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 \( t \in \{1, 2, 3, \ldots \} \), 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

\[
\pi(\mathbf{x}, \mathbf{c}, \beta, s(t)) = I(\beta^T \mathbf{x} + \epsilon > s(t)),
\]

where \( \beta, s \) are policy parameters at time-step \( t \), and \( \epsilon \) noise sampled from a mean-zero distribution \( G \). At time-step \( t + 1 \), an agent with type \( \nu \sim F \) will report covariates \( x(\mathbf{c}, \beta, s, \nu) \) 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

\[
\pi(\mathbf{x}, \mathbf{c}, \beta, s(t+1)) = I(\beta^T x + \epsilon > s(t+1)),
\]

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

Policy Loss

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

\[
L(\beta, s) = E_{\nu \sim F, x(\mathbf{c}, \beta, s, \nu)} \left[ \ell(\pi(\mathbf{x}, \mathbf{c}, \beta, s(t)), \nu) \right].
\]

Agent Behavior Model

Following Frankel & Kartik (2019), we assume each agent has a private type \( \nu \in \{0,1\} \). Let \( \mathbf{c} \) be observed and \( \gamma \) be a summary statistic. Agents myopically aim to maximize their utility with respect to a previous policy.

\[
\mathbf{n}(\mathbf{x}, \gamma, \nu) = -c_0(\mathbf{x} - \gamma, \nu) + \pi(\mathbf{x}, \mathbf{c}, \beta, s). \]

The agent best response is defined as

\[
x(\beta, s, \nu) = \arg\max_{\mathbf{x} \in \mathcal{X}} \mathbb{E}_{c,\beta,\gamma} \left[ n(\mathbf{x}, \mathbf{c}, \beta, s) \right].
\]

The agent’s score \( \beta^T x(\mathbf{c}, \beta, s, \nu) \) is visualized.

Equilibrium Policy Loss

At an equilibrium induced by a fixed \( \beta \), the level of competition is fixed over time. Let \( s(\beta) \) be the equilibrium threshold induced by \( \beta \). If \( s' = s(\beta) \), then we have that

\[
s(t+1) = q(P_{\beta}(s(t))) - \epsilon = s(t+1).
\]

The decision maker’s equilibrium policy loss is given by

\[
L_{eq}(\beta) = L(\beta, s(\beta), s(\beta)).
\]

Mean-Final Regime

We consider mean-field regime where there is an infinite number of agents. Let \( P_{\beta} \) be the distribution of agents with type \( s \) at time-step \( t \). Let \( \pi(\mathbf{x}, \mathbf{c}, \beta, s, \nu) \) be the best response. The mean-field equilibrium threshold \( s(\beta) \) under a fixed \( \beta \) satisfies \( s = q(P_{\beta}(s(t))). \)

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. \( \beta \).

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

\[
\frac{dL_{eq}}{d\beta} = \frac{d\ell}{d\beta} + \frac{d\ell}{d\pi} \cdot \frac{d\pi}{d\beta}
\]

Finite-Sample Approximation

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

\[
s_n(t+1) = q(P_{\beta, n}(s_n(t))) - \epsilon = s_n(t+1)
\]

As \( n, t \) grow large, we expect iterations to approximate the mean-field equilibrium threshold.

Learning Policies

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

Our Estimator

\( \triangleright \) For each agent \( i \), we perturb \( \beta, s \) as follows

\[
\beta_i = \beta + b_i \zeta, \quad \zeta \in (-1, 1)^d
\]

\( \triangleright \) Observe \( \ell, \pi \in \mathbb{R}^d \) - losses, treatment assignments

\( \triangleright \) Run OLS from perturbations to \( \ell, \pi \) to obtain regression coefficients \( \Gamma_{\ell, t}, \Gamma_{\pi, t} \)

\( \triangleright \) Kernel density estimate \( P_{\beta, n}(s(t)) \)

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

\[
\bigg( \ell_{eq, t_n}(\beta) \bigg) - \frac{d\ell_{eq, t_n}(\beta)}{d\beta}
\]

Simulation

We consider a population of agents including

Naturals - high \( \eta \), low \( \gamma \)
Gamers - low \( \eta \), high \( \gamma \)

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