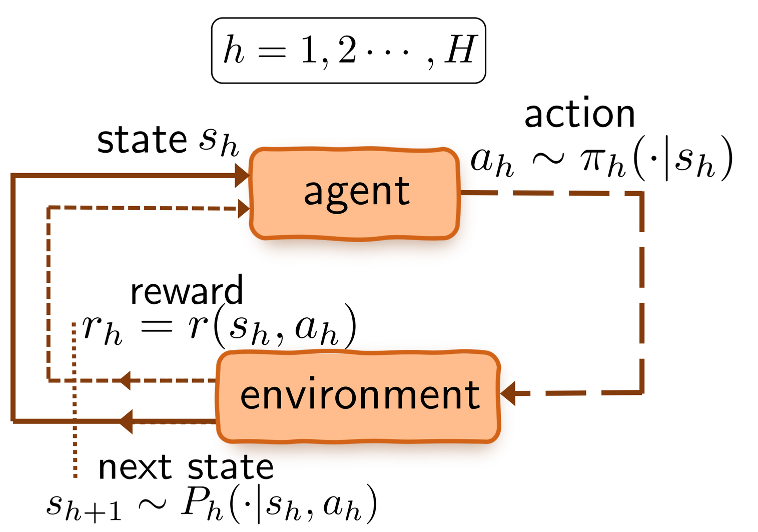
Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning

Gen Li Laixi Shi Yuxin Chen Yuantao Gu Yuejie Chi
Princeton CMU Princeton Tsinghua CMU

Reinforcement learning

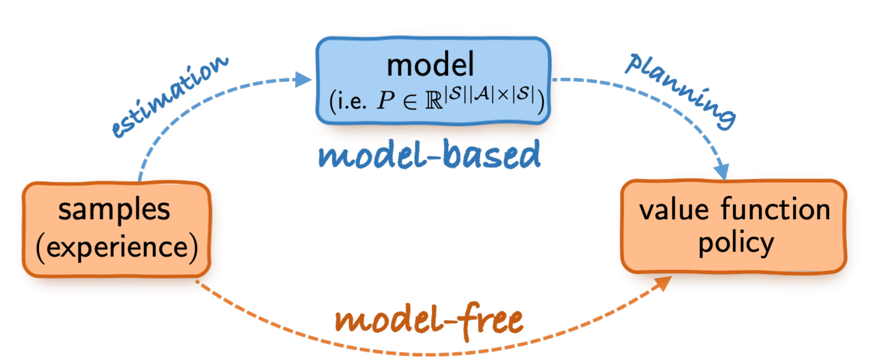
- **Episodic Markov decision process (MDP)**



- **Goal:** find the optimal policy $\pi^\star$ maximizing value function:

$$\forall s \in \mathcal{S}: \quad V_1^\pi(s) := \mathbb{E}\left[\sum_{h=1}^{H} r_h(s_h, a_h) \,\Big|\, s_1 = s\right]$$

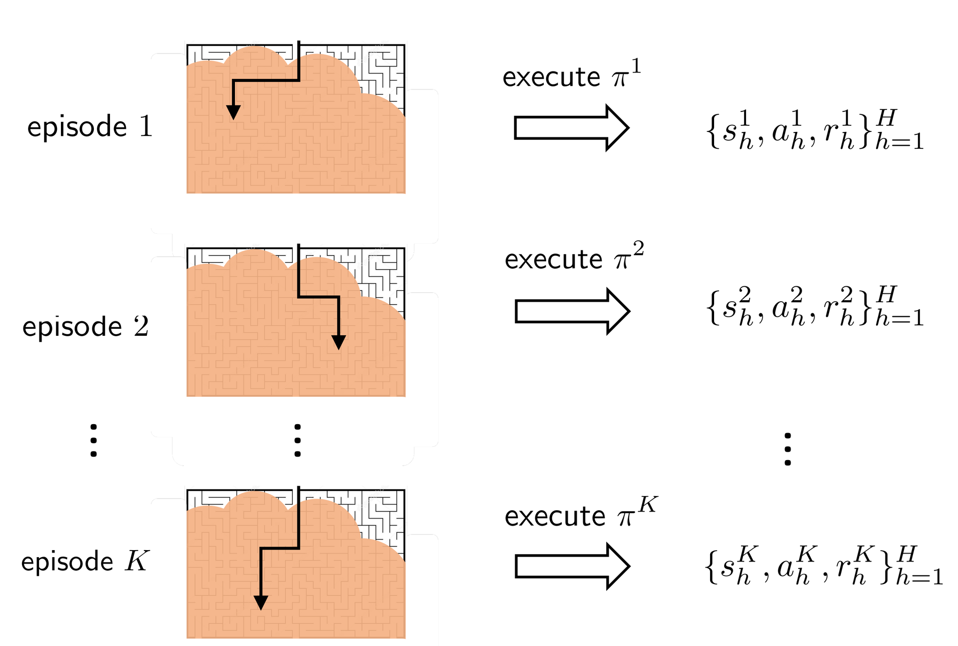
Model-based vs. model-free RL



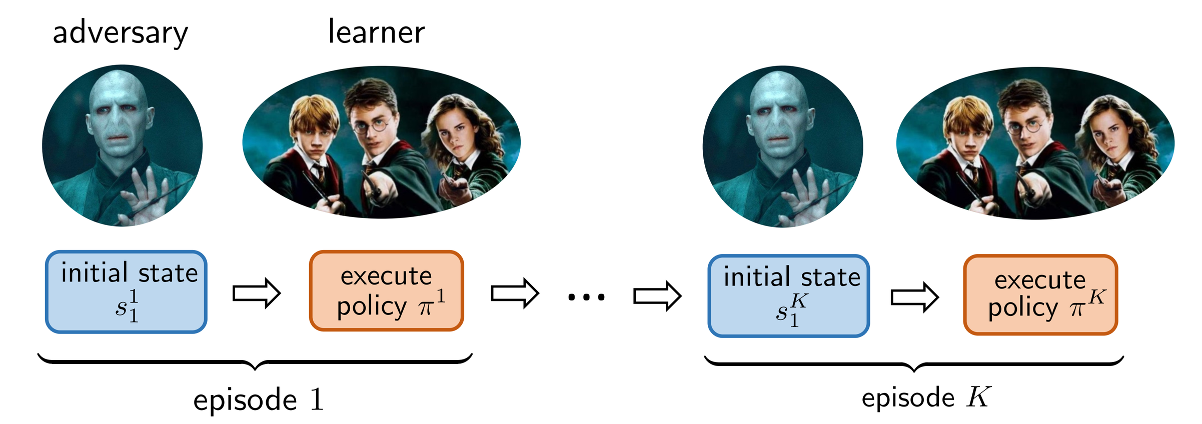
- **Model-based approach ("plug-in")**
  store transition kernel estimates $\to$ $O(S^2AH)$ *memory*
- **Model-free approach**
  maintain Q-estimates $\to$ $O(SAH)$ *memory*

Online RL

Sequentially execute MDP for K episodes, each consisting of H steps



Regret



Performance metric: given $\underbrace{\text{initial states } \{s_1^k\}_{k=1}^K}_{\text{chosen by nature/adversary}}$, define

$$\text{Regret}(\underbrace{T}_{\text{sample size: } KH}) := \sum_{k=1}^{K} \left(V_1^\star(s_1^k) - V_1^{\pi^k}(s_1^k)\right)$$

Lower bound (Domingues et al. '21)

$$\text{Regret}(T) \gtrsim \sqrt{H^2SAT}$$

Main result

Theorem: With high prob., Q-EarlySettled-Advantage achieves (up to log factor)

$$\text{Regret}(T) \lesssim \sqrt{H^2SAT} + H^6SA$$

with a memory complexity of $O(SAH)$

- regret-optimal with near-minimal burn-in cost $O(SA\text{poly}(H))$
- memory-efficient $O(SAH)$
- computationally efficient: runtime $O(T)$

Prior works

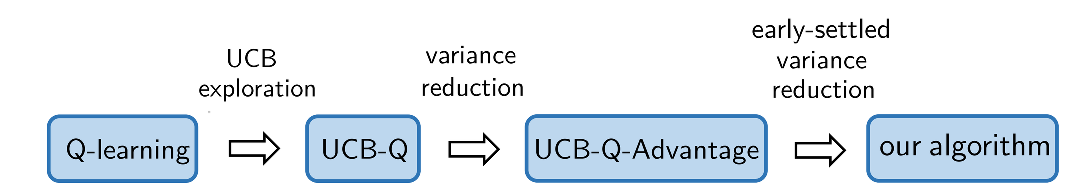
Algorithm	Regret	Memory
UCB-VI (Azar et al., 2017)	$\sqrt{H^2SAT} + H^4S^2A$	S^2AH
UCB-M-Q (Menard et al., 2021)	$\sqrt{H^2SAT} + H^4SA$	S^2AH
UCB-Q-Advantage (Zhang et al., 2020)	$\sqrt{H^2SAT} + H^8S^2A^{\frac{3}{2}}T^{\frac{1}{4}}$	SAH

Issues: (1) large burn-in cost; (2) large memory complexity

"Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning," G. Li, L. Shi, Y. Chen, Y. Gu, Y. Chi, arXiv:2110.04645, NeurIPS 2021

This work is supported in part by NSF, ONR, AFOSR, and ARO.

A glimpse of our algorithm design



Q-learning with UCB exploration (Jin et al., 2018)

$$Q_h(s_h, a_h) \leftarrow \underbrace{(1-\eta_k)Q_h(s_h, a_h) + \eta_k \mathcal{T}_k\left(Q_{h+1}\right)(s_h, a_h)}_{\text{classical Q-learning}} + \eta_k \underbrace{b_h(s_h, a_h)}_{\text{exploration bonus}}$$

- $b_h(s, a)$: upper confidence bound
  — *optimism in the face of uncertainty*
- inspired by UCB bandit algorithm (Lai, Robbins '85)

Issue: Regret$(T) \lesssim \sqrt{H^3SAT}$ $\Longrightarrow$ sub-optimal by a factor of $\sqrt{H}$

Reference-advantage decomposition (Zhang et al. '20)

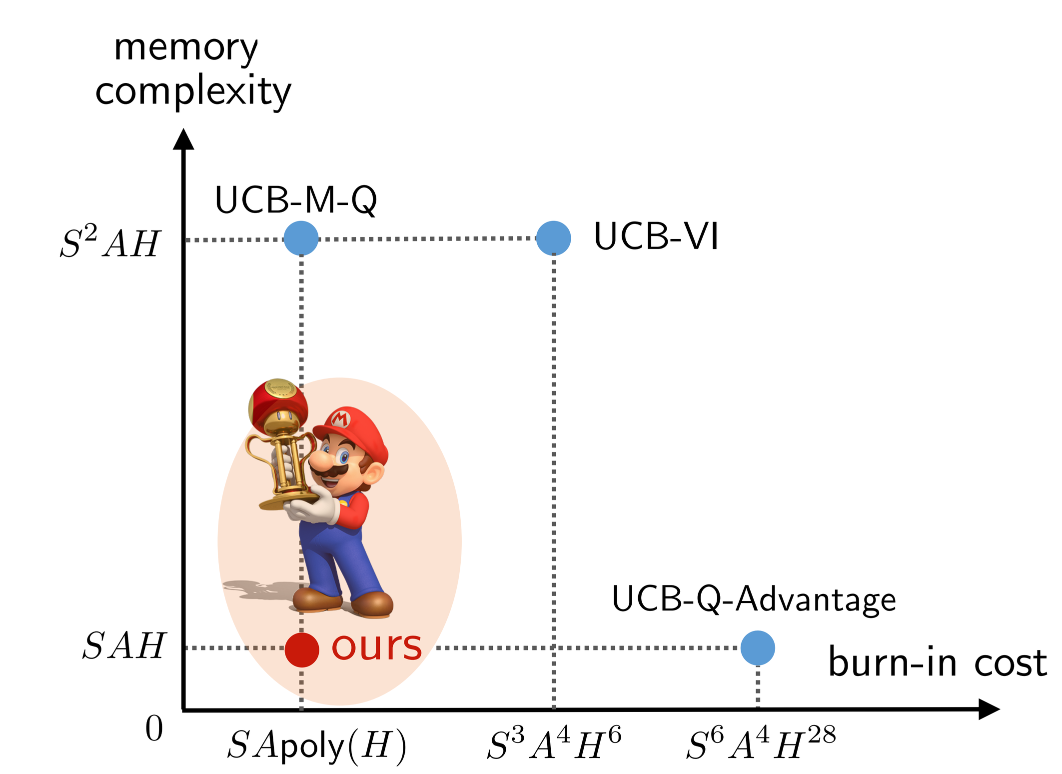
$$Q_h(s_h, a_h) \leftarrow (1-\eta_k)Q_h(s_h, a_h) + \eta_k \underbrace{b_h(s_h, a_h)}_{\text{UCB bonus}} + \eta_k\Big(\underbrace{\mathcal{T}_k(Q_{h+1}) - \mathcal{T}_k(\overline{Q}_{h+1})}_{\text{advantage}} + \underbrace{\widehat{\mathcal{T}}(\overline{Q}_{h+1})}_{\text{reference}}\Big)(s_h, a_h)$$

- Reference $\overline{Q}_h$, batch estimate $\widehat{\mathcal{T}}(\overline{Q}_{h+1})$: help reduce variability

Issue: *high burn-in cost* $O(S^6A^4H^{28})$

Q-EarlySettled-Advantage: maintains auxiliary sequences V_h^{UCB} & V_h^{LCB} to help settle the reference early

Concluding remarks



Model-free algorithms can simultaneously achieve

(1) regret optimality; (2) low burn-in cost; (3) memory efficiency