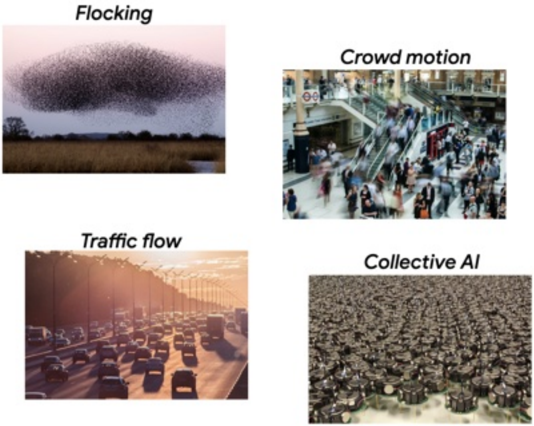
Reinforcement Learning for Finite Space Mean Field Type Games

Mathieu LAURIÈRE
NYU Shanghai
Joint work with Kai Shao, Jiacheng Shen, Chijie An (NYUSH)

CDC Workshop on
Large Population Teams: Control, Equilibria, and Learning
Milan, December 15, 2024

Outline

Many Agent Systems



[Image credits: Unsplash, Wikimedia Commons (Kilobots)]

with many strategic agents

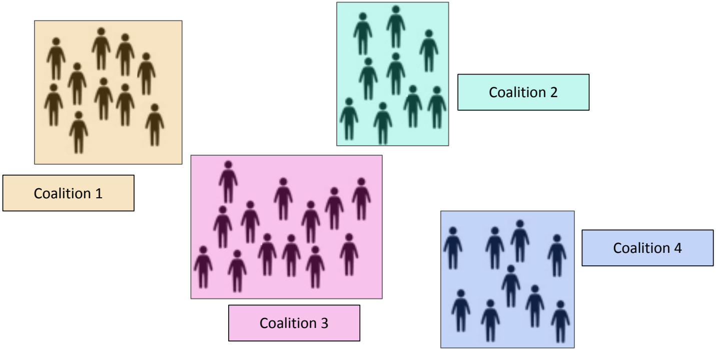
⇒ Game theory (here: dynamic & stochastic games) & Mean field approximation

Mean Field Models

- Fully non-cooperative: Nash equilibrium
  - Mean Field Games (MFGs) [Lasry and Lions, 2007] [Huang et al., 2006]
- Fully cooperative: social optimum
  - Mean Field Control (MFC) [Bensoussan et al., 2013]
  - Optimal control of McKean-Vlasov (MKV) dynamics [Carmona and Delarue, 2018]
  - Mean Field Markov Decision Processes (MFMDPs) [Motte and Pham, 2022], [Carmona et al., 2023]
- See [Bensoussan et al., 2013], [Gomes and Saúde, 2014], [Carmona and Delarue, 2018]
- Non-cooperative game between large (cooperative) coalitions/teams
  - Mean Field Type Games (MFTGs) [Tembine, 2014]

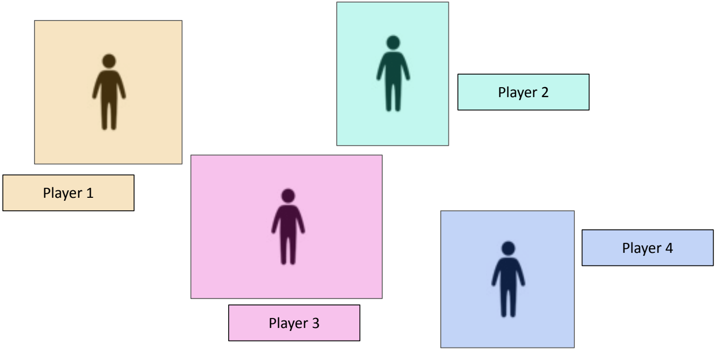
Intuition for MFTGs

Several **large coalitions**:



Intuition for MFTGs

Several **central players**, each of “mean field type”:



Intuition for MFTGs

- Finite number of "coalitions" (populations, groups)
- Each coalition has a large number of agents who cooperate (common objective)
- Agents of different coalitions are not cooperating
- Given the behavior of other coalitions, the agent of a given coalition are solving a social optimum problem
- Between coalitions: Nash equilibrium
- Other related concepts:
  - Multi-population MFGs, e.g. [Cirant, 2015], [Bensoussan et al., 2018]
  - Graphon games, e.g. [Parise and Ozdaglar, 2019], [Caines and Huang, 2019]
  - Mean field control games, e.g. [Angiuli et al., 2023] (infinite number of coalitions)

Intuition for MFTGs

- Finite number of "coalitions" (populations, groups)
- Each coalition has a large number of agents who cooperate (common objective)
- Agents of different coalitions are not cooperating
- Given the behavior of other coalitions, the agent of a given coalition are solving a social optimum problem
- Between coalitions: Nash equilibrium
- Other related concepts:
  - Multi-population MFGs, e.g. [Cirant, 2015], [Bensoussan et al., 2018]
  - Graphon games, e.g. [Parise and Ozdaglar, 2019], [Caines and Huang, 2019]
  - Mean field control games, e.g. [Angiuli et al., 2023] (infinite number of coalitions)

Some References on MFTGs

- Surveys and books: [Djehiche et al., 2017], [Tembine, 2017], [Barreiro-Gomez and Tembine, 2021]
- Applications: blockchain token economics [Barreiro-Gomez and Tembine, 2019], risk-sensitive control [Tembine, 2015] or more broadly in engineering [Barreiro-Gomez and Tembine, 2021]
- "Mean field games among teams" [Subramanian et al., 2023]
- "Team-against-team mean field problems" [Sanjari et al., 2023], [Yüksel and Başar, 2024]
- Special case: **zero-sum** MFTG [Cosso and Pham, 2019], [Carmona et al., 2020], [Başar and Moon, 2021], [Guan et al., 2024]
- **RL** for zero-sum LQ MFTGs [Carmona et al., 2020], [uz Zaman et al., 2024], [Zaman et al., 2024]: policy-gradient using the linear form of the optimal control
- Missing: **RL** methods for **general MFTGs**

Our Contributions

- Discrete-time, finite-state MFTGs as approximation of finite-player games
- Nash Q-Learning algorithm after quantization of simplex
- Deep RL algorithm based on DDPG
- Numerical experiments

Outline

Finite-Agent Model: Dynamics

- Game between m groups of many agents; each group: "coalition"
- In other words: m central players
- N_i denote the number of individual agents in coalition i
- $\Delta(S^i)$ and $\Delta(A^i)$ be the sets of probability distributions on S^i and A^i
- Agent j in coalition i has a state x_t^{ij} at time t
- The state of coalition i is characterized by the empirical distribution

$$\mu_t^{i,\bar{N}} = \frac{1}{N_i} \sum_{j=1}^{N_i} \delta_{x_t^{ij}} \in \Delta(S^i)$$

- The state of the whole population is characterized by the joint empirical distribution: $\mu_t^{\bar{N}} = (\mu_t^{1,\bar{N}}, \ldots, \mu_t^{m,\bar{N}})$
- The state of every agent $j \in [N_i]$ in coalition i evolves according to a transition kernel $p^i : S^i \times A^i \times \prod_{i'=1}^{m} \Delta(S^{i'}) \to \Delta(S^i)$
- If the agent takes action a_t^{ij} and the distribution is $\mu_t^{\bar{N}}$, then:

$$x_{t+1}^{ij} \sim p^i(\cdot | x_t^{ij}, a_t^{ij}, \mu_t^{\bar{N}})$$

and the agent obtains a reward $r_t^{ij} = r^i(x_t^{ij}, a_t^{ij}, \mu_t^{\bar{N}})$

Finite-Agent Model: Dynamics

- Game between m groups of many agents; each group: "coalition"
- In other words: m central players
- N_i denote the number of individual agents in coalition i
- $\Delta(S^i)$ and $\Delta(A^i)$ be the sets of probability distributions on S^i and A^i
- Agent j in coalition i has a state x_t^{ij} at time t
- The state of coalition i is characterized by the empirical distribution

$$\mu_t^{i,\bar{N}} = \frac{1}{N_i} \sum_{j=1}^{N_i} \delta_{x_t^{ij}} \in \Delta(S^i)$$

- The state of the whole population is characterized by the joint empirical distribution: $\mu_t^{\bar{N}} = (\mu_t^{1,\bar{N}}, \dots, \mu_t^{m,\bar{N}})$
- The state of every agent $j \in [N_i]$ in coalition i evolves according to a transition kernel $p^i : S^i \times A^i \times \prod_{i'=1}^{m} \Delta(S^{i'}) \to \Delta(S^i)$
- If the agent takes action a_t^{ij} and the distribution is $\mu_t^{\bar{N}}$, then:

$$x_{t+1}^{ij} \sim p^i(\cdot | x_t^{ij}, a_t^{ij}, \mu_t^{\bar{N}})$$

and the agent obtains a reward $r_t^{ij} = r^i(x_t^{ij}, a_t^{ij}, \mu_t^{\bar{N}})$

Finite-Agent Model: Dynamics

- Game between m groups of many agents; each group: "coalition"
- In other words: m central players
- N_i denote the number of individual agents in coalition i
- $\Delta(S^i)$ and $\Delta(A^i)$ be the sets of probability distributions on S^i and A^i
- Agent j in coalition i has a state x_t^{ij} at time t
- The state of coalition i is characterized by the empirical distribution

$$\mu_t^{i,\bar{N}} = \frac{1}{N_i} \sum_{j=1}^{N_i} \delta_{x_t^{ij}} \in \Delta(S^i)$$

- The state of the whole population is characterized by the joint empirical distribution: $\mu_t^{\bar{N}} = (\mu_t^{1,\bar{N}}, \dots, \mu_t^{m,\bar{N}})$
- The state of every agent $j \in [N_i]$ in coalition i evolves according to a transition kernel $p^i : S^i \times A^i \times \prod_{i'=1}^{m} \Delta(S^{i'}) \to \Delta(S^i)$
- If the agent takes action a_t^{ij} and the distribution is $\mu_t^{\bar{N}}$, then:

$$x_{t+1}^{ij} \sim p^i(\cdot | x_t^{ij}, a_t^{ij}, \mu_t^{\bar{N}})$$

and the agent obtains a reward $r_t^{ij} = r^i(x_t^{ij}, a_t^{ij}, \mu_t^{\bar{N}})$

Finite-Agent Model: Rewards

- All the agents in coalition i independently pick their actions according to a common policy $\pi^i : S^i \times \Delta(S^1) \times \cdots \times \Delta(S^m) \to \Delta(A^i)$, i.e., a_t^{ij} for all $j \in [N_i]$ are i.i.d. with distribution $\pi^i(\cdot|x_t^{ij}, \mu_t^{\bar{N}})$
- We denote by Π^i the set of such policies
- The **social reward** for the central player of population i is defined as:

$$J^{i,\bar{N}}(\pi^1, \ldots, \pi^m) = \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbb{E}\left[\sum_{t \geq 0} \gamma^t r_t^{ij}\right],$$

where $\gamma \in [0, 1)$ is a discount factor and $r_t^{ij} = r_t^i(x_t^{ij}, a_t^{ij}, \mu_t^{\bar{N}})$

Definition

A policy profile $(\pi_*^1, \ldots, \pi_*^m) \in \Pi^1 \times \cdots \times \Pi^m$ is a **Nash equilibrium** for the above finite-population game if: for all $i \in [m]$, for all $\pi^i \in \Pi^i$,

$$J^{i,\bar{N}}(\pi^i; \pi_*^{-i}) \leq J^{i,\bar{N}}(\pi_*^i; \pi_*^{-i}),$$

where π_*^{-i} denotes the vector of policies for central players in other coalitions except i.

Finite-Agent Model: Rewards

- All the agents in coalition i independently pick their actions according to a common policy $\pi^i : S^i \times \Delta(S^1) \times \cdots \times \Delta(S^m) \to \Delta(A^i)$, i.e., a_t^{ij} for all $j \in [N_i]$ are i.i.d. with distribution $\pi^i(\cdot | x_t^{ij}, \mu_t^{\bar{N}})$
- We denote by Π^i the set of such policies
- The **social reward** for the central player of population i is defined as:

$$J^{i,\bar{N}}(\pi^1, \ldots, \pi^m) = \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbb{E}\left[\sum_{t \geq 0} \gamma^t r_t^{ij} \right],$$

where $\gamma \in [0,1)$ is a discount factor and $r_t^{ij} = r_t^i(x_t^{ij}, a_t^{ij}, \mu_t^{\bar{N}})$

Definition

A policy profile $(\pi_*^1, \ldots, \pi_*^m) \in \Pi^1 \times \cdots \times \Pi^m$ is a **Nash equilibrium** for the above finite-population game if: for all $i \in [m]$, for all $\pi^i \in \Pi^i$,

$$J^{i,\bar{N}}(\pi^i; \pi_*^{-i}) \leq J^{i,\bar{N}}(\pi_*^i; \pi_*^{-i}),$$

where π_*^{-i} denotes the vector of policies for central players in other coalitions except i.

Finite-Agent Model: Rewards

- All the agents in coalition i independently pick their actions according to a common policy $\pi^i : S^i \times \Delta(S^1) \times \cdots \times \Delta(S^m) \to \Delta(A^i)$, i.e., a_t^{ij} for all $j \in [N_i]$ are i.i.d. with distribution $\pi^i(\cdot | x_t^{ij}, \mu_t^{\bar{N}})$
- We denote by Π^i the set of such policies
- The **social reward** for the central player of population i is defined as:

$$J^{i,\bar{N}}(\pi^1, \ldots, \pi^m) = \frac{1}{N_i} \sum_{j=1}^{N_i} \mathbb{E}\left[\sum_{t \geq 0} \gamma^t r_t^{ij}\right],$$

where $\gamma \in [0, 1)$ is a discount factor and $r_t^{ij} = r_t^i(x_t^{ij}, a_t^{ij}, \mu_t^{\bar{N}})$

Definition

A policy profile $(\pi_*^1, \ldots, \pi_*^m) \in \Pi^1 \times \cdots \times \Pi^m$ is a **Nash equilibrium** for the above finite-population game if: for all $i \in [m]$, for all $\pi^i \in \Pi^i$,

$$J^{i,\bar{N}}(\pi^i; \pi_*^{-i}) \leq J^{i,\bar{N}}(\pi_*^i; \pi_*^{-i}),$$

where π_*^{-i} denotes the vector of policies for central players in other coalitions except i.

MFTG: Mean Field

We let $N_i \to +\infty$

- State of coalition i: $\mu_t^{i,\bar{N}} \to \mu_t^i \in \Delta(S^i)$ for each $i \in [m]$
- State of the whole population: $\mu_t^{\bar{N}} \to \mu_t = (\mu_t^1, \dots, \mu_t^m) \in \Delta(S^1) \times \cdots \times \Delta(S^m)$
- We will refer to the limiting distributions as the **mean-field** distributions
- More rigorously: Propagation of chaos
- We expect all the agents' states to evolve independently, interacting only through the mean-field distributions
- A representative agent in mean-field coalition i has a state $x_t^i \in S^i$ which evolves according to: $x_{t+1}^i \sim p^i(\cdot|x_t^i, a_t^i, \mu_t)$, $a_t^i \sim \pi^i(\cdot|x_t^i, \mu_t)$, where $\pi^i \in \Pi^i$ is the policy for coalition i
- We consider that this policy is chosen by a **central player** and then applied by all the infinitesimal **agents** in coalition i.

MFTG: Mean Field

We let $N_i \to +\infty$

- State of coalition i: $\mu_t^{i,\bar{N}} \to \mu_t^i \in \Delta(S^i)$ for each $i \in [m]$
- State of the whole population: $\mu_t^{\bar{N}} \to \mu_t = (\mu_t^1, \ldots, \mu_t^m) \in \Delta(S^1) \times \cdots \times \Delta(S^m)$
- We will refer to the limiting distributions as the **mean-field** distributions
- More rigorously: Propagation of chaos
- We expect all the agents' states to evolve independently, interacting only through the mean-field distributions
- A representative agent in mean-field coalition i has a state $x_t^i \in S^i$ which evolves according to: $x_{t+1}^i \sim p^i(\cdot|x_t^i, a_t^i, \mu_t)$, $a_t^i \sim \pi^i(\cdot|x_t^i, \mu_t)$, where $\pi^i \in \Pi^i$ is the policy for coalition i
- We consider that this policy is chosen by a **central player** and then applied by all the infinitesimal **agents** in coalition i.

MFTG: Mean Field

We let $N_i \to +\infty$

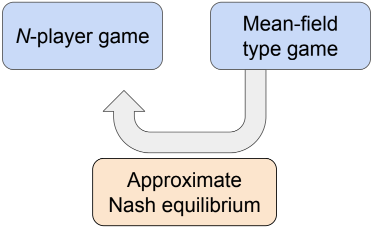
- State of coalition i: $\mu_t^{i,\bar{N}} \to \mu_t^i \in \Delta(S^i)$ for each $i \in [m]$
- State of the whole population: $\mu_t^{\bar{N}} \to \mu_t = (\mu_t^1, \dots, \mu_t^m) \in \Delta(S^1) \times \cdots \times \Delta(S^m)$
- We will refer to the limiting distributions as the **mean-field** distributions
- More rigorously: Propagation of chaos
- We expect all the agents' states to evolve independently, interacting only through the mean-field distributions
- A representative agent in mean-field coalition i has a state $x_t^i \in S^i$ which evolves according to: $x_{t+1}^i \sim p^i(\cdot|x_t^i, a_t^i, \mu_t)$, $a_t^i \sim \pi^i(\cdot|x_t^i, \mu_t)$, where $\pi^i \in \Pi^i$ is the policy for coalition i
- We consider that this policy is chosen by a **central player** and then applied by all the infinitesimal **agents** in coalition i.

MFTG: Rewards

- The total reward for coalition i is: $J^i(\pi^1, \dots, \pi^m) = \mathbb{E}\Big[\sum_{t\geq 0} \gamma^t r^i(x_t^i, a_t^i, \mu_t)\Big]$
- Goal: find a Nash equilibrium between the m central players.

Definition

A policy profile $(\pi_*^1, \dots, \pi_*^m) \in \Pi^1 \times \cdots \times \Pi^m$ is a **Nash equilibrium** for the above MFTG if: for all $i \in [m]$, for all $\pi^i \in \Pi^i$, $J^i(\pi^i; \pi_*^{-i}) \leq J^i(\pi_*^i; \pi_*^{-i})$, where π_*^{-i} denotes the vector of policies for players in other coalitions except i.



Approximate Equilibrium

Assumption

(a) *For each* $i \in [m]$*, the reward function* $r^i(x, a, \mu)$ *is bounded by a constant* $C_r > 0$ *and Lipschitz w.r.t.* μ *with constant* L_r*.*

(b) *The transition probability* $p(x'|x, a, \mu)$ *is* L_p*-Lipschitz continuous in* μ

(c) *The policies* $\pi(a|x, \mu)$ *satisfy the following Lipschitz bound:* $\|\pi(\cdot|x, \mu) - \pi(\cdot|x, \tilde{\mu})\|_1 \leq L_\pi d(\mu, \tilde{\mu})$ *for every* $x \in S^i$*, and* μ*,* $\tilde{\mu} \in \Delta(S^i)$*.*

Theorem (Approximate Nash equilibrium)

Let $(\pi_*^1, \ldots, \pi_*^m) \in \Pi^1 \times \cdots \times \Pi^m$ *be a Nash equilibrium for the MFTG. When the discount factor* γ *satisfies* $\gamma(1 + L_\pi + L_p) < 1$*, then*

$$\max_{\tilde{\pi}^i} J^{i,\bar{N}}(\tilde{\pi}^i; \pi_*^{-i}) \leq J^{i,\bar{N}}(\pi_*^i; \pi_*^{-i}) + \varepsilon(N),$$

for all $i \in [m]$*, with* $\varepsilon(N) = C \max_{i\in[m]} \left\{ |S^i| \sqrt{|A^i|} / \sqrt{N_i} \right\}$*, where* C *is a constant.*

- It justifies solving MFTGs because they provide an approximate solution for finite-agent games.
- Provides a rate of convergence, not just asymptotic convergence

Approximate Equilibrium

Assumption

(a) *For each* $i \in [m]$, *the reward function* $r^i(x, a, \mu)$ *is bounded by a constant* $C_r > 0$ *and Lipschitz w.r.t.* μ *with constant* L_r.

(b) *The transition probability* $p(x'|x, a, \mu)$ *is* L_p*-Lipschitz continuous in* μ

(c) *The policies* $\pi(a|x, \mu)$ *satisfy the following Lipschitz bound:* $\|\pi(\cdot|x, \mu) - \pi(\cdot|x, \tilde{\mu})\|_1 \le L_\pi d(\mu, \tilde{\mu})$ *for every* $x \in S^i$, *and* μ, $\tilde{\mu} \in \Delta(S^i)$.

Theorem (Approximate Nash equilibrium)

Let $(\pi_*^1, \dots, \pi_*^m) \in \Pi^1 \times \cdots \times \Pi^m$ *be a Nash equilibrium for the MFTG. When the discount factor* γ *satisfies* $\gamma(1 + L_\pi + L_p) < 1$, *then*

$$\max_{\tilde{\pi}^i} J^{i,\bar{N}}(\tilde{\pi}^i; \pi_*^{-i}) \le J^{i,\bar{N}}(\pi_*^i; \pi_*^{-i}) + \varepsilon(N),$$

for all $i \in [m]$, *with* $\varepsilon(N) = C \max_{i \in [m]} \left\{ |S^i| \sqrt{|A^i|} / \sqrt{N_i} \right\}$, *where* C *is a constant.*

- It justifies solving MFTGs because they provide an approximate solution for finite-agent games.
- Provides a rate of convergence, not just asymptotic convergence

Approximate Equilibrium

Assumption

(a) *For each* $i \in [m]$, *the reward function* $r^i(x, a, \mu)$ *is bounded by a constant* $C_r > 0$ *and Lipschitz w.r.t.* μ *with constant* L_r.

(b) *The transition probability* $p(x'|x, a, \mu)$ *is* L_p*-Lipschitz continuous in* μ

(c) *The policies* $\pi(a|x, \mu)$ *satisfy the following Lipschitz bound:* $\|\pi(\cdot|x, \mu) - \pi(\cdot|x, \tilde{\mu})\|_1 \leq L_\pi d(\mu, \tilde{\mu})$ *for every* $x \in S^i$, *and* μ, $\tilde{\mu} \in \Delta(S^i)$.

Theorem (Approximate Nash equilibrium)

Let $(\pi_*^1, \ldots, \pi_*^m) \in \Pi^1 \times \cdots \times \Pi^m$ *be a Nash equilibrium for the MFTG. When the discount factor* γ *satisfies* $\gamma(1 + L_\pi + L_p) < 1$, *then*

$$\max_{\tilde{\pi}^i} J^{i,\bar{N}}(\tilde{\pi}^i; \pi_*^{-i}) \leq J^{i,\bar{N}}(\pi_*^i; \pi_*^{-i}) + \varepsilon(N),$$

for all $i \in [m]$, *with* $\varepsilon(N) = C \max_{i \in [m]} \left\{ |S^i| \sqrt{|A^i|} / \sqrt{N_i} \right\}$, *where* C *is a constant.*

- It justifies solving MFTGs because they provide an approximate solution for finite-agent games.
- Provides a rate of convergence, not just asymptotic convergence

Reformulation with MFMDPs: Notations

- The expected one-step reward can be expressed as

$$\bar{r}^i(\mu_t, \bar{\pi}_t^i) = \sum_{x \in S^i} \mu_t^i(x) \sum_{a \in A^i} \bar{\pi}_t^i(a|x) r^i(x, a, \mu_t), \qquad \bar{\pi}_t^i = \pi_t^i(\cdot|\cdot, \mu_t)$$

- $\bar{S} = \bigtimes_{i=1}^m \bar{S}^i$ is the (mean-field) state space, where $\bar{S}^i = \Delta(S^i)$ is the (mean-field) state space of population i. The (mean-field) state is $\bar{s}_t = \mu_t \in \bar{S}$
- $\bar{A}^i = \Delta(A^i)^{|S^i|}$ is the (mean-field) action space
- $\bar{r}^i : \bar{S} \times \bar{A}^i \to \mathbb{R}$ is as defined above
- $\bar{F} = \bar{p} : \bar{S} \times \bar{A}^1 \times \cdots \times \bar{A}^m \to \bar{S}$ is defined such that: $\bar{p}(\bar{s}_t, \bar{a}_t^1, \ldots, \bar{a}_t^m) = \bar{s}_{t+1}$ where, if $\bar{s}_t = (\mu_t^1, \ldots, \mu_t^m)$ and $\bar{a}_t^i = \pi^i(\cdot|\cdot, \mu_t^i)$, then $\bar{s}_{t+1} = (\mu_{t+1}^1, \ldots, \mu_{t+1}^m)$
- The transitions of the mean-field state depends on all the central players' (mean-field) actions
- The i-th central player first chooses **(mean-field) policy** $\bar{\pi}^i : \bar{S} \to \bar{A}^i$
- When applied on μ_t, $\bar{\pi}^i(\mu_t)$ returns a policy for the individual agent, i.e., $\bar{\pi}^i(\mu_t) : S^i \ni x_t^i \mapsto \bar{\pi}^i(\mu_t, x_t^i) = \pi^i(\cdot|x_t^i, \mu_t) \in \Delta(A^i)$.
- This approach allows us to view the problem posed to the i-th central player as a **Mean Field MDP (MFMDP)**

Reformulation with MFMDPs: Notations

- The expected one-step reward can be expressed as

$$\bar{r}^i(\mu_t, \bar{\pi}_t^i) = \sum_{x \in S^i} \mu_t^i(x) \sum_{a \in A^i} \bar{\pi}_t^i(a|x) r^i(x, a, \mu_t), \qquad \bar{\pi}_t^i = \pi_t^i(\cdot|\cdot, \mu_t)$$

- $\bar{S} = \bigtimes_{i=1}^m \bar{S}^i$ is the (mean-field) state space, where $\bar{S}^i = \Delta(S^i)$ is the (mean-field) state space of population i. The (mean-field) state is $\bar{s}_t = \mu_t \in \bar{S}$
- $\bar{A}^i = \Delta(A^i)^{|S^i|}$ is the (mean-field) action space
- $\bar{r}^i : \bar{S} \times \bar{A}^i \to \mathbb{R}$ is as defined above
- $\bar{F} = \bar{p} : \bar{S} \times \bar{A}^1 \times \cdots \times \bar{A}^m \to \bar{S}$ is defined such that: $\bar{p}(\bar{s}_t, \bar{a}_t^1, \ldots, \bar{a}_t^m) = \bar{s}_{t+1}$ where, if $\bar{s}_t = (\mu_t^1, \ldots, \mu_t^m)$ and $\bar{a}_t^i = \pi^i(\cdot|\cdot, \mu_t^i)$, then $\bar{s}_{t+1} = (\mu_{t+1}^1, \ldots, \mu_{t+1}^m)$
- The transitions of the mean-field state depends on all the central players' (mean-field) actions
- The i-th central player first chooses **(mean-field) policy** $\bar{\pi}^i : \bar{S} \to \bar{A}^i$
- When applied on μ_t, $\bar{\pi}^i(\mu_t)$ returns a policy for the individual agent, i.e., $\bar{\pi}^i(\mu_t) : S^i \ni x_t^i \mapsto \bar{\pi}^i(\mu_t, x_t^i) = \pi^i(\cdot|x_t^i, \mu_t) \in \Delta(A^i)$.
- This approach allows us to view the problem posed to the i-th central player as a **Mean Field MDP (MFMDP)**

Reformulation with MFMDPs: Notations

- The expected one-step reward can be expressed as

$$\bar{r}^i(\mu_t, \bar{\pi}_t^i) = \sum_{x \in S^i} \mu_t^i(x) \sum_{a \in A^i} \bar{\pi}_t^i(a|x) r^i(x, a, \mu_t), \qquad \bar{\pi}_t^i = \pi_t^i(\cdot|\cdot, \mu_t)$$

- $\bar{S} = \bigtimes_{i=1}^m \bar{S}^i$ is the (mean-field) state space, where $\bar{S}^i = \Delta(S^i)$ is the (mean-field) state space of population i. The (mean-field) state is $\bar{s}_t = \mu_t \in \bar{S}$
- $\bar{A}^i = \Delta(A^i)^{|S^i|}$ is the (mean-field) action space
- $\bar{r}^i : \bar{S} \times \bar{A}^i \to \mathbb{R}$ is as defined above
- $\bar{F} = \bar{p} : \bar{S} \times \bar{A}^1 \times \cdots \times \bar{A}^m \to \bar{S}$ is defined such that: $\bar{p}(\bar{s}_t, \bar{a}_t^1, \ldots, \bar{a}_t^m) = \bar{s}_{t+1}$ where, if $\bar{s}_t = (\mu_t^1, \ldots, \mu_t^m)$ and $\bar{a}_t^i = \pi^i(\cdot|\cdot, \mu_t^i)$, then $\bar{s}_{t+1} = (\mu_{t+1}^1, \ldots, \mu_{t+1}^m)$
- The transitions of the mean-field state depends on all the central players' (mean-field) actions
- The i-th central player first chooses **(mean-field) policy** $\bar{\pi}^i : \bar{S} \to \bar{A}^i$
- When applied on μ_t, $\bar{\pi}^i(\mu_t)$ returns a policy for the individual agent, i.e., $\bar{\pi}^i(\mu_t) : S^i \ni x_t^i \mapsto \bar{\pi}^i(\mu_t, x_t^i) = \pi^i(\cdot|x_t^i, \mu_t) \in \Delta(A^i)$.
- This approach allows us to view the problem posed to the i-th central player as a **Mean Field MDP (MFMDP)**

Reformulation with MFMDPs: Notations

- The expected one-step reward can be expressed as

$$\bar{r}^i(\mu_t, \bar{\pi}_t^i) = \sum_{x \in S^i} \mu_t^i(x) \sum_{a \in A^i} \bar{\pi}_t^i(a|x) r^i(x, a, \mu_t), \qquad \bar{\pi}_t^i = \pi_t^i(\cdot|\cdot, \mu_t)$$

- $\bar{S} = \bigtimes_{i=1}^m \bar{S}^i$ is the (mean-field) state space, where $\bar{S}^i = \Delta(S^i)$ is the (mean-field) state space of population i. The (mean-field) state is $\bar{s}_t = \mu_t \in \bar{S}$
- $\bar{A}^i = \Delta(A^i)^{|S^i|}$ is the (mean-field) action space
- $\bar{r}^i : \bar{S} \times \bar{A}^i \to \mathbb{R}$ is as defined above
- $\bar{F} = \bar{p} : \bar{S} \times \bar{A}^1 \times \cdots \times \bar{A}^m \to \bar{S}$ is defined such that: $\bar{p}(\bar{s}_t, \bar{a}_t^1, \ldots, \bar{a}_t^m) = \bar{s}_{t+1}$ where, if $\bar{s}_t = (\mu_t^1, \ldots, \mu_t^m)$ and $\bar{a}_t^i = \pi^i(\cdot|\cdot, \mu_t^i)$, then $\bar{s}_{t+1} = (\mu_{t+1}^1, \ldots, \mu_{t+1}^m)$
- The transitions of the mean-field state depends on all the central players' (mean-field) actions
- The i-th central player first chooses **(mean-field) policy** $\bar{\pi}^i : \bar{S} \to \bar{A}^i$
- When applied on μ_t, $\bar{\pi}^i(\mu_t)$ returns a policy for the individual agent, i.e., $\bar{\pi}^i(\mu_t) : S^i \ni x_t^i \mapsto \bar{\pi}^i(\mu_t, x_t^i) = \pi^i(\cdot|x_t^i, \mu_t) \in \Delta(A^i)$.
- This approach allows us to view the problem posed to the i-th central player as a **Mean Field MDP (MFMDP)**

Reformulation with MFMDPs

Given policy profile $\bar{\boldsymbol{\pi}} = (\bar{\pi}^1, \ldots, \bar{\pi}^m)$, total reward of central player of coalition i:

$$\bar{v}^i_{\bar{\boldsymbol{\pi}}}(\bar{s}) = \bar{v}^i(\bar{s}, \bar{\boldsymbol{\pi}}) := \mathbb{E}_{\bar{\boldsymbol{\pi}}}\left[\sum_{t=0}^{\infty} \gamma^t \bar{r}^i(\bar{s}_t, \bar{a}^i_t) | \bar{s}_0 = \bar{s}\right]$$

where $\bar{s}_{t+1} \sim \bar{p}(\cdot|\bar{s}_t, \bar{a}^1_t, \ldots, \bar{a}^m_t)$, $\bar{a}^j_t \sim \bar{\pi}^i(\cdot|\bar{s}_t)$, $j = 1, \ldots, m,\ t \geq 0$.

Definition (Nash equilibrium for MFTG rephrased)

An MFTG **Nash equilibrium** $\bar{\boldsymbol{\pi}}_* = (\bar{\pi}^1_*, \ldots, \bar{\pi}^m_*)$ is such that for all $i = 1, \ldots, m$:
$\bar{v}^i(\bar{s}, \bar{\boldsymbol{\pi}}_*) \geq \bar{v}^i(\bar{s}, (\bar{\pi}^i, \bar{\boldsymbol{\pi}}^{-i}_*)), \forall \bar{s} \in \bar{S}, \forall \bar{\pi}^i \in \bar{\Pi}^i$.

Let $\bar{\boldsymbol{a}} = (\bar{a}^1, \ldots, \bar{a}^m)$, $\bar{\boldsymbol{\pi}}^{-i}(\mathrm{d}\bar{\boldsymbol{a}}^{-i}|\bar{s}) = \prod_{j\neq i} \bar{\pi}^j(\mathrm{d}\bar{a}^j|\bar{s})$, $\bar{\boldsymbol{a}}^{-i} \in \bar{A}^{-i} = \prod_{j\neq i} \bar{A}^j$.

Q-function for central player i: $\bar{Q}^i_{\bar{\boldsymbol{\pi}}}(\bar{s}, \bar{\boldsymbol{a}}) = \mathbb{E}_{\bar{\boldsymbol{\pi}}}\left[\sum_{t=0}^{\infty} \gamma^t \bar{r}^i(\bar{s}_t, \bar{a}^i) | \bar{s}_0 = \bar{s}, \bar{\boldsymbol{a}}_0 = \bar{\boldsymbol{a}}\right]$.

Definition

An **MDP for a central player** i against fixed policies $\bar{\pi}^{-i}$ of other players is a tuple $(\bar{S}, \bar{A}^i, \bar{p}_{\bar{\boldsymbol{\pi}}^{-i}}, \bar{r}_{\bar{\boldsymbol{\pi}}^{-i}}, \gamma)$ where

$$\bar{p}_{\bar{\boldsymbol{\pi}}^{-i}}(\bar{s}'|\bar{s}, \bar{a}^i) = \int_{\bar{A}^{-i}} \bar{p}(\bar{s}'|\bar{s}, \bar{\boldsymbol{a}}) \bar{\boldsymbol{\pi}}^{-i}(\mathrm{d}\bar{\boldsymbol{a}}^{-i}|\bar{s}), \quad \bar{r}_{\bar{\boldsymbol{\pi}}^{-i}}(\bar{s}, \bar{a}^i) = \bar{r}^i(\bar{s}, \bar{a}^i).$$

Reformulation with MFMDPs

Given policy profile $\bar{\boldsymbol{\pi}} = (\bar{\pi}^1, \dots, \bar{\pi}^m)$, total reward of central player of coalition i:

$$\bar{v}^i_{\bar{\boldsymbol{\pi}}}(\bar{s}) = \bar{v}^i(\bar{s}, \bar{\boldsymbol{\pi}}) := \mathbb{E}_{\bar{\boldsymbol{\pi}}}\left[\sum_{t=0}^{\infty} \gamma^t \bar{r}^i(\bar{s}_t, \bar{a}^i_t) | \bar{s}_0 = \bar{s}\right]$$

where $\bar{s}_{t+1} \sim \bar{p}(\cdot|\bar{s}_t, \bar{a}^1_t, \dots, \bar{a}^m_t)$, $\bar{a}^j_t \sim \bar{\pi}^i(\cdot|\bar{s}_t)$, $j = 1, \dots, m,\ t \geq 0$.

Definition (Nash equilibrium for MFTG rephrased)

An MFTG **Nash equilibrium** $\bar{\boldsymbol{\pi}}_* = (\bar{\pi}^1_*, \dots, \bar{\pi}^m_*)$ is such that for all $i = 1, \dots, m$:
$\bar{v}^i(\bar{s}, \bar{\boldsymbol{\pi}}_*) \geq \bar{v}^i(\bar{s}, (\bar{\pi}^i, \bar{\boldsymbol{\pi}}^{-i}_*)), \forall \bar{s} \in \bar{S}, \forall \bar{\pi}^i \in \bar{\Pi}^i$.

Let $\bar{\boldsymbol{a}} = (\bar{a}^1, \dots, \bar{a}^m)$, $\bar{\boldsymbol{\pi}}^{-i}(\mathrm{d}\bar{\boldsymbol{a}}^{-i}|\bar{s}) = \prod_{j\neq i} \bar{\pi}^j(\mathrm{d}\bar{a}^j|\bar{s})$, $\bar{\boldsymbol{a}}^{-i} \in \bar{A}^{-i} = \prod_{j\neq i} \bar{A}^j$.

Q-function for central player i: $\bar{Q}^i_{\bar{\boldsymbol{\pi}}}(\bar{s}, \bar{\boldsymbol{a}}) = \mathbb{E}_{\bar{\boldsymbol{\pi}}}\left[\sum_{t=0}^{\infty} \gamma^t \bar{r}^i(\bar{s}_t, \bar{a}^i) | \bar{s}_0 = \bar{s}, \bar{\boldsymbol{a}}_0 = \bar{\boldsymbol{a}}\right]$.

Definition

An **MDP for a central player** i against fixed policies $\bar{\pi}^{-i}$ of other players is a tuple $(\bar{S}, \bar{A}^i, \bar{p}_{\bar{\boldsymbol{\pi}}^{-i}}, \bar{r}_{\bar{\boldsymbol{\pi}}^{-i}}, \gamma)$ where

$$\bar{p}_{\bar{\boldsymbol{\pi}}^{-i}}(\bar{s}'|\bar{s}, \bar{a}^i) = \int_{\bar{A}^{-i}} \bar{p}(\bar{s}'|\bar{s}, \bar{\boldsymbol{a}}) \bar{\boldsymbol{\pi}}^{-i}(\mathrm{d}\bar{\boldsymbol{a}}^{-i}|\bar{s}), \quad \bar{r}_{\bar{\boldsymbol{\pi}}^{-i}}(\bar{s}, \bar{a}^i) = \bar{r}^i(\bar{s}, \bar{a}^i).$$

Reformulation with MFMDPs

Given policy profile $\bar{\boldsymbol{\pi}} = (\bar{\pi}^1, \ldots, \bar{\pi}^m)$, total reward of central player of coalition i:

$$\bar{v}^i_{\bar{\boldsymbol{\pi}}}(\bar{s}) = \bar{v}^i(\bar{s}, \bar{\boldsymbol{\pi}}) := \mathbb{E}_{\bar{\boldsymbol{\pi}}}\left[\sum_{t=0}^{\infty} \gamma^t \bar{r}^i(\bar{s}_t, \bar{a}^i_t) | \bar{s}_0 = \bar{s}\right]$$

where $\bar{s}_{t+1} \sim \bar{p}(\cdot|\bar{s}_t, \bar{a}^1_t, \ldots, \bar{a}^m_t)$, $\bar{a}^j_t \sim \bar{\pi}^i(\cdot|\bar{s}_t)$, $j = 1, \ldots, m,\ t \geq 0$.

Definition (Nash equilibrium for MFTG rephrased)

An MFTG **Nash equilibrium** $\bar{\boldsymbol{\pi}}_* = (\bar{\pi}^1_*, \ldots, \bar{\pi}^m_*)$ is such that for all $i = 1, \ldots, m$:
$\bar{v}^i(\bar{s}, \bar{\boldsymbol{\pi}}_*) \geq \bar{v}^i(\bar{s}, (\bar{\pi}^i, \bar{\boldsymbol{\pi}}^{-i}_*)), \forall \bar{s} \in \bar{S}, \forall \bar{\pi}^i \in \bar{\Pi}^i$.

Let $\bar{\boldsymbol{a}} = (\bar{a}^1, \ldots, \bar{a}^m)$, $\bar{\boldsymbol{\pi}}^{-i}(\mathrm{d}\bar{\boldsymbol{a}}^{-i}|\bar{s}) = \prod_{j \neq i} \bar{\pi}^j(\mathrm{d}\bar{a}^j|\bar{s})$, $\bar{\boldsymbol{a}}^{-i} \in \bar{A}^{-i} = \prod_{j \neq i} \bar{A}^j$.

Q-function for central player i: $\bar{Q}^i_{\bar{\boldsymbol{\pi}}}(\bar{s}, \bar{\boldsymbol{a}}) = \mathbb{E}_{\bar{\boldsymbol{\pi}}}\left[\sum_{t=0}^{\infty} \gamma^t \bar{r}^i(\bar{s}_t, \bar{a}^i) | \bar{s}_0 = \bar{s}, \bar{\boldsymbol{a}}_0 = \bar{\boldsymbol{a}}\right]$.

Definition

An **MDP for a central player** i against fixed policies $\bar{\boldsymbol{\pi}}^{-i}$ of other players is a tuple $(\bar{S}, \bar{A}^i, \bar{p}_{\bar{\boldsymbol{\pi}}^{-i}}, \bar{r}_{\bar{\boldsymbol{\pi}}^{-i}}, \gamma)$ where

$$\bar{p}_{\bar{\boldsymbol{\pi}}^{-i}}(\bar{s}'|\bar{s}, \bar{a}^i) = \int_{\bar{A}^{-i}} \bar{p}(\bar{s}'|\bar{s}, \bar{\boldsymbol{a}}) \bar{\boldsymbol{\pi}}^{-i}(\mathrm{d}\bar{\boldsymbol{a}}^{-i}|\bar{s}), \quad \bar{r}_{\bar{\boldsymbol{\pi}}^{-i}}(\bar{s}, \bar{a}^i) = \bar{r}^i(\bar{s}, \bar{a}^i).$$

Outline

Intuition

- Nash Q-learning for finite-player, finite-space games [Hu and Wellman, 2003]
- Depends on the policies of all the players
- Infinite horizon $\Rightarrow$ the equilibrium depend on infinite trajectories
- Recall: $\bar{Q}^i_{\bar{\pi}}(\bar{s}, \bar{a})$
- If we had DPP for $\bar{Q}^i$, we could use the MF Q-learning of [Carmona et al., 2023]
- But no DPP for $\bar{Q}^i$, unless we have the equilibrium policies of others
- Introduce an auxiliary Q-function (NashQ function) based on equilibrium policies
- Reduce the infinite-horizon game to a sequence on one-stage games

Intuition

- Nash Q-learning for finite-player, finite-space games [Hu and Wellman, 2003]
- Depends on the policies of all the players
- Infinite horizon $\Rightarrow$ the equilibrium depend on infinite trajectories
- Recall: $\bar{Q}^{i}_{\boldsymbol{\pi}}(\bar{s}, \bar{\boldsymbol{a}})$
- If we had DPP for $\bar{Q}^{i}$, we could use the MF Q-learning of [Carmona et al., 2023]
- But no DPP for $\bar{Q}^{i}$, unless we have the equilibrium policies of others
- Introduce an auxiliary Q-function (NashQ function) based on equilibrium policies
- Reduce the infinite-horizon game to a sequence on one-stage games

Intuition

- Nash Q-learning for finite-player, finite-space games [Hu and Wellman, 2003]
- Depends on the policies of all the players
- Infinite horizon $\Rightarrow$ the equilibrium depend on infinite trajectories
- Recall: $\bar{Q}^i_{\bar{\boldsymbol{\pi}}}(\bar{s}, \bar{\boldsymbol{a}})$
- If we had DPP for $\bar{Q}^i$, we could use the MF Q-learning of [Carmona et al., 2023]
- But no DPP for $\bar{Q}^i$, unless we have the equilibrium policies of others
- Introduce an auxiliary Q-function (NashQ function) based on equilibrium policies
- Reduce the infinite-horizon game to a sequence on one-stage games

Stage Game

Definition (Stage game and stage Nash equilibrium)

Consider as given a (mean-field) state $\bar{s} \in \bar{S}$ and a policy profile $\bar{\boldsymbol{\pi}} = (\bar{\pi}^1, \ldots, \bar{\pi}^m)$. The (mean-field) **stage game induced by $\bar{s}$ and $\bar{\boldsymbol{\pi}}$** is a static game in which player i takes an action $\bar{a}^i \in \bar{A}^i$, $i = 1, \ldots, m$ and gets the reward

$$\bar{Q}^i_{\bar{\boldsymbol{\pi}}}(\bar{s}, \bar{a}^1, \ldots, \bar{a}^m).$$

Player i is allowed to use a mixed strategy $\sigma^i \in \Delta(\bar{A}^i)$.

A **Nash equilibrium** for this stage game is a strategy profile $\boldsymbol{\sigma}_* = (\sigma^1_*, \ldots, \sigma^m_*)$ such that, for all $\sigma^i \in \Delta(\bar{A}^i)$,

$$\sigma^1_* \cdots \sigma^m_* \bar{Q}^i_{\bar{\boldsymbol{\pi}}}(\bar{s}) \geq \sigma^1_* \cdots \sigma^{i-1}_* \sigma^i \sigma^{i+1}_* \cdots \sigma^m_* \bar{Q}^i_{\bar{\boldsymbol{\pi}}}(\bar{s})$$

where we define

$$\sigma^1 \cdots \sigma^m \bar{Q}^i_{\bar{\boldsymbol{\pi}}}(\bar{s}) := \bar{r}^i(\bar{s}, \sigma^i) + \gamma \int_{\bar{S}} \int_{\bar{\boldsymbol{A}}} \bar{v}^i(\bar{s}', \bar{\boldsymbol{\pi}}) \bar{p}(\mathrm{d}\bar{s}' | \bar{s}, \bar{\boldsymbol{a}}) \boldsymbol{\sigma}(\mathrm{d}\bar{\boldsymbol{a}} | \bar{s}),$$

with $\bar{\boldsymbol{A}} = \bar{A}^1 \times \cdots \times \bar{A}^m$, $\boldsymbol{\sigma}(\mathrm{d}\bar{\boldsymbol{a}}|\bar{s}) = \prod_{i=1}^m \sigma^i(\mathrm{d}\bar{a}^i|\bar{s})$, and $\bar{r}^i(\bar{s}, \sigma^i) := \mathbb{E}_{\bar{a}^i \sim \sigma^i} \bar{r}^i(\bar{s}, \bar{a}^i)$.

Definition (Stage game and stage Nash equilibrium)

Consider as given a (mean-field) state $\bar{s} \in \bar{S}$ and a policy profile $\bar{\boldsymbol{\pi}} = (\bar{\pi}^1, \ldots, \bar{\pi}^m)$. The (mean-field) **stage game induced by $\bar{s}$ and $\bar{\boldsymbol{\pi}}$** is a static game in which player i takes an action $\bar{a}^i \in \bar{A}^i$, $i = 1, \ldots, m$ and gets the reward

$$\bar{Q}^i_{\bar{\boldsymbol{\pi}}}(\bar{s}, \bar{a}^1, \ldots, \bar{a}^m).$$

Player i is allowed to use a mixed strategy $\sigma^i \in \Delta(\bar{A}^i)$.

A **Nash equilibrium** for this stage game is a strategy profile $\boldsymbol{\sigma}_* = (\sigma^1_*, \ldots, \sigma^m_*)$ such that, for all $\sigma^i \in \Delta(\bar{A}^i)$,

$$\sigma^1_* \cdots \sigma^m_* \bar{Q}^i_{\bar{\boldsymbol{\pi}}}(\bar{s}) \geq \sigma^1_* \cdots \sigma^{i-1}_* \sigma^i \sigma^{i+1}_* \cdots \sigma^m_* \bar{Q}^i_{\bar{\boldsymbol{\pi}}}(\bar{s})$$

where we define

$$\sigma^1 \cdots \sigma^m \bar{Q}^i_{\bar{\boldsymbol{\pi}}}(\bar{s}) := \bar{r}^i(\bar{s}, \sigma^i) + \gamma \int_{\bar{S}} \int_{\bar{\boldsymbol{A}}} \bar{v}^i(\bar{s}', \bar{\boldsymbol{\pi}}) \bar{p}(\mathrm{d}\bar{s}' | \bar{s}, \bar{\boldsymbol{a}}) \boldsymbol{\sigma}(\mathrm{d}\bar{\boldsymbol{a}} | \bar{s}),$$

with $\bar{\boldsymbol{A}} = \bar{A}^1 \times \cdots \times \bar{A}^m$, $\boldsymbol{\sigma}(\mathrm{d}\bar{\boldsymbol{a}}|\bar{s}) = \prod_{i=1}^m \sigma^i(\mathrm{d}\bar{a}^i|\bar{s})$, and $\bar{r}^i(\bar{s}, \sigma^i) := \mathbb{E}_{\bar{a}^i \sim \sigma^i} \bar{r}^i(\bar{s}, \bar{a}^i)$.

Stage Game

Definition (NashQ function)

Given a Nash equilibrium $(\sigma_*^1, \ldots, \sigma_*^m)$, the **NashQ function** of player i is defined as:

$$\mathrm{Nash}\bar{Q}^i_{\bar{\boldsymbol{\pi}}}(\bar{s}) := \sigma_*^1 \cdots \sigma_*^m \bar{Q}^i_{\bar{\boldsymbol{\pi}}}(\bar{s}).$$

Proposition (Link between MFTG equilibrium and stage-game equilibrium)

The following statements are equivalent:

(i) $\bar{\boldsymbol{\pi}}_* = (\bar{\pi}_*^1, \ldots, \bar{\pi}_*^m)$ *is a Nash equilibrium for the MFTG with equilibrium payoff* $(\bar{v}^1_{\bar{\boldsymbol{\pi}}_*}, \ldots, \bar{v}^m_{\bar{\boldsymbol{\pi}}_*})$*;*

(ii) *For every* $\bar{s} \in \bar{S}$*,* $(\bar{\pi}_*^1(\bar{s}), \ldots, \bar{\pi}_*^m(\bar{s}))$ *is a Nash equilibrium in the stage game induced by state* $\bar{s}$ *and policy profile* $\bar{\boldsymbol{\pi}}_*$*.*

- Basis for RL algorithm: Nash Q-learning
- Requires solving stage-game
- We are going to discretize the simplexes

Stage Game

Definition (NashQ function)

Given a Nash equilibrium $(\sigma_*^1, \ldots, \sigma_*^m)$, the **NashQ function** of player i is defined as:

$$\mathrm{Nash}\bar{Q}^i_{\bar{\boldsymbol{\pi}}}(\bar{s}) := \sigma_*^1 \cdots \sigma_*^m \bar{Q}^i_{\bar{\boldsymbol{\pi}}}(\bar{s}).$$

Proposition (Link between MFTG equilibrium and stage-game equilibrium)

The following statements are equivalent:

(i) $\bar{\boldsymbol{\pi}}_* = (\bar{\pi}_*^1, \ldots, \bar{\pi}_*^m)$ *is a Nash equilibrium for the MFTG with equilibrium payoff* $(\bar{v}^1_{\bar{\boldsymbol{\pi}}_*}, \ldots, \bar{v}^m_{\bar{\boldsymbol{\pi}}_*})$*;*

(ii) *For every* $\bar{s} \in \bar{S}$, $(\bar{\pi}_*^1(\bar{s}), \ldots, \bar{\pi}_*^m(\bar{s}))$ *is a Nash equilibrium in the stage game induced by state* $\bar{s}$ *and policy profile* $\bar{\boldsymbol{\pi}}_*$*.*

- Basis for RL algorithm: Nash Q-learning
- Requires solving stage-game
- We are going to discretize the simplexes

Simplex Discretization

- $\bar{S}^i = \Delta(S^i)$ and $\Delta(A^i)$ are (finite-dimensional) simplexes; we endow them with the distances $d_{\bar{S}}(\bar{s}, \bar{s}') = \sum_{i\in[m]} d(\bar{s}^i, \bar{s'}^i) = \sum_{i\in[m]} \sum_{x\in S^i} |\mu^i(x) - \mu'^i(x)|$, and $d_{A^i}(\bar{a}^i(\bar{s}), \bar{a'}^i(\bar{s})) = \sum_{x,a} |\pi^i(a|x,\bar{s}) - \pi'^i(a|x,\bar{s})|$, where $\bar{s}^i = \mu^i$, $\bar{a}^i(\bar{s}) = \pi^i(\cdot|\cdot,\bar{s})$.
- In $\bar{A}^i = \{\bar{S} \to \Delta(A^i)\}$, we take $d_{\bar{A}^i}(\bar{a}^i, \bar{a'}^i) = \sup_{\bar{s}\in\bar{S}} d_{A^i}(\bar{a}^i(\bar{s}), \bar{a'}^i(\bar{s}))$.
- Quantization:
  - For $i = 1, \ldots, m$, let $\check{S}^i \subset \bar{S}^i$ and $\check{\Delta}(A^i) \subset \Delta(A^i)$ be finite approximations of $\bar{S}^i$ and $\Delta(A^i)$.
  - Mean-field finite spaces $\check{S} = \Pi_{i=1}^m \check{S}^i \subset \bar{S}$ and $\check{A}^i = \{\check{a}^i : \check{S} \to \check{\Delta}(A^i)\}$.
  - Let $\epsilon_S = \max_{\bar{s}\in\bar{S}} \min_{\check{s}\in\check{S}} d_{\bar{S}}(\bar{s}, \check{s})$ and $\epsilon_A = \max_i \max_{\bar{a}^i\in\bar{A}^i} \min_{\check{a}^i\in\check{A}^i} d_{\bar{A}^i}(\bar{a}^i, \check{a}^i)$, which characterize the fineness of the discretization.
  - The policy space of each player i is $\check{\Pi}^i = \{\check{\pi}^i : \check{S} \to \Delta(\check{A}^i)\}$.
- Projection operator $\mathrm{Proj}_{\check{S}} : \bar{S} \to \check{S}$, which maps $\bar{s}$ to the closest point in $\check{S}$.
- Transitions: Given a state $\check{s}_t$ and a joint action $(\check{a}_t^1, \ldots, \check{a}_t^m)$, we generate $\bar{s}_{t+1} = \bar{F}(\check{s}_t, \check{a}_t^1, \ldots \check{a}_t^m)$. Then, we project $\bar{s}_{t+1}$ back to $\check{S}$ and denote the projected state by $\check{s}_{t+1} = \mathrm{Proj}_{\check{S}}(\bar{s}_{t+1})$.

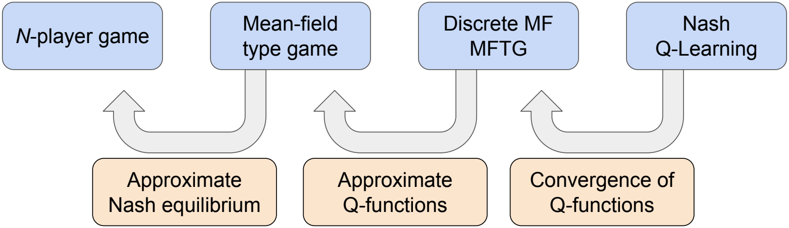
Simplex Discretization

- $\bar{S}^i = \Delta(S^i)$ and $\Delta(A^i)$ are (finite-dimensional) simplexes; we endow them with the distances $d_{\bar{S}}(\bar{s}, \bar{s}') = \sum_{i\in[m]} d(\bar{s}^i, \bar{s'}^i) = \sum_{i\in[m]} \sum_{x\in S^i} |\mu^i(x) - {\mu'}^i(x)|$, and $d_{A^i}(\bar{a}^i(\bar{s}), \bar{a'}^i(\bar{s})) = \sum_{x,a} |\pi^i(a|x, \bar{s}) - {\pi'}^i(a|x, \bar{s})|$, where $\bar{s}^i = \mu^i$, $\bar{a}^i(\bar{s}) = \pi^i(\cdot|\cdot, \bar{s})$.
- In $\bar{A}^i = \{\bar{S} \to \Delta(A^i)\}$, we take $d_{\bar{A}^i}(\bar{a}^i, \bar{a'}^i) = \sup_{\bar{s}\in\bar{S}} d_{A^i}(\bar{a}^i(\bar{s}), \bar{a'}^i(\bar{s}))$.
- Quantization:
  - For $i = 1, \ldots, m$, let $\check{S}^i \subset \bar{S}^i$ and $\check{\Delta}(A^i) \subset \Delta(A^i)$ be finite approximations of $\bar{S}^i$ and $\Delta(A^i)$.
  - Mean-field finite spaces $\check{S} = \Pi_{i=1}^m \check{S}^i \subset \bar{S}$ and $\check{A}^i = \{\check{a}^i : \check{S} \to \check{\Delta}(A^i)\}$.
  - Let $\epsilon_S = \max_{\bar{s}\in\bar{S}} \min_{\check{s}\in\check{S}} d_{\bar{S}}(\bar{s}, \check{s})$ and $\epsilon_A = \max_i \max_{\bar{a}^i\in\bar{A}^i} \min_{\check{a}^i\in\check{A}^i} d_{\bar{A}^i}(\bar{a}^i, \check{a}^i)$, which characterize the fineness of the discretization.
  - The policy space of each player i is $\check{\Pi}^i = \{\check{\pi}^i : \check{S} \to \Delta(\check{A}^i)\}$.
- Projection operator $\mathrm{Proj}_{\check{S}} : \bar{S} \to \check{S}$, which maps $\bar{s}$ to the closest point in $\check{S}$.
- Transitions: Given a state $\check{s}_t$ and a joint action $(\check{a}_t^1, \ldots, \check{a}_t^m)$, we generate $\bar{s}_{t+1} = \bar{F}(\check{s}_t, \check{a}_t^1, \ldots \check{a}_t^m)$. Then, we project $\bar{s}_{t+1}$ back to $\check{S}$ and denote the projected state by $\check{s}_{t+1} = \mathrm{Proj}_{\check{S}}(\bar{s}_{t+1})$.

Simplex Discretization

- $\bar{S}^i = \Delta(S^i)$ and $\Delta(A^i)$ are (finite-dimensional) simplexes; we endow them with the distances $d_{\bar{S}}(\bar{s}, \bar{s}') = \sum_{i \in [m]} d(\bar{s}^i, \bar{s'}^i) = \sum_{i \in [m]} \sum_{x \in S^i} |\mu^i(x) - \mu'^i(x)|$, and $d_{A^i}(\bar{a}^i(\bar{s}), \bar{a'}^i(\bar{s})) = \sum_{x,a} |\pi^i(a|x, \bar{s}) - \pi'^i(a|x, \bar{s})|$, where $\bar{s}^i = \mu^i$, $\bar{a}^i(\bar{s}) = \pi^i(\cdot|\cdot, \bar{s})$.
- In $\bar{A}^i = \{\bar{S} \to \Delta(A^i)\}$, we take $d_{\bar{A}^i}(\bar{a}^i, \bar{a'}^i) = \sup_{\bar{s} \in \bar{S}} d_{A^i}(\bar{a}^i(\bar{s}), \bar{a'}^i(\bar{s}))$.
- Quantization:
  - For $i = 1, \ldots, m$, let $\check{S}^i \subset \bar{S}^i$ and $\check{\Delta}(A^i) \subset \Delta(A^i)$ be finite approximations of $\bar{S}^i$ and $\Delta(A^i)$.
  - Mean-field finite spaces $\check{S} = \Pi_{i=1}^m \check{S}^i \subset \bar{S}$ and $\check{A}^i = \{\check{a}^i : \check{S} \to \check{\Delta}(A^i)\}$.
  - Let $\epsilon_S = \max_{\bar{s} \in \bar{S}} \min_{\check{s} \in \check{S}} d_{\bar{S}}(\bar{s}, \check{s})$ and $\epsilon_A = \max_i \max_{\bar{a}^i \in \bar{A}^i} \min_{\check{a}^i \in \check{A}^i} d_{\bar{A}^i}(\bar{a}^i, \check{a}^i)$, which characterize the fineness of the discretization.
  - The policy space of each player i is $\check{\Pi}^i = \{\check{\pi}^i : \check{S} \to \Delta(\check{A}^i)\}$.
- Projection operator $\text{Proj}_{\check{S}} : \bar{S} \to \check{S}$, which maps $\bar{s}$ to the closest point in $\check{S}$.
- Transitions: Given a state $\check{s}_t$ and a joint action $(\check{a}_t^1, \ldots, \check{a}_t^m)$, we generate $\bar{s}_{t+1} = \bar{F}(\check{s}_t, \check{a}_t^1, \ldots \check{a}_t^m)$. Then, we project $\bar{s}_{t+1}$ back to $\check{S}$ and denote the projected state by $\check{s}_{t+1} = \text{Proj}_{\check{S}}(\bar{s}_{t+1})$.

Theoretical Results



Algorithm 1: Discretized Nash Q-learning for MFTG (**DNashQ-MFTG**)

Inputs: A series of learning rates $\alpha_t \in (0,1)$, $t \geq 0$, and exploration levels ϵ_t, $t \geq 0$
Outputs: Nash Q-functions $\check{Q}^i_N$ for $i = 1, \ldots, m$
Initialization: $\check{Q}^i_{0,0}(\check{s}, \check{a}^1, \ldots, \check{a}^m) = 0$ for all $\check{s} \in \check{S}$ and $\check{a}^i \in \check{A}^i$;
for $k = 0, 1, \ldots, N-1$ **do**
  Initialize state $\check{s}_0$
  for $t = 0, \ldots, T-1$ **do**
    Generate a random number $\zeta_t \sim \mathcal{U}[0,1]$
    if $\zeta_t \geq \epsilon_t$ **then**
      Solve the stage game $\check{Q}^i_{k,t}(\check{s}_t)$ and get strategy profile $(\check{\pi}^{i,1}_*, \ldots, \check{\pi}^{i,m}_*)$ for $i = 1, \ldots, m$
      Sample $\check{a}^i_t \sim \check{\pi}^{i,i}_*$ for $i = 1, \ldots, m$
    else
      Sample action $\check{a}^i_t$ uniformly from $\check{A}^i$ for $i = 1, \ldots, m$
    end if
    Observe $r^1_t, \ldots, r^m_t$, $\check{a}^1_t, \ldots, \check{a}^m_t$, and $\check{s}_{t+1} = \mathrm{Proj}_{\check{S}}(\bar{F}(\check{s}_t, \check{a}^1_t, \ldots, \check{a}^m_t))$
    Solve the stage game $\check{Q}^i_{k,t}(\check{s}_{t+1})$ and get strategy profile $(\check{\pi}'^{i,1}_*, \ldots, \check{\pi}'^{i,m}_*)$ for $i = 1, \ldots, m$
    Compute $\mathrm{Nash}\check{Q}^i_{k,t}(\check{s}_{t+1}) = \check{\pi}'^{i,1}_* \ldots \check{\pi}'^{i,m}_* \check{Q}^i_{k,t}(\check{s}_{t+1})$
    Copy $\check{Q}^i_{k,t+1} = \check{Q}^i_{k,t}$ for $i = 1, \ldots, m$ and update $\check{Q}^i_{k,t+1}$ by:
$$\check{Q}^i_{k,t+1}(\check{s}_t, \check{a}^1, \ldots, \check{a}^m) = (1-\alpha_t)\check{Q}^i_{k,t}(\check{s}_t, \check{a}^1, \ldots, \check{a}^m) + \alpha_t(r^i_t + \beta \mathrm{Nash}\check{Q}^i_{k,t}(\check{s}_{t+1}))$$
  end for
  Copy $\check{Q}^i_{k+1,0} = \check{Q}^i_{k,T-1}$ for $i = 1 \ldots . m$
end for

Convergence

Recall:

$$\check{Q}^i_{t+1}(\check{s}, \check{a}^1, \ldots, \check{a}^m) = (1-\alpha_t)\check{Q}^i_t(\check{s}, \check{a}^1, \ldots, \check{a}^m) + \alpha_t(\bar{r}^i_t + \beta \mathrm{Nash}\check{Q}^i_t(\check{s}')),$$

where

$$\mathrm{Nash}\check{Q}^i_t(\check{s}') = \check{\pi}^{i,1}_* \cdots \check{\pi}^{i,m}_* \check{Q}^i_t(\check{s}'),$$

with $\check{\pi}^{i,j}_*$ obtained by solving the one-stage game with rewards $(\check{Q}^1_t(\check{s}'), \ldots, \check{Q}^m_t(\check{s}'))$.

Theorem (NashQ-learning convergence)

Under suitable assumptions (see [Hu and Wellman, 2003, Yang et al., 2018]), $\check{Q}_t = (\check{Q}^1_t, \ldots, \check{Q}^m_t)$ converges to the Nash equilibrium Q-functions $\check{Q}_{\check{\boldsymbol{\pi}}_*} = (\check{Q}^1_{\check{\boldsymbol{\pi}}_*}, \ldots, \check{Q}^m_{\check{\boldsymbol{\pi}}_*})$.

Then: focus on the difference between the approximated Nash Q-function, $\check{Q}^i_t(\mathrm{Proj}_{\check{S}}(\bar{s}), \mathrm{Proj}_{\check{A}^1}(\bar{a}^1), \ldots, \mathrm{Proj}_{\check{A}^m}(\bar{a}^m))$ and the true Nash Q-function, $\bar{Q}^i_{\bar{\pi}_*}(\bar{s}, \bar{a}^1 \ldots \bar{a}^m)$, in the infinite space $\bar{S} \times \bar{A}^i \times \cdots \times \bar{A}^m$

Convergence

Recall:

$$\check{Q}^i_{t+1}(\check{s}, \check{a}^1, \ldots, \check{a}^m) = (1-\alpha_t)\check{Q}^i_t(\check{s}, \check{a}^1, \ldots, \check{a}^m) + \alpha_t(\bar{r}^i_t + \beta \mathrm{Nash}\check{Q}^i_t(\check{s}')),$$

where

$$\mathrm{Nash}\check{Q}^i_t(\check{s}') = \check{\pi}^{i,1}_* \cdots \check{\pi}^{i,m}_* \check{Q}^i_t(\check{s}'),$$

with $\check{\pi}^{i,j}_*$ obtained by solving the one-stage game with rewards $(\check{Q}^1_t(\check{s}'), \ldots, \check{Q}^m_t(\check{s}'))$.

Theorem (NashQ-learning convergence)

Under suitable assumptions (see [Hu and Wellman, 2003, Yang et al., 2018]), $\check{Q}_t = (\check{Q}^1_t, \ldots, \check{Q}^m_t)$ converges to the Nash equilibrium Q-functions $\check{Q}_{\check{\boldsymbol{\pi}}_*} = (\check{Q}^1_{\check{\boldsymbol{\pi}}_*}, \ldots, \check{Q}^m_{\check{\boldsymbol{\pi}}_*})$.

Then: focus on the difference between the approximated Nash Q-function, $\check{Q}^i_t(\mathrm{Proj}_{\check{S}}(\bar{s}), \mathrm{Proj}_{\check{A}^1}(\bar{a}^1), \ldots, \mathrm{Proj}_{\check{A}^m}(\bar{a}^m))$ and the true Nash Q-function, $\bar{Q}^i_{\bar{\boldsymbol{\pi}}_*}(\bar{s}, \bar{a}^1 \ldots \bar{a}^m)$, in the infinite space $\bar{S} \times \bar{A}^i \times \cdots \times \bar{A}^m$

Approximation Analysis

Assumption

(a) *For each i, $\bar{r}^i$ is bounded and Lipschitz continuous w.r.t. $(\bar{s}_t, \bar{a}_t^i)$ with constant $L_{\bar{r}^i}$. $\bar{F}$ is Lipschitz continuous w.r.t. $(\bar{s}, \bar{a}^1, \ldots, \bar{a}^m)$ with constant $L_{\bar{F}}$ in expectation.*

(b) *$\bar{v}^i_{\bar{\boldsymbol{\pi}}}$ is Lipschitz continuous w.r.t. $\bar{s}$ with constant $L_{\bar{v}_{\bar{\boldsymbol{\pi}}}}$.*

Notation: $\mathrm{Proj}(\bar{s}, \bar{a}^1 \ldots \bar{a}^m) = (\mathrm{Proj}_{\check{S}}(\bar{s}), \mathrm{Proj}_{\check{A}^1}(\bar{a}^1), \ldots, \mathrm{Proj}_{\check{A}^m}(\bar{a}^m))$.

Theorem (Discrete problem analysis)

Let $\epsilon > 0$. Suppose there is a unique pure policy $\bar{\boldsymbol{\pi}}_*^p$ for the MFTG for each i and $\bar{s} \in \bar{S}$, the function $v^i_{\bar{\boldsymbol{\pi}}_*^p}(\bar{s})$ is a global optimal point for the stage game $\bar{Q}^i_{\bar{\boldsymbol{\pi}}_*^p}(\bar{s})$. Then, if t is large enough, for each i, $\bar{s} \in \bar{S}$, $i = 1, 2, \cdots$, we have

$$|\check{Q}_t^i(\mathrm{Proj}(\bar{s}, \bar{a}^1 \ldots \bar{a}^m)) - \bar{Q}^i_{\bar{\boldsymbol{\pi}}_*^p}(\bar{s}, \bar{a}^1 \ldots \bar{a}^m)| \leq \epsilon',$$

where

$$\epsilon' = \epsilon(t) + C_1 \epsilon_A + C_2 \epsilon_S,$$

with $\epsilon(t) \to 0$ as $t \to +\infty$, ϵ_S and ϵ_A defined above, respectively, $C_1 = \frac{1}{1-\gamma}(L_{\bar{r}^i} + \gamma L_{\bar{v}^i_{\bar{\boldsymbol{\pi}}_*}} L_{\bar{F}} m)$ and $C_2 = \frac{\gamma}{1-\gamma} L_{\bar{v}^i_{\bar{\boldsymbol{\pi}}_*}} + L_{\bar{r}^i} + \gamma L_{\bar{v}^i_{\bar{\boldsymbol{\pi}}_*}} L_{\bar{F}}$.

Nash Q-Learning

Pros and cons:

- Convergence proof under suitable assumptions
- But requires solving a stage-game at each iteration
- Easy for games with finite and small spaces
- Hard for games with large or even continuous spaces
- Quantization is not scalable

Outline

Deep RL

Another idea:

- Do not discretize the simplexes
- Keep mean field state space as a continuous space
- Use deep RL to learn a policy as a NN
- Train one NN per central player (coalition)
- Evaluate convergence using exploitability
- For now, no proof of convergence

MFTG DDPG

Algorithm 2: DDPG for MFTG

Inputs: A number of episodes N; a length T for each episode; a minibatch size N_{batch}; a learning rate τ.
Outputs: Policy functions for each central player represented by $\pi^i_{\omega_i}$.
Initialize parameters θ_i and ω_i for critic networks $Q^i_{\theta_i}$ and actor networks $\pi^i_{\omega_i}$, $i = 1, ..., m$
Initialize $\theta'_i \leftarrow \theta_i$ and $\omega'_i \leftarrow \omega_i$ for target networks $Q^{i'}_{\theta'_i}$ and $\pi^{i'}_{\omega'_i}$, $i = 1, ..., m$
Initialize replay buffer R_{buffer}
for $k = 0, 1, ..., N-1$ **do**
  Initialize distribution $\bar{s}_0$
  for $t = 0, 1, \ldots, T-1$ **do**
    Select actions $\bar{a}^i_t = \pi^i_{\omega_i}(\bar{s}_t) + \epsilon_t$, where ϵ_t is the exploration noise, for $i = 1, ..., m$
    Execute $\bar{a}^i_t$, observe reward $\bar{r}^i(\bar{s}_t, \bar{a}^i_t)$, for $i = 1, ..., m$
    Observe $\bar{s}_{t+1}$
    Store transition $(\bar{s}_t, \bar{a}^1_t, ..., \bar{a}^m_t, \bar{r}^1_t, ..., \bar{r}^m_t, \bar{s}_{t+1})$ in R_{buffer}
    Sample a random minibatch of N_{batch} transitions $(\bar{s}_j, \bar{a}^1_j, ..., \bar{a}^m_j, \bar{r}^1_j, ..., \bar{r}^m_j, \bar{s}_{j+1})$ from R_{buffer}
    Set $y^i_j = \bar{r}^i_j + \gamma Q^{i'}_{\theta'_i}(\bar{s}_{j+1}, \pi^{i'}_{\omega'_i}(\bar{s}_{j+1}))$ for $i = 1, ..., m$, $j = 1, ..., N_{\text{batch}}$
    Update the critic networks by minimizing the loss:
    $L^i(\theta_i) = \frac{1}{N_{\text{batch}}} \sum_j (y^i_j - Q^i_{\theta_i}(\bar{s}_j, \bar{a}^i_j))^2$, for $i = 1, ..., m$
    Update the actor policies using the sampled policy gradients $\nabla_{\omega_i} v^i$, for $i = 1, ..., m$:

$$\nabla_{\omega_i} v^i(\omega_i) \approx \frac{1}{N_{\text{batch}}} \sum_j \nabla_{\bar{a}^i} Q^i_{\theta_i}(\bar{s}_j, \pi^i_{\omega_i}(\bar{s}_j)) \nabla_{\omega_i} \pi^i_{\omega_i}(\bar{s}_j)$$

    Update target networks: $\theta'_i \leftarrow \tau\theta_i + (1-\tau)\theta'_i$, $\omega'_i \leftarrow \tau\omega_i + (1-\tau)(\omega'_i)$, for $i = 1, ..., m$.
  end for
end for

Outline

Experimental Setup

- **Metrics:**
  - Testing rewards of each central player
  - Exploitability
- **Training and testing sets:**
  - Training set: randomly generated tuples of distributions
  - Testing set: a finite number of tuples of distributions that are not in the training set
- **Baseline:**
  - No baseline for our problems
  - Independent Learning-Mean Field Type Game (IL-MFTG): ablation study (hide the distribution of the other population)
- **Games:** 5 examples in the paper
- **Improvement:** Average exploitability improvement of at least 30% in each game

Exploitability

Definition

The **exploitability** of a policy profile $(\pi^1, \dots, \pi^m) \in \Pi^1 \times \cdots \times \Pi^m$ is

$$\mathcal{E}(\pi^1, \dots, \pi^m) = \sum_{i=1}^m \mathcal{E}^i(\pi^1, \dots, \pi^m),$$

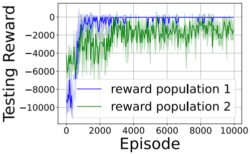
where the i-th central player's exploitability is:

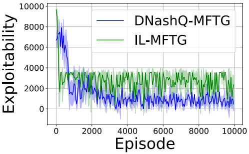
$$\mathcal{E}^i(\pi^1, \dots, \pi^m) = \max_{\tilde{\pi}^i \in \Pi^i} J^i(\tilde{\pi}^i; \pi^{-i}) - J^i(\pi^i; \pi^{-i}).$$

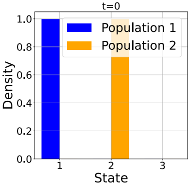
- $\mathcal{E}^i(\pi^1, \dots, \pi^m) =$ how much player i can be better off by deviating from π^i
- $\mathcal{E}(\pi^1, \dots, \pi^m) = 0$ iff $(\pi^1, \dots, \pi^m)$ is a Nash equilibrium for the MFTG

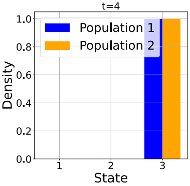
Example 1: 1D Population Matching Grid Game

- There are $m = 2$ populations
- Agent's state space: 3-state 1D grid world
- Actions: moving left, staying, and moving right, with individual noise perturbing the movements.
- Rewards: encourage Coalition 1 to stay where it is initialized but also to consider avoiding Coalition 2, and encourage Coalition 2 to match Coalition 1.





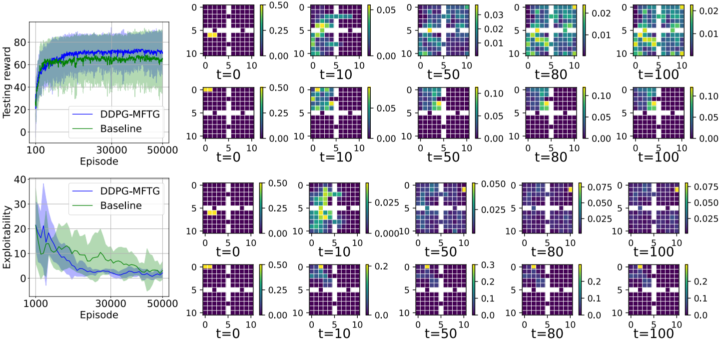




Ex. 1: Left and middle: averaged testing rewards and exploitabilities resp. (mean $\pm$ standard deviation). Right: one realization of population evolution at $t = 0$ and 4 for one testing distribution.

Example 2: Four-room with crowd aversion

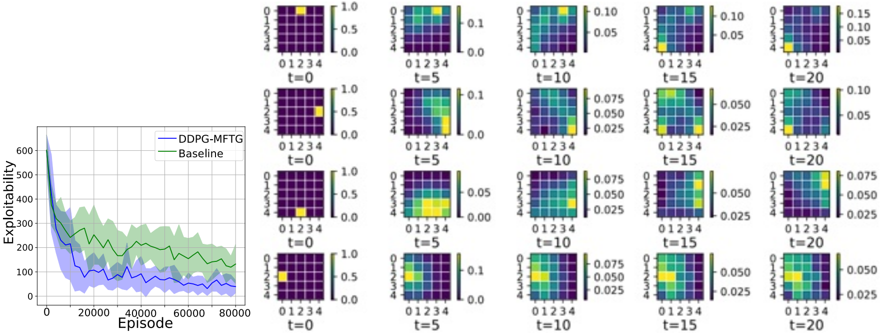
- There are $m = 2$ populations
- Agent's state space: 2D grid world composed of 4 connected rooms of size 5×5
- The policies' inputs are thus of dimension $2 \times 4 \times 5 \times 5 = 200$
- Rewards: encourage the two populations to spread while avoiding each other; Coalition 2 has a penalty for going to rooms other than the one she started in.



Ex. 2: Left, top and bottom: averaged testing rewards and exploitabilities resp. (mean $\pm$ stddev). Right, two top rows: distribution evolution of the two populations with our method. Right two bottom rows: distribution evolution with the baseline. Color bars indicate density values.

Example 3: Predator-prey 2D with 4 groups

- There are $m = 4$ populations.
- Player's state space is a 5×5-state 2D grid world with walls on the boundaries
- Rewards:Coalition 1 is a predator of Coalition 2, Coalition 2 avoids Coalition 1 and chases Coalition 3, which avoids Coalition 2 while chasing Coalition 4. Coalition 4 tries to avoid Coalition 3. There is also a cost for moving.



Ex. 3: Left: averaged exploitabilities (mean $\pm$ stddev). Right: populations evolution, one coalition per row and one time per column: $t = 0, 5, 10, 15, 20$. Color bars indicate density values.

Exploitability Results

We summarize the improvement brought by our method compared with the corresponding baseline, in each example.
The quantities are:

- **Baseline Exploitability:** The baseline's mean value (as described in the paper).
- **Our Exploitability:** Our method's mean value (as described in the paper).
- **Improvement:** The percentage improvement is calculated as:

$$\text{Improvement (percentage)} = \frac{\text{Baseline} - \text{Ours}}{\text{Baseline}} \times 100.$$

	Ex. 1	**Ex. 2**	**Ex. 3**	**Ex. 4**	**Ex. 5**
Baseline Exploit.	2355.35	3.13	131.43	2.69	6.93
Our Exploit.	471.40	2.16	38.75	1.39	3.14
Improvement	79.98%	31.0%	70.52%	48.3%	54.69%

Comparison of baseline and our exploitability metrics across the 5 examples described in the text, along with percentage improvement.

Outline

Summary and Perspectives

Summary:

- Finite space MFTGs: Nash equilibrium between mean field coalitions
- Reformulation in terms of MFMDP
- Nash Q-learning after quantization of the simplexes
- Deep RL without discretization

Paper: https://arxiv.org/abs/2409.18152

Perspectives:

- Problem setting: more complex settings, information structure, ...
- Numerical analysis: Proof of convergence for deep RL
- Numerical experiments: More complex examples and deep RL methods
- Continuous space (beyond LQ)

Thank you!

References I

[Angiuli et al., 2023] Angiuli, A., Detering, N., Fouque, J.-P., Lin, J., et al. (2023).
Reinforcement learning algorithm for mixed mean field control games.
Journal of Machine Learning, 2(2).

[Barreiro-Gomez and Tembine, 2019] Barreiro-Gomez, J. and Tembine, H. (2019).
Blockchain token economics: A mean-field-type game perspective.
IEEE Access, 7:64603–64613.

[Barreiro-Gomez and Tembine, 2021] Barreiro-Gomez, J. and Tembine, H. (2021).
Mean-field-type Games for Engineers.
CRC Press.

[Başar and Moon, 2021] Başar, T. and Moon, J. (2021).
Zero-sum differential games on the Wasserstein space.
Communications in Information and Systems, 21(2):219–251.

[Bensoussan et al., 2013] Bensoussan, A., Frehse, J., and Yam, P. (2013).
Mean field games and mean field type control theory, volume 101.
Springer.

[Bensoussan et al., 2018] Bensoussan, A., Huang, T., and Laurière, M. (2018).
Mean field control and mean field game models with several populations.
Minimax Theory and its Applications, 3(2):173–209.

References II

[Caines and Huang, 2019] Caines, P. E. and Huang, M. (2019).
Graphon mean field games and the GMFG equations: ε-Nash equilibria.
In *2019 IEEE 58th conference on decision and control (CDC)*, pages 286–292. IEEE.

[Carmona and Delarue, 2018] Carmona, R. and Delarue, F. (2018).
Probabilistic Theory of Mean Field Games with Applications I-II.
Springer.

[Carmona et al., 2020] Carmona, R., Hamidouche, K., Laurière, M., and Tan, Z. (2020).
Policy optimization for linear-quadratic zero-sum mean-field type games.
In *2020 59th IEEE Conference on Decision and Control (CDC)*, pages 1038–1043. IEEE.

[Carmona et al., 2023] Carmona, R., Laurière, M., and Tan, Z. (2023).
Model-free mean-field reinforcement learning: mean-field MDP and mean-field Q-learning.
The Annals of Applied Probability, 33(6B):5334–5381.

[Cirant, 2015] Cirant, M. (2015).
Multi-population mean field games systems with neumann boundary conditions.
Journal de Mathématiques Pures et Appliquées, 103(5):1294–1315.

[Cosso and Pham, 2019] Cosso, A. and Pham, H. (2019).
Zero-sum stochastic differential games of generalized McKean–Vlasov type.
Journal de Mathématiques Pures et Appliquées, 129:180–212.

References III

[Djehiche et al., 2017] Djehiche, B., Tcheukam, A., and Tembine, H. (2017).
Mean-field-type games in engineering.
AIMS Electronics and Electrical Engineering, 1(1):18–73.

[Gomes and Saúde, 2014] Gomes, D. A. and Saúde, J. (2014).
Mean field games models—a brief survey.
Dynamic Games and Applications, 4:110–154.

[Guan et al., 2024] Guan, Y., Afshari, M., and Tsiotras, P. (2024).
Zero-sum games between mean-field teams: Reachability-based analysis under mean-field sharing.
In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 9731–9739.

[Heinrich et al., 2015] Heinrich, J., Lanctot, M., and Silver, D. (2015).
Fictitious self-play in extensive-form games.
In *International conference on machine learning*, pages 805–813. PMLR.

[Hu and Wellman, 2003] Hu, J. and Wellman, M. P. (2003).
Nash Q-learning for general-sum stochastic games.
Journal of machine learning research, 4(Nov):1039–1069.

[Huang et al., 2006] Huang, M., Malhamé, R. P., and Caines, P. E. (2006).
Large population stochastic dynamic games: closed-loop McKean-Vlasov systems and the Nash certainty equivalence principle.
Commun. Inf. Syst., 6(3):221–251.

[Lasry and Lions, 2007] Lasry, J.-M. and Lions, P.-L. (2007).
Mean field games.
Jpn. J. Math., 2(1):229–260.

[Motte and Pham, 2022] Motte, M. and Pham, H. (2022).
Mean-field markov decision processes with common noise and open-loop controls.
The Annals of Applied Probability, 32(2):1421–1458.

[Parise and Ozdaglar, 2019] Parise, F. and Ozdaglar, A. (2019).
Graphon games.
In *Proceedings of the 2019 ACM Conference on Economics and Computation*, pages 457–458.

[Perrin et al., 2020] Perrin, S., Pérolat, J., Laurière, M., Geist, M., Elie, R., and Pietquin, O. (2020).
Fictitious play for mean field games: Continuous time analysis and applications.
Advances in Neural Information Processing Systems.

[Sanjari et al., 2023] Sanjari, S., Saldi, N., and Yüksel, S. (2023).
Nash equilibria for exchangeable team against team games and their mean field limit.
In *2023 American Control Conference (ACC)*, pages 1104–1109. IEEE.

[Subramanian et al., 2023] Subramanian, J., Kumar, A., and Mahajan, A. (2023).
Mean-field games among teams.
arXiv preprint arXiv:2310.12282.

[Tembine, 2014] Tembine, H. (2014).
Tutorial on mean-field-type games.
In *19th world congress of the international federation of automatic control (IFAC), Cape Town, South Africa*, pages 24–29.

[Tembine, 2015] Tembine, H. (2015).
Risk-sensitive mean-field-type games with Lp-norm drifts.
Automatica, 59:224–237.

[Tembine, 2017] Tembine, H. (2017).
Mean-field-type games.
AIMS Math, 2(4):706–735.

[uz Zaman et al., 2024] uz Zaman, M. A., Koppel, A., Laurière, M., and Başar, T. (2024).
Independent RL for cooperative-competitive agents: A mean-field perspective.
arXiv preprint arXiv:2403.11345.

References VI

[Yang et al., 2018] Yang, Y., Luo, R., Li, M., Zhou, M., Zhang, W., and Wang, J. (2018).
Mean field multi-agent reinforcement learning.
In *Proceedings of ICML*.

[Yüksel and Başar, 2024] Yüksel, S. and Başar, T. (2024).
Information dependent properties of equilibria: Existence, comparison, continuity and team-against-team games.
In *Stochastic Teams, Games, and Control under Information Constraints*, pages 395–436. Springer.

[Zaman et al., 2024] Zaman, M. A. U., Laurière, M., Koppel, A., and Başar, T. (2024).
Robust cooperative multi-agent reinforcement learning: A mean-field type game perspective.
In *6th Annual Learning for Dynamics & Control Conference*, pages 770–783. PMLR.