Reinforcement Learning in Stationary Mean-field Games

Jayakumar Subramanian & Aditya Mahajan

ECE & CIM, McGill University and GERAD

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

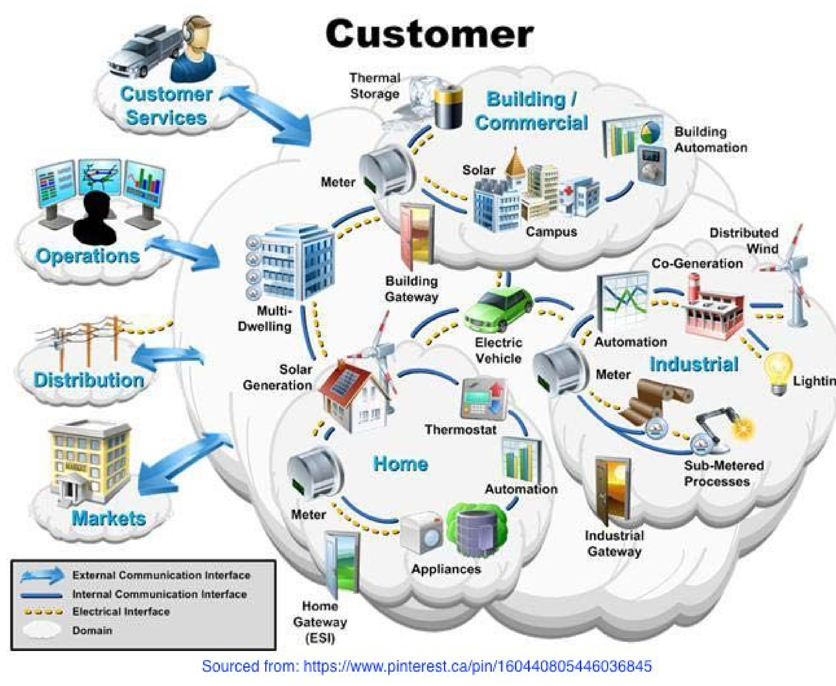

Fig. 1: Smart Grid - Demand Response

Fig. 2: Financial Markets

Solution concept

- Mean-field equilibrium—competitive agents.
- Mean-field social-welfare optimal policy—cooperative agents.
- Extension to stationary mean-field games:
  - Stationary mean-field equilibrium (SMFE)
  - Stationary social-welfare optimal policy (SMF-SO)

Our contribution

- Generalization of these solution concepts to their local variants using bounded rationality based arguments.
- Development of policy gradient based reinforcement learning algorithms to predict these solution concepts.
- Proof of convergence of these algorithms to the right solution concept under mild technical conditions.

Mean field game (MFG) model

- Agent set: $N := \{1, 2, \dots, n\}$ homogeneous agents;
- State and action spaces for each agent: $\mathcal{X}, \mathcal{A}$ (finite and identical for all agents);
- Empirical mean field (or population average):
  $Z_t(x) = \frac{1}{n} \sum_{i \in N} \mathbb{1}\{X^i_t = x\}, \quad \forall x \in \mathcal{X}$.
- Dynamical state evolution for each agent $i \in N$ (decoupled by mean-field):
  $X^i_{t+1} \sim P(X^i_t, A^i_t, Z_t)$.
- Per-step reward to agent i (decoupled by mean-field):
  $R^i_t = r(X^i_t, A^i_t, Z_t, X^i_{t+1})$.

Stationary MFG model

1. **Time homogeneous policy**: All agents follow a time-homogeneous, stochastic policy, $\pi_t = \pi\colon \mathcal{X} \to \Delta(\mathcal{A})$ for all t.
2. **Stationarity of mean-field**: When all agents follow a policy $\pi \in \Pi$, the mean-field of states $\{Z_t\}_{t \geqslant 0}$ converges almost surely to a constant limit: $z = \Phi(z, \pi)$.
3. **Agent's performance evaluation**: Agents evaluate their performance by assuming infinite population stationary mean-field:
   $V_{\pi,z}(x) = \mathbb{E}_{\substack{A^i_t \sim \pi(X^i_t) \\ X^i_{t+1} \sim P(X^i_t, A^i_t, z)}} \Big[\sum_{t=0}^{\infty} \gamma^t r(X^i_t, A^i_t, z, X^i_{t+1}) \Big| X^i_0 = x \Big]$.

Stationary MF equilibrium (SMFE)

A stationary mean-field equilibrium (SMFE) is a pair of policy $\pi \in \Pi$ and mean-field $z \in \Delta(\mathcal{X})$ which satisfies the following two properties:

1. *Sequential rationality:* For any other policy π', $V_{\pi,z}(x) \geqslant V_{\pi',z}(x), \quad \forall x \in \mathcal{X}$.
2. *Consistency:* The mean-field z is stationary under policy π, i.e., $z = \Phi(z, \pi)$.

Stationary MF social-welfare optimal policy (SMF-SO)

A policy $\pi \in \Pi$ is stationary mean-field social welfare optimal (SMF-SO) if it satisfies the following property:

- *Optimality:* For any other policy $\pi' \in \Pi$, $V_{\pi,z}(x) \geqslant V_{\pi',z'}(x), \quad \forall x \in \mathcal{X}$, where z and z' are the stationary mean-field distributions: $z = \Phi(z, \pi) \quad \textit{and} \quad z' = \Phi(z', \pi')$.

Local SMFE (LSMFE)

A local stationary mean-field equilibrium (LSMFE) is a pair of policy $\pi_\theta \in \Pi$ and mean-field $z \in \Delta(\mathcal{X})$ which satisfies the following two properties:

1. *Local sequential rationality:* $\partial J_{\pi_\theta, z} / \partial\theta = 0$.
2. *Consistency:* $z = \Phi(z, \pi_\theta)$.

Local SMF-SO (LSMF-SO)

A policy $\pi_\theta \in \Pi$ is local stationary mean-field social welfare optimal (LSMF-SO) if it satisfies the following property:

- *Local optimality:* $dJ_{\pi_\theta, z_\theta} / d\theta = 0$, where z_θ is the stationary mean-field distribution corresponding to π_θ, i.e., satisfies $z_\theta = \Phi(z_\theta, \pi_\theta)$.

RL algorithm for learning LSMFE

Suppose $G_{\theta,z}$ is an unbiased estimator of $\partial J_{\pi_\theta, z} / \partial\theta$. Then, we start with an initial guess $\theta_0 \in \Theta$ and $z_0 \in \Delta(\mathcal{X})$ and at each step of the iteration, update the guess (θ_k, z_k) using two-timescale stochastic gradient ascent:

$$z_{k+1} = z_k + \beta_k \big[\hat{\Phi}(z_k, \pi_{\theta_k}) - z_k \big]; \quad \theta_{k+1} = \big[\theta_k + \alpha_k G_{\theta_k, z_k} \big]_\Theta$$

where $[\,\cdot\,]_\Theta$ denotes projection on Θ, learning rates $\{\alpha_k, \beta_k\}_{k \geqslant 0}$ are chosen s.t.: $\sum \alpha_k = \infty$, $\sum \beta_k = \infty$, $\sum (\alpha_k^2 + \beta_k^2) < \infty$, $\lim_{k\to\infty} \alpha_k = 0$, $\lim_{k\to\infty} \beta_k = 0$, $\lim_{k\to\infty} \alpha_k / \beta_k = 0$.

Stationary mean-field estimation

$\hat{\Phi}(z, \pi)$ is an unbiased approximation of $\Phi(z, \pi)$ which is generated using a mini-batch of m samples $(X^j, A^j, Y^j)_{j=1}^m$ where $X^j \sim z$, $A^j \sim \pi(\cdot | X^j)$, and $Y^j \sim P(X^j, A^j, z)$ and set

$$\hat{\Phi}(z, \pi)(y) = \frac{1}{m} \sum_{j=1}^{m} \mathbb{1}\{Y^j = y\}.$$

Likelihood ratio based gradient estimate

$$\frac{\partial J_{\theta,z}}{\partial\theta} = \mathbb{E}_{X \sim \xi_0} \left[\frac{\partial V_{\theta,z}(X)}{\partial\theta} \right]$$

$$\frac{\partial V_{\theta,z}(x)}{\partial\theta} = \mathbb{E}_{A_t \sim \pi_\theta(X_t)} \Big[\sum_{t=0}^{\infty} \gamma^t \Lambda^t_\theta V_{\pi_\theta, z}(X_t) \;\Big|\; X_0 = x \Big],$$

where $\Lambda^t_\theta = \nabla_\theta \log[\pi_\theta(A_t \mid X_t)]$.

RL algorithm for learning LSMF-SO

Suppose T_θ is an unbiased estimator for $dJ_{\pi_\theta, z_\theta} / d\theta$, where z_θ is the fixed point of $z = \Phi(z, \pi_\theta)$. Then, we start with an initial guess $\theta_0 \in \Theta$, and at each step of the iteration, update the guess using stochastic gradient ascent:

$$\theta_{k+1} = \big[\theta_k + \alpha_k T_{\theta_k} \big]_\Theta$$

Numerical examples

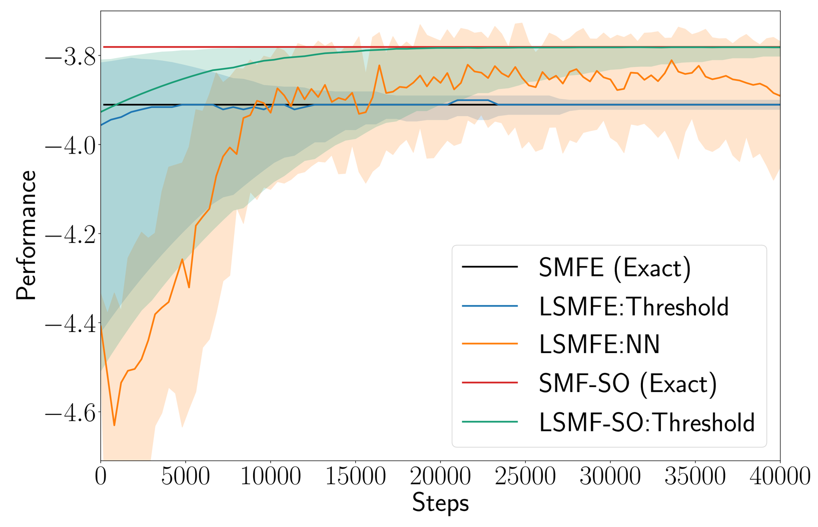

Fig. 3: Malware spread

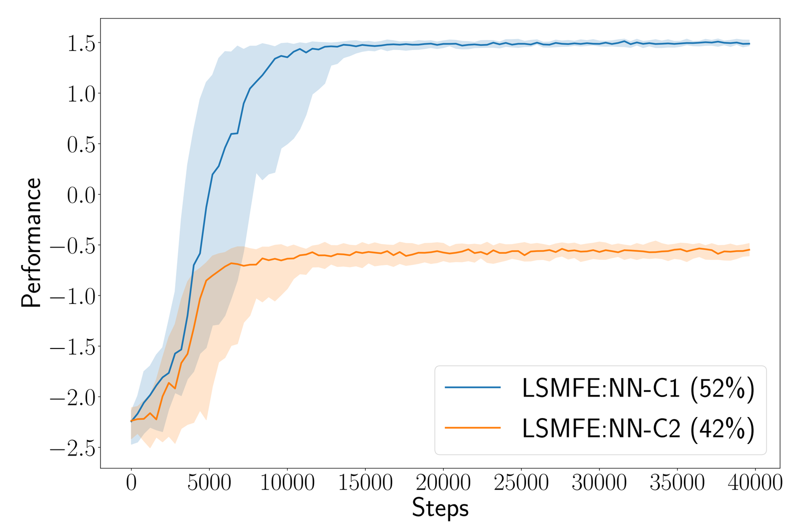

Fig. 4: Product investments

LaTeX TikZposter