

Workshop on Aligning Reinforcement Learning
Experimentalists and Theorists

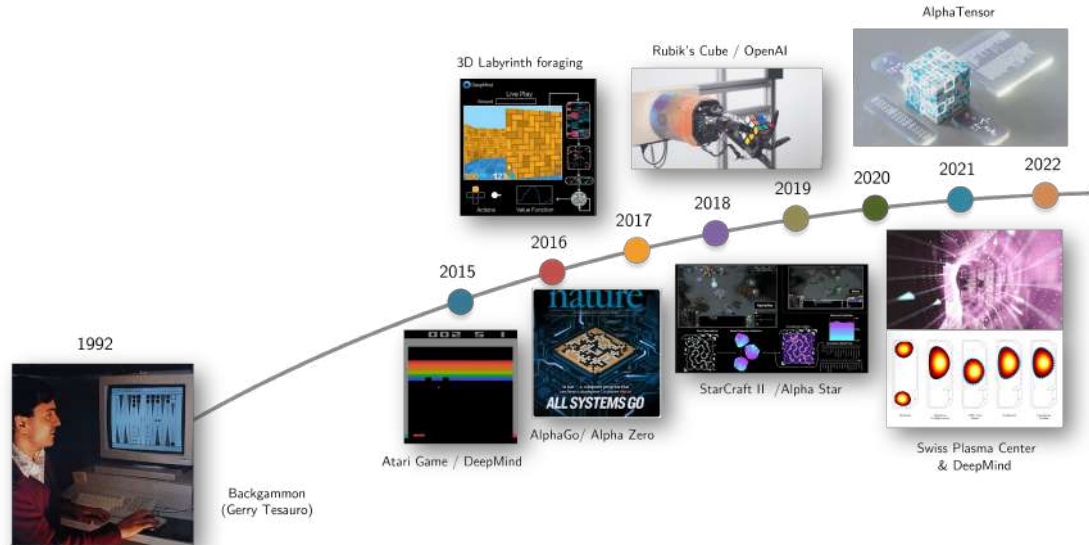
July 26, 2024

The Rise of Reinforcement Learning from One to Many

Niao He

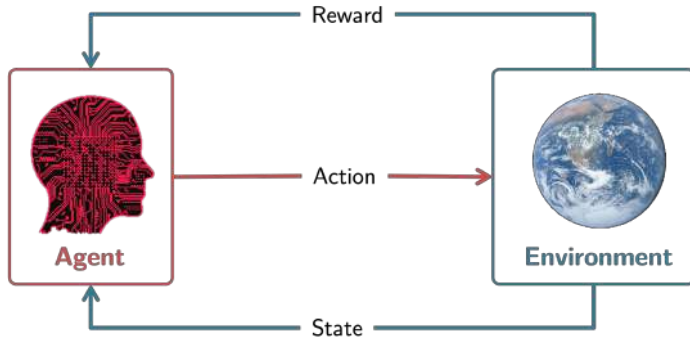


Reinforcement Learning (RL) in the Limelight



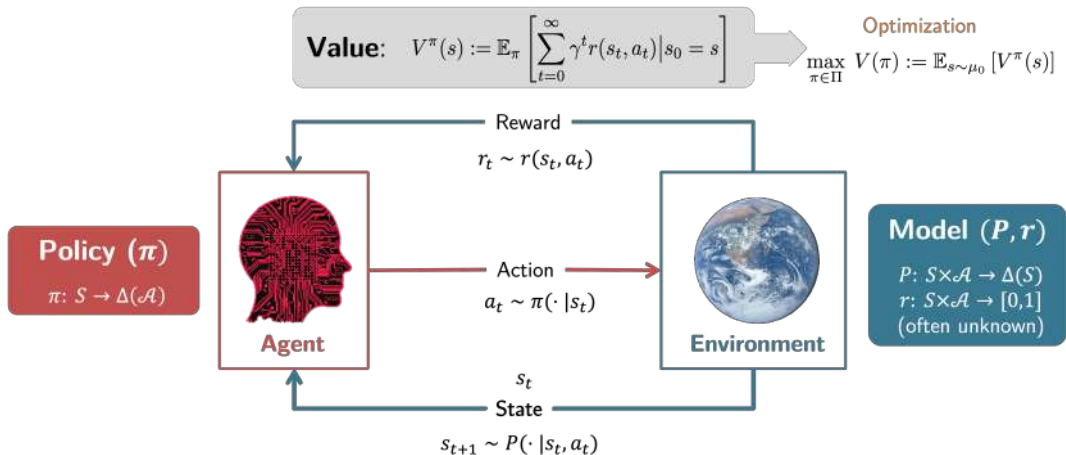
RL in a Nutshell

make good decisions by learning from experiences in an uncertain environment

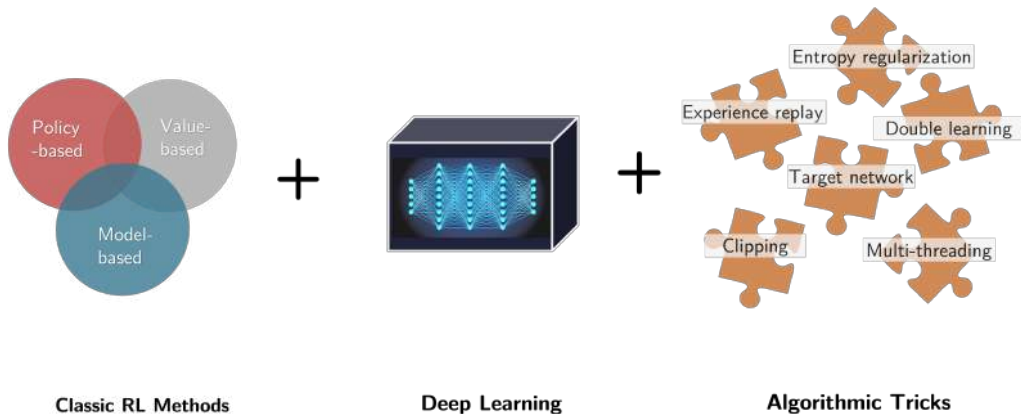


RL in Mathematical Framework

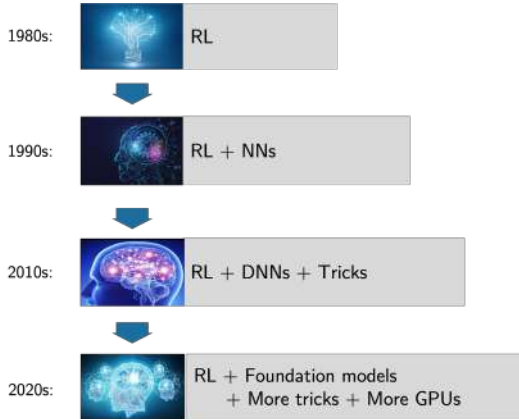
Goal: learning the optimal policy by maximizing (discounted) cumulative rewards



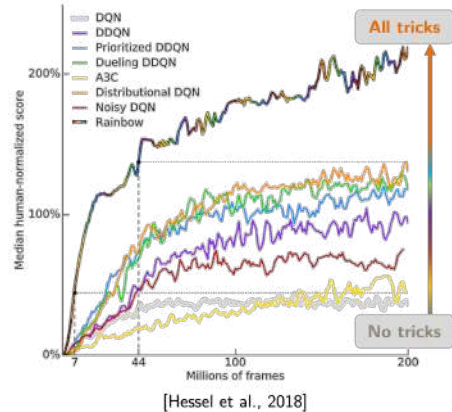
The Nuts and Bolts of Modern RL Solution



Grand Challenge I: Lack of Principles



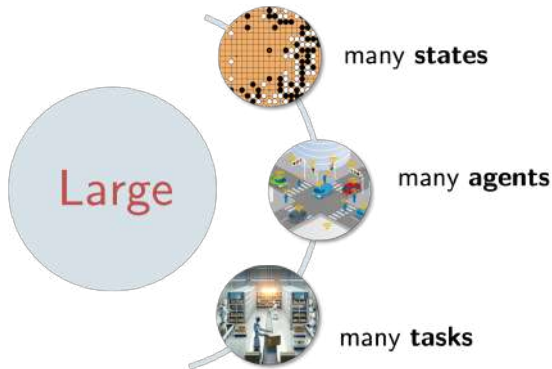
The more, the better ?



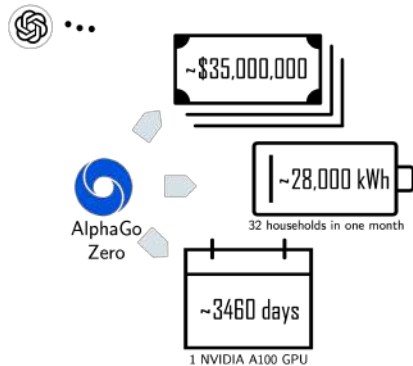
Hessel et al. Rainbow: Combining Improvements in Deep Reinforcement Learning. AAAI, 2018.

Heseel et al. Muesli: Combining Improvements in Policy Optimization. ICML, 2021.

Grand Challenge II: Need for Scalability



How much does it cost to train AlphaGo Zero?



This Talk: RL from One Agent to Many Agents

Single Agent



Policy Optimization

$$\max_{\pi} V(\pi)$$

Two Agents



Min-Max Optimization

$$\max_{\pi_1} \min_{\pi_2} V(\pi_1, \pi_2)$$

Finite Agents



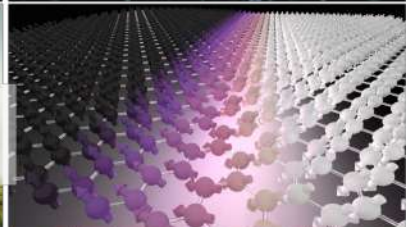
Equilibrium Optimization

$$\max_{\pi_i \in \Pi_i} V_i(\pi_i, \pi_{-i}), \forall i = 1, \dots, N$$

- ▶ Unlike the cooperative setting¹, the goal is seeking to compute the equilibrium.
- ▶ Significant progress in structured settings, e.g. zero-sum Markov Games, Markov Potential Games.
- ▶ Our focus: many-agent setting.



**Applications involving many
interacting decision makers are
everywhere.**



Curses of Many-agent RL



Curse of multiagency

The size of state-action space **scales exponentially** with the number of agents.

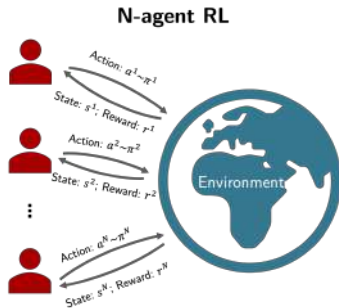
Curse of tractability

Computing Nash equilibrium for general-sum games is **PPAD-complete**.
[Daskalakis, Goldberg, Papadimitriou, 2008]

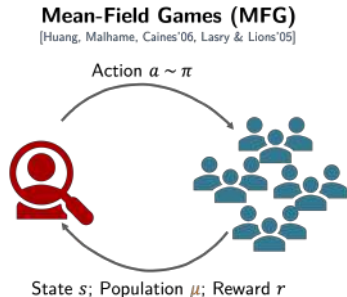
Curse of independent learning

Even in 3-player Markov game, it is **computational and statistically intractable** for agents to play independently via no-regret learning.
[Foster, Golowich, Kakade, 2023]

Breaking the Curses with Mean-Field Games!



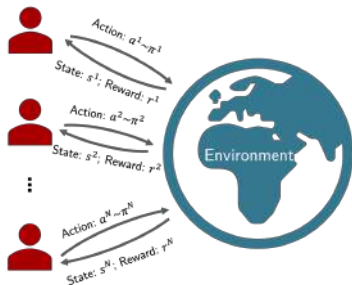
Symmetrization
 $N \rightarrow \infty$



for agent $i = 1, 2, \dots, N$
 $(s^i)' \sim P^i(\cdot | s^1, s^2, \dots, s^N, a^1, a^2, \dots, a^N)$
 $r^i \sim R^i(s^1, s^2, \dots, s^N, a^1, a^2, \dots, a^N)$

for a **representative** agent
 $s' \sim P(\cdot | s, a, \mu)$
 $r \sim R(s, a, \mu)$

N-player Dynamic Games



N-Player DGs (finite horizon)

- ▶ N agents, horizon H , finite state/action sets \mathcal{S}, \mathcal{A}
- ▶ State-action configuration at time h :
 $\rho_h = (s_h^i, a_h^i)_{i=1}^N \in (\mathcal{S} \times \mathcal{A})^N$
- ▶ Dynamics (for agent i):
 $s_0^i \sim \mu_0, s_{h+1}^i \sim P^i(\cdot | s_h^i, a_h^i, \rho_h^{-i}), a_h^i \sim \pi_h^i(\cdot | s_h^i)$
- ▶ Reward: $r_h^i = R^i(s_h^i, a_h^i, \rho_h^{-i})$

Policy $\boldsymbol{\pi}^* = (\pi^1, \dots, \pi^N)$ is an ϵ -Nash equilibrium policy if

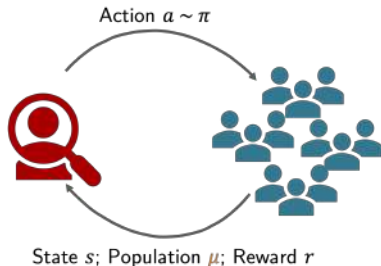
$$\underbrace{\mathcal{E}_N(\boldsymbol{\pi}^*)}_{\text{exploitability}} := \max_{i=1, \dots, N} \left\{ \max_{\pi^i} V^i(\pi^i, \boldsymbol{\pi}^{*-i}) - V^i(\boldsymbol{\pi}^*) \right\} \leq \epsilon.$$

- ▶ Markov policy (agent i): $\pi^i = \{\pi_h^i\}_{h=0}^{H-1}$
- ▶ Value function (agent i): $V^i(\boldsymbol{\pi}) := \mathbb{E} \left[\sum_{h=0}^{H-1} R^i(s_h^1, a_h^1, \dots, s_h^N, a_h^N) \right]$

Mean-Field Games ($N \rightarrow \infty$, exact symmetry $P^i \equiv P, R^i \equiv R$)

Finite-Horizon MFGs

- ▶ Infinite agents, horizon H , finite state/action sets \mathcal{S}, \mathcal{A}
- ▶ Population distribution: $\mu = \{\mu_h\}_{h=0}^{H-1} \in \Delta(\mathcal{S} \times \mathcal{A})^H$
- ▶ Dynamics: $s_0 \sim \mu_0$,
 $s_{h+1} \sim P(\cdot | s_h, a_h, \mu_h), a_h \sim \pi_h(\cdot | s_h)$
- ▶ Reward: $r_h = R(s_h, a_h, \mu_h)$



Policy π^* is an **MFG equilibrium** if

$$V(\pi^*, \mu^{\pi^*}) = \max_{\pi} V(\pi, \mu^{\pi^*}).$$

- ▶ Markov policy: $\pi = \{\pi_h\}_{h=0}^{H-1}$
- ▶ Mean field: $\mu^{\pi} = \{\mu_h\}_{h=0}^{H-1}$ is the population evolution induced by π that $\mu_{h+1} = \Gamma_P(\mu_h, \pi)$
- ▶ Value function: $V(\pi, \mu) := \mathbb{E} \left[\sum_{h=0}^{H-1} R(s_h, a_h, \mu_h) \right]$

Mean-Field Games Extension (for arbitrary N agents)

Every DG induces a MFG by **symmetrization** + **Lipschitz extension** + **averaging**.

Induced MFG

Given an N -player DG $\mathcal{G} = (\mathcal{S}, \mathcal{A}, \rho_0, N, H, \{P^i\}_{i=1}^N, \{R^i\}_{i=1}^N)$, define the MFG $(\mathcal{S}, \mathcal{A}, N, H, P, R)$ with rewards and dynamics

$$P(s, a, \mu) = \frac{1}{N} \sum_{i=1}^N \text{Ext} \left(\text{Sym} \left(P^i(s, a, \cdot) \right) \right) (\mu),$$
$$R(s, a, \mu) = \frac{1}{N} \sum_{i=1}^N \text{Ext} \left(\text{Sym} \left(R^i(s, a, \cdot) \right) \right) (\mu).$$

- ▶ **Symmetrization:** Define $\text{Sym}(f) : \mathcal{X}^K \rightarrow \mathbb{R}^D$ as $\text{Sym}(f)(\mathbf{x}) = \frac{1}{K!} \sum_{g:\text{permutation}} f(g(\mathbf{x}))$;
- ▶ **Lipschitz extension:** Given an L -Lipschitz function $f : U \rightarrow V$, for $U \subset \Delta(\mathcal{S} \times \mathcal{A})$, there exists an L -Lipschitz extension $\text{Ext}(f) : \Delta(\mathcal{S} \times \mathcal{A}) \rightarrow V$, which agrees with f on U .

Roadmap: Fundamental Questions

I. Approximation

How good is the MFG approximation to the N-agent RL?



II. Tractability

What are the computational and statistical complexities of solving MFGs?

III. Independent Learning

Can N agents learn independently to achieve the desired equilibrium?

I. Approximation Guarantees

How good is the MFG approximation to the N -agent game?

- ▶ Exploitability: $\mathcal{E}_N(\boldsymbol{\pi}^* =: [\pi_{\text{MFG}}^*, \dots, \pi_{\text{MFG}}^*]) = \max_{i=1, \dots, N} [\max_{\pi^i} V^i(\pi^i, \boldsymbol{\pi}^{*, -i}) - V^i(\boldsymbol{\pi}^*)]$
- ▶ Existing results only established asymptotic results in the stationary setting under some symmetry assumptions:
 - Homogeneous agents: [Saldi, Basar, and Raginsky 2018; Anahtarci, Kariksiz, and Saldi 2022]
 - Multi-population MFGs: [Subramanian et al. 2020; Pérolat et al. 2022]
 - Graphon MFGs: [Parise and Ozdaglar 2019; Caines and Huang 2019]
- ▶ Our focus: **finite N , finite H , heterogeneous agents**

I. Approximation Guarantees

Theorem (Yardim & H.'24)

Let π_{MFG}^* be an MFG equilibrium of the induced MFG. Then the N -agent exploitability under the policy $\pi^* = (\pi_{\text{MFG}}^*, \dots, \pi_{\text{MFG}}^*)$ satisfies:

$$\mathcal{E}_N(\pi^*) \leq \mathcal{O} \left(\frac{H^2(1 - L^H)|\mathcal{S}||\mathcal{A}|}{(1 - L)\sqrt{N}} + \alpha \frac{H^2(1 - L^H)}{(1 - L)} + \beta H \right).$$

L : Lipschitz constant of population flow operator:

$$\Gamma_P(\mu, \pi)(s', a') := \sum_{s \in \mathcal{S}, a \in \mathcal{A}} \mu(s, a) P(s' | s, a, \mu) \pi(a' | s').$$

α : degree of dynamic heterogeneity among agents:

$$\max_{\substack{i \in [N], s, a \in \mathcal{S} \times \mathcal{A} \\ \mu \in \Delta_{\mathcal{S} \times \mathcal{A}}}} \max_{\substack{\boldsymbol{\rho} \in (\mathcal{S} \times \mathcal{A})^{N-1} \\ \sigma(\boldsymbol{\rho}) = \mu}} \|P^i(s, a, \boldsymbol{\rho}) - P(s, a, \mu)\|_1$$

β : degree of reward heterogeneity among agents.

Takeaway: MFG approximation is good for stable dynamics and is robust to small violations of symmetry.

I. Approximation Guarantees (cont'd)

- ▶ **Exact symmetry:** If $P^1 = P^2 = \dots = P^N, R^1 = R^2 = \dots = R^N$, we have $\alpha = \beta = 0$ and this yields the first non-asymptotic bound of MFG approximation: $\mathcal{E}_N(\pi^*) \leq \mathcal{O}\left(\frac{L^H}{\sqrt{N}}\right)$.
- ▶ **Lower Bound:** There exists $\mathcal{S}, \mathcal{A}, P, R$ such that for all $H, N > 0$, the FH-MFG has unique NE π_H^* , and if every agent in the DG plays π_H^* , they suffer from exploitability $\mathcal{E}^N(\pi_H^*, \dots, \pi_H^*) \geq \Omega(H)$, unless $N \geq \Omega(2^H)$.
- ▶ **Infinite-horizon stationary MFG:** If population evolution operator Γ_P is non-expansive, i.e. $L \leq 1$, we can obtain approximation bound: $\mathcal{O}\left(\frac{(1-\gamma)^{-3}}{\sqrt{N}}\right)$, where γ is the discount factor. Without the non-expansiveness assumption, there exists N -agent dynamic game where the mean-field solution suffers from an exploitability of order $\Omega(N^{-\log_2 \gamma^{-1}})$.

B. Yardim, A. Goldman, and N. He. When is Mean-Field Reinforcement Learning Tractable and Relevant? AAMAS 2024.

B. Yardim and N. He. Exploiting Approximate Symmetry in Dynamic Games for Efficient Multi-Agent Reinforcement Learning. 2024.

Roadmap: Fundamental Questions

I. Approximation

How good is the MFG approximation to the N-agent RL?



II. Tractability

What are the computational and statistical complexities of solving MFGs?

III. Independent Learning

Can N agents learn independently to achieve the desired equilibrium?

II.1. Computational Tractability

Is it computationally easier to solve MFGs than N -agent games (assuming known models)?

	MFG Type	Key Assumptions	Complexity
[Guo et al. 2019] [Anahtarci et al. 2022] [Cui and Koepl 2021]	Stat-MFG	Lipschitz P, R + Regularization + Contractive Γ_P	$\mathcal{O}(\log \varepsilon^{-1})$ single-agent RL
[Zhang et al. 2024]	FH-MFG	Monotone R, μ -independent P	$\mathcal{O}(\varepsilon^{-2})$ policy eval.
Our results	Stat-MFG FH-MFG FH-MFG	Lipschitz P, R Lipschitz $R + \mu$ -independent P Linear $R + \mu$ -independent P	PPAD-complete PPAD-complete PPAD-complete

II.2. Statistical Tractability

Putting computation aside, is it statistically tractable to learn MFGs (**with unknown model**) in general?

- ▶ **Sample efficiency**: how many samples are sufficient to learn an ϵ -NE given a hypothesis class?
- ▶ Rich theory for single-agent RL since Kearns and Singh (1998), for example:

Finite-horizon MDPs	Sample efficiency	Representative work
Tabular	$\text{poly}(\mathcal{S} , \mathcal{A} , H, \frac{1}{\epsilon})$	UCBVI [Azar, Osband, and Munos 2017] Q-learning with UCB [Jin et al. 2018]
Function approximation	$\text{poly}(\text{comp}(\mathcal{F}), H, \frac{1}{\epsilon})$	Eluder Dimension [Russo and Van Roy 2013] Bellman Rank [Jiang et al. 2017] Bellman Eluder Dimension [Jin, Liu, and Miryoosefi 2021] ...

II.2. Statistical Tractability

Theorem (Huang, H., Krause'24)

Given a model class \mathcal{M} with $|\mathcal{M}| < +\infty$, assume that

- (i) Realizability: $M^* \in \mathcal{M}$
- (ii) Lipschitz continuity in density: P, R are L_P, L_R -Lipschitz in μ .

There exists an iterative model-elimination-based algorithm that requires at most

$$\tilde{O}(\text{poly}(\text{dimPE}(\mathcal{M}), H, 1 + L_P, 1 + L_R, \frac{1}{\epsilon}, \log \frac{|\mathcal{M}|}{\delta}))$$

samples to return an ϵ -Nash when learning the FH-MFG with probability at least $1 - \delta$.

- ▶ **dimPE(\mathcal{M})** is a new notion of “Model-Based Eluder Dimension”, which is small for many cases:
 - Tabular MFG: $\text{dimPE}(\mathcal{M}) \leq |S| \cdot |\mathcal{A}|$
 - Linear MFG: assume $P(s'|s, a, \mu) = \phi(s, a)^\top U(\mu) \psi(s')$ with $\phi(s, a) \in \mathbb{R}^d$, then $\text{dimPE}(\mathcal{M}) \leq d$.
- ▶ **Takeaway:** Learning MFG is statistically tractable under minimal structural assumptions.

Roadmap: Fundamental Questions

I. Approximation

How good is the MFG approximation to the N-agent RL?



II. Tractability

What are the computational and statistical complexities of solving MFGs?

III. Independent Learning

Can N agents learn independently to achieve the desired equilibrium?

III. Independent Learning

Can N agents learn independently to achieve the desired Nash equilibrium?

	No population manipulation	Single path	N -agent simulator	Independent learning
Guo et al., 2019	No	No	No	No
Anahtarci et al., 2022	No	No	No	No
Subramanian et al., 2019	No	No	No	No
Xie et al., 2021	No	Yes	No	No
Zaman et al, 2022	No	Yes	No	No
Decentralized PMD	Yes	Yes	Yes	Yes

Table 1. Summary of existing RL approaches for learning stationary MFGs.

III. Independent Learning

- ▶ **Goal:** Learning from samples by **simulating the N -agent game**.
- ▶ **Main idea:** Repeat the following iteration.

1. **Estimate Q -values:** Each agent, keeping their policies fixed, performs **TD learning** for a given number of steps;
2. **Policy update:** Agents simultaneously perform **policy mirror ascent** update using Q -value estimates:

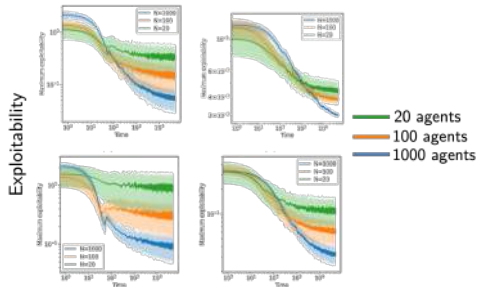
$$\text{(PMD) : } \pi(s) \leftarrow \arg \max_{u \in \Delta_{\mathcal{A}}} \langle u, Q^{\pi}(s, \cdot) \rangle + \underbrace{\tau \cdot h(u)}_{\text{regularizer}} - \frac{1}{\eta_t} \underbrace{D(u; \pi(s))}_{\text{Bregman divergence}} .$$

III. Independent Learning

Theorem (informal)

Under additional structural assumptions, N agents running (regularized) Policy Mirror Descent independently converges ϵ -close to (regularized) Nash equilibrium, requiring only $\text{poly}(1/\epsilon)$ sample complexity, up to the approximation error.

- ▶ Theoretical guarantees hold for
 - Stationary MFGs with regularization [YCGH'23]
 - Stateless MFGs with monotone rewards [YCH'23]
 - FH-MFGs with monotone rewards and μ -independent dynamics [YH'24]
- ▶ **Note:** The abstraction of MFG is used merely as a proof technique, not in the algorithm.

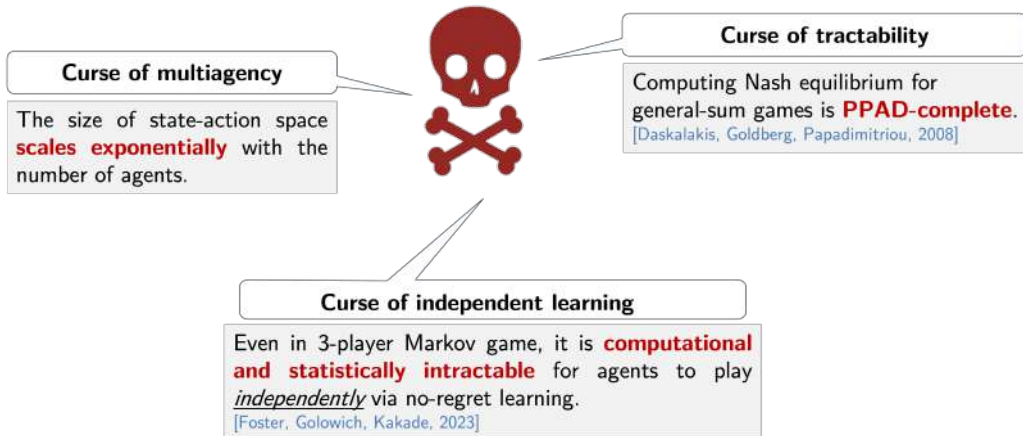


B. Yardim, S. Cayci, M. Geist, and N. He. Policy Mirror Ascent for Efficient and Independent Learning in Mean-Field Games. ICML 2023.

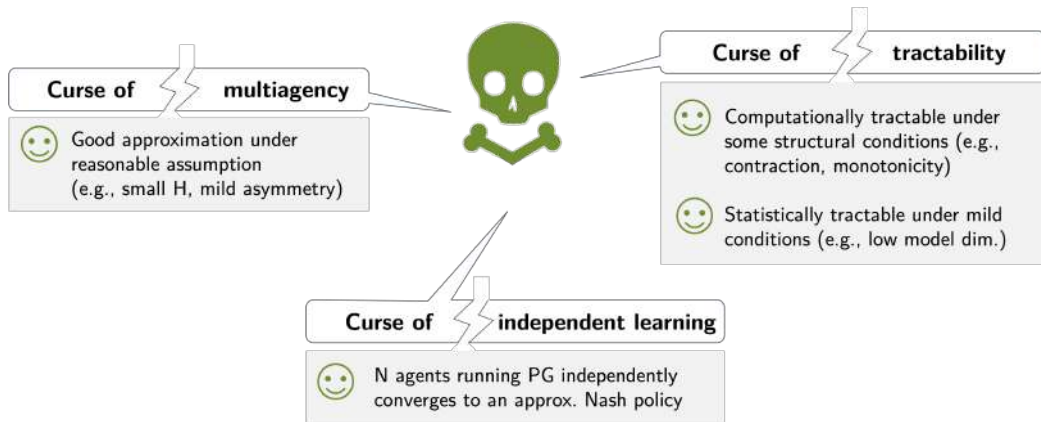
B. Yardim, S. Cayci, and N. He. Stateless Mean-Field Games: A Framework for Independent Learning with Large Populations. EWRL 2023.

B. Yardim and N. He. Exploiting Approximate Symmetry in Dynamic Games for Efficient Multi-Agent Reinforcement Learning. AAMAS Workshop, 2024.

Revisit: Curses of Many-agent RL



Summary: MFG is a promising solution for many-agent RL!



Open Questions: computational-statistical gaps? beyond benign asymmetry? beyond MFG approx.?

Collaborators and Related Papers



Batuhan Yardim
(ETH Zurich)



Jiawei Huang
(ETH Zurich)



Semih Cayci
(RWTH Aachen)



Matthieu Geist
(Cohere, ex Google)



Andreas Krause
(ETH Zurich)



- YCGH'23** B. Yardim , S. Cayci, M. Geist, and N. He. Policy Mirror Ascent for Efficient and Independent Learning in Mean-Field Games. ICML 2023.
- HYH'24** J. Huang, B. Yardim, and N. He. On the Statistical Efficiency of Mean-Field RL with General Function Approximation. AISTATS 2024.
- YGH'24** B. Yardim, A. Goldman, and N. He. When is Mean-Field Reinforcement Learning Tractable and Relevant? International Conference on Autonomous Agents and Multiagent Systems (AAMAS), 2024.
- HHK'24** J. Huang, N. He, and A. Krause. Model-Based RL for Mean-Field Games is not Statistically Harder than Single-Agent RL. ICML 2024.
- YH'24** B. Yardim and N. He. Exploiting Approximate Symmetry in Dynamic Games for Efficient Multi-Agent Reinforcement Learning. AAMAS Workshop on Optimization and Learning in Multiagent Systems, 2024.

Thank You!

References I

-  Anahtarci, Berkay, Can Deha Kariksiz, and Naci Saldi (2022). “Q-learning in regularized mean-field games”. In: [Dynamic Games and Applications](#), pp. 1–29.
-  Caines, Peter E and Minyi Huang (2019). “Graphon mean field games and the GMFG equations: ε -Nash equilibria”. In: [2019 IEEE 58th conference on decision and control \(CDC\)](#). IEEE, pp. 286–292.
-  Cui, Kai and Heinz Koepl (2021). “Approximately solving mean field games via entropy-regularized deep reinforcement learning”. In: [International Conference on Artificial Intelligence and Statistics](#). PMLR, pp. 1909–1917.
-  Guo, Xin et al. (2019). “Learning mean-field games”. In: [Advances in Neural Information Processing Systems](#) 32.
-  Parise, Francesca and Asuman Ozdaglar (2019). “Graphon games”. In: [Proceedings of the 2019 ACM Conference on Economics and Computation](#), pp. 457–458.
-  Pérolat, Julien et al. (2022). “Scaling Mean Field Games by Online Mirror Descent”. In: [Proceedings of the 21st International Conference on Autonomous Agents and Multiagent Systems](#), pp. 1028–1037.
-  Saldi, Naci, Tamer Basar, and Maxim Raginsky (2018). “Markov–Nash equilibria in mean-field games with discounted cost”. In: [SIAM Journal on Control and Optimization](#) 56.6, pp. 4256–4287.

References II

-  Subramanian, Sriram Ganapathi et al. (2020). “Multi Type Mean Field Reinforcement Learning”. In: Proceedings of the 19th International Conference on Autonomous Agents and MultiAgent Systems. AAMAS '20. Auckland, New Zealand: International Foundation for Autonomous Agents and Multiagent Systems, pp. 411–419. ISBN: 9781450375184.
-  Zhang, Fengzhuo et al. (2024). “Learning Regularized Monotone Graphon Mean-Field Games”. In: Advances in Neural Information Processing Systems 36.