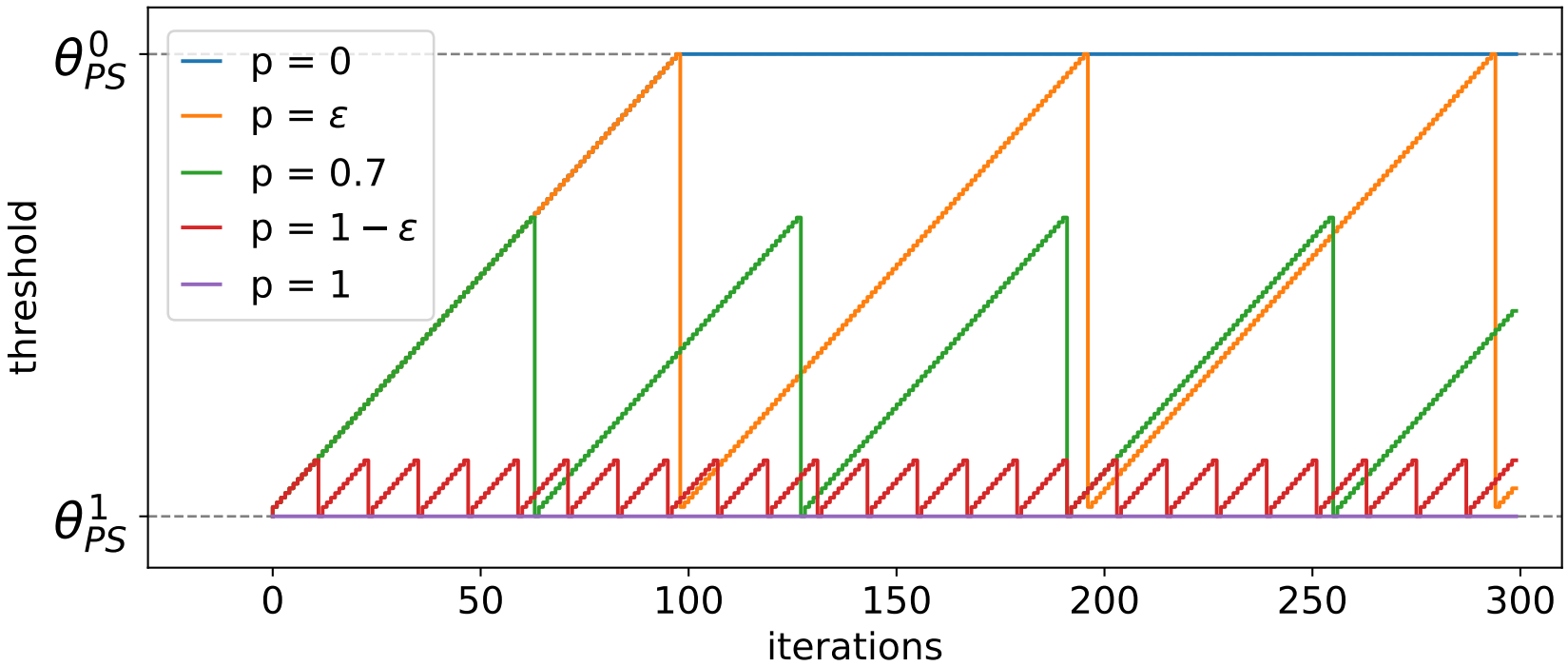
Result II: Retraining Methods are Non-Robust under SM

Common algorithmic approach: repeatedly retrain the classifier weights to be optimal on the data distribution induced by the previous classifier.

Result II: Retraining Methods are Non-Robust under SM

Common algorithmic approach: repeatedly retrain the classifier weights to be optimal on the data distribution induced by the previous classifier.

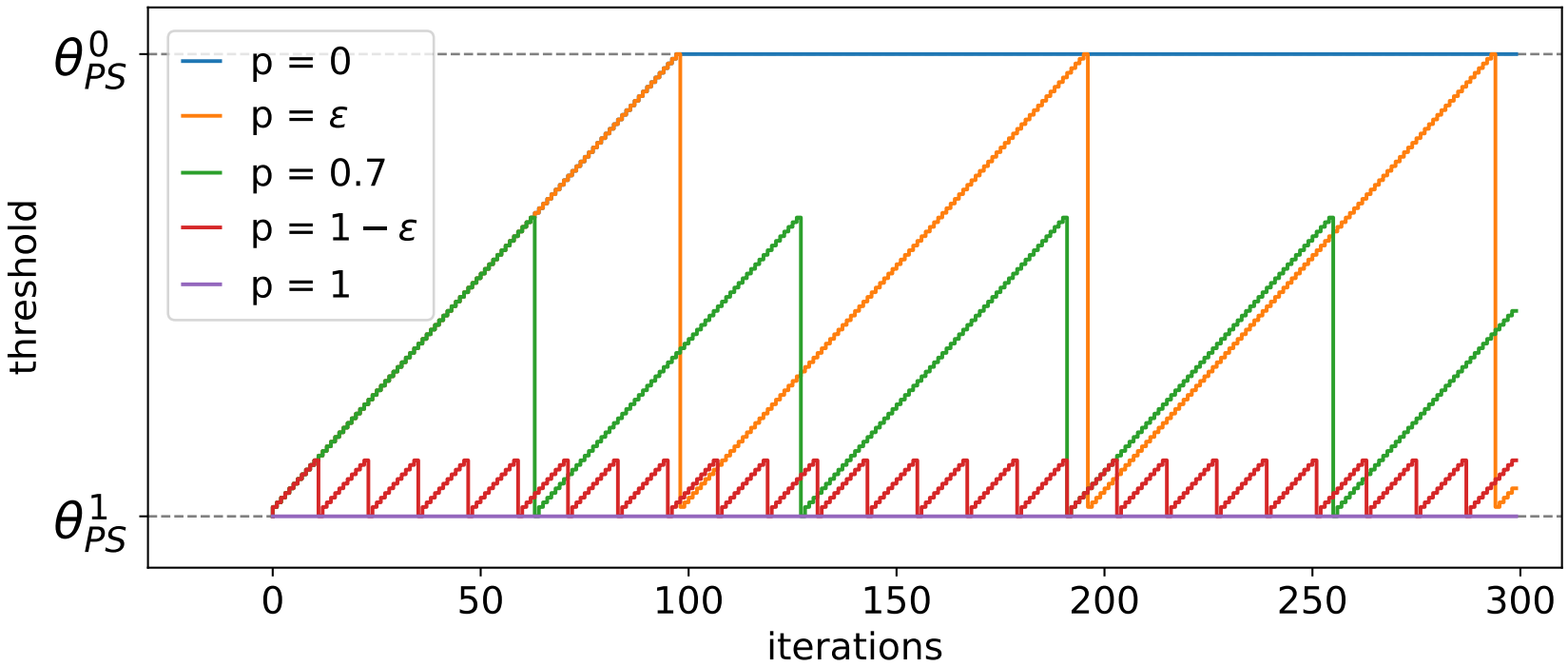
Retraining on mixed populations with p fraction of non-strategic agents:



Result II: Retraining Methods are Non-Robust under SM

Common algorithmic approach: repeatedly retrain the classifier weights to be optimal on the data distribution induced by the previous classifier.

Retraining on mixed populations with p fraction of non-strategic agents:



Repeated Retraining breaks down with ϵ fraction of non-strategic agents

Result III: SM Maximizes Social Burden at Optimality

Alternate algorithmic approach: Use anticipated distribution shifts from standard microfoundations to compute the "optimal point":

$$\operatorname{argmin}_{\theta \in \Theta} \mathbb{E}_{(x,y)\sim\tilde{\mathcal{D}}(\theta)} \mathbb{1}\left\{y \neq f_\theta(x)\right\}.$$

Result III: SM Maximizes Social Burden at Optimality

Alternate algorithmic approach: Use anticipated distribution shifts from standard microfoundations to compute the "optimal point":

$$\operatorname{argmin}_{\theta \in \Theta} \mathbb{E}_{(x,y)\sim\tilde{\mathcal{D}}(\theta)} \mathbb{1}\left\{y \neq f_\theta(x)\right\}.$$

We show standard microfoundations lead to *extreme solutions*:

Proposition (Informal)

The "optimal points" induced by SM maximize negative externalities (i.e. social burden) within a large family of alternate models for agent behavior.

Selecting Alternative Microfoundations

Step 1: We describe two natural properties to guide this search:

1. *Aggregate smoothness:* aggregate distribution map must be smooth
   - Guarantees the robust existence of fixed points of retraining.
2. *Expenditure constraint*: agents expend no more on gaming than the utility of a positive outcome.
   - Helps capture realistic agent-level responses and limits social burden.

Step 2: Using these properties as a guide, we propose noisy response:

Definition (Informal)

Noisy response captures *imperfect agents* using ideas from *smoothed analysis*. The idea is to add random perturbations (in a careful way).

We show that noisy response satisfies a number of desirable properties.