CMSC 35401: The Interplay of Economics and ML (Winter 2024)

Performative Prediction: Strategic Learning from the Macro Lens

Instructor: Haifeng Xu

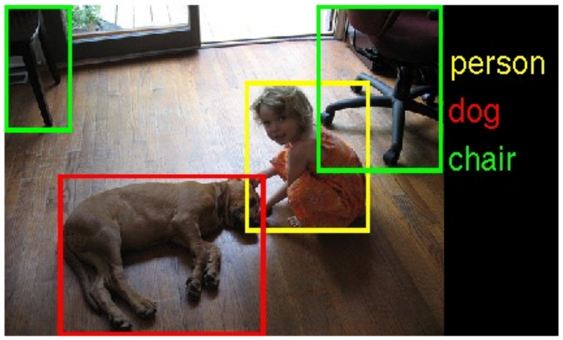
Slides partially adapted from a tutorial by Celestine Mendler-Dünner at NeurIPS'23

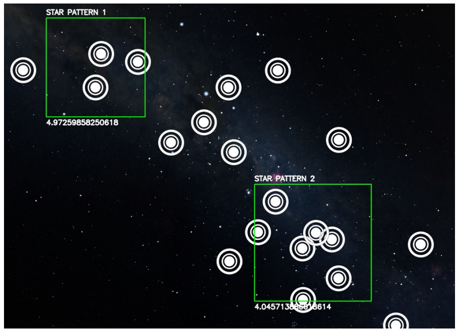
Outline

- The Motivation and Model
- From Prediction to Power

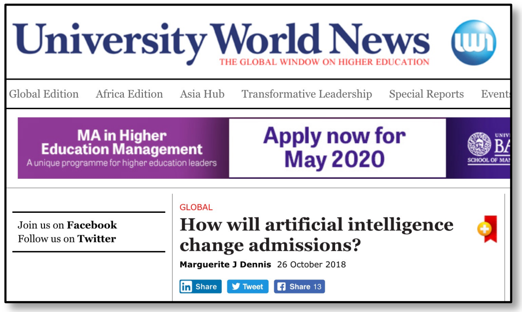
Learning has Varied Effects in Varied Contexts

- Learning in objective context is mostly descriptive



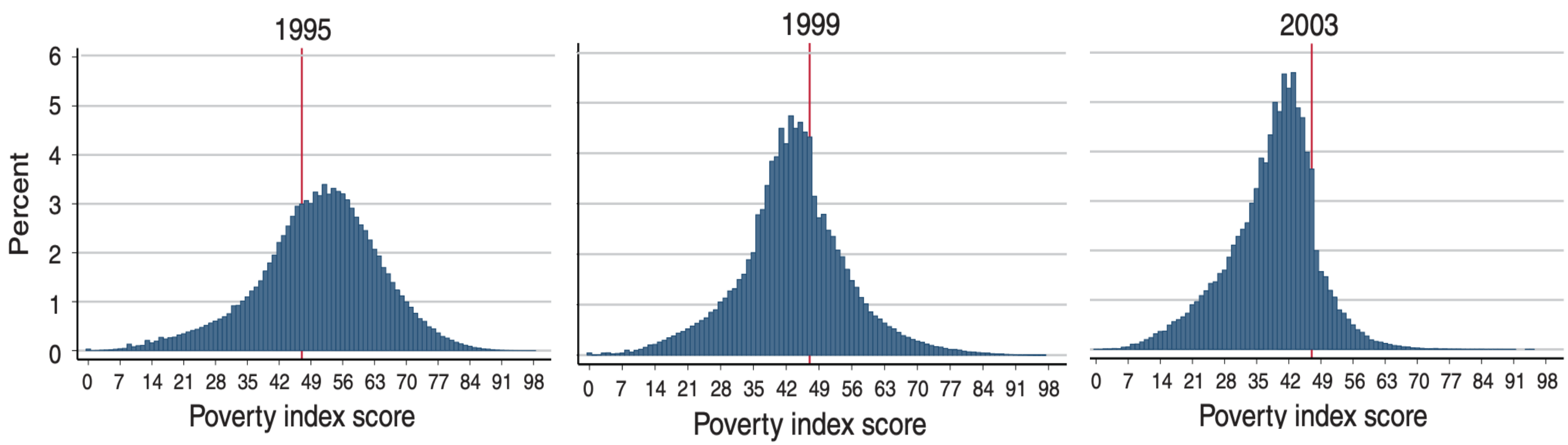


- Learning in economic/societal contexts is causative
  - It affects downstream audience's behaviors, decisions



Examples of Prediction in Societal Contexts

- Poverty index prediction, and people's response

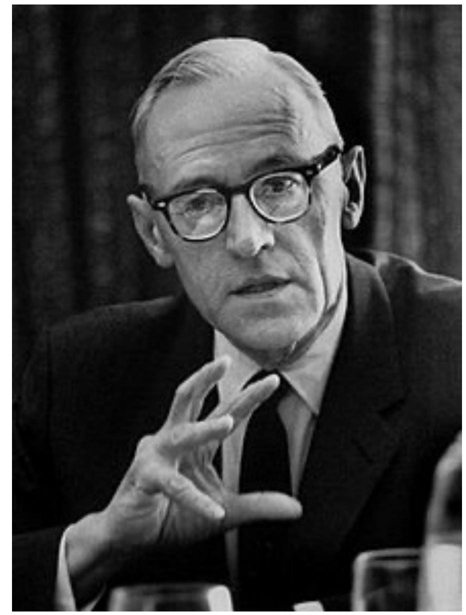


Source:
Camacho and Conover
AMERICAN ECONOMIC JOURNAL: ECONOMIC POLICY

Examples of Prediction in Societal Contexts

"Forecasts that can affect the predicted events … are one of the most difficult and central problems that the theory of prediction has to offer"

"Prediction cannot be caried out using economic theory and statistic alone"



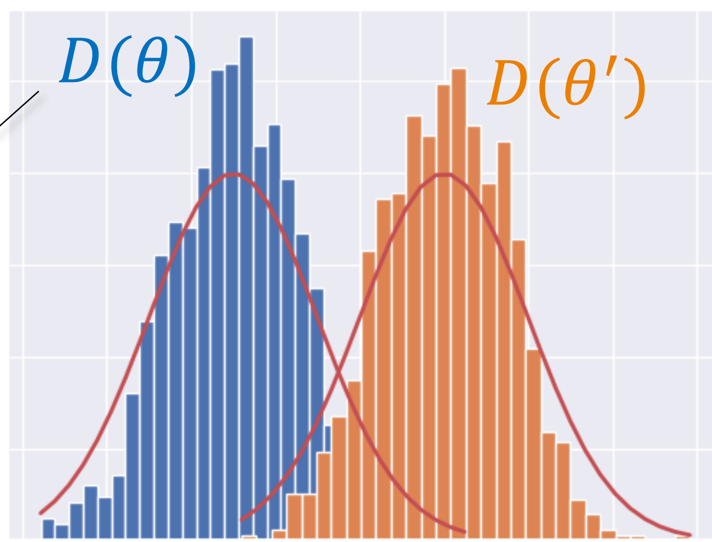
Oskar Morgenstern, 1928
(founder of game theory)

Performative Prediction [Perdomo et al., ICML'20]

<u>In essence</u>

- Avoids micro-level agent incentive modeling
- Instead, model entire population's responses as macro-level distribution shift

Data distribution under model A_θ

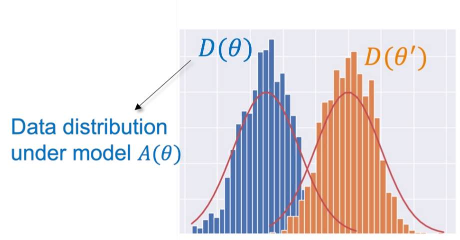


Performative Prediction [Perdomo et al., ICML'20]

Formal Model:

- Want to train model $A_\theta(x): X \to Y$, with parameter θ
  - E.g., $A_\theta(x) = \mathbb{I}(\theta \cdot x \geq 0)$ could be the class of linear classifiers
- Compute expected loss
  - E.g., $\text{loss}(A_\theta(x), y) = \mathbb{I}(A_\theta(x) \neq y)$

$$\text{Loss} = \mathbb{E}_{(x,y)\sim D(\theta)}[\text{loss}(A_\theta(x), y)]$$



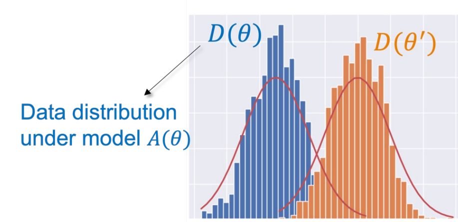
Performative Prediction [Perdomo et al., ICML'20]

<u>Formal Model:</u>

- Want to train model $A_\theta(x)\colon X \to Y$, with parameter θ
  - E.g., $A_\theta(x) = \mathbb{I}(\theta \cdot x \geq 0)$ could be the class of linear classifiers
- Compute expected loss
  - E.g., $\mathrm{loss}(A_\theta(x), y) = \mathbb{I}(A_\theta(x) \neq y)$

$$\text{Loss} = \mathbb{E}_{(x,y)\sim D\cancel{(\theta)}}[\mathrm{loss}(A_\theta(x), y)]$$

Q: How is this different from standard machine learning?



Performative Prediction [Perdomo et al., ICML'20]

Formal Model:

- ➢ Want to train model $A_\theta(x): X \to Y$, with parameter θ
  - E.g., $A_\theta(x) = \mathbb{I}(\theta \cdot x \geq 0)$ could be the class of linear classifiers
- ➢ Compute expected loss
  - E.g., $\text{loss}(A_\theta(x), y) = \mathbb{I}(A_\theta(x) \neq y)$

$$\text{Loss} = \mathbb{E}_{(x,y)\sim D(\theta)}[\text{loss}(A_\theta(x), y)]$$

This distribution dependence on model θ is called performativity.

- ✓ Hence model has causal influence on target distribution
- ✓ Strategic behaviors, self-fulfilling prophecy are examples, but this is a general and macro-level model at population level
- ✓ A special example of distribution shift and causality

Performative Prediction [Perdomo et al., ICML'20]

<u>Formal Model:</u>

➢ Want to train model $A_\theta(x): X \to Y$, with parameter θ
- E.g., $A_\theta(x) = \mathbb{I}(\theta \cdot x \geq 0)$ could be the class of linear classifiers

➢ Compute expected loss
- E.g., $\text{loss}(A_\theta(x), y) = \mathbb{I}(A_\theta(x) \neq y)$

$$\text{Loss} = \mathbb{E}_{(x,y)\sim D(\theta)}[\text{loss}(A_\theta(x), y)]$$

Performativity is a known concept in Econ, finance and public policy


Investopedia

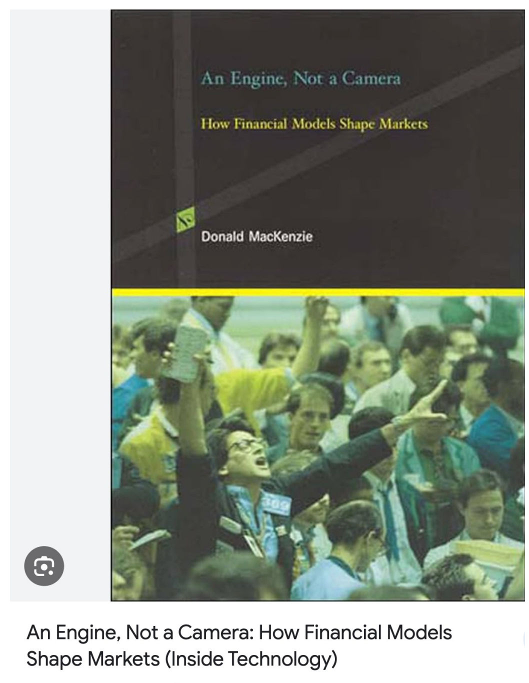
ECONOMY › ECONOMICS

Performativity: What It Is, How It Works, Evidence

By ADAM HAYES Updated October 02, 2023
Reviewed by SOMER ANDERSON
Fact checked by VIKKI VELASQUEZ

What Is Performativity in Economics?

The performativity thesis suggests that economic or financial models, rather than objectively measuring some aspect of reality, instead help shape that aspect of reality to the form that the model describes. That is, performativity describes the notion that economic theory does not merely describe the world as it appears but has the capacity to act upon the world and in doing so *make* the economy—and the agents within it—appear more like the theory itself.



Performative Prediction [Perdomo et al., ICML'20]

Formal Model:

- Want to train model $A_\theta(x): X \to Y$, with parameter θ
  - E.g., $A_\theta(x) = \mathbb{I}(\theta \cdot x \geq 0)$ could be the class of linear classifiers
- Compute expected loss
  - E.g., $\text{loss}(A_\theta(x), y) = \mathbb{I}(A_\theta(x) \neq y)$

$$\text{Loss} = \mathbb{E}_{(x,y)\sim D(\theta)}[\text{loss}(A_\theta(x), y)]$$

Performativity is a known concept in Econ, finance and public policy

What Does it Mean to Say that Economics is Performative?

Michel Callon

(July 2006)

Forthcoming in: D. MacKenzie, F. Muniesa and L. Siu (Eds.), *Do Economists Make Markets? On the Performativity of Economics*, Princeton University Press.

Performative Prediction [Perdomo et al., ICML'20]

Formal Model:

- Want to train model $A_\theta(x): X \to Y$, with parameter θ
  - E.g., $A_\theta(x) = \mathbb{I}(\theta \cdot x \geq 0)$ could be the class of linear classifiers
- Compute expected loss
  - E.g., $\text{loss}(A_\theta(x), y) = \mathbb{I}(A_\theta(x) \neq y)$

$$\theta^* = \text{argmax}_\theta \quad \mathbb{E}_{(x,y)\sim D(\theta)}[\text{loss}(A_\theta(x), y)]$$

- Find θ^* that minimizes loss

Key Challenges for Finding the Optimal Model

Challenge 1: Complex loss function due to distribution shift

- Convexity is crucial for optimization, but unclear how to capture "convex" properties of $D(\theta)$: $\Theta \rightarrow \text{Distributions}$

$$\theta^* = \text{argmax}_\theta \quad \mathbb{E}_{(x,y)\sim D(\theta)}[\text{loss}(A_\theta(x), y)]$$

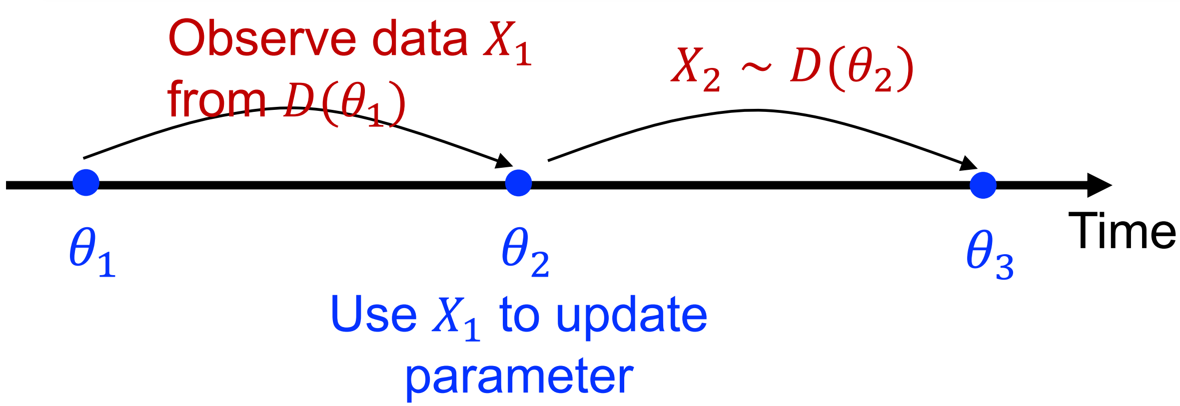
Key Challenges for Finding the Optimal Model

Challenge 1: Complex loss function due to distribution shift

- Convexity is crucial for optimization, but unclear how to capture "convex" properties of $D(\theta)\colon \Theta \to \mathrm{Distributions}$

$$\theta^* = \mathrm{argmax}_\theta \quad \mathbb{E}_{(x,y)\sim D(\theta)}[\mathrm{loss}(A_\theta(x), y)]$$

Challenge 2: delayed feedback, mis-matched data and training objective



This is called "re-training", used widely by many leading RSs

Re-training

Repeat the following for $t = 1,2,\cdots$

- Deploy model A_{θ_t}
- Observe data set X_t drawn from population distribution $D(\theta_t)$
- Update parameter to θ_{t+1} by minimizing empirical risk over X_t

$$\theta_{t+1} = \mathrm{argmax}_\theta \sum_{(x,y)\in X_t} [\mathrm{loss}(A_\theta(x), y)]$$

Re-training

Repeat the following for $t = 1,2,\cdots$

- Deploy model A_{θ_t}
- Observe data set X_t drawn from population distribution $D(\theta_t)$
- Update parameter to θ_{t+1} by minimizing empirical risk over X_t

$$\theta_{t+1} = \text{argmax}_\theta \sum_{(x,y)\in X_t} [\text{loss}(A_\theta(x), y)] \qquad (1)$$

Compare with original (most desirable) optimization:

$$\theta^* = \text{argmax}_\theta \mathbb{E}_{(x,y)\sim D(\theta)}[\text{loss}(A_\theta(x), y)] \qquad (2)$$

- (1) does not account for distribution shift.
  - Why? Besides recent samples X_t, we know nothing about $D(\theta_{t+1})$
  - This is the mis-match between data and objective

Re-training

Repeat the following for $t = 1,2,\cdots$

- Deploy model A_{θ_t}
- Observe data set X_t drawn from population distribution $D(\theta_t)$
- Update parameter to θ_{t+1} by minimizing empirical risk over X_t

$$\theta_{t+1} = \mathrm{argmax}_{\theta} \sum_{(x,y)\in X_t} [\mathrm{loss}(A_\theta(x), y)]$$

Mis-match between $X_t \sim D(\theta_t)$ and θ_{t+1} inspires another training algorithm – Gradient Descent

$$\theta_{t+1} = \theta_t - \gamma \frac{\partial \sum_{(x,y)\in X_t} [\mathrm{loss}(A_\theta(x),y)]}{\partial \theta}$$

Why? We already know $X_t \sim D(\theta_t)$ and θ_{t+1} are mis-matched, so we do not want θ_{t+1} to be too different from θ_t

Under Assumptions, Retraining Converges

We say distribution mapping is α-sensitive if for all θ, θ',

$$\mathrm{Wasserstein}\big(\mathrm{D}(\theta), \mathrm{D}(\theta')\big) \leq \alpha \big|\big|\theta - \theta'\big|\big|_2$$

That is, parameter change does not lead to dramatic distribution shift

Theorem [Perdomo et al., ICML'20]: If the loss function is γ-strongly convex and β-smooth in data, and $D(\theta)$ is not too sensitive ($\alpha < \gamma/\beta$), then retraining converges to a stable point at a linear rate.

A point $\bar{\theta}$ is stable if

$$\bar{\theta} = \mathrm{argmax}_\theta \ \mathbb{E}_{(x,y)\sim D(\bar{\theta})}[\mathrm{loss}(A_\theta(x), y)]$$

Fix distribution

$$\theta^* = \mathrm{argmax}_\theta \ \mathbb{E}_{(x,y)\sim D(\theta)}[\mathrm{loss}(A_\theta(x), y)]$$

Recall optimal model

Account for distribution shift

Under Assumptions, Retraining Converges

We say distribution mapping is α-sensitive if for all θ, θ',

$$\text{Wasserstein}\big(\mathrm{D}(\theta), \mathrm{D}(\theta')\big) \leq \alpha\big||\theta - \theta'|\big|_2$$

That is, parameter change does not lead to dramatic distribution shift

Theorem [Perdomo et al., ICML'20]: If the loss function is γ-strongly convex and β-smooth in data, and $D(\theta)$ is not too sensitive ($\alpha < \gamma/\beta$), then retraining converges to a stable point at a linear rate.

A point $\bar{\theta}$ is stable if

$$\bar{\theta} = \text{argmax}_\theta \ \mathbb{E}_{(x,y)\sim D(\bar{\theta})}[\text{loss}(A_\theta(x), y)]$$

A Nash equilibrium

$$\theta^* = \text{argmax}_\theta \ \mathbb{E}_{(x,y)\sim D(\theta)}[\text{loss}(A_\theta(x), y)]$$

Recall optimal model

The Stackelberg Equ.

Under Assumptions, Retraining Converges

We say distribution mapping is α-sensitive if for all θ, θ',

$$\mathrm{Wasserstein}\big(\mathrm{D}(\theta), \mathrm{D}(\theta')\big) \leq \alpha \big||\theta - \theta'|\big|_2$$

That is, parameter change does not lead to dramatic distribution shift

Theorem [Perdomo et al., ICML'20]: If the loss function is γ-strongly convex and β-smooth in data, and $D(\theta)$ is not too sensitive ($\alpha < \gamma/\beta$), then retraining converges to a stable point at a linear rate.

A point $\bar{\theta}$ is stable if

$$\bar{\theta} = \mathrm{argmax}_{\theta}\ \mathbb{E}_{(x,y)\sim D(\bar{\theta})}[\mathrm{loss}(A_\theta(x), y)]$$

$$\theta^* = \mathrm{argmax}_{\theta}\ \mathbb{E}_{(x,y)\sim D(\theta)}[\mathrm{loss}(A_\theta(x), y)]$$

Recall the performatively-optimal model

By HW2, Problem 2(4), θ^* is always better than any stable point $\bar{\theta}$!

Under Assumptions, Retraining Converges

We say distribution mapping is α-sensitive if for all θ, θ',

$$\text{Wasserstein}\big(\mathrm{D}(\theta), \mathrm{D}(\theta')\big) \leq \alpha \big|\big|\theta - \theta'\big|\big|_2$$

That is, parameter change does not lead to dramatic distribution shift

Theorem [Perdomo et al., ICML'20]: If the loss function is γ-strongly convex and β-smooth in data, and $D(\theta)$ is not too sensitive ($\alpha < \gamma/\beta$), then retraining converges to a stable point at a linear rate.

Question 1: how much better can performatively-optimal θ^* be than a stable point $\bar{\theta}$?

Ans: can be much better (easy to find examples)

Under Assumptions, Retraining Converges

We say distribution mapping is α-sensitive if for all θ, θ',

$$\mathrm{Wasserstein}\big(\mathrm{D}(\theta), \mathrm{D}(\theta')\big) \leq \alpha \big|\big|\theta - \theta'\big|\big|_2$$

That is, parameter change does not lead to dramatic distribution shift

Theorem [Perdomo et al., ICML'20]: If the loss function is γ-strongly convex and β-smooth in data, and $D(\theta)$ is not too sensitive ($\alpha < \gamma/\beta$), then retraining converges to a stable point at a linear rate.

Question 2: how to get to the performatively-optimal θ^* then?

Ans: can be achieved by (very tailored) algorithms that directly optimizes the true "performative loss"

- Best known convergence speed is $T^{1/d}$ which is not ideal
  [Jagadeesan et al. ICML'22]

Under Assumptions, Retraining Converges

We say distribution mapping is α-sensitive if for all θ, θ',

$$\mathrm{Wasserstein}\big(\mathrm{D}(\theta), \mathrm{D}(\theta')\big) \le \alpha\left|\left|\theta - \theta'\right|\right|_2$$

That is, parameter change does not lead to dramatic distribution shift

Theorem [Perdomo et al., ICML'20]: If the loss function is γ-strongly convex and β-smooth in data, and $D(\theta)$ is not too sensitive ($\alpha < \gamma/\beta$), then retraining converges to a stable point at a linear rate.

Remarks.

- Proof idea is to show the re-training procedure is a contracting mapping, which always reduces $|\theta_{t+1} - \theta_t|$
- A special case is when $\alpha = 0$, which is the standard ML problem
- Gradient descent can be similarly shown to work with similar guarantee

Main Open Computational Problems

Problem 1: $\alpha < \gamma/\beta$ is a very strong assumption – how to achieve convergence under weaker assumptions?

Theorem [Perdomo et al., ICML'20]: If the loss function is γ-strongly convex and β-smooth in data, and $D(\theta)$ is not too sensitive ($\alpha < \gamma/\beta$), then retraining converges to a stable point at a linear rate.

Main Open Computational Problems

Problem 1: $\alpha < \gamma/\beta$ is a very strong assumption – how to achieve convergence under weaker assumptions?

Problem 2: how to achieve fast convergence to performatively optimal model θ^*, under realistic conditions (e.g., sample access to data, weaker loss function assumptions, etc.)

Problem 3: achieve faster algorithms for specific application domains by leveraging its structures.

- E.g., performative foundation model training, which affects downstream users' fine-tuning

Outline

- The Motivation and Model
- From Prediction to Power

Prediction as an Engine not a Camera

- Loss of performative prediction captures two aspects:

$$\text{Loss}\big(\theta, D(\theta)\big) = \mathbb{E}_{(x,y)\sim D(\theta)}[\text{loss}(A_\theta(x), y)]$$

Prediction as an Engine not a Camera

- Loss of performative prediction captures two aspects:

$$\text{Loss}(\theta, D(\theta)) \;=\; \underbrace{\text{Loss}\left(\theta, D(\bar{\theta})\right)}_{\text{Loss from optimizing given data } D(\bar{\theta})} + \underbrace{\left[\text{Loss}(\theta, D(\theta)) - \text{Loss}\left(\theta, D(\bar{\theta})\right)\right]}_{\text{Loss from steering } D(\bar{\theta}) \text{ to desirable population } D(\theta)}$$

Steering happens quite often in e-commerce (leads to anti-trust concerns)

FTC vs Amazon
"…shoppers consequently face less relevant search results and are steered toward more expensive products. Amazon deliberately steers shoppers away from offers that are not featured in the Buy Box"

Prediction as an Engine not a Camera

- Loss of performative prediction captures two aspects:

$$\text{Loss}\big(\theta, D(\theta)\big) = \underbrace{\text{Loss}\left(\theta, D(\bar{\theta})\right)}_{\text{Loss from optimizing given data } D(\bar{\theta})} + \underbrace{\left[\text{Loss}\big(\theta, D(\theta)\big) - \text{Loss}\left(\theta, D(\bar{\theta})\right)\right]}_{\text{Loss from steering (current) } D(\bar{\theta}) \text{ to induced (future) population } D(\theta)}$$

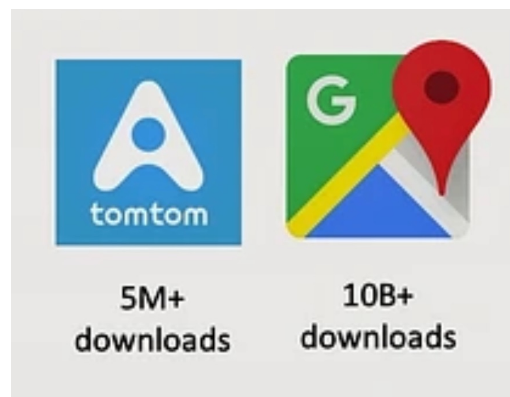
Steering happens quite often in e-commerce (leads to anti-trust concerns)

EU vs Google
"…The general court [of the EU] finds that, by favoring its own comparison shopping service on its general results pages …by means of ranking algorithms, Google departed from competition on the merits"

Performativity and Power

The ability to steer depends on power.

- The more market power you have, the more you can steer population behaviors
- Hence, more powerful/dominating firms have more steering power, and faster convergence to performative optimal (which may be bad), and also more concerns of anti-trust due to large deviation from current population



Thank You

Haifeng Xu

University of Chicago

haifengxu@uchicago.edu