Information Design and Price Mechanisms in Platforms for Cyber-Physical Systems with Learning Agents

UW ECE: Lillian Ratliff, Baosen Zhang; Stanford MS&E: Ramesh Johari
CPS: 1931718, 1931696 PI Emails: {ratliffl,zhangbao}@uw.edu, rjohari@stanford.edu

Introduction

A number of large-scale cyber-physical systems (CPS) today are effectively digitally-mediated markets. These CPS impose fundamental physical constraints: these include environmental uncertainty, spatial heterogeneity, temporal delays, and resource constraints. A technology-enabled platform digitally mediates transactions between buyers and sellers with the ability to shape the interaction of these agents.

Challenges:

Cyber and physical worlds and inherently meshed and digital platforms that support market-based services are entangled with physically constrained systems. To design optimal platforms, we must consider the following:

- Agent behavior can be quite complex
- They possess limited information and receive limited feedback
- They learn continuously over time to discover optimal strategies

Approaches:

We consider two levers available to a platform:

- Information feedback mechanisms
- Price-setting Mechanisms

We approach the problem from two viewpoints:

- From the agents, we characterize the collective behavior of dynamic learning agents
- The platform design policies for both information disclosure and pricing changes. We are especially interested in how these two complement each other.

Students and Postdocs:

- Tanner Fiez, Benjamin Chasnov, Yuanyuan Shi at UW
- Mohammad Rasouli at Stanford

Learning in Stackelberg Games

- Works on learning in commonly overlook the hierarchical decision-making structure.
- Studying Stackelberg games provide insights into optimization landscapes of zero-sum games.
- We show that deterministic gradient updates only converge to Stackelberg equlibria

Model

- Consider zero-sum games on continuous action spaces (X_1, X_2) and cost $(f, -f)$
- The leader (player 1) and the follower (player 2) solves the following problems:

$$\min_{x_1 \in X_1} \{f_1(x_1, x_2) \mid x_2 \in \arg\min_{y \in X_2} f_2(x_1, y)\}$$

$$\min_{x_2 \in X_2} f_2(x_1, x_2)$$

- We are interested in differential Stackelberg equilibria: the joint strategy $x^* = (x_1^*, x_2^*)$, where $Df_1(x_1^*) = 0, D_2 f_x(x_2^*) = 0, D^2 f_1(x^*) > 0, D_2^2 f_2(x^*) > 0$, where D_i is the partial derivative and D is the total derivative

Learning Dynamics

- Each player follows gradient dynamics

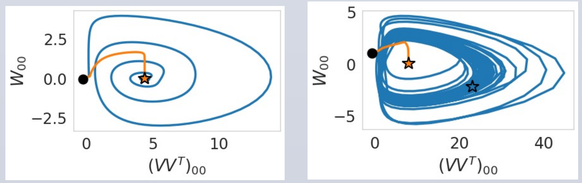
$$x_{i,k+1} = x_{i,k} - \gamma_{i,k} h_{\mathcal{S},i}(x_k)$$

- We assume the full information setting where the gradients can be computed exactly

Results

Under mild assumptions, deterministic updates converge to a Stackelberg equilibrium at the rate of $O(\epsilon^k)$, where ϵ is a constant and k is the number of iterations [1]

- This can be much faster than simultaneous updates when players moves at the same time



Learning covariance matrix. Orange is the Stackelberg setting (no oscillation) and blue is simultaneous gradient descents

[1] Fiez, Chasnov, and Ratliff. "Implicit learning dynamics in stackelberg games: Equilibria characterization, convergence analysis, and empirical study." In ICML, 2020

Learning in Cournot Games

- Cournot games are used to model many socio-economic systems where players learn and compete without the full information
- They are not no-regret games
- We show that policy gradient dynamics converge to Nash equilibria

Model

- Consider N players producing a homogenous good, each with action space $x_i \geq 0$
- The profit of player i is

$$\pi_i(x_1, \ldots, x_N) = x_i \cdot p\left(\sum_{j=1}^{N} x_j\right) - C_i(x_i)$$

where p is the market price function and C is a cost

- We are interested in Nash equilibria, a vector $x \geq 0$, where

$$\pi_i(x_i^*, \mathbf{x}_{-\mathbf{i}}^*) \geq \pi_i(\tilde{x}_i, \mathbf{x}_{-\mathbf{i}}^*), \quad \text{for all} \quad \tilde{x}_i$$

Learning Dynamics

- At each time step, player i draws an action in the form of $a_i \sim \pi_{\theta_i}(\cdot) = (\theta_i + X_i)^+$, where X_i is a random variable. E.g., X_i Gaussian gives the standard Gaussian policy gradient
- The mean θ_i is updated based only on the gradient of the expected return of player i

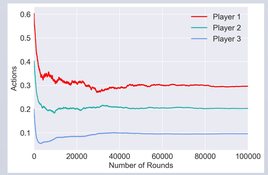
Results

The policy gradient updates converges exponentially quickly to a Nash equilibrium if [2]:

- The price function is linear
- There are only two players

We conjecture that this result holds for more general settings

3-player symmetric game



3-player asymmetric game

[2] Shi and Zhang. "Multi-agent reinforcement learning in cournot games." In IEEE Conference on Decision and Control (CDC), 2020

Adaptive Experimental Design

- Suppose there are two algorithms that you want to compare over time
- The industry practice is to randomly switch between the two and see how they do:

- Not a good idea because of interference: each one change the system states seen by the other

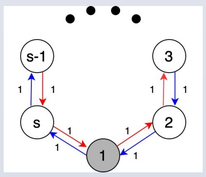
Model

- Consider two different Markov chains, indexed by $l = 1,2$ evolving on a common space S, defined by transition matrices

$$P(l) = (P(l, x, y) : x, y \in \mathcal{S}), l = 1, 2$$

- A policy is a sequence of random variables $A = (A_n, n \geq 0), A_n = \{1,2\}$ that determines which chain to run
- The stationary rewards are $\alpha(2), a(1)$
- Goal: design a policy and an estimator to estimate the treatment effect $\alpha = \alpha(2) - a(1)$

Example



Chain 1 is red, and chain 2 is blue. Rewards are only earned in state 1 for each chain. The reward distribution is Bernoulli$(q(l))$ for chain l

- A naïve sampling policy leads to an estimation variance that scales with s, the number of states
- A joint policy gives estimators with variance that does not grow with s

Results

- The idea is to leverage cooperative exploration: using one chain to drive the states where we want to sample using the other chain
- We use the maximum likelihood estimator (MLE)
- The policy can be designed using a convex optimization problem, and the MLE is both consistent and efficient [3]

[3] Glynn, Johari, and Rasouli. "Adaptive experimental design with temporal interference: A maximum likelihood approach." NeurIPS 2020