Policy Learning with Competing Agents

Roshni Sahoo[*] Stefan Wager[†]

[*] Department of Computer Science, Stanford University
[†] Graduate School of Business, Stanford University

Introduction

Decision makers often aim to learn a treatment assignment policy under a **capacity constraint** on the number of agents that they can treat. When agents can respond **strategically** to such policies, **competition** arises, complicating the estimation of the effect of the policy. *Examples: college admissions, job hiring.*

Treatment Assignment Model

Let $q \in (0,1)$. At each time step $t \in \{1, 2, 3...\}$, the decision maker assigns treatments to $1-q$ proportion of a target population based on observed covariates $\mathbf{x} \in \mathcal{X}$. At time t, the decision maker's policy is

$$\pi(\mathbf{x}, \epsilon; \boldsymbol{\beta}, s^t) = \mathbb{I}(\boldsymbol{\beta}^T\mathbf{x} + \epsilon > s^t),$$

where $\boldsymbol{\beta}, s^t$ are policy parameters at time-step t, and ϵ noise sampled from a mean-zero distribution G. At time-step $t+1$, an agent with type $\nu \sim F$ will report covariates $\mathbf{x}(\boldsymbol{\beta}, s^t, \nu)$ to the decision maker, reacting strategically to the policy deployed in time step t. At time-step $t+1$, the decision maker's policy is

$$\pi(\mathbf{x}, \epsilon; \boldsymbol{\beta}, s^{t+1}) = \mathbb{I}(\boldsymbol{\beta}^T\mathbf{x} + \epsilon > s^{t+1}),$$

where s^{t+1} is determined by the q-th quantile of marginal distribution of $\boldsymbol{\beta}^T\mathbf{x}_i(\boldsymbol{\beta}, s^t, \nu) + \epsilon$.

Policy Loss

The decision maker observes a loss $\ell(\pi, \nu)$ if they assign a treatment $\pi \in \{0,1\}$ to an agent with type ν. The **population policy loss** at time-step $t+1$ is $L(\boldsymbol{\beta}, s^t, s^{t+1})$, where

$$L(\boldsymbol{\beta}, s, r) = \mathbb{E}_{\nu\sim F, \epsilon\sim G}\left[\ell(\pi(\mathbf{x}(\boldsymbol{\beta}, s, \nu), \epsilon; \boldsymbol{\beta}, r), \nu)\right].$$

Agent Behavior Model

Following Frankel & Kartik (2019), we assume each agent has a **private** type $\nu = (\boldsymbol{\eta}, \boldsymbol{\gamma}) \sim F$.

$\boldsymbol{\eta} \in \mathcal{X}$ - raw covariates.

$\boldsymbol{\gamma} \in \mathcal{G}$ - ability to modify their covariates.

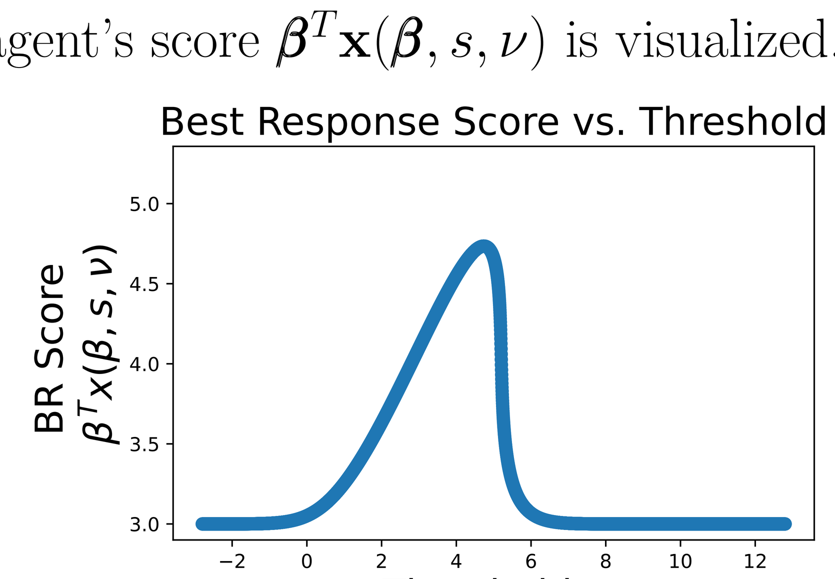
Agents **myopically** aim to maximize their utility with respect to a previous policy.

$$u(\mathbf{x}; \boldsymbol{\beta}, s, \nu) = -\underbrace{c_\nu(\mathbf{x} - \boldsymbol{\eta}; \boldsymbol{\gamma})}_{\text{cost of deviating from } \boldsymbol{\eta}} + \underbrace{\pi(\mathbf{x}, \epsilon; \boldsymbol{\beta}, s)}_{\text{reward}}.$$

The agent **best response** is defined as

$$\mathbf{x}(\boldsymbol{\beta}, s, \nu) = \underset{\mathbf{x}\in\mathcal{X}}{\operatorname{argmax}}\, \mathbb{E}_{\epsilon\sim G}\left[u(\mathbf{x}; \boldsymbol{\beta}, s, \nu)\right].$$

The agent's score $\boldsymbol{\beta}^T\mathbf{x}(\boldsymbol{\beta}, s, \nu)$ is visualized.



Equilibrium Policy Loss

At an **equilibrium** induced by a fixed $\boldsymbol{\beta}$, the level of competition is **fixed over time**. Let $s(\boldsymbol{\beta})$ be the equilibrium threshold induced by $\boldsymbol{\beta}$. If $s^t = s(\boldsymbol{\beta})$, then we have that

$$s^{t+1}, s^{t+2} \cdots = s(\boldsymbol{\beta}).$$

The decision maker's **equilibrium policy loss** is given by $L_{\text{eq}}(\boldsymbol{\beta}) = L(\boldsymbol{\beta}, s(\boldsymbol{\beta}), s(\boldsymbol{\beta}))$.

Mean-Field Regime

We consider **mean-field regime** where there is an infinite number of agents. Let $P_{\boldsymbol{\beta},s}$ be the distribution over scores when agents best respond to $\boldsymbol{\beta}, s$, and let $q(P_{\boldsymbol{\beta},s})$ be its q-th quantile. The level of competition evolves via **deterministic fixed-point iteration**.

$$s^{t+1} = q(P_{\boldsymbol{\beta},s^t}) \quad t = 1, 2, \ldots$$

The mean-field equilibrium threshold $s(\boldsymbol{\beta})$ under a fixed $\boldsymbol{\beta}$ satisfies $s = q(P_{\boldsymbol{\beta},s})$.

Mean-Field Equilibrium Theorem

When the variance of the noise distribution G is sufficiently high, the mean-field equilibrium threshold **exists** and is **unique** and **varies smoothly** w.r.t. $\boldsymbol{\beta}$.

Implication: $L_{\text{eq}}(\boldsymbol{\beta})$ is differentiable! This enables learning optimal policies via **gradient descent**.

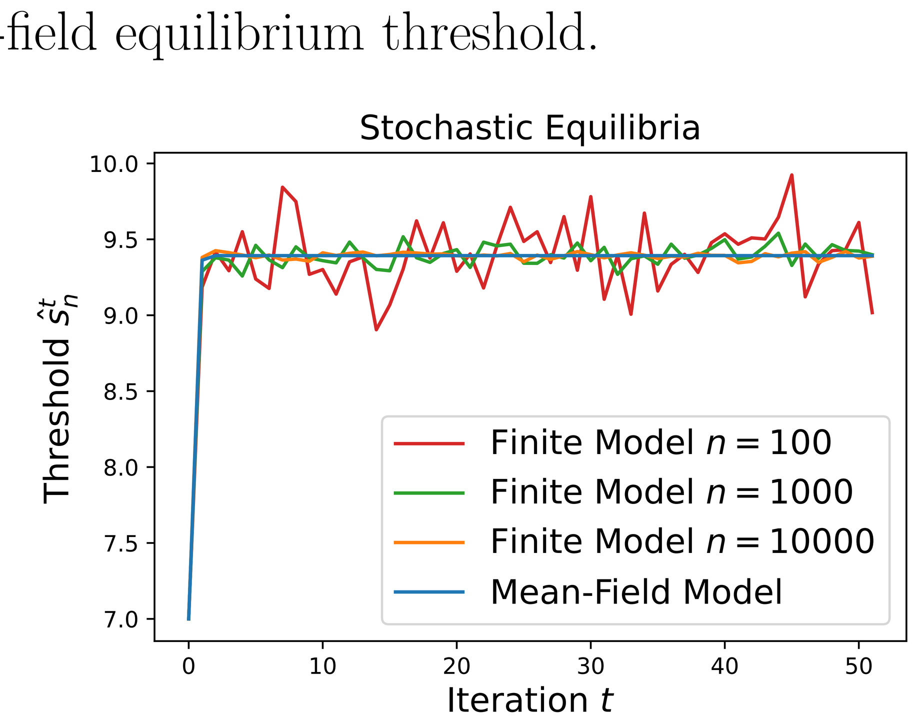
$$\underbrace{\frac{dL_{\text{eq}}}{d\boldsymbol{\beta}}}_{\text{policy effect}} = \underbrace{\frac{\partial L}{\partial \boldsymbol{\beta}}}_{\text{model effect}} + \underbrace{\left(\frac{\partial L}{\partial s} + \frac{\partial L}{\partial r}\right) \cdot \frac{\partial s}{\partial \boldsymbol{\beta}}}_{\text{equilibrium effect}}.$$

Finite-Sample Approximation

We consider the regime with a **finite** number of agents. Let $P^n_{\boldsymbol{\beta},s}$, $q(P^n_{\boldsymbol{\beta},s})$ be the *empirical* distribution over scores when agents best respond to $\boldsymbol{\beta}, s$ and its q-th quantile. The level of competition oscillates via **stochastic fixed-point iteration**.

$$\hat{s}_n^{t+1} = q(P^n_{\boldsymbol{\beta},\hat{s}_n^t}) \quad t = 1, 2 \ldots.$$

As n, t grow large, we expect iterates to approximate the mean-field equilibrium threshold.



Stochastic Equilibria Theorem

Let $\epsilon, \delta \in (0,1)$. Let $q(P_{\boldsymbol{\beta},s})$ be a contraction in s with Lipschitz constant κ. Let $k = \lceil \frac{\log(\frac{\epsilon}{2S})}{\log \kappa} \rceil$. For t such that $t \geq k$ and n such that

$$n \geq \frac{2}{\epsilon^2(1-\kappa)^2 D^2} \log(\frac{2k}{\delta}),$$

we have that

$$P(|\hat{s}_n^t - s(\boldsymbol{\beta})| \geq \epsilon) \leq \delta.$$

Learning Policies

Following Wager & Xu (2020), we can estimate $\frac{dL_{\text{eq}}}{d\boldsymbol{\beta}}$ in finite samples without disturbing the equilibrium via mean-zero perturbations.

Our Estimator

- ▷ For each agent i, we perturb $\boldsymbol{\beta}, s$ as follows

$$\boldsymbol{\beta}_i = \boldsymbol{\beta} + b_n \boldsymbol{\zeta}_i \quad \boldsymbol{\zeta} \in \{-1, 1\}^d$$
$$s_i = s + b_n \zeta_i \quad \zeta \in \{-1, 1\}^d.$$

- ▷ Observe $\boldsymbol{\ell}, \boldsymbol{\pi} \in \mathbb{R}^n$ - losses, treatment assignments.
- ▷ Run OLS from perturbations to ℓ, π to obstain regression coefficients $\hat{\Gamma}^n_{\ell,\boldsymbol{\beta}}$, $\hat{\Gamma}^n_{\ell,s,\ell,r}$, $\hat{\Gamma}^n_{\pi,s}$, $\hat{\Gamma}^n_{\pi,\boldsymbol{\beta}}$.
- ▷ Kernel density estimate $p^n_{\boldsymbol{\beta},s,b}(r)$.

$$\underbrace{\hat{\Gamma}^t_n}_{\frac{dL_{\text{eq}}}{d\boldsymbol{\beta}}} = \underbrace{\hat{\Gamma}^n_{\ell,\boldsymbol{\beta}}}_{\frac{\partial L}{\partial \boldsymbol{\beta}}} + \underbrace{\hat{\Gamma}^n_{\ell,s,\ell,r}}_{\frac{\partial L}{\partial s} + \frac{\partial L}{\partial r}} \cdot \underbrace{\left(\frac{1}{p^n_{\boldsymbol{\beta},\hat{s}^t_n,b_n}(\hat{s}^t_n) - \hat{\Gamma}^n_{\pi,s}} \cdot \hat{\Gamma}^n_{\pi,\boldsymbol{\beta}}\right)}_{\frac{\partial s}{\partial \boldsymbol{\beta}}}.$$

Consistency Theorem

Let $\{t_n\}$ be an increasing sequence $t_n \to \infty$. There exists a sequence $\{b_n\}$ such that $b_n \to 0$ so that

$$\hat{\Gamma}^{t_n}_n(\boldsymbol{\beta}) \xrightarrow{p} \frac{dL_{\text{eq}}}{d\boldsymbol{\beta}}(\boldsymbol{\beta}).$$

Simulation

We consider a population of agents including

Naturals - high $\boldsymbol{\eta}$, low $\boldsymbol{\gamma}$.

Gamers - low $\boldsymbol{\eta}$, high $\boldsymbol{\gamma}_1$.

The decision maker earns a loss of $-\boldsymbol{\eta}_1$ on agents they accept. The naive policy $\boldsymbol{\beta} = [1, 0]$ accepts many gamers and earns suboptimal policy loss. Our estimator enables learning the optimal policy!