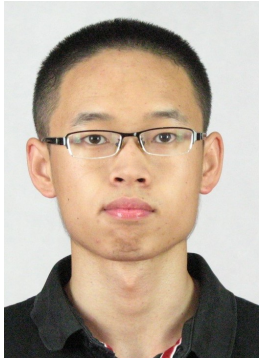
Multi-agent Reinforcement Learning: Statistical and Optimization Perspectives

Yuejie Chi

Carnegie Mellon University

Cornell University
October 2022

My wonderful collaborators



Gen Li

UPenn

Yuxin Chen

UPenn

Yuting Wei

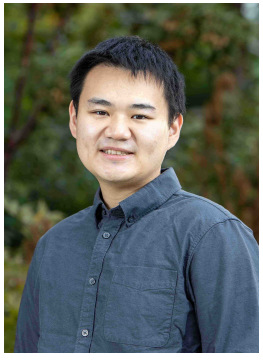
UPenn

Shicong Cen

CMU

Lin Xiao

Meta AI



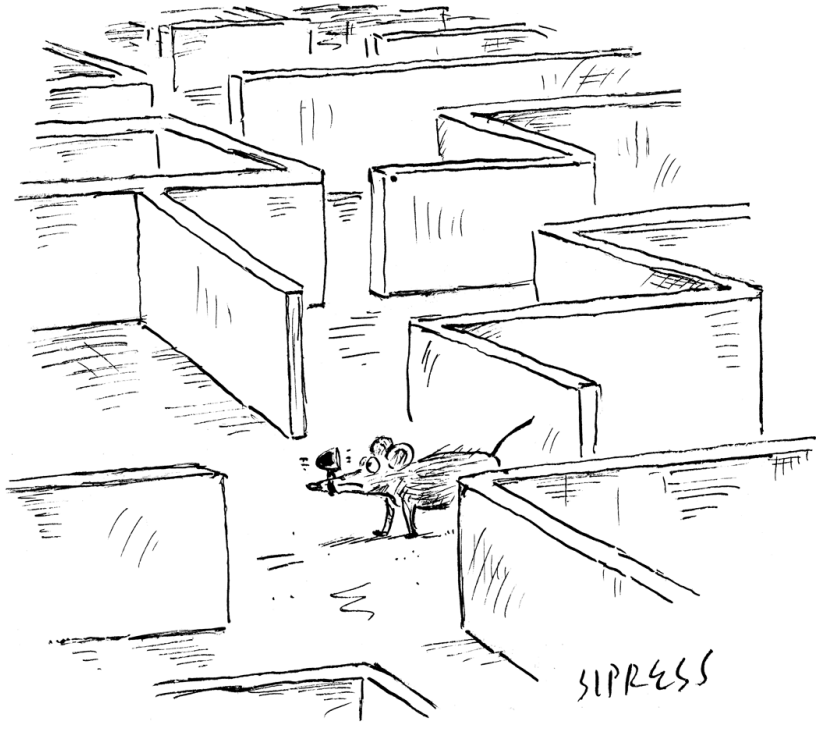
Simon Du

UW

Reinforcement learning (RL)

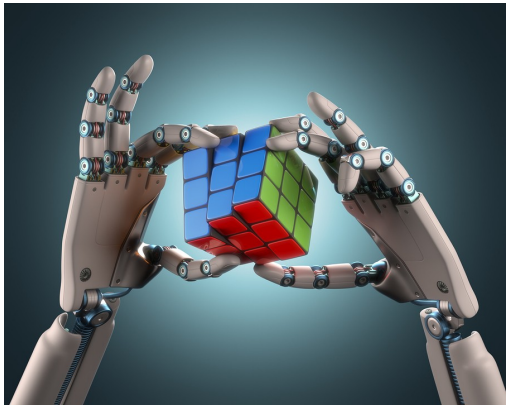
In RL, an agent learns by interacting with an environment.

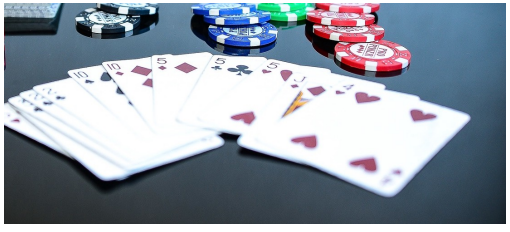
- unknown environments
- maximize total rewards
- trial-and-error
- sequential and online

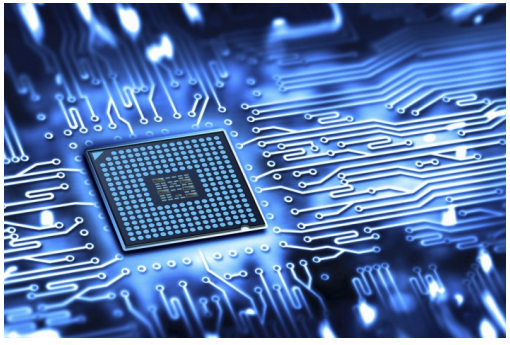


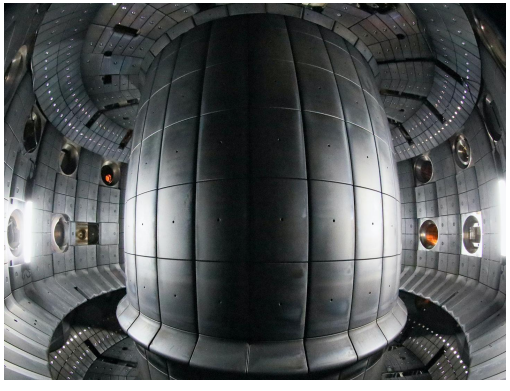
"Recalculating ... recalculating ..."

Recent successes in RL



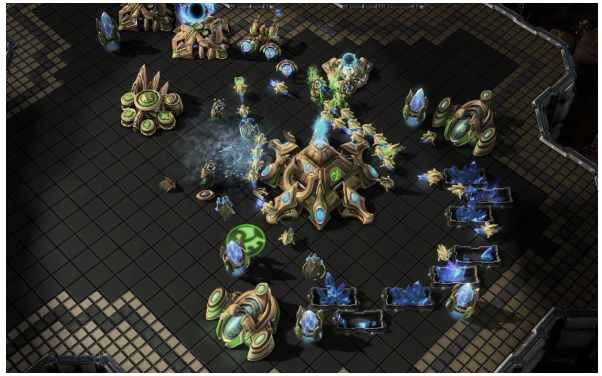






RL holds great promise in the next era of artificial intelligence.

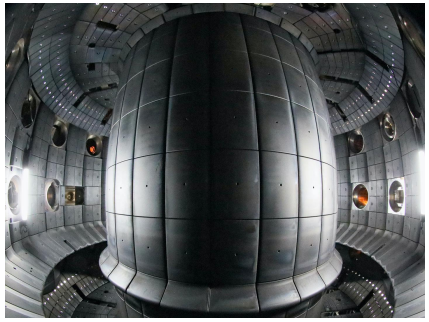
Multi-agent reinforcement learning (MARL)



To collaborate or to compete, that is the question.

Sample efficiency

Collecting data samples might be expensive or time-consuming



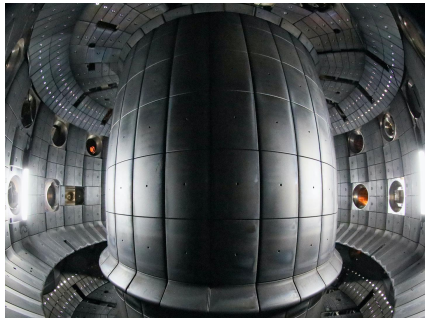
nuclear plant

autonomous driving

online ads

Sample efficiency

Collecting data samples might be expensive or time-consuming



nuclear plant

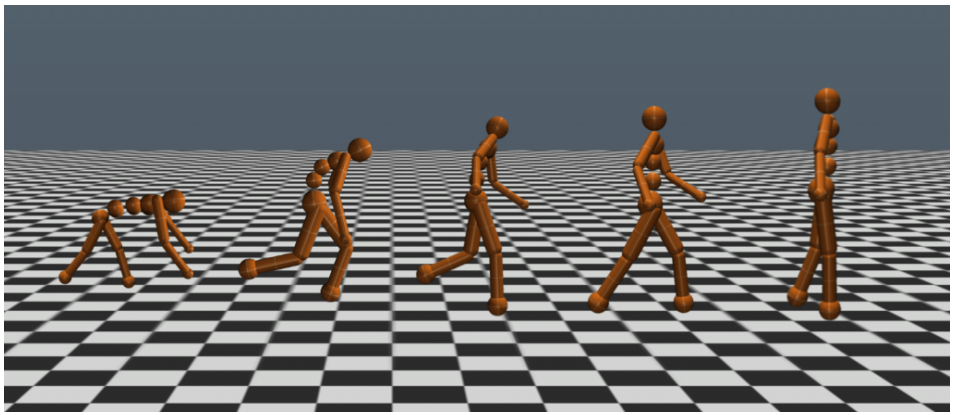
autonomous driving

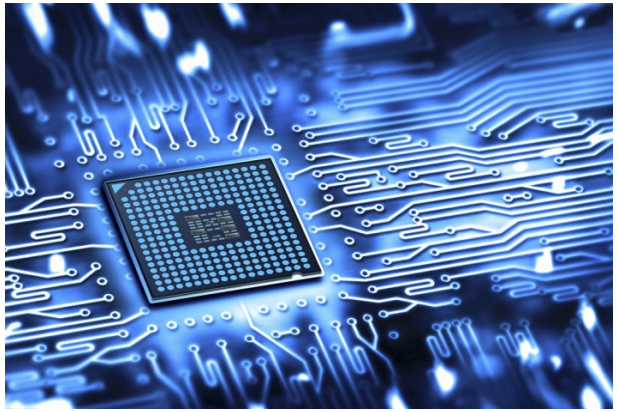
online ads

Calls for design of sample-efficient RL algorithms!

Computational efficiency

Running RL algorithms might take a long time and space

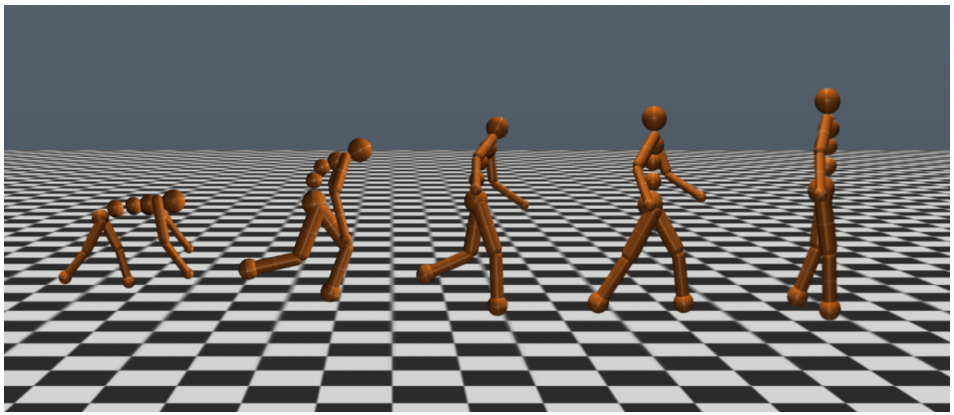


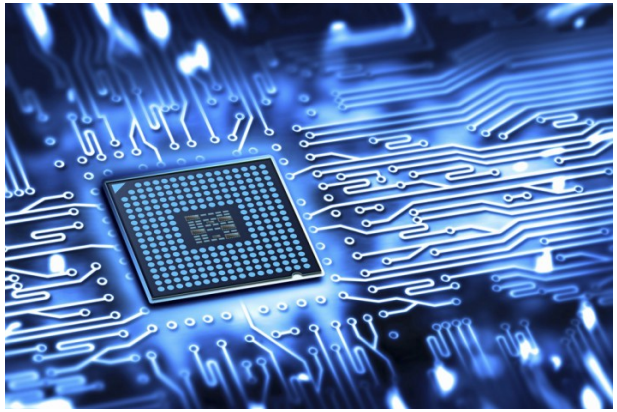


many CPUs / GPUs / TPUs + computing hours

Computational efficiency

Running RL algorithms might take a long time and space





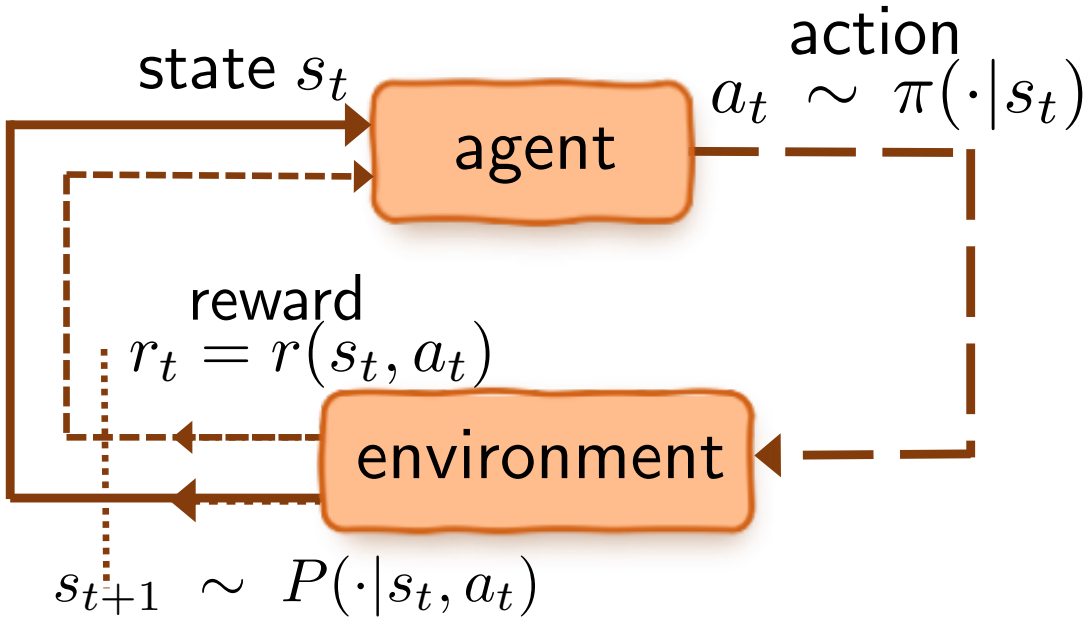
many CPUs / GPUs / TPUs + computing hours

Calls for computationally efficient RL algorithms!

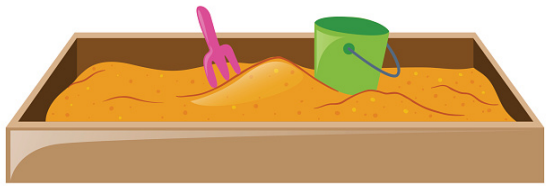
From asymptotic to non-asymptotic analyses

Non-asymptotic analyses are key to understand sample and computational efficiency in modern RL.

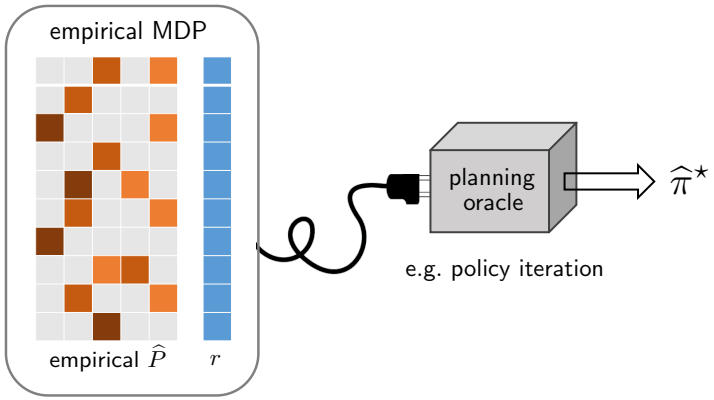
Recent advances in single-agent RL

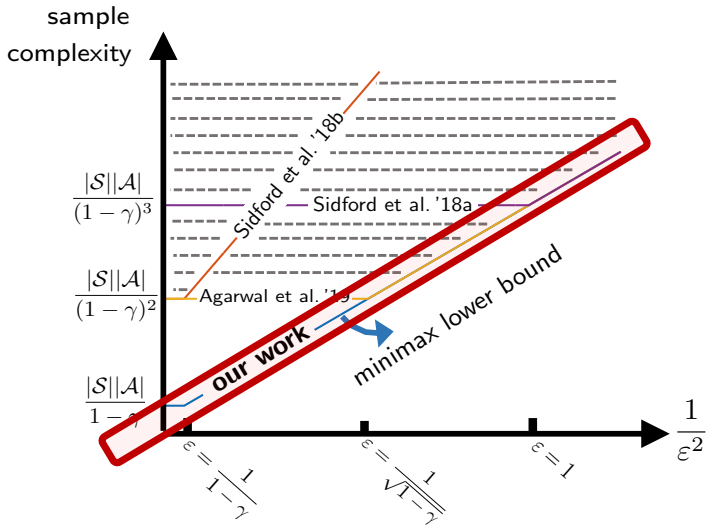


The playground: Markov decision processes



Recent advances in single-agent RL: model-based

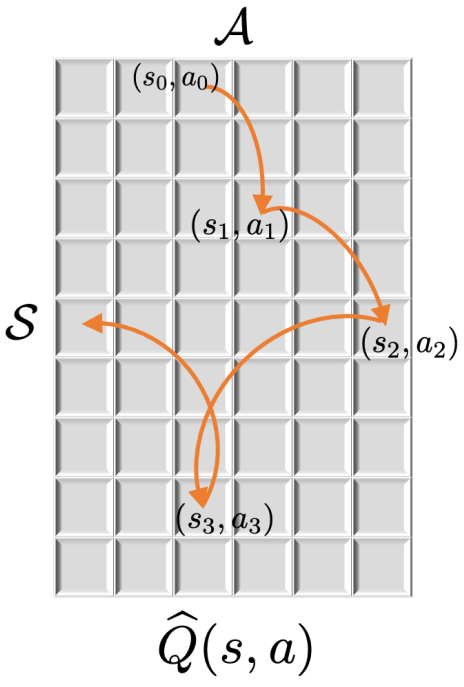


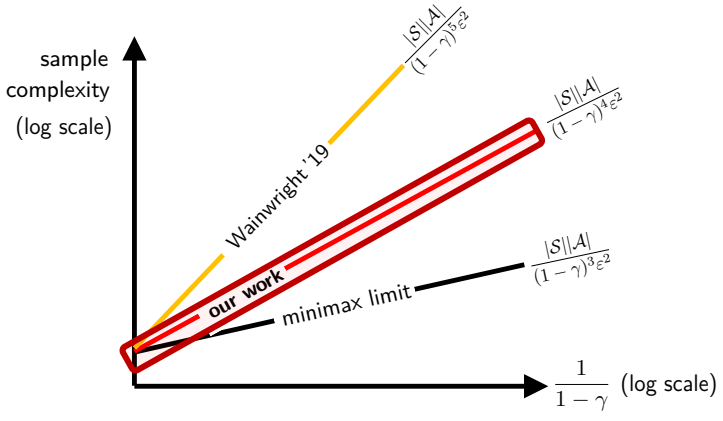


Plug-in estimators are minimax-optimal

(Sidford et al., 2018; Agarwal et al., 2019; Wang 2019; Li et al., 2020)

Recent advances in single-agent RL: value-based

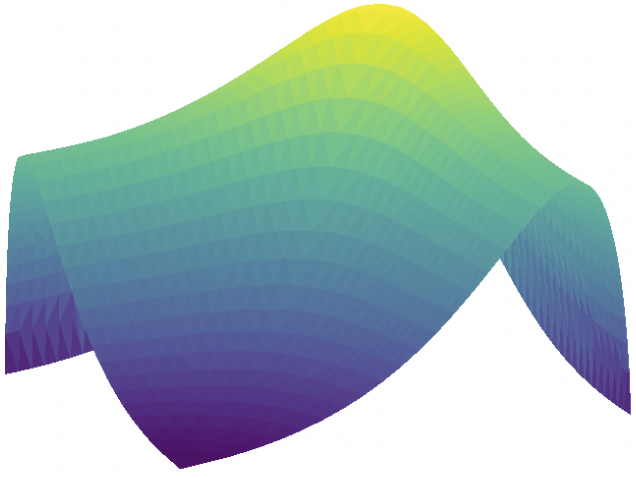


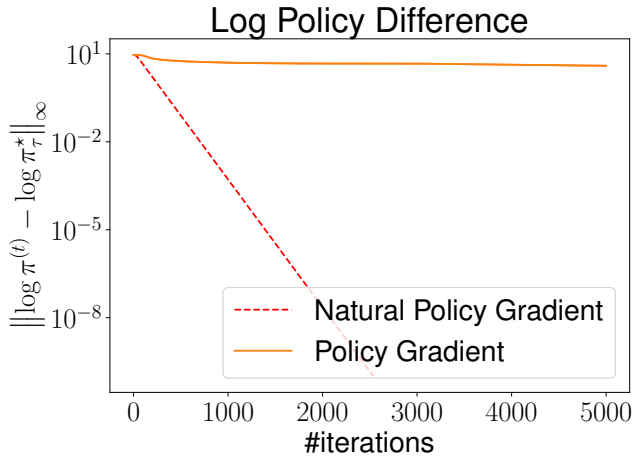


Q-learning is not minimax-optimal

(Even-Dar and Mansour, 2013; Wainwright, 2019; Chen et al., 2020; Li et al., 2021)

Recent advances in single-agent RL: policy-based



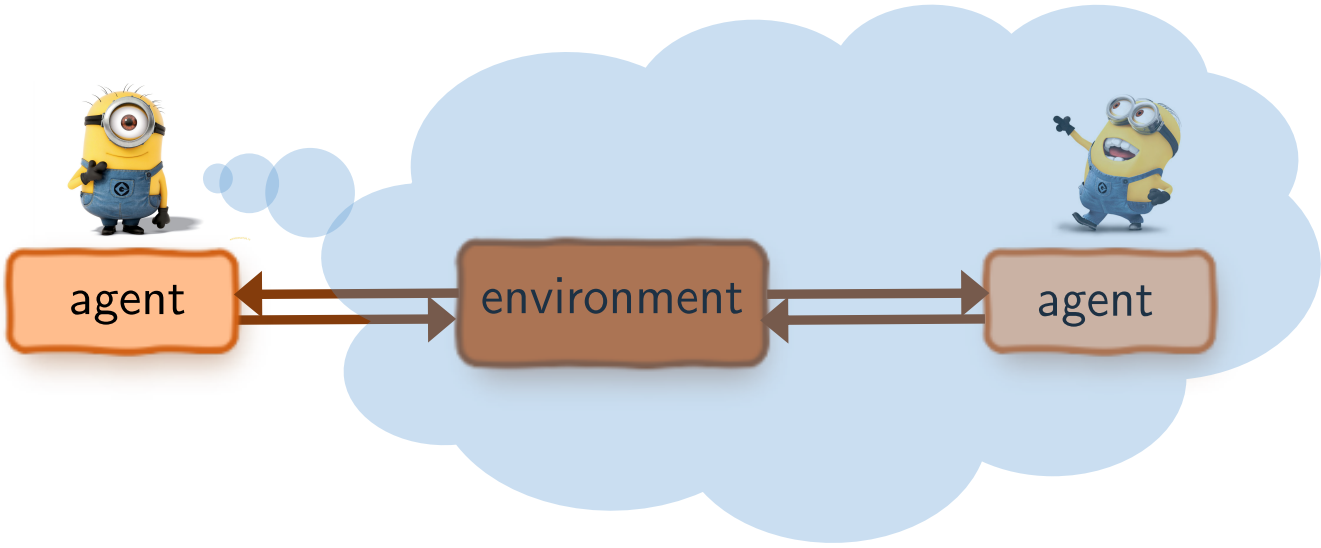


Global convergence of policy gradient methods

(Agarwal et al., 2019; Mei et al., 2020; Cen et al., 2020; Lan, 2021; Xiao, 2022)

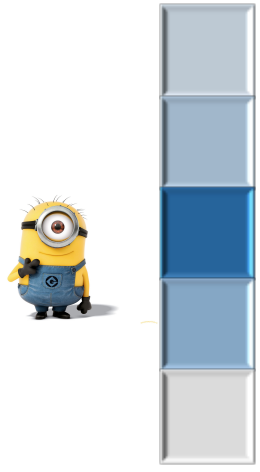
Challenges in MARL: nonstationarity

Challenges in MARL: nonstationarity

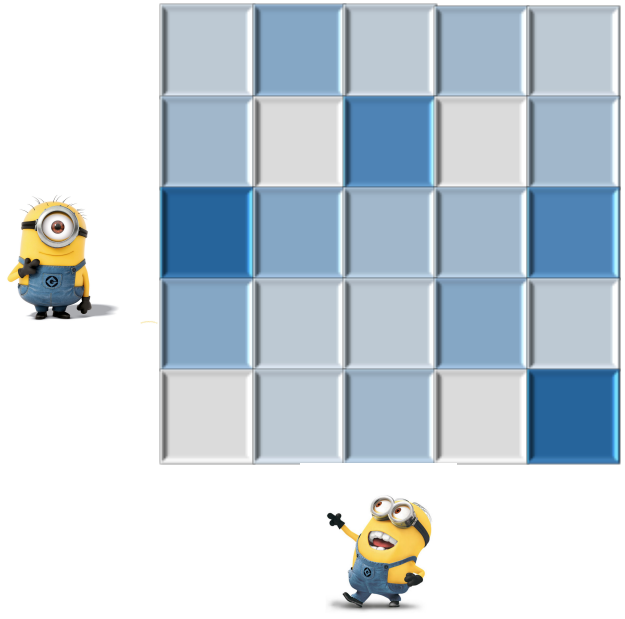


From a single-agent perspective:
the environment is **time-varying** and **nonstationary**!

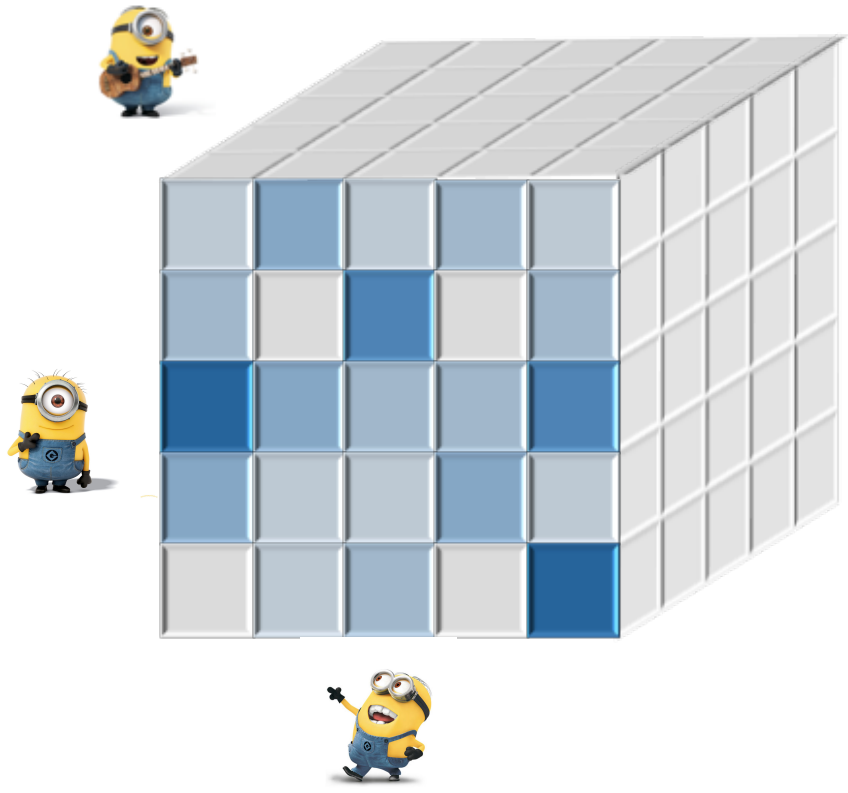
Challenges in MARL: curse of multiple agents



Challenges in MARL: curse of multiple agents

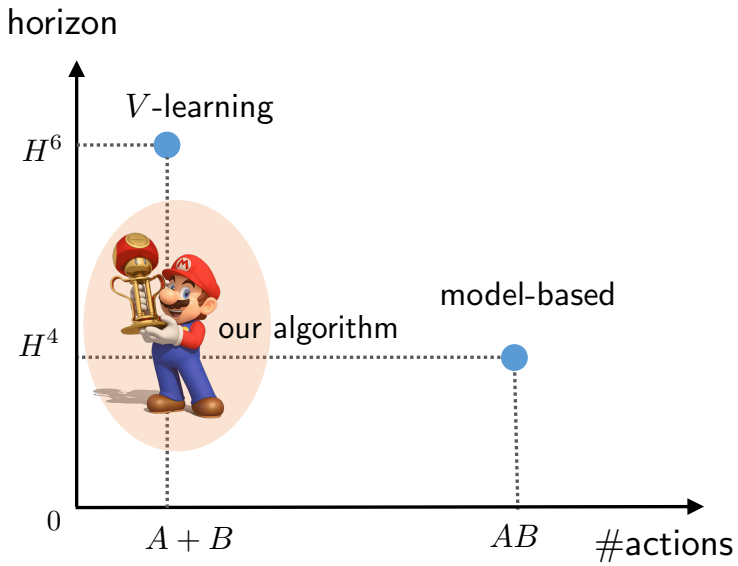


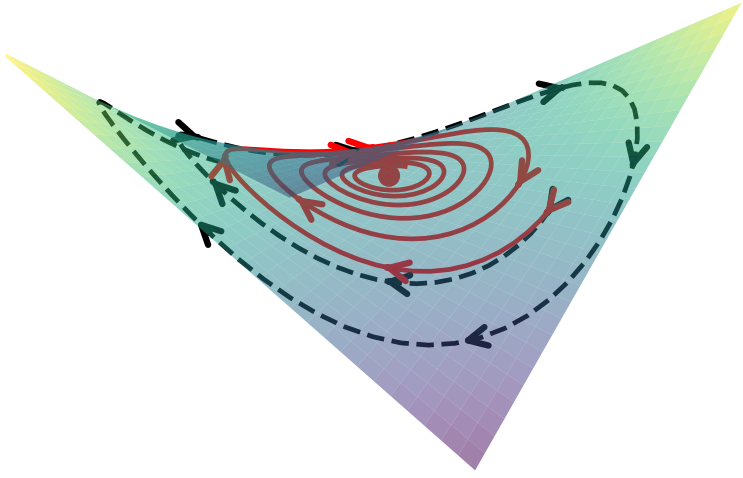
Challenges in MARL: curse of multiple agents



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

This talk: two-player zero-sum Markov games



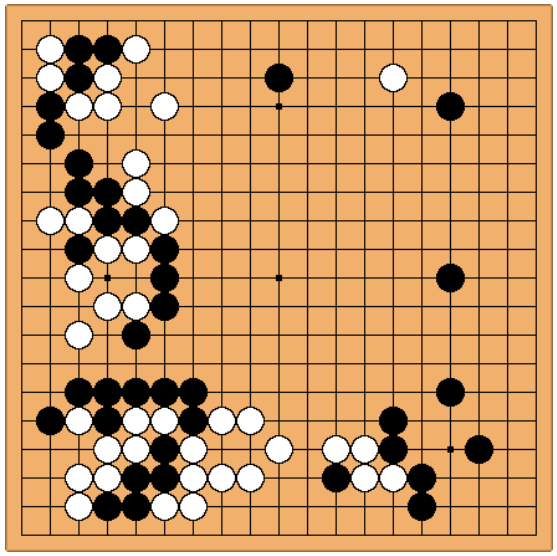


Statistical perspective:
Minimax optimality under the generative model

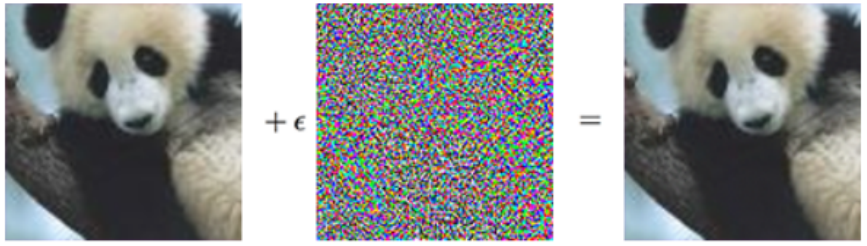
Optimization perspective:
Last-iterate convergence of policy optimization

Backgrounds: two-player zero-sum Markov games

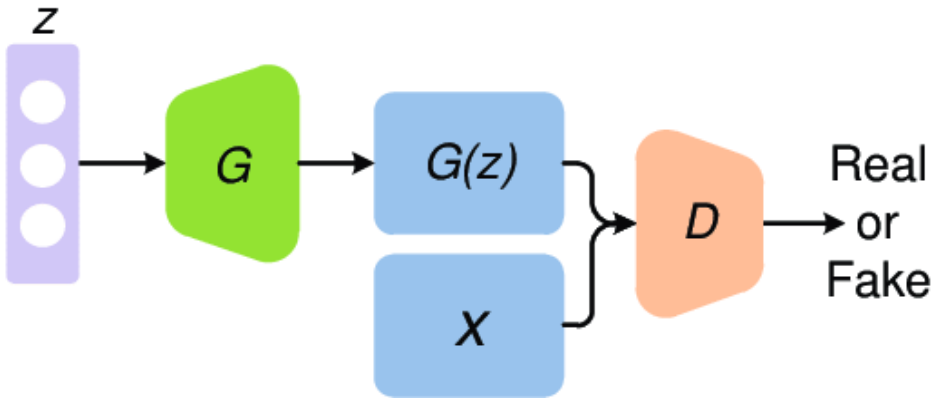
Competitive games



Go

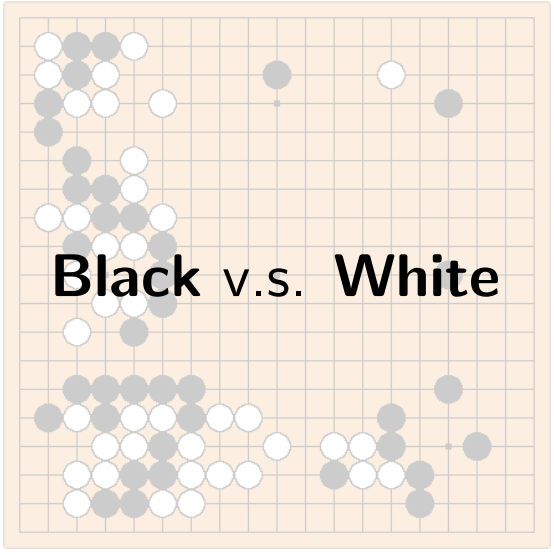


Adversarial Training

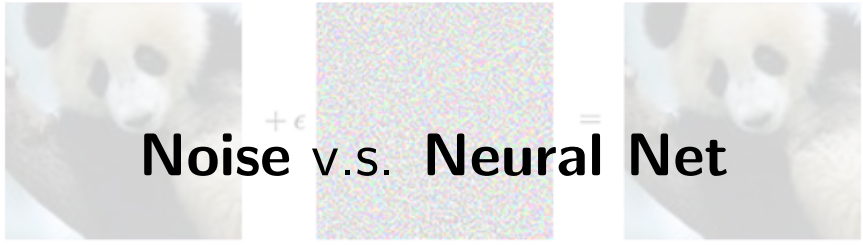


Generative Adversarial Networks

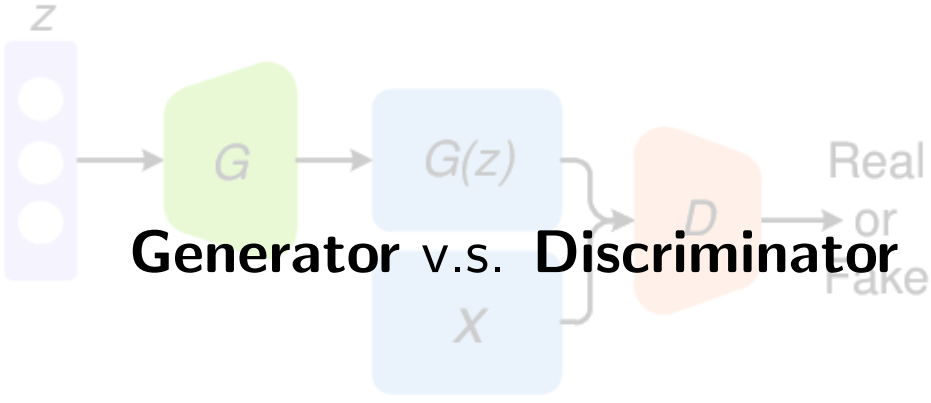
Competitive games



Black v.s. **White**

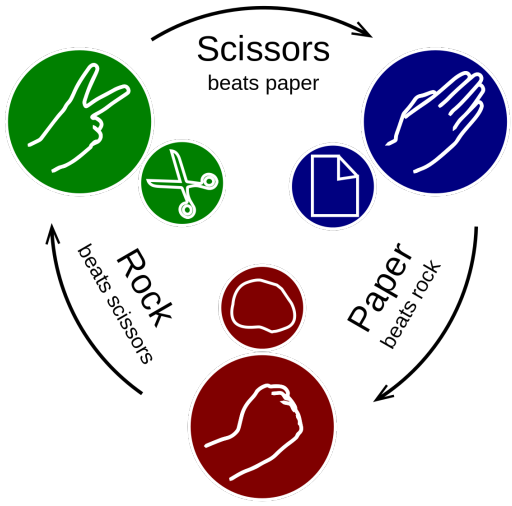


Noise v.s. **Neural Net**



Generator v.s. **Discriminator**

Zero-sum two-player matrix game



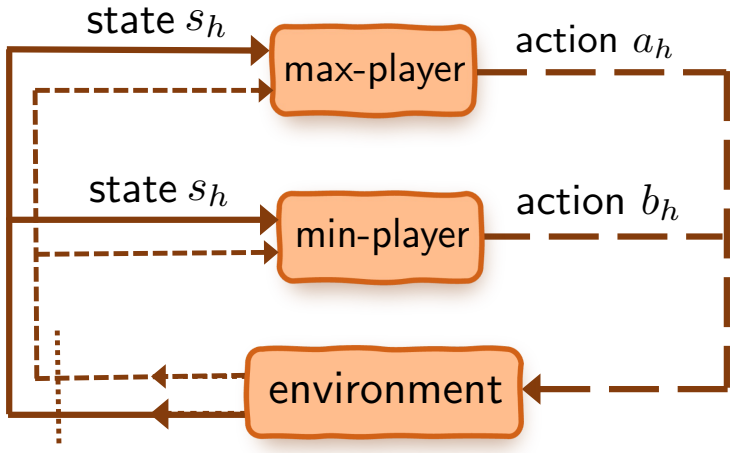
	0	-1	1
	1	0	-1
	-1	1	0

Zero-sum two-player matrix game

$$\max_{\mu \in \Delta(\mathcal{A})} \min_{\nu \in \Delta(\mathcal{B})} \mu^\top A \nu$$

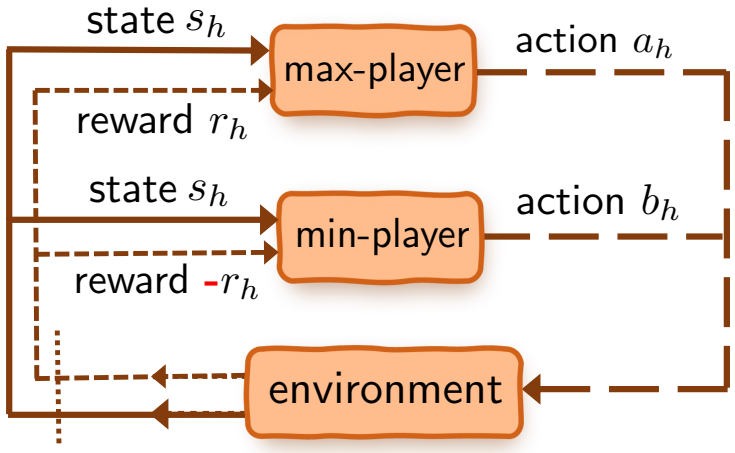
- $\mathcal{A}$, $\mathcal{B}$: action space of the two players;
- $\mu \in \Delta(\mathcal{A})$, $\nu \in \Delta(\mathcal{B})$: policies of the two players;
- $\Delta(\mathcal{A})$, $\Delta(\mathcal{B})$: set of probability distribution over $\mathcal{A}$, $\mathcal{B}$;
- $A \in \mathbb{R}^{|\mathcal{A}| \times |\mathcal{B}|}$: payoff matrix.

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



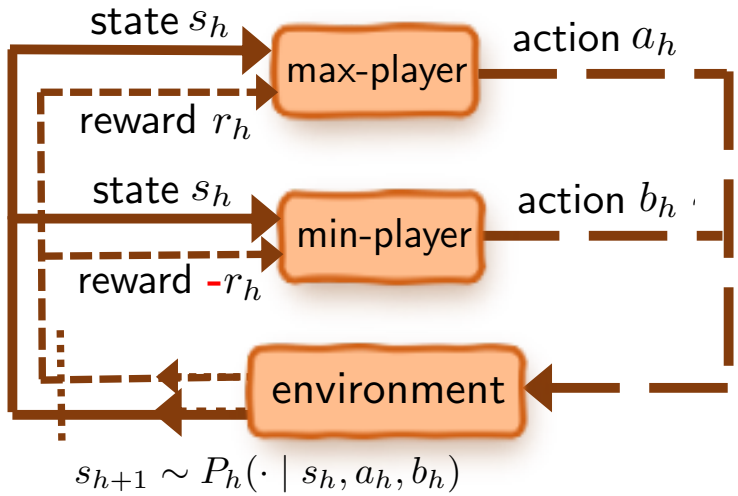
- $\mathcal{S}$: shared state space
- H: horizon
- $\mathcal{A}$: action space of max-player
- $\mathcal{B}$: action space of min-player

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



- $\mathcal{S}$: shared state space
- $\mathcal{A}$: action space of max-player
- H: horizon
- $\mathcal{B}$: action space of min-player
- immediate reward: max-player $r_h(s, a, b) \in [0, 1]$
  min-player $-r_h(s, a, b)$

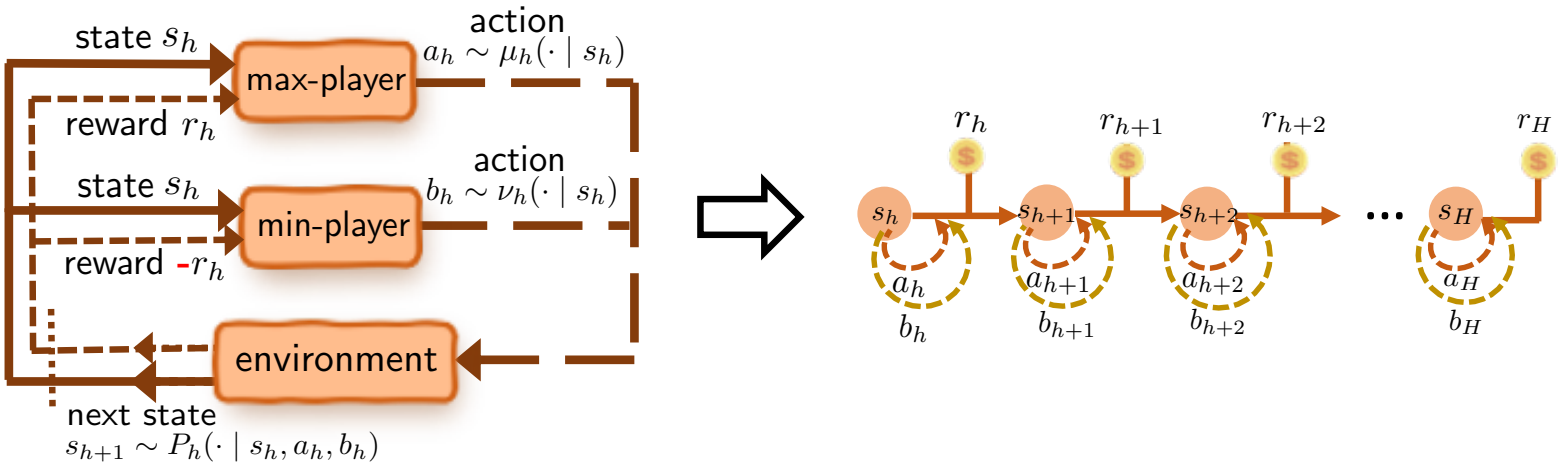
Two-player zero-sum Markov games (finite-horizon)



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

Value function of policy pair

μ: policy of max-player; ν: policy of min-player

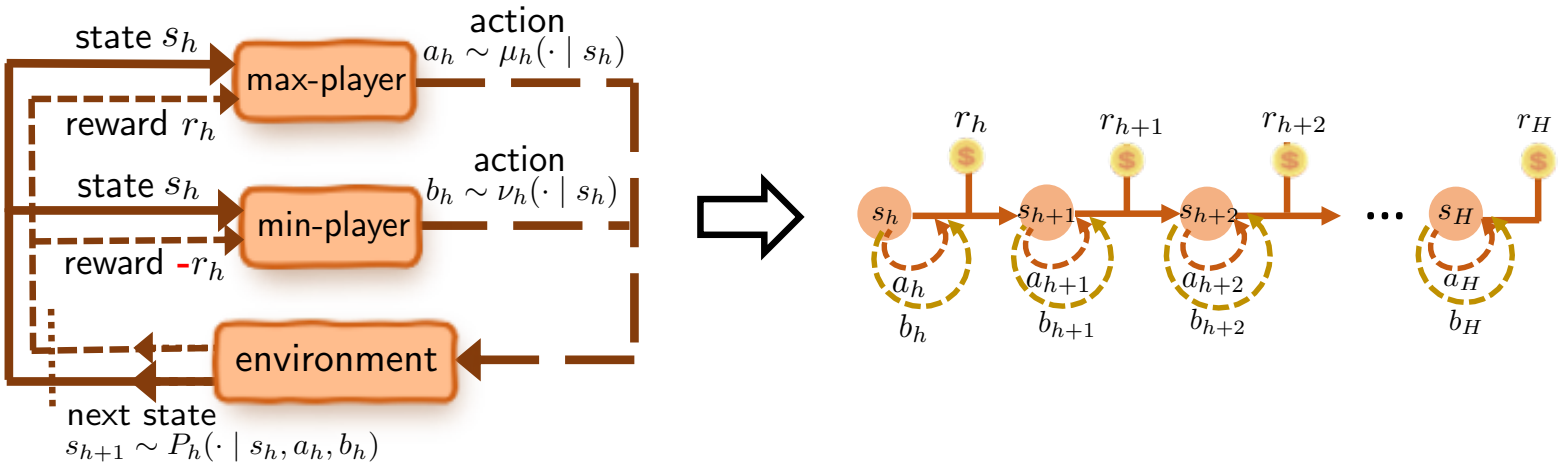


Value function of policy pair (μ, ν):

$$V^{\mu,\nu}(s) := \mathbb{E}\left[\sum_{t=1}^{H} r_t(s_t, a_t, b_t) \,\middle|\, s_1 = s\right]$$

Value function of policy pair

μ: policy of max-player; ν: policy of min-player

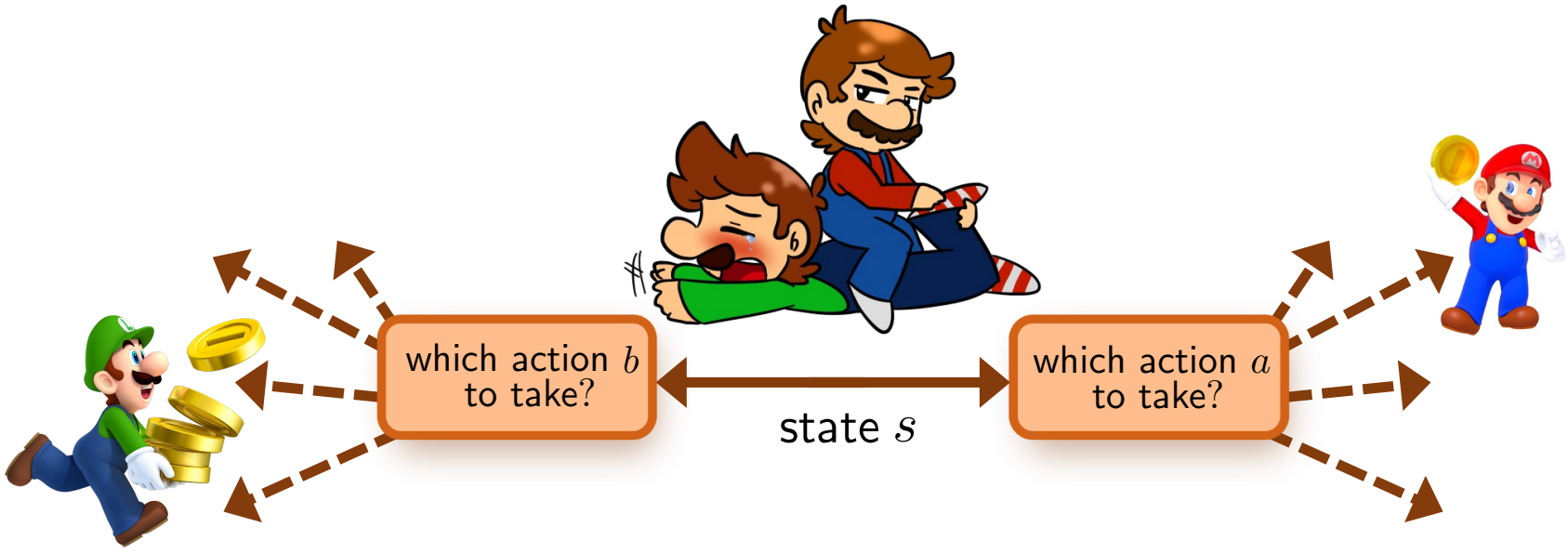


Value function of policy pair (μ, ν):

$$V^{\mu,\nu}(s) := \mathbb{E}\left[\sum_{t=1}^{H} r_t(s_t, a_t, b_t) \,\middle|\, s_1 = s\right]$$

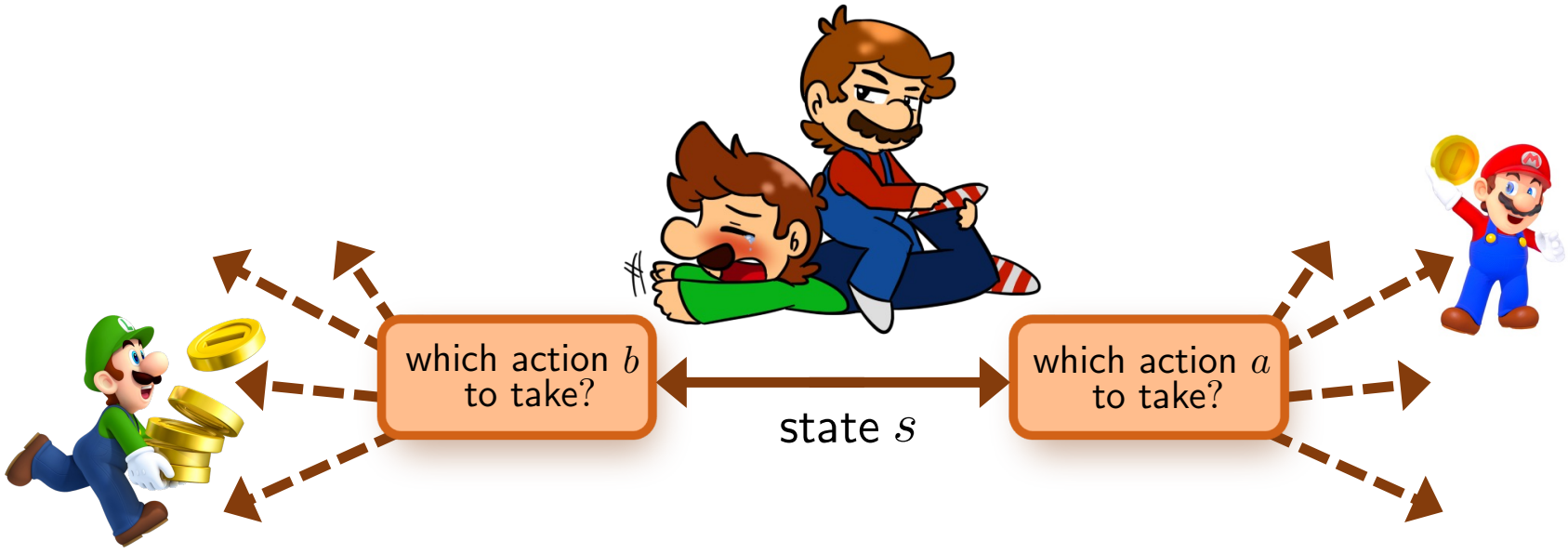
- $\{(a_t, b_t, s_{t+1})\}$: generated when max-player and min-player execute policies μ and ν *independently (i.e. no coordination)*

Target policy



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

Target policy



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

Zero-sum two-player Markov game

$$\max_{\mu \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \min_{\nu \in \Delta(\mathcal{B})^{|\mathcal{S}|}} V^{\mu,\nu}(s)$$

Nash equilibrium (NE)

John von Neumann

John Nash

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

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

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

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 equilibrium (NE)

John von Neumann

John Nash

An ϵ-NE policy pair $(\widehat{\mu}, \widehat{\nu})$ obeys

$$\max_{\mu} V^{\mu,\widehat{\nu}} - \epsilon \leq V^{\widehat{\mu},\widehat{\nu}} \leq \min_{\nu} V^{\widehat{\mu},\nu} + \epsilon$$

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

Nash equilibrium (NE)

John von Neumann

John Nash

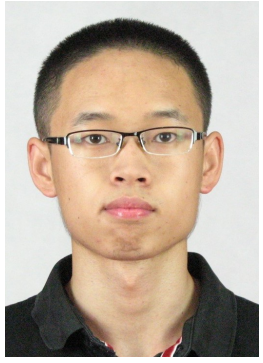
An ϵ-NE policy pair $(\widehat{\mu}, \widehat{\nu})$ obeys

$$\max_{\mu} V^{\mu,\widehat{\nu}} - \epsilon \leq V^{\widehat{\mu},\widehat{\nu}} \leq \min_{\nu} V^{\widehat{\mu},\nu} + \epsilon$$

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

Goal: efficiently learn the NE statistically and computationally

A statistical perspective: Minimax-optimal sample complexity under the generative model



Gen Li
UPenn

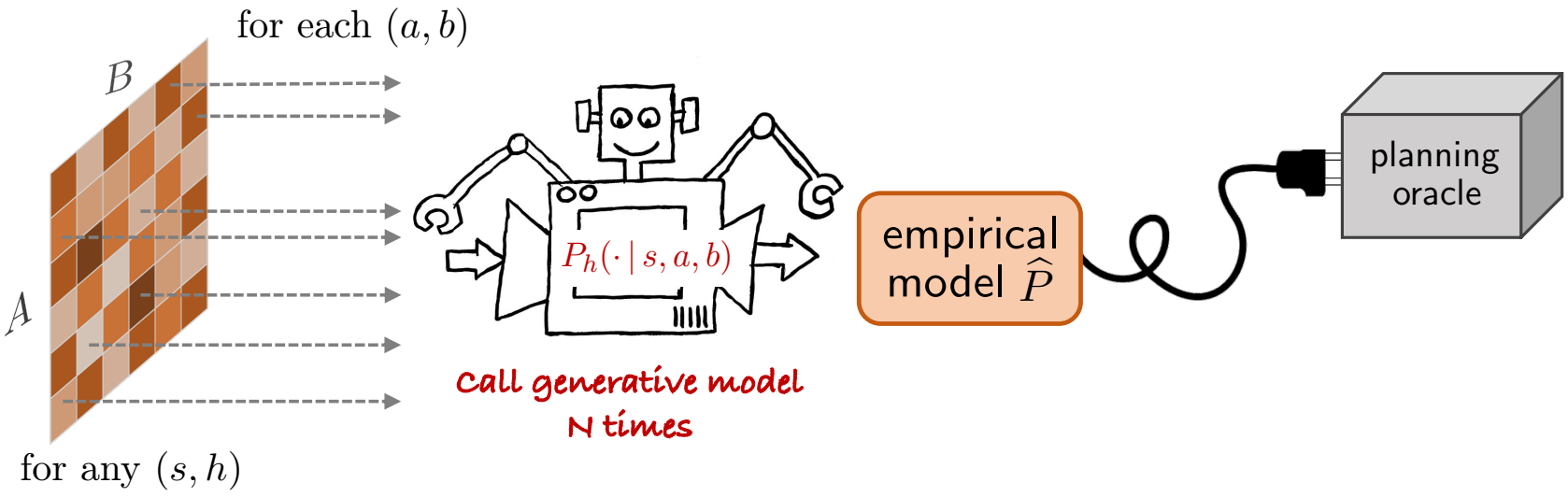
Yuxin Chen
UPenn

Yuting Wei
UPenn

"Minimax-optimal multi-agent RL in Markov games with a generative model,"
G. Li, Y. Chi, Y. Wei, Y. Chen, NeurIPS 2022

Model-based approach w/ non-adaptive sampling

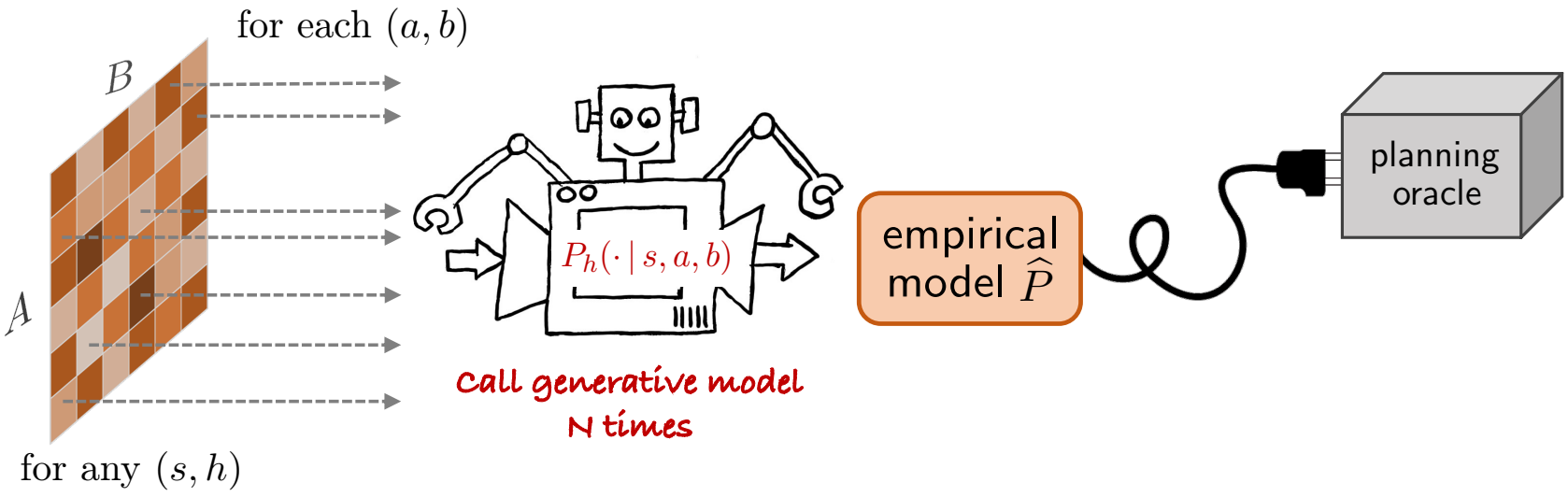
(Zhang et al., 2020)



1. for each (s, a, b, h), call generative models N times

Model-based approach w/ non-adaptive sampling

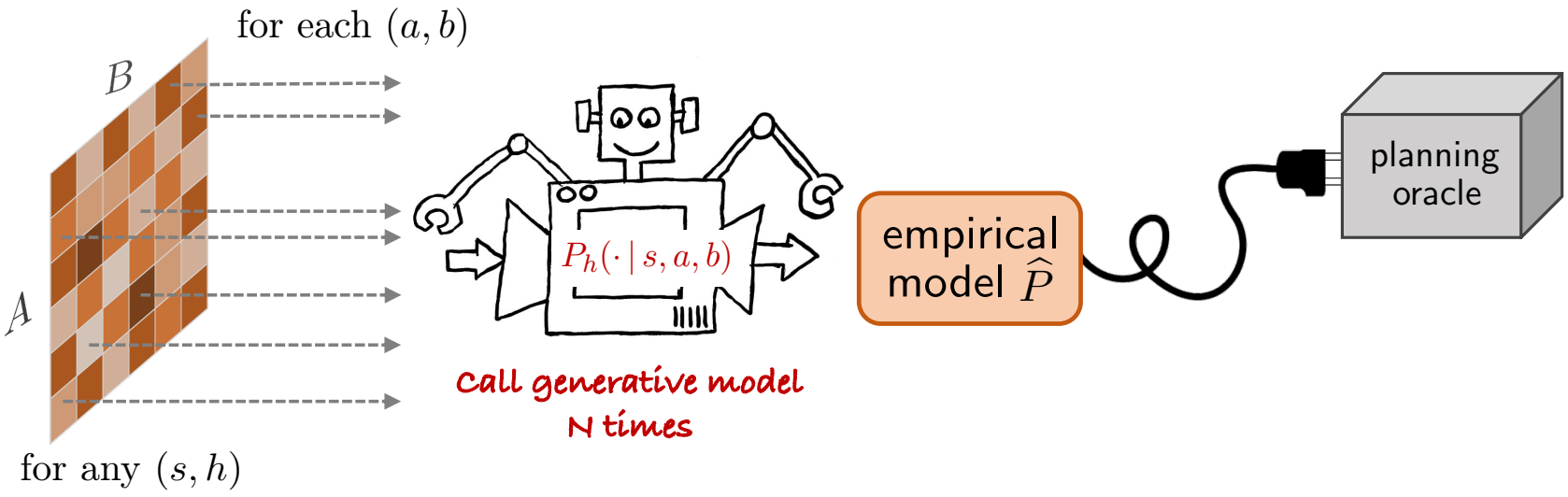
(Zhang et al., 2020)



1. for each (s, a, b, h), call generative models N times
2. build empirical model $\widehat{P}$

Model-based approach w/ non-adaptive sampling

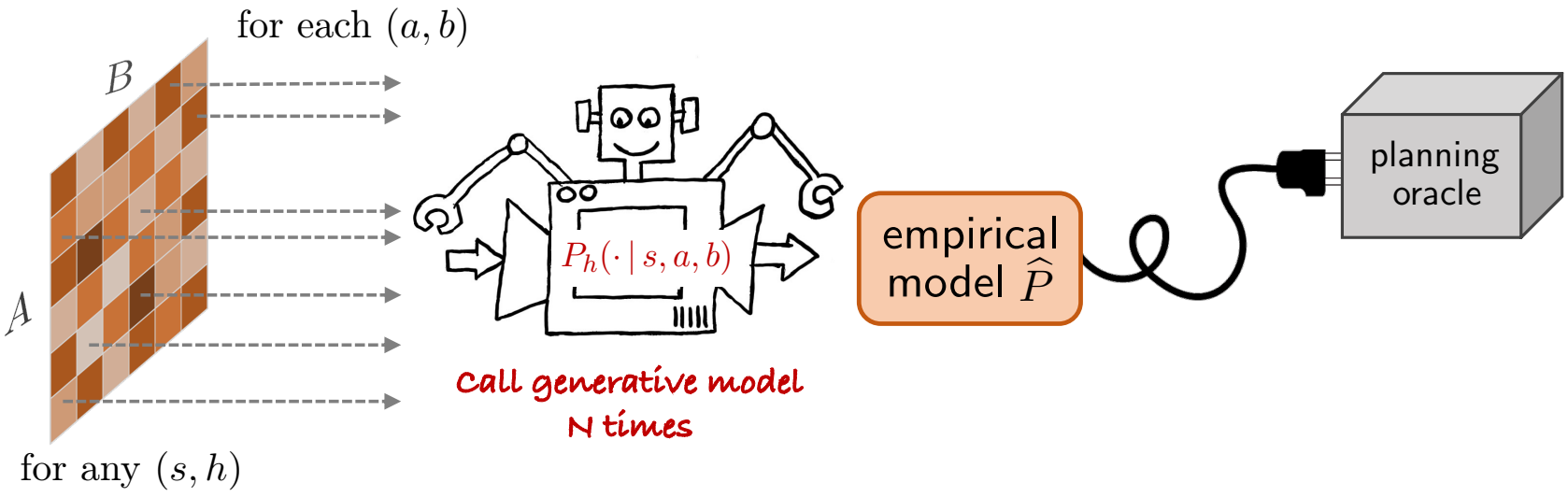
(Zhang et al., 2020)



1. for each (s, a, b, h), call generative models N times
2. build empirical model $\widehat{P}$, and run classical planning algorithms

Model-based approach w/ non-adaptive sampling

(Zhang et al., 2020)

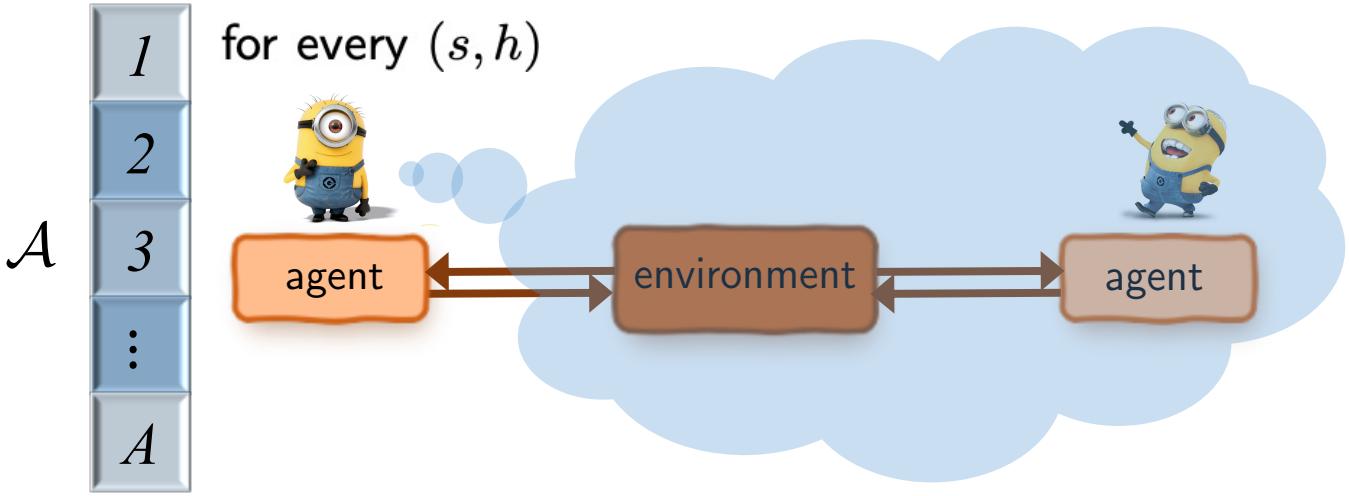


1. for each (s, a, b, h), call generative models N times
2. build empirical model $\widehat{P}$, and run classical planning algorithms

sample complexity: $\frac{H^4 SAB}{\epsilon^2}$

Breaking the curse of multi-agents?

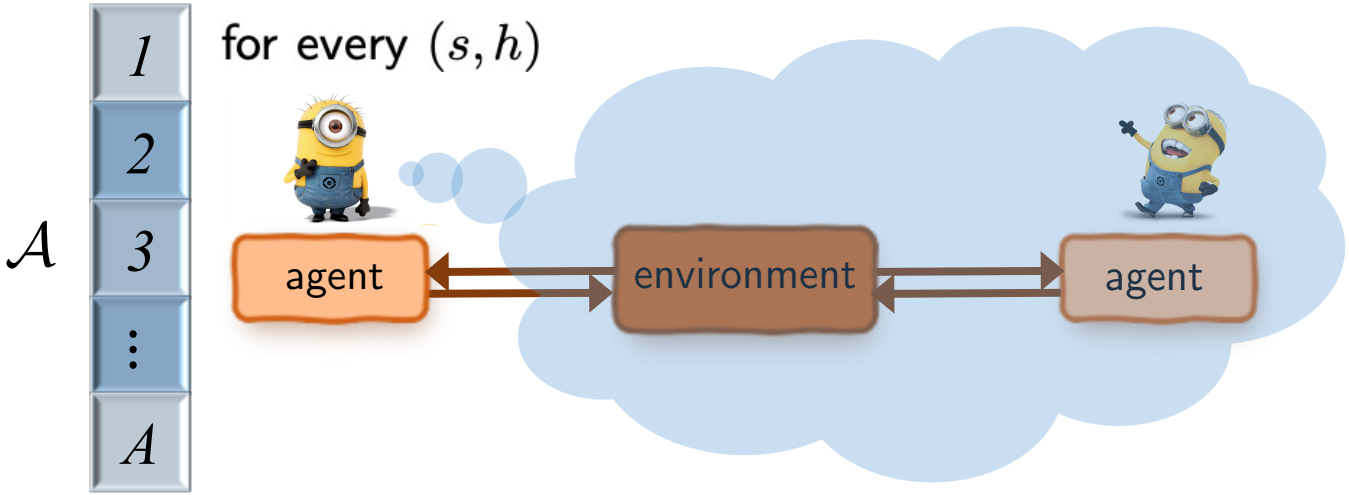
(Song, Mei, Bai, 2021; Jin et al., 2021; Basar et al., 2021)



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

Breaking the curse of multi-agents?

(Song, Mei, Bai, 2021; Jin et al., 2021; Basar et al., 2021)

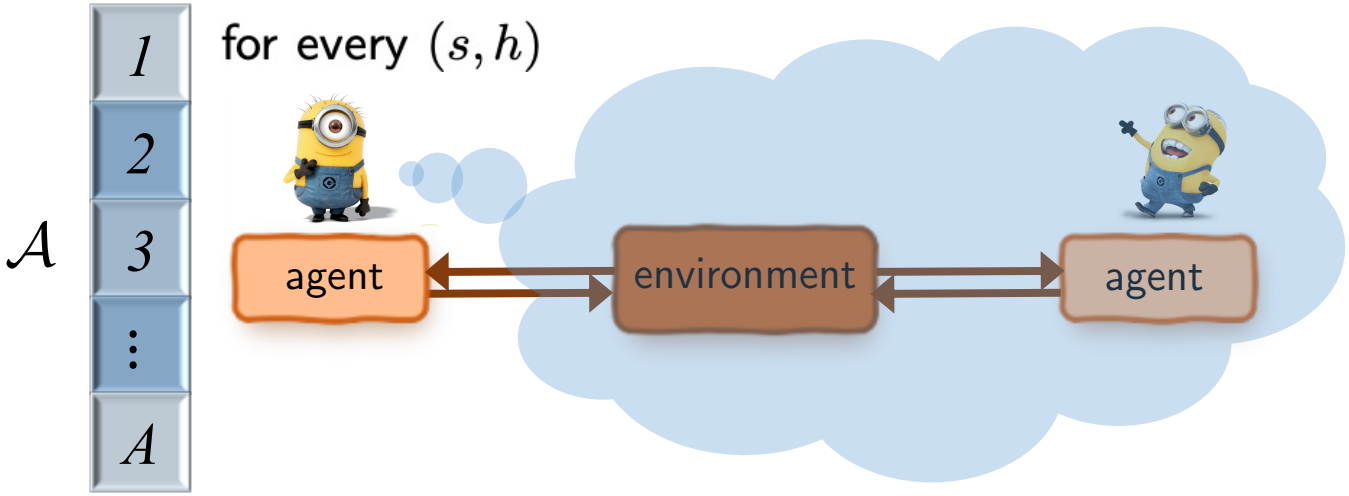


V-learning (online setting): MARL meets adversarial learning: for the max-player, for $h = 1, \ldots, H$

1. *adaptive sampling:* sampling $\mathcal{A}$ based on $\mu_h(\cdot|s)$

Breaking the curse of multi-agents?

(Song, Mei, Bai, 2021; Jin et al., 2021; Basar et al., 2021)

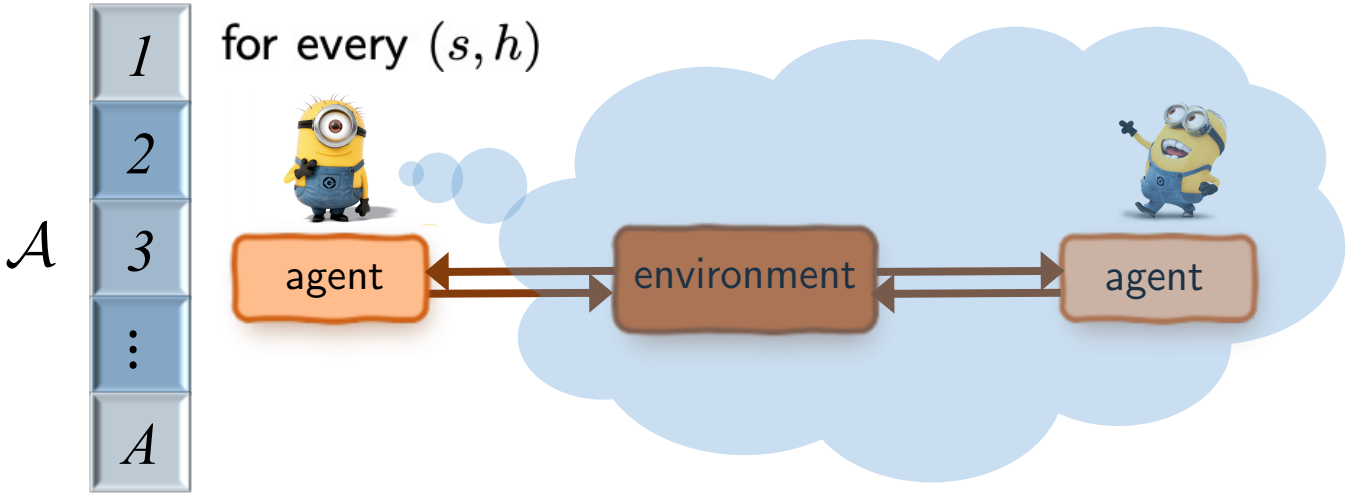


V-learning (online setting): MARL meets adversarial learning: for the max-player, for $h = 1, \ldots, H$

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

Breaking the curse of multi-agents?

(Song, Mei, Bai, 2021; Jin et al., 2021; Basar et al., 2021)

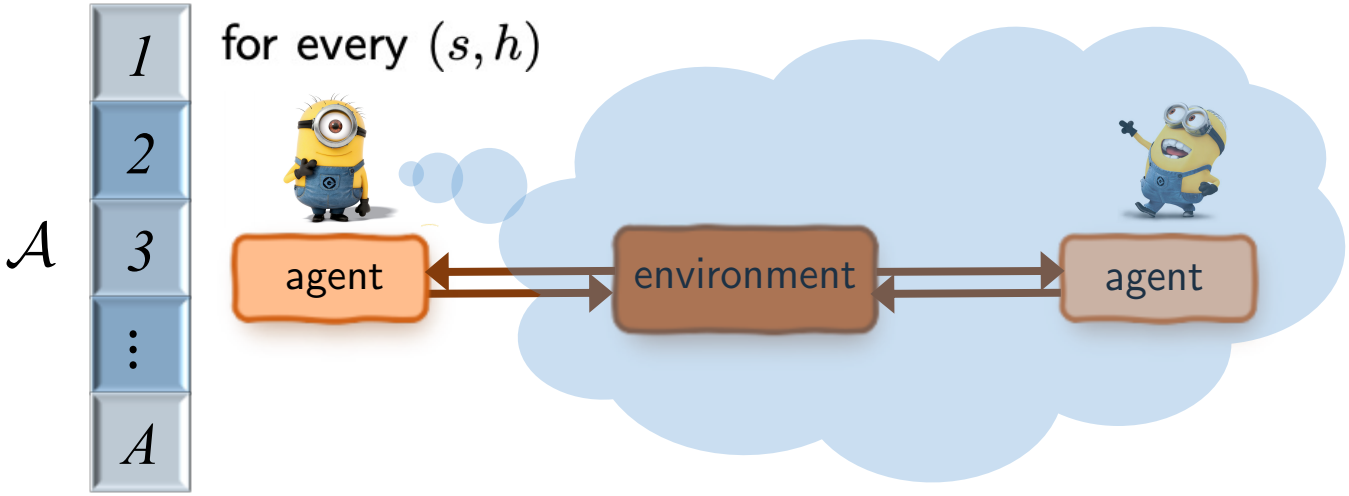


V-learning (online setting): MARL meets adversarial learning: for the max-player, for $h = 1, \ldots, H$

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 learning subroutine*, e.g. FTRL

Breaking the curse of multi-agents?

(Song, Mei, Bai, 2021; Jin et al., 2021; Basar et al., 2021)

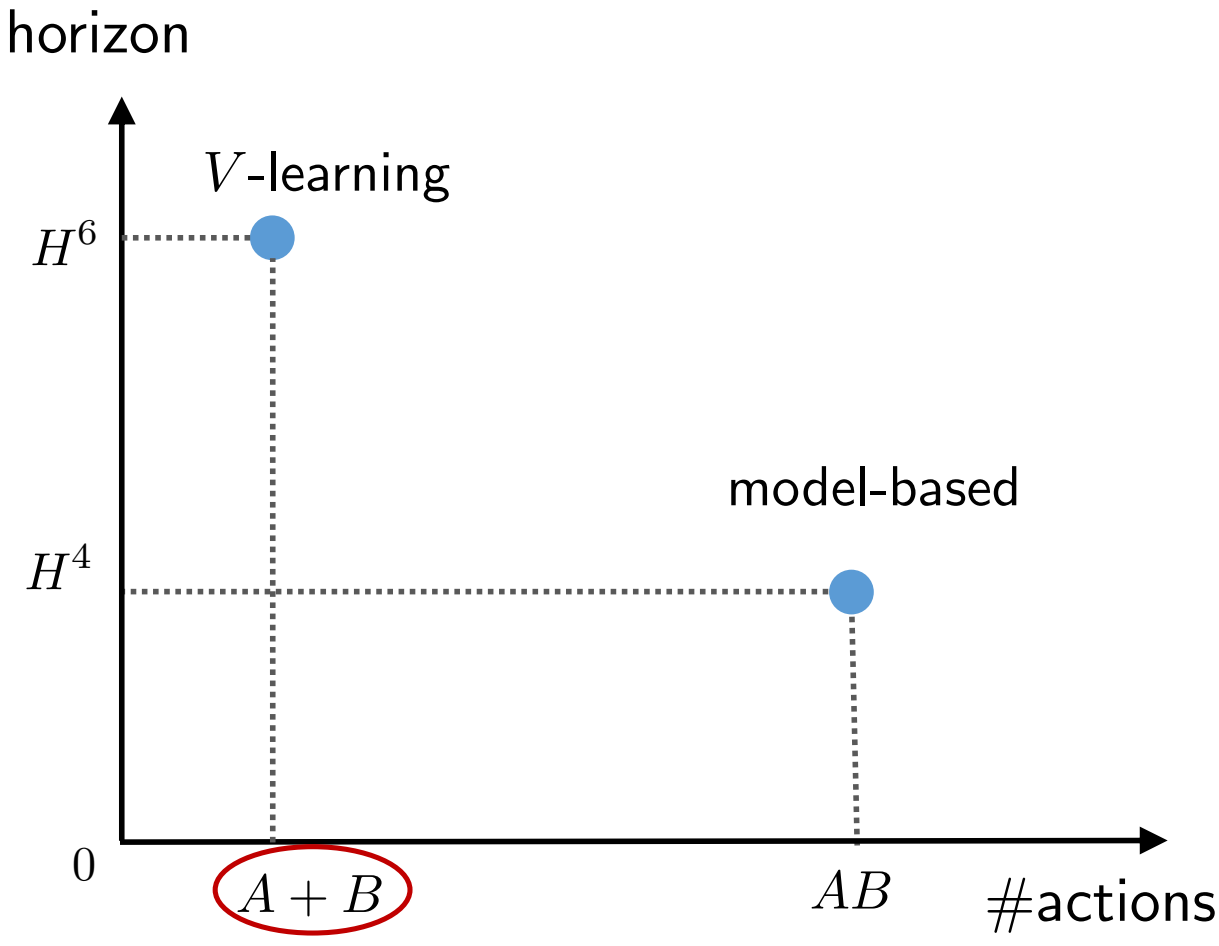


V-learning (online setting): MARL meets adversarial learning: for the max-player, for $h = 1, \ldots, H$

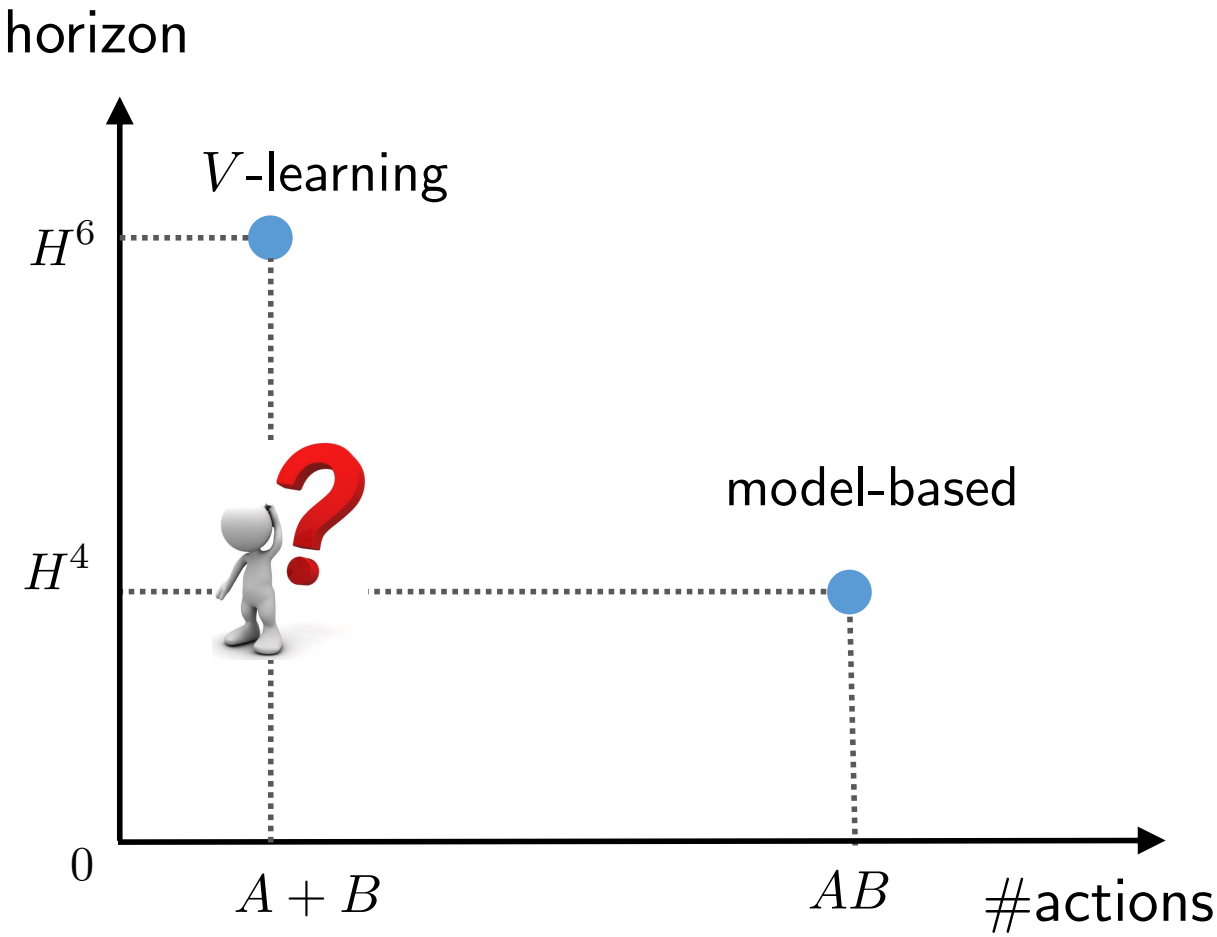
1. *adaptive sampling:* sampling $\mathcal{A}$ based on $\mu_h(\cdot|s)$
2. estimate V-function only with *Hoeffding bonus* (of size S)
3. update policy via *adversarial learning subroutine*, e.g. FTRL

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

Summary of prior arts



Summary of prior arts

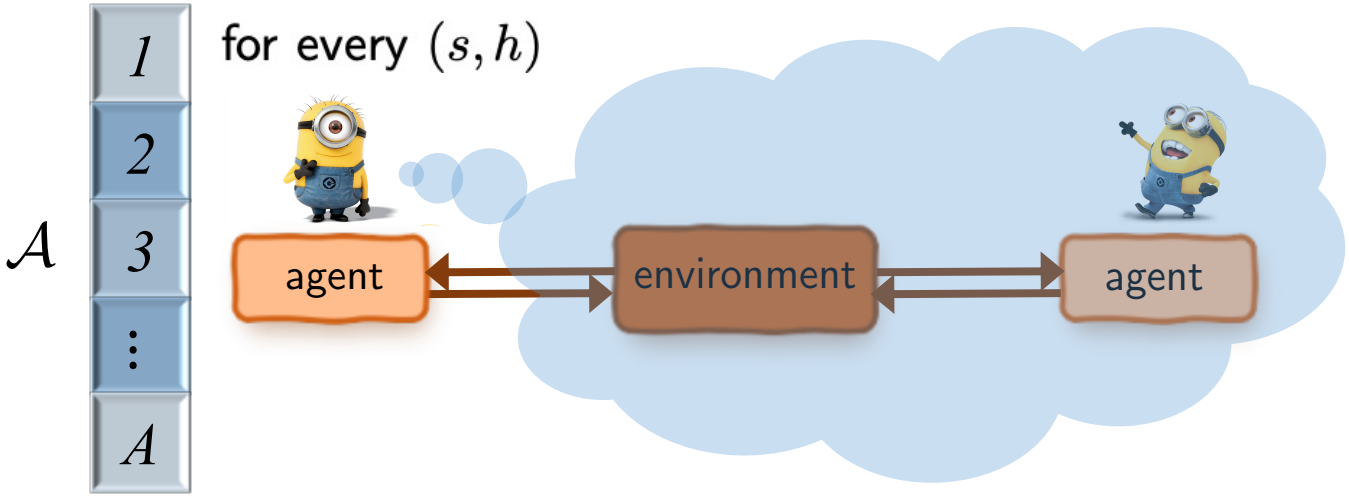


Can we simultaneously overcome

curse of multi-agents & barrier of long horizon?

Our algorithm (with a generative model)

(Li et al., NeurIPS 2022)

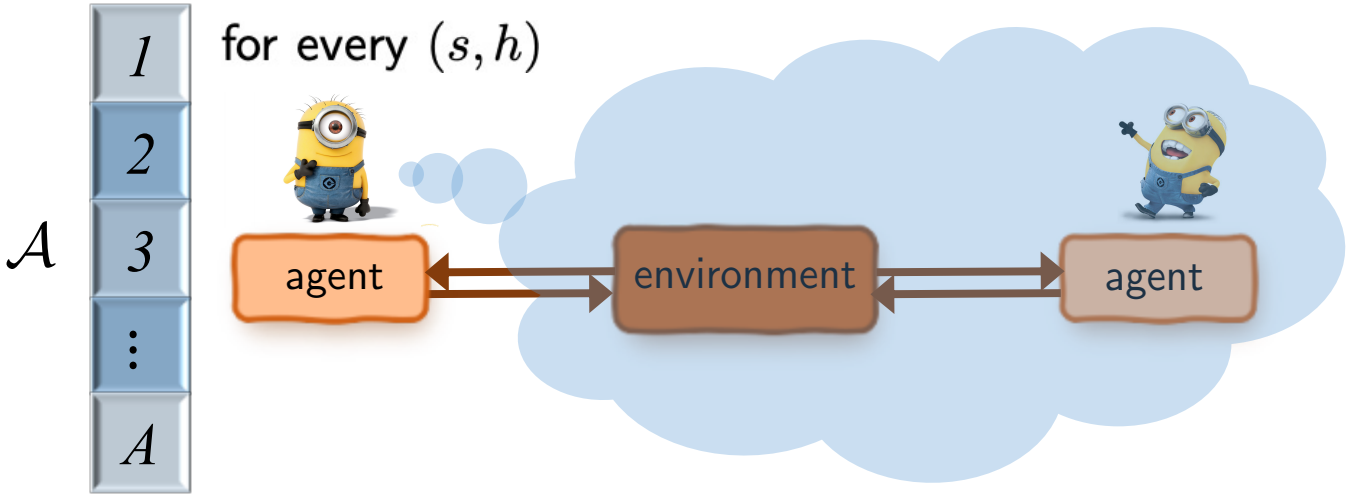


Nash-Q-FTRL (ours): for the max-player, for $h = H, \ldots, 1$

- collect $k = 1, \ldots, K$ samples:

Our algorithm (with a generative model)

(Li et al., NeurIPS 2022)

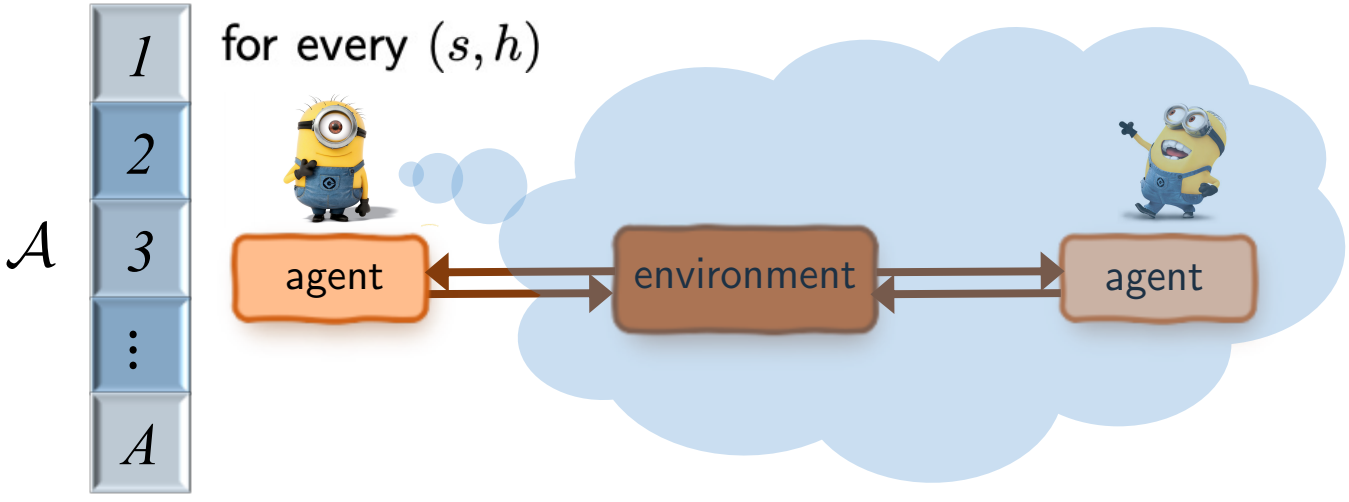


Nash-Q-FTRL (ours): for the max-player, for $h = H, \ldots, 1$

- collect $k = 1, \ldots, K$ samples:
  1. *adaptive sampling:* sample $\mathcal{A}$ based on $\mu_h^k(\cdot|s)$

Our algorithm (with a generative model)

(Li et al., NeurIPS 2022)

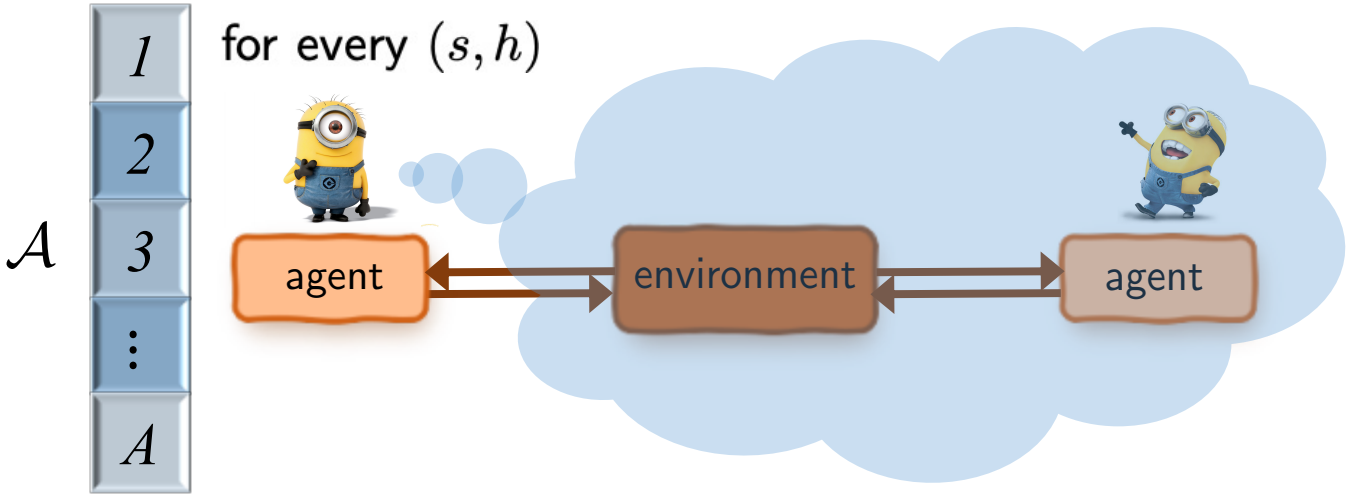


Nash-Q-FTRL (ours): for the max-player, for $h = H, \ldots, 1$

- 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

Our algorithm (with a generative model)

(Li et al., NeurIPS 2022)

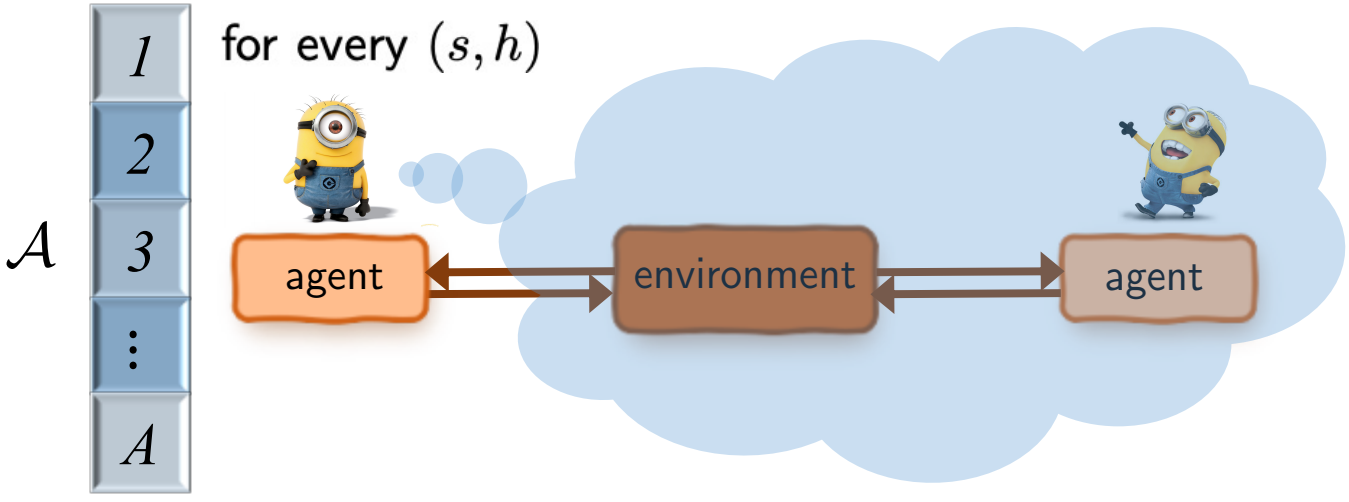


Nash-Q-FTRL (ours): for the max-player, for $h = H, \ldots, 1$

- 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 FTRL

Our algorithm (with a generative model)

(Li et al., NeurIPS 2022)



Nash-Q-FTRL (ours): 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 FTRL
- output a Markov policy μ_h and V_h with Bernstein bonuses

Main result: two-player zero-sum Markov games

Theorem (Li, Chi, Wei, Chen '22)

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

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

Main result: two-player zero-sum Markov games

Theorem (Li, Chi, Wei, Chen '22)

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

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

- **minimax lower bound:** $\widetilde{\Omega}\left(\frac{H^4 S(A+B)}{\epsilon^2}\right)$
- breaks curse of multi-agents & long-horizon barrier at once!

Main result: two-player zero-sum Markov games

Theorem (Li, Chi, Wei, Chen '22)

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

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

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

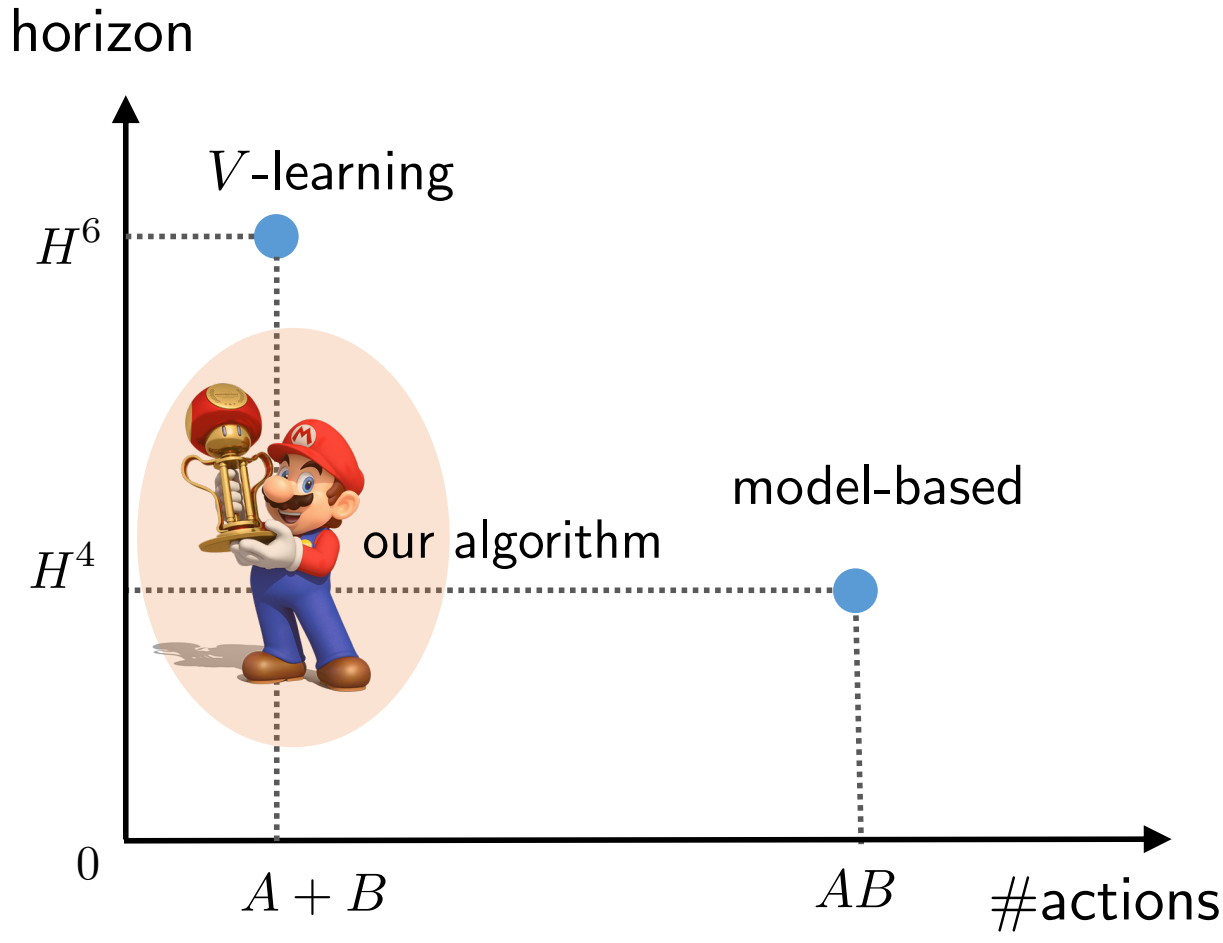
Main result: two-player zero-sum Markov games

Theorem (Li, Chi, Wei, Chen '22)

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

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

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



Our algorithm breaks curses of multi-agents and long-horizon barrier simultaneously!

Extension: multi-player general-sum Markov games

Theorem (Li, Chi, Wei, Chen '22)

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

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

Extension: multi-player general-sum Markov games

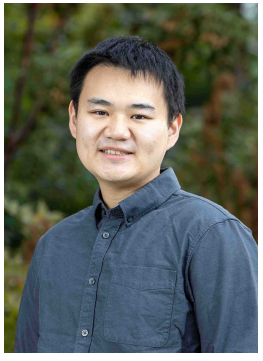
Theorem (Li, Chi, Wei, Chen '22)

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

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

- **minimax lower bound:** $\widetilde{\Omega}\left(\frac{H^4 S \max_i A_i}{\epsilon^2}\right)$
- near-optimal when the number of players m is fixed

An optimization lens: last-iterate convergence of policy optimization with entropy regularization



Shicong Cen — CMU

Yuting Wei — UPenn

Lin Xiao — Meta AI

Simon Du — UW

"Fast policy extragradient methods for competitive games with entropy regularization," S. Cen, Y. Wei, Y. Chi, NeurIPS 2021.

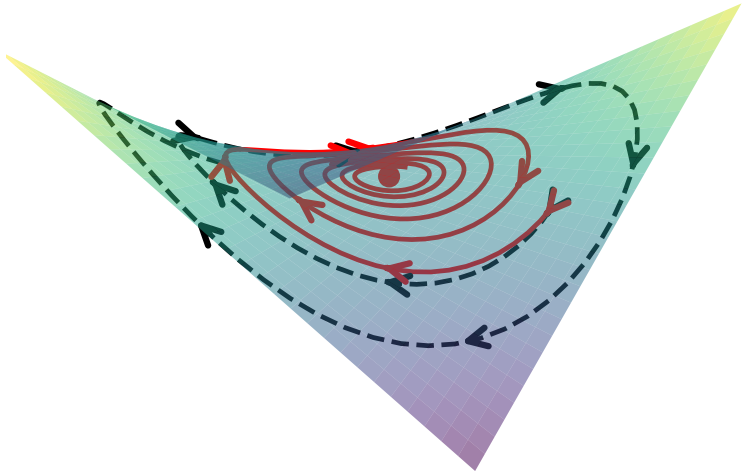
"Faster last-iterate convergence of policy optimization in zero-sum Markov games," S. Cen, Y. Chi, S. Du, L. Xiao, 2022.

Policy optimization: saddle-point optimization

Zero-sum two-player Markov game

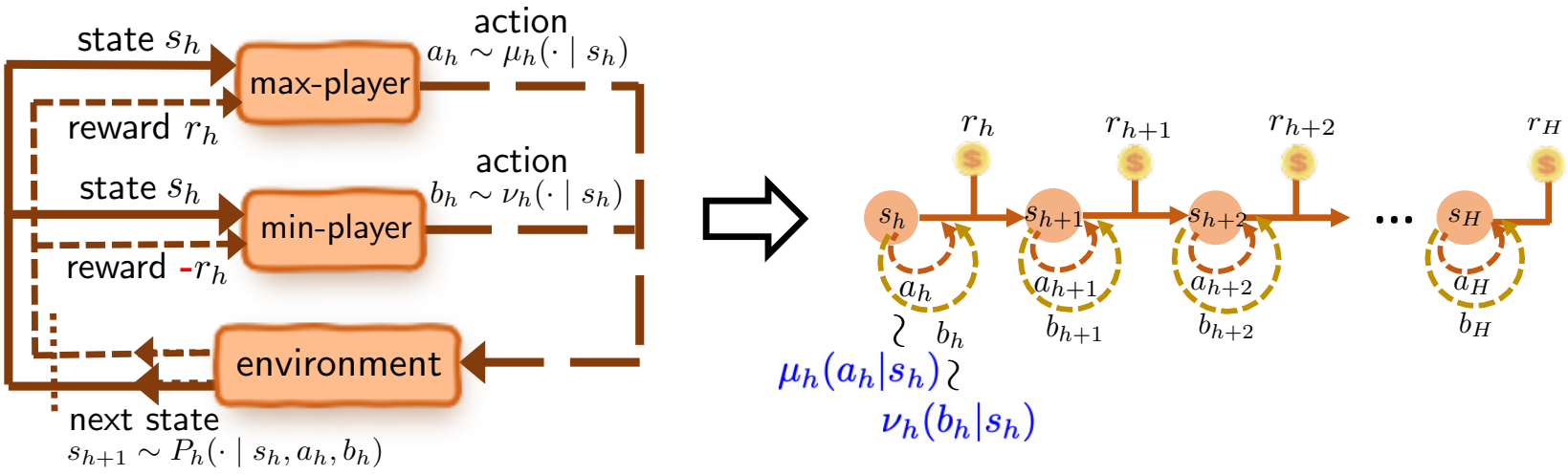
Given an initial state distribution $s \sim \rho$, find policy π such that

$$\max_{\mu \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \min_{\nu \in \Delta(\mathcal{B})^{|\mathcal{S}|}} V^{\mu,\nu}(\rho) := \mathbb{E}_{s\sim\rho}[V^{\mu,\nu}(s)]$$



Can we design a policy optimization method that guarantees fast *last-iterate* convergence?

Entropy regularization in MARL

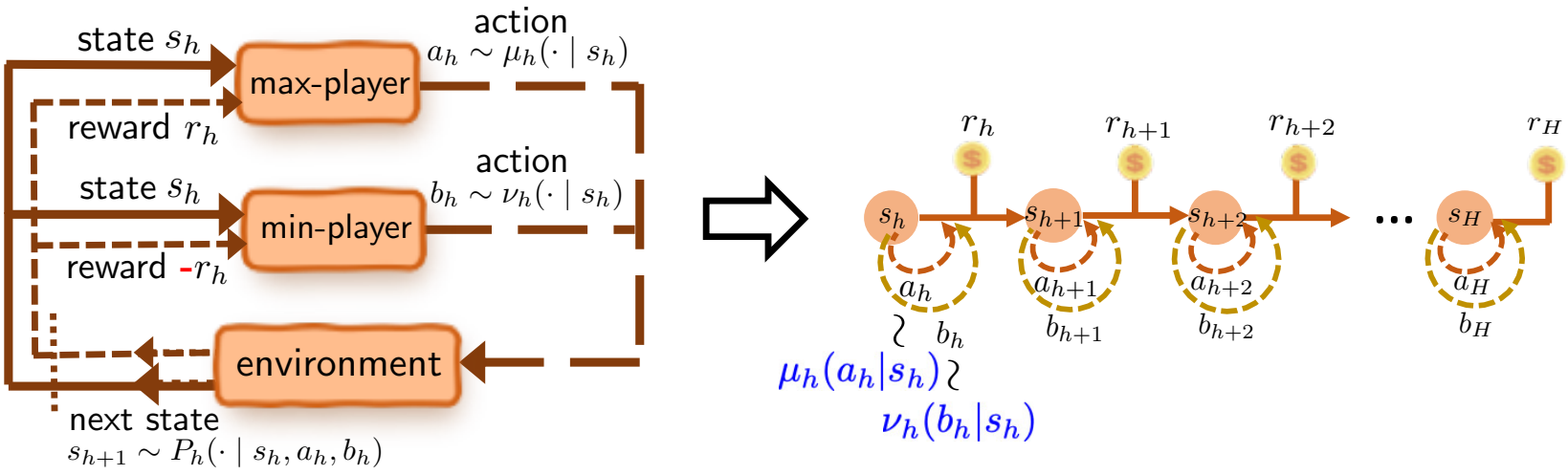


Promote the stochasticity of the policy pair using the **"soft"** value function (Williams and Peng, 1991; Cen et al., 2020):

$$V_\tau^{\mu,\nu}(s) := \mathbb{E}\left[\sum_{h=1}^{H}\left(r_t + \tau\mathcal{H}(\mu_t(\cdot|s_t) - \tau\mathcal{H}(\nu_t(\cdot|s_t)\right)\Big| s_0 = s\right],$$

where $\mathcal{H}$ is the Shannon entropy, and $\tau \geq 0$ is the reg. parameter.

Entropy regularization in MARL



Promote the stochasticity of the policy pair using the **"soft"** value function (Williams and Peng, 1991; Cen et al., 2020):

$$V_\tau^{\mu,\nu}(s) := \mathbb{E}\left[\sum_{h=1}^{H}\left(r_t + \tau\mathcal{H}(\mu_t(\cdot|s_t) - \tau\mathcal{H}(\nu_t(\cdot|s_t)\right)\Big| s_0 = s\right],$$

where $\mathcal{H}$ is the Shannon entropy, and $\tau \geq 0$ is the reg. parameter.

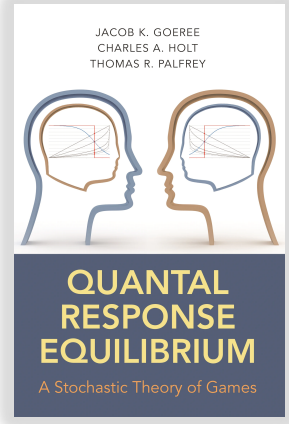
$$\max_{\mu\in\Delta(\mathcal{A})^{|\mathcal{S}|}} \min_{\nu\in\Delta(\mathcal{B})^{|\mathcal{S}|}} V_\tau^{\mu,\nu}(\rho)$$

Quantal response equilibrium (QRE)

Quantal response equilibrium (McKelvey and Palfrey, 1995)

The quantal response equilibrium (QRE) is the policy pair $(\mu_\tau^\star, \nu_\tau^\star)$ *that is the unique solution to*

$$\max_{\mu \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \min_{\nu \in \Delta(\mathcal{B})^{|\mathcal{S}|}} V_\tau^{\mu,\nu}(\rho).$$



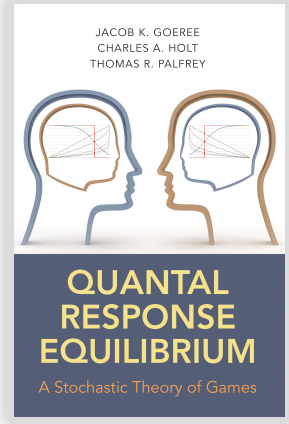
- Unlike NE, QRE assumes bounded rationality: action probability follows the logit function.

Quantal response equilibrium (QRE)

Quantal response equilibrium (McKelvey and Palfrey, 1995)

The quantal response equilibrium (QRE) is the policy pair $(\mu_\tau^\star, \nu_\tau^\star)$ *that is the unique solution to*

$$\max_{\mu \in \Delta(\mathcal{A})^{|\mathcal{S}|}} \min_{\nu \in \Delta(\mathcal{B})^{|\mathcal{S}|}} V_\tau^{\mu,\nu}(\rho).$$



- Unlike NE, QRE assumes bounded rationality: action probability follows the logit function.

Translating to an ϵ-NE: setting $\tau \asymp \widetilde{O}(\epsilon/H)$.

Soft value iteration

Soft value iteration: for $h = H, \dots, 1$

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

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

Soft value iteration

Soft value iteration: for $h = H, \dots, 1$

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

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

Soft value iteration

Soft value iteration: for $h = H, \dots, 1$

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

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

Entropy-regularized matrix game

$$\max_{\mu \in \Delta(\mathcal{A})} \min_{\nu \in \Delta(\mathcal{B})} \mu^\top A \nu + \tau\mathcal{H}(\mu) - \tau\mathcal{H}(\nu)$$

A prelude: entropy-regularized matrix game

Optimistic multiplicative weights update (OMWU) method

(Related to OMD, Rakhlin and Sridharan, 2013): for $t = 0, 1, \cdots,$

$$\text{predict}: \begin{cases} \bar{\mu}^{(t+1)} \propto [\mu^{(t)}]^{1-\eta\tau} \exp\left([A\bar{\nu}^{(t)}]/\tau\right)^{\eta\tau} \\ \bar{\nu}^{(t+1)} \propto [\nu^{(t)}]^{1-\eta\tau} \exp\left(-[A^\top \bar{\mu}^{(t)}]/\tau\right)^{\eta\tau} \end{cases}$$

$$\text{update}: \begin{cases} \mu^{(t+1)} \propto [\mu^{(t)}]^{1-\eta\tau} \exp\left([A\bar{\nu}^{(t+1)}]/\tau\right)^{\eta\tau} \\ \nu^{(t+1)} \propto [\nu^{(t)}]^{1-\eta\tau} \exp\left(-[A^\top \bar{\mu}^{(t+1)}]/\tau\right)^{\eta\tau} \end{cases}$$

A prelude: entropy-regularized matrix game

Optimistic multiplicative weights update (OMWU) method

(Related to OMD, Rakhlin and Sridharan, 2013): for $t = 0, 1, \cdots$,

$$\text{predict}: \begin{cases} \bar{\mu}^{(t+1)} \propto [\mu^{(t)}]^{1-\eta\tau} \exp\left([A\bar{\nu}^{(t)}]/\tau\right)^{\eta\tau} \\ \bar{\nu}^{(t+1)} \propto [\nu^{(t)}]^{1-\eta\tau} \exp\left(-[A^\top\bar{\mu}^{(t)}]/\tau\right)^{\eta\tau} \end{cases}$$

$$\text{update}: \begin{cases} \mu^{(t+1)} \propto [\mu^{(t)}]^{1-\eta\tau} \exp\left([A\bar{\nu}^{(t+1)}]/\tau\right)^{\eta\tau} \\ \nu^{(t+1)} \propto [\nu^{(t)}]^{1-\eta\tau} \exp\left(-[A^\top\bar{\mu}^{(t+1)}]/\tau\right)^{\eta\tau} \end{cases}$$

Theorem (Cen, Wei, Chi, 2021)

Suppose that $\eta \le \min\left\{\frac{1}{2\tau+2\|A\|_\infty}, \frac{1}{4\|A\|_\infty}\right\}$, *then for all* $t \ge 0$, *the last-iterate converges to* ϵ*-QRE within* $\widetilde{O}\left(\frac{1}{\eta\tau}\log\frac{1}{\epsilon}\right)$ *iterations.*

Linear, last-iterate convergence to the QRE!

Soft value iteration via nested-loop OMWU

Soft value iteration: for $h = H, \dots, 1$

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

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

Soft value iteration via nested-loop OMWU

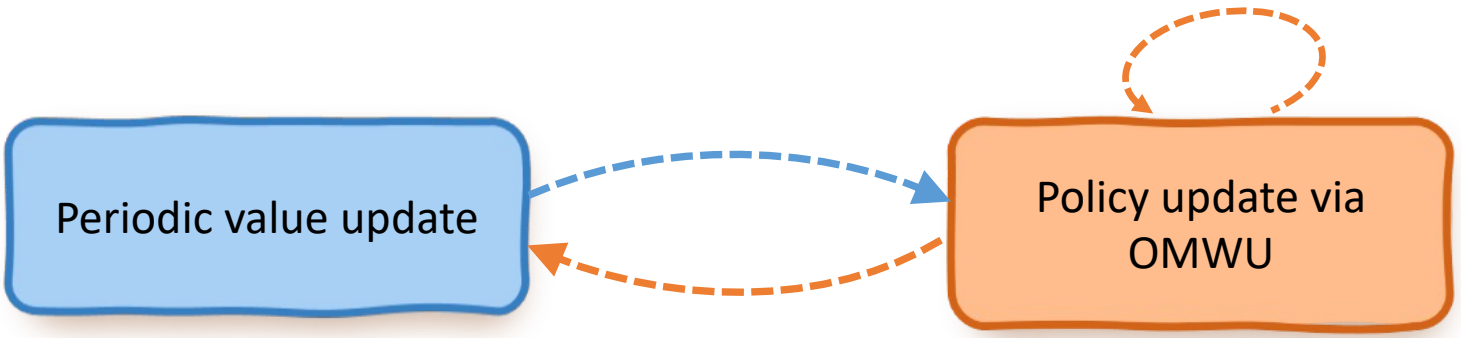
Soft value iteration: for $h = H, \dots, 1$

$$Q_h(s,a,b) \leftarrow r_h(s,a,b) + \cdot \mathop{\mathbb{E}}_{s' \sim P_h(\cdot|s,a,b)} \Big[\underbrace{\max_{\mu} \min_{\nu} \mu(s')^\top Q_{h+1}(s')\nu(s') + \tau\mathcal{H}(\mu(s')) - \tau\mathcal{H}(\nu(s'))}_{\text{Entropy-regularized matrix game}} \Big],$$

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

Nested-loop approach:

$(\mu_h^{(t)}, \nu_h^{(t)}) \leftarrow \texttt{OMWU}(Q_h)$



Soft value iteration via nested-loop OMWU

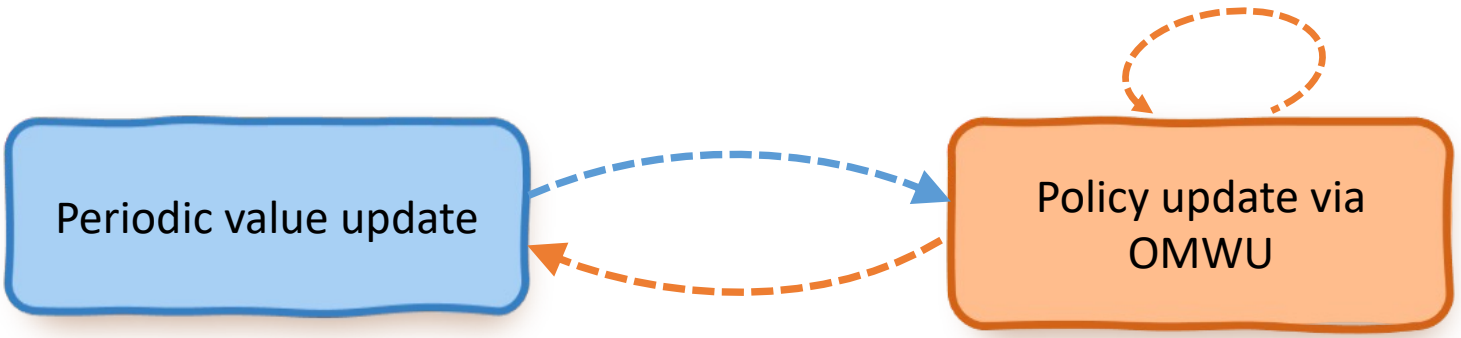
Soft value iteration: for $h = H, \dots, 1$

$$Q_h(s,a,b) \leftarrow r_h(s,a,b) + \cdot \mathop{\mathbb{E}}_{s' \sim P_h(\cdot|s,a,b)} \Big[\underbrace{\max_{\mu} \min_{\nu} \mu(s')^\top Q_{h+1}(s')\nu(s') + \tau\mathcal{H}(\mu(s')) - \tau\mathcal{H}(\nu(s'))}_{\text{Entropy-regularized matrix game}} \Big],$$

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

Nested-loop approach:

$(\mu_h^{(t)}, \nu_h^{(t)}) \leftarrow \texttt{OMWU}(Q_h)$



However, not easy to use in online settings...

A two-timescale single-loop approach?

Soft value iteration: for $h = H, \dots, 1$

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

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

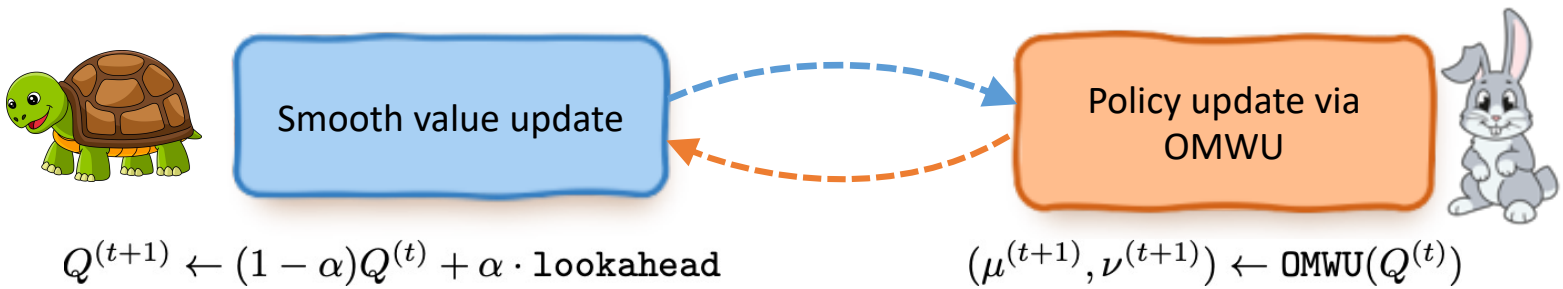
A two-timescale single-loop approach?

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

$$Q_h(s,a,b) \leftarrow r_h(s,a,b) + \cdot \mathop{\mathbb{E}}_{s' \sim P_h(\cdot|s,a,b)} \Big[\underbrace{\max_{\mu} \min_{\nu} \mu(s')^\top Q_{h+1}(s')\nu(s') + \tau\mathcal{H}(\mu(s')) - \tau\mathcal{H}(\nu(s'))}_{\text{Entropy-regularized matrix game}} \Big],$$

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

Single-loop, two-timescale approach:



Main result: episodic setting

Theorem (Cen, Chi, Du, Xiao, 2022)

The last-iterate of the two-timescale single-loop algorithm finds an ϵ-QRE in

$$\widetilde{O}\left(\frac{H^2}{\tau}\log\frac{1}{\epsilon}\right)$$

iterations, corresponding to $\widetilde{O}\left(\frac{H^3}{\epsilon}\right)$ iterations for finding an ϵ-NE.

- First last-iterate convergence result for the episodic setting.
- **Almost dimension-free:** independent of the size of the state-action space.

Main result: discounted setting

Theorem (Cen, Chi, Du, Xiao, 2022)

For the infinite-horizon γ-discounted setting, the last-iterate of the single-loop algorithm finds an ϵ-QRE in

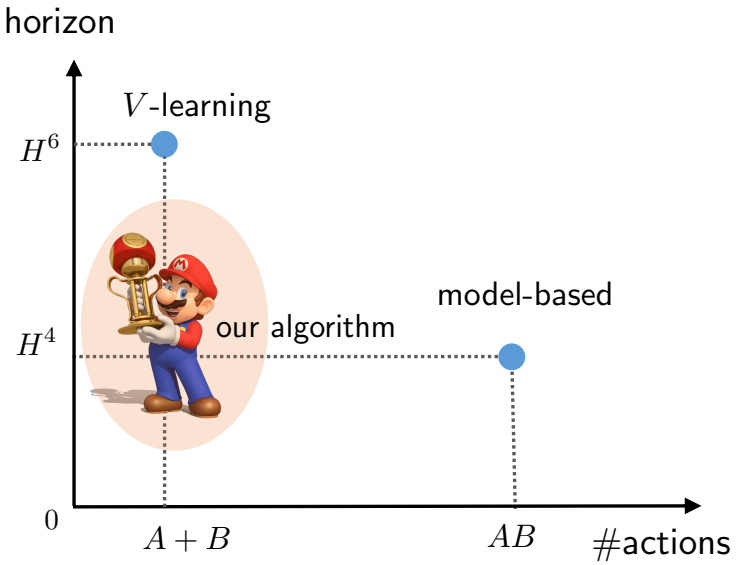
$$\widetilde{O}\left(\frac{S}{(1-\gamma)^4\tau}\log\frac{1}{\epsilon}\right)$$

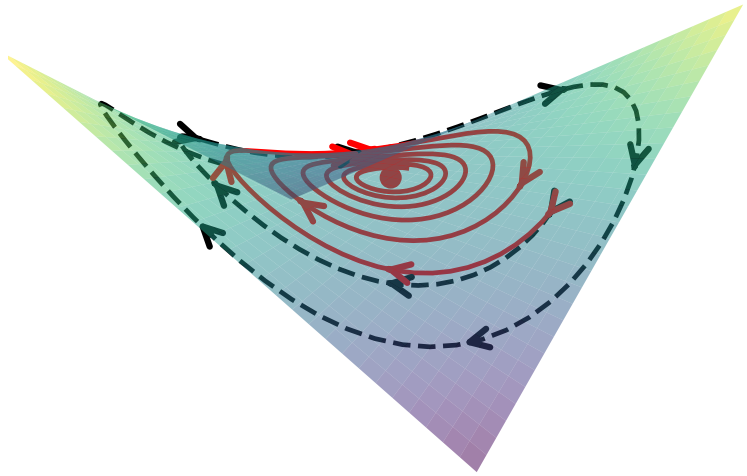
iterations, and in $\widetilde{O}\left(\frac{S}{(1-\gamma)^5\epsilon}\right)$ iterations for finding an ϵ-NE.

- This significantly improves upon the prior art $\widetilde{O}\left(\frac{S^5(A+B)^{1/2}}{(1-\gamma)^{16}c^4\epsilon^2}\right)$ of (Wei et al., 2021) and $\widetilde{O}\left(\frac{S^2\|1/\rho\|^5}{(1-\gamma)^{14}c^4\epsilon^3}\right)$ of (Zeng et al., 2022) in *all* parameter dependencies.

Concluding remarks

Concluding remarks





Understanding MARL: confluence of optimization, learning, statistics, control and game theory!

Future directions:

- function approximation
- constrained MARL
- offline RL
- many more...

References

- Minimax-optimal multi-agent RL in zero-sum Markov games with a generative model, arXiv:2208.10458, NeurIPS 2022.
- Faster Last-iterate Convergence of Policy Optimization in Zero-Sum Markov Games, arXiv:2210.01050.
- Fast policy extragradient methods for competitive games with entropy regularization, arXiv:2105.15186, NeurIPS 2021.

Thank you!

https://users.ece.cmu.edu/~yuejiec/