A policy gradient algorithm to compute boundedly rational stationary mean field equilibria

Jayakumar Subramanian & Aditya Mahajan

ECE & CIM, McGill University and GERAD

Mean field games: Large number of small, anonymous agents with negligible individual impact

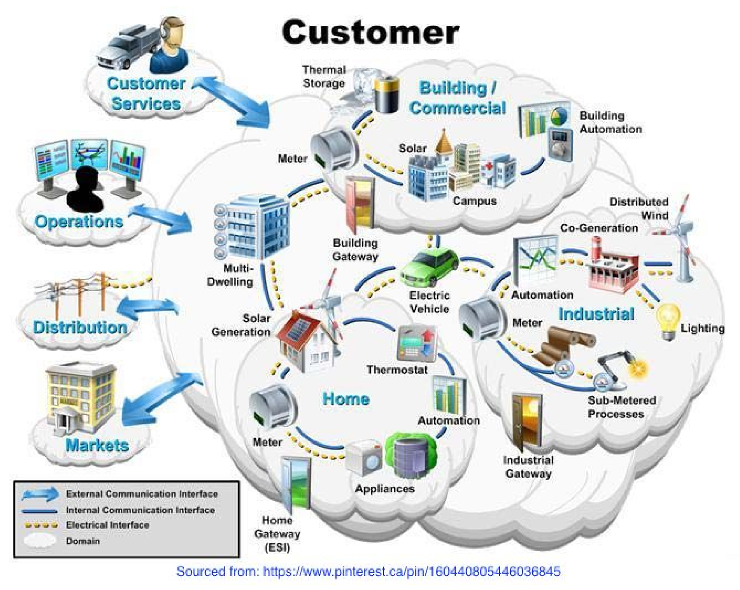

Fig. 1: Smart Grid - Demand Response

Fig. 2: Financial Markets

Solution concept

- Mean field equilibrium and its refinements are standard solution concepts in mean field games.

Our contribution

- Definition of an equilibrium for stationary mean field games based on bounded rationality.
- This equilibrium is a generalization of Nash equilibrium and mean field equilibrium.
- Development of a policy gradient based algorithm to predict this equilibrium.

Mean field game model

- Agent set: $N := \{1, \dots, n\}$ agents;
- State and action spaces for each agent: $\mathcal{X}, \mathcal{A}$ (finite and identical for all agents);
- Dynamical state evolution for each agent $i \in N$:

$$\mathbb{P}[X^i_{t+1} = x^i \mid \boldsymbol{X}_{1:t}, \boldsymbol{A}_{1:t}] = \mathbb{P}[X^i_{t+1} = x^i \mid X^i_t, A^i_t] =: P(x^i \mid X^i_t, A^i_t);$$

- Empirical mean field (or population average): $\xi_t \in \Delta(\mathcal{X})$, given by:

$$\xi_t(x) = \frac{1}{n}\sum_{i \in N} \mathbb{1}\{X^i_t = x\}, \quad \forall x \in \mathcal{X}.$$

- Per-step payoff to agent i: $u(X^i_t, A^i_t, \xi_t)$

Key assumptions

1. An agent uses only its current state to pick actions: $\mu^i_t : \mathcal{X} \to \Delta(\mathcal{A})$ and $A^i_t \sim \mu^i_t(X^i_t)$.
2. μ^i_t does not depend on time.
3. All agents play identical policies. Thus $\boldsymbol{\mu} = \{\mu, \mu, \dots, \mu\}$.
4. Each agent assumes that the population average is stationary. Thus agent i's assessment of its payoff is:

$$V^i_{\mu,\pi}(x) = \mathbb{E}_{A^i_t \sim \mu(X^i_t)}\Big[\sum_{t=0}^{\infty} \gamma^t u(X^i_t, A^i_t, \pi) \;\Big|\; X^i_0 = x\Big].$$

5. We consider parametrized policies μ_θ, where $\theta \in \Theta$ (a closed, convex space).

Stationary mean field equilibrium (SMFE)

SMFE is a pair of a belief $\pi \in \Delta(\mathcal{X})$ and a policy $\mu : \mathcal{X} \to \Delta(\mathcal{A})$, which satisfies the following two properties:

1. **Sequential Rationality**: For any other policy $\tilde{\mu} : \mathcal{X} \to \Delta(\mathcal{A})$,
   $V_{\mu,\pi}(x) \geqslant V_{\tilde{\mu},\pi}(x), \quad \forall x \in \mathcal{X}.$
2. **Consistency**: The belief π is stationary under policy μ, i.e., $\pi = \mathtt{StatDist}(\pi, \mu)$.

Gradient based SMFE (∇-SMFE)

∇-SMFE is a pair of belief $\pi \in \Delta(\mathcal{X})$ and a parametrized policy $\mu_\theta : \mathcal{X} \to \Delta(\mathcal{A})$, where $\theta \in \Theta$, which satisfies the following two properties:

1. **Gradient based sequential rationality**: Let $V_{\theta,\pi}$ be agents' payoff assessment. Then, $\nabla_\theta V_{\theta,\pi} = 0$.
2. **Consistency**: The belief π is stationary under policy μ_θ, i.e., $\pi = \mathtt{StatDist}(\pi, \mu_\theta)$.

Policy gradient based algorithm: Main proposition

- If $\theta_{k+1} = [\theta_k + \alpha_k G_{\theta_k}]_\Theta$ converges to a limit θ^* along any sample path, then $(\theta^*, \pi_{\theta^*})$ is a ∇-SMFE.
- Likelihood ratio based gradient estimate:

$$G_{\theta_k} = \mathbb{E}_{X \sim \xi_0}[\nabla_\theta V_{\theta,\pi}(X)], \text{ where } \nabla_\theta V_{\theta,\pi}(x) = \mathbb{E}_{A_t \sim \mu_\theta(X_t)}\Big[\sum_{\sigma=0}^{\infty} \Lambda^\sigma_\theta V_{\theta,\pi}(X_\sigma) \;\Big|\; X_0 = x\Big].$$

- Simultaneous perturbation based gradient estimate:

$$G_{\theta_k} = \eta(J_{\theta+\beta\eta,\pi} - J_{\theta-\beta\eta,\pi})/2\beta$$

Policy improvement

input : θ_0 : Initial parameter; K : # iterations; ξ_0 : initial mean field dist; B : burn-in period; n_p : # particles

for *iterations* $k = 1 : K$ **do**
- $\pi_k = \mathtt{StatDist}(\xi_0, \mu_{\theta_k}, B, n_p)$
- $G_{\theta_k} = \mathtt{PolicyGradient}(\theta_k, \xi_0, \pi_k)$
- $\theta_{k+1} \leftarrow [\theta_k + \alpha_k G_{\theta_k}]_\Theta$

return θ_{K+1}

Stationary distribution

input : ξ_0 : Initial dist; θ : parameter; B : burn-in period; n_p : # particles

for $i = 1 : n_p$ **do**
- $x^i_0 \sim \xi_0$
- **for** $t = 0 : B$ **do**
  - $a^i_t \sim \mu_\theta$; $x^i_{t+1} \sim P(\cdot | x^i_t, a^i_t)$;

for $x \in \mathcal{X}$ **do**
- $\pi(x) = \frac{1}{n_p}\sum_{i=1}^{n_p} \mathbb{1}\{x^i_{B+1} = x\}$

return π

Example: Malware spread in networks

- Dynamics ($\{\eta_t\}_{t \geqslant 0}$: i.i.d. process):

$$X^i_{t+1} = \begin{cases} X^i_t + (1 - X^i_t)\eta_t, & \text{for } A^i_t = 0, \\ 0, & \text{for } A^i_t = 1, \end{cases}.$$

- The per-step payoff is:

$$u(x, a, \xi) = -(k + \bar{\xi})x - \lambda a;$$

$\bar{\xi}$ is the mean of ξ and k, λ are given constants.

- We consider threshold policies with $\Theta = [0, 1]$:

$$\mu_\theta(x) = \begin{cases} 0, & \text{if } x < \theta, \\ 1, & \text{if } x \geqslant \theta. \end{cases}$$

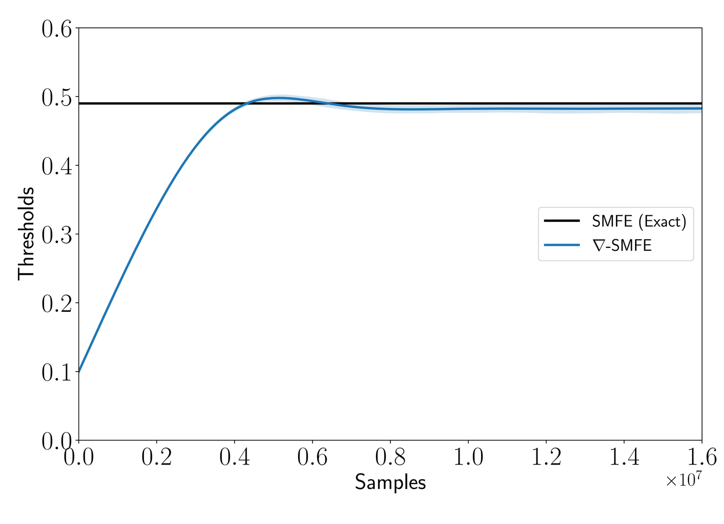

Fig. 3: Thresholds evolution

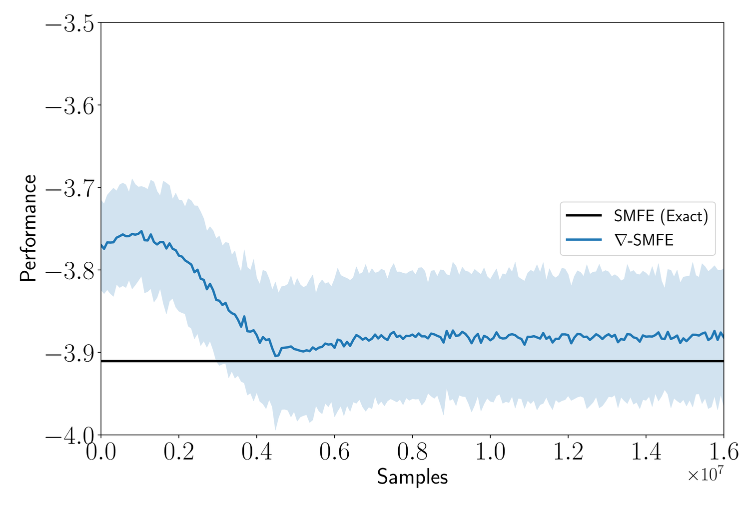

Fig. 4: Performance evolution

Conclusions

- In this work an RL algorithm is used for planning. This implies that the iterates in our algorithm are not representative of the learning dynamics of individual agents.
  - For this to be an RL algorithm, each agent would have to make an assumption on all other agents' behaviour in the learning phase.
  - This coordination in learning is not easily justified in a competitive game with strategic agents, where the agents can try and influence their opponents during learning.
- Although we presented only policy based algorithms, bounded rationality can also be modelled using a critic only variant with function approximation.