Reinforcement learning in stationary mean-field games

Aditya Mahajan
McGill University

Based on work with Jayakumar Subramanian (Adobe Research)

Machine Learning and Mean-field games seminars
23rd Nov 2021


Mean-field interactions arise in various applications

The importance of mean-field interactions have led to various models of mean-field games and teams.

Excellent overview in the previous two talks in this series!

Outline

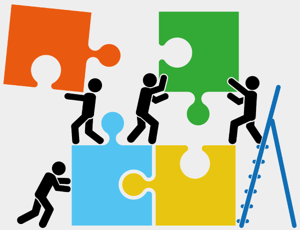
System Model

- Mean-field models
  - Stationary mean-field equilibrium
  - Stationary mean-field social optimality
  - Local solution concepts

Outline

System Model

- Mean-field models
  - Stationary mean-field equilibrium
  - Stationary mean-field social optimality
  - Local solution concepts



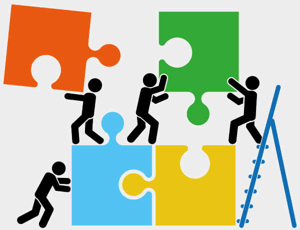
RL for MF

- RL for SMFE
- RL for SMF-SO

Outline

System Model

- Mean-field models
  - Stationary mean-field equilibrium
  - Stationary mean-field social optimality
  - Local solution concepts



RL for MF

- RL for SMFE
- RL for SMF-SO

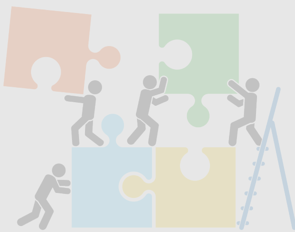
Numerical examples

- Malware spread in networks

Outline

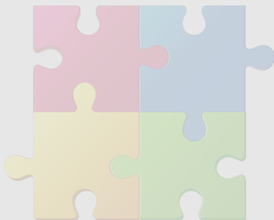
System Model

- Mean-field models
  - Stationary mean-field equilibrium
  - Stationary mean-field social optimality
  - Local solution concepts



RL for MF

- RL for SMFE
- RL for SMF-SO



Numerical examples

- Malware spread in networks

System Model

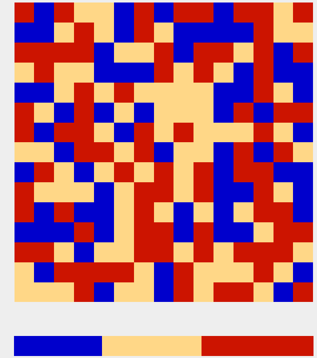
Population of homogeneous agents

- n homogeneous agents.
- State space $\mathcal{S}$; action space $\mathcal{A}$.
- $(S^i_t, A^i_t) \in \mathcal{S} \times \mathcal{A}$. State and action of agent i at time t.

System Model

Population of homogeneous agents

- n homogeneous agents.
- State space $\mathcal{S}$; action space $\mathcal{A}$.
- $(S^i_t, A^i_t) \in \mathcal{S} \times \mathcal{A}$. State and action of agent i at time t.



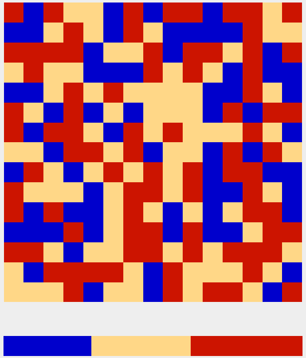
Mean-field coupling

- **Mean-field**: $Z_t(s) = \dfrac{1}{n} \sum_{i \in N} \mathbb{1}\{S^i_t = s\}$.

System Model

Population of homogeneous agents

- n homogeneous agents.
- State space $\mathcal{S}$; action space $\mathcal{A}$.
- $(S^i_t, A^i_t) \in \mathcal{S} \times \mathcal{A}$. State and action of agent i at time t.



Mean-field coupling

- **Mean-field**: $Z_t(s) = \dfrac{1}{n} \sum_{i \in N} \mathbb{1}\{S^i_t = s\}$.
- Dynamics: $S^i_{t+1} \sim P(S^i_t, A^i_t, Z_t)$
- Per-step reward: $R^i_t = r(S^i_t, A^i_t, Z_t, S^i_{t+1})$.

Utility of each agent

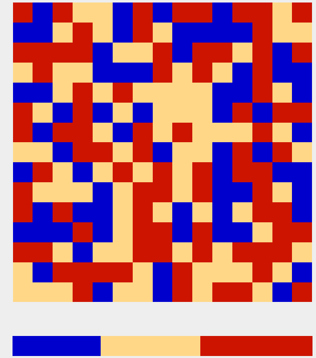
- Utility of agent i

$$V^i(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R^i_t \,\middle|\, S^i_0 = s \right].$$

System Model

Population of homogeneous agents

- n homogeneous agents.
- State space $\mathcal{S}$; action space $\mathcal{A}$.
- $(S_t^i, A_t^i) \in \mathcal{S} \times \mathcal{A}$. State and action of agent i at time t.



Mean-field coupling

- **Mean-field**: $Z_t(s) = \dfrac{1}{n} \sum_{i \in N} \mathbb{1}\{S_t^i = s\}$.
- Dynamics: $S_{t+1}^i \sim P(S_t^i, A_t^i, Z_t)$
- Per-step reward: $R_t^i = r(S_t^i, A_t^i, Z_t, S_{t+1}^i)$.
- If all agents play a Markov policy $\pi_t \colon \mathcal{S} \to \Delta(\mathcal{A})$:

$$Z_{t+1}(s') = \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} Z_t(s) \pi_t(a|s) P(s'|s, a, Z_t)$$

Utility of each agent

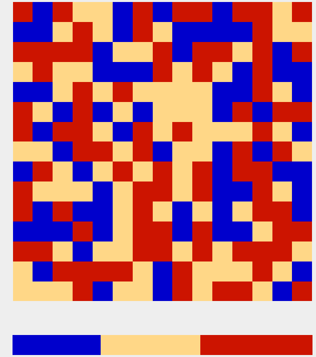
- Utility of agent i

$$V^i(s) = \mathbb{E}\Bigg[\sum_{t=0}^{\infty} \gamma^t R_t^i \,\Bigg|\, S_0^i = s\Bigg].$$

System Model

Population of homogeneous agents

- n homogeneous agents.
- State space $\mathcal{S}$; action space $\mathcal{A}$.
- $(S^i_t, A^i_t) \in \mathcal{S} \times \mathcal{A}$. State and action of agent i at time t.



Mean-field coupling

- **Mean-field**: $Z_t(s) = \dfrac{1}{n} \sum_{i \in N} \mathbb{1}\{S^i_t = s\}$.
- Dynamics: $S^i_{t+1} \sim P(S^i_t, A^i_t, Z_t)$
- Per-step reward: $R^i_t = r(S^i_t, A^i_t, Z_t, S^i_{t+1})$.
- If all agents play a Markov policy $\pi_t \colon \mathcal{S} \to \Delta(\mathcal{A})$:

$$Z_{t+1}(s') = \sum_{s \in \mathcal{S}} \sum_{a \in \mathcal{A}} Z_t(s) \pi_t(a|s) P(s'|s, a, Z_t)$$

$$\boxed{Z_{t+1} = \Phi(Z_t, \pi_t)}$$

Discrete-time Fokker-Plank eqn

Utility of each agent

- Utility of agent i

$$V^i(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R^i_t \,\middle|\, S^i_0 = s \right].$$

Solution concept

Stationary mean-field equilibrium (SMFE)

- Solution concept proposed by Weintraub, Benkard, and Van Roy (2005, 2008)
  . . . and extended by Adlakha, Johari, and Weintraub (2010).
- Contemporaneous to the other "evolutive" solution concept for mean-field games
  - Huang, Malhame, Caines (2003, 2006)
  - Larsy and Lions (2005)

Solution concept

Stationary mean-field equilibrium (SMFE)

- Solution concept proposed by Weintraub, Benkard, and Van Roy (2005, 2008)
  . . . and extended by Adlakha, Johari, and Weintraub (2010).
- Contemporaneous to the other "evolutive" solution concept for mean-field games
  - Huang, Malhame, Caines (2003, 2006)
  - Larsy and Lions (2005)

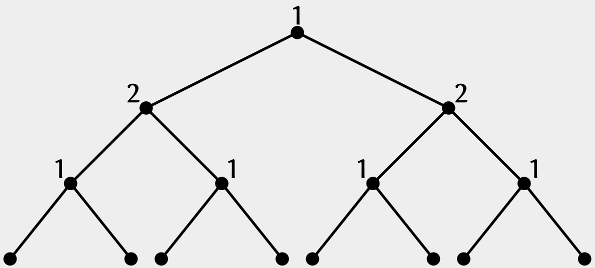
Interpreting SMFE

- Presented as an approximation to Markov perfect equilibrium
  . . . of a game where agents observe the state of all players
- The equilibrium in "evolutive" mean-field game is also typically presented as an approximation to Markov perfect equilibrium.

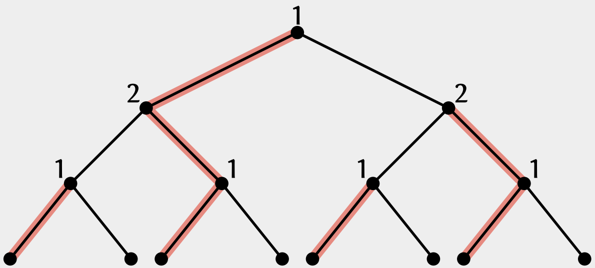
An alternative view:

SMFE is a sequential equilibrium of a game with imprefect information.

Review: Extensive form games with perfect information



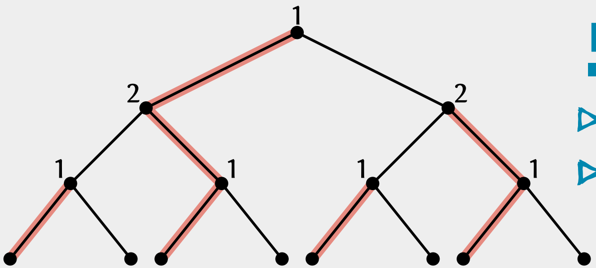
Review: Extensive form games with perfect information



Normal-form reduction

–	–	–	–	–
–	–	–	–	–
–	–	–	–	–
–	–	–	–	–
–	–	–	–	–

Review: Extensive form games with perfect information



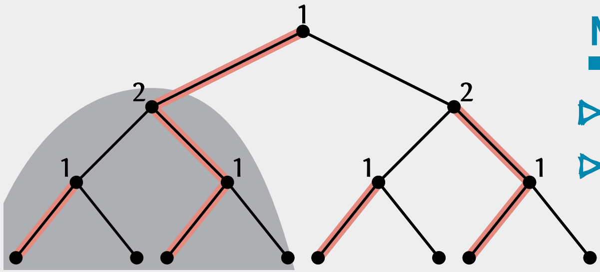
Nash equilibrium

- Reduce the extensive form game to a normal form game.
- A NE strategy of the normal form game is a NE of the extensive form game.
  - Not ideal, because gives rise to equilibrium which are based on non-credible threats.

Normal-form reduction

–	–	–	–	–
–	–	–	–	–
–	–	–	–	–
–	–	–	–	–
–	–	–	–	–

Review: Extensive form games with perfect information



Nash equilibrium

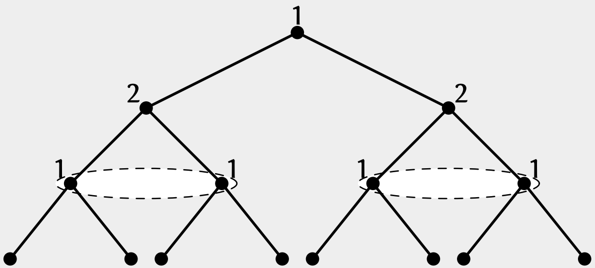
- Reduce the extensive form game to a normal form game.
- A NE strategy of the normal form game is a NE of the extensive form game.
  - Not ideal, because gives rise to equilibrium which are based on non-credible threats.

Subgame-perfect equilibrium

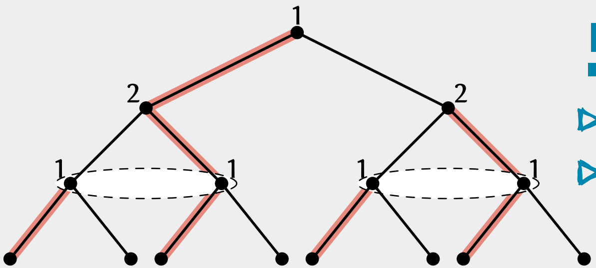
- A strategy profile which is a NE of every subgame
- Can be solved by dynamic programming

- Special case: Markov perfect equilibrium

Review: Extensive form games with imperfect information



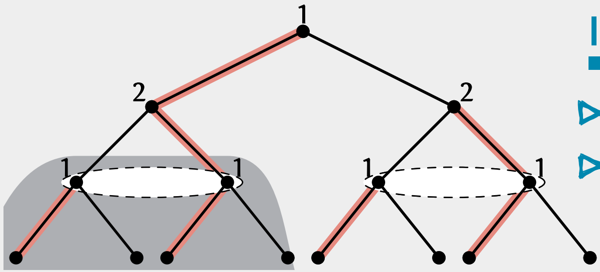
Review: Extensive form games with imperfect information



Information sets

- Nodes of a game tree which are indistinguishable to a player
- Must play the same move at all nodes in an information set.

Review: Extensive form games with imperfect information

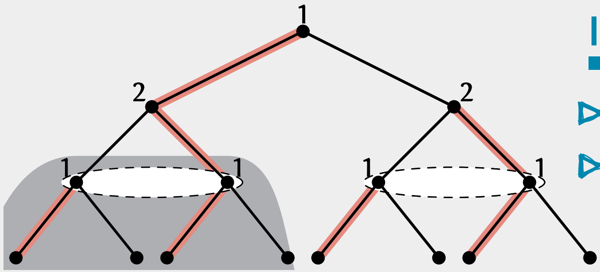


Information sets

- Nodes of a game tree which are indistinguishable to a player
- Must play the same move at all nodes in an information set.
  - How to evaluate performance of a sub-tree?

    Need to have belief on all nodes in an information set.

Review: Extensive form games with imperfect information



Information sets

- Nodes of a game tree which are indistinguishable to a player
- Must play the same move at all nodes in an information set.
  - How to evaluate performance of a sub-tree?

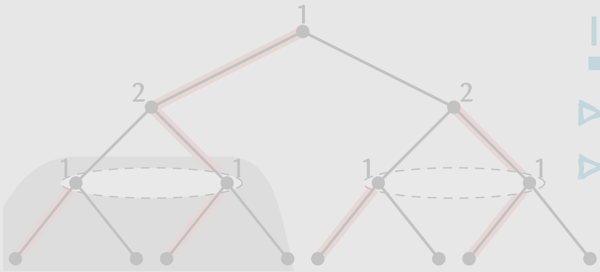
    Need to have belief on all nodes in an information set.

Sequential equilibrium (Kreps and Wilson, 1982)

A strategy profile and a belief system which satisfy:

- Sequential rationality: If we evaluate performance according to beliefs, then in each subgame, each player is playing a NE.
- Consistency: The beliefs are Bayes consistent with the strategy.

Review: Extensive form games with imperfect information



Information sets

- Nodes of a ga... a player
- Must play the... tion set.
  - How to ev...

  Need to h... on set.

DP doesn't work!
Difficult to compute both for finite and infinite horizon models

Sequential equilibrium (Kreps and Wilson, 1982)

A strategy profile and a belief system which satisfy:

- **Sequential rationality:** If we evaluate performance according to beliefs, then in each subgame, each player is playing a NE.
- **Consistency:** The beliefs are Bayes consistent with the strategy.

A squential equilibrium for mean-field game

Stationary mean-field equilibrium (SMFE)

- A strategy profile $\pi\colon \mathcal{S} \to \Delta(\mathcal{A})$ and a mean-field belief z such that:
- Consistency: $z = \Phi(z, \pi)$

A squential equilibrium for mean-field game

Stationary mean-field equilibrium (SMFE)

- A strategy profile $\pi\colon \mathcal{S} \to \Delta(\mathcal{A})$ and a mean-field belief z such that:
- Consistency: $z = \Phi(z, \pi)$

Stationarity of beliefs

A squential equilibrium for mean-field game

Stationary mean-field equilibrium (SMFE)

- A strategy profile $\pi \colon \mathcal{S} \to \Delta(\mathcal{A})$ and a mean-field belief z such that:
- Consistency: $z = \Phi(z, \pi)$

Stationarity of beliefs

- Evaluation of performance (same for all i)

$$V^i_{\pi,z}(s) = \mathbb{E}_{\substack{A^i_t \sim \pi(S^i_t) \\ S^i_{t+1} \sim P(S^i_t, A^i_t, z)}} \left[\sum_{t=0}^{\infty} \gamma^t r(S^i_t, A^i_t, z, S^i_{t+1}) \,\middle|\, S^i_0 = s \right].$$

A squential equilibrium for mean-field game

Stationary mean-field equilibrium (SMFE)

- A strategy profile $\pi \colon \mathcal{S} \to \Delta(\mathcal{A})$ and a mean-field belief z such that:
- **Consistency**: $z = \Phi(z, \pi)$
- Evaluation of performance (same for all i)

$$V^i_{\pi,z}(s) = \mathbb{E}_{\substack{A^i_t \sim \pi(S^i_t) \\ S^i_{t+1} \sim P(S^i_t, A^i_t, z)}} \left[\sum_{t=0}^{\infty} \gamma^t r(S^i_t, A^i_t, z, S^i_{t+1}) \,\middle|\, S^i_0 = s \right].$$

Stationarity of beliefs

Evaluate performance according to belief

A squential equilibrium for mean-field game

Stationary mean-field equilibrium (SMFE)

- A strategy profile $\pi \colon \mathcal{S} \to \Delta(\mathcal{A})$ and a mean-field belief z such that:
- **Consistency**: $z = \Phi(z, \pi)$

Stationarity of beliefs

- Evaluation of performance (same for all i)

$$V^i_{\pi,z}(s) = \mathbb{E}_{\substack{A^i_t \sim \pi(S^i_t) \\ S^i_{t+1} \sim P(S^i_t, A^i_t, z)}} \left[\sum_{t=0}^{\infty} \gamma^t r(S^i_t, A^i_t, z, S^i_{t+1}) \,\middle|\, S^i_0 = s \right].$$

Evaluate performance according to belief

- **Sequential rationality:** For any other strategy π'

$$V_{\pi,z}(s) \geq V_{\pi',z}(s)$$

A squential equilibrium for mean-field game

Stationary mean-field equilibrium (SMFE)

- A strategy profile $\pi\colon \mathcal{S} \to \Delta(\mathcal{A})$ and a mean-field belief z such that:
- **Consistency**: $z = \Phi(z, \pi)$
- Evaluation of performance (same for all i)

$$V^i_{\pi,z}(s) = \mathbb{E}_{\substack{A^i_t \sim \pi(S^i_t) \\ S^i_{t+1} \sim P(S^i_t, A^i_t, z)}} \left[\sum_{t=0}^{\infty} \gamma^t r(S^i_t, A^i_t, z, S^i_{t+1}) \,\middle|\, S^i_0 = s \right].$$

- **Sequential rationality**: For any other strategy π'

$$V_{\pi,z}(s) \geq V_{\pi',z}(s)$$

Stationarity of beliefs

Evaluate performance according to belief

Social optimality

Stationary mean-field social-welfare optimality (SMF-SO)

- Consider the setting where the players are cooperative.
- Performance of a generic agent (same as before)

$$V^i_{\pi,z}(s) = \mathbb{E}_{\substack{A^i_t \sim \pi(S^i_t) \\ S^i_{t+1} \sim P(S^i_t, A^i_t, z)}} \left[\sum_{t=0}^{\infty} \gamma^t r(S^i_t, A^i_t, z, S^i_{t+1}) \,\middle|\, S^i_0 = s \right].$$

Social optimality

Stationary mean-field social-welfare optimality (SMF-SO)

- Consider the setting where the players are cooperative.
- Performance of a generic agent (same as before)

$$V^i_{\pi,z}(s) = \mathbb{E}_{\substack{A^i_t \sim \pi(S^i_t) \\ S^i_{t+1} \sim P(S^i_t, A^i_t, z)}} \left[\sum_{t=0}^{\infty} \gamma^t r(S^i_t, A^i_t, z, S^i_{t+1}) \,\middle|\, S^i_0 = s \right].$$

- Optimality

$$V_{\pi,z}(s) \geqslant V_{\pi',z'}(s)$$

Social optimality

Stationary mean-field social-welfare optimality (SMF-SO)

- Consider the setting where the players are cooperative.
- Performance of a generic agent (same as before)

$$V^i_{\pi,z}(s) = \mathbb{E}_{\substack{A^i_t \sim \pi(S^i_t) \\ S^i_{t+1} \sim P(S^i_t, A^i_t, z)}} \left[\sum_{t=0}^{\infty} \gamma^t r(S^i_t, A^i_t, z, S^i_{t+1}) \,\middle|\, S^i_0 = s \right].$$

- Optimality

$$V_{\pi,z}(s) \geq V_{\pi',z'}(s)$$

$$z = \Phi(z, \pi) \qquad z' = \Phi(z', \pi')$$

Social optimality

Stationary mean-field social-welfare optimality (SMF-SO)

- Consider the setting where the players are cooperative.
- Performance of a generic agent (same as before)

$$V^i_{\pi,z}(s) = \mathbb{E}_{\substack{A^i_t \sim \pi(S^i_t) \\ S^i_{t+1} \sim P(S^i_t, A^i_t, z)}} \left[\sum_{t=0}^{\infty} \gamma^t r(S^i_t, A^i_t, z, S^i_{t+1}) \,\middle|\, S^i_0 = s \right].$$

- Optimality

$$V_{\pi,z}(s) \geq V_{\pi',z'}(s)$$

$$z = \Phi(z, \pi) \qquad z' = \Phi(z', \pi')$$

Equilibrium and social optimality are different

- For equilibrium, deviation in policy does not change the stationary mean-field (single player is deviating)
- For optimality, deviation in policy changes the stationary mean-field (entire population is deviating)

Agents with bounded rationality

Global solution	Solution concepts require global search over all policies

Agents with bounded rationality

Global solution	Solution concepts require global search over all policies
Curse of dimensionality	Verification requires computation of value functions

Agents with bounded rationality

Global solution	Solution concepts require global search over all policies
Curse of dimensionality	Verification requires computation of value functions

Use local search over parameterized policies

Agents with bounded rationality

Global solution	Solution concepts require global search over all policies
Curse of dimensionality	Verification requires computation of value functions

Use local search over parameterized policies

Use function approximation

Local versions of the solution concepts

Local versions of the solution concepts

Preliminaries

- Scalarize returns: Assume $s_0^i \sim \xi_0$ (start state distribution, independent across agents)

$$J_{\pi,z} = \mathbb{E}_{S_0 \sim \xi}[V_{\pi,z}(S_0)]$$

- Parameterize policies: π_θ where $\theta \in \Theta$ [closed compact set] (e.g., softmax)

Local versions of the solution concepts

Preliminaries

- Scalarize returns: Assume $s_0^i \sim \xi_0$ (start state distribution, independent across agents)

$$J_{\pi,z} = \mathbb{E}_{S_0 \sim \xi}[V_{\pi,z}(S_0)]$$

- Parameterize policies: π_θ where $\theta \in \Theta$ [closed compact set] (e.g., softmax)

Local SMFE (LSMFE)

LSMFE is a pair (π_θ, z) that satisfies:

- Local sequential rationality:

$$\frac{\partial J_{\pi_\theta, z}}{\partial \theta} = 0$$

- Consistency: $z = \Phi(z, \pi_\theta)$

Local versions of the solution concepts

Preliminaries

- Scalarize returns: Assume $s_0^i \sim \xi_0$ (start state distribution, independent across agents)

$$J_{\pi,z} = \mathbb{E}_{S_0 \sim \xi}[V_{\pi,z}(S_0)]$$

- Parameterize policies: π_θ where $\theta \in \Theta$ [closed compact set] (e.g., softmax)

Local SMFE (LSMFE)

LSMFE is a pair (π_θ, z) that satisfies:

- Local sequential rationality:

$$\frac{\partial J_{\pi_\theta, z}}{\partial \theta} = 0$$

- Consistency: $z = \Phi(z, \pi_\theta)$

Local SMF-SO (LSMF-SO)

LSMF-SO is a policy π_θ that satisfies:

- Local optimality:

$$\frac{d J_{\pi_\theta, z_\theta}}{d\theta} = 0$$

where $z_\theta = \Phi(z_\theta, \pi_\theta)$

Comparison of the two local solution concepts

$$\frac{\mathrm{d}J_{\pi,z}}{\mathrm{d}\theta} = \frac{\partial J_{\pi,z}}{\partial \pi}\frac{\partial \pi}{\partial \theta} + \frac{\partial J_{\pi,z}}{\partial z}\frac{\partial z}{\partial \theta}$$

Comparison of the two local solution concepts

$$\boxed{\frac{dJ_{\pi,z}}{d\theta}} = \frac{\partial J_{\pi,z}}{\partial \pi}\frac{\partial \pi}{\partial \theta} + \frac{\partial J_{\pi,z}}{\partial z}\frac{\partial z}{\partial \theta}$$

Local optimality

Comparison of the two local solution concepts

$$\boxed{\frac{dJ_{\pi,z}}{d\theta}} = \boxed{\frac{\partial J_{\pi,z}}{\partial \pi}\frac{\partial \pi}{\partial \theta}} + \frac{\partial J_{\pi,z}}{\partial z}\frac{\partial z}{\partial \theta}$$

$$\frac{\partial J_{\pi_\theta,z}}{\partial \theta}$$

Local optimality

Sequential rationality

Comparison of the two local solution concepts

$$\boxed{\frac{dJ_{\pi,z}}{d\theta}} = \boxed{\frac{\partial J_{\pi,z}}{\partial \pi}\frac{\partial \pi}{\partial \theta}} + \boxed{\frac{\partial J_{\pi,z}}{\partial z}}\frac{\partial z}{\partial \theta}$$

$$\frac{dJ_{\pi,z}}{d\theta} \longleftrightarrow \text{Local optimality}$$

$$\frac{\partial J_{\pi,z}}{\partial \pi}\frac{\partial \pi}{\partial \theta} \longleftrightarrow \frac{\partial J_{\pi_\theta,z}}{\partial \theta} \longleftrightarrow \text{Sequential rationality}$$

$$\frac{\partial J_{\pi,z}}{\partial z} \longleftrightarrow \neq 0$$

Local optimality

Sequential rationality

Comparison of the two local solution concepts

$$\frac{dJ_{\pi,z}}{d\theta} = \frac{\partial J_{\pi,z}}{\partial \pi}\frac{\partial \pi}{\partial \theta} + \frac{\partial J_{\pi,z}}{\partial z}\frac{\partial z}{\partial \theta}$$

Local optimality

$$\frac{\partial J_{\pi_\theta,z}}{\partial \theta}$$

Sequential rationality

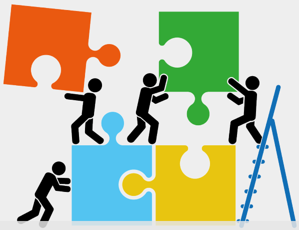
$\neq 0$

LSMFE $\neq$ LSMF-SO

Outline

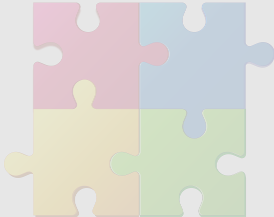
System Model

- Mean-field models
  - Stationary mean-field equilibrium
  - Stationary mean-field social optimality
  - Local solution concepts



RL for MF

- RL for SMFE
- RL for SMF-SO



Numerical examples

- Malware spread in networks

RL for SMFE (strategic agents)

Two-time scale algorithm

- Update policy parameters: $\theta_{k+1} = [\theta_k + \alpha_k G_{\theta_k, z_k}]_\Theta$

RL for SMFE (strategic agents)

Two-time scale algorithm

- Update policy parameters: $\theta_{k+1} = [\theta_k + \alpha_k G_{\theta_k, z_k}]_\Theta$

Unbiased estimator of $\partial J_{\theta,z}/\partial\theta$

RL for SMFE (strategic agents)

Two-time scale algorithm

- Update policy parameters: $\theta_{k+1} = [\theta_k + \alpha_k \boxed{G_{\theta_k, z_k}}]_\Theta$

  Unbiased estimator of $\partial J_{\theta,z}/\partial\theta$

- Update mean-field: $z_{k+1} = z_k + \beta_k[\hat{\Phi}(z_k, \pi_{\theta_k}) - z_k]$

RL for SMFE (strategic agents)

Two-time scale algorithm

- Update policy parameters: $\theta_{k+1} = [\theta_k + \alpha_k G_{\theta_k, z_k}]_\Theta$

  Unbiased estimator of $\partial J_{\theta,z}/\partial\theta$

- Update mean-field: $z_{k+1} = z_k + \beta_k [\hat{\Phi}(z_k, \pi_{\theta_k}) - z_k]$

  Unbiased est. of $\Phi(z_k, \pi_{\theta_k})$

RL for SMFE (strategic agents)

Two-time scale algorithm

- Update policy parameters: $\theta_{k+1} = [\theta_k + \alpha_k G_{\theta_k, z_k}]_\Theta$
- Update mean-field: $z_{k+1} = z_k + \beta_k [\hat{\Phi}(z_k, \pi_{\theta_k}) - z_k]$
- Two-timescale conditions: $\dfrac{\alpha_k}{\beta_k} \longrightarrow 0$ + (standard Robbins-Monro conditions)

RL for SMFE (strategic agents)

Two-time scale algorithm

- Update policy parameters: $\theta_{k+1} = [\theta_k + \alpha_k G_{\theta_k, z_k}]_\Theta$ Slower timescale

- Update mean-field: $z_{k+1} = z_k + \beta_k [\hat{\Phi}(z_k, \pi_{\theta_k}) - z_k]$ Faster timescale

- Two-timescale conditions: $\dfrac{\alpha_k}{\beta_k} \longrightarrow 0$ + (standard Robbins-Monro conditions)

RL for SMFE (strategic agents)

Two-time scale algorithm

- Update policy parameters: $\theta_{k+1} = [\theta_k + \alpha_k G_{\theta_k, z_k}]_\Theta$

Slower timescale

- Update

cale

Convergence result: Under standard technical conditions,

$(\theta_k, z_k) \to$ LSMFE

- Two-timescale conditions: $\dfrac{\alpha_k}{\beta_k} \to 0$ + (standard Robbins-Monro conditions)

Practical considerations

Unrolling two timescales

- It is hard to make two timescale algos work in practice.
- For every iteration of the slow timescale (update of θ_k),

  . . . run multiple rollouts of the fast timescale (update of z_k).
- Equivalent to estimating $\Phi(z, \pi_\theta)$ in a particle-filter like approach

Practical considerations

Unrolling two timescales

- It is hard to make two timescale algos work in practice.
- For every iteration of the slow timescale (update of θ_k),

  . . . run multiple rollouts of the fast timescale (update of z_k).

Estimating gradients

- Likelihood ratio based estimates

$$\frac{\partial J_{\pi_\theta, z}}{\partial \theta} = \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t \nabla_\theta \log \pi_\theta(A^i_t | S^i_t) V_{\pi_\theta, z}(S^i_t) \;\Big|\; S_0 \sim \xi_0\Big]$$

Practical considerations

Unrolling two timescales

- It is hard to make two timescale algos work in practice.
- For every iteration of the slow timescale (update of θ_k),

  . . . run multiple rollouts of the fast timescale (update of z_k).

Estimating gradients

- Likelihood ratio based estimates

$$\frac{\partial J_{\pi_\theta, z}}{\partial \theta} = \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t \nabla_\theta \log \pi_\theta(A_t^i | S_t^i) V_{\pi_\theta, z}(S_t^i) \;\Big|\; S_0 \sim \xi_0\Big]$$

- Simultaneous perturbation based estimates

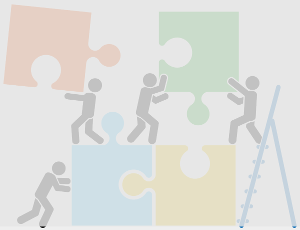
$$G_{\theta, z} = \frac{\eta}{2c}\big(J_{\theta + c\eta, z} - J_{\theta - c\eta, z}\big) \qquad \begin{bmatrix} \text{SPSA: } \eta_i \sim \text{Unif}(\pm 1) \\ \text{SFSA: } \eta_i \sim \mathcal{N}(0, I) \end{bmatrix}$$

Similar ideas work for LSMF-SO
(except we don't have likelihood
ratio based gradient estimates)

Outline

System Model

- Mean-field models
  - Stationary mean-field equilibrium
  - Stationary mean-field social optimality
  - Local solution concepts



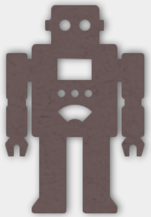
RL for MF

- RL for SMFE
- RL for SMF-SO

Numerical examples

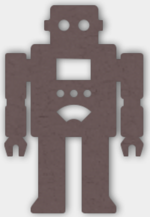
- Malware spread in networks

Example 1: Malware spread in networks

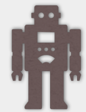


Healthy (0)

Example 1: Malware spread in networks

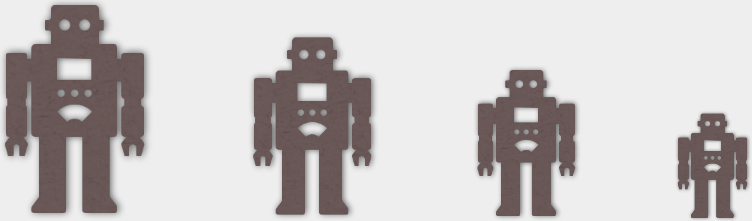


Healthy (0)



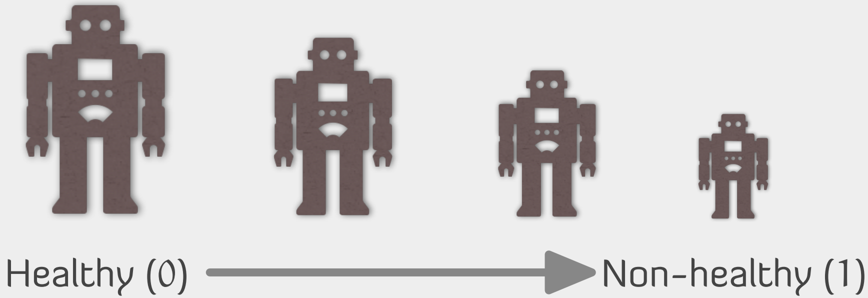
Non-healthy (1)

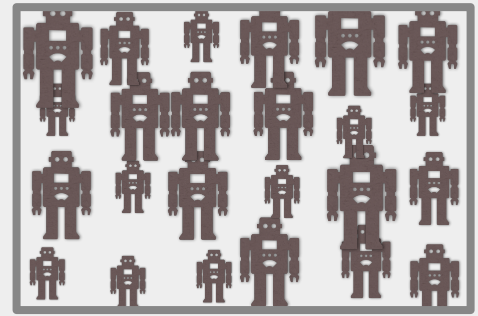
Example 1: Malware spread in networks



Healthy (0) → Non-healthy (1)

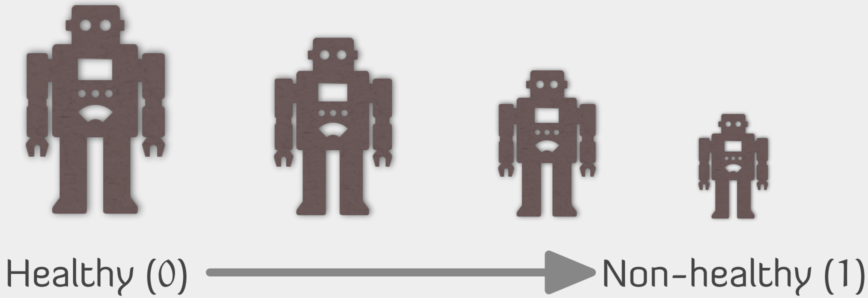
Example 1: Malware spread in networks

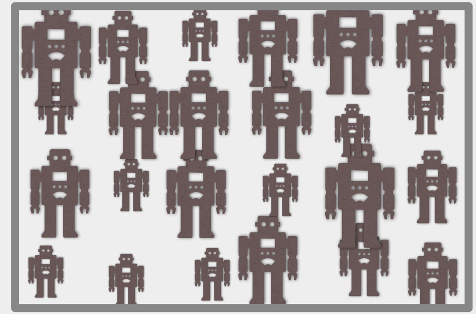




Mean-field

Example 1: Malware spread in networks

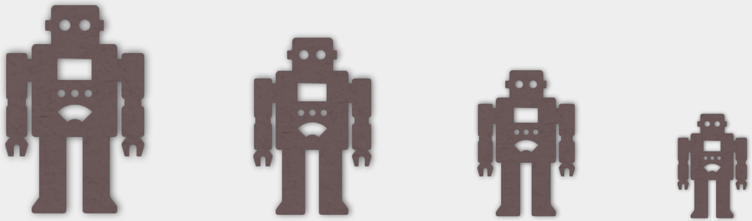




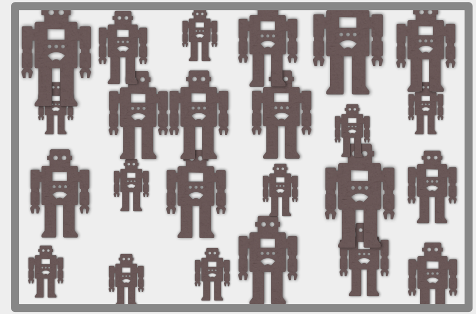
Action - 0 (Do nothing)

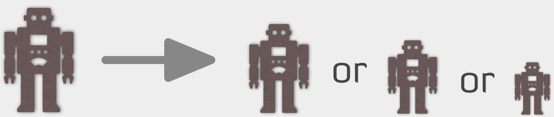
Action - 1 (Repair)

Example 1: Malware spread in networks

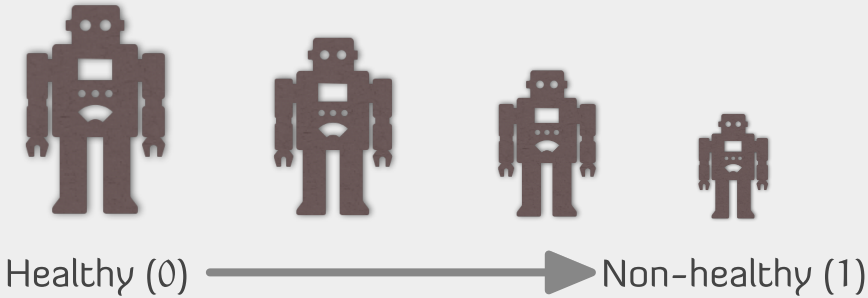


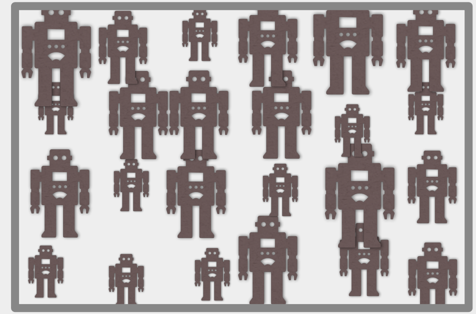
Healthy (0) → Non-healthy (1)



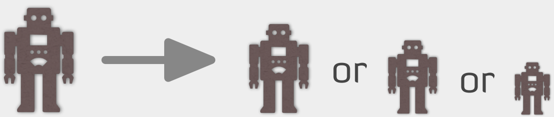
Action - 0 (Do nothing)	Action - 1 (Repair)
	

Example 1: Malware spread in networks

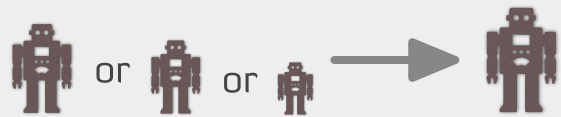




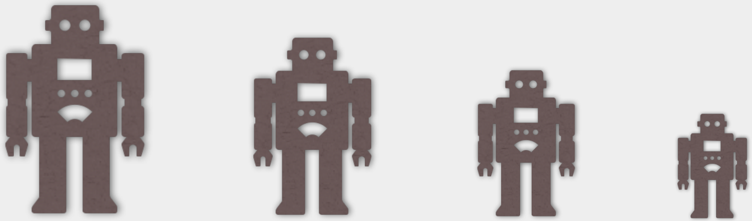
Action - 0 (Do nothing)

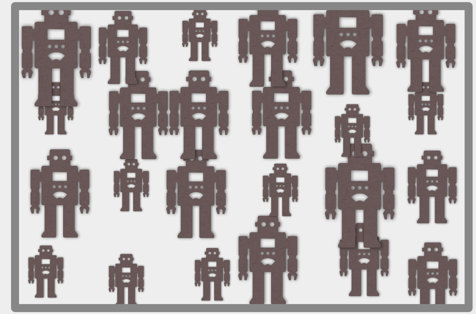


Action - 1 (Repair)



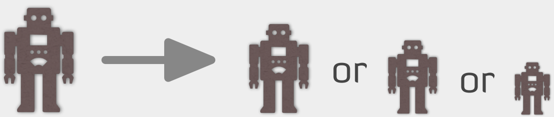
Example 1: Malware spread in networks

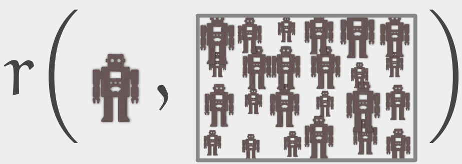




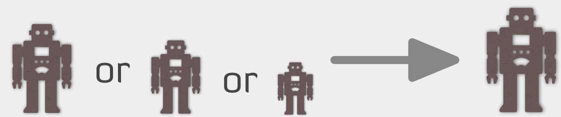
Healthy (0) → Non-healthy (1)

Action - 0 (Do nothing)

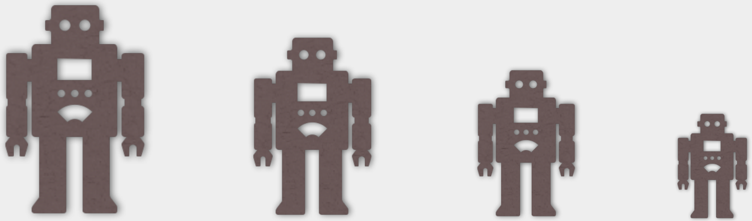




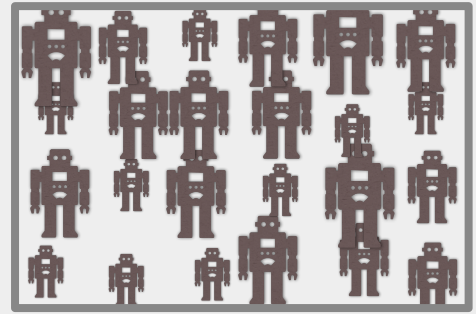
Action - 1 (Repair)



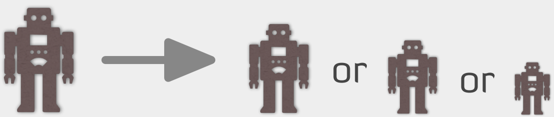
Example 1: Malware spread in networks

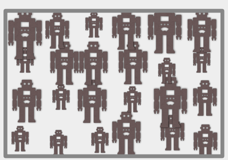


Healthy (0) → Non-healthy (1)

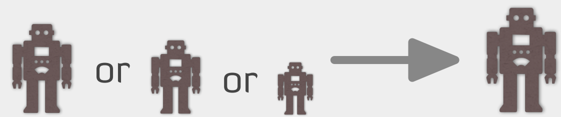


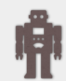
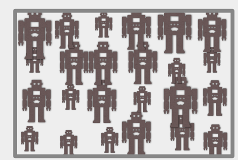
Action - 0 (Do nothing)



$r\Big($  ,  $\Big)$

Action - 1 (Repair)



$r\Big($  ,  $\Big) + \boxed{\text{Repair}}$

Example 1: Malware spread in networks

Salient features

- Representative model for problems with positive externalities.
- Reward: $r(S^i, A^i, Z) = -(k + \langle Z \rangle) S^i - \lambda A^i$ where $\langle Z \rangle = \int s Z(s) ds$.
- Known that SMFE is unique and is a threshold-based strategy: Repair when $S^i_t \geqslant \tau$.

Example 1: Malware spread in networks

Salient features

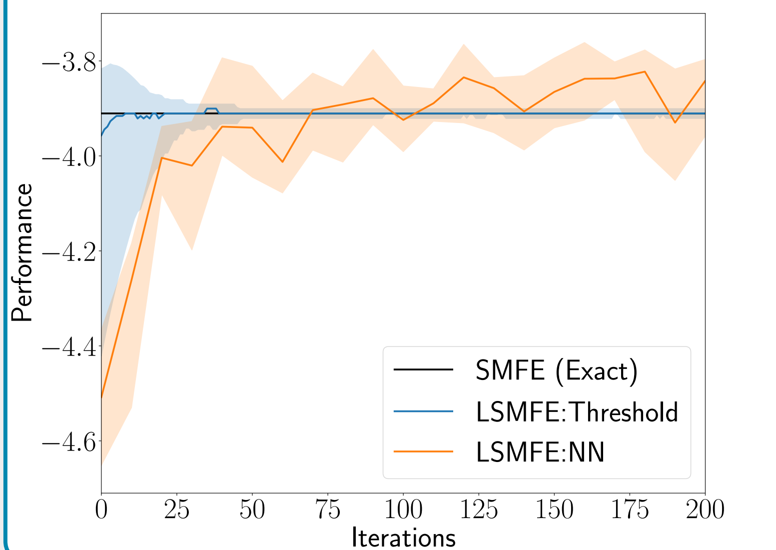
- Representative model for problems with positive externalities.
- Reward: $r(S^i, A^i, Z) = -(k + \langle Z \rangle)S^i - \lambda A^i$ where $\langle Z \rangle = \int sZ(s)ds$.
- Known that SMFE is unique and is a threshold-based strategy: Repair when $S^i_t \geqslant \tau$.

Policy parameterizations

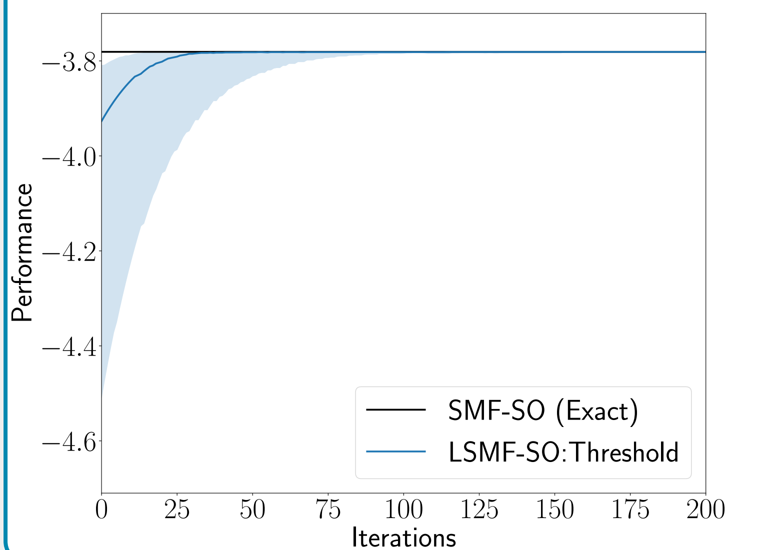
- Threshold based policies with $\tau \in [0, 1]$. Update τ using SPSA.
- Neural network based policies. Compute gradient using REINFORCE.

Results: Performance

Strategic agents

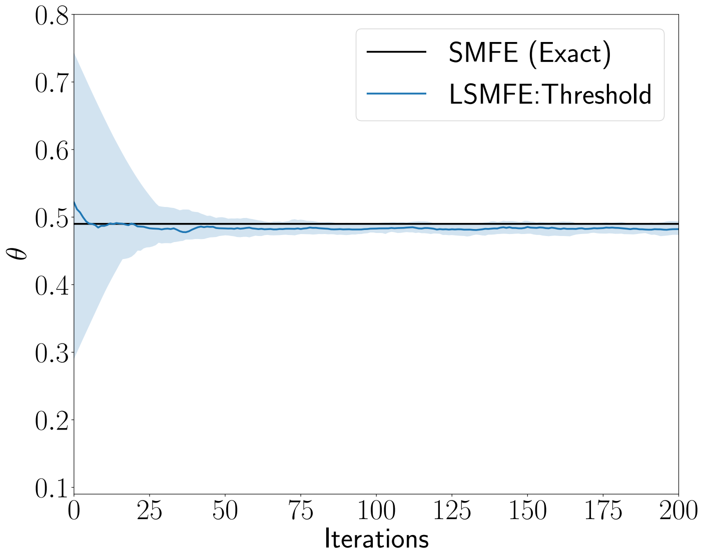


Cooperative agents

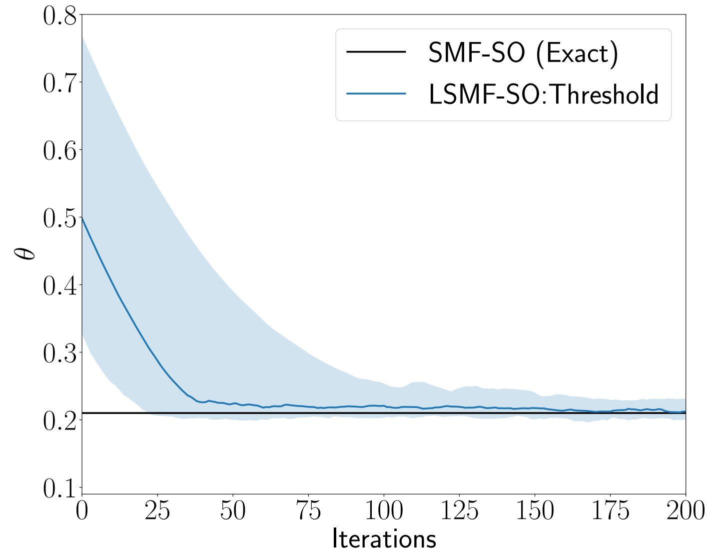


Results: Thresholds

Strategic agents

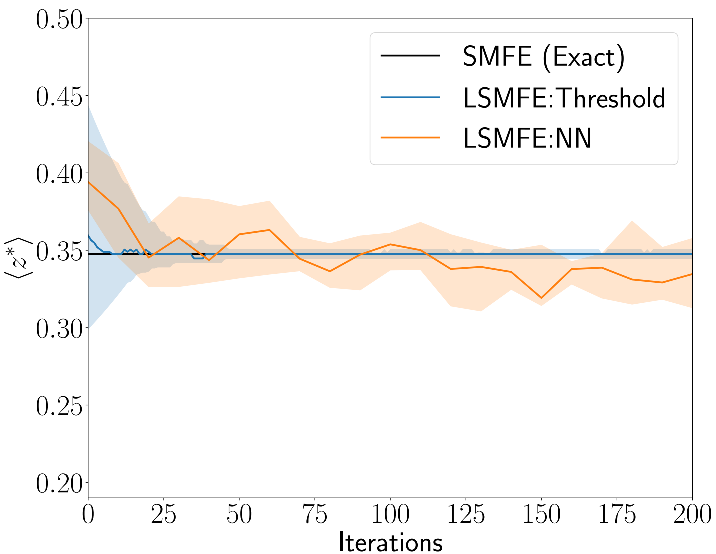


Cooperative agents

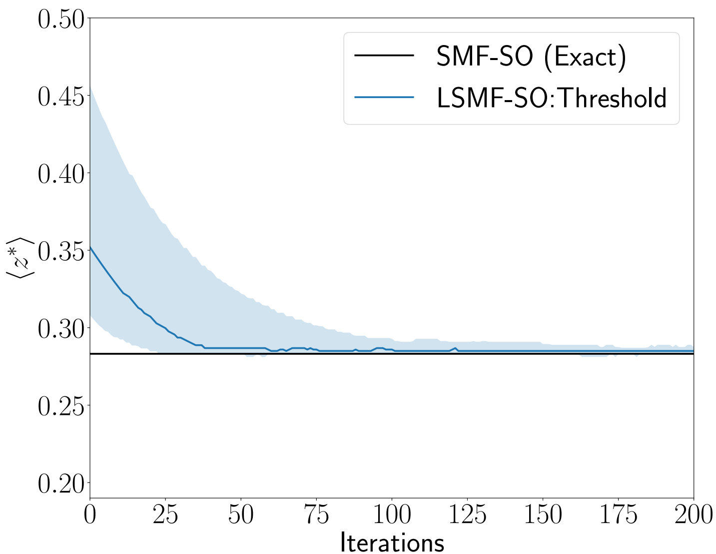


Results: Stationary mean-fields

Strategic agents



Cooperative agents



Example 2: Investment in product quality

Model (adapted from Weintraub, Benkard, Van Roy (2010)

- Models investment decisions of firms in a fragmented market.
- Each firm has p products.
- State space: $[0,1]^p$ (indicating quality of each product)
- Action space: $\{0,1\}^p$ (indicating investment decision in each product)
- Mean-field coupled dynamics and reward models.

Example 2: Investment in product quality

Model (adapted from Weintraub, Benkard, Van Roy (2010)

- Models investment decisions of firms in a fragmented market.
- Each firm has p products.
- State space: $[0, 1]^p$ (indicating quality of each product)
- Action space: $\{0, 1\}^p$ (indicating investment decision in each product)
- Mean-field coupled dynamics and reward models.

Simulation details

- Consider $p = 3$ products.
- Neural networks based policy parameterization.
- Cluster the tails of the trajectories

Example 2: Investment in product quality

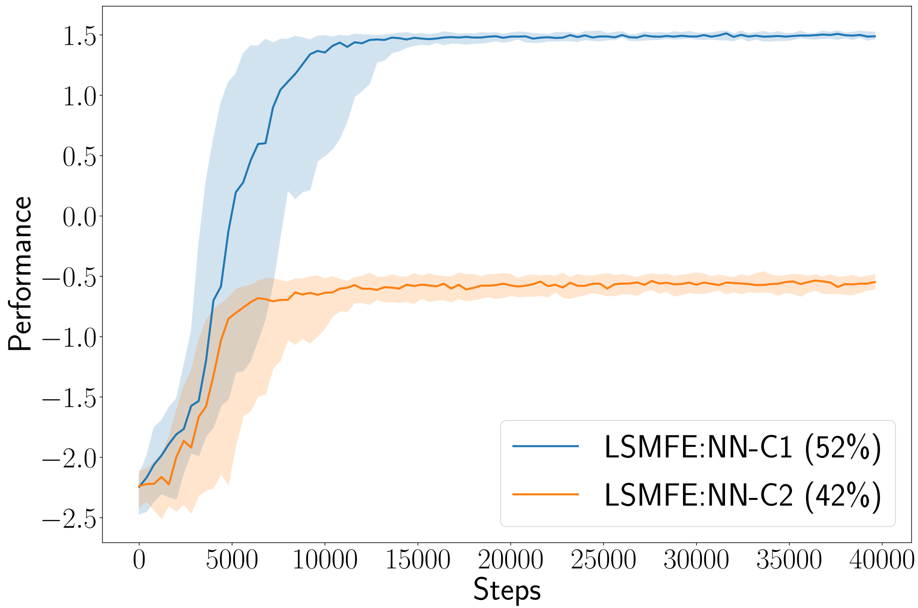
Model (adapte

- Models invest
- Each firm has
- State space:
- Action space:
- Mean-field co

Simulation det

- Consider $p =$
- Neural netwo
- Cluster the ta

Strategic agents



Conclusion

Takeaway message: Learning in large games and teams can be easier than small and medium ones.

Conclusion

Takeaway message: Learning in large games and teams can be easier than small and medium ones.

Stationary Mean-field games

- Provide a different view of looking at mean-field games.
- Arguments easily extend to heterogeneous population (agents with multiple types).

Conclusion

Takeaway message: Learning in large games and teams can be easier than small and medium ones.

Stationary Mean-field games

- Provide a different view of looking at mean-field games.
- Arguments easily extend to heterogeneous population (agents with multiple types).

Comparison with "evolutive" mean-field games

- Both planning and learning solutions have lower complexity than the "evolutive" counterpart.
- But require stronger conditions for existence of equilibrium.

- **email**: aditya.mahajan@mcgill.ca
- **web**: http://cim.mcgill.ca/~adityam

Thank you

Funding

- NSERC Discovery

References

- http://dl.acm.org/citation.cfm?id=3331700