Foundations of Reinforcement Learning

Multi-agent RL: sample complexity

Yuejie Chi

Department of Electrical and Computer Engineering

Carnegie Mellon University

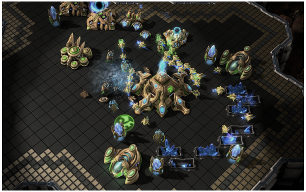
Spring 2023

Outline

Background: finite-horizon two-player zero-sum Markov games

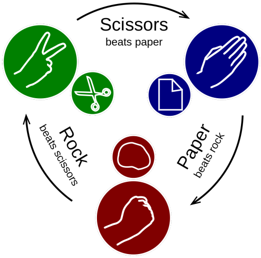
Statistical perspective: sample complexity

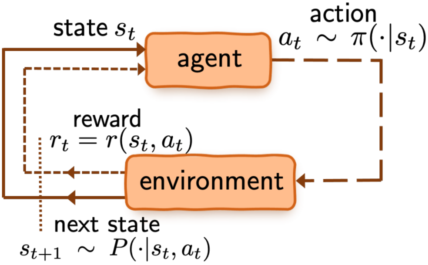
Multi-agent reinforcement learning (MARL)



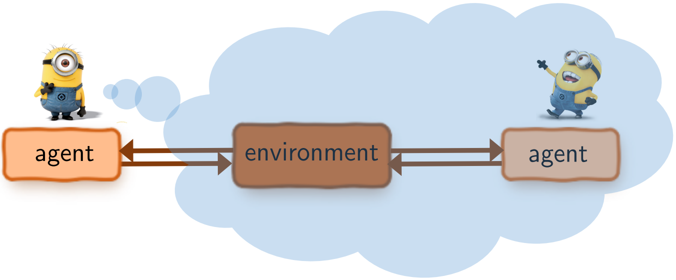
To collaborate or to compete, that is the question.

MARL = Game theory + RL



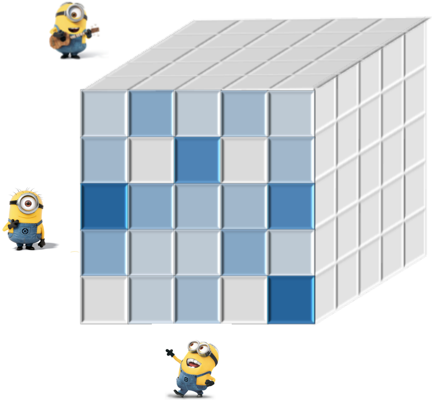


Challenges in MARL: nonstationarity



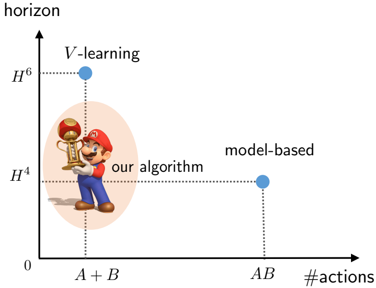
From a single-agent perspective:
the environment is **time-varying** and **nonstationary**!

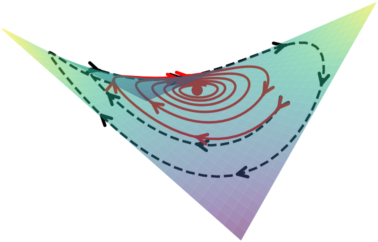
Challenges in MARL: curse of multiple agents



The explosion of choices:
The joint action space grows **exponentially** with the agents!

Two-player zero-sum Markov games



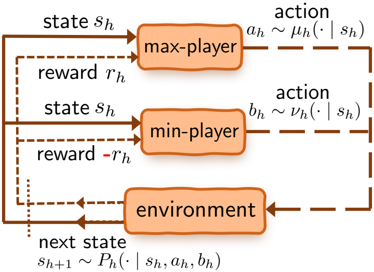


Statistical perspective:
this lecture

Optimization perspective:
next lecture

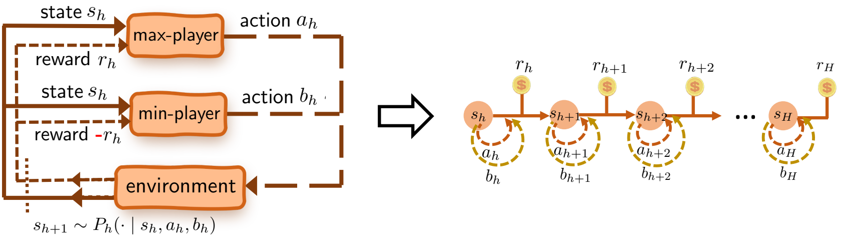
Background: finite-horizon two-player zero-sum Markov games

Two-player zero-sum Markov games (finite-horizon)



- $\mathcal{S}$: shared state space
- H: horizon
- $\mathcal{A} = [A]$: action space of max-player
- $\mathcal{B} = [B]$: action space of min-player
- immediate reward: max-player $r_h(s,a,b) \in [0,1]$
  min-player $-r_h(s,a,b)$
- $\mu = \{\mu_h\}$: policy of max-player; $\nu = \{\nu_h\}$: policy of min-player
- $P_h(\cdot \mid s,a,b)$: unknown transition probabilities

Value function



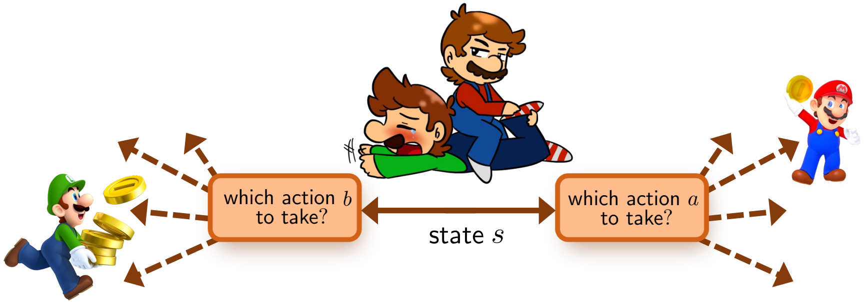
Value function of policy pair (μ, ν):

$$V_h^{\mu,\nu}(s) := \mathbb{E}\left[\sum_{t=h}^{H} r_t(s_t, a_t, b_t) \,\middle|\, s_h = s\right]$$

$$Q_h^{\mu,\nu}(s, a, b) := \mathbb{E}\left[\sum_{t=h}^{H} r_t(s_t, a_t, b_t) \,\middle|\, s_t = s, a_t = a, b_t = b\right]$$

- $\{(a_t, b_t, s_{t+1})\}$: generated when max-player and min-player execute policies μ and ν *independently (i.e. no coordination)*

Optimal policy?



- Each agent seeks **optimal policy** maximizing her own value
- But two agents have conflicting goals ...

Compromise: Nash equilibrium (NE)

John von Neumann

John Nash

An NE policy pair $(\mu^\star, \nu^\star)$ obeys

$$\max_{\mu} V^{\mu,\nu^\star} = V^{\mu^\star,\nu^\star} = \min_{\nu} V^{\mu^\star,\nu}$$

- no unilateral deviation is beneficial
- no coordination between two agents (they act *independently*)

Nash value iteration (finite-horizon)

Nash value iteration: for $h = H, \ldots, 1$

$$Q_h(s,a,b) \longleftarrow r_h(s,a,b) + \mathop{\mathbb{E}}_{s' \sim P_h(\cdot|s,a,b)} \Bigg[\underbrace{\max_{\mu(s)} \min_{\nu(s)} \mu(s')^\top Q_{h+1}(s') \nu(s')}_{\text{matrix game}} \Bigg],$$

where $Q_h(s) = [Q_h(s,\cdot,\cdot)] \in \mathbb{R}^{A \times B}$.

- The matrix game can be solved efficiently (see next lecture).
- Requires knowledge of the transition kernel $P_h(\cdot|s,a,b)$.

How do we learn the NE without access to the model?

Aside: infinite-horizon discounted setting

Value function of policy pair (μ, ν):

$$V^{\mu,\nu}(s) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_t(s_t, a_t, b_t) \,\Big|\, s_0 = s\right]$$

$$Q^{\mu,\nu}(s,a,b) := \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t, b_t) \,\Big|\, s_0 = s, a_0 = a, b_0 = b\right]$$

where $\gamma \in [0,1)$ is the discount factor.

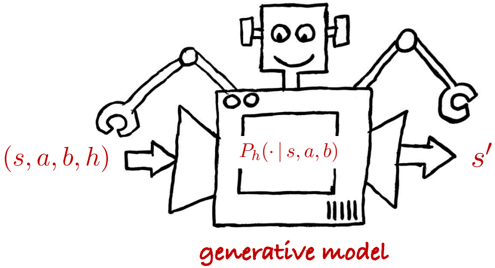
Nash value iteration:

$$Q(s,a,b) \longleftarrow r_h(s,a,b) + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a,b)}\Bigg[\underbrace{\max_{\mu(s)} \min_{\nu(s)} \mu(s')^\top Q(s') \nu(s')}_{\text{matrix game}}\Bigg],$$

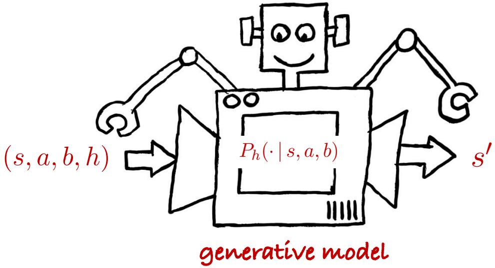
where $Q(s) = [Q(s,\cdot,\cdot)] \in \mathbb{R}^{A \times B}$.

Statistical perspective: sample complexity

A generative model / simulator



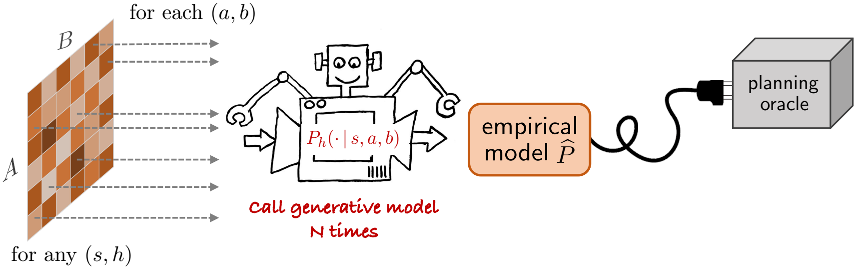
One can query generative model w/ state-action-step tuple (s, a, b, h), and obtain $s' \overset{\text{ind.}}{\sim} P_h(s' \mid s, a, b)$



Question: *how many samples are sufficient to learn an* $\underbrace{\varepsilon\text{-Nash policy pair}}_{\max_\mu V^{\mu,\hat{\nu}} - \varepsilon \leq V^{\hat{\mu},\hat{\nu}} \leq \min_\nu V^{\hat{\mu},\nu} + \varepsilon}$ *?*

Model-based approach (non-adaptive sampling)

— [Zhang et al., 2020]

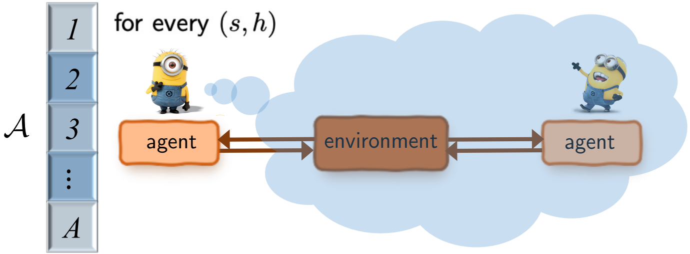


1. for each (s, a, b, h), call generative models N times
2. build empirical model $\widehat{P}$, and run "plug-in" methods

sample complexity: $\frac{H^4 SAB}{\varepsilon^2}$ — **curse of multiagents!**

Breaking the curse of multi-agents?

— [Jin et al., 2021, Song et al., 2021, Mao and Başar, 2022]

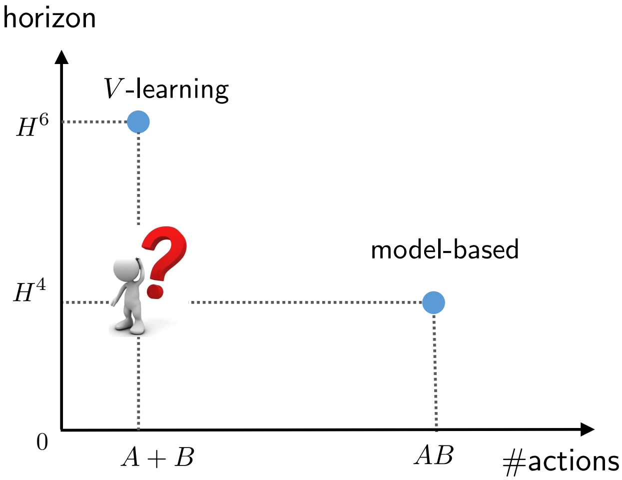


V-learning (online setting): MARL meets adversarial learning: for the max-player, for $h = 1, \ldots, H$

1. *adaptive sampling:* sampling $\mathcal{A}$ based on $\mu_h(\cdot|s)$
2. estimate V-function only with *Hoeffding bonus* (of size S)
3. update policy via *adversarial bandit learning subroutine*

sample complexity: $\frac{H^6 S(A+B)}{\varepsilon^2}$

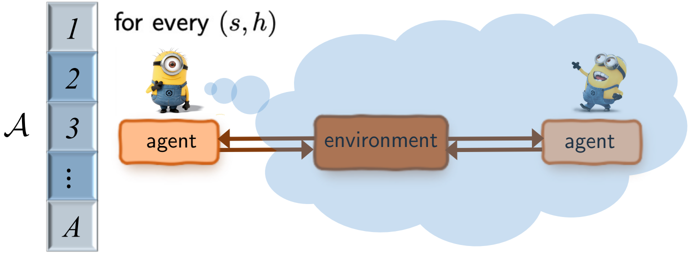
Summary so far



Can we simultaneously overcome curse of multi-agents & barrier of long horizon?

Improved algorithm (with a generative model)

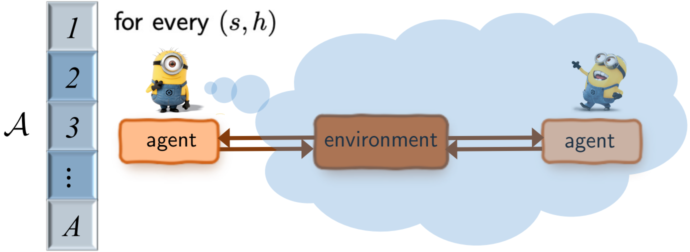
— [Li et al., 2022]



Nash-Q-FTRL: for the max-player, for $h = H, \ldots, 1$

- collect $k = 1, \ldots, K$ samples:
    1. *adaptive sampling:* sample $\mathcal{A}$ based on $\mu_h^k(\cdot|s)$
    2. estimate single-agent Q-function $Q_h(s, \cdot)$ via Q-learning
    3. update policy $\mu_h^{k+1}(\cdot|s)$ via adversarial bandit learning subroutine
- output a Markov policy μ_h and V_h with Bernstein bonuses

Single-side estimate via adaptive sampling



One-sided Q-function estimation via adaptive sampling

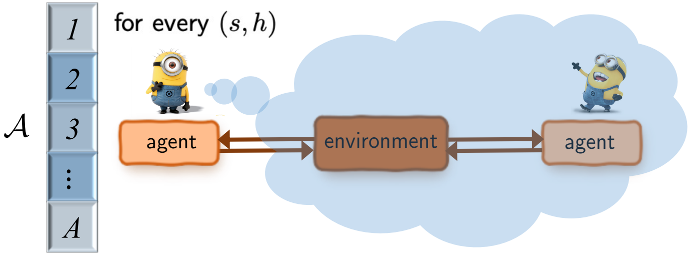
- e.g. $Q(s, a)$ as opposed to $Q(s, a, b)$
- draw an independent sample based on current policy iterates:

$$b_{h,s,a} \sim \nu_h(\cdot|s), \qquad s'_{h,s,a} \sim P_h(s, a, b_{h,s,a})$$

  instead of sampling over all $b \in \mathcal{B}$.

- update the one-sided Q-function via the Q-learning update rule

Adversarial learning via FTRL



Policy update via adversarial learning routine

- Given the one-sided Q-estimate $Q_h^k(s,a)$, update the policy via `Follow-the-Regularized-Leader (FTRL)` (with entropy regularization):

$$\mu_h^{k+1} = \arg\max_{\pi} \left\{ \left\langle \pi, Q_h^k(s,a) \right\rangle + \frac{1}{\eta_{k+1}} \mathcal{H}(\pi) \right\} \propto \exp\left(\eta_{k+1} Q_h^k(s,a)\right)$$

This is exponential weight update.

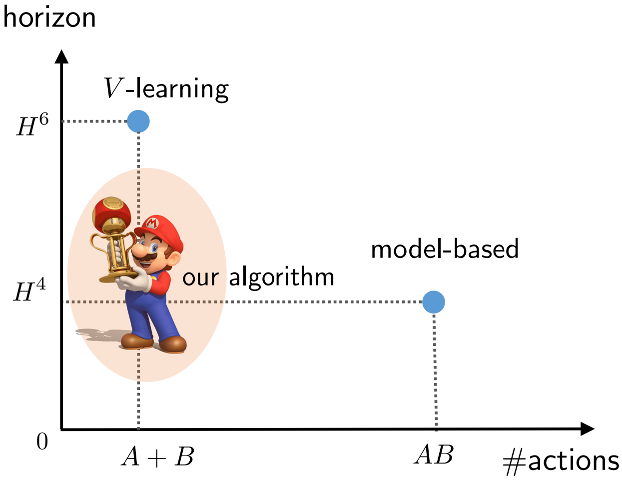
Main result: two-player zero-sum Markov games

Theorem 1 ([Li et al., 2022])

For any $0 < \varepsilon \leq H$, the policy pair $(\widehat{\mu}, \widehat{\nu})$ returned by Nash-Q-FTRL is ε-Nash, with sample complexity at most

$$\widetilde{O}\left(\frac{H^4 S(A+B)}{\varepsilon^2}\right).$$

- **minimax lower bound:** $\widetilde{\Omega}\left(\frac{H^4 S(A+B)}{\varepsilon^2}\right)$
- breaks curse of multi-agents & long-horizon barrier at once!
- full ε-range (no burn-in cost)
- other features: Markov policy, decentralized, . . .



Nash-Q-FTRL breaks curses of multi-agents and long-horizon barrier simultaneously!

Extension: multi-player general-sum Markov games

- Learning NE in general-sum games is computationally infeasible (i.e., PPAD-complete)
- Instead, focusing on learning the *coarse correlated equilibrium (CCE)*. A joint policy π is said to be a CCE if

$$V_{i,1}^{\pi}(s) \geq V_{i,1}^{\star,\pi_{-i}}(s), \qquad \text{for all } (s,i) \in \mathcal{S} \times [m].$$

- A key distinction from the definition of NE lies in the fact that it allows the policy to be correlated across the players.

Extension: multi-player general-sum Markov games

Theorem 2 ([Li et al., 2022])

For any $0 < \varepsilon \leq H$, the joint policy $\widehat{\pi}$ returned by the proposed algorithm is ε-CCE, with sample complexity at most

$$\widetilde{O}\left(\frac{H^4 S \sum_i A_i}{\varepsilon^2}\right)$$

- **minimax lower bound:**

$$\widetilde{\Omega}\left(\frac{H^4 S \max_i A_i}{\varepsilon^2}\right)$$

- near-optimal when the number of players m is fixed

References I

Jin, C., Liu, Q., Wang, Y., and Yu, T. (2021).
V-learning–a simple, efficient, decentralized algorithm for multiagent RL.
arXiv preprint arXiv:2110.14555.

Li, G., Chi, Y., Wei, Y., and Chen, Y. (2022).
Minimax-optimal multi-agent RL in Markov games with a generative model.
In *Advances in Neural Information Processing Systems.*

Mao, W. and Başar, T. (2022).
Provably efficient reinforcement learning in decentralized general-sum Markov games.
Dynamic Games and Applications, pages 1–22.

Song, Z., Mei, S., and Bai, Y. (2021).
When can we learn general-sum Markov games with a large number of players sample-efficiently?
arXiv preprint arXiv:2110.04184.

Zhang, K., Kakade, S., Basar, T., and Yang, L. (2020).
Model-based multi-agent RL in zero-sum Markov games with near-optimal sample complexity.
Advances in Neural Information Processing Systems, 33:1166–1178.