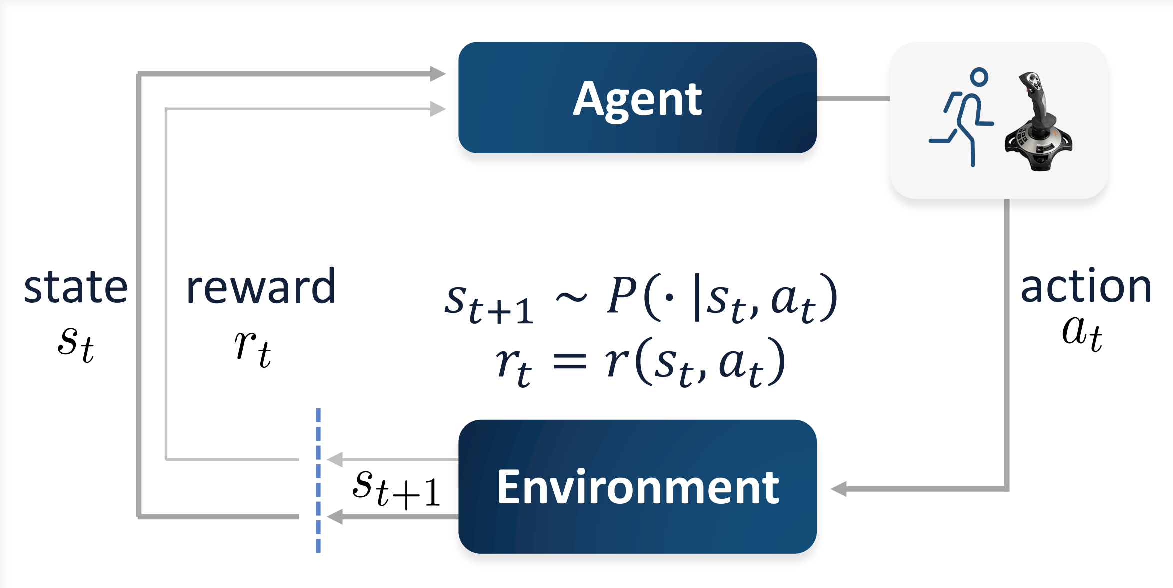
Model-Free Reinforcement Learning: from Clipped Pseudo-Regret to Sample Complexity

Zihan Zhang   Yuan Zhou   Xiangyang Ji

2021/7/21

Discounted MDP



Infinite horizon
with discounted factor $\gamma < 1$

A policy π :
π: States(S) $\rightarrow$ Actions (A), $a = \pi(s)$

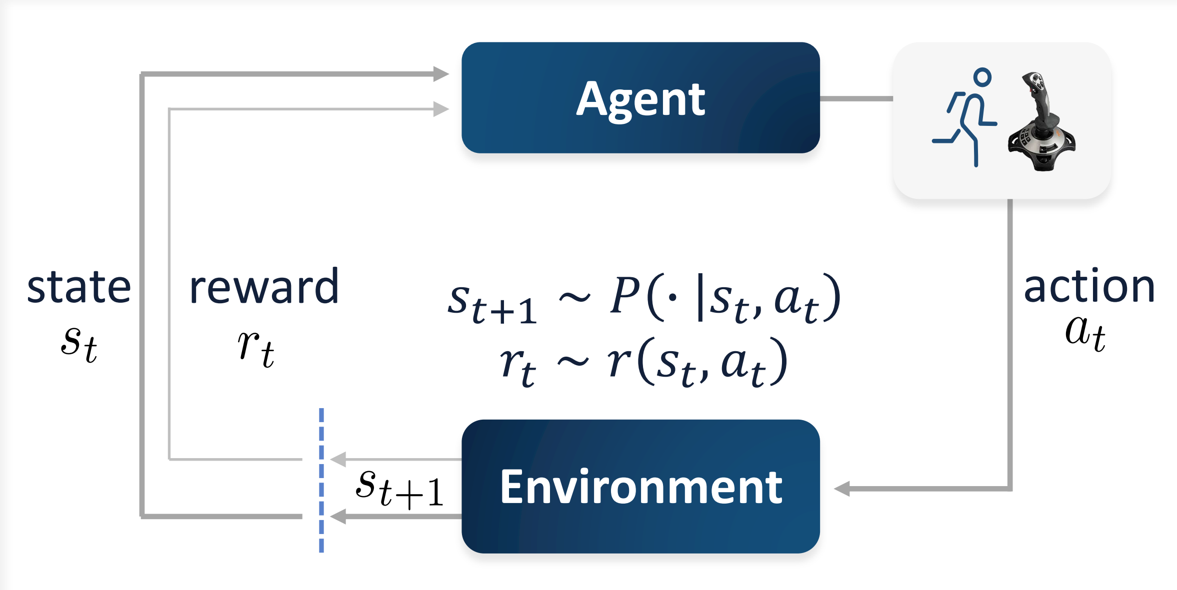
Goal: maximize value function

$$V^{\pi}(s) = \mathbb{E}\left[\sum_{t\geq 0} \gamma^t r_{t+1} \,\middle|\, s_1 = s, \pi\right]$$

$$Q^{\pi}(s,a) = \mathbb{E}\left[\sum_{t\geq 0} \gamma^t r_{t+1} \,\middle|\, s_1 = s, a_1 = a, \pi\right]$$

$V^*(Q^*) = V^{\pi^*}(Q^{\pi^*})$: value ($Q$) function of opt policy

ϵ-Sample Complexity



Given $\epsilon \in (0, \frac{1}{1-\gamma})$, ϵ -sample complexity is the number of steps when an ϵ-suboptimal policy is executed:

$$\sum_{t \geq 1} \mathbf{I}[V^{\pi_t}(s_t) < V^*(s_t) - \epsilon]$$

- the number of trials needed to learn an ϵ-optimal policy

Main Result

- **Theorem 1:** With variance reduction: For any $\epsilon \in \left(0, \frac{(1-\gamma)^{14}}{S^2 A^2}\right]$ and $\delta > 0$, with probability $1-\delta$, the (ϵ, p)-sample complexity of UCB-Multistage is bounded by

$$\tilde{O}\left(\frac{SA\ln(1/\delta)}{\epsilon^2(1-\gamma)^3}\right)$$

- **Theorem 2**: Without variance reduction: For any $\epsilon \in \left(0, \frac{1}{1-\gamma}\right]$ and $\delta > 0$, with probability $1-\delta$, the (ϵ, p)-sample complexity of UCB-Multistage is bounded by

$$\tilde{O}\left(\frac{SA\ln(1/\delta)}{\epsilon^2(1-\gamma)^{5.5}}\right)$$

Existing Results

	Algorithm	Sample Complexity	Space Complexity
Model-based	R-max [Kakade,2003]	$\tilde{O}(S^2 A \ln(1/\delta)\,\epsilon^{-3}(1-\gamma)^{-6})$	$O(S^2 A)$
	MoRmax [Szita & Szepesvari 2010]	$\tilde{O}(SA \ln(1/\delta)\,\epsilon^{-2}(1-\gamma)^{-6})$	
	UCRL-γ [Lattimore & Hutter, 2012]	$\tilde{O}(S^2 A \ln(1/\delta)\,\epsilon^{-2}(1-\gamma)^{-3})$	
Model-free	Infinite Q-learning with UCB [Dong et al., 2019]	$\tilde{O}(SA \ln(1/\delta)\,\epsilon^{-2}(1-\gamma)^{-7})$	$O(SA)$
	UCB-Multistage-Advantage (our result)	$\tilde{O}(SA \ln(1/\delta)\,\epsilon^{-2}(1-\gamma)^{-3})$ for $\epsilon < (SA)^{-2}(1-\gamma)^{14}$	
	UCB-Multistage (our result)	$\tilde{O}(SA \ln(1/\delta)\,\epsilon^{-2}(1-\gamma)^{-5.5})$	
	Delayed Q-learning [Strehl et al., 2006]	$\tilde{O}(SA \ln(1/\delta)\,\epsilon^{-4}(1-\gamma)^{-8})$	
	Median-PAC (Pazis et al., 2016)	$\tilde{O}(SA \ln(1/\delta)\,\epsilon^{-2}(1-\gamma)^{-4})$	$O(SA\epsilon^{-2}(1-\gamma)^{-4})$
Lower bound	[Lattiore & Hutter, 2012]	$\Omega(SA\epsilon^{-2}(1-\gamma)^{-3})$	

All bounds are in Big-O / Big-Omega and ignore logarithmic factors.

Pseudo-Regret

- Pseudo-regret vector: $\phi_t(s) = V_t(s) - r\big(s, \pi_t(s)\big) - \gamma P_{s,\pi_t(s)} V_t$
- Assuming V_t is always optimistic, i.e., $V_t \geq V^*$

$$V^*(s_t) - V^{\pi_t}(s_t) \leq V_t(s_t) - V^{\pi_t}(s_t) = \gamma P_{\pi_t}(V_t - V^{\pi_t}) + \phi_t = \sum_{i=0}^{\infty} (\gamma P_{\pi_t})^i \phi_t$$

- $V^*(s_t) - V^{\pi_t}(s_t) > \epsilon$ implies that $\mathbf{1}_{s_t} \sum_{i=0}^{\infty} (\gamma P_{\pi_t})^i \phi_t > \epsilon$

- Assuming $\pi_{t+i} = \pi_t$ for $1 \leq i \leq H \coloneqq \max\left\{\frac{\ln(8/((1-\gamma)\epsilon))}{\ln(1/\gamma)}, \frac{1}{1-\gamma}\right\}$

$$\mathbf{1}_{s_t} \sum_{i=0}^{\infty} (\gamma P_{\pi_t})^i \phi_t \leq \mathrm{E}\left[\sum_{i=0}^{H-1} \gamma^i \phi_t(s_{t+i})\right] + \frac{\epsilon}{8}$$

V_t: the value function at time t

P_{π_t}: the transition matrix of π_t

$\mathbf{1}_s$: $[0,0,\dots,1,\dots,0]^T$ (1 is at the s-th coordinate)

Pseudo-Regret

- $V^*(s_t) - V^{\pi_t}(s_t) > \epsilon$ implies $\sum_{i=0}^{H-1} \gamma^i \phi_t(s_{t+i}) > \frac{7\epsilon}{8}$ in expectation

- $\sum_{i=0}^{H-1} \gamma^i \phi_t(s_{t+i}) > \frac{7\epsilon}{8}$ implies $\sum_{i=0}^{H-1} \gamma^i \operatorname{clip}(\phi_t(s_{t+i}), \frac{\epsilon}{8}) \geq \frac{3\epsilon}{4}$

- With concentration inequalities in hand, it suffices to bound

$$\sum_{t \geq 1} \sum_{i=0}^{H-1} \gamma^i \operatorname{clip}(\phi_t(s_{t+i}), \frac{\epsilon}{8}) \approx H \sum_{t \geq 1} \operatorname{clip}(\phi_t(s_t), \frac{\epsilon}{8})$$

- The sample complexity is then bounded by

$$\frac{4H}{3\epsilon} \sum_{t \geq 1} \operatorname{clip}(\phi_t(s_t), \frac{\epsilon}{8})$$

$\operatorname{clip}(x, y) := x \mathbb{I}[x \geq y]$

Stage-Based Framework

- Let $e_1 = H, e_{i+1} = \lfloor (1+1/H)e_i \rfloor$; $L = \left\{\sum_{i=1}^{j} e_i \,\middle|\, j \geq 1\right\}$: the grid marking the end of the stages
- Algorithm only updates $Q_t(s,a), V_t(s)$ when $n_h(s,a) \in L$

For episode t $= 1, 2, 3, \ldots$:

$\pi_t \leftarrow$ greedy policy according to Q; execute π_t

For h = 1, 2, 3, …, H: If $n_t(s_t, a_t) \in L$ then

$(s,a) \leftarrow (s_t, a_t)$

$Q_t(s,a) \leftarrow \min\left\{r_h(s,a) + \frac{1}{\check{n}_t(s,a)} \sum_{\ell \in \check{n}_t(s,a)} V_l(s_{l+1}) + b_t(s,a), Q_{t-1}(s,a)\right\}$

$b_t(s,a) = \tilde{\Theta}\left(H \cdot \check{n}_t(s,a)^{-1/2}\right)$

$V_t(s) \leftarrow \max_a Q_t(s,a)$

$\check{n}_t(s,a)$: set of the episodes of the latest *completed* stage for (s, a) by step t (or the size of the set)

V_t, Q_t: the V, Q vectors at the beginning of step t

The stage-based framework is used in our previous work [Zhang, Zhou and Ji, 2020]

Multi-Stage Learning

By the update rule $\phi_t(s_t) \leq 2b_t(s_t, a_t) + \gamma P_{s_t,a_t}(V_{\rho_t(s_t,a_t))} - V_t)$

bonus term | gap of value function

Observation

- Limitation of model-free learning: only can remember recent value function;
- Stage-based update: regret depends on the difference between the **remembered** value function and **current** value function;

Solution

- Accelerated updates: reduce the gap of time → reduce the gap of value function

$\rho_t(s, a)$: the time the last stage of (s, a) starts.

Thank You