ETH zürich

Department of Computer Science

On the Sample Efficiency of Reinforcement Learning for Mean-Field Games

Jiawei Huang

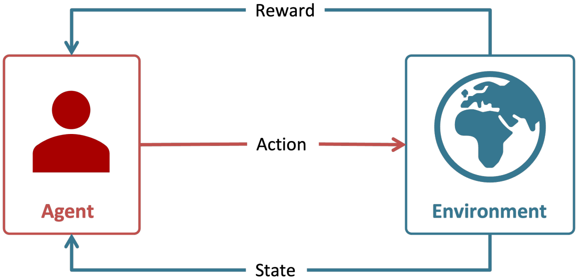
January 13, 2025

Outline

Outline

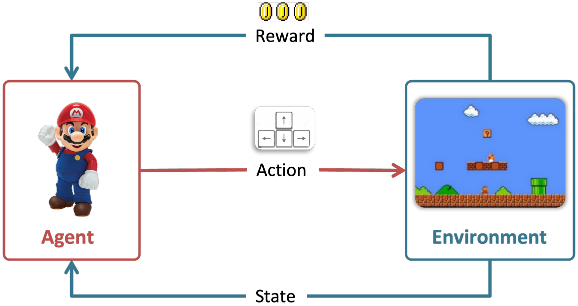
Reinforcement Learning (RL) in a Nutshell

- Learn to **make good decisions** from interactions with an **uncertain environment**.



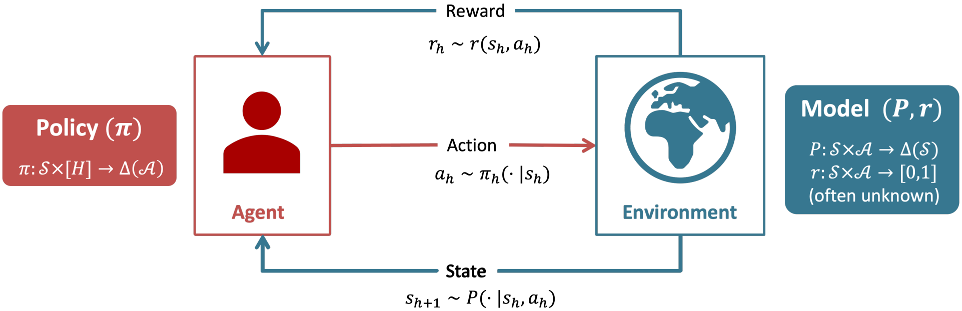
Example

- Learn to **make good decisions** from interactions with an **uncertain environment**.



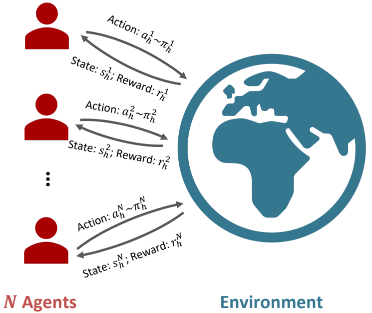
Mathematical Framework for Finite-Horizon RL

Goal: $\max_{\pi} J(\pi) = \mathbb{E}_{\pi}[\sum_{h=1}^{H} r_h]$



From One to Many: the Multi-Agent RL Setup

Multi-Agent RL

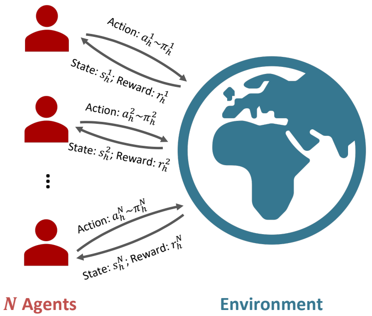


For agent $n = 1,2,\ldots,N$

$$(s_h^n)' \sim P^n\left(\cdot \,|s_h^1,\ldots,s_h^N,a_h^1,\ldots,a_h^N\right)$$
$$r_h^n \sim r^n(s_h^1,\ldots,s_h^N,a_h^1,\ldots,a_h^N)$$

Challenges in Large-Population Multi-Agent RL

Multi-Agent RL



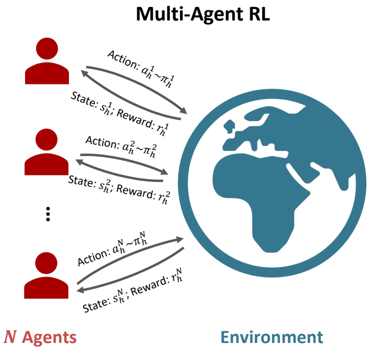
For agent $n = 1,2, \dots, N$

$$(s_h^n)' \sim P^n\left(\cdot \,|s_h^1, \dots, s_h^N, a_h^1, \dots, a_h^N\right)$$
$$r_h^n \sim r^n(s_h^1, \dots, s_h^N, a_h^1, \dots, a_h^N)$$

Curse of Multi-Agency

- The complexity of the system scales exponentially as the number of agents.

Challenges in Large-Population Multi-Agent RL



For agent $n = 1,2, \dots, N$

$$(s_h^n)' \sim P^n\left(\cdot \mid s_h^1, \dots, s_h^N, a_h^1, \dots, a_h^N\right)$$

$$r_h^n \sim r^n(s_h^1, \dots, s_h^N, a_h^1, \dots, a_h^N)$$

Curse of Multi-Agency

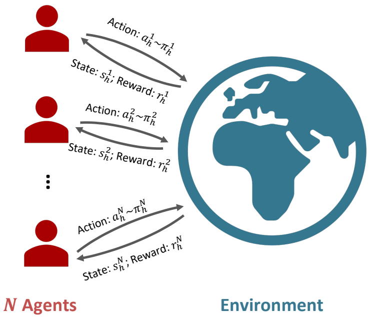
- The complexity of the system scales exponentially as the number of agents.

Curse of Computational Intractability

- Different from single-agent RL, we are interested in **Nash Equilibrium (NE)** policies.
  - At NE, no agent has incentives to deviate from their current policy.

Challenges in Large-Population Multi-Agent RL

Multi-Agent RL



For agent $n = 1,2, \dots, N$

$$(s_h^n)' \sim P^n\left(\cdot \,|s_h^1, \dots, s_h^N, a_h^1, \dots, a_h^N\right)$$
$$r_h^n \sim r^n(s_h^1, \dots, s_h^N, a_h^1, \dots, a_h^N)$$

Curse of Multi-Agency

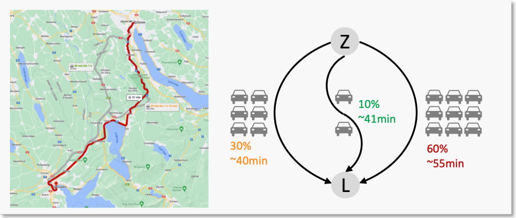
- The complexity of the system scales exponentially as the number of agents.

Curse of Computational Intractability

- Different from single-agent RL, we are interested in **Nash Equilibrium (NE)** policies.
  - At NE, no agent has incentives to deviate from their current policy.
- Computing NE for is PPAD-complete even for three players (Daskalakis et al., 2009).

Breaking the Curses by the Blessing of Symmetricity

Zurich → Luzern: Which route shall I choose?

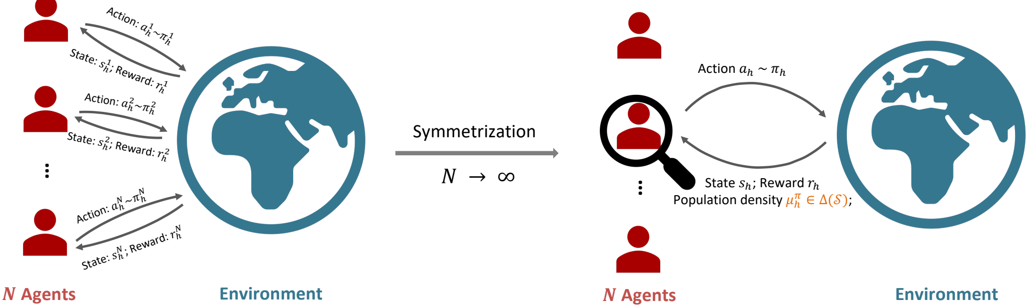


- **Agents**: drivers/cars;
- **Actions**: which routes to choose;
- The more drivers in one route, the longer time it takes;
- **Special Structure**: Large population and symmetric agents.
    - **Not important**: which agent take which route?
    - **Important**: what proportion of agents take each route?

Breaking the Curses by the Blessing of Symmetricity

Multi-Agent RL

Mean-Field Games (MFGs)
[Huang, Malhame, Caines'06, Lasry & Lions'05]



For agent $n = 1,2, \dots, N$

$$(s_h^n)' \sim P^n\left(\cdot \mid s_h^1, \dots, s_h^N, a_h^1, \dots, a_h^N\right)$$
$$r_h^n \sim r^n(s_h^1, \dots, s_h^N, a_h^1, \dots, a_h^N)$$

For a representative agent

$$s_h' \sim P\left(\cdot \mid s_h, a_h, \mu_h^\pi\right)$$
$$r_h \sim r(s_h, a_h, \mu_h^\pi)$$

Breaking the Curses by the Blessing of Symmetricity

Breaking the Curse of Multi-Agency

- transition and reward functions no longer depend on the number of agents.

Breaking the Curse of Computational Intractability

- NE can be computed efficienctly under some conditions (known transition/reward)
  - Contractivity (Guo et al., 2019; Yardim et al., 2023)
  - Monotonicity (Perolat et al., 2021; Zhang et al., 2024)
  - Sub-modularity (Dianetti et al., 2021)
  - ...

Outline

(Finite-Horizon) Mean-Field Games

Basic Setup

- $M := (\mathcal{S}, \mathcal{A}, \mu_1, H, \mathbb{P}, r)$
- $\mathcal{S}$ and $\mathcal{A}$: state and action space;
- $\mu_1 \in \Delta(\mathcal{S})$: initial state distribution;
- H: finite horizon;
- $\mathbb{P} := \{\mathbb{P}_h\}_{h=1}^{H}$, $r := \{r_h\}_{h=1}^{H}$: non-stationary transition and reward functions.

(Finite-Horizon) Mean-Field Games

Policy and Agents-Environment Interaction

- $M := (\mathcal{S}, \mathcal{A}, \mu_1, H, \mathbb{P}, r)$
- All the agents share a non-stationary policy $\pi := \{\pi_h\}_{h=1}^{H}$, $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$;
- Only need to focus on a representative agent

(Finite-Horizon) Mean-Field Games

Policy and Agents-Environment Interaction

- $M := (\mathcal{S}, \mathcal{A}, \mu_1, H, \mathbb{P}, r)$
- All the agents share a non-stationary policy $\pi := \{\pi_h\}_{h=1}^{H}$, $\pi_h : \mathcal{S} \to \Delta(\mathcal{A})$;
- Only need to focus on a representative agent
- Start with $s_1 \sim \mu_1$, for $h = 1, ..., H$
    - Take action $a_h \sim \pi_h(\cdot|s_h)$
    - Observe next state $s_{h+1} \sim \mathbb{P}_h(\cdot|s_h, a_h, \mu_h^\pi)$, and reward $r_h \leftarrow r_h(s_h, a_h, \mu_h^\pi)$
    - State density involves

$$\mu_1^\pi(\cdot) := \mu_1(\cdot)$$
$$\mu_{h+1}^\pi(\cdot) := \sum_{s_h \in \mathcal{S}, a_h \in \mathcal{A}} \mu_h^\pi(s_h)\pi_h(a_h|s_h)\mathbb{P}_h(\cdot|s_h, a_h, \mu_h^\pi).$$

(Finite-Horizon) Mean-Field Games

Learning objective: the Nash Equilibrium (NE)

- $M := (\mathcal{S}, \mathcal{A}, \mu_1, H, \mathbb{P}, r)$
- **Definition**: total return of a deviating agent taking π' while the other stick to π:

$$J_M(\pi', \pi) := \mathbb{E}\left[\sum_{h=1}^{H} r_h \middle| _{s_{h+1}\sim\mathbb{P}_h(\cdot|s_h,a_h,\mu_h^{\pi}),\ r_h\leftarrow r_h(s_h,a_h,\mu_h^{\pi})}^{\forall h,\ a_h\sim\pi'_h(\cdot|s_h),}\right].$$

- Policy π_M^{NE} is a NE of M if:

$$\forall\pi,\quad J_M(\pi_M^{\text{NE}}, \pi_M^{\text{NE}}) \geq J_M(\pi, \pi_M^{\text{NE}}). \qquad \text{(No incentive to deviate)}$$

- Policy $\widehat{\pi}_M^{\text{NE}}$ is an ε-NE of M if:

$$\forall\pi,\quad J_M(\widehat{\pi}_M^{\text{NE}}, \widehat{\pi}_M^{\text{NE}}) \geq J_M(\pi, \widehat{\pi}_M^{\text{NE}}) - \varepsilon. \qquad (\varepsilon\text{-incentive to deviate})$$

Key Question to Address in this Work

Practical Considerations

- **Model Uncertainty**
  - True MFGs model ($\mathbb{P}$ and r) may be unknown.
  - Need to estimate from interaction samples.
  - Generating samples can be costly (sample complexity matters).

Key Question to Address in this Work

Practical Considerations

- **Model Uncertainty**
- **Function Approximation**
  - Rich state and action spaces (large $\mathcal{S}, \mathcal{A}$)
  - Model/value functions depend on density ($\in$ uncountable set).

Key Question to Address in this Work

Practical Considerations

- **Model Uncertainty**
- **Function Approximation**
  - Rich state and action spaces (large $\mathcal{S}, \mathcal{A}$)
  - Model/value functions depend on density ($\in$ uncountable set).
  - To estimate model/value functions
    - Tabular representation is not efficient (scales as $|\mathcal{S}|, |\mathcal{A}|$ and covering number of $\Delta(\mathcal{S})$)
    - We need function approximations (e.g. neural networks).

Key Question to Address in this Work

Practical Considerations

- **Model Uncertainty**
- **Function Approximation**
  - Rich state and action spaces (large $\mathcal{S}, \mathcal{A}$)
  - Model/value functions depend on density ($\in$ uncountable set).
  - To estimate model/value functions
    - Tabular representation is not efficient (scales as $|\mathcal{S}|, |\mathcal{A}|$ and covering number of $\Delta(\mathcal{S})$)
    - We need function approximations (e.g. neural networks).
  - Theoretical formulation:
    - A set of functions are available to approximate true model/optimal value.
    - The sample complexity would depend on complexity of function class.

Key Question to Address in this Work

Practical Considerations

- **Model Uncertainty**
- **Function Approximation**

Literature Previous to our work and Limitations

	Unknown Model ?	Non-Tabular Setting ?	Other Remarks
(Huang et al., 2006) (Lasry and Lions, 2007) (Bensoussan et al., 2013)	✗	✗	
(Guo et al., 2019) (Perolat et al., 2021)	✓	✗	Require additional structural assumptions
(Pasztor et al., 2021)	✓	✓	Mean-Field Control Setting (“Cooperative” MFGs)

Key Question to Address in this Work

What is the **sample complexity** for solving **NE** in MFGs with RL
with **general function approximation**?

Challenges

- How to do strategic exploration?
- Due to MFGs' special structure, previous results in single-agent RL or Markov Games are not directly appliable.

Outline

Setting and Assumptions

Model-Based Function Approximation Setting

- For convenience, assume true reward r^* is known (can be extended to unknown reward setting)
- A model function class $\mathcal{M} = \{M_1, M_2, ..., M_{|\mathcal{M}|}\}$ is available, $M_i := \{\mathbb{P}_{M_i,h}\}_{h\in[H]}$.

Setting and Assumptions

Model-Based Function Approximation Setting

- For convenience, assume true reward r^* is known (can be extended to unknown reward setting)
- A model function class $\mathcal{M} = \{M_1, M_2, ..., M_{|\mathcal{M}|}\}$ is available, $M_i := \{\mathbb{P}_{M_i,h}\}_{h\in[H]}$.

Assumptions

1. **Realizability**: The true model $M^* := \{\mathbb{P}_{M^*,h}\}_{h\in[H]} \in \mathcal{M}$
2. **Lipschitz Continuity in Density**: $\forall M \in \mathcal{M}, \forall h, s_h, a_h, \forall \mu, \mu' \in \Delta(\mathcal{S})$

$$\begin{aligned}\|\mathbb{P}_{M,h}(\cdot|s_h, a_h, \mu) - \mathbb{P}_{M,h}(\cdot|s_h, a_h, \mu')\|_1 \leq & L_T\|\mu - \mu'\|_1, \\ |r^*_h(s_h, a_h, \mu) - r^*_h(s_h, a_h, \mu')| \leq & L_r\|\mu - \mu'\|_1.\end{aligned}$$

Setting and Assumptions

Model-Based Function Approximation Setting

- For convenience, assume true reward r^* is known (can be extended to unknown reward setting)
- A model function class $\mathcal{M} = \{M_1, M_2, ..., M_{|\mathcal{M}|}\}$ is available, $M_i := \{\mathbb{P}_{M_i,h}\}_{h\in[H]}$.

Assumptions

1. **Realizability**: The true model $M^* := \{\mathbb{P}_{M^*,h}\}_{h\in[H]} \in \mathcal{M}$
2. **Lipschitz Continuity in Density**: $\forall M \in \mathcal{M}, \forall h, s_h, a_h, \forall \mu, \mu' \in \Delta(\mathcal{S})$

$$\begin{aligned}\|\mathbb{P}_{M,h}(\cdot|s_h, a_h, \mu) - \mathbb{P}_{M,h}(\cdot|s_h, a_h, \mu')\|_1 \leq & L_T\|\mu - \mu'\|_1, \\ |r^*_h(s_h, a_h, \mu) - r^*_h(s_h, a_h, \mu')| \leq & L_r\|\mu - \mu'\|_1.\end{aligned}$$

Data Collection Oracle (Centralized MFGs)

- Given any two policies π and π', we assume an oracle can return a trajectory generated by

$$a_h \sim \pi'_h(\cdot|s_h),\ r_h \leftarrow r^*_h(s_h, a_h, \mu^\pi_{M^*,h}),\ s_{h+1} \sim \mathbb{P}_{M^*,h}(\cdot|s_h, a_h, \mu^\pi_{M^*,h}).$$

- $\approx$ the trajectory of one agent taking π' while the others take π in finite N-agent system.
- Sample complexity := number of queries to the oracle

Function Approximation Complexity Measure

Rich Literature in Single-Agent Setting

- Eluder Dimension (Levy et al., 2022; Osband and Van Roy, 2014; Russo and Van Roy, 2013)
- Bellman Rank/Witness Rank (Jiang et al., 2017; Sun et al., 2019)
- Bellman Eluder Dimension (Jin et al., 2021)
- Low-Rank MDP (Agarwal et al., 2020; Uehara et al., 2021)
- Bilinear Rank (Du et al., 2021)
- Decision to Estimation Coefficient (Foster et al., 2021)
- Coverage Coefficient (Xie et al., 2022)
- ...

Function Approximation Complexity Measure

For Mean-Field model class $\mathcal{M}$, we get inspired from

- Eluder Dimension (Levy et al., 2022; Osband and Van Roy, 2014; Russo and Van Roy, 2013)
    - Denote as dimE$(\mathcal{M})$;
    - (To make life easier, we omit its formal definition here).
    - Similar to VC-dimension, measures the complexity (expressive power) of $\mathcal{M}$.

Function Approximation Complexity Measure

For Mean-Field model class $\mathcal{M}$, we get inspired from

- Eluder Dimension (Levy et al., 2022; Osband and Van Roy, 2014; Russo and Van Roy, 2013)
    - Denote as dimE$(\mathcal{M})$;
    - (To make life easier, we omit its formal definition here).
    - Similar to VC-dimension, measures the complexity (expressive power) of $\mathcal{M}$.

Is Sample Complexity Scaling with Complexity of $\mathcal{M}$ Good Enough?

Function Approximation Complexity Measure

For Mean-Field model class $\mathcal{M}$, we get inspired from

- Eluder Dimension (Levy et al., 2022; Osband and Van Roy, 2014; Russo and Van Roy, 2013)
    - Denote as dimE$(\mathcal{M})$;
    - (To make life easier, we omit its formal definition here).
    - Similar to VC-dimension, measures the complexity (expressive power) of $\mathcal{M}$.

Is Sample Complexity Scaling with Complexity of $\mathcal{M}$ Good Enough?

- Different from single-agent setting, the transition functions are defined on $\mathcal{S} \times \mathcal{A} \times \Delta(\mathcal{S})$.
- The complexity of $\mathcal{M}$ can be extremely high:
    - In the worst cases, dimE$(\mathcal{M})$ is exponential in $\exp(|\mathcal{S}|)$.
- Can we do better?

Main Result

Theorem (Informal)

Given a Mean-Field model class $\mathcal{M}$, satisfying Realizability and Lipschitz continuity assumptions, learning an ε-NE with probability $1-\delta$ only consumes samples at most:

$$\widetilde{O}(\mathit{Poly}(\mathit{dimPE}(\mathcal{M}), H, 1+L_T, L_r, \frac{1}{\varepsilon}, \log\frac{|\mathcal{M}|}{\delta}))$$

A New Complexity Measure: Partial Model-Based Eluder Dimension (dimPE$(\mathcal{M})$)

- Given an arbitrary policy π, define

$$\mathcal{M}_{|\pi} := \{M_{|\pi} | M \in \mathcal{M}\}$$

  with $M_{|\pi} := \{\mathbb{P}_{M,h}(\cdot|\cdot,\cdot,\mu^{\pi}_{M,h})\}_{h\in[H]}$.
- dimPE$(\mathcal{M}) := \max_\pi$ dimE$(\mathcal{M}_{|\pi})$.
- Essentially, dimPE$(\mathcal{M})$ measures the complexity of the single-agent model class $\mathcal{M}_{|\pi}$ for some (adversarially) chosen π.

Main Result

Theorem (Informal)

Given a Mean-Field model class $\mathcal{M}$, satisfying Realizability and Lipschitz continuity assumptions, learning an ε-NE with probability $1-\delta$ only consumes samples at most:

$$\widetilde{O}(\mathit{Poly}(\mathit{dimPE}(\mathcal{M}), H, 1+L_T, L_r, \frac{1}{\varepsilon}, \log\frac{|\mathcal{M}|}{\delta}))$$

Interpretation: Model-Based RL for MFGs is not Statistically Harder than Single-Agent RL

Main Result

Theorem (Informal)

Given a Mean-Field model class $\mathcal{M}$, satisfying Realizability and Lipschitz continuity assumptions, learning an ε-NE with probability $1-\delta$ only consumes samples at most:

$$\widetilde{O}(\textit{Poly}(\textit{dimPE}(\mathcal{M}), H, 1+L_T, L_r, \frac{1}{\varepsilon}, \log\frac{|\mathcal{M}|}{\delta}))$$

Interpretation: Model-Based RL for MFGs is not Statistically Harder than Single-Agent RL

- dimPE$(\mathcal{M}) \leq |\mathcal{S}||\mathcal{A}|$
  - Tabular MFGs is sample-efficient in general.

Main Result

Theorem (Informal)

Given a Mean-Field model class $\mathcal{M}$, satisfying Realizability and Lipschitz continuity assumptions, learning an ε-NE with probability $1-\delta$ only consumes samples at most:

$$\widetilde{O}(\textit{Poly}(\textit{dimPE}(\mathcal{M}), H, 1+L_T, L_r, \frac{1}{\varepsilon}, \log\frac{|\mathcal{M}|}{\delta}))$$

Interpretation: Model-Based RL for MFGs is not Statistically Harder than Single-Agent RL

- dimPE$(\mathcal{M}) \leq |\mathcal{S}||\mathcal{A}|$
  - Tabular MFGs is sample-efficient in general.
- Linear dynamics: $\mathbb{P}_{M,h}(s_{h+1}|s_h,a_h,\mu_h) = \phi(s_h,a_h)^\top U_h(\mu_h)\psi(s_{h+1})$
  - $\phi \in \mathbb{R}^d, U \in \mathbb{R}^{d\times d'}, \psi \in \mathbb{R}^{d'}$.
  - In general $d' \gg d$; dimPE$(\mathcal{M}) = \widetilde{O}(d)$, while dimE$(\mathcal{M}) = \widetilde{O}(d')$.

Main Result

Theorem (Informal)

Given a Mean-Field model class $\mathcal{M}$, satisfying Realizability and Lipschitz continuity assumptions, learning an ε-NE with probability $1-\delta$ only consumes samples at most:

$$\widetilde{O}(\textit{Poly}(\textit{dimPE}(\mathcal{M}), H, 1+L_T, L_r, \frac{1}{\varepsilon}, \log\frac{|\mathcal{M}|}{\delta}))$$

Interpretation: Model-Based RL for MFGs is not Statistically Harder than Single-Agent RL

- dimPE$(\mathcal{M}) \leq |\mathcal{S}||\mathcal{A}|$
  - Tabular MFGs is sample-efficient in general.
- Linear dynamics: $\mathbb{P}_{M,h}(s_{h+1}|s_h, a_h, \mu_h) = \phi(s_h, a_h)^\top U_h(\mu_h)\psi(s_{h+1})$
  - $\phi \in \mathbb{R}^d, U \in \mathbb{R}^{d\times d'}, \psi \in \mathbb{R}^{d'}$.
  - In general $d' \gg d$; dimPE$(\mathcal{M}) = \widetilde{O}(d)$, while dimE$(\mathcal{M}) = \widetilde{O}(d')$.
- Not computationally efficient for now.

Outline

A Model-Elimination Based Algorithms

Algorithm Sketch

For $k = 1, 2, ...$, (start with $\mathcal{M}^1 := \mathcal{M}$)

1. Find a desired policy π^k
2. Construct $\mathcal{M}^k_{|\pi^k} := \{M_{|\pi^k} | M \in \mathcal{M}^k\}$.
   i.e. fix the density with π^k for each $M \in \mathcal{M}^k$.
3. Collect samples and $\mathcal{M}^{k+1} \leftarrow \{M \in \mathcal{M}^k | M_{|\pi^k} \approx M^*_{\pi^k}\}$.

A Model-Elimination Based Algorithms

Algorithm Sketch

For $k = 1, 2, ...$, (start with $\mathcal{M}^1 := \mathcal{M}$)

1. Find a desired policy π^k
2. Construct $\mathcal{M}^k_{|\pi^k} := \{M_{|\pi^k} | M \in \mathcal{M}^k\}$.
   i.e. fix the density with π^k for each $M \in \mathcal{M}^k$.
3. Collect samples and $\mathcal{M}^{k+1} \leftarrow \{M \in \mathcal{M}^k | M_{|\pi^k} \approx M^*_{\pi^k}\}$
   – the only step we collect samples
   – essentially a single-agent model elimination procedure
   – all the agent take π^k except one doing exploration

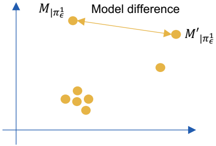
A Model-Elimination Based Algorithms

Algorithm Sketch

For $k = 1, 2, ...$, (start with $\mathcal{M}^1 := \mathcal{M}$)

1. Find a **desired policy** π^k – the key step
2. Construct $\mathcal{M}^k_{|\pi^k} := \{M_{|\pi^k} | M \in \mathcal{M}^k\}$.
   i.e. fix the density with π^k for each $M \in \mathcal{M}^k$.
3. Collect samples and $\mathcal{M}^{k+1} \leftarrow \{M \in \mathcal{M}^k | M_{|\pi^k} \approx M^*_{\pi^k}\}$
   – the only step we collect samples
   – essentially a single-agent model elimination procedure
   – all the agent take π^k except one doing exploration

Key Step: How to Choose π^k for Fast Elimination?

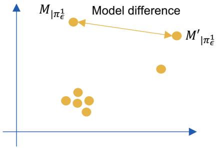


Visualization of $\mathcal{M}^k_{|\pi_\epsilon^1}$

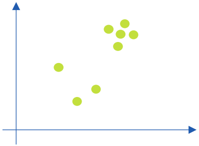
(dots represent models in $\mathcal{M}^k_{|\pi_\epsilon^1}$)

Consider ϵ-cover for policy space $\Pi_\epsilon := \{\pi_\epsilon^1, \pi_\epsilon^2 \ldots\}$

Key Step: How to Choose π^k for Fast Elimination?

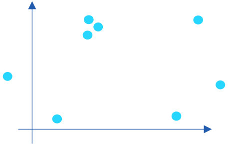


Visualization of $\mathcal{M}^k_{|\pi_\epsilon^1}$
(dots represent models in $\mathcal{M}^k_{|\pi_\epsilon^1}$)



Visualization of $\mathcal{M}^k_{|\pi_\epsilon^2}$
(dots represent models in $\mathcal{M}^k_{|\pi_\epsilon^2}$)

...



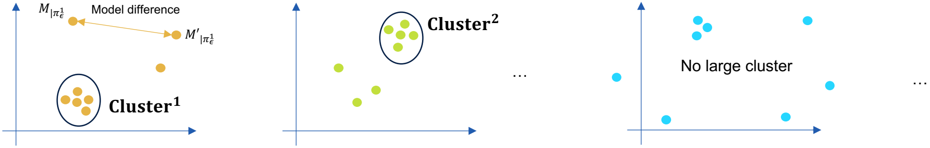
Visualization of $\mathcal{M}^k_{|\pi_\epsilon^i}$
(dots represent models in $\mathcal{M}^k_{|\pi_\epsilon^i}$)

...

Consider ϵ-cover for policy space $\Pi_\epsilon \coloneqq \{\pi_\epsilon^1, \pi_\epsilon^2 \ldots\}$

Key Step: How to Choose π^k for Fast Elimination?

Case 1: Non-concentrated setting



Visualization of $\mathcal{M}^k_{|\pi_\epsilon^1}$ (dots represent models in $\mathcal{M}^k_{|\pi_\epsilon^1}$)

Visualization of $\mathcal{M}^k_{|\pi_\epsilon^2}$ (dots represent models in $\mathcal{M}^k_{|\pi_\epsilon^2}$)

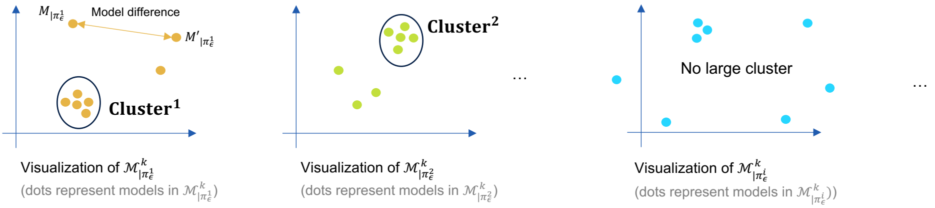
Visualization of $\mathcal{M}^k_{|\pi_\epsilon^i}$ (dots represent models in $\mathcal{M}^k_{|\pi_\epsilon^i}$))

Consider ϵ-cover for policy space $\Pi_\epsilon \coloneqq \{\pi_\epsilon^1, \pi_\epsilon^2 \ldots\}$

- $\exists \pi_\varepsilon^i \in \Pi_\varepsilon$, s.t. no $O(\varepsilon)$-cluster with more than $\frac{|\mathcal{M}^k|}{2}$ models.

Key Step: How to Choose π^k for Fast Elimination?

Case 1: Non-concentrated setting

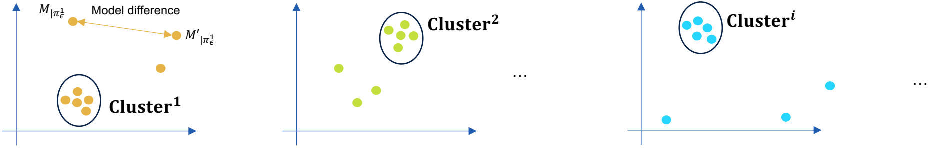


Consider ϵ-cover for policy space $\Pi_\epsilon \coloneqq \{\pi_\epsilon^1, \pi_\epsilon^2 \ldots\}$

- $\exists \pi_\varepsilon^i \in \Pi_\varepsilon$, s.t. no $O(\varepsilon)$-cluster with more than $\frac{|\mathcal{M}^k|}{2}$ models.
- By choosing $\pi^k \leftarrow \pi_\varepsilon^i$, only models surrounds $M^*_{|\pi_\varepsilon^i}$ remains
- Therefore, $|\mathcal{M}^{k+1}| \leq \frac{|\mathcal{M}^k|}{2}$.

Key Step: How to Choose π^k for Fast Elimination?

Case 2: Concentrated setting



Visualization of $\mathcal{M}^k_{|\pi_\epsilon^1}$ (dots represent models in $\mathcal{M}^k_{|\pi_\epsilon^1}$)

Visualization of $\mathcal{M}^k_{|\pi_\epsilon^2}$ (dots represent models in $\mathcal{M}^k_{|\pi_\epsilon^2}$)

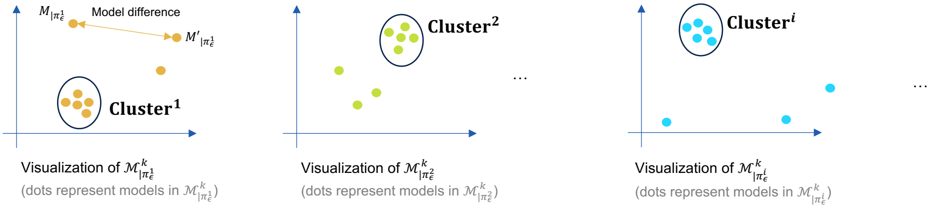
Visualization of $\mathcal{M}^k_{|\pi_\epsilon^i}$ (dots represent models in $\mathcal{M}^k_{|\pi_\epsilon^i}$)

Consider ϵ-cover for policy space $\Pi_\epsilon \coloneqq \{\pi_\epsilon^1, \pi_\epsilon^2 \ldots\}$

- $\forall \pi_\varepsilon^i \in \Pi_\varepsilon$, there exists an $O(\varepsilon)$-cluster with more than $\frac{|\mathcal{M}^k|}{2}$ models.
- Thanks to Lipschitz continuity
  1. **Local alignment lemma**: If $M_{|\pi} \approx M^*_{|\pi}$ and $\pi \approx$ NE of M, then $\pi \approx$ NE of M^*
  2. **"Fixed point" structure**: $\exists \pi_\varepsilon^i \in \Pi_\varepsilon$, s.t. $\pi_\varepsilon^i \approx$ NE of all models in that $O(\varepsilon)$-cluster.

Key Step: How to Choose π^k for Fast Elimination?

Case 2: Concentrated setting

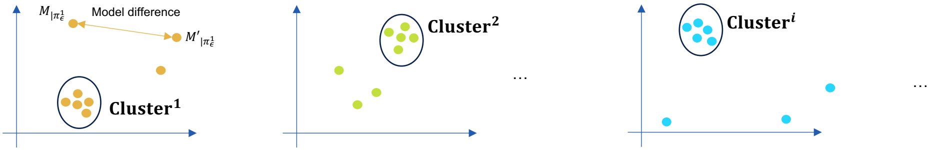


Consider ϵ-cover for policy space $\Pi_\epsilon \coloneqq \{\pi_\epsilon^1, \pi_\epsilon^2 \dots\}$

By choosing $\pi^k = \pi_\varepsilon^i$, run model-elimination and get $\mathcal{M}^{k+1}$:

Key Step: How to Choose π^k for Fast Elimination?

Case 2: Concentrated setting



Visualization of $\mathcal{M}^k_{|\pi_\epsilon^1}$ (dots represent models in $\mathcal{M}^k_{|\pi_\epsilon^1}$)

Visualization of $\mathcal{M}^k_{|\pi_\epsilon^2}$ (dots represent models in $\mathcal{M}^k_{|\pi_\epsilon^2}$)

Visualization of $\mathcal{M}^k_{|\pi_\epsilon^i}$ (dots represent models in $\mathcal{M}^k_{|\pi_\epsilon^i}$)

Consider ϵ-cover for policy space $\Pi_\epsilon := \{\pi_\epsilon^1, \pi_\epsilon^2 \dots\}$

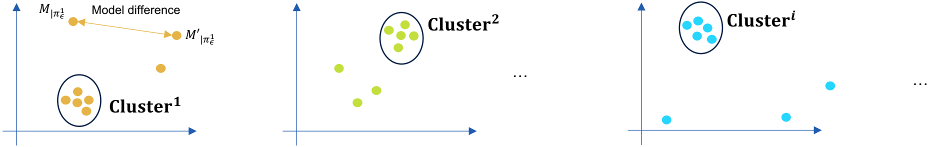
By choosing $\pi^k = \pi_\varepsilon^i$, run model-elimination and get $\mathcal{M}^{k+1}$:

- If $\texttt{Cluster}^i \cap \mathcal{M}^{k+1} \neq \emptyset$:

  $M^*_{|\pi_\varepsilon^i} \in \texttt{Cluster}^i$, and therefore, $\pi_\varepsilon^i \approx$ NE of M^*.

Key Step: How to Choose π^k for Fast Elimination?

Case 2: Concentrated setting



Visualization of $\mathcal{M}^k_{|\pi_\epsilon^1}$
(dots represent models in $\mathcal{M}^k_{|\pi_\epsilon^1}$)

Visualization of $\mathcal{M}^k_{|\pi_\epsilon^2}$
(dots represent models in $\mathcal{M}^k_{|\pi_\epsilon^2}$)

Visualization of $\mathcal{M}^k_{|\pi_\epsilon^i}$
(dots represent models in $\mathcal{M}^k_{|\pi_\epsilon^i}$)

Consider ϵ-cover for policy space $\Pi_\epsilon \coloneqq \{\pi_\epsilon^1, \pi_\epsilon^2 \ldots\}$

By choosing $\pi^k = \pi_\varepsilon^i$, run model-elimination and get $\mathcal{M}^{k+1}$:

- If $\texttt{Cluster}^i \cap \mathcal{M}^{k+1} \neq \emptyset$:
  $M^*_{|\pi_\varepsilon^i} \in \texttt{Cluster}^i$, and therefore, $\pi_\varepsilon^i \approx$ NE of M^*.
- If $\texttt{Cluster}^i \cap \mathcal{M}^{k+1} = \emptyset$:
  $|\mathcal{M}^{k+1}| \leq \frac{|\mathcal{M}^k|}{2}$ because of the size of that cluster.

Put Everything Together

Case 1: Non-concentrated setting

- Every time $|\mathcal{M}^{k+1}| \leq \frac{|\mathcal{M}^k|}{2}$

Case 2: Concentrated setting

- Either find a NE or $|\mathcal{M}^{k+1}| \leq \frac{|\mathcal{M}^k|}{2}$.

Put Everything Together

Case 1: Non-concentrated setting

- Every time $|\mathcal{M}^{k+1}| \leq \frac{|\mathcal{M}^k|}{2}$

Case 2: Concentrated setting

- Either find a NE or $|\mathcal{M}^{k+1}| \leq \frac{|\mathcal{M}^k|}{2}$.

Conclusion

- $O(\log |\mathcal{M}|)$ elimination steps at most.
- Each elimination costs Poly$(\text{dimE}(\mathcal{M}^k_{|\pi^k})) =$ Poly$(\text{dimPE}(\mathcal{M}))$ samples.
- Q.E.D.

Outline

Summary

Take Aways

- A new complexity measure: Partial Model-Based Eluder Dimension;
- A new model elimination based RL algorithm for centralized MFGs;

Under realizability and Lipschitz conditions, Model-Based RL for centralized MFGs is not Statistically Harder than Single-Agent RL.

Future Directions

- Computational efficiency;
- Decentralized setting;
- Equilibrium selection, steering, mechanism design.

Collaborators and Related Papers

Batuhan Yardim
(ETH Zurich)

Niao He
(ETH Zurich)

Andreas Krause
(ETH Zurich)

AISTATS 2024 J. Huang, B. Yardim, and N. He. On the Statistical Efficiency of Mean-Field Reinforcement Learning with General Function Approximation

ICML 2024 J. Huang, N. He. and A. Krause. Model-Based RL for Mean-Field Games is not Statistically Harder than Single-Agent RL

Thanks!

References I

Agarwal, A., Kakade, S., Krishnamurthy, A., and Sun, W. (2020). Flambe: Structural complexity and representation learning of low rank mdps. *Advances in neural information processing systems*, 33:20095–20107.

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

Daskalakis, C., Goldberg, P. W., and Papadimitriou, C. H. (2009). The complexity of computing a nash equilibrium. *Communications of the ACM*, 52(2):89–97.

Dianetti, J., Ferrari, G., Fischer, M., and Nendel, M. (2021). Submodular mean field games: Existence and approximation of solutions. *The Annals of Applied Probability*, 31(6):2538–2566.

Du, S., Kakade, S., Lee, J., Lovett, S., Mahajan, G., Sun, W., and Wang, R. (2021). Bilinear classes: A structural framework for provable generalization in rl. In *International Conference on Machine Learning*, pages 2826–2836. PMLR.

Foster, D. J., Kakade, S. M., Qian, J., and Rakhlin, A. (2021). The statistical complexity of interactive decision making. *arXiv preprint arXiv:2112.13487*.

References II

Guo, X., Hu, A., Xu, R., and Zhang, J. (2019). Learning mean-field games. *Advances in neural information processing systems*, 32.

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.

Jiang, N., Krishnamurthy, A., Agarwal, A., Langford, J., and Schapire, R. E. (2017). Contextual decision processes with low bellman rank are pac-learnable. In *International Conference on Machine Learning*, pages 1704–1713. PMLR.

Jin, C., Liu, Q., and Miryoosefi, S. (2021). Bellman eluder dimension: New rich classes of rl problems, and sample-efficient algorithms. *Advances in neural information processing systems*, 34:13406–13418.

Lasry, J.-M. and Lions, P.-L. (2007). Mean field games. *Japanese journal of mathematics*, 2(1):229–260.

Levy, O., Cassel, A., Cohen, A., and Mansour, Y. (2022). Eluder-based regret for stochastic contextual mdps.

References III

Osband, I. and Van Roy, B. (2014). Model-based reinforcement learning and the eluder dimension. *Advances in Neural Information Processing Systems*, 27.

Pasztor, B., Bogunovic, I., and Krause, A. (2021). Efficient model-based multi-agent mean-field reinforcement learning. *arXiv preprint arXiv:2107.04050*.

Perolat, J., Perrin, S., Elie, R., Laurière, M., Piliouras, G., Geist, M., Tuyls, K., and Pietquin, O. (2021). Scaling up mean field games with online mirror descent. *arXiv preprint arXiv:2103.00623*.

Russo, D. and Van Roy, B. (2013). Eluder dimension and the sample complexity of optimistic exploration. *Advances in Neural Information Processing Systems*, 26.

Sun, W., Jiang, N., Krishnamurthy, A., Agarwal, A., and Langford, J. (2019). Model-based rl in contextual decision processes: Pac bounds and exponential improvements over model-free approaches. In *Conference on learning theory*, pages 2898–2933. PMLR.

Uehara, M., Zhang, X., and Sun, W. (2021). Representation learning for online and offline rl in low-rank mdps. *arXiv preprint arXiv:2110.04652*.

Xie, T., Foster, D. J., Bai, Y., Jiang, N., and Kakade, S. M. (2022). The role of coverage in online reinforcement learning. *arXiv preprint arXiv:2210.04157*.

References IV

Yardim, B., Cayci, S., Geist, M., and He, N. (2023). Policy mirror ascent for efficient and independent learning in mean field games. In *International Conference on Machine Learning*, pages 39722–39754. PMLR.

Zhang, F., Tan, V., Wang, Z., and Yang, Z. (2024). Learning regularized monotone graphon mean-field games. *Advances in Neural Information Processing Systems*, 36.