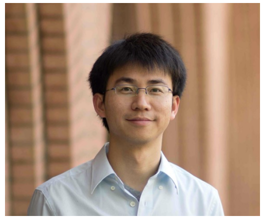
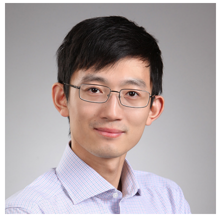
ICML 2022
International Conference On Machine Learning

Plan Better Amid Conservatism: Offline Multi-Agent Reinforcement Learning

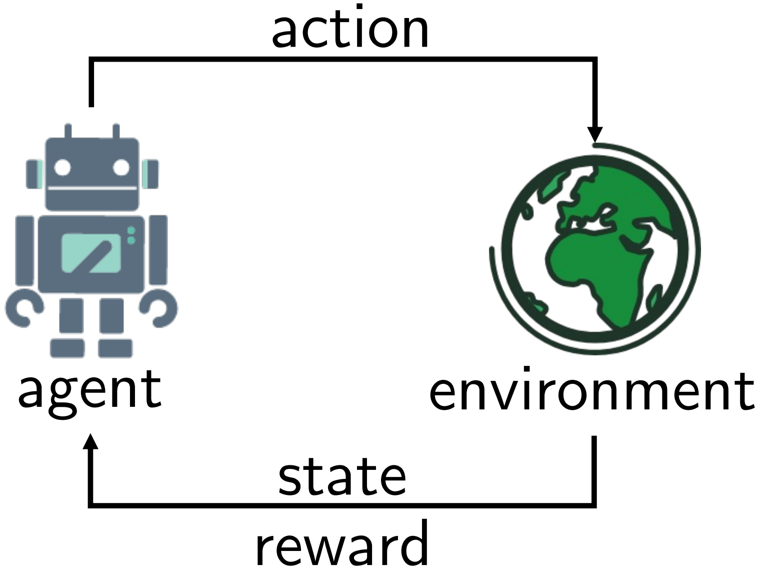
Ling Pan[1], Longbo Huang[1], Tengyu Ma[2], Huazhe Xu[2]

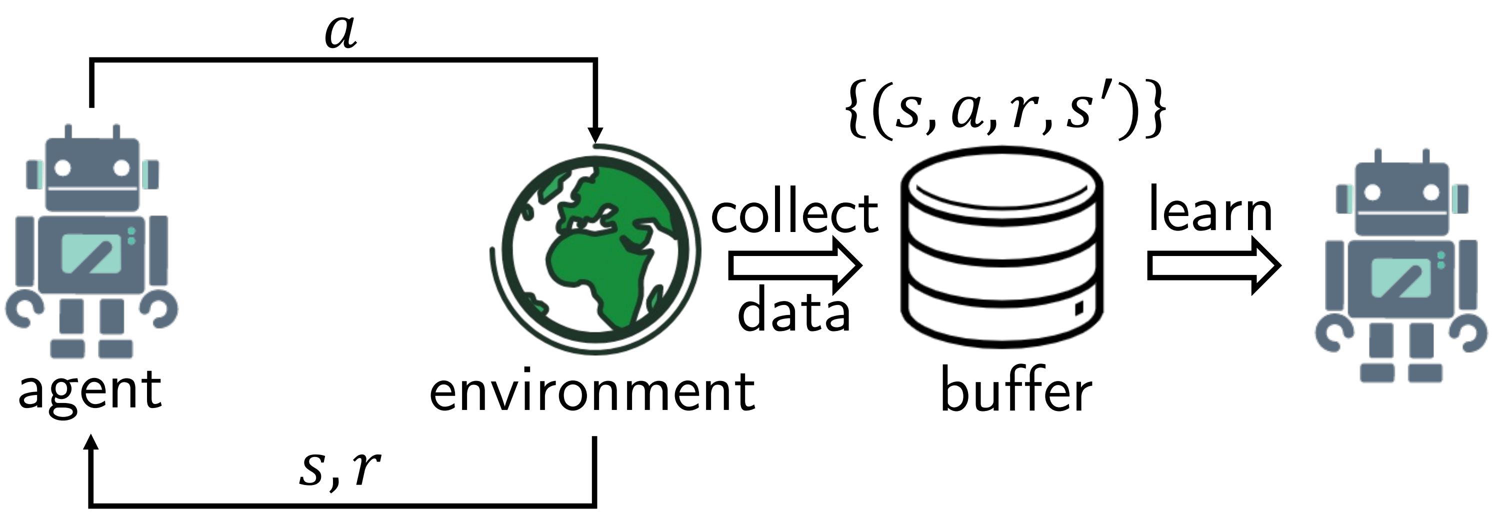
[1] Institute for Interdisciplinary Information Sciences, Tsinghua University

[2] Stanford University

Offline Reinforcement Learning



Reinforcement learning



Offline Reinforcement learning

- Key challenge
  - Distribution shift
  - Extrapolation error

Offline Reinforcement Learning

- Existing Approaches
  - Behavior regularization: TD3+Behavior Cloning (Fujimoto et al., 2021) …
    - Add a behavior cloning term to the policy update of TD3

$$\pi = \operatorname{argmax}_{\pi} \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\lambda Q\big(s, \pi(s)\big) - (\pi(s) - a)^2\right]$$

-- Largely depends on the quality of the dataset

Offline Reinforcement Learning

- Existing Approaches
  - Behavior regularization: TD3+Behavior Cloning (Fujimoto et al., 2021) …
    - Add a behavior cloning term to the policy update of TD3

$$\pi = \text{argmax}_{\pi} \mathbb{E}_{(s,a)\sim\mathcal{D}}\left[\lambda Q\big(s, \pi(s)\big) - (\pi(s) - a)^2\right]$$

-- Largely depends on the quality of the dataset

  - Critic regularization: Conservative Q-Learning (Kumar et al. 2020) …
    - Based on a conservative estimation of the Q-function

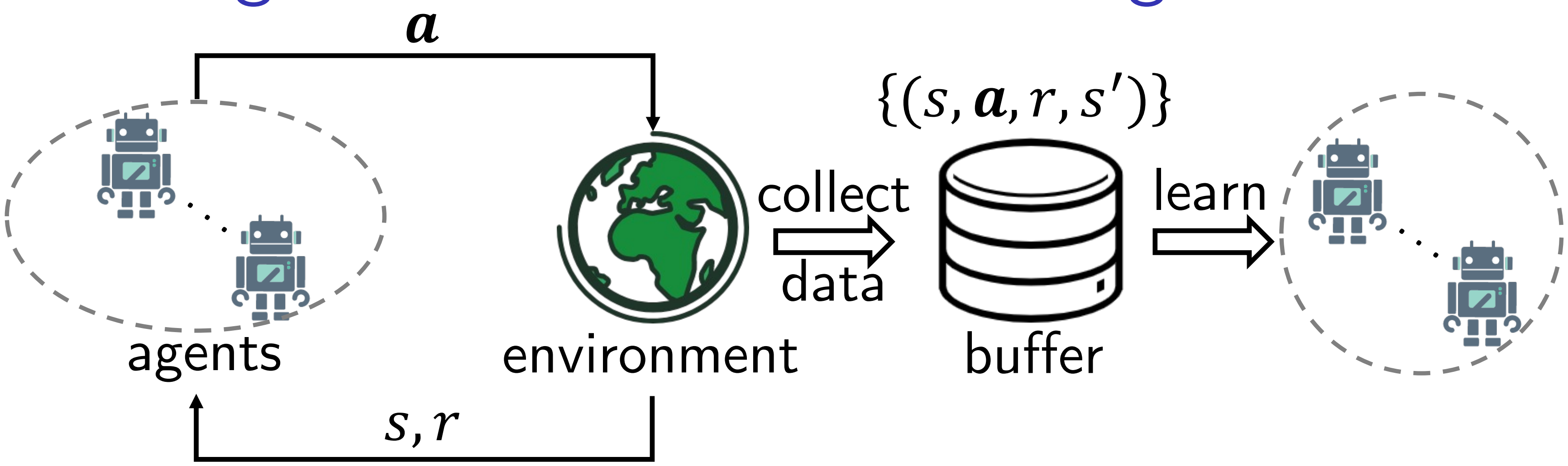
minimize Q-values of (s,a) sampled from a uniform distribution/the policy

$$\mathbb{E}_{\mathcal{D}_i}[(Q_i(o_i, a_i) - y_i)^2] + \alpha \mathbb{E}_{\mathcal{D}_i}\left[\log \sum_{a_i} \exp(Q_i(o_i, a_i)) - \mathbb{E}_{a_i \sim \hat{\pi}_{\beta_i}(a_i|o_i)}[Q_i(o_i, a_i)]\right]$$

maximize Q-values for (s,a) in the dataset to be large

-- The performance degrades dramatically with an increasing number of agents

Offline Multi-Agent Reinforcement Learning

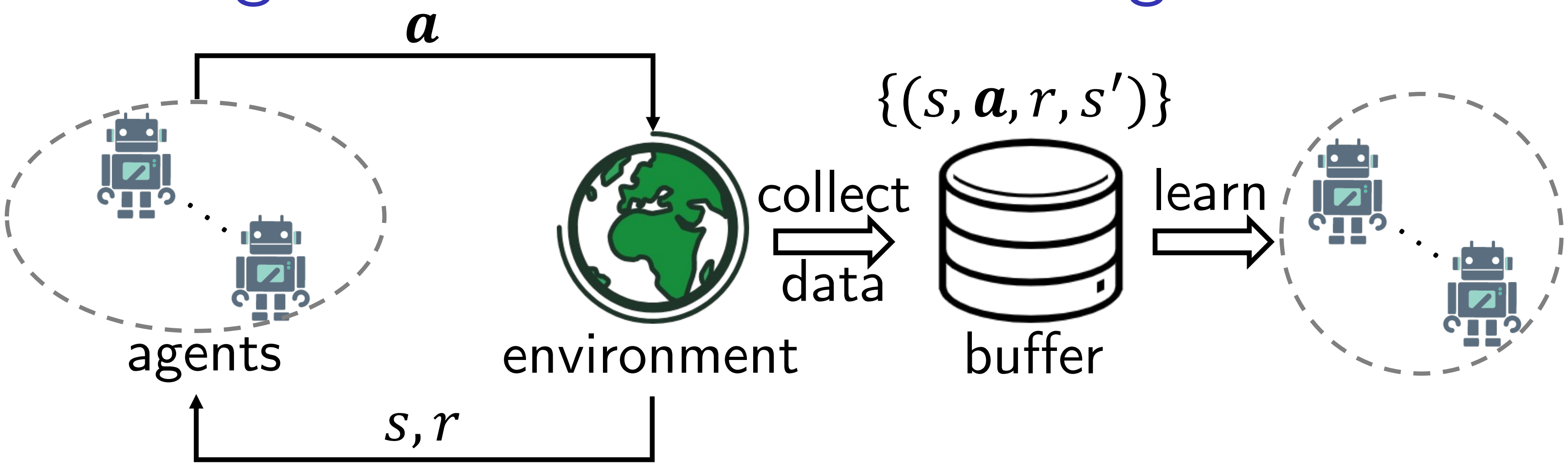


- Multi-agent actor-critic
  - Centralized value function
    - Multi-agent DDPG (MADDPG) [Lowe et al., 2017]
      - Critic i: $Q_i(s, a_1, \cdots, a_n)$

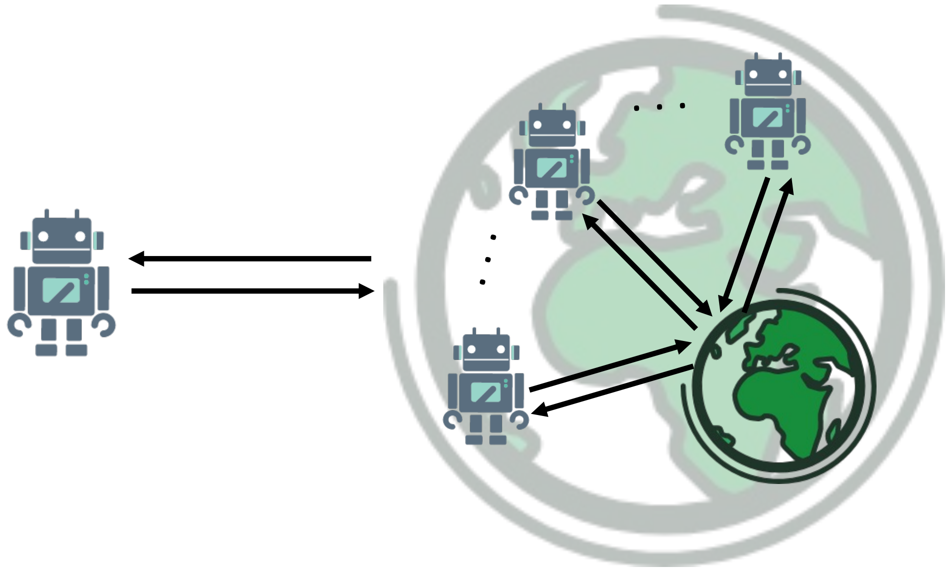
        $\mathcal{L}(\theta_i) = \mathbb{E}_{\mathcal{D}}[(Q_i(s, a_1, \cdots, a_n) - y_i)^2]$, where $y_i = r_i + \gamma \bar{Q}_i(s', a'_1, \cdots, a'_n)|_{a'_j = \bar{\pi}_j(o'_j)}$
      - Actor i: $\pi_i(o_i)$

        $\nabla_{\varphi_i} J(\pi_i) = \mathbb{E}_{\mathcal{D}}\left[\nabla_{\varphi_i} \pi_i(a_i|o_i) \nabla_{a_i} Q_i(s, a_1, \cdots, a_n)|_{a_i = \pi_i(o_i)}\right]$

Offline Multi-Agent Reinforcement Learning



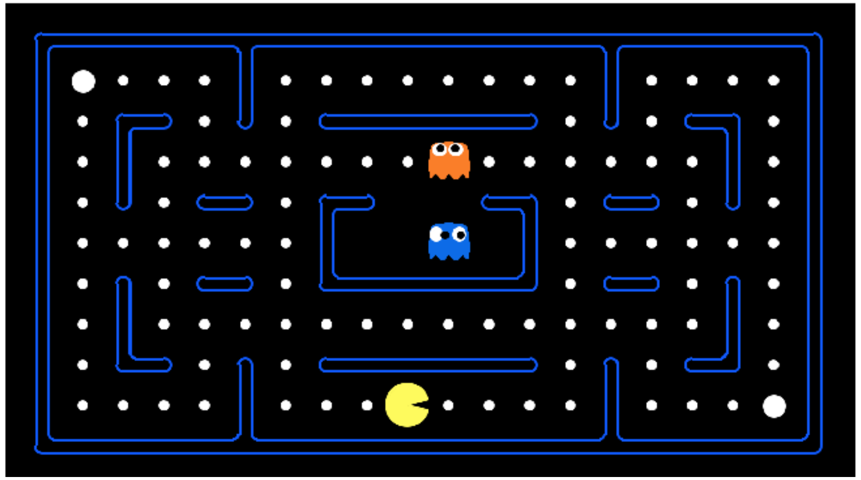
- Multi-agent actor-critic
  - Centralized value function
    - Multi-agent DDPG (MADDPG) [Lowe et al., 2017]
  - Decentralized value function
    - Independent DDPG (IDDPG) [de Witt et al., 2020]
      - Critic i: $Q_i(o_i, a_i)$
      - Actor i: $\pi_i(o_i)$



[Figure based on Jakob Foerster's talk]

From (Offline) Single-Agent RL to Multi-Agent RL

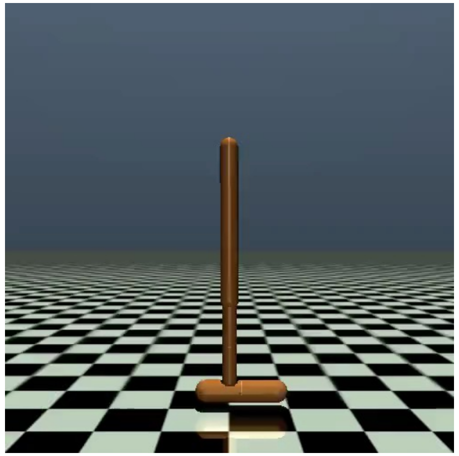
Online



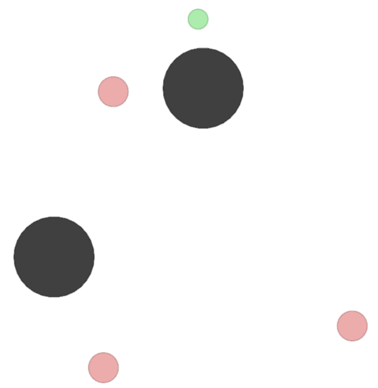
PPO

Independent PPO or Multi-Agent PPO

Offline



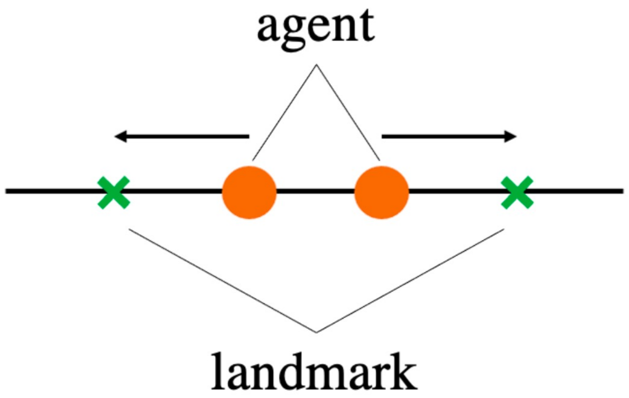
CQL



Offline Multi-Agent Reinforcement Learning

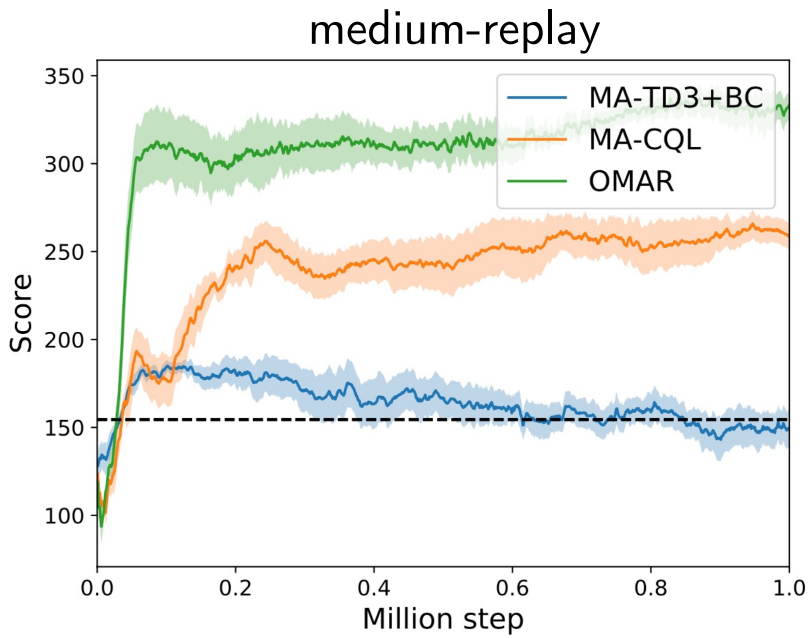
- A motivating example

Task



The Spread environment
$(n \geq 1)$

Baselines

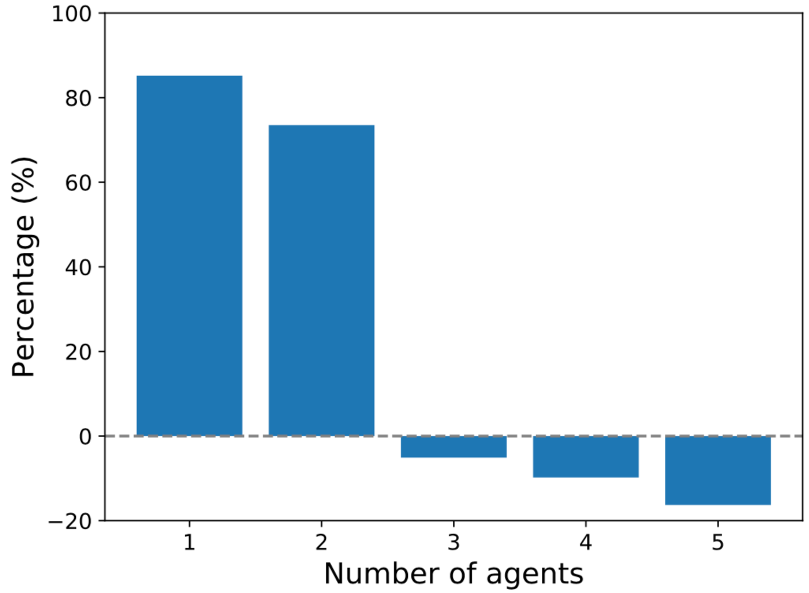


- Multi-agent setting:
  - **Cooperate** to **cover** all landmarks

- Multi-agent TD3+BC (behavior cloning)
  - Largely depends on the quality of the dataset

Offline Multi-Agent Reinforcement Learning

- A motivating example



- Multi-agent CQL

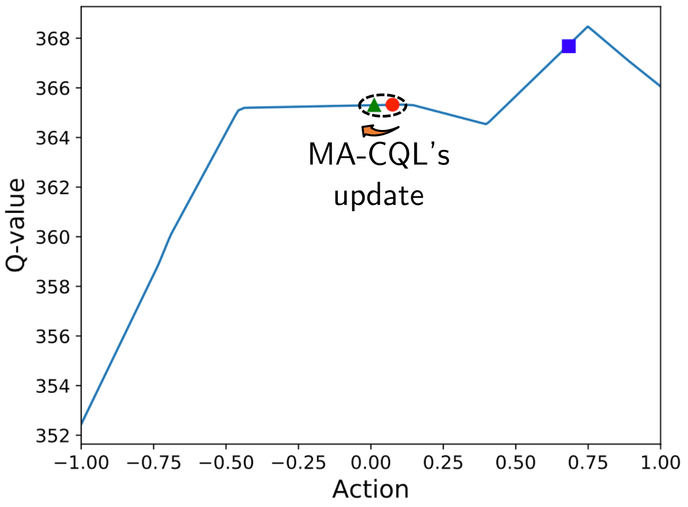
Performance improvement percentage of MA-CQL over the behavior policy with varying number of agents

$$\frac{score_{CQL} - score_{behavior}}{score_{behavior}} \times 100$$

➢ The performance of CQL degrades dramatically with an increasing number of agents.

Offline Multi-Agent Reinforcement Learning

- Key issue



● Predicted action from the MA-CQL agent
▲ Updated predicted action by MA-CQL
■ Updated predicted action by OMAR

The policy gets stuck in a bad local optimum.

- First-order policy gradient method is prone to local optima
- The agent can fail to globally optimize the conservative value function well
- Lead to suboptimal, uncoordinated learning behavior

The problem is **exacerbated** severely in the *offline* **multi-agent** setting!

Offline Multi-Agent Reinforcement Learning

- Requires ***each*** of the agent to learn a good policy for a ***successful joint policy***.

One fails to learn a good policy

⬇

Fails to cooperate with others

⬇

Leads to **uncoordinated global failure**

Offline Multi-Agent RL with Actor Rectification (OMAR)

- Idea

the action provided by the zeroth-order optimizer

$$\min \mathbb{E}_{\mathcal{D}_i}\left[(1-\tau)Q_i\big(o_i, \pi_i(o_i)\big) - \tau\left(\pi_i(o_i) - \hat{a}_i\right)^2\right]$$

Escape from bad local optima

  - Zeroth-order optimizer: $\hat{a}_i = \operatorname{argmax}_{a_i \sim \mathcal{N}} Q_i(o_i, a_i)$
  - Behavior cloning (TD3+BC): $\hat{a}_i \sim \mathcal{D}_i$

Offline Multi-Agent RL with Actor Rectification (OMAR)

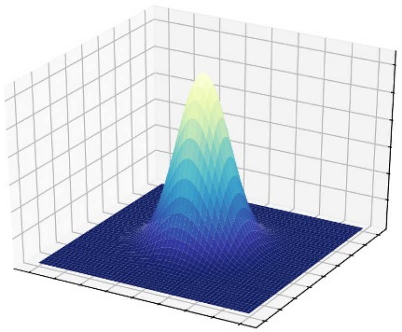
- Idea

the action provided by the zeroth-order optimizer

$$\min \mathbb{E}_{\mathcal{D}_i}\left[(1-\tau)Q_i\big(o_i, \pi_i(o_i)\big) - \tau\big(\pi_i(o_i) - \hat{a}_i\big)^2\right]$$

- Zeroth-order optimizer (evolution strategy)

For agent i



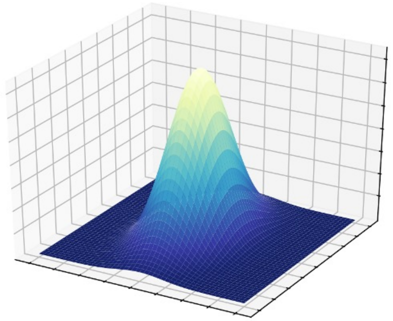
$\mathcal{N}(\mu_i, \sigma_i)$

$\sim a_i^1$ … $\sim a_i^K$

Q_i^1 … Q_i^K

$$\mu_i' \leftarrow \frac{\sum_{k=1}^{K} \exp\left(\beta Q_i^k\right) a_i^k}{\sum_{m=1}^{K} \exp\left(\beta Q_i^m\right)}$$

$$\sigma_i' \leftarrow \sqrt{\sum_{k=1}^{K}\left(a_i^k - \mu_i\right)^2}$$



$\mathcal{N}(\mu_i', \sigma_i')$

Sample K candidate actions → **Evaluate** Q-values → **Update** the sampling distribution

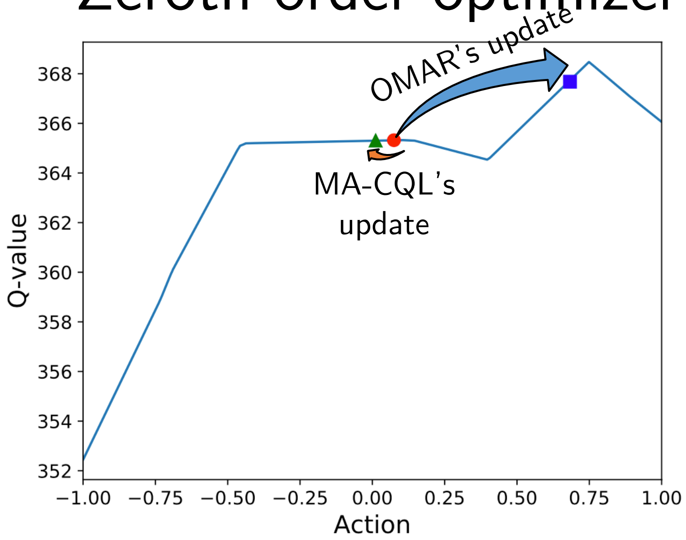
Offline Multi-Agent RL with Actor Rectification (OMAR)

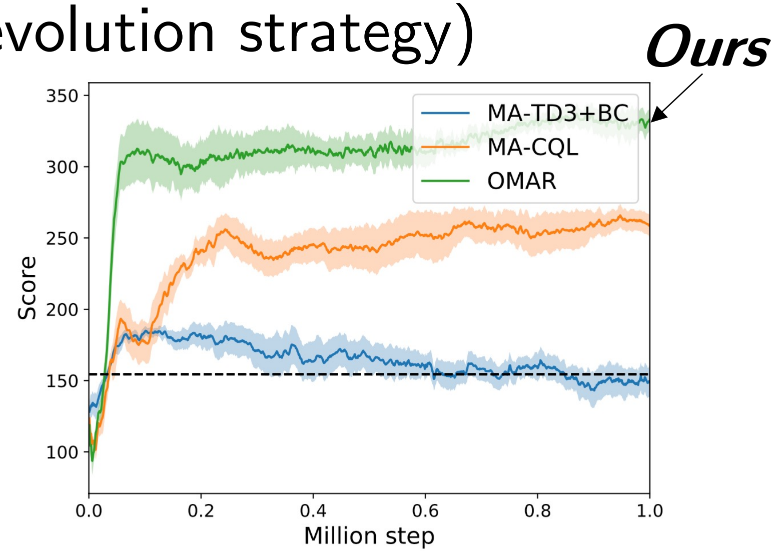
- Idea

the action provided by the zeroth-order optimizer

$$\min \mathbb{E}_{\mathcal{D}_i}\left[(1-\tau)Q_i\big(o_i,\pi_i(o_i)\big) - \tau\big(\pi_i(o_i) - \hat{a}_i\big)^2\right]$$

- Zeroth-order optimizer (evolution strategy)





- ● Predicted action from the MA-CQL agent
- ▲ Updated predicted action by MA-CQL
- ■ Updated predicted action by OMAR

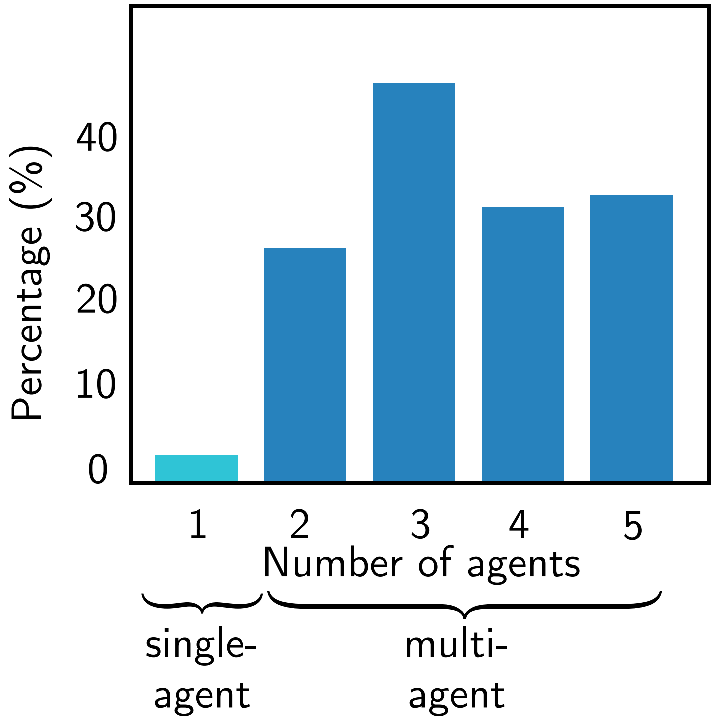
👍 Better leverage the **global** information in the critic

👍 Help the actor to **escape** from the **bad local optima**

➢ Safe policy improvement guarantee

Offline Multi-Agent RL with Actor Rectification (OMAR)

- Is OMAR effective with an increasing number of agents?

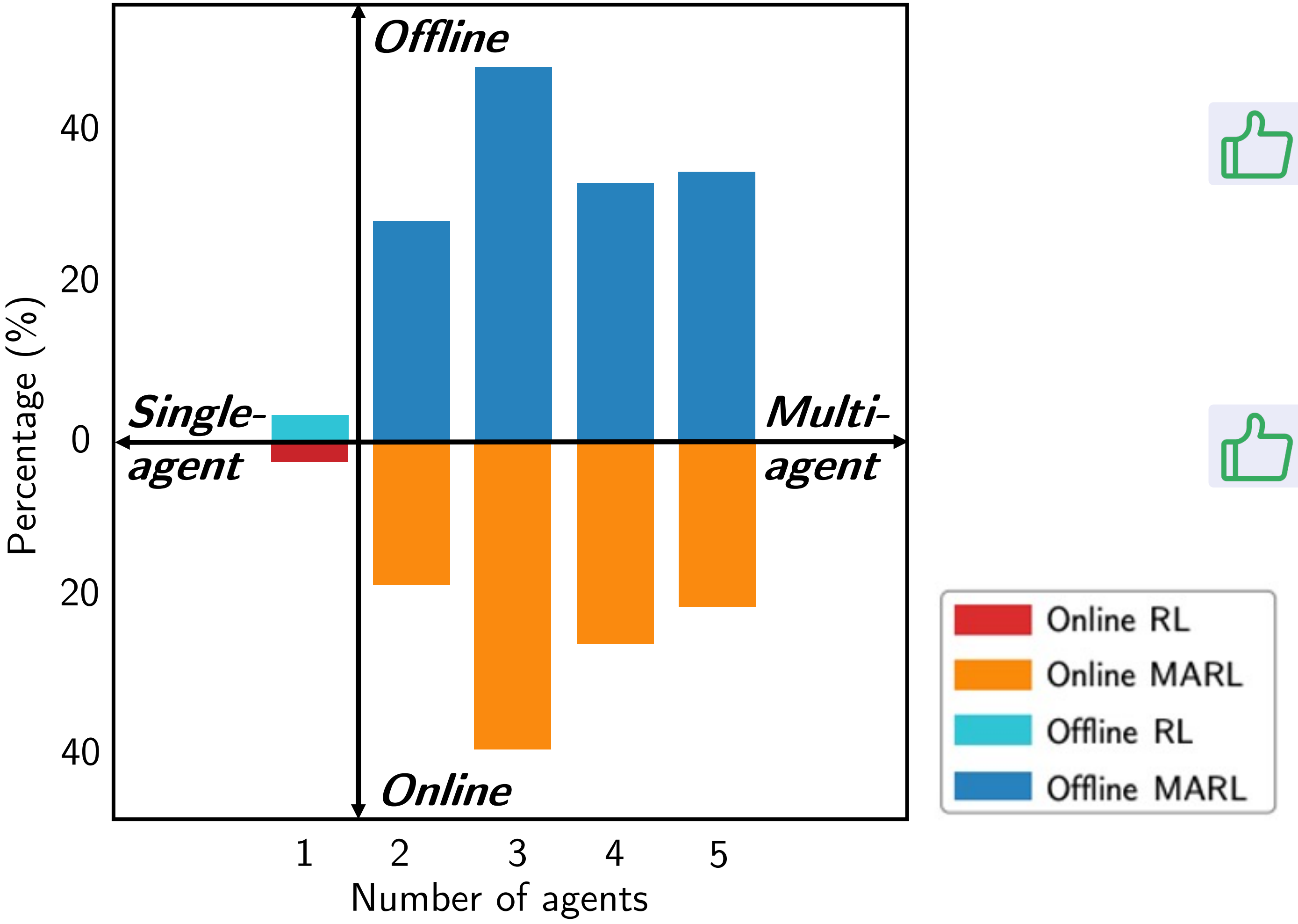


Performance improvement percentage of OMAR over MA-CQL with varying number of agents

- Multi-agent > Single-agent

Offline Multi-Agent RL with Actor Rectification (OMAR)

- Is OMAR effective in online/offline, single/multi-agent settings?



Generally applicable in all settings

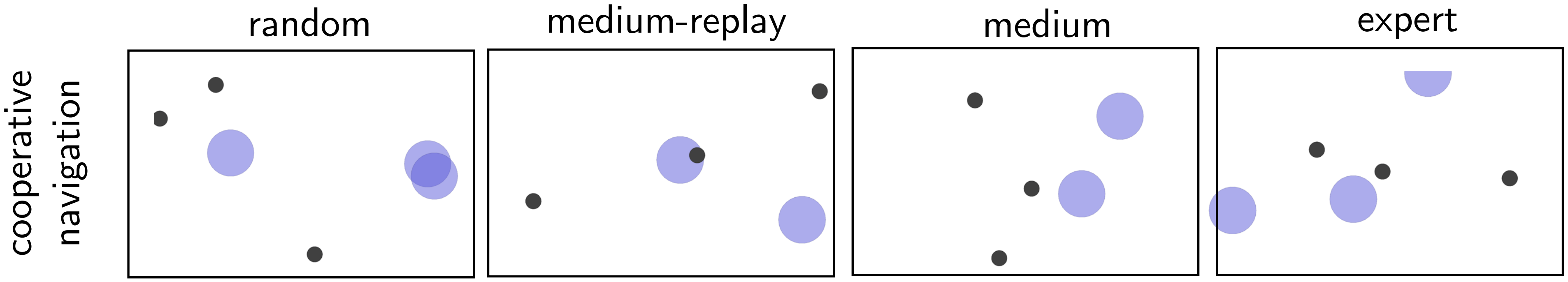
- Multi-agent > Single-agent
- Offline > Online

Most significant in the offline MARL case

Experiments

- Multi-agent particle environments

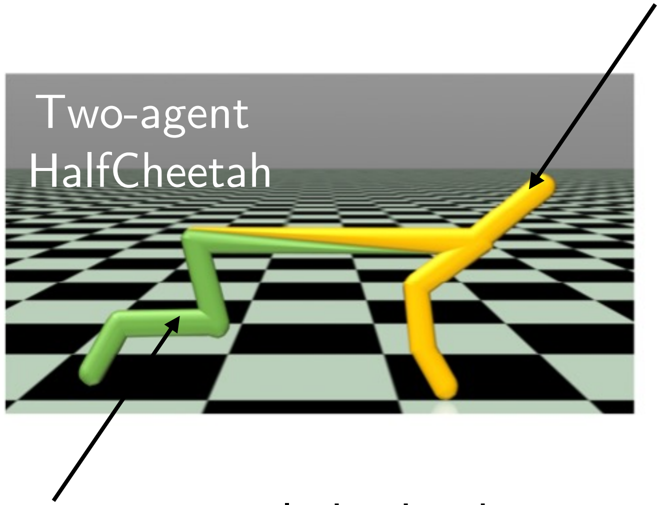
		MA-ICQ	MA-TD3+BC	MA-CQL	OMAR
Random	Cooperative navigation	6.3 ± 3.5	9.8 ± 4.9	24.0 ± 9.8	$\mathbf{34.4} \pm 5.3$
	Predator-prey	2.2 ± 2.6	5.7 ± 3.5	5.0 ± 8.2	$\mathbf{11.1} \pm 2.8$
	World	1.0 ± 3.2	2.8 ± 5.5	0.6 ± 2.0	$\mathbf{5.9} \pm 5.2$
Medium -replay	Cooperative navigation	13.6 ± 5.7	15.4 ± 5.6	20.0 ± 8.4	$\mathbf{37.9} \pm 12.3$
	Predator-prey	34.5 ± 27.8	28.7 ± 20.9	24.8 ± 17.3	$\mathbf{47.1} \pm 15.3$
	World	12.0 ± 9.1	17.4 ± 8.1	29.6 ± 13.8	$\mathbf{42.9} \pm 19.5$
Medium	Cooperative navigation	29.3 ± 5.5	29.3 ± 4.8	34.1 ± 7.2	$\mathbf{47.9} \pm 18.9$
	Predator-prey	63.3 ± 20.0	65.1 ± 29.5	61.7 ± 23.1	$\mathbf{66.7} \pm 23.2$
	World	71.9 ± 20.0	73.4 ± 9.3	58.6 ± 11.2	$\mathbf{74.6} \pm 11.5$
Expert	Cooperative navigation	104.0 ± 3.4	108.3 ± 3.3	98.2 ± 5.2	$\mathbf{114.9} \pm 2.6$
	Predator-prey	113.0 ± 14.4	115.2 ± 12.5	93.9 ± 14.0	$\mathbf{116.2} \pm 19.8$
	World	109.5 ± 22.8	110.3 ± 21.3	71.9 ± 28.1	$\mathbf{110.4} \pm 25.7$



Experiments

- Multi-agent MuJoCo

agent 1: control the front joints

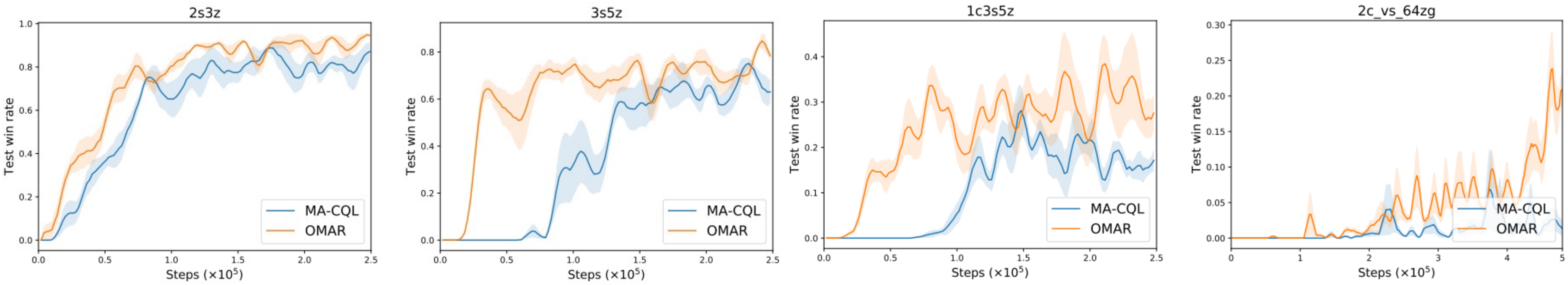


agent 2: control the back joints

	Random	Medium-reply	Medium	Expert
MA-ICQ	7.4 ± 0.0	35.6 ± 2.7	73.6 ± 5.0	110.6 ± 3.3
MA-TD3+BC	7.4 ± 0.0	27.1 ± 5.5	75.5 ± 3.7	**114.4** ± 3.8
MA-CQL	7.4 ± 0.0	41.2 ± 10.1	50.4 ± 10.8	64.2 ± 24.9
OMAR	**15.4** ± 12.3	**57.7** ± 5.1	**80.7** ± 10.2	**113.5** ± 4.3

Experiments

- StarCraft II Micromanagement Benchmark

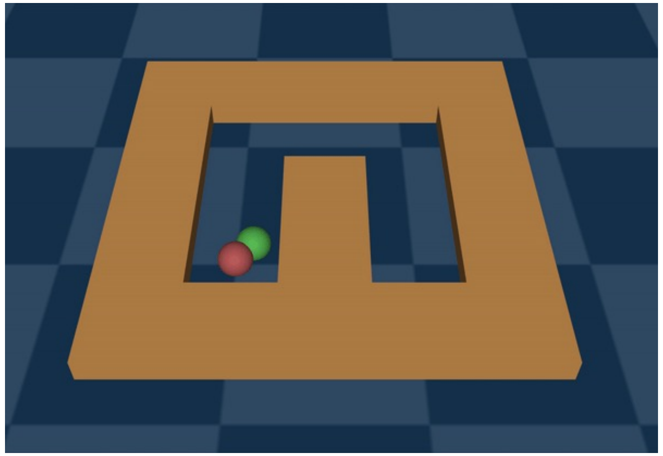


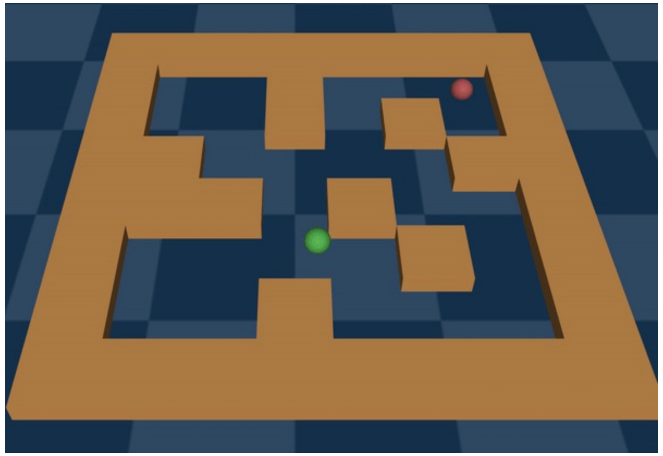
The average performance gain of OMAR over MA-CQL is 76.7%.

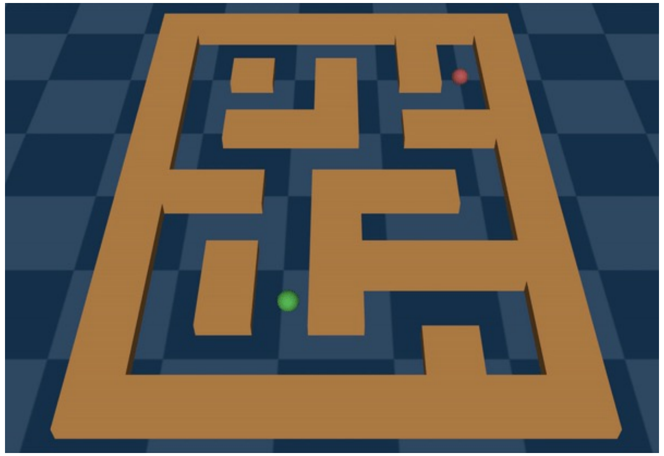
Experiments

- D4RL

	umaze	medium	large
TD3+BC	41.1 ± 4.9	75.5 ± 27.1	103.9 ± 31.4
ICQ	4.8 ± 3.8	13.0 ± 7.9	9.2 ± 20.0
CQL	109.8 ± 23.9	106.4 ± 11.0	94.6 ± 44.6
OMAR	$\mathbf{124.7} \pm 7.6$	$\mathbf{125.7} \pm 12.3$	$\mathbf{157.7} \pm 12.3$







OMAR is compatible for single-agent control.

Thank you!

Q & A