Dynamic Games with Networked Players and Learning Algorithms
--A Tutorial--

Tamer Başar
University of Illinois Urbana-Champaign

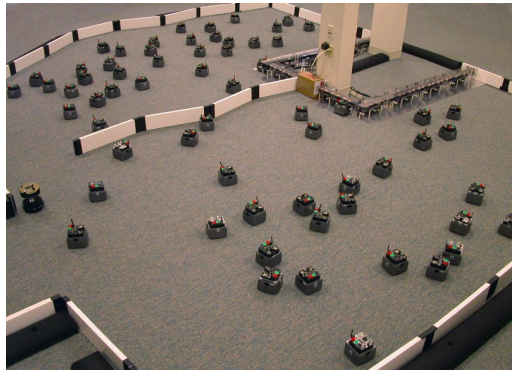
20th ISDGA, Valladolid, Spain
July 10, 2024

Outline

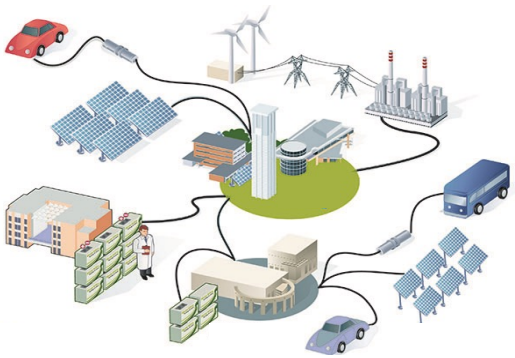
- A general introduction to multi-agent dynamical systems
- Appropriateness of the framework of (stochastic) dynamic games and underlying challenges arising due to informational asymmetry
- Hierarchy in decision making and strategic information transmission
- The role of mean-field games (MFGs) to alleviate the challenges in the high population regime, and the associated solution concept of mean-field equilibrium (MFE)
- Computation of MFE along with learning schemes
- Digression: A brief intro to RL and actor-critic algorithms for single and multi-agent systems
- Zero-order stochastic optimization (ZSO) based RL and finite sample guarantees
- Illustration through multi-population LQ-MFGs—*consensus and dissensus*
- What lies in the future

Multi-agent Dynamical Systems

- Multi-agent systems (MASs) are ubiquitous



Robotics



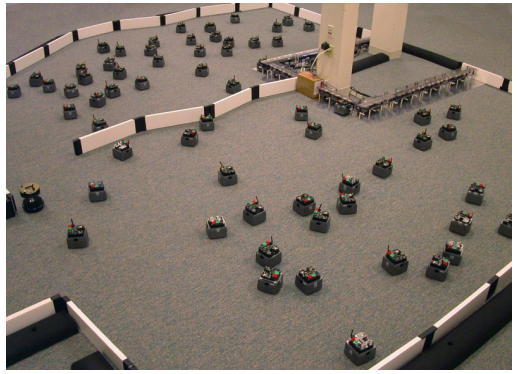
Smart Grid

Unmanned Aerial Vehicles

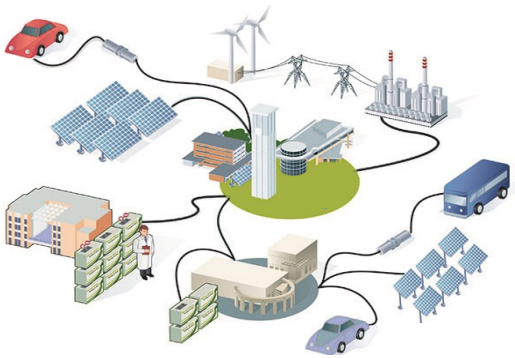
MOBA Video Games

Multi-agent Dynamical Systems

- Multi-agent systems (MAS) are ubiquitous



Robotics



Smart Grid

Unmanned Aerial Vehicles

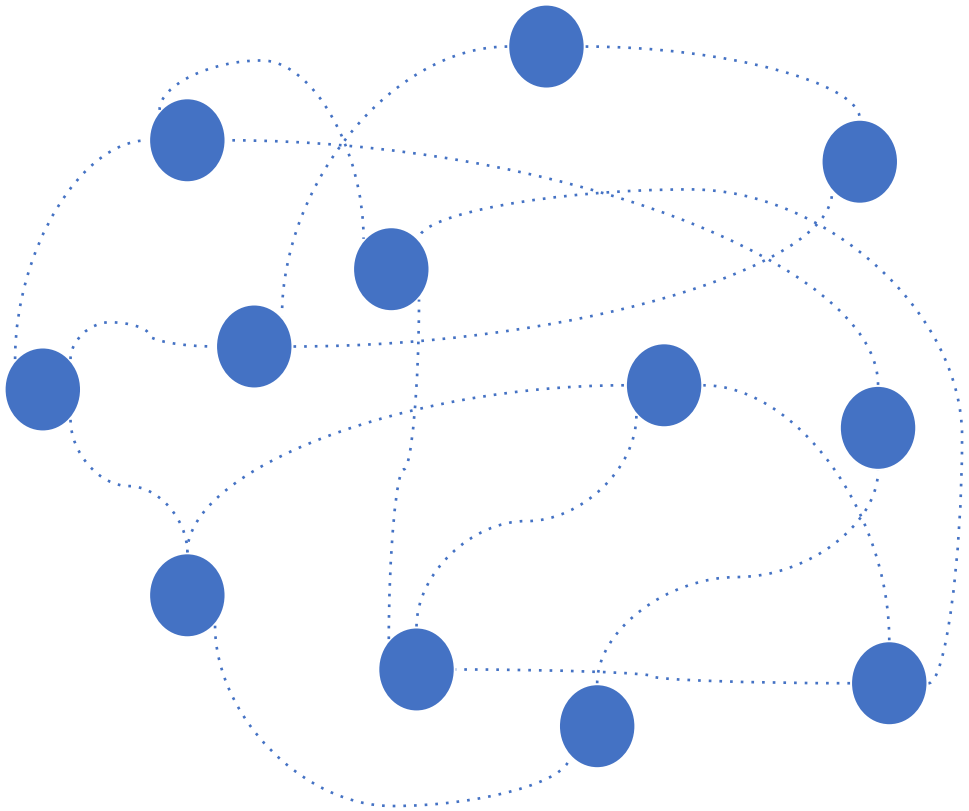
MOBA Video Games

Other selected applications:

- Mobile sensor networks
- Distributed optimization (with topological and informational constraints)
- Social networks (evolution of opinions)

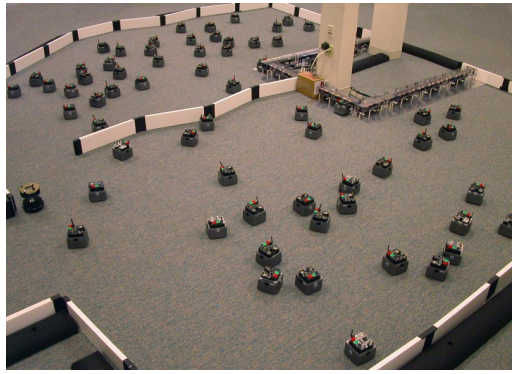
Opinion dynamics leading to consensus/dissensus

Nodes representing opinions held by different agents, who interact with their neighbors and update their opinions, either in line with others' (leading to consensus) OR maintaining a distance with others' (leading to dissensus)

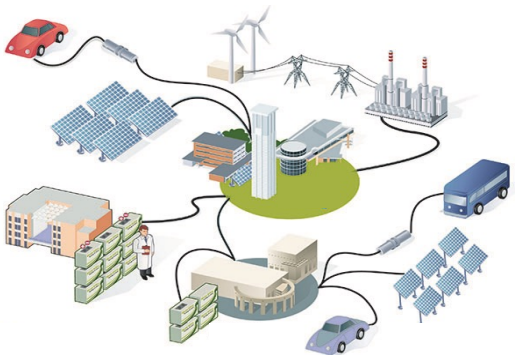


Multi-agent Dynamical Systems

- Multi-agent systems (MAS) are ubiquitous



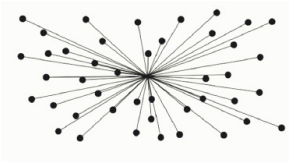
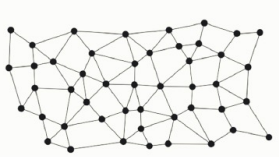
Robotics



Smart Grid

Unmanned Aerial Vehicles

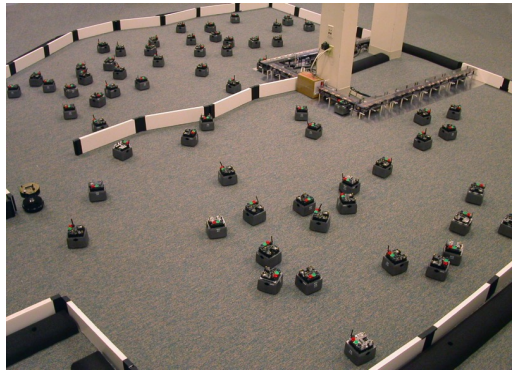
MOBA Video Games

- Centralized protocol  vs Decentralized protocol 

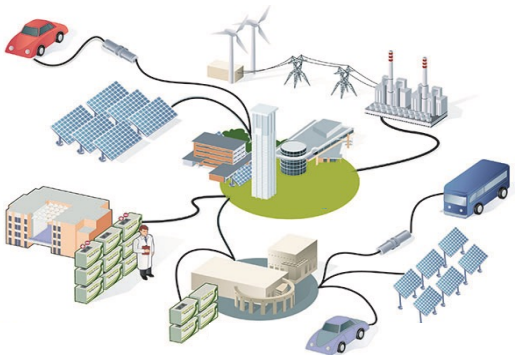
  - Central controller may not exist or may not be desirable in many MAS applications
  - Advantages of decentralization: i) resilient to attacks; ii) scalable; iii) privacy-preserving; iv) use of only local information

Multi-agent Dynamical Systems

- Multi-agent systems (MASs) are ubiquitous



Robotics



Smart Grid

Unmanned Aerial Vehicles

MOBA Video Games

Advantages of decentralization also bring along several challenges because of interactions of multiple agents under informational asymmetry and misalignment of objectives, and the need to **learn** for performance improvement in a nonstationary environment (using e.g., the machinery of reinforcement learning).[1,2,3]

****

[1]K. Zhang, Z. Yang, TB, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," *Handbook of Reinforcement Learning and Control,* Studies in Systems, Decision and Control 325, Springer Nature, 2021, pp. 321-384.

[2]K. Zhang, Z. Yang, TB, "Decentralized multi-agent reinforcement learning with networked agents: Recent advances," *Frontiers of Information Technology & Electronic Engineering and Control,* 22(6):802-814, 2021.

[3]K. Zhang, Z. Yang, H. Liu, T. Zhang, TB, "Finite-sample analysis for decentralized batch multi-agent RL with networked agents," IEEE TAC, 69(12):5925-5940, Dec. 2021

Toward a dynamic game-theoretic setting

An appropriate framework for a systematic study of such multi-agent dynamical systems in an uncertain environment, with informational and possibly resource constraints, with robustness considerations built in, and with generally different objectives by the agents is provided by *stochastic noncooperative dynamic game theory*.

Game theory as a modeling and computational framework[1,2]

- Game theory provides the right modeling and computational framework to capture interactions among multiple interacting agents/players/actors (physical, economic, social, and even biological) with possibly misaligned objectives (zero-sum or nonzero-sum).
- Dynamic game theory provides a richer framework capturing evolution of these interactions over time, where information structures (who knows what, and when), memory restrictions (how deep into the past do agents recall), resource constraints, their allocation and utilization over time, and tradeoffs between short-term and long-term goals play important roles.

****

[1]TB, G.J. Olsder, *Dynamic Noncooperative Game Theory*, SIAM, 1999.

[2]TB, G. Zaccour, *Handbook of Dynamic Game Theory, Vols I & II*, Springer, 2018

Game theory as a modeling and computational framework

- Game theory provides the right modeling and computational framework to capture interactions among multiple interacting agents/players/actors (physical, economic, social, and even biological) with possibly misaligned objectives (zero-sum or nonzero-sum).
- Dynamic game theory provides a richer framework capturing evolution of these interactions over time, where information structures (who knows what, and when), memory restrictions (how deep into the past do agents recall), resource constraints, their allocation and utilization over time, and tradeoffs between short-term and long-term goals play important roles.
- Multiple solution concepts exist (Nash, Stackelberg, Markov perfect equilibrium, etc.) tailored to the scenario at hand, and the roles of different players in the decision-making process (symmetric, hierarchical, etc.)
- With infinite population of players, interactions among players take a different meaning, leading to mean-field games and the associated solution concept of mean-field equilibrium.

General NZS SDGs with networked agents

- $x_t = (x_{1,t}, \ldots, x_{N,t})$; $x_{i,(t+1)} = f_{it}(x_{it}, u_{it}, c_{it}(x_{-i,t}, u_{-i,t}), w_{i,t})$, $i = 1, \ldots, N$ (state dynamics)

  an underlying network that governs connections, neighborhood interactions ($\mathcal{N}_{i,t}$ for i)

General NZS SDGs with networked agents

- $x_t = (x_{1,t}, \ldots, x_{N,t})$; $x_{i,(t+1)} = f_{it}(x_{it}, u_{it}, c_{it}(x_{-i,t}, u_{-i,t}), w_{i,t})$, $i = 1, \ldots, N$ (state dynamics)

  an underlying network that governs connections, neighborhood interactions ($\mathcal{N}_{i,t}$ for i)

  E.g. $c_{it}(x_{-i,t}, u_{-i,t}) = c_{it}(x_{j,t}, u_{j,t}; j \in \mathcal{N}_{i,t})$

  Or even, $c_{it}(x_{-i,t}, u_{-i,t}) = (1/|\mathcal{N}_{i,t}|) \sum_{j \in \mathcal{N}(i,t)} x_{j,t}$ (*average of states of all players neighbor to* i)

General NZS SDGs with networked agents

- $x_t = (x_{1,t}, \ldots, x_{N,t})$; $x_{i,(t+1)} = f_{it}(x_{it}, u_{it}, c_{it}(x_{-i,t}, u_{-i,t}), w_{i,t})$, $i = 1, \ldots, N$ (state dynamics)

  an underlying network that governs connections, neighborhood interactions ($\mathcal{N}_{i,t}$ for i)

- *Information structures* (control policy of player i: $\gamma_i \in \Gamma_i$)

  Closed-loop perfect state: $u_{i,t} = \gamma_i(t; x_s, s=1, \ldots, t)$

  Partial (local) state: $u_{i,t} = \gamma_i(t; x_{i,s}, s=1, \ldots, t)$

  Measurement feedback: $u_{i,t} = \gamma_i(t; y_{i,s}, s=1, \ldots, t)$, $y_{i,t} = h_{i,t}(x_{i,t}, x_{-i,t}, v_{i,t})$

  AND many others (s.a. failing links and channels, costs on access and usage of information; memory restrictions, …)*

---

*S. Aggarwal, TB, D. Maity, "Linear quadratic zero-sum differential games with intermittent and costly sensing," *IEEE Control Systems Letters* (L-CSS), June 2024.

General NZS SDGs with networked agents

- $x_t = (x_{1,t}, \ldots, x_{N,t})$; $x_{i,(t+1)} = f_{it}(x_{it}, u_{it}, c_{it}(x_{-i,t}, u_{-i,t}), w_{i,t})$, $i = 1, \ldots, N$ (state dynamics)

  an underlying network that governs connections, neighborhood interactions ($\mathcal{N}_{i,t}$ for i)

- *Information structures* (control policy of player i: $\gamma_i \in \Gamma_i$)

  Closed-loop perfect state: $u_{i,t} = \gamma_i(t; x_s, s=1, \ldots, t)$

  Partial (local) state: $u_{i,t} = \gamma_i(t; x_{i,s}, s=1, \ldots, t)$

  Measurement feedback: $u_{i,t} = \gamma_i(t; y_{i,s}, s=1, \ldots, t)$, $y_{i,t} = h_{i,t}(x_{i,t}, x_{-i,t}, v_{i,t})$

- Policy space Γ_i for player i would also reflect action and communication constraints (such as quantization, frequency of interactions, opportunistic sensing, and disruptions due to intermittent failure of channels)

General NZS SDGs with networked agents

- $x_t = (x_{1,t}, \ldots, x_{N,t})$; $x_{i,(t+1)} = f_{it}(x_{it}, u_{it}, c_{it}(x_{-i,t}, u_{-i,t}), w_{i,t})$, $i = 1, \ldots, N$ (state dynamics)

  an underlying network that governs connections, neighborhood interactions ($\mathcal{N}_{i,t}$ for i)

- *Information structures* (control policy of player i: $\gamma_i \in \Gamma_i$)

  Closed-loop perfect state: $u_{i,t} = \gamma_i(t; x_s, s=1, \ldots, t)$

  Partial (local) state: $u_{i,t} = \gamma_i(t; x_{i,s}, s=1, \ldots, t)$

  Measurement feedback: $u_{i,t} = \gamma_i(t; y_{i,s}, s=1, \ldots, t)$, $y_{i,t} = h_{i,t}(x_{i,t}, x_{-i,t}, v_{i,t})$

- *Loss function* for player i (over $t = 1, \ldots, T$) -- T could be ∞

  $L_i(x_{[1,T]}, u_{[1,T]}) = (1/T) \sum_{t \in [1,T]} g_{i,t}(x_{it}, u_{it}, k_{it}(x_{-i,t}, u_{-i,t}))$

  Take expectations (for horizon [1,T]) with $u = \gamma(\cdot)$: $J_i(\gamma_i, \gamma_{-i})$ [normal form of game]

General NZS SDGs with networked agents

- $x_t = (x_{1,t}, \ldots, x_{N,t})$; $\quad x_{i,(t+1)} = f_{it}(x_{it}, u_{it}, c_{it}(x_{-i,t}, u_{-i,t}), w_{i,t})$, $\quad i = 1, \ldots, N$ (state dynamics)

  an underlying network that governs connections, neighborhood interactions ($\mathcal{N}_{i,t}$ for i)

- *Information structures* (control policy of player i: $\gamma_i \in \Gamma_i$)

  Closed-loop perfect state: $u_{i,t} = \gamma_i(t; x_s, s=1, \ldots, t)$

  Partial (local) state: $u_{i,t} = \gamma_i(t; x_{i,s}, s=1, \ldots, t)$

  Measurement feedback: $u_{i,t} = \gamma_i(t; y_{i,s}, s=1, \ldots, t)$, $\quad y_{i,t} = h_{i,t}(x_{i,t}, x_{-i,t}, v_{i,t})$

- *Loss function* for player i (over $t = 1, \ldots, T$) -- T could be ∞

  $L_i(x_{[1,T]}, u_{[1,T]}) = (1/T) \sum_{t \in [1,T]} g_{i,t}(x_{it}, u_{it}, k_{it}(x_{-i,t}, u_{-i,t}))$

  Take expectations (for horizon [1,T]) with $u = \gamma(\cdot)$: $\quad J_i(\gamma_i, \gamma_{-i})$ [normal form of game]

- *Nash equilibrium* γ^*: $\quad J_i(\gamma_i^*, \gamma_{-i}^*) \leq J_i(\gamma_i, \gamma_{-i}^*) \quad \forall\, \gamma_i \in \Gamma_i$, $i = 1, \ldots, N$

General NZS SDGs with networked agents

- $x_t = (x_{1,t}, \ldots, x_{N,t})$; $\quad x_{i,(t+1)} = f_{it}(x_{it}, u_{it}, c_{it}(x_{-i,t}, u_{-i,t}), w_{i,t})$, $\quad i = 1, \ldots, N$ (state dynamics)

  an underlying network that governs connections, neighborhood interactions ($\mathcal{N}_{i,t}$ for i)

- *Information structures* (control policy of player i: $\gamma_i \in \Gamma_i$)

  Closed-loop perfect state: $u_{i,t} = \gamma_i(t; x_s, s=1, \ldots, t)$

  Partial (local) state: $u_{i,t} = \gamma_i(t; x_{i,s}, s=1, \ldots, t)$

  Measurement feedback: $u_{i,t} = \gamma_i(t; y_{i,s}, s=1, \ldots, t)$, $\quad y_{i,t} = h_{i,t}(x_{i,t}, x_{-i,t}, v_{i,t})$

- *Loss function* for player i (over $t = 1, \ldots, T$) -- T could be ∞

  $L_i(x_{[1,T]}, u_{[1,T]}) = (1/T) \sum_{t \in [1,T]} g_{i,t}(x_{it}, u_{it}, k_{it}(x_{-i,t}, u_{-i,t}))$

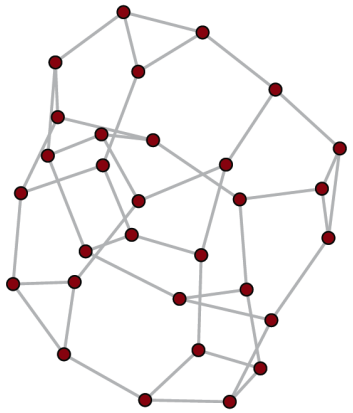
  Take expectations (for horizon [1,T]) with $u = \gamma(\cdot)$: $\quad J_i(\gamma_i, \gamma_{-i})$ [normal form of game]

- *ϵ-Nash equilibrium* γ^ϵ: $\quad J_i(\gamma_i^\epsilon, \gamma_{-i}^\epsilon) \le J_i(\gamma_i, \gamma_{-i}^\epsilon) + \epsilon \quad \forall\, \gamma_i \in \Gamma_i$, $i = 1, \ldots, N$

An Alternative Multi-Agent MDP Framework

- Setting: Networked Multi-agent MDP
  - $\langle \mathcal{S}, \{\mathcal{A}^i\}_{i\in\mathcal{N}}, P, \{R^i\}_{i\in\mathcal{N}}, \{\mathcal{G}_t\}_{t\geq 0}\rangle$ with the communication network $\{\mathcal{G}_t = (\mathcal{N}, \mathcal{E}_t)\}_{t\geq 0}$
  - Each agent i has individual policy $\pi^i_{\theta^i} : \mathcal{S} \to \mathcal{P}(\mathcal{A}^i)$
  - One goal: maximize the globally averaged return $\max_\theta J(\theta) = \lim_{T\to\infty} \frac{1}{T}\mathbb{E}\left(\sum_{t=0}^{T-1} \frac{1}{N}\sum_{i\in\mathcal{N}} r^i{}_{t+1} \middle| \pi_\theta\right)$

- Other settings
  - Fully competitive (e.g. 2-agent zero-sum Markov games, $R^1 \equiv -R^2$)
  - Mixed setting (no restriction on relationships among agents)
  - A special case of mixed setting is teams against teams
  - General solution concept is Nash equilibrium
  - There is also interest in correlated equilibrium and course correlated equilibrium*



A fully decentralized network with heterogeneous agents

*W. Mao, TB, "Provably efficient reinforcement learning in decentralized general-sum Markov games." *Dynamic Games and Applications*, 13(1):165-186, March 2023.

Back to
General NZS SDGs with networked agents

- $x_t = (x_{1,t}, \ldots, x_{N,t})$; $\quad x_{i,(t+1)} = f_{it}(x_{it}, u_{it}, c_{it}(x_{-i,t}, u_{-i,t}), w_{i,t})$, $\quad i = 1, \ldots, N$ (state dynamics)

  an underlying network that governs connections, neighborhood interactions ($\mathcal{N}_{i,t}$ for i)

- *Information structures* (control policy of player i: $\gamma_i \in \Gamma_i$)

  Closed-loop perfect state: $u_{i,t} = \gamma_i(t; x_s, s=1, \ldots, t)$

  Partial (local) state: $u_{i,t} = \gamma_i(t; x_{i,s}, s=1, \ldots, t)$

  Measurement feedback: $u_{i,t} = \gamma_i(t; y_{i,s}, s=1, \ldots, t)$, $\quad y_{i,t} = h_{i,t}(x_{i,t}, x_{-i,t}, v_{i,t})$

- *Loss function* for player i (over $t = 1, \ldots, T$) -- T could be ∞

  $L_i(x_{[1,T]}, u_{[1,T]}) = (1/T) \sum_{t \in [1,T]} g_{i,t}(x_{it}, u_{it}, k_{it}(x_{-i,t}, u_{-i,t}))$

  Take expectations (for horizon [1,T]) with $u = \gamma(\cdot)$: $\quad J_i(\gamma_i, \gamma_{-i})$ [normal form of game]

- *ϵ-Nash equilibrium* γ^ϵ: $\quad J_i(\gamma_i^\epsilon, \gamma_{-i}^\epsilon) \le J_i(\gamma_i, \gamma_{-i}^\epsilon) + \epsilon \quad \forall\, \gamma_i \in \Gamma_i$, $i = 1, \ldots, N$

Computation of NE

- **Policy iteration** (involves iteration based on individual best-response functions):[1,2,3,4]

  $\gamma_i^{t+1} \in Br_i(\gamma_{-i}^{t})$, for each i, and t=0,1,…. respecting the network structure, including comm

  Fixed point of this multiple-valued map is NE

  Existence, uniqueness can be established under various structural assumptions[1,2,3,4]

  Convergence of various types of iterations (synchronous, asynchronous, sequential, randomness)

****

[1]TB, "Decentralized multicriteria optimization of linear stochastic systems," TAC, April 1978.

[2]TB, "An eqm theory for multi-person DM with multiple probabilistic models," TAC, February1985.

[3]TB, S. Li, "Distributed algorithms for the computation of NE in linear SDGs," *SICON', May 1989*

[4]S.Yüksel, TB, Stochastic Teams, Games and Control under Information Constraints, Springer 2024

Computation of NE

- **Policy iteration** (involves iteration based on individual best-response functions):[1,2,3,4]

  $\gamma_i^{t+1} \in Br_i(\gamma_{-i}^{t})$, for each i, and t=0,1,…. respecting the network structure including comm

  Fixed point of this multiple-valued map is NE

  Existence, uniqueness can be established under various structural assumptions[1,2,3,4]

- *Iteration in the action space* (involves expansion of individual sigma fields)

  Leading to possible asymptotic agreement, depending on richness of iteration[5]

****

[1]TB, "Decentralized multicriteria optimization of linear stochastic systems," TAC, April 1978.

[2]TB, "An eqm theory for multi-person DM with multiple probabilistic models," TAC, Feb 1985.

[3]TB, S. Li, "Distributed algorithms for the computation of NE in linear SDGs," *SICON', May 1989*

[4]S.Yüksel, TB, Stochastic Teams, Games and Control under Information Constraints, Springer 2024

[5]S. Li, TB, "Asymptotic agreement and convergence of asynchronous algorithms," TAC, July 1987

Computation of NE

- **Relaxation** (involves memory in the iteration based on individual best-response functions):[6]

  $\gamma_i^{t+1} \in Br_i(\gamma_{-i}^s, s=t, t-1, .., t-d_i)$, for each i and t, where d_i is depth of memory for Player i

  Leading to improvement in convergence, such as faster rates

****

[6]TB, "Relaxation techniques and asynchronous algorithms for on-line computation of noncooperative equilibria," *J Economic Dynamics and Control*, 11:531-549, Dec 1987.

Computation of NE

- **Relaxation** (involves memory in the iteration based on individual best-response functions):[6]

  $\gamma_i^{t+1} \in Br_i(\gamma_{-i}^s, s=t, t-1, .., t-d_i)$, for each i and t, where d_i is depth of memory for Player i

  Leading to improvement in convergence, such as faster rates

- *Extremum Seeking* (sinusoidal excitation signals to estimate gradients and Hessians)

  Leading to convergence (asymptotic and finite-time) to NE with and w/o model info[7,8,9]

****

[6]TB, "Relaxation techniques and asynchronous algorithms for on-line computation of noncooperative equilibria," *J Economic Dynamics and Control*, 11:531-549, Dec 1987.

[7]P.Frihaof, M. Krstic, TB, "Nash equilibrium seeking in noncooperative games," *TAC, May 2012.*

[8]T.R.Oliveira, V.H.P. Rodrigues, M.Krstic, TB, "Nash equilibrium seeking in quadratic noncooperative games under two delayed information-sharing schemes," *JOTA, 191:700-735, 2021*

[9]J.I..Poveda, M. Krstic, TB, "Fixed-time NE seeking in time-varying networks," *TAC, April 2023*

With additional *zero*th player, as leader (L)

- $x_t = (x_{0,t}, x_{1,t}, \ldots, x_{N,t})$; $\quad x_{i,(t+1)} = f_{it}(x_{it}, x_{0t}, u_{0t}, u_{it}, c_{it}(x_{-i,t}, u_{-i,t}), w_{i,t}), \quad i = 1, \ldots, N$

  $x_{0,(t+1)} = f_{0t}(x_{0t}, u_{0t}, c_{0t}(x_{-0,t}, u_{-0,t}), w_{0,t})$

With additional *zero*th player, as leader (L)

- $x_t = (x_{0,t}, x_{1,t}, \ldots, x_{N,t})$; $\quad x_{i,(t+1)} = f_{it}(x_{it}, x_{0t}, u_{0t}, u_{it}, c_{it}(x_{-i,t}, u_{-i,t}), w_{i,t})$, $\quad i = 1, \ldots, N$

  $x_{0,(t+1)} = f_{0t}(x_{0t}, u_{0t}, c_{0t}(x_{-0,t}, u_{-0,t}), w_{0,t})$

- *Information structures* (control policy of player i: $\gamma_i \in \Gamma_i$) – as before
- *Loss function* for player i (over t = 1, . . ., T) -- T could be ∞

  $L_i(x_{[1,T]}, u_{[1,T]}) = (1/T) \sum_{t \in [1,T]} g_{i,t}(x_{it}, u_{it}, k_{it}(x_{-i,t}, u_{-i,t}))$, $\quad i = 0, 1, \ldots, N$

  c_{it} and k_{it} again respect underlying network topology

  Take expectations (for horizon [1,T]) with $u = \gamma(\cdot)$: $\quad J_i(\gamma_i, \gamma_{-i})$ [normal form of game]

With additional *zero*th player, as leader (L)

- $x_t = (x_{0,t}, x_{1,t}, \ldots, x_{N,t})$; $\quad x_{i,(t+1)} = f_{it}(x_{it}, x_{0t}, u_{0t}, u_{it}, c_{it}(x_{-i,t}, u_{-i,t}), w_{i,t})$, $\quad i = 1, \ldots, N$

  $x_{0,(t+1)} = f_{0t}(x_{0t}, u_{0t}, c_{0t}(x_{-0,t}, u_{-0,t}), w_{0,t})$

- *Information structures* (control policy of player i: $\gamma_i \in \Gamma_i$) – as before
- *Loss function* for player i (over t = 1, . . ., T) -- T could be ∞

  $L_i(x_{[1,T]}, u_{[1,T]}) = (1/T) \sum_{t \in [1,T]} g_{i,t}(x_{it}, u_{it}, k_{it}(x_{-i,t}, u_{-i,t}))$, $\quad i = 0, 1, \ldots, N$

  Take expectations (for horizon [1,T]) with $u = \gamma(\cdot)$: $\quad J_i(\gamma_i, \gamma_{-i})$ [normal form of game]

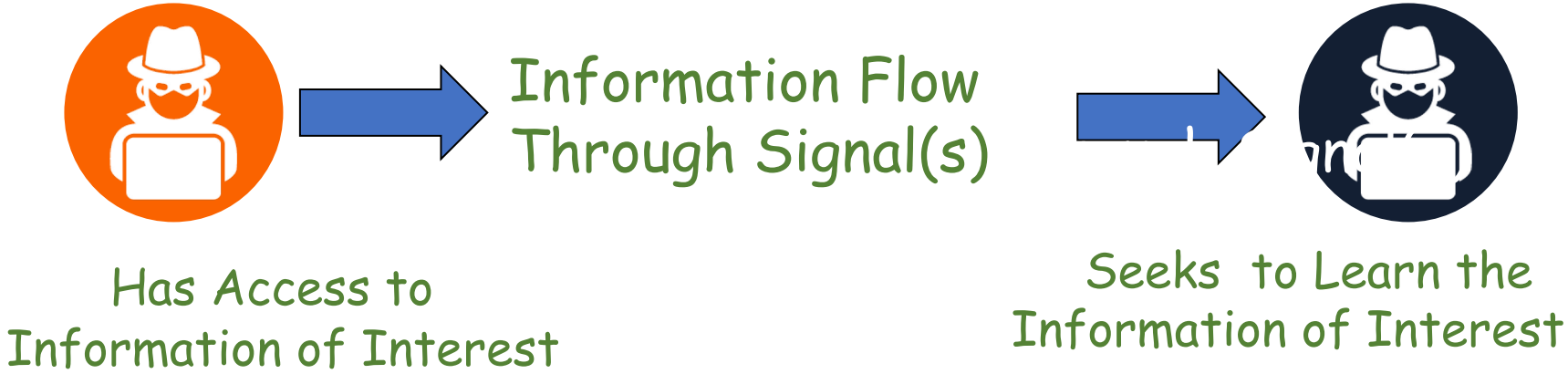
- *Stackelberg equilibrium (SE)* γ^*:

$$J_0(\gamma_0^*, \gamma_{-0}^*) \leq \sup_{\beta \in R(\mu)} J_0(\mu, \beta) \quad \forall\, \mu \in \Gamma_0$$

  where $R(\mu)$ is the NE reaction set of N followers to each announced policy $\mu \in \Gamma_0$ of L.

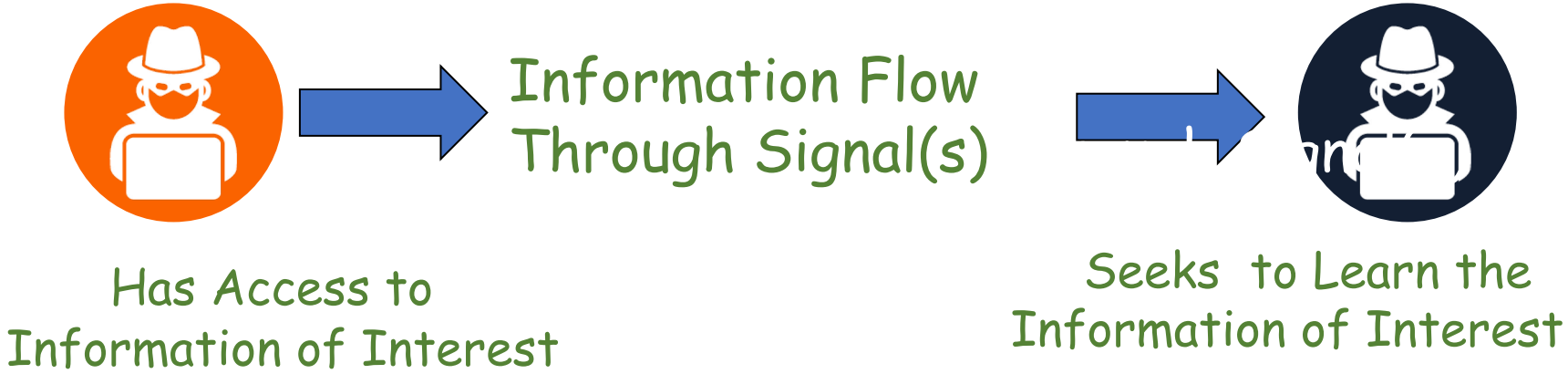
  Note: This is a '*pessimistic*' SE if $R(\mu)$ is not a singleton.

An important recent application: Strategic Information Transmission



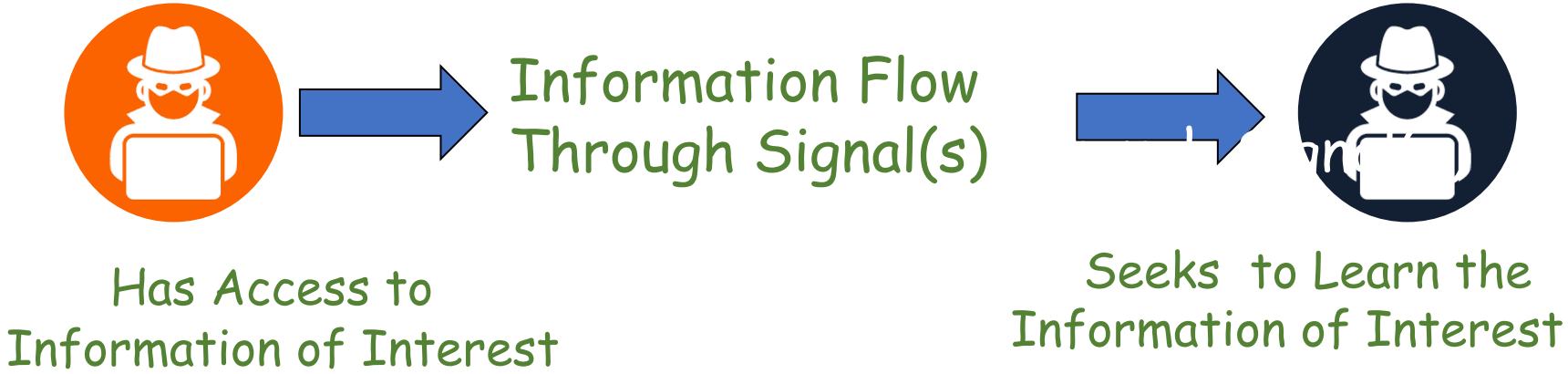
Question: How to craft the information of interest to induce the receiving agent to perceive it as desired (by S) even when R is aware of the relationship between the received signal and the information of interest?

Mathematical Ingredients of SIT

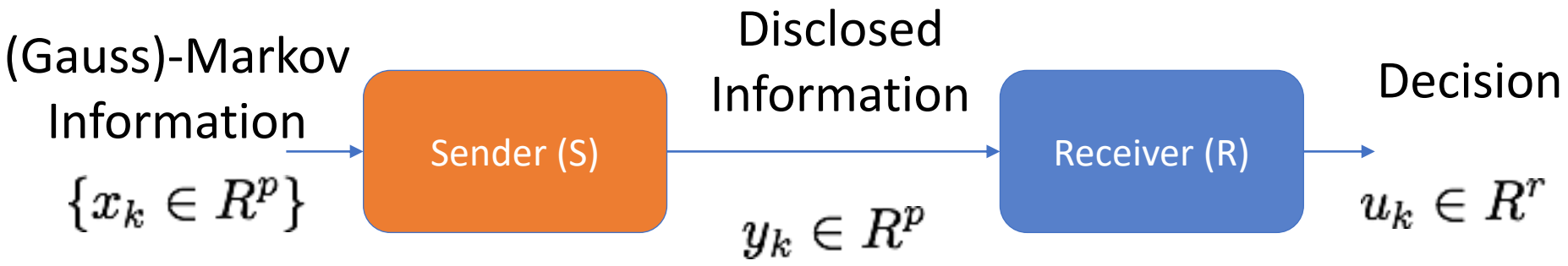


- Noncooperative intelligent agents (S and R)
- Different objectives
  - R wants to learn information of interest
  - S wants R to perceive that information as he desires
- Information of interest is drawn from a distribution
- Distribution and objectives are common knowledge

Game-theoretic Solutions



- Lack of cooperation and non-aligned objectives lead to a nonzero-sum dynamic game with asymmetric information
- Possible solution concepts:
  - Nash equilibrium (symmetric for S and R)
  - Stackelberg equilibrium (S leader and R follower)
- Structurally different equilibrium policies



R **observes** η_k and y_k

S selects Borel-measurable policy:

$$\eta_k : R^{kp} \to R^p$$

$$y_k = \eta_k(x_1, \ldots, x_k)$$

S's objective:

$$\min_{\eta_1,\ldots,\eta_n} E\left\{\sum_{k=1}^{n} \|Q_{S,k}x_k - R_{S,k}u_k\|^2\right\}$$

R selects Borel-measurable policy:

$$\gamma_k : R^{kp} \to R^r$$

$$u_k = \gamma_k(y_1, \ldots, y_k)$$

R's objective:

$$\min_{\gamma_1,\ldots,\gamma_n} E\left\{\sum_{k=1}^{n} \|Q_{R,k}x_k - R_{R,k}u_k\|^2\right\}$$

1.MO. Sayin, E. Akyol, TB, "Hierarchical multistage Gaussian signaling games in noncooperative communication and control systems", Automatica, Sept 2019.
2. MO. Sayin, TB, "Persuasion-based robust sensor design against attackers with unknown control objectives," TAC, Oct 2021.
3.MO. Sayin, TB, "Bayesian Persuasion With State-Dependent Quadratic Cost Measures," IEEE TAC, 67(3):1241-52, 2022

SIT Parallels Mechanism Design

Just as in mechanism design, or rather paralleling it, strategic information transmission (or deception) aims at having the receiving agent act (based on the distorted/biased information it receives) in a way aligned with the objective of the sending agent.

It is to control/manipulate the perception, and make the other side act in line with the intensions of the sender of information.

SIT Parallels Mechanism Design
--both entail inducement of behavior*--

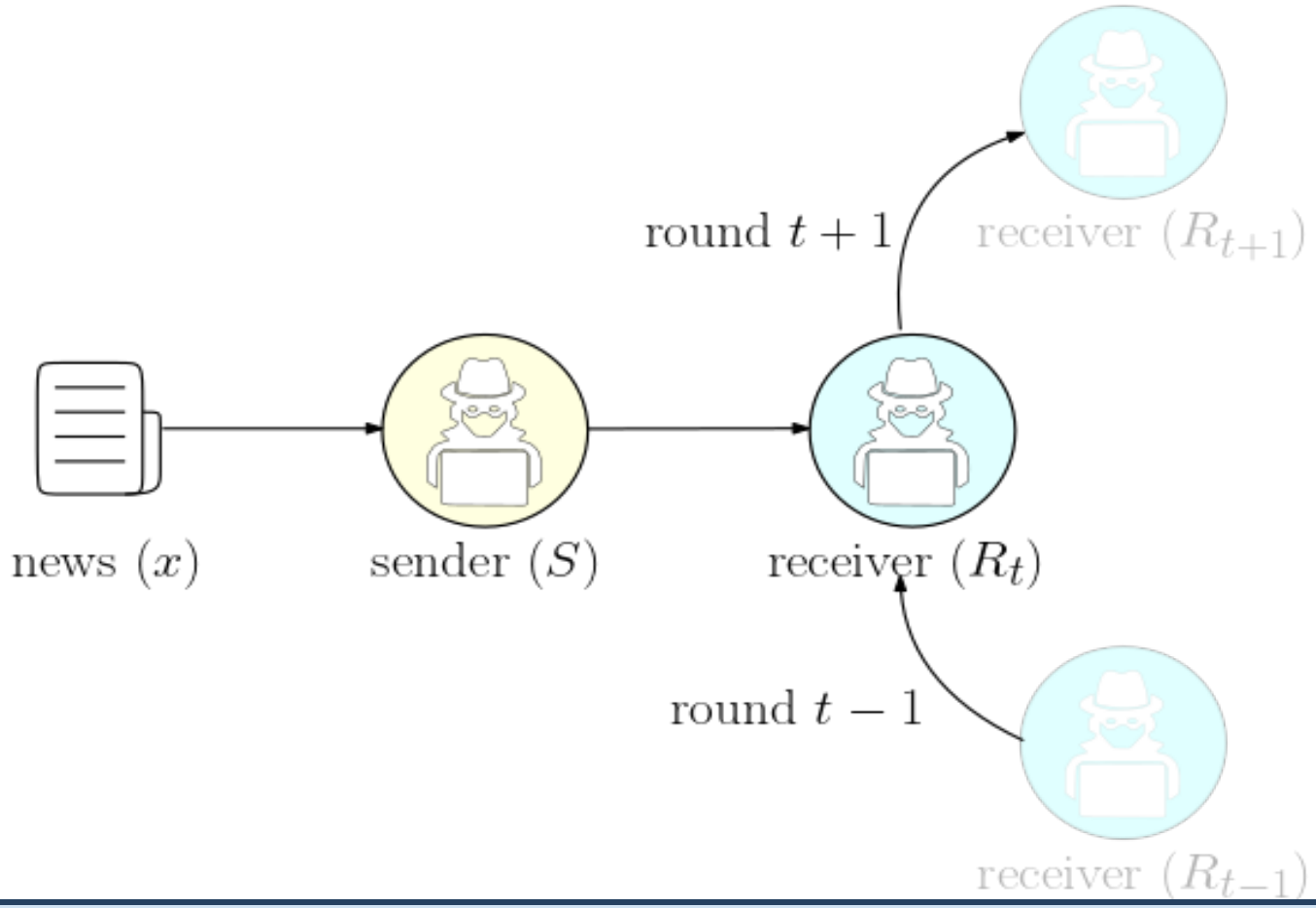
While strategic information transmission (SIT) entails control/manipulation of the perception by shaping the information transmitted mechanism design entails (within a *principal-agent* framework) manipulation of the utility function of the agent so that the agent's act/decision based on optimization of his/her utility is aligned (to the extent possible) with the objective of the *principal.*

---

* TB, "Inducement of desired behavior via soft policies," *International Game Theory Review, Special Issue on Game Theory and Optimization*, June 2024

Online SIT

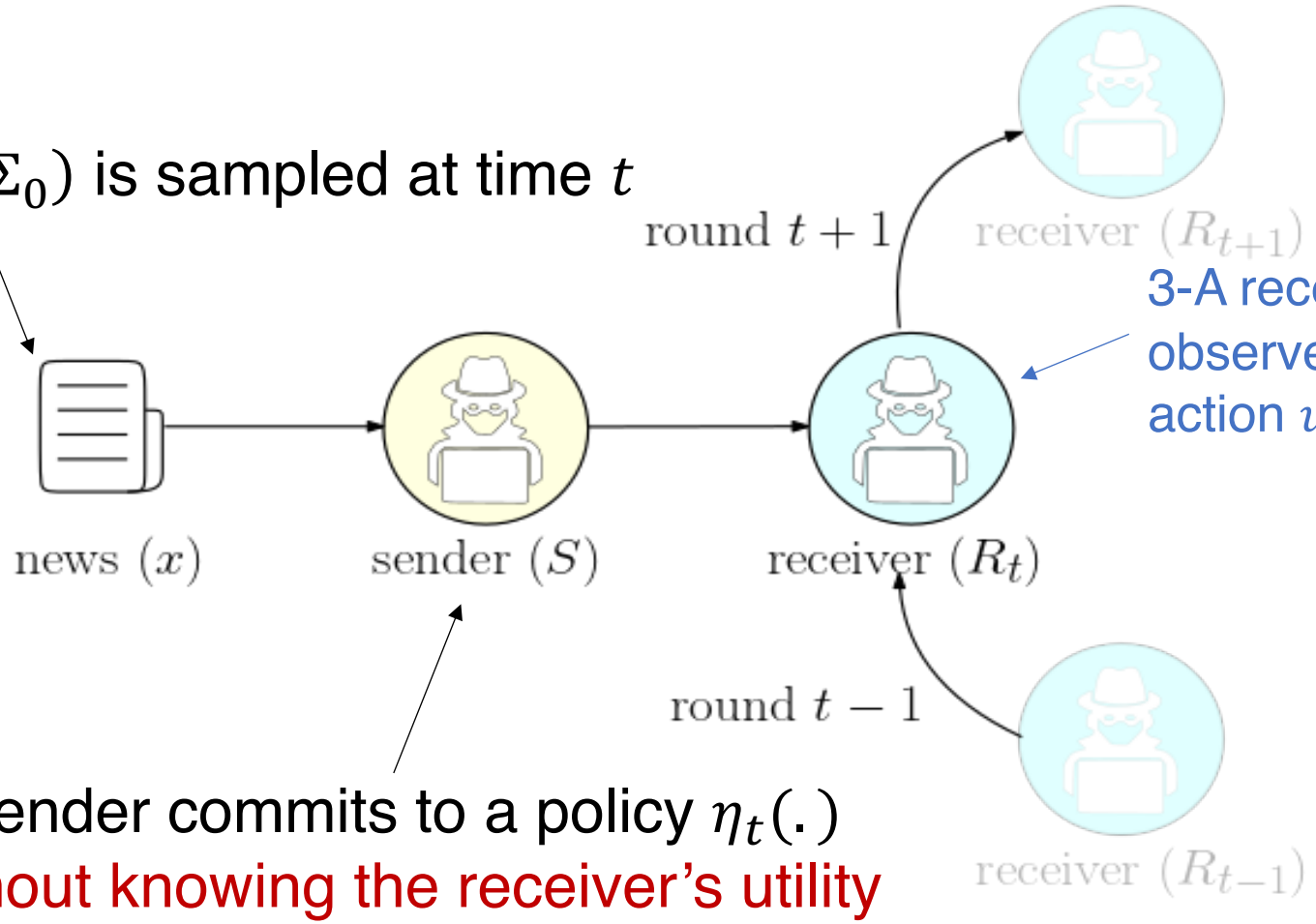
(Velicheti, Bastopcu, Etesami, TB, "Learning how to strategically disclose information, ACC'24)



The sender plays a repeated game with receiver of unknown type

Online Strategic Information Transmission

2-A state $x_t \sim N(0, \Sigma_0)$ is sampled at time t



3-A receiver enters the system, observes y_t and takes an action u_t and leaves

1-Sender commits to a policy $\eta_t(.)$
without knowing the receiver's utility

4- Costs: sender gets a stage cost $\|Q_S x_t + R_S u_t\|^2$ and receiver gets a cost $\left\|Q_{R_t} x_t + R_{R_t} u_t\right\|^2$

Online Strategic Information Transmission (SIT)

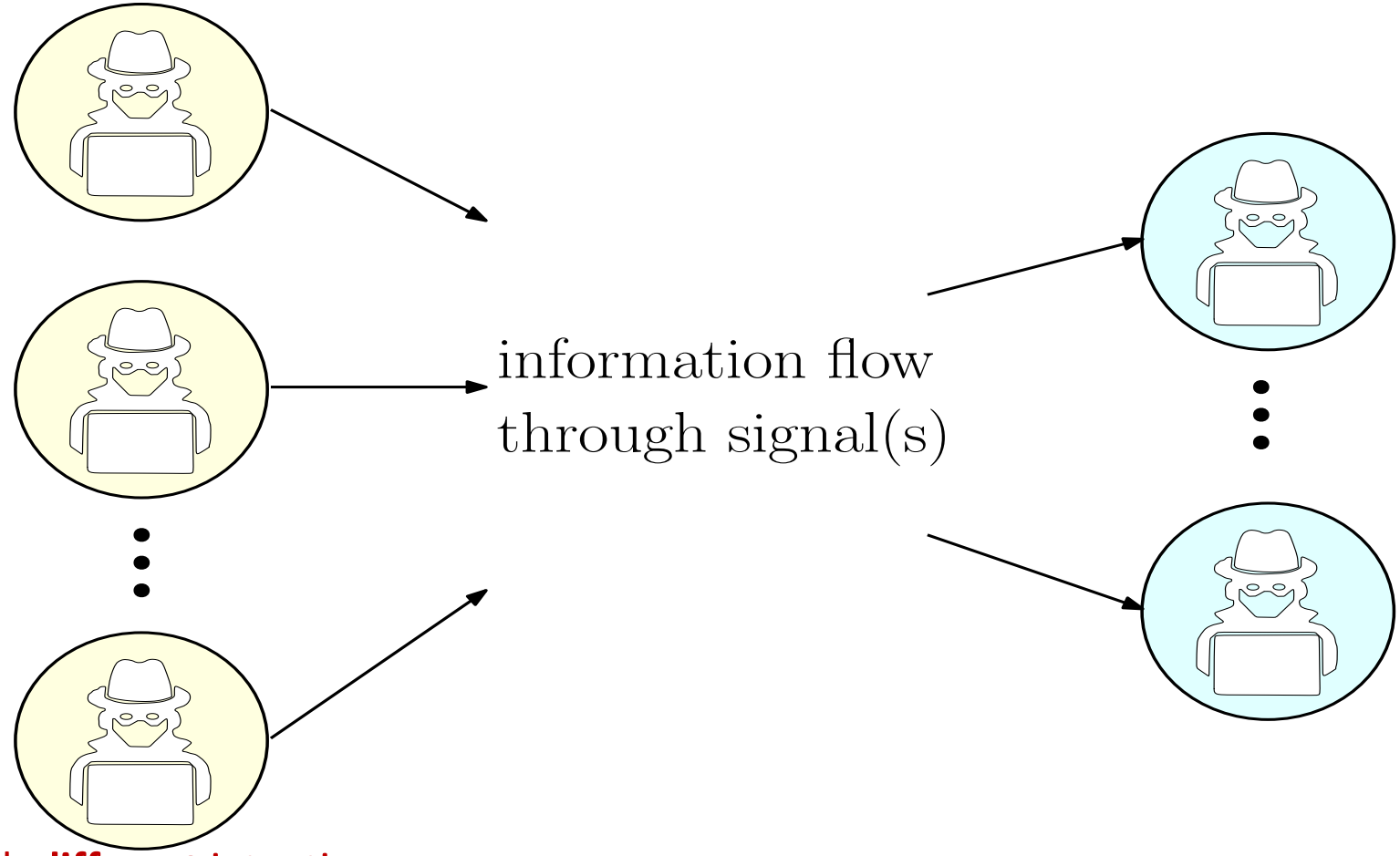
- We have considered:
  - **Full Feedback:** Sender observes the receiver's true type ($\theta_t = \{Q_{R_t}, R_{R_t}\}$)
  - **Bandit Feedback:** Sender only observes their cost as a result of their signaling policy at time t

Goal:

Design an online algorithm (an algorithm which bases its decision only on the past history) which performs as good as a signaling policy designed with known receiver utilities

Extensions exist to multiple Ss and multiple Rs

- Multiple Senders (leaders) to multiple Receivers (followers) with different intentions



with **different** intentions

R. K. Velicheti, M. Bastopcu, TB, "Value of Information in Games with Multiple Strategic Information Providers," available at arXiv:2306.14886

Other scenarios in going from Single S-R to Groups of Ss and Rs

- Multiple Senders crafting an information of interest along different objectives and sending to different groups (types) of Receivers with some overlap, allowing networking within groups
- Studying impact of manipulated information at different scales and layers, taking into account intra-group behavior at moderate- and high-population levels
- Introducing mismatch between the probabilistic outlooks of Ss and Rs – different a priori distributions
- Persuasion capacity in dynamic and/or noisy environments with multiple senders and/or multiple receivers, and associated computational tools

Back to Challenges in deriving NE and SE

- Almost complete theory for NE under CL perfect state (PS) information for all players (recursive computation in the spirit of DP), as well as under OL information (à la maximum principle). Also, an almost complete theory for SE when leader has OL information and followers either CLPS or OL. SE when leader has CLPS information is challenging through a direct approach, but an indirect approach with connections to mechanism design is feasible.[1,2]

****

[1]TB, G.J. Olsder. *Dynamic Noncooperative Game Theory,* SIAM, 1999.

[2]TB, "Affine incentive schemes for stochastic systems with dynamic information," *SIAM J Control & Optimiz,* 22(2):199-210, 1984.

Challenges in deriving NE and SE (model based)

- Almost complete theory for NE under CL perfect state (PS) information for all players (recursive computation in the spirit of DP), as well as under OL information (à la maximum principle). Also, an almost complete theory for SE when leader has OL information and followers either CLPS or OL. SE when leader has CLPS information is challenging through a direct approach, but an indirect approach with connections to mechanism design is feasible.[1,2]
- Other *dynamic (asymmetric)* information structures (s.a. local state, decentralized, measurement feedback): Extremely challenging! Possibly infinite-dimensional (even if, e.g., NE exists and is unique), even in linear-quadratic (LQ) NZS SDGs.[1]

****

[1]TB, G.J. Olsder. *Dynamic Noncooperative Game Theory,* SIAM, 1999.

[2]TB, "Affine incentive schemes for stochastic systems with dynamic information," *SIAM J Control & Optimiz,* 22(2):199-210, 1984.

Challenges in deriving NE and SE (model based)

- Almost complete theory for NE under CL perfect state (PS) information for all players (recursive computation in the spirit of DP), as well as under OL information (à la maximum principle). Also, an almost complete theory for SE when leader has OL information and followers either CLPS or OL. SE when leader has CLPS information is challenging through a direct approach, but an indirect approach with connections to mechanism design is feasible.[1,2]
- Other *dynamic (asymmetric)* information structures (s.a. local state, decentralized, measurement feedback): Extremely challenging! Possibly infinite-dimensional (even if, e.g., NE exists and is unique), even in linear-quadratic (LQ) NZS SDGs.[1]
- A few exceptions exist when asymmetric information can be decomposed into common and private information—game can then be "lifted" to one with only common information[3]

****

[1]TB, G.J. Olsder. *Dynamic Noncooperative Game Theory,* SIAM, 1999.

[2]TB, "Affine incentive schemes for stochastic systems with dynamic information," *SIAM J Control & Optimiz,* 22(2):199-210, 1984.

[3]A. Gupta, A. Nayyar, C. Langbort, TB, "Common information based Markov Perfect Equilibria for linear Gaussian games with asymmetric information," *SIAM J Control & Optimiz,* 52(5):3228-3260, 2014.

Challenges in deriving NE and SE (model based)

- Almost complete theory for NE under CL perfect state (PS) information for all players (recursive computation in the spirit of DP), as well as under OL information (à la maximum principle). Also, an almost complete theory for SE when leader has OL information and followers either CLPS or OL. SE when leader has CLPS information is challenging through a direct approach, but an indirect approach with connections to mechanism design is feasible.[1,2]
- Other *dynamic (asymmetric)* information structures (s.a. local state, decentralized, measurement feedback): Extremely challenging! Possibly infinite-dimensional (even if, e.g., NE exists and is unique), even in linear-quadratic (LQ) NZS SDGs.[1]
- There are also issues (computational and otherwise) in scaling up NE to a high population of players, even under symmetric information[4,5]

****

[1]TB, G.J. Olsder, *Dynamic Noncooperative Game Theory,* SIAM, 1999.

[4]TB, R. Srikant, "A Stackelberg network game with a large number of followers," *JOTA,* 115(3):479-490, Dec 2002..

[5]E.Altman, TB, R. Srikant, "NE for combined flow control and routing in networks: Asymptotic behavior for a large number of users," *TAC,* 46(6):917-9300, June 2002.

Challenges in deriving NE and SE (model based)

- Almost complete theory for NE under CL perfect state (PS) information for all players (recursive computation in the spirit of DP), as well as under OL information (à la maximum principle). Also, an almost complete theory for SE when leader has OL information and followers either CLPS or OL. SE when leader has CLPS information is challenging through a direct approach, but an indirect approach with connections to mechanism design is feasible.
- Other *dynamic* information structures (s.a. local state, decentralized, measurement feedback): Extremely challenging! Possibly infinite-dimensional (even if, e.g., NE exists and is unique), even in linear-quadratic (LQ) NZS SDGs.
- Why? Strategic interaction!
- Each player (agent) has to *second guess* the information available to other players (and not to her) in her active neighborhood, as it could be useful to her in improving her performance. If all players are doing so, then this leads to an *infinite recursion* -- asymptotically *learning* relevant information through direct measurement of others' actions or through their impact on state or performance. Even obtaining *approximate* NE is a formidable task (such as placing restrictions on the dimensions of policies).

Any way out to overcome this challenge?

This is particularly important for (and relevant to) multi-agent systems where a relatively large number of agents with differing individual (local) objectives and having access to only local (decentralized) information interact with each other toward a common (global) goal or slightly misaligned goals.

Mean-Field Games Approach[1,2,3] is the Answer

- Lift the N-player game up to an appropriate infinite-population one (assuming one exists, with possibly multiple populations, also respecting neighborhood relationships).
- Each generic agent faces a stochastic control problem, confronting a (sub)population which is exogenous to the agent, and not affected by the agent's actions.

****

[1]J.-M. Lasry, P.-L. Lions, "Mean field games," *Japan J. Math,* 2(1):229-260, 2007.

[2]M. Huang, P.E. Caines, R.P.. Malhamé, "Large population cost-coupled LQG problems with nonuniform agents: Individual-mass behavior and decentralized ϵ-Nash equilibria," *IEEE TAC*, 52(9):1560-1571, 2007.

[3]N. Saldi, TB, M. Raginsky, "Approximate Nash equilibria in partially observed stochastic games with mean-field interactions," *Mathematics of. Operations Research*, 44(3):1006-1033, 2019.

Mean-Field Games Approach is the Answer

- Lift the N-player game up to an appropriate infinite-population one (assuming one exists, with possibly multiple populations, also respecting neighborhood relationships).
- Each generic agent faces a stochastic control problem, confronting a (sub)population which is exogenous to the agent, and not affected by the agent's actions.

Digression

- E.g., in local state dynamics $x_{i,(t+1)} = f_{it}(x_{it}, u_{it}, c_{it}(x_{-i,t}, u_{-i,t}), w_{i,t})$, $i = 1, \ldots, N$,

  $c_{it}(x_{-i,t}, u_{-i,t}) = (1/|\mathcal{N}_{i,t}|) \sum_{j\in\mathcal{N}(i,t)} x_{j,t}$ where $|\mathcal{N}_{i,t}|$ is very large, even $\to \infty$
- *And/or in loss function* for player i (over $t = 1, \ldots, T$) -- T could be ∞

  $L_i(x_{[1,T]}, u_{[1,T]}) = (1/T) \sum_{t \in [1,T]} g_{i,t}(x_{it}, u_{it}, k_{it}(x_{-i,t}, u_{-i,t}))$,

  $k_{it}(x_{-i,t}, u_{-i,t}) = (1/|\mathcal{N}_{i,t}|) \sum_{j\in\mathcal{N}(i,t)} x_{j,t}$ where $|\mathcal{N}_{i,t}|$ is very large, even $\to \infty$
- Here c_{it} and k_{it} are stochastic processes exogenous to agent (player) i
- Also, aggregate actions ($u_{j,t}$) of neighboring agents could enter the formulation

Mean-Field Games Approach is the Answer

- Lift the N-player game up to an appropriate infinite-population one (assuming one exists, with possibly multiple populations, also respecting neighborhood relationships).
- Each generic agent faces a stochastic control problem, confronting a (sub)population which is exogenous to the agent, and not affected by the agent's actions.

Digression

- Such structures arise in many problems of interest that involve a large number of players, such as congestion control, sharing of common resources, and consensus formation[4,5]

---

[4]TB, "A consensus problem in mean field setting with noisy measurements of target," Proc. 2018 ACC, pp. 6521-6526.

[5]H.Shen and TB, "Pricing under information asymmetry for a large population of users," Telecommunication Systems, 47(1-2):123-136, June 2011.

Mean-Field Games Approach[1,2,3] is the Answer

- Lift the N-player game up to an appropriate infinite-population one (assuming one exists, with possibly multiple populations, also respecting neighborhood relationships).
- Each generic agent faces a stochastic control problem, confronting a (sub)population which is exogenous to the agent, and not affected by the agent's actions.
- Generate the state process (and other possible aggregate quantities) under these optimal control policies and require consistency (entails solving a FP eq).
- Together with the optimal control policies, this leads to mean-field equilibrium (MFE).

****

[1]J.-M. Lasry, P.-L. Lions, "Mean field games," *Japan J. Math,* 2(1):229-260, 2007.

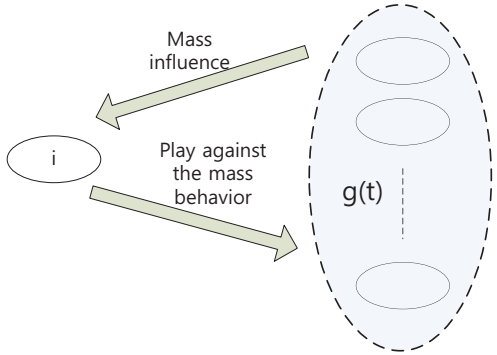
[2]M. Huang, P.E. Caines, R.P.. Malhamé, "Large population cost-coupled LQG problems with nonuniform agents: Individual-mass behavior and decentralized ϵ-Nash equilibria," *IEEE TAC*, 52(9):1560-1571, 2007.

[3]N. Saldi, TB, M. Raginsky, "Approximate Nash equilibria in partially observed stochastic games with mean-field interactions," *Mathematics of Operations Research*, 44(3):1006-1033, 2019.

Mean-Field Games Approach is the Answer

- Lift the N-player game up to an appropriate infinite-population one (assuming one exists, with possibly multiple populations, also respecting neighborhood relationships).
- Each generic agent faces a stochastic control problem, confronting a (sub)population which is exogenous to the agent, and not affected by the agent's actions.
- Generate the state process (and other possible aggregate quantities) under these optimal control policies and require consistency (entails solving a FP eq).
- Together with the optimal control policies, this leads to mean-field equilibrium (MFE).

Schematically for a single population with exogenous process g:



- Stochastic control problem for generic agent leads to an optimal policy, say μ^*, that depends on g (and only local information for the agent)
- Use that policy in the state equation of the generic agent, and find g so that it is consistent with the emerging state process (FP)—g^*
- (μ^*, g^*) constitutes the MFE

Mean-Field Games Approach is the Answer

- Lift the N-player game up to an appropriate infinite-population one (assuming one exists, with possibly multiple populations, also respecting neighborhood relationships).
- Each generic agent faces a stochastic control problem, confronting a (sub)population which is exogenous to the agent, and not affected by the agent's actions.
- Generate the state process (and other possible aggregate quantities) under these optimal control policies and require consistency (entails solving a FP eq).
- Together with the optimal control policies, this leads to mean-field equilibrium (MFE).
- It is possible to build in robustness through a risk-sensitive formulation (working with exponentiated loss functions for the agents)—connection to introducing an adversary agent, with generic agent now facing a zero-sum stochastic dynamic game.[4,5,6]

****

[4]H. Tembine, Q. Zhu, TB, "Risk-sensitive mean field games," *IEEE TAC, 59*(4):835-850, April 2014.

[5]J. Moon, TB, "Linear-quadratic risk-sensitive and robust mean-field games," *IEEE TAC*, 62(3):1062-1077, March 2017; --"Risk-sensitive mean field games via the stochastic maximum principle," *Dynamic Games and Applications*, 9:1100-1125, 2019.

[6]N. Saldi, TB, M. Raginsky, "Approximate Markov-Nash equilibria for discrete-time risk-sensitive mean-field games," *Mathematics of Operations Research*, 45(4):1596-1620, Nov 2020.

Mean-Field Games Approach is the Answer

- Lift the N-player game up to an appropriate infinite-population one (assuming one exists, with possibly multiple populations, also respecting neighborhood relationships).
- Each generic agent faces a stochastic control problem, confronting a (sub)population which is exogenous to the agent, and not affected by the agent's actions.
- Generate the state process (and other possible aggregate quantities) under these optimal control policies and require consistency (entails solving a FP eq).
- Together with the optimal control policies, this leads to mean-field equilibrium (MFE)
- Finally, study the relationship between finite N and infinite N solutions—leading to ε-NE, thus resolving the formidable task of obtaining approximate NE for games with asymmetric information (such as local measurements only), where $\varepsilon \to 0$ as $N \to \infty$.

Mean-Field Games with Stackelberg Leader[7]

- Stackelberg leader (L) faces an infinite-population of followers.
- For each policy of L, the followers play a mean-field Nash game as before, generating the Nash reaction set.
- L optimizes his expected cost on that reaction set, which in turn leads to followers' corresponding policies and the exogenous process at the followers' level.
- These, together with L's policy, lead to mean-field equilibrium (MFE).
- Study (as before) the relationship between finite (N) population of followers and infinite N solutions—with the policies in the MFE leading to $(\varepsilon_S,\varepsilon_F)$ Stackelberg-NE, where ε_S provides the level of approximation to L's objective, and ε_F determines the level of approximation to each of N followers in the Nash game, where both $\varepsilon_S \to 0$ and $\varepsilon_F \to 0$ as $N \to \infty$.

****

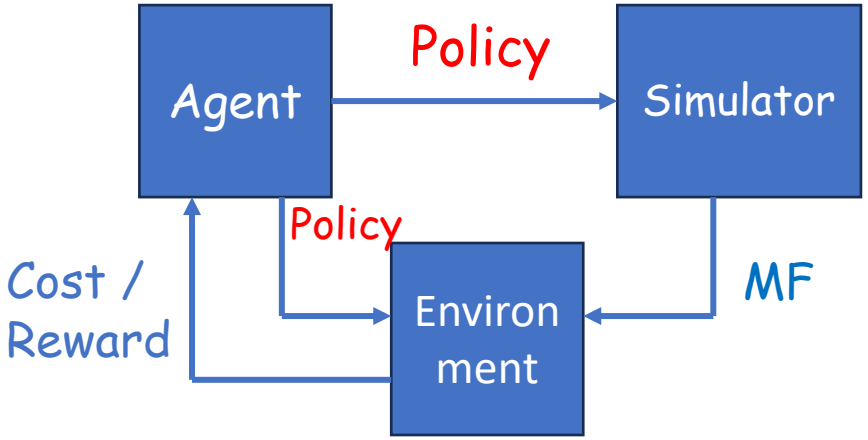
[7]J. Moon, TB, "Linear quadratic mean field Stackelberg differential games," Automatica, 97:200-213, Nov 2018.

Computational Aspects

- MFE-based approximate NE policy is scalable
- Computation of the mean field at NE requires solution of a fixed-point equation, which requires full modelling knowledge
- One way around this is for each agent to interact with a central coordinator (simulator) who collects state values and/or policies of the agents, computes the mean field, and broadcasts to all agents, who then update their policies based on the received MF, ... and so on. With a finite number, L, of different populations of agents, L different MFs are computed.

Computational Aspects

- MFE-based approximate NE policy is scalable
- Computation of the mean field at NE requires solution of a fixed-point equation, which requires full modelling knowledge
- One way around this is for each agent to interact with a central coordinator (simulator) who collects state values and/or policies of the agents, computes the mean field, and broadcasts to all agents, who then update their policies based on the received MF, ... and so on. With a finite number, L, of different populations of agents, L different MFs are computed.



Single population schematic:
Generic Agent (A) interacts with the Simulator (S) and the Environment (E), feeding policy and/or state values.
S computes the MF, feeding it to E, where cost/reward of A is generated and sent to A, who updates its policy based on some optimization algorithm.

Computational Aspects with Learning

- What if the agents do not know their own models? Then bring in RL for each agent into the iterative/learning process
- Parametrize the policies and optimize over the parameters, using e.g., policy gradient, respecting also computation and communication constraints[1,2,3]
- When explicit form of the agent's objective function is not available, its gradient can be computed only approximately, using e.g., **zero-order stochastic optimization (ZSO)**
- This will require further study of finite sample guarantees for the underlying algorithms
- It may be possible to avoid use of a central coordinator (simulator)[4]

****

[1]T. Li, G. Peng, Q. Zhu, TB, "The confluence of networks, games, and learning: A game-theoretic framework for multiagent decision making over networks," *IEEE Control Systems Magazine, 42*(4):35-67, August 2022.

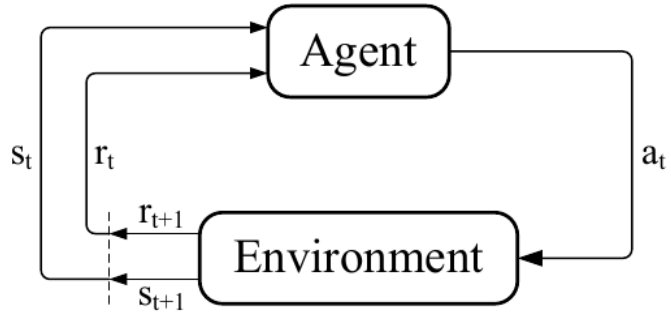
[2]T. Chen, K. Zhang, G.B. Giannakis, TB, "Communication-efficient policy gradient methods for distributed reinforcement learning," *IEEE TCNS*, 9(2):917-929, June 2022.

[3]B. Hu, K. Zhang, N. Li, M. Mesbahi, M. Fazel,TB, "Toward a theoretical foundation of policy optimization for learning control policies," *Annual Review of Control, Robotics, and Autonomous Systems*, 6:123-158, 2023.

[4]M.A. uz Zaman, S. Bhatt, A. Koppel TB, "Oracle-free reinforcement learning in mean-field games along a single sample path," Proc. 25th Internat Conf AI & Statistics (AISTATS 2023), Valencia, Spain, April 25-27, 2023.

Digression: Reinforcement Learning (RL)*

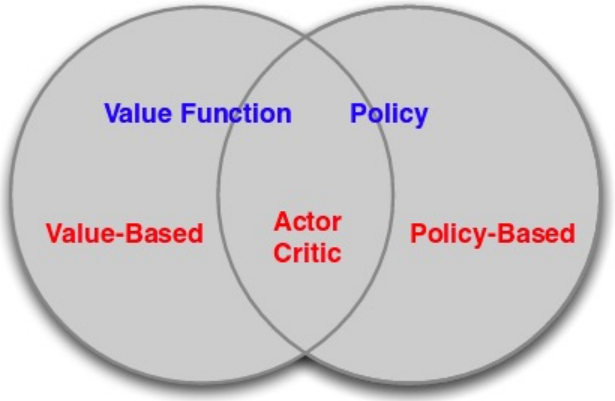
- Reinforcement Learning: solving Markov decision processes/optimal control problems without knowing the model
- Goal: maximize accumulated/time-average reward



$$\text{e.g.,} \quad \max_{\pi} \quad J(\pi) = \lim_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}(r_{t+1})$$

- Algorithms:
  - Critic-only: e.g., Q-learning
  - Actor-only: e.g., policy gradient
  - Actor-critic enjoys both advantages

		Cons.
Value-base		erge with function approx.; ardly handle continuous actions
Policy-base		riance of gradient estimate



Source: *Reinforcement learning Lecture Notes*, David Silver, 2016

*K. Zhang, Z. Yang, TB, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," *Handbook of Reinforcement Learning and Control, Springer*, pages 321-384, 2021

Actor-critic Algorithm

- Parametrize policy $\pi : \mathcal{S} \to \mathcal{P}(\mathcal{A})$ by π_θ with parameter $\theta \in \mathbb{R}^m$
- Basis: Policy Gradient Theorem [Sutton et. al. '00]

baseline

$$\nabla_\theta J(\theta) = \mathbb{E}_{s\sim\rho_\theta, a\sim\pi_\theta}\left\{\nabla_\theta \log \pi_\theta(s,a) \cdot [Q_\theta(s,a) - b(s)]\right\}$$

Relative Q-function

$$Q_\theta(s,a) = \sum_{t\geq 0} \mathbb{E}\left[\bar{r}_{t+1} - J(\theta) \mid s_0 = s, a_0 = a, \pi_\theta\right]$$

- Critic: policy evaluation with function approximation $Q(s,a) \to Q(s,a;\omega)$

Average running reward → $\mu_{t+1} = (1-\beta_{\omega,t}) \cdot \mu_t + \beta_{\omega,t} \cdot r_{t+1}$

Temporal difference (TD) error → $\delta_t = r_{t+1} - \mu_t + Q(s_{t+1}, a_{t+1}; \omega_t) - Q(s_t, a_t; \omega_t)$

TD update → $\omega_{t+1} = \omega_t + \beta_{\omega,t} \cdot \delta_t \cdot \nabla_\omega Q_t(\omega_t)$

- Actor: policy improvement

Advantage function → $A_t = Q(s_t, a_t; \omega_t) - \int_{a\in\mathcal{A}} \pi_{\theta_t}(s_t, a) Q(s_t, a; \omega_t) da$

Policy gradient update → $\theta_{t+1} = \theta_t + \beta_{\theta,t} \cdot A_t \cdot \nabla_\theta \log \pi_{\theta_t}(s_t, a_t)$

Convergence of AC Algorithms

- A sound full theoretical understanding of the convergence is still lacking, e.g. finite sample analysis, convergence guarantees with nonlinear functions, such as DNNs
- Asymptotic convergence can be shown [Konda & Borkar, ’99][Konda & Tsitsiklis, ’03][Bhatnagar et al., ’09]
  - Online update: faster critic and slower actor
  - Linear function approximation (FA) for the critic
  - Technique: two-time-scale stochastic approximation [Borkar ’08]
  - Single-time-scale online AC with linear FA [Castro & Meir, ’10]
- An important question: How good is the resulting policy?

Decentralized MARL AC*

- Policy Gradient Theorem for Multi-agent MDP

$$\nabla_{\theta^i} J(\theta) = \mathbb{E}_{s\sim d_\theta, a\sim\pi_\theta}\left[\nabla_{\theta^i}\log\pi^i_{\theta^i}(s,a^i)\cdot Q_\theta(s,a)\right] = \mathbb{E}_{s\sim d_\theta, a\sim\pi_\theta}\left[\nabla_{\theta^i}\log\pi^i_{\theta^i}(s,a^i)\cdot A_\theta(s,a)\right]$$
$$= \mathbb{E}_{s\sim d_\theta, a\sim\pi_\theta}\left[\nabla_{\theta^i}\log\pi^i_{\theta^i}(s,a^i)\cdot A^i_\theta(s,a)\right].$$

where $A^i_\theta(s,a) = Q_\theta(s,a) - \sum_{a^i\in\mathcal{A}^i}\pi^i_{\theta^i}(s,a^i)\cdot Q_\theta(s,a^i,a^{-i})$ ← local advantage function for agent i

- Decentralized multi-agent actor-critic (preserves privacy)
  - Critic-step:

Local temporal difference (TD) update

$$\begin{cases} \mu^i_{t+1} = (1-\beta_{\omega,t})\cdot\mu^i_t + \beta_{\omega,t}\cdot r^i_{t+1} \\ \delta^i_t = r^i_{t+1} - \mu^i_t + Q_{t+1}(\omega^i_t) - Q_t(\omega^i_t) \\ \widetilde{\omega}^i_t = \omega^i_t + \beta_{\omega,t}\cdot\delta^i_t\cdot\nabla_\omega Q_t(\omega^i_t) \end{cases}$$

Consensus update → $\omega^i_{t+1} = \sum_{j\in\mathcal{N}} c_t(i,j)\cdot\widetilde{\omega}^j_t$

  – Actor-step:

$$\psi^i_t = \nabla_{\theta^i}\log\pi^i_{\theta^i_t}(s_t,a^i_t)$$

$$A^i_t = Q_t(\omega^i_t) - \sum_{a^i\in\mathcal{A}^i}\pi^i_{\theta^i_t}(s_t,a^i)\cdot Q(s_t,a^i,a^{-i}_t;\omega^i_t)$$

Local policy gradient → $\theta^i_{t+1} = \theta^i_t + \beta_{\theta,t}\cdot A^i_t\cdot\psi^i_t$

*K. Zhang, Z. Yang, TB, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," *Handbook of Reinforcement Learning and Control, Springer*, pages 321-384, 2021

MA AC Algorithm Features

- Two-time-scale update (first critic, then actor)

$$\sum_t \beta_{\omega,t} = \sum_t \beta_{\theta,t} = \infty, \quad \sum_t \beta_{\omega,t}^2 + \beta_{\theta,t}^2 < \infty$$

$$\beta_{\theta,t} = o(\beta_{\omega,t})$$

$\lim_{t\to\infty} \beta_{\omega,t+1} \cdot \beta_{\omega,t}^{-1} = 1$, and same for $\beta_{\theta,t}$

- Favorable properties
  - No need to share either the policy or the reward of other agents
  - But have to share actions (there is a state-value based one that avoids this, using the result that TD error unbiasedly estimates the advantage function)
  - Memory-efficient: Online algorithm updates
  - Lower-variance: EPG + function approximation
- Applicable to both finite and continuous action spaces

Further on the MA AC Algorithm

- Basic technique: two-timescale stochastic approximation
- Critic-step: most existing proof techniques on distributed/consensus optimization do not apply
  - TD update is not a stochastic gradient of a fixed and well-defined function
  - Noises are correlated in online RL algorithms

$$\widetilde{\omega}_t^i = \omega_t^i + \beta_{\omega,t} \cdot \delta_t^i \cdot \nabla_\omega Q_t(\omega_t^i)$$

$$\omega_{t+1}^i = \sum_{j \in \mathcal{N}} c_t(i,j) \cdot \widetilde{\omega}_t^j$$

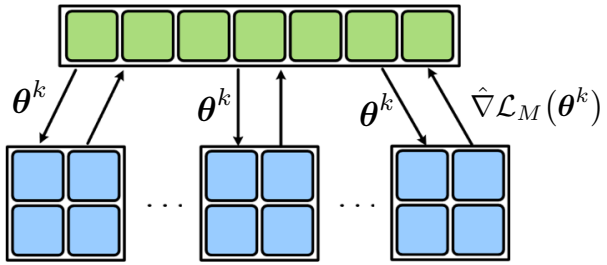
- As a byproduct, we have stability, i.e., almost sure boundedness, of this consensus stochastic sequence
- The same type of convergence guarantees as single-agent actor-critic

Competitive MARL & Finite-sample Analysis

- Related paper: K. Zhang, Z. Yang, H. Liu, T. Zhang, T. Başar, "Finite-sample analysis for decentralized batch multi-agent RL with networked agents," IEEE TAC, 69(12):5925-5940, Dec. 2021
  - Finite-sample performance analysis for decentralized batch MARL
  - Mixed setting: zero-sum setting for two teams where each team is a networked MDP as discussed; special case is zero-sum Markov games
  - Algorithm: fitted-Q iteration that can be implemented in a decentralized way
  - Results: quantified how the Q-function estimation error depends on
    - # of samples and # of iterations
    - Error caused by decentralized computation
    - Expressiveness of the function class for Q-function approx.

MARL with Efficient Communication

- Communication-efficient MARL [CZGB21][LZYWBSL19]
  - [CZGB21]: MARL policy gradient method, but agents only transmit the local gradients occasionally by checking certain conditions
  - [LZYWBSL19]: actor-critic as before, but only randomly transmit one entry in each consensus step



- [CZGB21] T. Chen, K. Zhang, G. B. Giannakis, and TB, "Communication-efficient policy-gradient methods for distributed reinforcement learning," *IEEE TCNS, 2022*.
- [LZYWBSL19] L. Lin, K. Zhang, Z. Yang, Z. Wang, TB, R. Sandhu, and J. Liu, "A Communication-Efficient Multi-Agent Actor-Critic Algorithm for Distributed Reinforcement Learning," *IEEE CDC*, 2019.

RL for LQ Zero-sum Dynamic Games

Strong connection (equivalence) between LQ ZSDGs, risk-sensitive control, and H_∞ (robust) control

- Policy gradient methods for Nash-equilibrium seeking, with finite-sample convergence guarantees
- Provably convergent derivative-free policy-optimization (PO)
- Inner-loop & outer-loop optimization of underlying min max problem
- Major challenge: all descent directions do not work→ implicit regularization

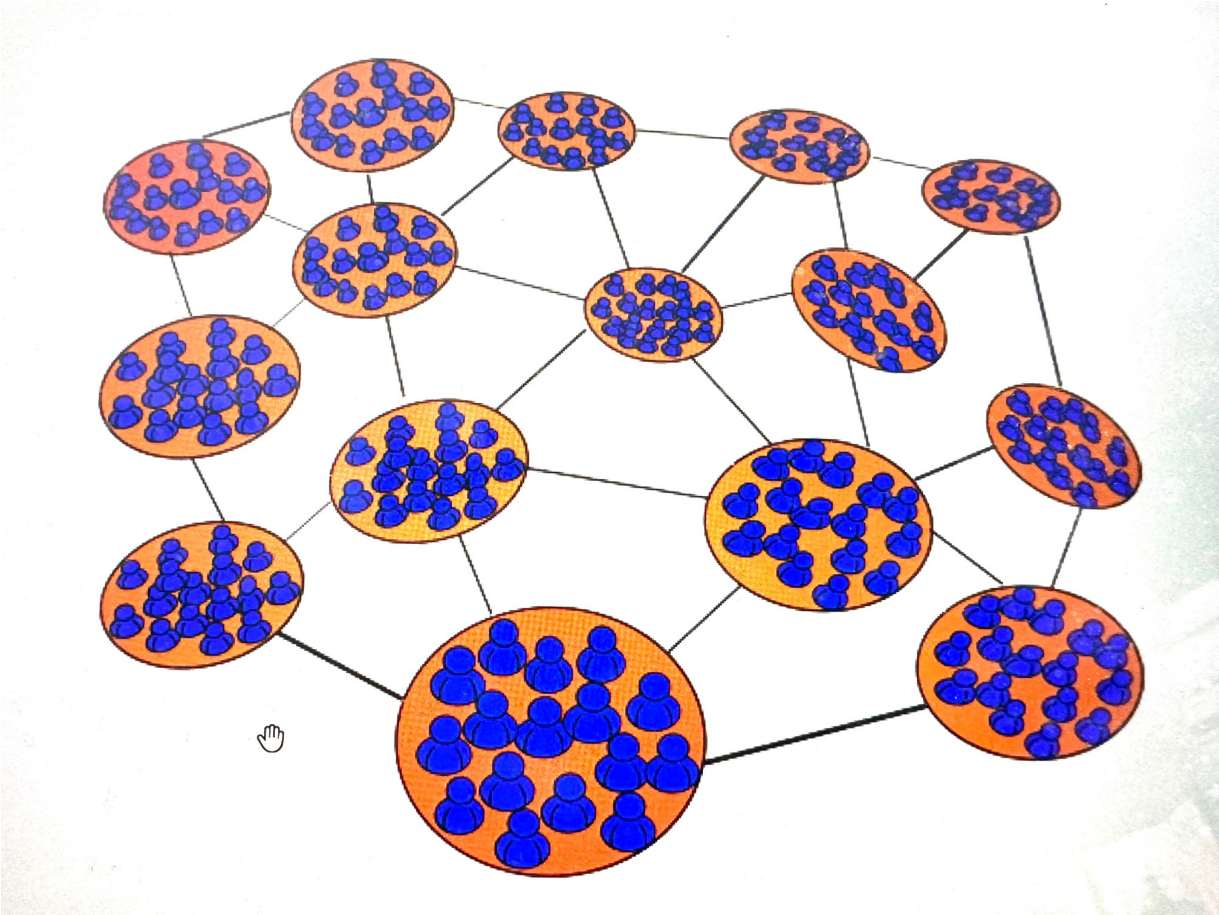
- K. Zhang, X Zhang, B. Hu, TB, "Derivative-free policy optimization for linear risk-sensitive and robust control design: Implicit regularization and sample complexity," *NeurIPS 2021; Dec 6-14, 2021*
- K. Zhang, B. Hu, TB, "Policy optimization for H-2 linear control with H-infinity robustness guarantee: Implicit regularization and global convergence," *SIAM J Control and Optimization*, 59(6):4081-4110, 2021

Back to MFGs:
An Illustration on Learning for MFGs

Framework of a specific class of Linear-Quadratic Mean Field Games with multiple types of agents (multiple populations), with coexistence of consensus and dissensus.

A specific framework for LQ MFGs

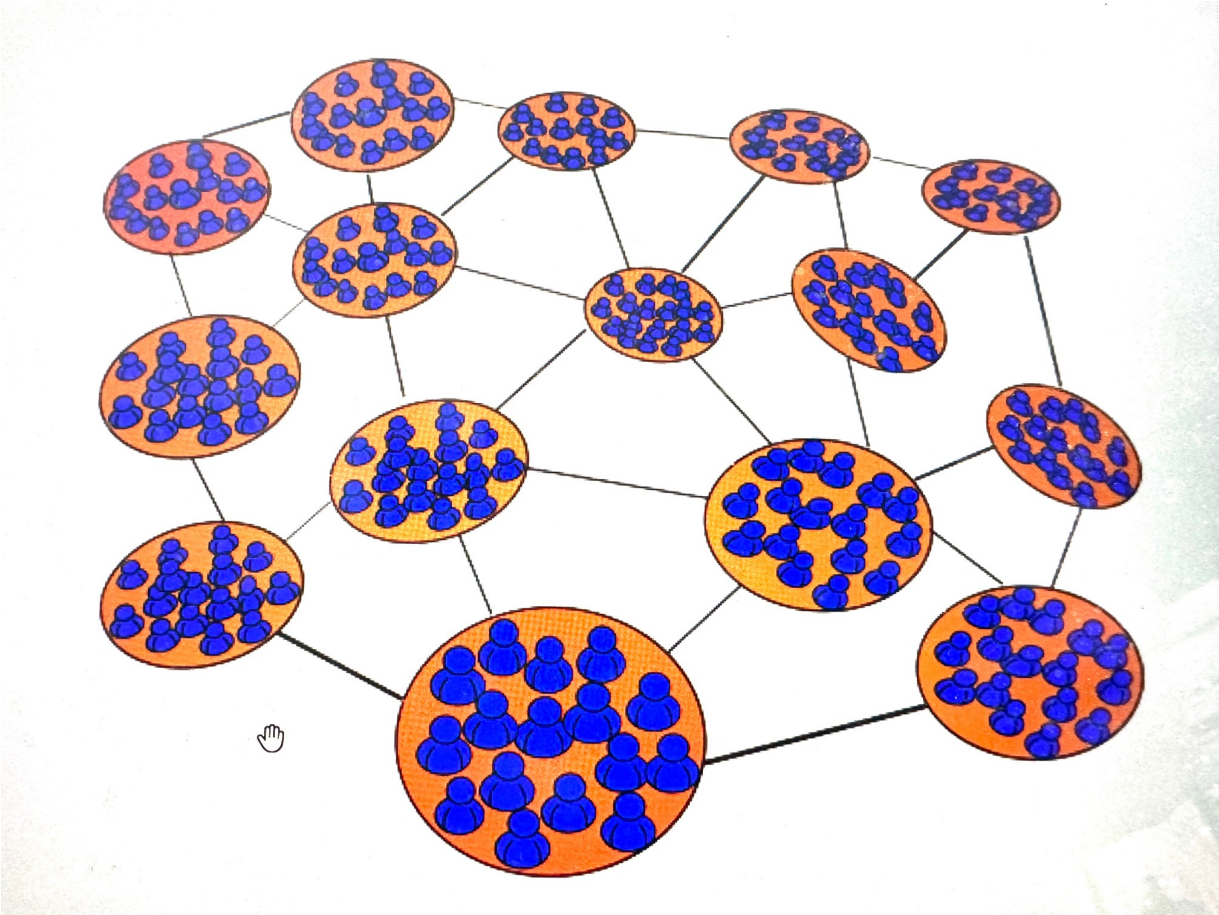
Multiple types (populations) of agents where `like' ones want to stay close to each other (consensus) whereas different populations want to have some separation between them (dissensus).



A specific framework for LQ MFGs

Multiple types (populations) of agents where `like' ones want to stay close to each other (consensus) whereas different populations want to have some separation between them (dissensus).

How can we accommodate/capture consensus and dissensus within a single large-scale decision-making formulation?

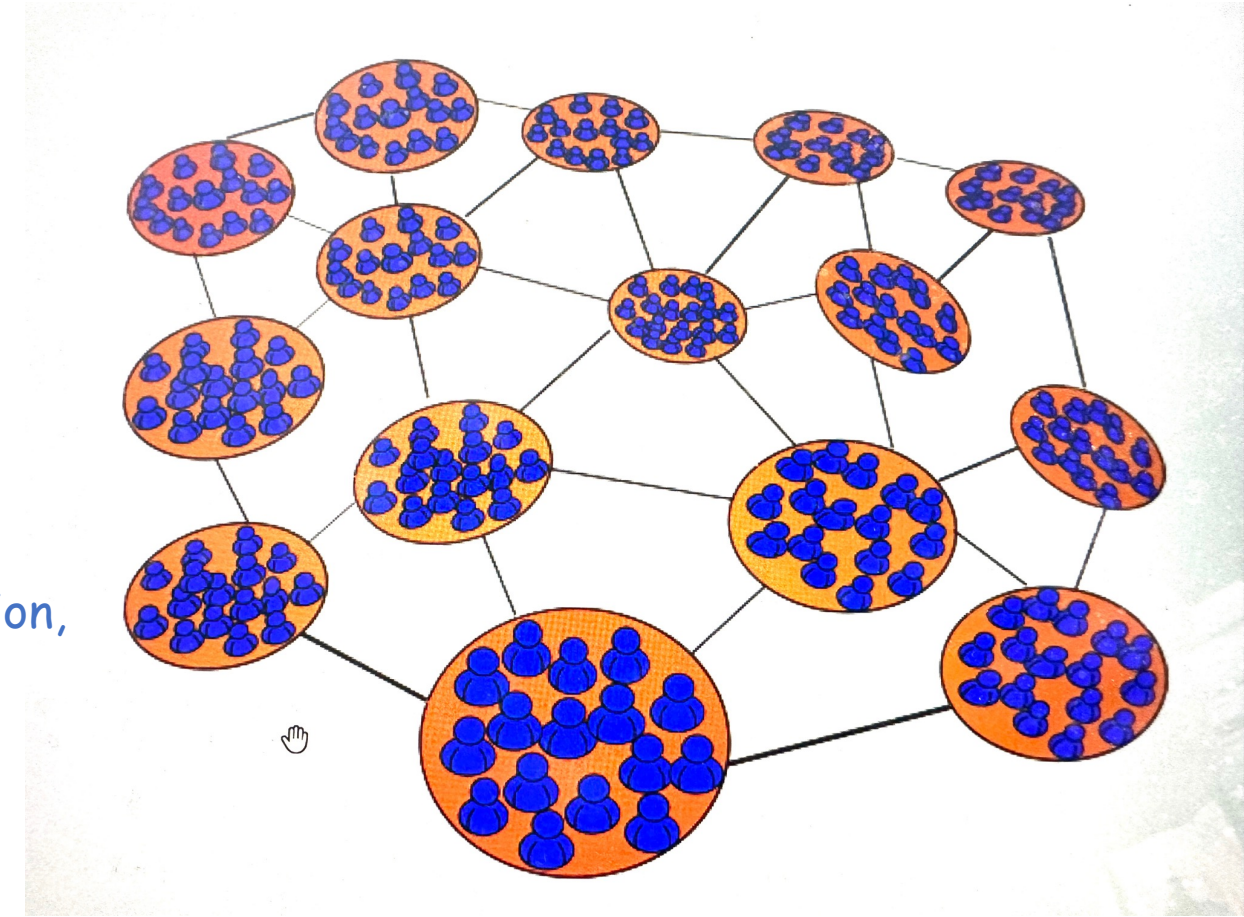


A specific framework for LQ MFGs

Multiple types (populations) of agents where `like' ones want to stay close to each other (consensus) whereas different populations want to have some separation between them (dissensus).

How can we accommodate/capture consensus and dissensus within a single large-scale decision-making formulation?

The answer is: a game-theoretic formulation, by building into the objective functions of different agents their preferences and attitudes toward others in <u>different populations</u>

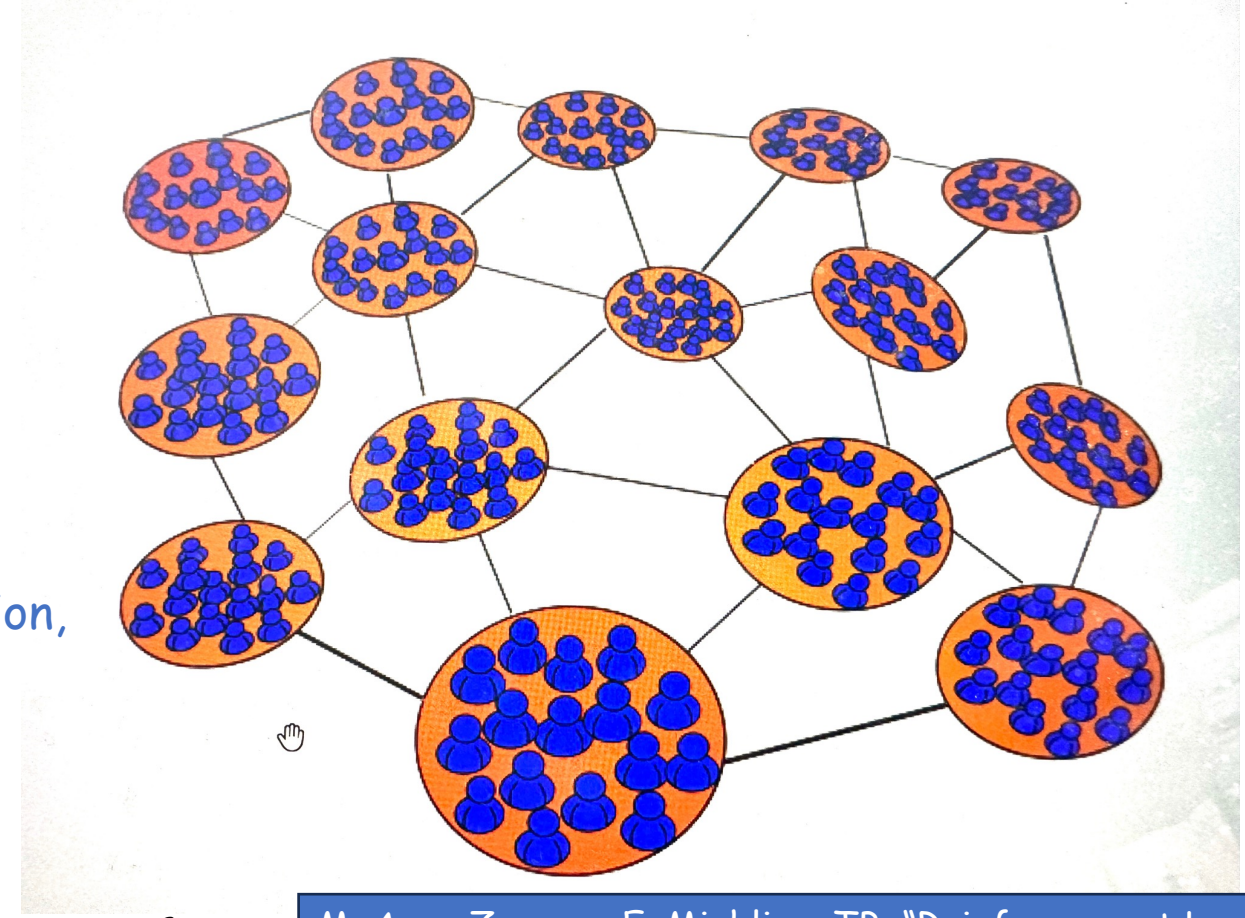


A specific framework for LQ MFGs

Multiple types (populations) of agents where `like' ones want to stay close to each other (consensus) whereas different populations want to have some separation between them (dissensus).

How can we accommodate/capture consensus and dissensus within a single large-scale decision-making formulation?

The answer is: a game-theoretic formulation, by building into the objective functions of different agents their preferences and attitudes toward others in <u>different populations</u>



AND given that we generally have large numbers of agents in each population, this calls for an analysis based on <u>MFGs</u>

M. A. uz Zaman, E. Miehling, TB, "Reinforcement Learning for Non-Stationary Discrete-Time Linear-Quadratic Mean-Field Games in Multiple Populations," *Dynamic Games and Applications*, 13:118-164, 2023

First: Single-Population MFGs (finite N)

<u>N agent game ($N < \infty$)</u>

- Agent $n \in [N]$ has linear dynamics:

$$Z_{t+1}^n = AZ_t^n + BU_t^n + W_t^n,$$

where W_t^n is independent zero-mean Gaussian noise

- Agent n has local information: $I_t^n := (I_{t-1}^n, U_{t-1}^n, Z_t^n) \quad I_0^n = Z_0^n$
- Agents have coupled cost functions ($Q \geq 0, C_U > 0$ and $C_Z \geq 0$)

$$J_n^{(N)}(\pi^n, \pi^{-n}) := \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\underbrace{\boxed{\|Z_t^n\|_Q^2 + \|U_t^n\|_{C_U}^2}}_{\text{Regulation}} + \underbrace{\boxed{\left\|Z_t^n - \frac{1}{N-1}\sum_{n'\neq n} Z_t^{n'}\right\|_{C_Z}^2}}_{\text{Consensus}}\right]$$

Single-Population MFGs (infinite N)

<u>Mean-Field game ($N \to \infty$)</u>

- Consider *generic* agent, with linear dynamics

$$Z_{t+1} = AZ_t + BU_t + W_t,$$

where W_t is independent zero-mean Gaussian noise

- Agent has local information, $I_t = (I_{t-1}, U_{t-1}, Z_t) \in \mathcal{I}_t$, $I_0 = Z_0$
- *Mean-field (MF)* $\bar{Z} := (\bar{Z}_0, \bar{Z}_1, \dots)$(aggregate behavior of agents)
- Agent has MF-coupled cost function:

$$J(\phi, \bar{Z}) = \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\underbrace{\|Z_t\|_Q^2 + \|U_t\|_{C_U}^2}_{\text{Regulation}} + \underbrace{\|Z_t - \bar{Z}_t\|_{C_Z}^2}_{\text{Consensus}}\right]$$

Single-Population MFGs (infinite N)

<u>Mean-Field Game ($N \to \infty$)</u>

- MF Equilibrium is a controller/trajectory pair $(\phi, \bar{Z})$
- Define two operators:
  - <u>(Consistency)</u> $\Lambda : \Phi \to \bar{\mathcal{Z}}$, $\bar{Z}$ consistent with ϕ if,

$$\bar{Z}_{t+1} = A\bar{Z}_t + B\phi_t(\bar{Z}_t)$$

  - <u>(Optimality)</u> $\Psi : \bar{\mathcal{Z}} \to \Phi$, ϕ optimal for $\bar{Z}$ if,

$$\Psi(\bar{Z}) = \underset{\phi \in \Phi}{\operatorname{argmin}}\, J(\phi, \bar{Z})$$

Unique FP of $\bar{Z}$ = $\boldsymbol{\Lambda}(\phi)$ and $\phi \in \boldsymbol{\Psi}(\bar{Z})$ → MFE $(\phi^*, \bar{Z}^*)$

Single-Population MFGs (finite N approx)

<u>Mean-Field Game ($N \to \infty$)</u>

- MF Equilibrium is a controller/trajectory pair $(\phi, \bar{Z})$
- Define two operators:
  - <u>(Consistency)</u> $\Lambda : \Phi \to \bar{\mathcal{Z}}$, $\bar{Z}$ consistent with ϕ if,

$$\bar{Z}_{t+1} = A\bar{Z}_t + B\phi_t(\bar{Z}_t)$$

  - <u>(Optimality)</u> $\Psi : \bar{\mathcal{Z}} \to \Phi$, ϕ optimal for $\bar{Z}$ if,

$$\Psi(\bar{Z}) = \operatorname*{argmin}_{\phi \in \Phi} J(\phi, \bar{Z})$$

Unique FP of $\bar{Z} = \Lambda(\phi)$ and $\phi \in \Psi(\bar{Z})$ → MFE $(\phi^*, \bar{Z}^*)$

In the N-agent game, if each agent uses ϕ^*, ϵ-NE, $\epsilon \sim O(1/\sqrt{N})$

Multi-Population MFGs

<u>L populations, with N_l agents in population $l \in [L]$</u>

- Each agent $n \in N_l, l \in [L]$ has linear and uncoupled dynamics:

$$Z_{t+1}^{n,l} = A^l Z_t^{n,l} + B^l U_t^{n,l} + W_t^{n,l}$$

where $W_t^{n,l}$ is independent zero-mean Gaussian noise

- Each agent has access to only her local state and action with full memory
- Agents have coupled cost functions (with neighboring populations):

Regulation

$$J_{n,l}^{(\mathsf{N})}\left((\pi^{n,l}, \pi^{-n,l}), \pi^{-l}\right) := \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\Bigg[\boxed{\left\|Z_t^{n,l}\right\|_{Q^l}^2 + \left\|U_t^{n,l}\right\|_{C_U^l}^2}$$

$$+ \boxed{\left\|Z_t^{n,l} - \frac{1}{N_l - 1} \sum_{\substack{n' \in [N_l] \\ n' \neq n}} Z_t^{n',l}\right\|_{C_Z^{ll}}^2} + \boxed{\sum_{\substack{k \in \mathcal{L}_l \\ k \neq l}} \left\|Z_t^{n,l} - \left(\beta^{lk} + \frac{1}{N_k} \sum_{n' \in [N_k]} Z_t^{n',k}\right)\right\|_{C_Z^{lk}}^2}\Bigg]$$

Intra-Population consensus

Inter-Population consensus (dissensus)

Multi-Population MFGs

<u>Multi-Population MFG ($N_l \to \infty$)</u>

- Generic agent $l \in [L]$ has linear and uncoupled dynamics,

$$Z_{t+1}^l = A^l Z_t^l + B^l U_t^l + W_t^l,$$

where W_t^l is independent Gaussian noise.

- Each generic agent has local information as before
- Agent costs coupled through mean-field $\bar{\mathbf{Z}} := (\bar{Z}^1, \dots, \bar{Z}^L)$,

Regulation

$$J_l(\phi^l, \bar{\mathbf{Z}}) := \limsup_{T\to\infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\Big[\boxed{\|Z_t^l\|_{Q^l}^2 + \|U_t^l\|_{C_U^l}^2}$$

Inter-Population consensus (dissensus)

$$+ \boxed{\left\|Z_t^l - \bar{Z}_t^l\right\|_{C_Z^{ll}}^2} + \boxed{\sum_{\substack{k\in\mathcal{L}_l \\ k\neq l}} \left\|Z_t^l - \left(\beta^{lk} + \bar{Z}_t^k\right)\right\|_{C_Z^{lk}}^2}\Big]$$

Intra-Population consensus

Multi-Population MFGs

<u>Multi-Population MFG ($N_l \to \infty$)</u>

- MFE in Multi-Population MFG now has L components
- Define two operators:
  - <u>(Consistency)</u> $\Lambda : \Phi \to \bar{\mathscr{Z}}$, $\bar{\mathbf{Z}}$ consistent with $\boldsymbol{\phi}$ if,

$$\bar{Z}^l_{t+1} = A^l \bar{Z}^l_t + B^l \phi^l_t(\bar{Z}^l_t), \;\; \bar{Z}^l_0 = \nu^l_0 \;\; \text{for all } l \in [L]$$

  - <u>(Optimality)</u> $\Psi : \bar{\mathscr{Z}} \to \Phi$, $\boldsymbol{\phi}$ optimal for $\bar{\mathbf{Z}}$ if,

$$\psi^l(\bar{\mathrm{Z}}) = \underset{\phi^l \in \Phi^l}{\operatorname{argmin}} \, J_l(\phi^l, \bar{\mathrm{Z}})$$

Unique FP of $\bar{Z} = \boldsymbol{\Lambda}(\phi)$ and $\phi \in \boldsymbol{\Psi}(\bar{Z})$ → MFE $(\phi^*, \bar{Z}^*)$

Characterization of MFE

MFE for Multi-Population LQ-MFG

- Existence & Uniqueness of MFE can be established under
  - (A^l, B^l) controllable, $(A^l, \sqrt{Q^l})$ observable $\forall l \in [L]$.

Proposition 1 *An MFE* $(\phi^*, \bar{Z}^*)$ *exists and is unique. Furthermore, the equilibrium controller* $\phi^* = (\phi^{1*}, \ldots, \phi^{L*})$ *and the equilibrium mean-field* $\bar{Z}^* = (\bar{Z}_1^*, \bar{Z}_2^*, \ldots)$ *take the following forms:*

1. $\phi^{l*}(Z_t^l, \bar{Z}_t^*) = -K_{l,1}^* \begin{bmatrix} Z_t^l \\ \bar{Z}_t^* \end{bmatrix} - K_{l,2}^* = \boxed{-K_{l,1}^{1*} Z_t^l - K_{l,1}^{2*} \bar{Z}_t^*} - \boxed{K_{l,2}^*}$ *for all* $l \in [L]$
2. $\bar{Z}_{t+1}^* = \boxed{\mathsf{F}^* \bar{Z}_t^*} + \boxed{\mathsf{C}^*}$

for some matrices $K_{l,1}^* = (K_{l,1}^{1*}, K_{l,1}^{2*})$, $K_{l,2}^*$, F^* *and* C^*.

Linear terms

Offset terms

Approximate NE with finite N_l

<u>Multi-Population MFG (finite N_l)</u>

- ε-Nash guarantee for MFE

Theorem 2 *Let $\tilde{\phi}$ denote the collection of controllers where each agent n in each population l employs the equilibrium controller of the corresponding generic agent. Then, for all $n \in [N_l]$, $l \in [L]$,*

$$J_{n,l}^{(\mathsf{N})}(\tilde{\phi}) \leq \inf_{\pi^{n,l} \in \Pi^l} J_{n,l}^{(\mathsf{N})}((\pi^{n,l}, \tilde{\phi}^{-n,l}), \tilde{\phi}^{-l}) + \mathcal{O}\Big(1/\sqrt{\min_{k \in \mathcal{L}_l} N_k}\Big) \qquad (14)$$

where $J_{n,l}^{(\mathsf{N})}(\cdot)$ is the cost function of (10), $\tilde{\phi}^{-n,l}$ denotes the elements of $\tilde{\phi}$ corresponding to population l excluding agent n, and $\tilde{\phi}^{-l}$ denotes the elements of $\tilde{\phi}$ excluding population l.

Toward RL: Formulation as Drifted-LQR

- Without loss of generality consider affine mean-fields: $\bar{\mathsf{Z}}_{t+1} = \mathsf{F}\ \bar{\mathsf{Z}}_t + \mathsf{C} + \omega_t$
- Noise due to imperfect simulation
- Define extended state of agent $l \in [L]$: $\mathsf{X}_t^l := (Z_t^l, \bar{\mathsf{Z}}_t) \in \mathbb{R}^{m(L+1)}$
- The dynamics of extended state is:

$$\mathsf{X}_{t+1}^l = \bar{\mathsf{A}}^l \mathsf{X}_t^l + \bar{B}^l U_t^l + \bar{\mathsf{C}} + \bar{\mathsf{W}}_t^l$$

$$\bar{\mathsf{A}}^l = \begin{bmatrix} A^l & 0 \\ 0 & \mathsf{F} \end{bmatrix}, \bar{B}^l = \begin{bmatrix} B^l \\ 0 \end{bmatrix}, \bar{\mathsf{C}} = \begin{bmatrix} 0 \\ \mathsf{C} \end{bmatrix}, \bar{\mathsf{W}}_t^l = \begin{bmatrix} W_t^l \\ \omega_t \end{bmatrix}$$

- The cost of agent l (LQR with drift):

$$J_l(\phi^l, \bar{\mathsf{Z}}) := \limsup_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E}\left[\left\|\mathsf{X}_t^l - \bar{\beta}^l\right\|_{\bar{\mathsf{Q}}^l}^2 + \left\|U_t^l\right\|_{C_U^l}^2\right]$$

Toward RL: Formulation as Drifted-LQR

- Search in the space of affine controllers *w.l.o.g.:*

$$\phi_t^l(\mathsf{X}_t^l) = -K_{l,1}\mathsf{X}_t^l - K_{l,2}$$

- Cost under affine controller:

$$J_l((K_{l,1}, K_{l,2}), \bar{\mathsf{Z}}) = J_l^1(K_{l,1}, \bar{\mathsf{Z}}) + J_l^2((K_{l,1}, K_{l,2}), \bar{\mathsf{Z}}) + \bar{\beta}^{l\top}\bar{\mathsf{Q}}^l\bar{\beta}^l$$

- If the control offset and dynamics offset are 0:

$$J_l((K_{l,1}, 0), \bar{\mathsf{Z}}) = J_l^1(K_{l,1}, \bar{\mathsf{Z}}) + \bar{\beta}^{l\top}\bar{\mathsf{Q}}^l\bar{\beta}^l$$

- Hence, the linear and offset terms can be learned in a decoupled manner and independently
- Cost J_l^1 (J_l^2) satisfies local smoothness, Lipschitz property, and gradient domination (strong convexity)

RL for Multi-Population MFGs

- The RL algorithm is divided into two parts for each agent $l \in [L]$:
  - first part estimates linear terms
  - second part estimates offset terms
- In both parts, each agent uses ZSO to estimate controller parameters
- Central simulator computes the mean-field under the set of controllers
- First part deals with linear terms in controllers and mean-fields
  - by keeping the offset terms zero
- Second part deals with offset terms in the controllers and mean-fields

Central Simulator

- Simulator simulates the behavior of a single agent in each population as an estimate for the mean-field of that population
- Under controller ϕ^l the mean-field of population l is

$$\bar{Z}^l_{t+1} = A^l \bar{Z}^l_t + B^l \phi^l \left(\begin{bmatrix} \bar{Z}^l_t \\ \bar{\mathsf{Z}}_t \end{bmatrix} \right) + \omega^l_t$$

- In the first part of algorithm, linear controller (control offset $= 0$)
  - As a result, linear mean-field:

$$\bar{\mathsf{Z}}_{t+1} = \mathsf{F}\ \bar{\mathsf{Z}}_t + \omega_t$$

- In the second part of algorithm, the controller is affine
  - As a result, affine mean-field:

$$\bar{\mathsf{Z}}_{t+1} = \mathsf{F}\ \bar{\mathsf{Z}}_t + \mathsf{C} + \omega_t$$

ZSO based RL Algorithm

Algorithm 1: RL for Multi-Population LQ-MFGs

Input: Number of iterations: S_1, S_2

Initialize: $(K_{l,1}^{(1)})_{l\in[L]}$ with stabilizing $K_{l,1}^{(1,1)}$ and $K_{l,1}^{(1,2)} = 0$, $K_{l,2}^{(0)} = 0$, $\bar{\mathsf{Z}}^{(1)} = 0$

for $s \in \{1, \ldots, S_1 - 1\}$ **do**

▷ Each generic agent performs ZSO to update $K_{l,1}^{(s+1)}$

$$K_{l,1}^{(s+1)} = ZSO((K_{l,1}^{(1)}, K_{l,2}^{(0)}), 1, \bar{\mathsf{Z}}^{(s)}, R_1, r_1, \eta_1, k_1)$$

Simulator uses $(K_{l,1}^{(s+1)}, K_{l,2}^{(0)})$ to obtain $\bar{\mathsf{Z}}^{(s+1)}$.

end for

Initialize: $K_{l,2}^{(1)}$

for $s \in \{1, \ldots, S_2\}$ **do**

▷ Each generic agent performs ZSO to obtain $K_{l,2}^{(s+1)}$

$$K_{l,2}^{(s+1)} = ZSO((K_{l,1}^{(S_1)}, K_{l,2}^{(1)}), 2, \bar{\mathsf{Z}}^{(s+S_1)}, R_2, r_2, \eta_2, k_2)$$

Simulator uses $(K_{l,1}^{(S_1)}, K_{l,2}^{(s+1)})$ to obtain $\bar{\mathsf{Z}}^{(S_1+s+1)}$.

end for

Output: $(K_{l,1}^{(S_1)}, K_{l,2}^{(S_2)})_{l\in[L]}, \bar{\mathsf{Z}}^{(S_1+S_2)}$

ZSO updates linear controller $\forall l \in [L]$

Simulator updates linear MF

Estimating Linear Terms

ZSO updates control offset $\forall l \in [L]$

Simulator updates MF offset

Estimating Offset Terms

ZSO Algorithm

- ZSO is a stochastic gradient descent algorithm (SGD)
- SGD is performed using smoothed gradient of a function f

ZSO pseudocode

- Initialize x
- For $r \in \{1, \dots, R\}$

  \\ *Generate k perturbations with norm r*

  - Generate $e^i \sim \mathbb{S}^1(r), \forall\, i \in \{1, \dots, k\}$
  - Compute smoothed gradient $\widehat{\nabla} f$

$$\widehat{\nabla} f(x) = \frac{1}{k} \sum_{l=1}^{k} \frac{mL}{r^2} f(x + e^l) e^l$$

  - $x \leftarrow x - \eta \widehat{\nabla} f(x)$

Finite Sample guarantees for ZSO (Linear)

- Finite sample convergence for ZSO in *first* part of Algorithm

- Specifies values for:
  - Number of iterations R_1
  - Smoothing radius r_1
  - Mini-batch size k_1
  - Learning rate η_1

➔ High confidence bound on the estimation error ϵ_1

- Depends on properties of J_l^1

(such as Lipschitz constant, smoothness constant, gradient domination constant, and local radius)

Lemma 1 *For a given linear mean-field trajectory* $\bar{Z}$ *and* $\epsilon_1, \delta_1 > 0$, *if the smoothing radius* r_1, *the learning rate* η_1 *and the mini-batch size* k_1 *are chosen such that*

$$r_1 = \frac{1}{8\varphi_1^l} \min\left(\theta_1^l \mu^l \sqrt{\frac{\epsilon_1}{240}}, \frac{1}{\varphi_1^l}\sqrt{\frac{\epsilon_1 \mu^l}{30}}\right), \eta_1 = \min\left(1, \frac{1}{8\varphi_1^l}, \frac{\rho_1^l}{\sqrt{\mu^l}/32 + \varphi_1^l + \lambda_1^l}\right)$$

$$k_1 = 1024 \frac{(mL)^2}{r_1^2}\left(J_l(K_i^{(0)}) + \frac{\lambda_1^l}{\rho_1}\right)^2 \log\left(\frac{2mL}{\delta}\right)\frac{1}{\mu^l \epsilon_1}$$

and the number of iterations is $R_1 = \frac{8}{\eta_1 \mu^l} \log\left(\frac{2(J_l^1(K_{l,1}^{(1)}) - J_l^1(K_{l,1}^*))}{\epsilon_1}\right)$, *then*

$$J_l^1(K_{l,1}^{(R_1)}) - J_l^1(\bar{K}_{l,1}^*) \leq \frac{\epsilon_1}{2},$$

$$\|K_{l,1}^{(R_1)} - \bar{K}_{l,1}^*\|_F \leq \sigma_{\min}^{-1}(\bar{\Sigma}^l)\sigma_{\min}^{-1}(C_U^l)\frac{\epsilon_1}{2}$$

with probability at least $1 - \delta_1 R_1$, *and the control gain* $\bar{K}_{l,1}^* = \operatorname{argmin}_{K_{l,1}} J_l^1(K_{l,1}, \bar{Z})$.

Finite Sample guarantees for ZSO (Offset)

- Finite sample convergence for ZSO in *second* part of Algorithm

- Specifies values for:
  - Number of iterations R_2
  - Smoothing radius r_2
  - Mini-batch size k_2
  - Learning rate η_2

➔ High confidence bound on the estimation error ϵ_2

- Depends on properties of J_l^2

- Provides convergence of controller to arbitrary threshold

Lemma 2 *For a given affine mean-field trajectory* $\bar{\mathsf{Z}}$, *control gain* $K_{l,1}$ *and* $\epsilon_2, \delta_2 > 0$, *if the smoothing radius* r_2, *the learning rate* η_2 *and the mini-batch size* k_2 *are chosen such that*

$$r_2 = \min\left(1, \rho_2^l, \frac{\nu^l \epsilon_2}{32\varphi_2^l \lambda_2^l}\right), \quad \eta_2 = \min\left(\frac{1}{\varphi_2^l}, \rho_2^l\left(\frac{\nu^l}{32} + \varphi_2^l + \lambda_2^l\right)^{-1}\right)$$

$$k_2 = 1024\frac{m^2}{r_2^2}\left(J_l(K_{l,2}^{(0)}) + \frac{\lambda_2^l}{\rho_2^l}\right)^2 \log\left(\frac{2m}{\delta}\right) \max\left(\frac{1}{\nu^l \epsilon_2}, \left(\frac{\lambda_2^l}{\nu^l \epsilon_2}\right)^2\right)$$

and the number of inner loop iterations is

$$R_2 = \frac{1}{\nu^l \eta_2} \log\left(\frac{4(J_l^2((K_{l,1}, K_{l,2}^{(1)}), \bar{\mathsf{Z}}) - J_l^2((K_{l,1}, \bar{K}_{l,2}^*), \bar{\mathsf{Z}}))}{\epsilon_2}\right)$$

then the difference between the output cost $J_l^2((K_{l,1}, K_{l,2}^{(R_2)}), \bar{\mathsf{Z}})$ *and the optimal cost* $J_l^2((K_{l,1}, \bar{K}_{l,2}^*), \bar{\mathsf{Z}})$ *is*

$$J_l^2((K_{l,1}, K_{l,2}^{(R_2)}), \bar{\mathsf{Z}}) - J_l^2((K_{l,1}, \bar{K}_{l,2}^*), \bar{\mathsf{Z}}) \leq \epsilon_2/2,$$

$$\|K_{l,2}^{(R_2)} - \bar{K}_{l,2}^*\|_2 \leq \sqrt{\frac{\epsilon_2}{\nu^l}}$$

with probability at least $1 - \delta_2 R_2$, *and* $\bar{K}_{l,2}^* = \operatorname{argmin}_{K_{l,2}} J_l^2((K_{l,1}, K_{l,2}), \bar{\mathsf{Z}})$.

Finite Sample guarantees for ZSO-RL

- Finite sample convergence for RL algorithm under standard assumptions

- Specifies values for:
  - Number of iterations of first and second parts of algo S_1 and S_2
  - Convergence and confidence bounds for ZSO algorithms $\varepsilon_1, \delta_2, \varepsilon_2$ and δ_2

- Provides convergence of MFE to arbitrary threshold **with high confidence**

Theorem 3 *If the outer loop iterations S_1 and S_2 are defined such that,*

$$S_1 = \frac{1}{1-T_1}\log\left(\frac{2\|\mathsf{F}^{(1)} - \mathsf{F}^*\|_F}{\epsilon}\right), \quad S_2 = \frac{1}{1-T_2}\log\left(\frac{2\|\bar{\mathsf{C}}^{(1)} - \bar{\mathsf{C}}^*\|_2}{\epsilon}\right), \quad (26)$$

$\epsilon_1, \delta_1, \epsilon_2, \delta_2$ are defined s.t.

$$\epsilon_1 = \frac{(1-T_1)\epsilon}{\|\mathsf{B}\|_F \sum_{l\in[L]} \sigma_{\min}^{-1}(\bar{\Sigma}^l)\sigma_{\min}^{-1}(C_U^l)}, \delta_1 = \frac{\delta}{S_1 R_1}, \epsilon_2 = \epsilon^2, \delta_2 = \frac{\delta}{S_2 R_2} \quad (27)$$

and the parameters $r_1, r_2, \eta_1, \eta_2, k_1, k_2, R_1, R_2$ are defined as in the statements of Lemmas 1 and 2, then, the error between the approximate MFE $\left(\left(K_{l,1}^{(S_1)}, K_{l,2}^{(S_2)}\right)_{l\in[L]}, \bar{\mathsf{Z}}^{(S_1+S_2)}\right)$ and the MFE $\left(\left(K_l^*\right)_{l\in[L]}, \bar{\mathsf{Z}}^*\right)$ is

$$\|\mathsf{F}^{(S_1)} - \mathsf{F}^*\|_F \le \epsilon, \quad \|K_{l,1}^{(S_1)} - K_{l,1}^*\|_F \le D_l^2\epsilon \quad (28)$$

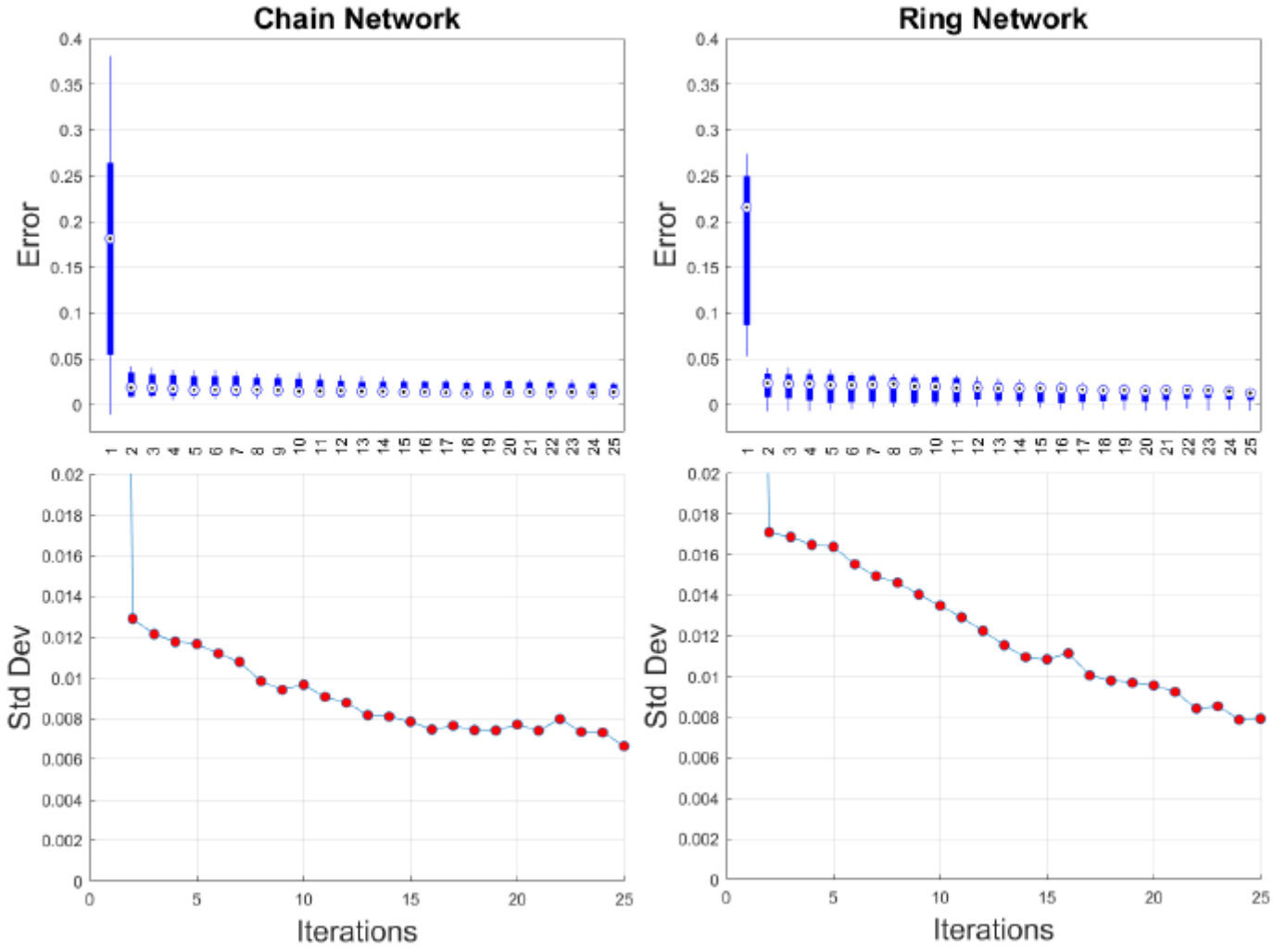
and $\mathsf{F}^{(s)}$ are stable $\forall s \in [S_1]$ with probability at least $1-\delta$. Furthermore if $\epsilon \le \min\left(1, \frac{1-T_2}{2D^3\|\mathsf{B}\|_2}\right)$, then

$$\|\mathsf{C}^{(S_2)} - \mathsf{C}^*\|_2 \le D^4\epsilon, \quad \|K_{l,2}^{(S_2+1)} - K_{l,2}^*\|_2 \le D_l^5\epsilon, \forall l \in [L] \quad (29)$$

with probability at least $1-2\delta$.

Numerical Analysis

- Simulation of the algorithm for 3 populations in a
  - Chain network
  - Ring network
- Scalar state and action spaces
- Plots show estimation error and std. dev. of the MFE (10 runs and 25 iterations each of the algorithm; 1500 iterations and rollouts in ZSO)
- Fast initial drop and steady decline (consistent with exact MFE update)



Conclusion—What lies ahead (MFGs)

MFG framework provides a versatile setting for addressing some complex decision-making problems in MASs.

Several fruitful research opportunities exist toward broadening its applicability:

- Robustness through risk-sensitive objective functions
- Imperfect local state measurements for agents
- Populations not fixed in advance, but formed through clustering mechanisms
- What if agents do not obey the rules of the algorithm: irrational behavior and stubbornness
- Hierarchical decision structures and incentivization toward truthful revelation
- More general (nonlinear) models, and parametrization for learning MFE

Selected Challenging Problems in MARL (1/2)

- Partially observable setting: decentralized POMDP (Dec-POMDP)
  - Each agent has to maintain beliefs over the global state and actions of other agents as opposed to the case in POMDPs where beliefs are on states
  - Most existing algorithms are either not scalable or not backed by theory
- Partially observed stochastic games (POSG)
  - Yet another level of difficulty, due to the need to generate beliefs over the opponents' policies → complicated (complex) information state
- Non-stationarity
  - For both Dec-POMDP and POSG, an agent confronts non-stationary/time-varying environment due to actions of other agents

Selected Challenging Problems in MARL (2/2)

- MARL for zero-sum/general-sum SGs with function approximation
  - A global convergence theory for policy-based methods is still lacking, even for the LQ setting. Naïve policy gradient fails to converge
- Heterogeneous / Adversarial Agents (Safe/Robust MARL)
  - Not all agents may strictly follow the underlying model, and some may be compromised or attacked during the operation
  - Need to develop safe and robust algorithms that can accommodate (or be resilient to) such uncertainties
- Theoretical basis for Deep MARL
  - Some initial promising work for single-agent Deep RL, but none for MARL

Selected Additional References (1/2)

- M. L. Littman. “Markov games as a framework for multi-agent reinforcement learning,” *In Machine learning proceedings,* pages 157–163. Elsevier, 1994.
- C. Boutilier. “Planning, learning and coordination in multi-agent decision processes,” *In Conference on Theoretical Aspects of Rationality and Knowledge,* pages 195–210, 1996.
- M. L. Littman. “Friend-or-foe Q-learning in general-sum games,” *In International Conference on Machine Learning,* volume 1, pages 322–328, 2001.
- Y. Shoham, R. Powers, and T. Grenager. “Multi-agent reinforcement learning: A critical survey,” *Technical Report,* 2003.
- X. Wang and T. Sandholm. “Reinforcement learning to play an optimal Nash equilibrium in team Markov games,” *In Advances in Neural Information Processing Systems*, pages 1603–1610, 2003.
- J. Hu and M.P. Wellman. “Nash Q-learning for general-sum stochastic games,” *Journal of Machine Learning Research,* 4(11):1039–1069, 2003.
- T. Smith and R. Simmons. “Heuristic search value iteration for POMDPs,” *In Proceedings of the Conference on Uncertainty in Artificial Intelligence*, pages 520–527, 2004.
- D. Silver and J. Veness. “Monte-Carlo planning in large POMDPs,” *In Advances in Neural Information Processing Systems,* pages 2164–2172, 2010.
- S. Kar, J.M. Moura, and H.V. Poor. “QD-learning: A collaborative distributed strategy for multi-agent reinforcement learning through consensus + innovations,” *IEEE Transactions on Signal Processing*, 61(7):1848–1862, 2013.
- J. Foerster, Y.M. Assael, N. Freitas, and S. Whiteson. “Learning to communicate with deep multi-agent reinforcement learning,” *In Advances in Neural Information Processing Systems*, pages 2137–2145, 2016.
- J. Perolat, F. Strub, B. Piot, and O. Pietquin. “Learning Nash equilibrium for general-sum Markov games from batch data.,” *arXiv preprint arXiv:1606.08718*, 2016.

Selected Additional References (2/2)

- R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch. “Multi-agent actor-critic for mixed cooperative competitive environments,” *arXiv preprint arXiv:1706.02275*, 2017.
- G. Arslan and S. Yüksel, “Decentralized Q-learning for stochastic teams and games,” *IEEE Transactions on Automatic Control*, vol. 62, no. 4, pp. 1545–1558, 2017.
- A. Mathkar and V. S. Borkar. “Distributed reinforcement learning via gossip,” *IEEE Transactions on Automatic Control* , 62(3):1465–1470, 2017.
- J. K. Gupta, M. Egorov, and M. Kochenderfer. “Cooperative multi- agent control using deep reinforcement learning,” *In International Conference on Autonomous Agents and Multiagent Systems,* pages 66–83. Springer, 2017.
- E. V. Mazumdar, M. I Jordan, and S. S. Sastry. “On finding local Nash equilibria (and only local Nash equilibria) in zero-sum games,” *arXiv preprint arXiv:1901.00838,* 2019.
- C. Jin, P. Netrapalli, and M. I Jordan. “Minmax optimization: Stable limit points of gradient descent ascent are locally optimal,” *arXiv preprint arXiv:1902.00618*, 2019.
- K. Zhang, X Zhang, B. Hu, and T. Başar. Derivative-free policy optimization for linear risk-sensitive and robust control design: Implicit regularization and sample complexity. *Proc. 38th International Conference on Machine Learning: Workshop on Reinforcement Learning Theory (ICML 2021; July 19-23, 2021; virtual), PMLR 139, 2021*. .
- M. Sayin, K. Zhang, D.S. Leslie, T. Başar, and A. Ozdaglar. Decentralized Q-learning in zero-sum Markov games. *Proc. 35th Conference on Neural Information Processing Systems (NeurIPS 2021; December 6-14, 2021; virtual), Sydney, Australia.*
- M.A. uz Zaman, M. Lauriere, A. Koppel, and T. Başar. Robust multi-agent reinforcement learning: A mean-field perspective. *Proc. 6th Annual Learning for Dynamics and Control Conference (L4DC 2024),* Oxford, England, July 15-17, 2024, PMLR 242:770-783.
- W. Mao, H. Qiu, C. Wang, H. Franke, Z.T. Kalbarczyk, and T. Başar. Convergence to (coarse) correlated equilibria in full-information general-sum Markov games. Proc. 6th Annual L4DCl Conference (L4DC 2024), Oxford, July 2024, PMLR 242:361-374.