RL and Equilibria for Multi-Agent Dynamical Systems in the High Population Regime

Tamer Başar
University of Illinois Urbana-Champaign

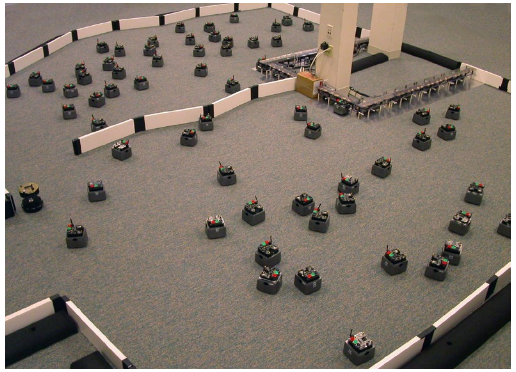
IMSI Workshop on Foundations of MA & MF RL
University of Chicago
May 18-22, 2026

Outline

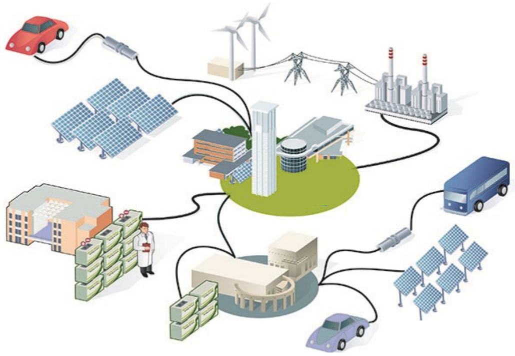
- Multi-agent dynamical systems and appropriateness of the framework of (stochastic) dynamic games along with underlying challenges arising due to informational asymmetry
- The role of mean-field games (MFGs) to alleviate the challenges in the high population regime, and the associated solution concept of mean-field equilibrium (MFE)
- Computation of MFE along with learning schemes
- Zero-order stochastic optimization (ZSO) based RL and finite sample guarantees
- Challenges in addressing networked games with delayed (spatial and information) coupling
- Approximating MFE in infinite horizon via finite-horizon MFE, and associated benefits in development of learning schemes for the former
- Illustration through multi-population LQ-MFGs—*consensus and dissensus*
- What lies in the future

Multi-agent Dynamical Systems

- Multi-agent systems (MAS) are ubiquitous



Robotics



Smart Grid

Unmanned Aerial Vehicles

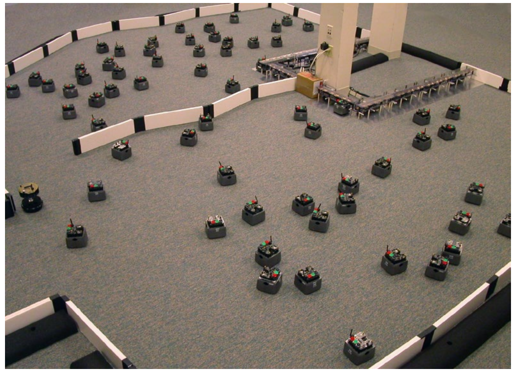
MOBA Video Games

Other selected applications:

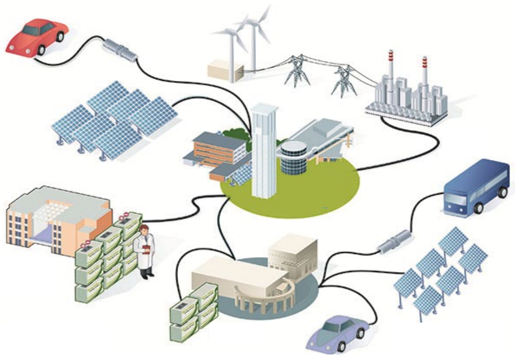
- Mobile sensor networks
- Distributed optimization (with topological and informational constraints)
- Social networks (evolution of opinions)

Multi-agent Dynamical Systems

- Multi-agent systems (MAS) are ubiquitous



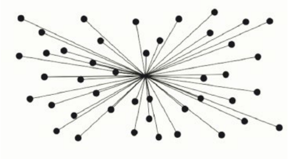
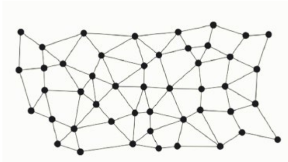
Robotics



Smart Grid

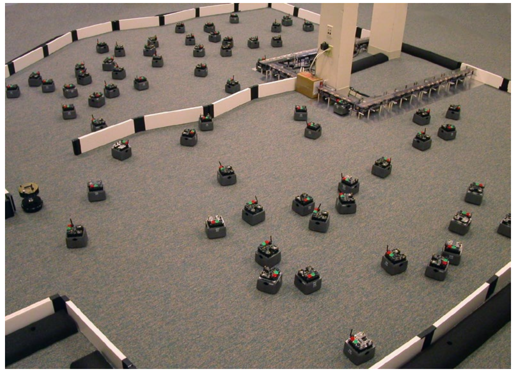
Unmanned Aerial Vehicles

MOBA Video Games

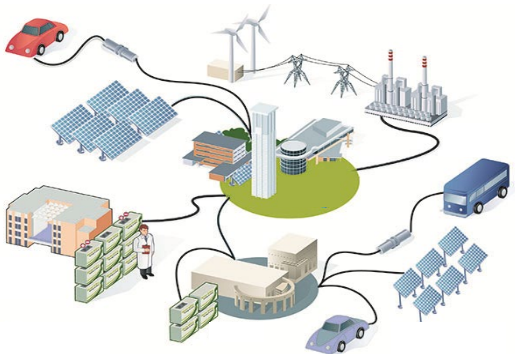
- Centralized protocol  vs Decentralized protocol  ✔
  - **Central** controller may not exist or may not be desirable in many MAS applications
  - Advantages of **decentralization**: i) resilient to attacks; ii) scalable; iii) privacy-preserving; iv) use of only local information

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, 1998 (1st version Academic Press, 1982).

[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

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

  an underlying network that governs connections, neighborhood interactions ($\mathcal{N}_{i,t}$ for i)
- *Information structures* (control policy of player i: $\boldsymbol{\gamma}_i \in \boldsymbol{\Gamma}_i$)

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

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

  Measurement feedback: $u_{i,t} = \boldsymbol{\gamma}_i(t;\, y_{i,s}, s=1, \ldots, t)$, $\quad y_{i,t} = h_{i,t}(x_{i,t}, x_{-i,t}, v_{i,t})$
- *Loss function* for player i (over t = 1, . . ., T) -- T could be ∞

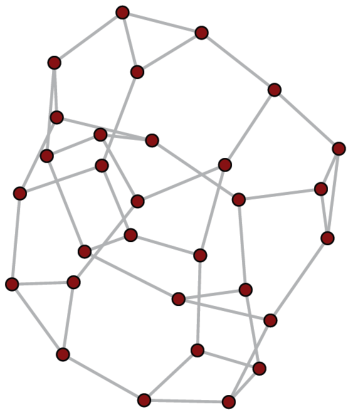
  $L_i(x_{[1,T]}, u_{[1,T]}) = (1/T) \sum_{t \in [1,T]} g_{i,t}(x_{it}, u_{it}, k_{it}(x_{-i,t}, u_{-i,t}))$

  Take expectations (for horizon [1,T]) with $u = \boldsymbol{\gamma}(\cdot)$: $\quad J_i(\boldsymbol{\gamma}_i, \boldsymbol{\gamma}_{-i})$ [normal form of game]
- *ϵ-Nash equilibrium* $\boldsymbol{\gamma}^{\epsilon}$: $\; J_i(\boldsymbol{\gamma}_i^{\epsilon}, \boldsymbol{\gamma}_{-i}^{\epsilon}) \leq J_i(\boldsymbol{\gamma}_i, \boldsymbol{\gamma}_{-i}^{\epsilon}) + \epsilon \quad \forall\, \boldsymbol{\gamma}_i \in \boldsymbol{\Gamma}_i$, $i = 1, \ldots, N$

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.

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.Frihauf, 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 *zeroth* 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: $\boldsymbol{\gamma}_i \in \boldsymbol{\Gamma}_i$) – 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 $u = \boldsymbol{\gamma}(\cdot)$: $\quad J_i(\boldsymbol{\gamma}_i, \boldsymbol{\gamma}_{-i})$ [normal form of game]
- *Stackelberg equilibrium (SE)* $\boldsymbol{\gamma}^*$:

$$J_0(\boldsymbol{\gamma}_0^*, \boldsymbol{\gamma}_{-0}^*) \leq \sup_{\beta \in R(\mu)} J_0(\mu, \boldsymbol{\beta}) \quad \forall\, \mu \in \boldsymbol{\Gamma}_0$$

  where $R(\mu)$ is the NE reaction set of N followers to each announced policy $\mu \in \boldsymbol{\Gamma}_0$ of L.

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

Challenges in deriving NE

- 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).[1]
- 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.[1]

****

[1]TB, G.J. Olsder. *Dynamic Noncooperative Game Theory*, SIAM, 1998.

Challenges in deriving NE

- 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).
- 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).

Responding to the challenge: Mean-Field Games[1,2,3]

- 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 $N \to \infty$.

****

[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.

Responding to the challenge: Mean-Field Games

- 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.

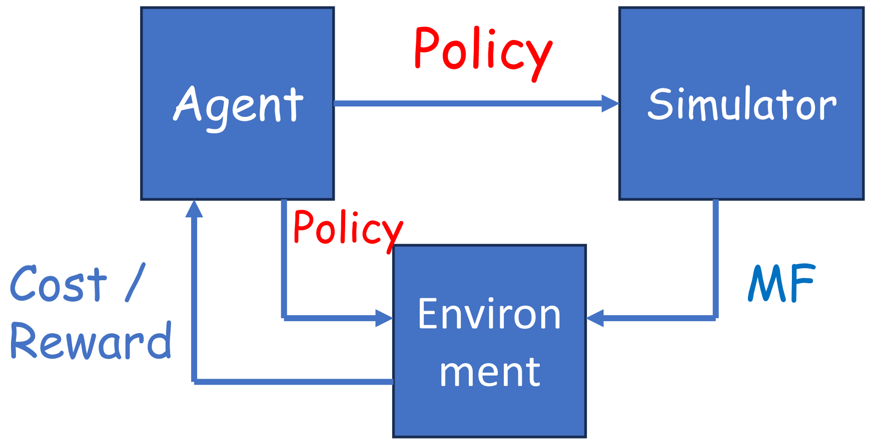
[6]N. Saldi, TB, M. Raginsky, "Approximate Markov-Nash equilibria for discrete-time risk-sensitive mean-field games," *Mathematics of Operations Research*, 45(4):1596-1620, November 2020.

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.

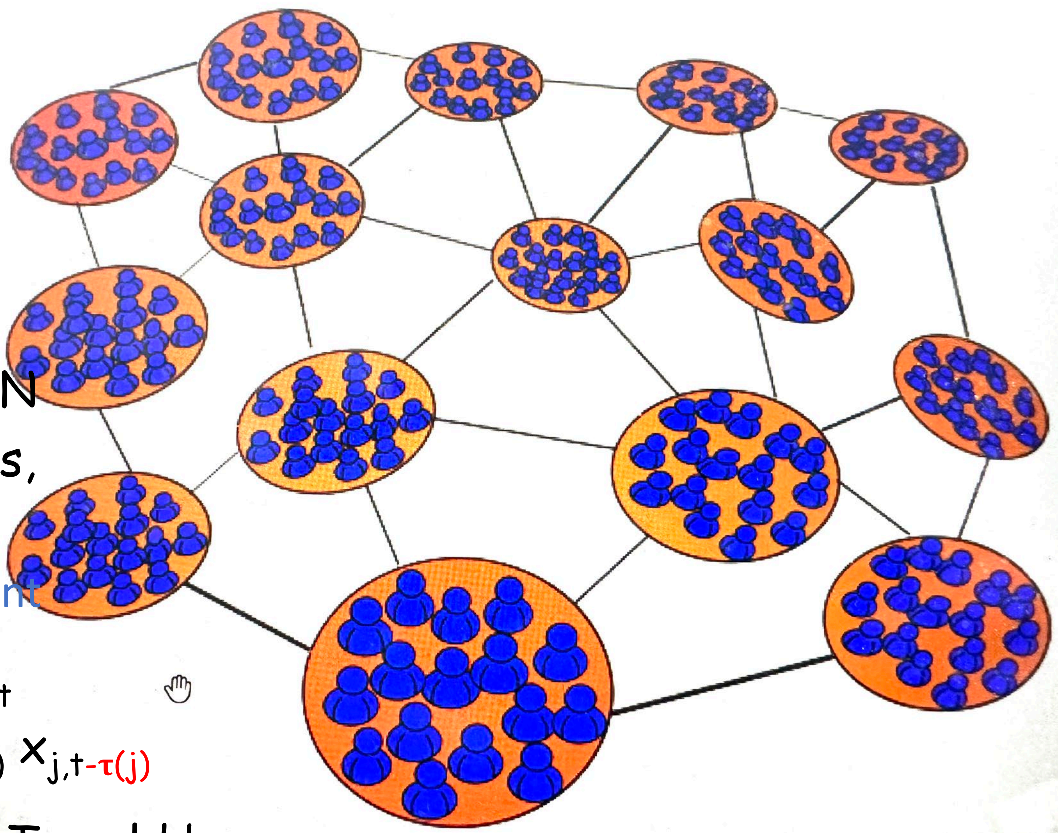
[2]T. Chen, K. Zhang, G.B. Giannakis, TB, "Communication-efficient policy gradient methods for distributed reinforcement learning," *IEEE TCNS*, 9(2):917-929, June 2022.

[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.

Switching Gears I: Networked Games with Delayed Coupling

Multiple dynamical systems, each one controlled by one or multiple agents with tight connections, but different dynamical systems are coupled loosely and with delayed connections as well as delayed information flow.



$x_t = (x_{1,t}, \ldots, x_{N,t})$; with state dynamics

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

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

Coupling delays could be > 1, and be system dependent

E.g. $c_{it}(x_{j,t-\tau(j)}, u_{j,t-\tau(j)}; j \in \mathcal{N}_{i,t})$, $\tau(j) > 0$, $j \in \mathcal{N}_{i,t}$

Or even, $c_{it}(x_{-i,t-\tau(i)}, u_{-i,t-\tau(i)}) = (1/|\mathcal{N}_{i,t}|) \sum_{j\in\mathcal{N}(i,t)} x_{j,t-\tau(j)}$

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

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

➔ *Looking for Nash equilibrium (or ϵ-Nash equilibrium)*

Delayed coupling leads to complexity even with perfect state information for all players

- Say 2 players (u,v), $x_t = (x_{1,t}, x_{2,t})$; $x_{t+2} = f_t(x_{t+1}, x_t, u_t, v_t)$, $+ w_t$ (high-order state dynamics)
  $\{w_t\}$ an i.i.d. process with a prob dist that assigns +ve prob mass to every open subset of $\mathbb{R}^n$
- Under standard stage-additive cost functions for the players, the standard approach to derivation of state-feedback NE policies does not apply*

*TB, "Informational uniqueness of closed-loop Nash equilibria for a class of nonstandard dynamic games," *JOTA*, 464(4):409-419, 1985

Delayed coupling leads to complexity even with perfect state information for all players

- Say 2 players (u,v), $x_t = (x_{1,t}, x_{2,t})$; $x_{t+2} = f_t(x_{t+1}, x_t, u_t, v_t), + w_t$ (high-order state dynamics) $\{w_t\}$ an i.i.d. process with a prob dist that assigns +ve prob mass to every open subset of $\mathbb{R}^n$
- Under standard stage-additive cost functions for the players, the standard approach to derivation of state-feedback NE policies does not apply*
- Every NE is *informationally unique* and can be obtained through an iterative process that sweeps the time interval twice, in forward and retrograde time. And these policies entail full state memory, as opposed to those for games with 1st-order dynamics.*
- NOT implementable in infinite horizon, and also when there is a high population of players

*TB, "Informational uniqueness of closed-loop Nash equilibria for a class of nonstandard dynamic games," *JOTA*, 464(4):409-419, 1985

High population of agents: MFE is manageable **

$x_t = (x_{1,t}, \ldots, x_{N,t})$; with state dynamics

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

an underlying network that governs connections,

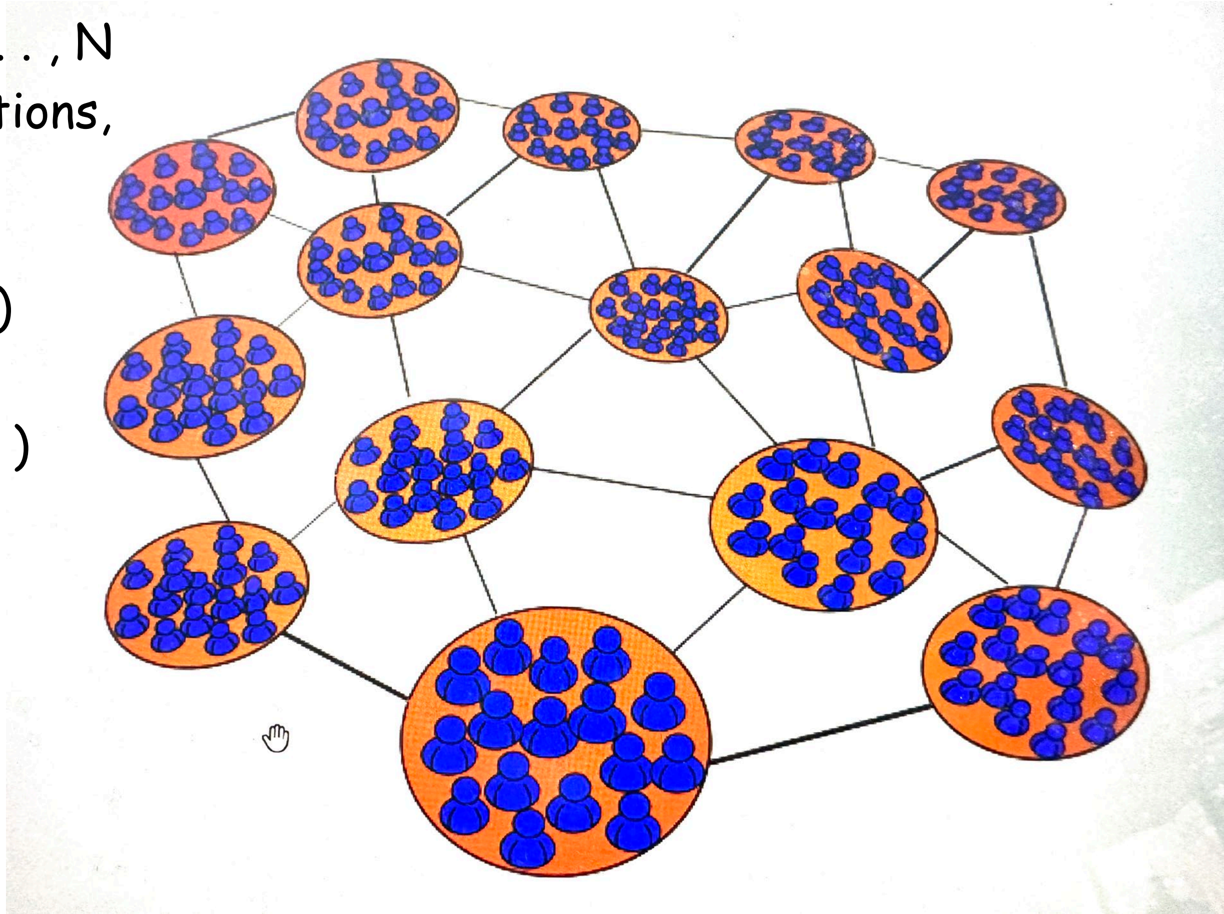
neighborhood interactions ($\mathcal{N}_{i,t}$ for i)

$c_{it}(x_{-i,t-1}, u_{-i,t-1}) = (1/|\mathcal{N}_{i,t}|) \sum_{j\in\mathcal{N}(i,t)} x_{j,t-1}$

Loss function for player i (over t = 1, . . ., T)

$L_i(x_{[1,T]}, u_{[1,T]}) =$

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



** TB, "Nash equilibria of networked games with delayed coupling in the high population regime," *Central European J Operations Research*, 33(2):415-427, June 2025.

High population of agents: MFE is manageable **

$x_t = (x_{1,t}, \ldots, x_{N,t})$; with state dynamics

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

an underlying network that governs connections,

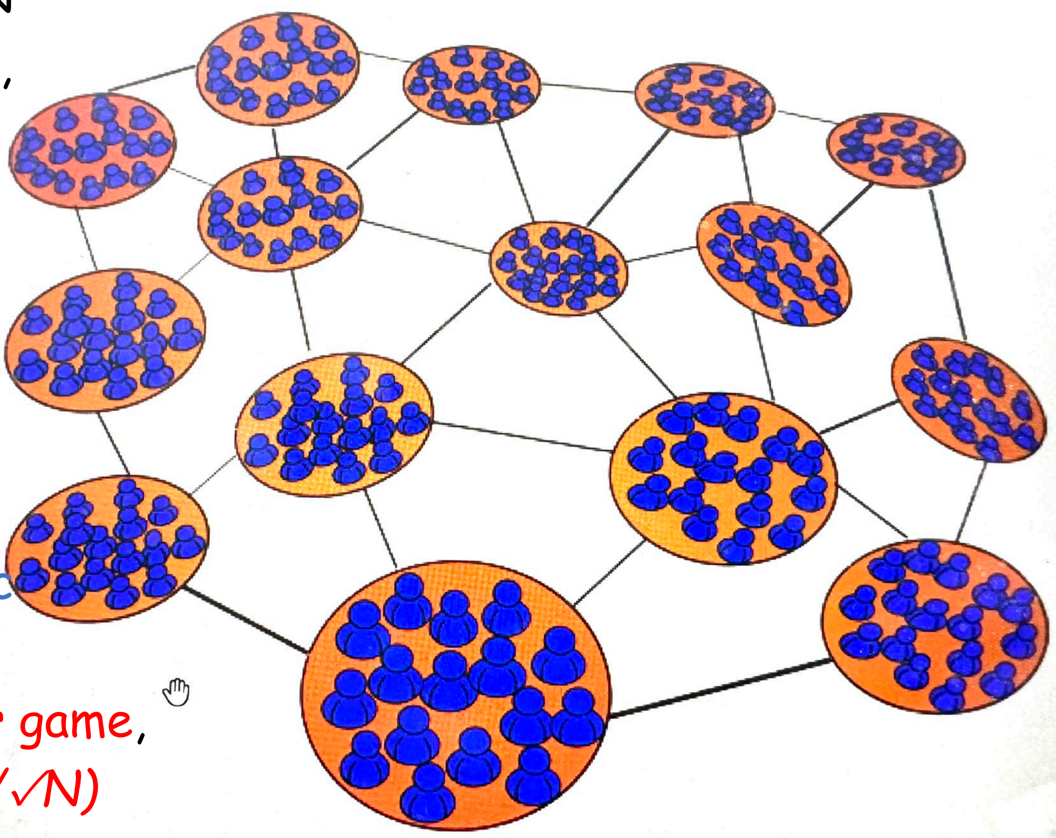
neighborhood interactions ($\mathcal{N}_{i,t}$ for i)

$c_{it}(x_{-i,t-1}, u_{-i,t-1}) = (1/|\mathcal{N}_{i,t}|) \sum_{j \in \mathcal{N}(i,t)} x_{j,t-1}$

Loss function for player i (over t = 1, . . ., T)

$L_i(x_{[1,T]}, u_{[1,T]}) =$

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



- MFE policies of agents, $\{{}^*\gamma_t(x_t)\}$, use current values of local states and some deterministic terms generated by delayed difference eqs
- $\{{}^*\gamma_t(x_t)\}$, when used by players in an N-player game, provides an *ϵ-Nash equilibrium, with $\epsilon = O(1/\sqrt{N})$*

**TB, "Nash equilibria of networked games with delayed coupling in the high population regime," *Central European J Operations Research*, 33(2):415-427, June 2025.

Shifting Gears II. "Closeness" in MFGs[5]

- Q: Can finite-horizon (FH) MFE effectively approximate infinite-horizon (IH) MFE?
- Answer is yes, under fairly general conditions, when cost functions are discounted, using the general setting of [6] for discrete-time MFGs.
- Any accumulation point of FH MFE, under weak convergence, as T increases, constitutes a non-stationary IH MFE. In turn, the latter can be shown to have stationary IH MFE as an accumulation point, thus making FH MFE an effective approximator for stationary IH MFE and establishing robustness of MFE with respect to the time horizon.
- Learning schemes developed for FH MFGs can be used for corresponding IH MFGs
- One byproduct: when two MFGs have FH MFE that are close under the 1-Wasserstein metric, the corresponding stationary MFE are also close under the same metric.

****

[5]U. Aydin, TB, N. Saldi, "Approximation of discrete-time infinite-horizon mean-field equilibria via finite-horizon mean-field equilibria," *Proc IEEE CDC*, Dec 2025, pp. 305-310. Extended version as arXiv:2509.01039, September 1, 2025.

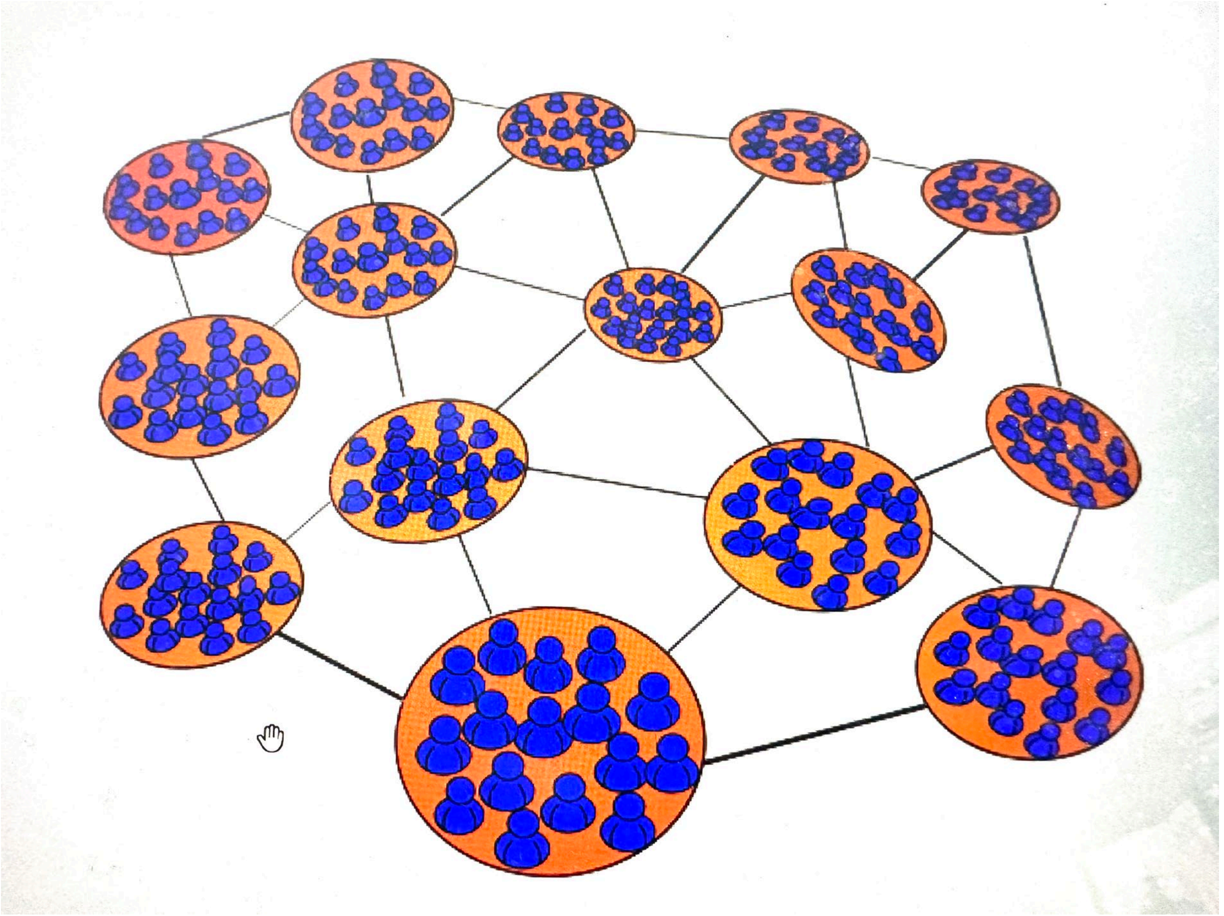
[6]N. Saldi, TB, M. Raginsky, "Markov-Nash equilibria in mean-field games with discounted cost," *SIAM J Control and Optimization*, 56(6):4256-4287, 2018.

An Illustration of MFE & RL in Multi-Population MFGs

Earlier discussion on MFE and learning schemes now put into action within the 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

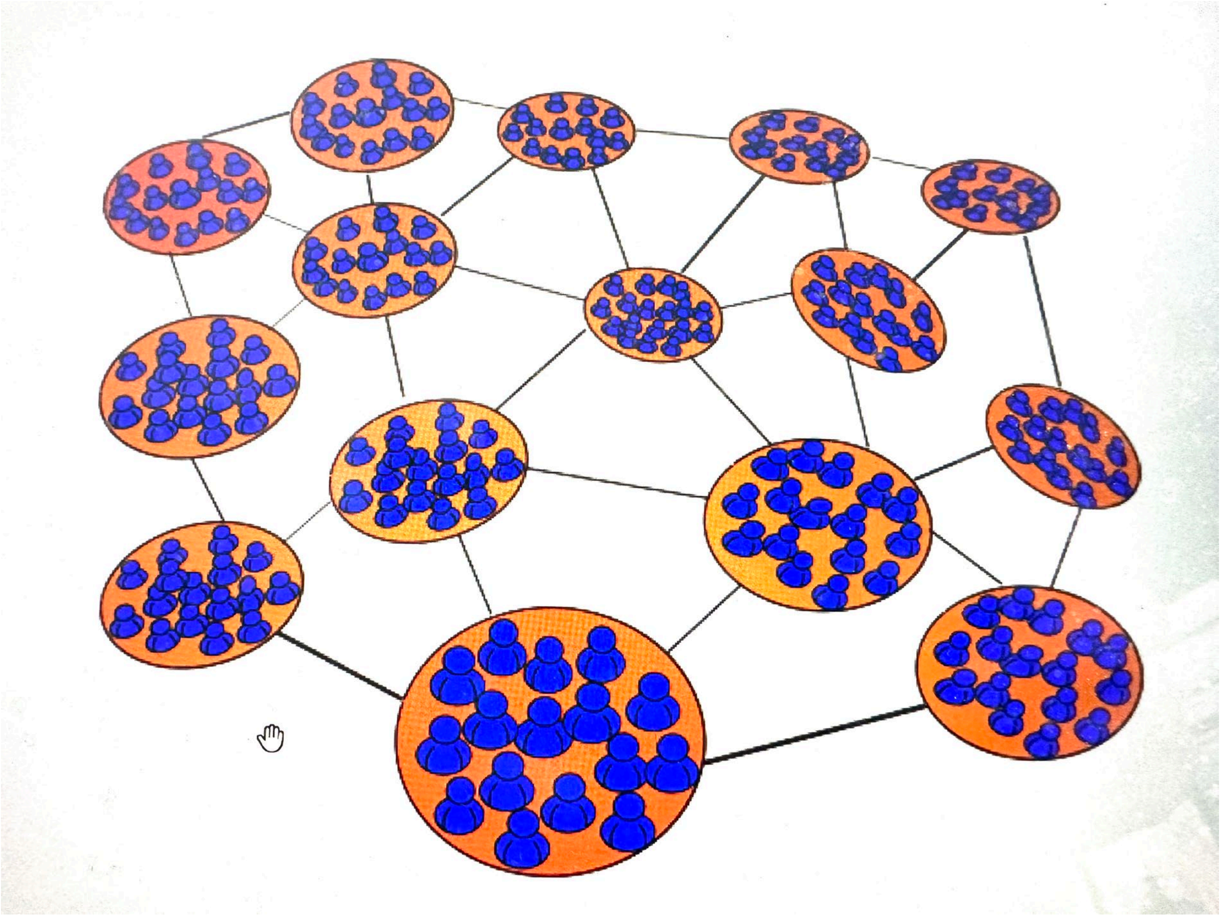
Multiple types (populations) of agents where `like' ones want to stay close to each other (consensus) whereas different populations want to have some separation between them (dissensus).



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?

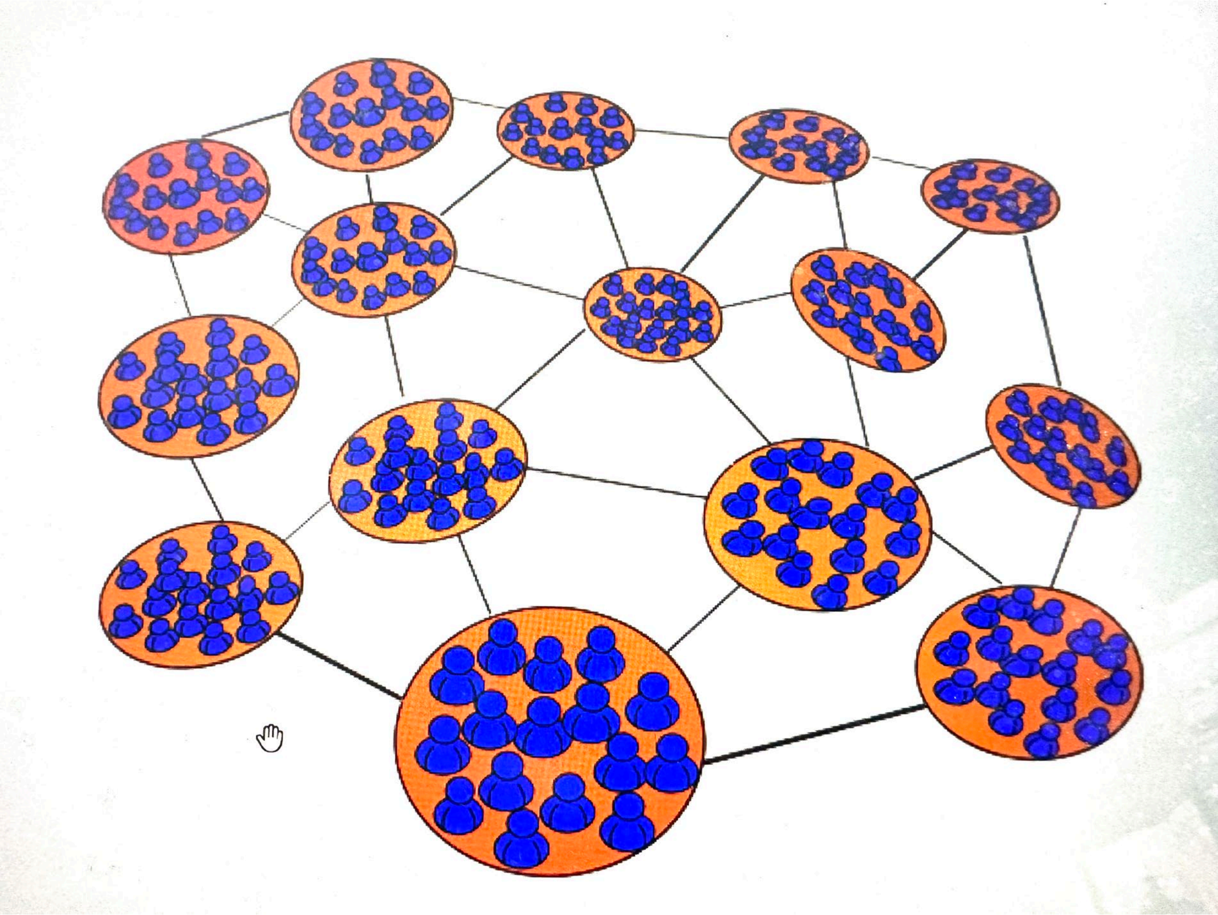


A specific framework for LQ MFGs

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 different populations

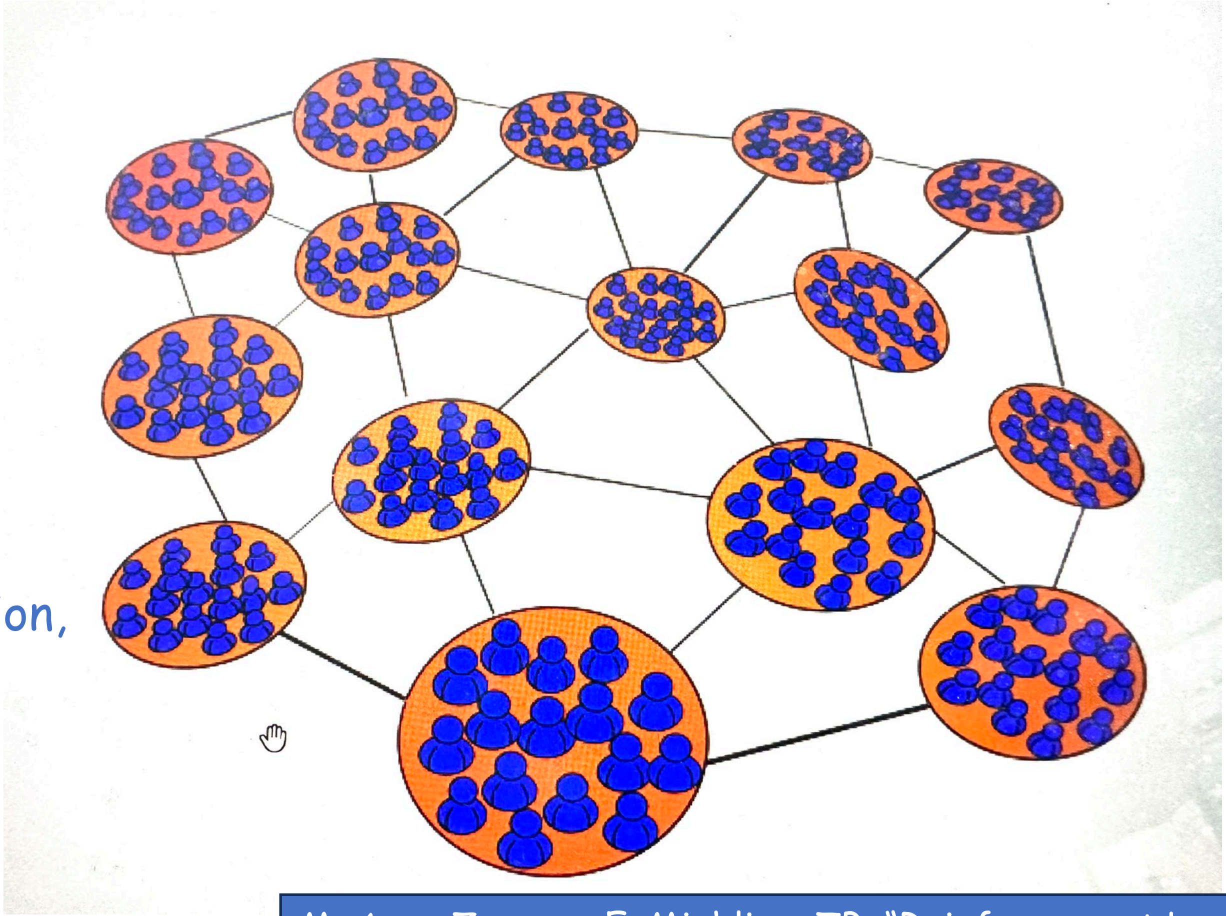


A specific framework for LQ MFGs

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 different populations



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

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 $C_Z \geq 0$)

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

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:

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

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)

<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 $\bar{Z} = \boldsymbol{\Lambda}(\phi)$ and $\phi \in \boldsymbol{\Psi}(\bar{Z})$ → 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

L populations, with N_l agents in population $l \in [L]$

- 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):

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

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

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 $\bar{\mathbf{Z}} := (\bar{Z}^1, \ldots, \bar{Z}^L)$,

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

$$+ \underbrace{\boxed{\left\|Z_t^l - \bar{Z}_t^l\right\|_{C_Z^{ll}}^2}}_{\text{Intra-Population consensus}} + \underbrace{\boxed{\sum_{\substack{k\in\mathcal{L}_l \\ k\neq l}} \left\|Z_t^l - \left(\beta^{lk} + \bar{Z}_t^k\right)\right\|_{C_Z^{lk}}^2\Bigg]}}_{\text{Inter-Population consensus (dissensus)}}$$

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,

$$\bar{Z}_{t+1}^{l} = A^{l}\bar{Z}_{t}^{l} + B^{l}\phi_{t}^{l}(\bar{Z}_{t}^{l}), \;\; \bar{Z}_{0}^{l} = \nu_{0}^{l} \;\; \text{for all } l \in [L]$$

  - <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

<u>MFE for Multi-Population LQ-MFG</u>

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

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

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

Multi-Population MFG (finite N_l)

- 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 $n \in [N_l]$, $l \in [L]$,*

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

where $J_{n,l}^{(\mathrm{N})}(\cdot)$ 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):

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

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

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

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

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

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

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

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 $K_{l,1}^{(1,2)} = 0$, $K_{l,2}^{(0)} = 0$, $\bar{\mathrm{Z}}^{(1)} = 0$
for $s \in \{1, \ldots, S_1 - 1\}$ **do**
  $\triangleright$ `Each generic agent performs ZSO to update` $K_{l,1}^{(s+1)}$

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

  Simulator uses $(K_{l,1}^{(s+1)}, K_{l,2}^{(0)})$ to obtain $\bar{\mathrm{Z}}^{(s+1)}$.
end for
Initialize: $K_{l,2}^{(1)}$
for $s \in \{1, \ldots, S_2\}$ **do**
  $\triangleright$ `Each generic agent performs ZSO to obtain` $K_{l,2}^{(s+1)}$

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

  Simulator uses $(K_{l,1}^{(S_1)}, K_{l,2}^{(s+1)})$ to obtain $\bar{\mathrm{Z}}^{(S_1+s+1)}$.
end for
Output: $(K_{l,1}^{(S_1)}, K_{l,2}^{(S_2)})_{l \in [L]}, \bar{\mathrm{Z}}^{(S_1+S_2)}$

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

<u>ZSO pseudocode</u>

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

  \\ *Generate k perturbations with norm r*

  - Generate $e^i \sim \mathbb{S}^1(r), \forall\, i \in \{1, \ldots, k\}$
  - 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* $\bar{\mathsf{Z}}$ *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* $\bar{K}_{l,1}^* = \mathrm{argmin}_{K_{l,1}} J_l^1(K_{l,1}, \bar{\mathsf{Z}})$.

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,*

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

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

$$\epsilon_1 = \frac{(1-T_1)\epsilon}{\|\mathsf{B}\|_F \sum_{l\in[L]} \sigma_{\min}^{-1}(\bar{\Sigma}^l)\sigma_{\min}^{-1}(C_U^l)}, \delta_1 = \frac{\delta}{S_1R_1}, \epsilon_2 = \epsilon^2, \delta_2 = \frac{\delta}{S_2R_2} \tag{27}$$

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 $\left((K_{l,1}^{(S_1)}, K_{l,2}^{(S_2)})_{l\in[L]}, \bar{\mathsf{Z}}^{(S_1+S_2)}\right)$ and the MFE $\left((K_l^*)_{l\in[L]}, \bar{\mathsf{Z}}^*\right)$ is

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

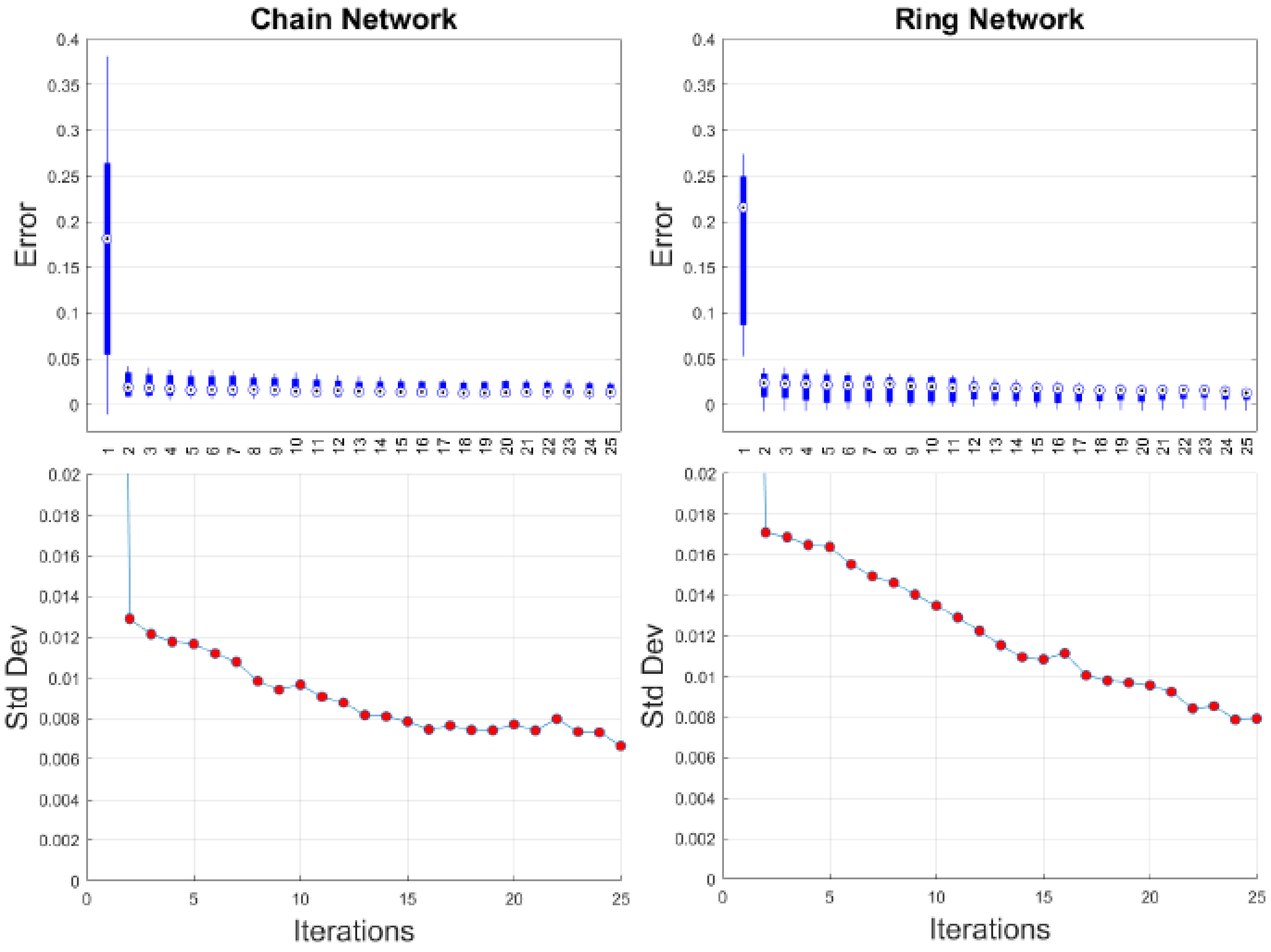
and $\mathsf{F}^{(s)}$ are stable $\forall s \in [S_1]$ with probability at least $1-\delta$. Furthermore if $\epsilon \leq \min\left(1, \frac{1-T_2}{2D^3\|\mathsf{B}\|_2}\right)$, then

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

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

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

Three Recent Books of Relevance

- S. Yüksel and T. Başar, *Stochastic Teams, Games and Control under Information Constraints*, Systems & Control: Foundations and Applications Series, Birkhäuser, Boston, July 2024
- T. Başar, B. Djehiche, H. Tembine, *Mean-Field-Type Game Theory I: Foundations and New Directions*, Springer International Publishing AG, Switzerland, February 2026
- T. Başar, B. Djehiche, H. Tembine, *Mean-Field-Type Game Theory II: Applications*, Springer International Publishing AG, Switzerland, April 2026

Thanks !