Understanding Model Expressivity for Learning in Strategic Environments

Tinashe Handina and Eric Mazumdar

Caltech

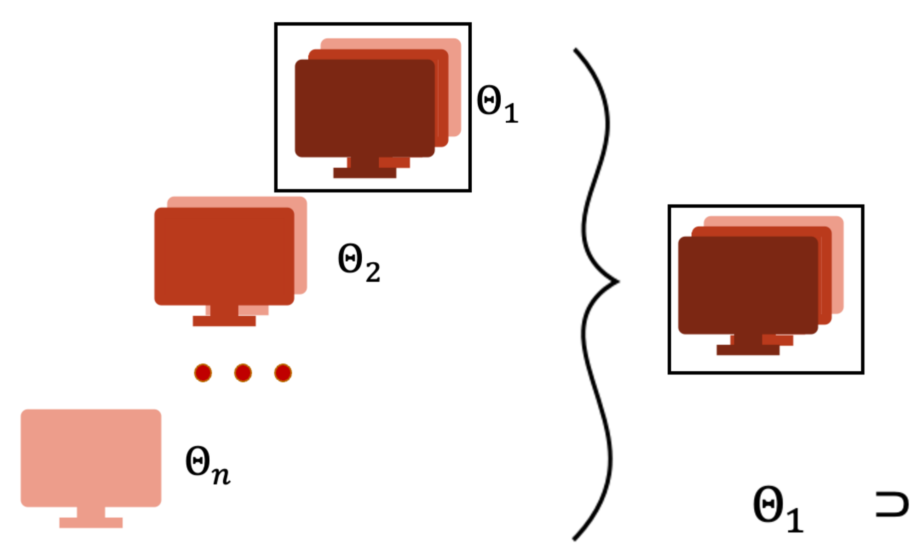
Question: How do strategic interactions affect the relationship between model class expressivity and equilibrium performance?

Results: Strategic interactions can yield a non-trivial relationship between model class expressivity and model performance at equilibrium

Implications: The choice of model class expressivity for models deployed in strategic environments should be treated as a strategic action

Model

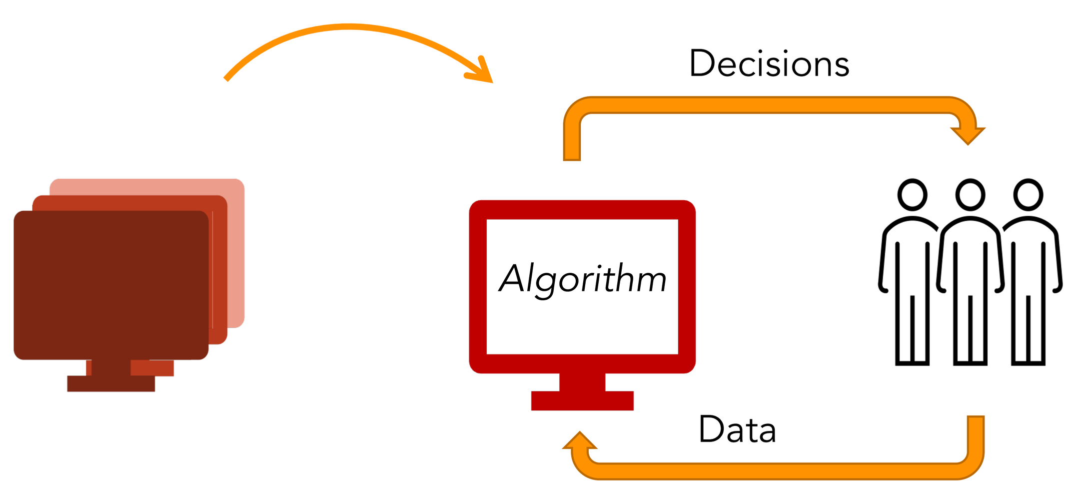
Phase 1: Design phase



Learner **selects** a **model class** from the menu of model classes.

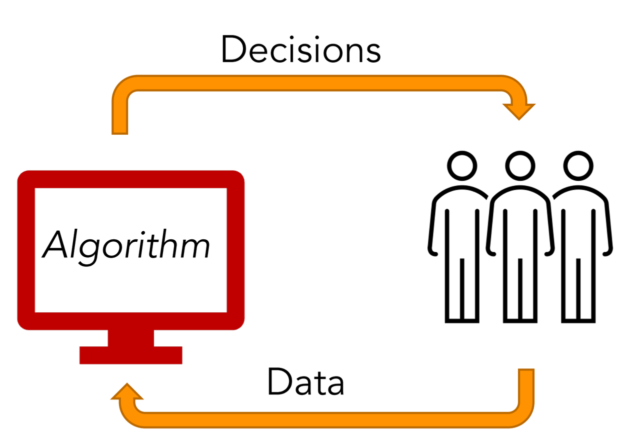
$\Theta_1 \supset \Theta_2 \supset \cdots \supset \Theta_n$

Phase 2: Training and deployment



Learner **trains** models, **deploys** into the **environment,** and retrains given environmental response to model.

Phase 3: Equilibrium



Learner and agents **settle into an** equilibrium.

Model class selection in games

Should the learner always select the most expressive model class if they want the best equilibrium outcome?

YES: If the environment is stationary or the environment and learner are in a Stackelberg game where the learner leads

Stationary environments: The environment has only one action. (i.e., $\mathcal{E} = \{e\}$). Equilibrium is (θ^*, e^*) such that

$$\theta^* = argmin_{\theta \in \Theta_i} f_{learner}(\theta, e)$$

Stackelberg environments – learner leads: [1]
Equilibrium is a joint strategy (θ^*, e^*) such that

$$\theta^* = argmin_{\theta \in \Theta_i} f_{learner}(\theta, BR_e(\theta))$$

and $e^* = BR_e(\theta^*) = argmin_{e \in \mathcal{E}} f_{environment}(\theta^*, e)$

NO: If the environment and learner are in a Stackelberg game where the learner follows or in a Nash game:

Stackelberg environments – learner follows: [2]
Equilibrium is a joint strategy (θ^*, e^*) such that

$$e^* = argmin_{e \in \mathcal{E}} f_{environment}(BR_l(e), e)$$

and $\theta^* = BR_l(e^*) = argmin_{\theta \in \Theta_i} f_{learner}(\theta, e^*)$

General Nash environments: [3]
Equilibrium is a joint strategy (θ^*, e^*) such that

$$f_{environment}(\theta^*, e') \geq f_{environment}(\theta^*, e^*) \;\forall\, e' \in \mathcal{E}$$
$$f_{learner}(\theta', e^*) \geq f_{learner}(\theta^*, e^*) \;\forall\, \theta' \in \Theta_i$$

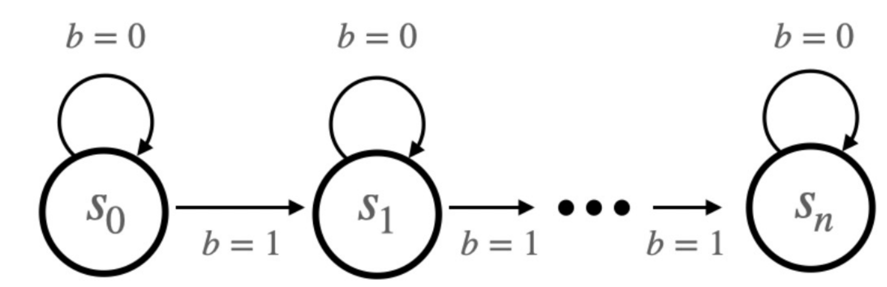
TAKEAWAY: Optimizing over more expressive model classes can lead to worse equilibrium outcomes when learning in strategic environments

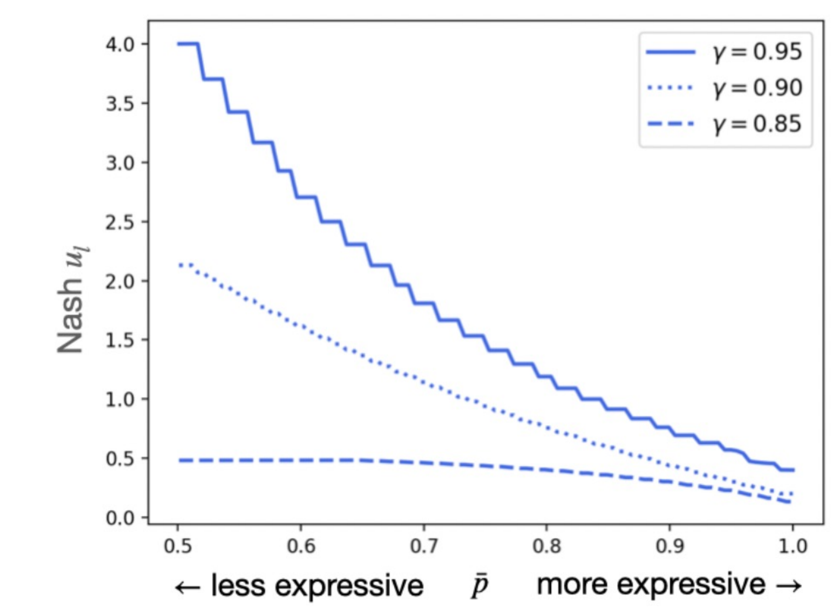
INFORMAL THEOREM: *For a two-player monotone game G satisfying some game regularity assumptions, if equilibrium (θ^*, e^*) in $\Theta \times \mathcal{E}$ is not Pareto optimal, then there exists a restriction of the learner's action set such that the restricted game G′ has a Nash equilibrium (θ', e') with:*

$$f_{learner}(\theta', e') < f_{learner}(\theta^*, e^*)$$

Examples

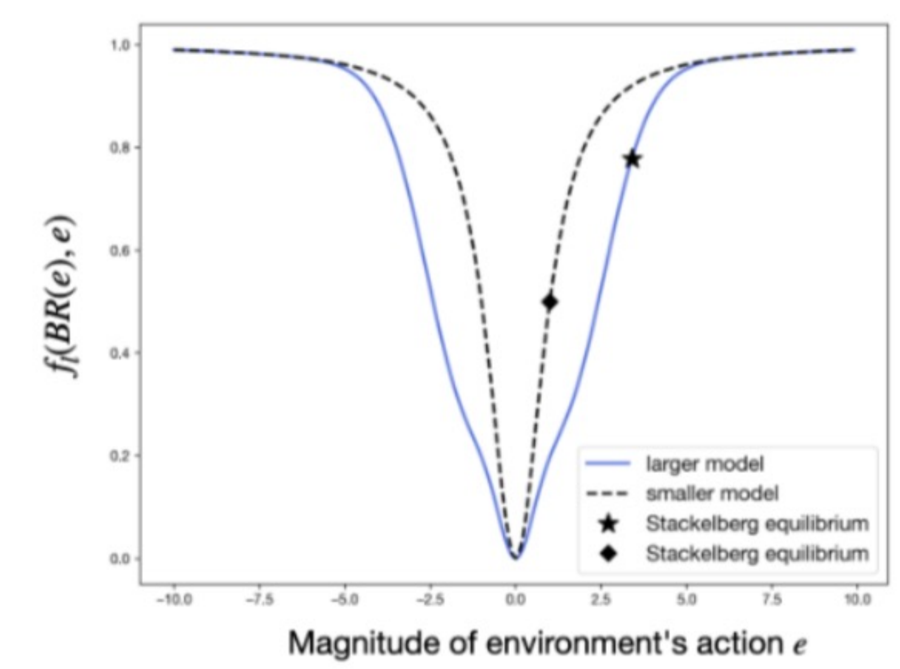
Multi-Agent Reinforcement Learning





The learner in this instance selects a policy class from $\{\Theta_k\}_{k=1}^N$ where $\Theta_k := 1 - p_k \leq \theta \leq p_k$ for $0.5 \leq p_k \leq 1$. We construct payoffs in a manner that results in decreasing performance with respect to expressivity of the model class.

Strategic Regression



A learner selects between two model classes:

$$\Theta_1 := \theta_1^{\top} x + \theta_2 \exp(-||x||^2) \quad \text{or} \quad \Theta_2 := \theta^{\top} x$$

Agents can collectively strategically manipulate their data by adding some deviation e. The payoff at the Stackelberg equilibrium is higher when the learner makes use of the smaller model class Θ_2 instead of Θ_1

[1] Moritz Hardt, Nimrod Megiddo, Christos Papadimitriou, and Mary Wootters. Strategic classification. In Innovations in Theoretical Computer Science, page 111–122, 2016
[2] M. Jagielski, A. Oprea, B. Biggio, C. Liu, C. Nita-Rotaru, and B. Li. Manipulating machine learning: Poisoning attacks and countermeasures for regression learning. In 2018 IEEE Symposium on Security and Privacy (SP). IEEE Computer Society, 2018
[3] Meena Jagadeesan, Michael Jordan, Jacob Steinhardt, and Nika Haghtalab. Improved bayes risk can yield reduced social welfare under competition. In Thirty-seventh Conference on Neural Information Processing Systems, 2023