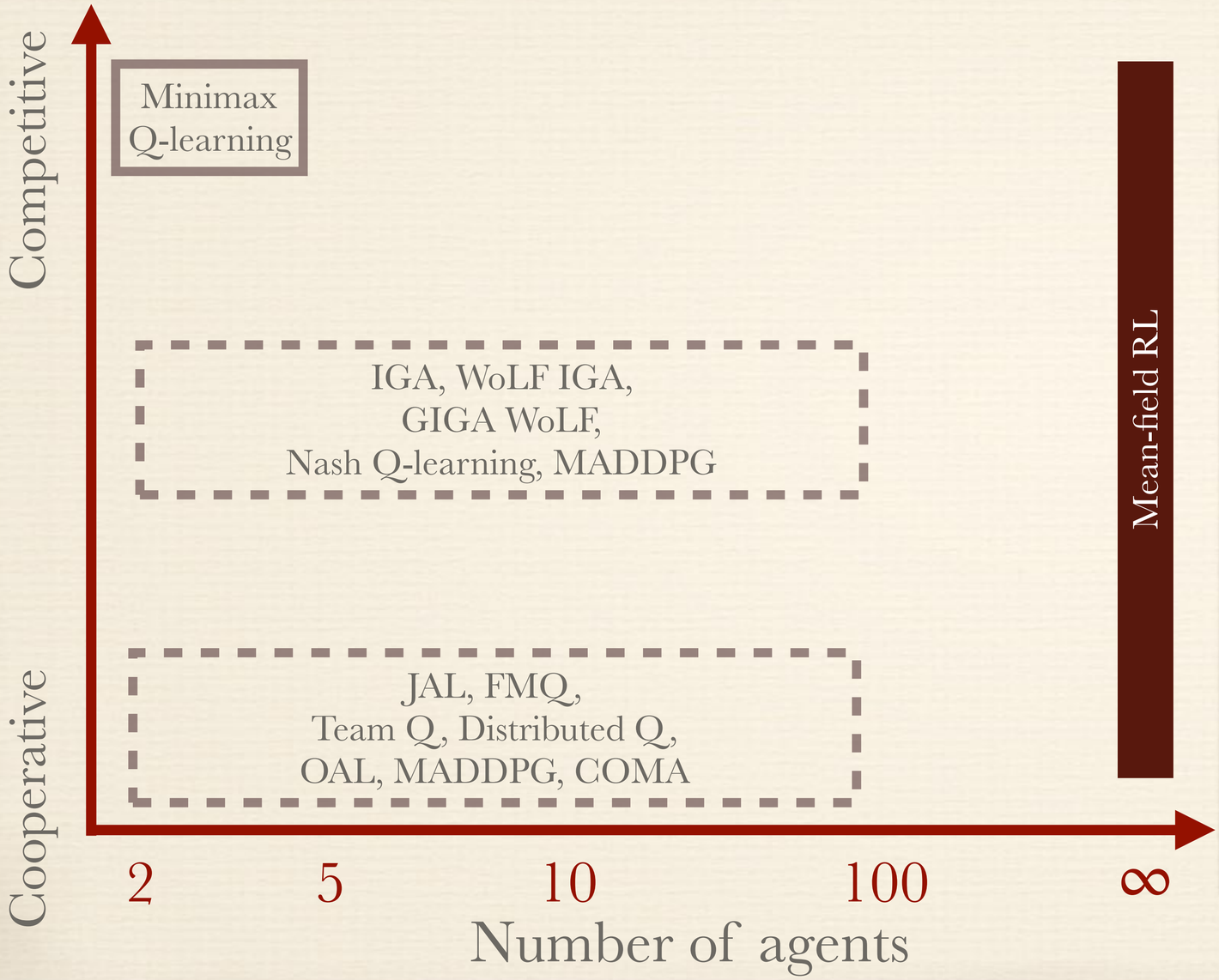
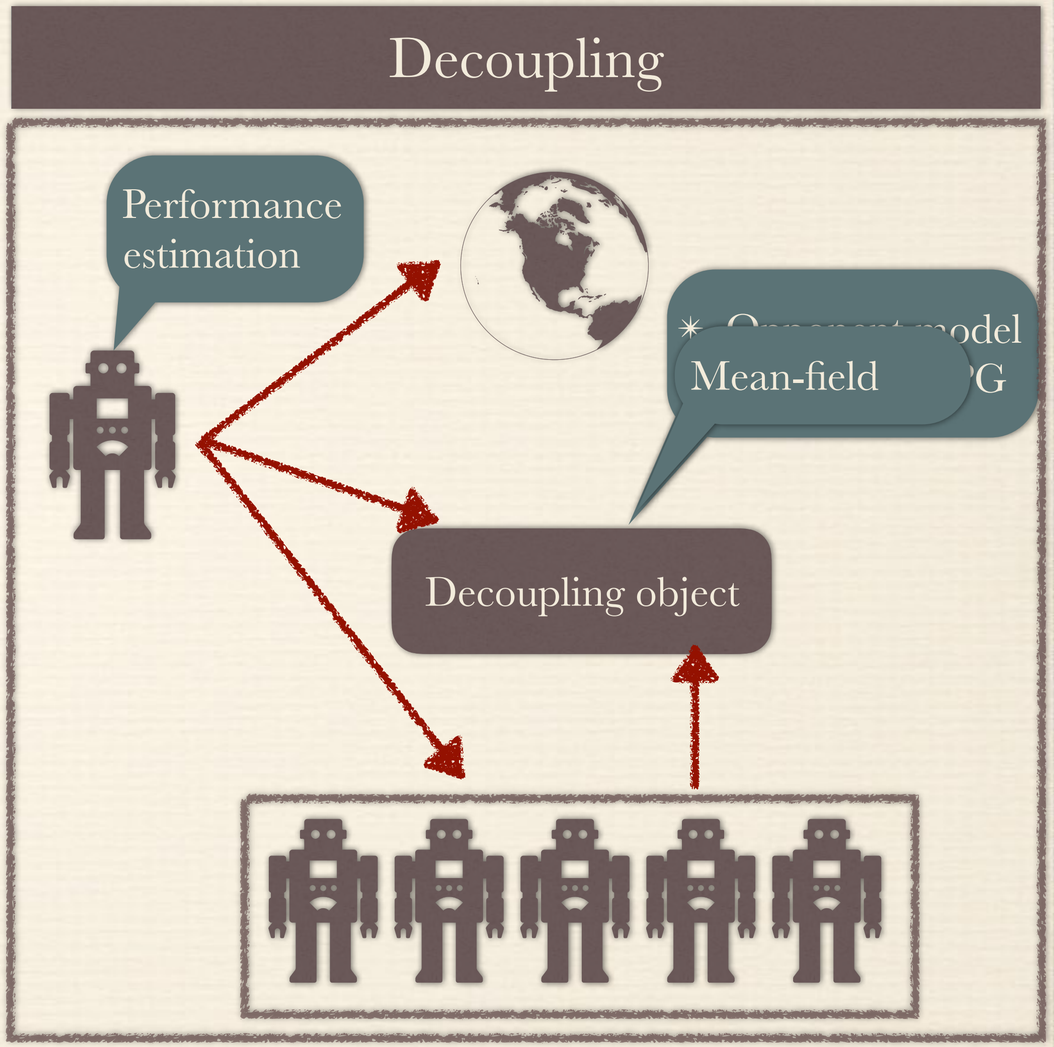
Reinforcement Learning in Stationary Mean-field Games

AAMAS 2019, 15 May 2019, Montreal
Jayakumar Subramanian and Aditya Mahajan
ECE, McGill University

MARL landscape





Mean-field game model
[Weintraub et al. (2005), Huang et al. (2006), Lasry & Lions (2007)]

- Inspired by the mean-field theory in statistical physics
- Large number of agents $N = \{1,2,\ldots,n\}$
- Agents divided into homogeneous sub-populations [Regularity] $X_t^i \in \mathcal{X}, A_t^i \in \mathcal{A}$
- Agents are anonymous [Symmetry]
- Each agent is ‘small’ or ‘inconsequential’ [Infinite population limit]
- Population of agent replaced by their empirical distribution of states and actions

Our work

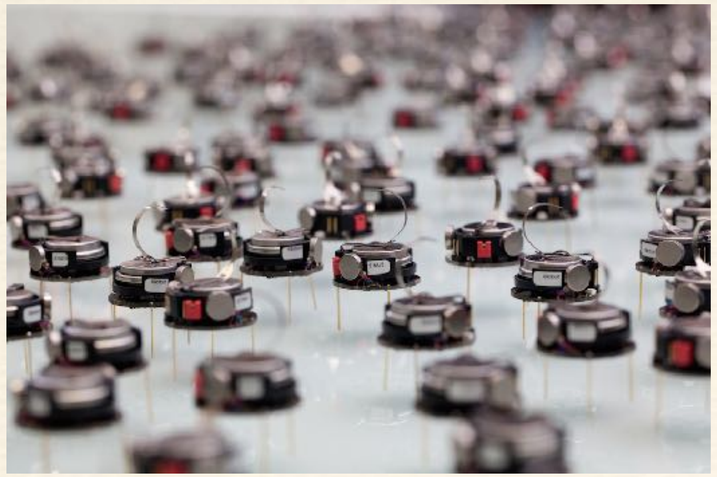
We consider:

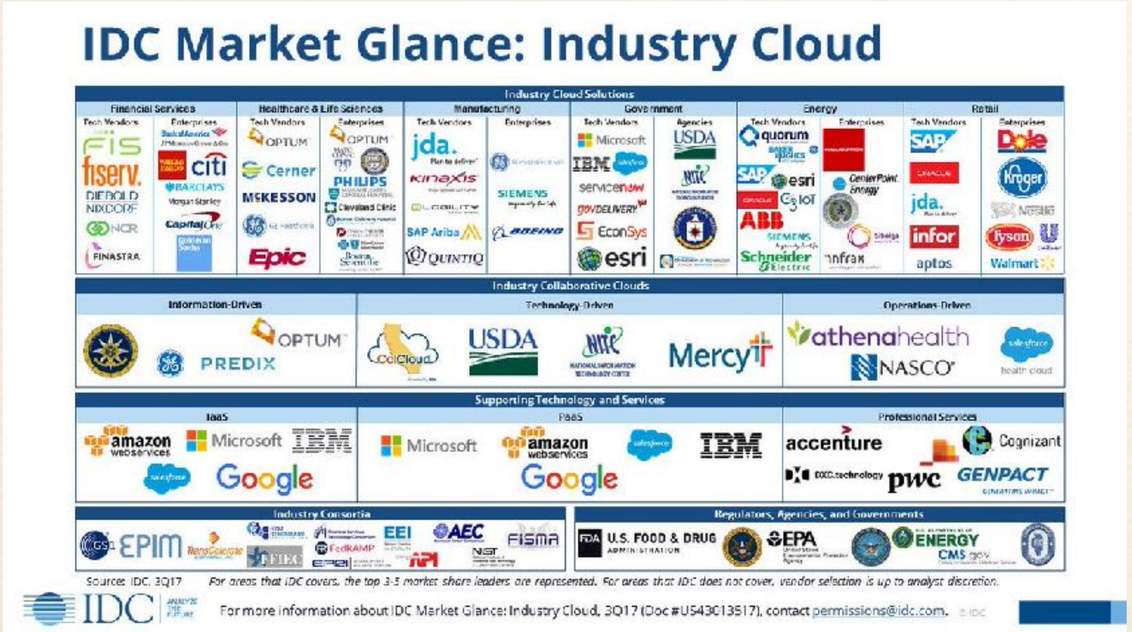
- Stationary mean-field game model [Weintraub, Benkard and Van Roy (2005)]
- Two cases: non-cooperative and cooperative

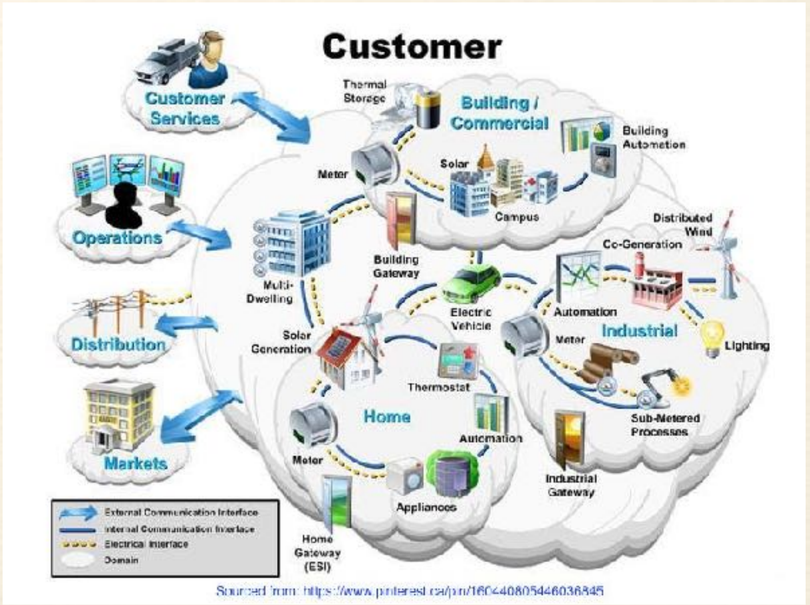
and propose:

- Bounded rationality based local solution concepts for both cases
- RL algorithms to determine these solution concepts for both cases

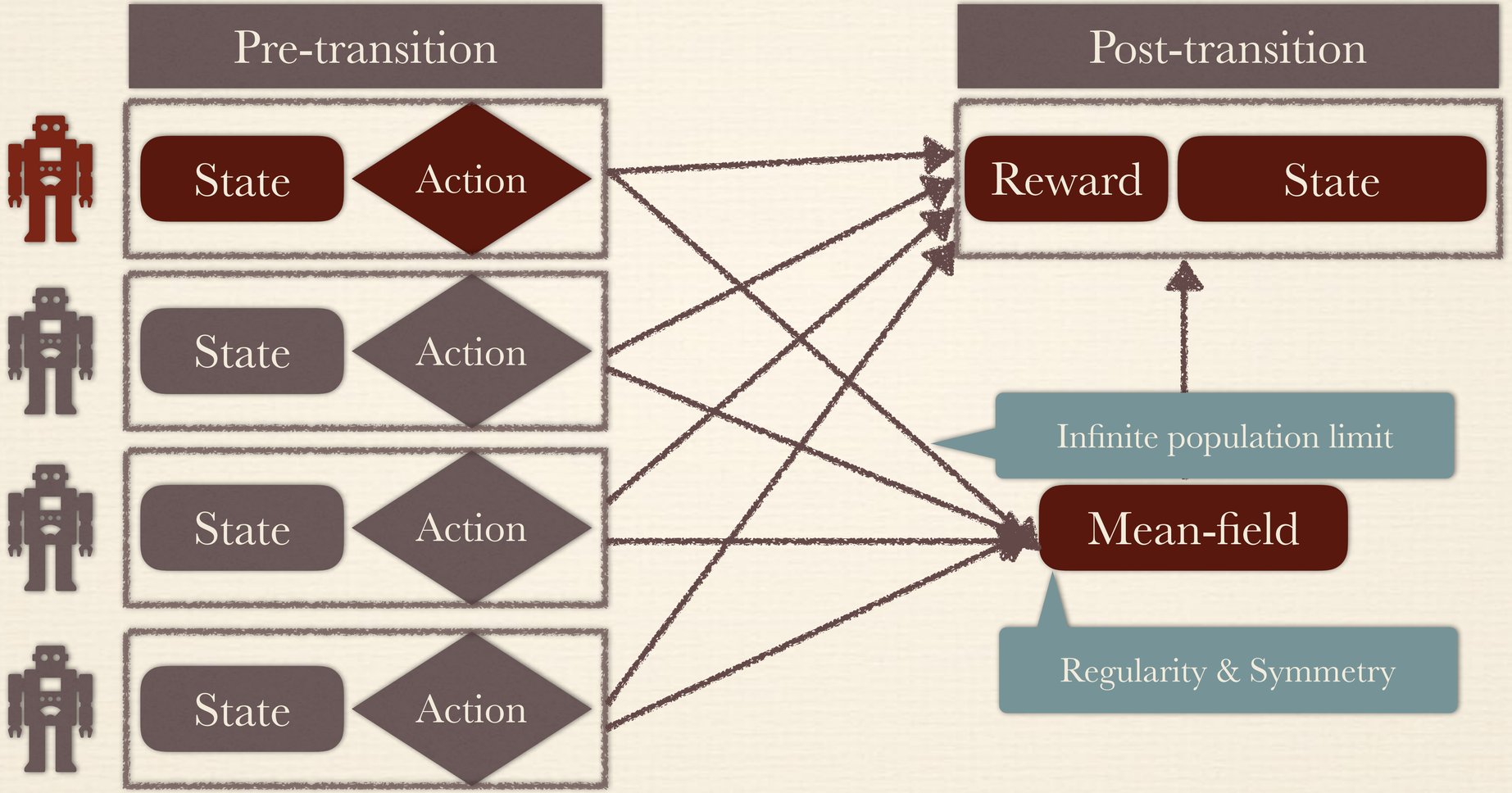
Applications







Mean-field game model



Mean-field games (MFG)

Mean-field

$$Z_t(x,a) = \frac{1}{n}\sum_{i\in N} \mathbf{1}\{X_t^i = x, A_t^i = a\}, \forall (x,a) \in \mathcal{X}\times\mathcal{A}$$

Mean-field games [Huang et al. (2006), Lasry and Lions (2007)]

$$\mathbb{P}(X_{t+1}^i \,|\, \mathbf{X}_t = \mathbf{x}_t, \mathbf{A}_t = \mathbf{a}_t)$$

coupled only through mean-field

$$= \mathbb{P}(X_{t+1}^i \,|\, X_t^i = x_t^i, A_t^i = a_t^i, \mathbf{Z_t = \xi(\mathbf{x}_t, \mathbf{a}_t)})$$

$$\mathbb{E}[R_t^i \,|\, \mathbf{X}_t = \mathbf{x}_t, \mathbf{A}_t = \mathbf{a}_t]$$

coupled only through mean-field

$$= \mathbb{E}[R_t^i \,|\, X_t^i = x_t^i, A_t^i = a_t^i, \mathbf{Z_t = \xi(\mathbf{x}_t, \mathbf{a}_t)}]$$

Features in MFG

- Infinite population limit: empirical mean-field converges to statistical mean-field.

$$Z_t(x) = \frac{1}{n}\sum_{i \in N} \mathbf{1}\{X_t^i = x\} \approx \frac{1}{n}\sum_{i \in N} \mathbb{P}(X_t^i = x), \forall (x) \in \mathcal{X}\ .$$

- All agents follow identical time varying Markov (stochastic) policies:

$$\pi_t : \mathcal{X} \to \Delta(\mathcal{A})$$

- Statistical mean-field evolution (McKean-Vlasov):

$$Z_{t+1}(y) = \sum_{x \in \mathcal{X}} \sum_{a \in \mathcal{A}} Z_t(x)\pi_t(a \mid x)P(y \mid x, a, Z_t) = \Phi(Z_t, \pi_t), \forall y \in \mathcal{X}\ .$$

MFG-RL landscape

- Model based (Stochastic adaptive control): Kizilkale and Caines (2012)
- Q-learning for MFG (periodic solutions): Yin et al., (2014)
- Potential mean-field games: Mguni et al., (2018)
- Inverse MFG RL: Yang et al., (2018)
- MFG RL using mean-field of actions: Yang et al., (2018)

Stationary MFG (SMFG)

Additional assumptions over MFG

1. Time homogeneous policy:

$$\pi : \mathcal{X} \to \Delta(\mathcal{A})$$

2. Stationary mean-field:

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

3. Performance evaluation:

$$V^i_{\pi,z}(x) = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(X^i_t, A^i_t, z, X^i_{t+1}) \,|\, X^i_0 = x]$$

SMFG solution concept: non-cooperative

Stationary mean-field equilibrium (SMFE)

A SMFE is a pair of policy and mean-field (π, z) which satisfies:

Sequential rationality	$V^i_{\pi,z}(x) \geq V^i_{\pi',z}(x), \forall \pi', \forall x \in \mathcal{X}$.
Consistency	$z = \Phi(z, \pi)$.

Bounded rationality

Global nature	Solution concepts require global search over policies	Local search in a parametrised space
Curse of dimensionality	Verification requires computation of value functions that suffer from the curse of dimensionality	Use of function approximation

Local SMFE

Local stationary mean-field equilibrium (LSMFE)

A LSMFE is a pair of policy and mean-field (π_θ, z) which satisfies:

Sequential rationality	$\frac{\partial J_{\theta,z}}{\partial \theta} = 0$, where $J_{\theta,z} = \mathbb{E}_{X\sim\xi_0}[V_{\pi_\theta,z}(X)]$
Consistency	$z = \Phi(z, \pi_\theta)$.

Global vs. local solution concepts

- MFG: Unique SMFE $\not\Rightarrow$ Unique LSMFE

Sufficient conditions for unique LSMFE (that agrees with SMFE)

1. **SMFE is unique**
2. $J_{\theta,z}$ is concave in $\theta, \forall z$

RL problem formulation

Initial state distribution: $X_0 \sim \xi_0 \in \Delta(\mathcal{X})$

Parametrized policy: $\pi_\theta : \mathcal{X} \to \Delta(\mathcal{A})$

Non-cooperative game objective: $J^i_{\theta,z} = \mathbb{E}^{\pi_\theta}[\sum_{t=0}^{\infty} \gamma^t R^i_t \,|\, z]$

LSMFE algorithm

Two-timescale stochastic gradient ascent [Borkar, 1997]

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

2. $\theta_{k+1} = [\theta_k + \alpha_k[G_{\theta_k, z_k}]_\Theta$

$\hat{\Phi}(z, \pi)$ is an unbiased estimator of $\Phi(z, \pi)$

$G_{\theta,z}$ is an unbiased estimator of $\dfrac{\partial J_{\pi_\theta, z}}{\partial \theta}$

$$\sum_k \alpha_k = \infty, \sum_k \beta_k = \infty, \sum_k (\alpha_k^2 + \beta_k^2) < \infty$$

$$\lim_{k\to\infty} \alpha_k = 0, \lim_{k\to\infty} \beta_k = 0, \lim_{k\to\infty} \alpha_k/\beta_k = 0$$

Learning rate conditions

Estimation of $\Phi(z, \pi_\theta)$ and $G_{\theta,z}$.

$\Phi(z, \pi_\theta)$ is estimated in a particle-filter like approach

Policy gradient estimation

Likelihood ratio based

$$\partial J_{\pi_\theta,z} = \mathbb{E}[\sum_{t=0}^{\infty} \nabla_\theta \log[\pi_\theta(A_t \mid X_t)] V_{\pi_\theta,z}(X_t) \mid X_0 \sim \xi_0]$$

Simultaneous perturbation based

$$G_{\theta,z} = \frac{\eta}{2c}(J_{\pi^+,z} - J_{\pi^-,z}),$$

$\eta_i \sim$ Rademacher(± 1) : *SPSA*;

$\eta_i \sim$ Normal(0,I) : *SFSA* .

Convergence

Proposition: If the following conditions are satisfied:

1. $\Phi(z, \pi_\theta)$, $\partial J_{\pi_\theta,z}/\partial\theta$ are Lispschitz in θ, z.
2. $\hat{\Phi}(z, \pi_\theta)$, $G_{\theta,z}$ are unbiased estimators of $\Phi(z, \pi_\theta)$, $\partial J_{\pi_\theta,z}/\partial\theta$.
3. Estimation error: $G_{\theta,z} - \partial J_{\pi_\theta,z}/\partial\theta$ has bounded variance.
4. $\dot{z} = \Phi(z, \pi_\theta) - z$ has a unique globally asymptotially stable eq. point: $f(\theta), \forall\theta$
5. $f(\theta)$ is Lipschitz in θ.

Then, almost surely:

1. $\|z_n - f(\theta_n)\| \to 0$ as $n \to \infty$
2. $\{\theta_n\} \to$ asymptotic pseudotrajectory of semiflow induced $\dot{\theta} = \partial J_{\pi_\theta,z}/\partial\theta$.
3. Iteration converges to a LSMFE.

Comparison between SMFG solution concepts

Fixed mean-field (unilateral change)

Sequential rationality

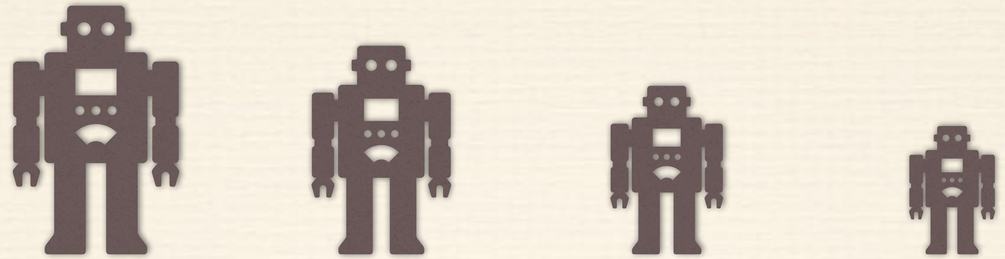
$$V^i_{\pi,z}(x) \geq V^i_{\pi',z}(x), \forall \pi', \forall x \in \mathcal{X}$$

Optimality

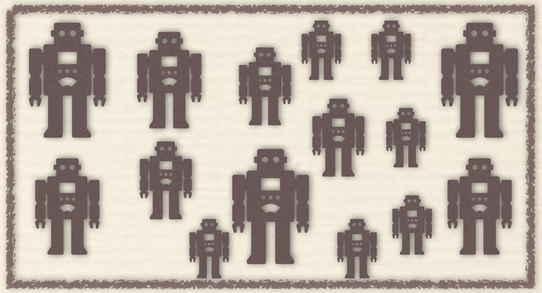
$$V^i_{\pi,z}(x) \geq V^i_{\pi',z'}(x), \forall \pi', \forall x \in \mathcal{X}$$

Policy dependent mean-field (concurrent change)

Numerical example: Malware spread



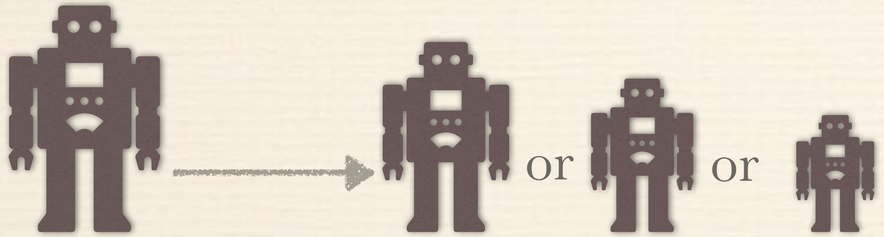
Healthy (0) → Non-healthy (1)

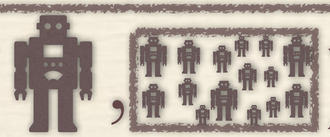


Mean-field

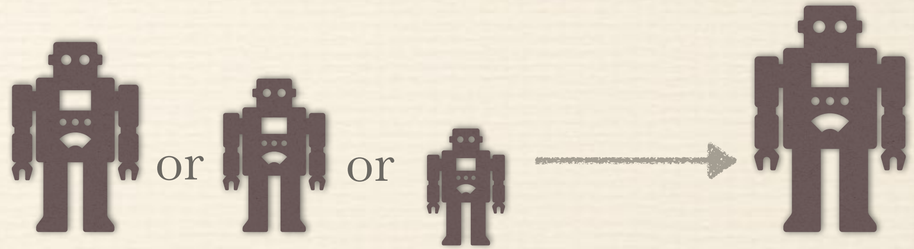
Transition & Reward

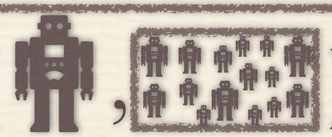
Action - 0 (Do nothing)

State: 

Reward: r ()

Action - 1 (Repair)

State: 

Reward: r () + Repair

Numerical example: Product Investments

Investment decisions of firms in a fragmented market

- Each agent (firm) produces n_p products
- State of each firm: $X_t^{i,j} \in [0,1], j \in \{1,\ldots,n_p\}$ (normalized product quality)
- Action — investment decision: $\in \{0,1\}^{n_p}$

Transition

$$X_{t+1}^{i,j} = \begin{cases} \omega_t(1 - X_t^{i,j}), & \text{if } \langle Z^j \rangle < q \ \& \ A_t^{i,j} = 1, \\ 0.5\omega_t(1 - X_t^{i,j}), & \text{if } \langle Z^j \rangle \geq q \ \& \ A_t^{i,j} = 1, \\ X_t^{i,j}, & \text{if } A_t^{i,j} = 0, \end{cases}$$

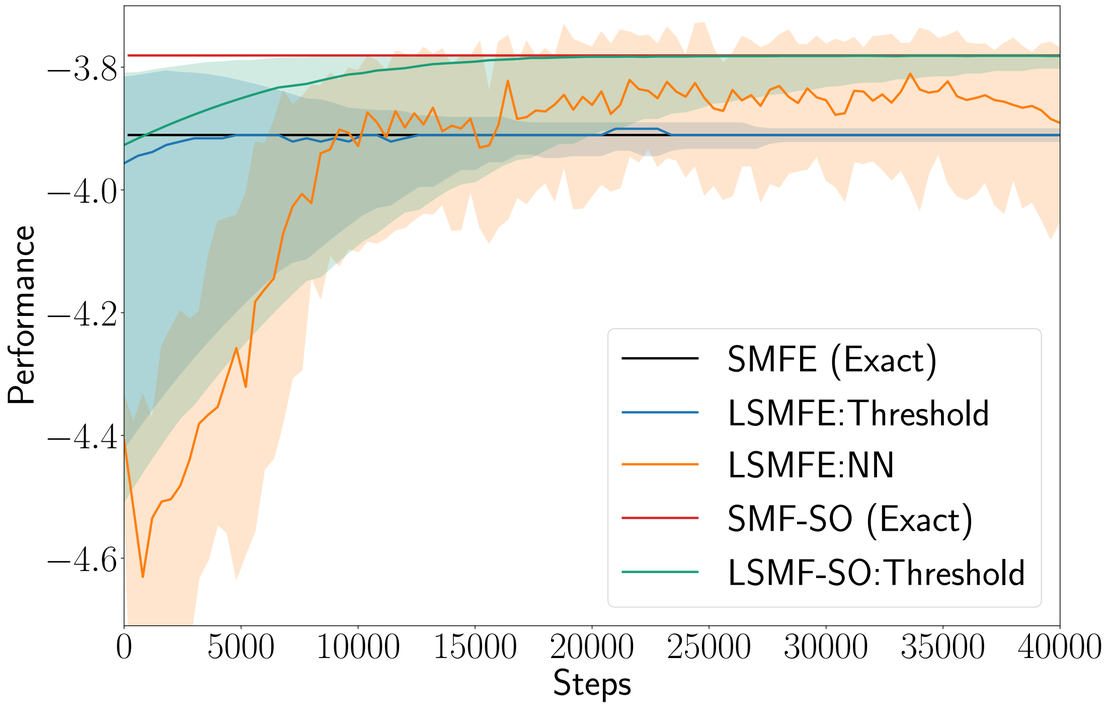
ω_t: [0,1] -valued i.i.d.process with probability density f, and $\langle Z_t^j \rangle$: mean of Z_t^j

Reward

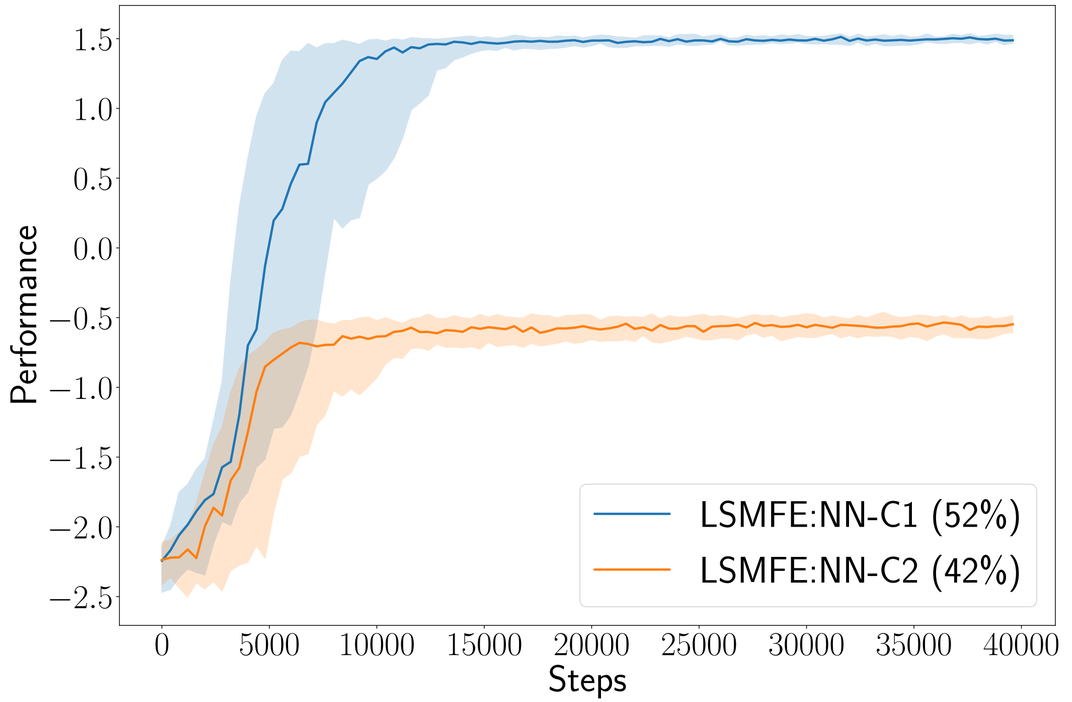
$$r(X_t^i, A_t^i, Z_t^i) = \sum_{j=1}^{n_p} \left[d^j X_t^{i,j} - c^j \langle Z_t^j \rangle - \lambda A_t^{i,j} \right]$$

Results

Malware spread



Investment example



Summary

Presented use of MFG framework for MARL

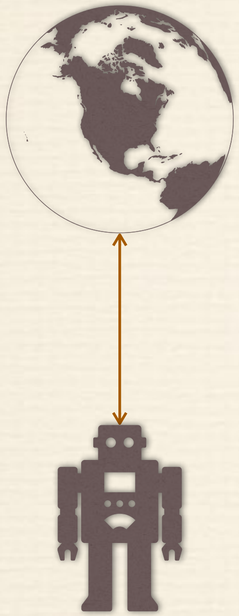
Proposed two local solution concepts for Stationary MFG

Proposed two RL algorithms for these solution concepts & presented their convergence conditions

Thanks

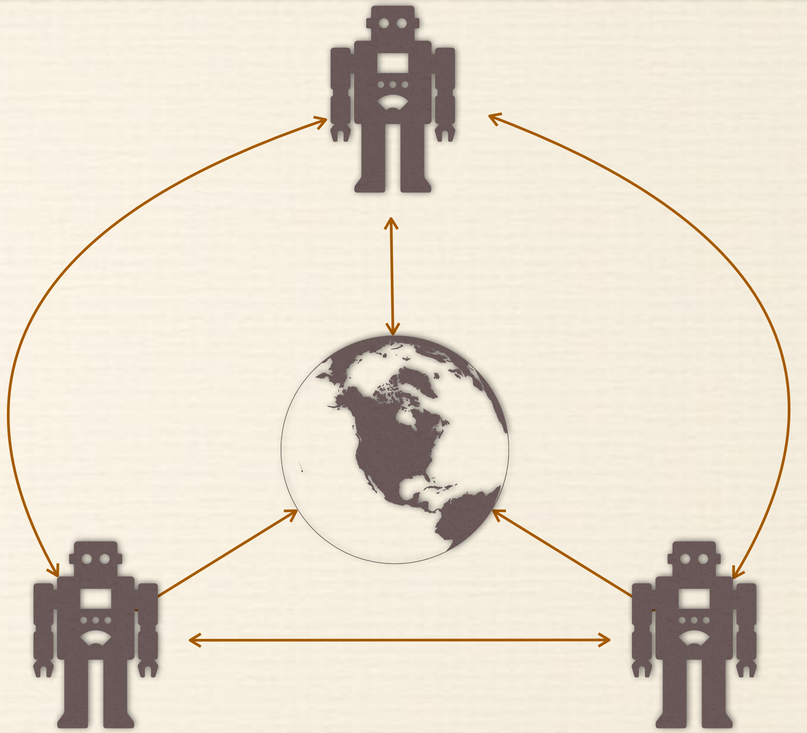
Background

RL



Policy that maximizes return over a horizon

MARL



Policy that maximizes return over a horizon simultaneously for all agents

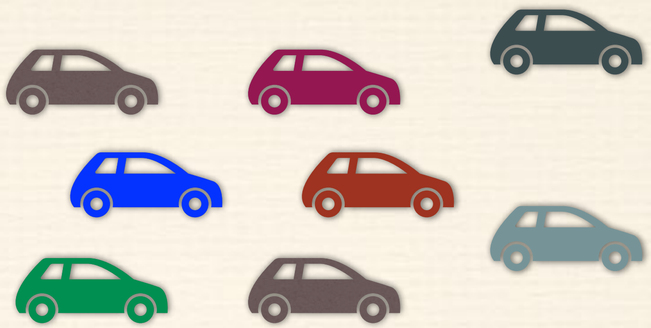
MARL vs. RL - an example

Self-driving car

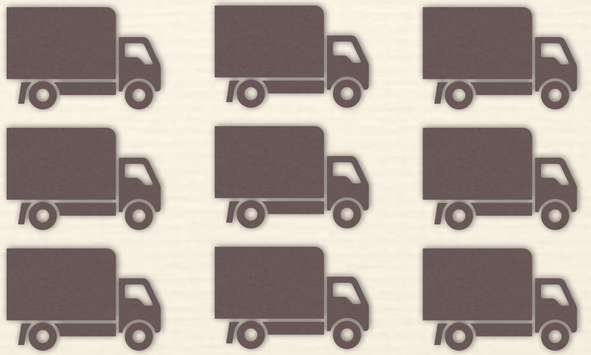
Non-cooperative

Cooperative

Multiple self-driving cars



Platoon of trucks



(Game theory)

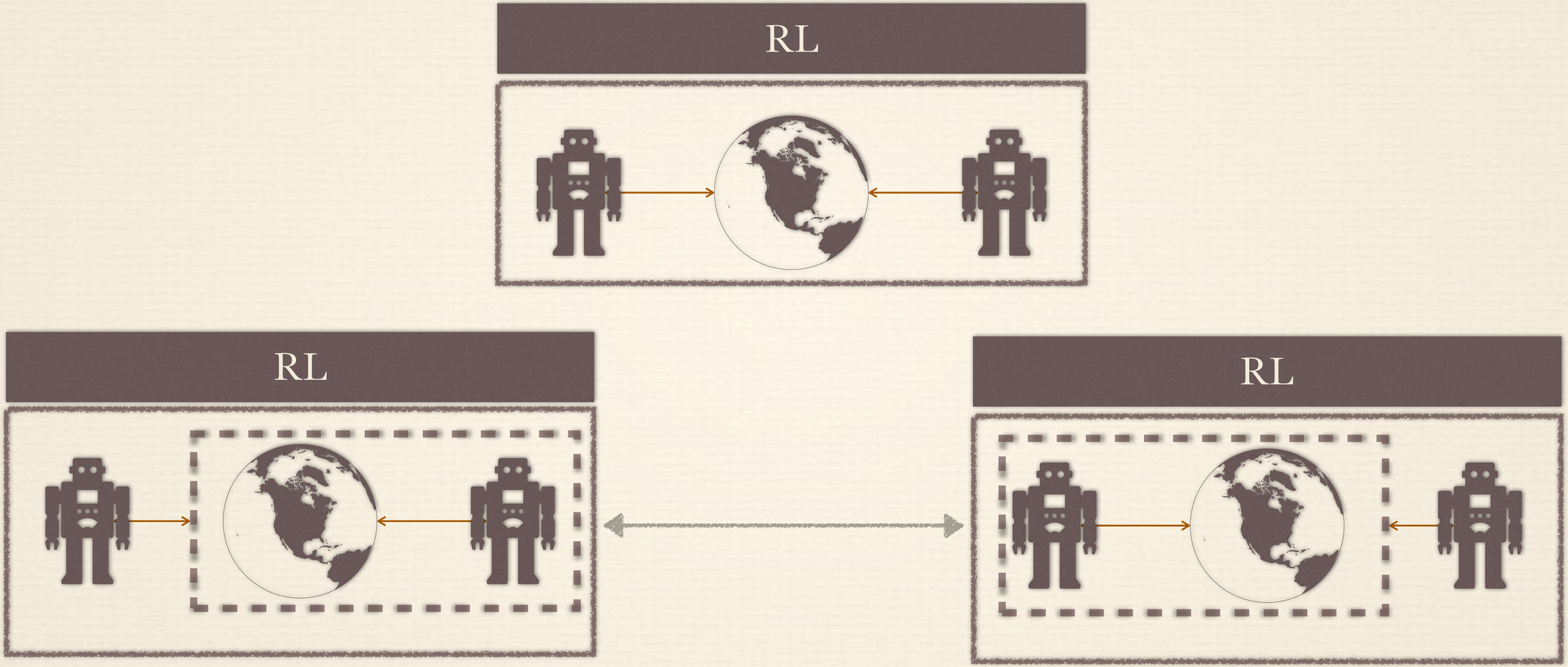
Nash equilibrium

Vector objective

Social welfare optimal

(Team theory)

Non-stationarity



Regularity & Symmetry

Regularity

All agents have the same state and action spaces

Symmetry

$$\mathbb{P}(\mathbf{X}_{t+1} = \mathbf{x}_{t+1} \,|\, \mathbf{X}_t = \mathbf{x}_t, \mathbf{A}_t = \mathbf{a}_t) = \mathbb{P}(\mathbf{X}_{t+1} = \mathbf{x}_{t+1} \,|\, \mathbf{X}_t = \sigma\mathbf{x}_t, \mathbf{A}_t = \sigma\mathbf{a}_t)$$

$$\mathbb{E}[\mathbf{R}_t \,|\, \mathbf{X}_t = \mathbf{x}_t, \mathbf{A}_t = \mathbf{a}_t] = \mathbb{E}[\mathbf{R}_t \,|\, \mathbf{X}_t = \sigma\mathbf{x}_t, \mathbf{A}_t = \sigma\mathbf{a}_t]$$

MFG solution concept: non-cooperative

Mean-field equilibrium (MFE)

A MFE is a pair of policy sequence and mean-field sequence $(\pi_{1:t}, z_{1:t})$ which satisfies:

Sequential rationality	$V^i_{t,\pi_{1:t},z_{1:t}}(x) \geq V^i_{t,\pi'_{1:t},z_{1:t}}(x), \forall \pi'_{1:t}, \forall x \in \mathcal{X}$.
Consistency	$z_{t+1} = \Phi(\pi_t, z_t), \forall t$.

LSMFE Proof outline

1. Image space of Φ is bounded $\Rightarrow$ estimation error $\hat{\Phi}(z, \pi) - \Phi(z, \pi)$ is bounded.

2. Learning rate conditions + conditions in proposition + Point 1. $\Rightarrow$
Conditions of Theorem 6 in Kushner and Yin (2003) are satisfied.

3. Iteration converges to $\partial J_{\pi_{\theta^*}, z^*} / \partial \theta = 0$ and $z^* = f(\theta^*) = \Phi(z^*, \theta^*)$.

Iteration converges to a LSMFE