Sharp Variance-Dependent Bounds in Reinforcement Learning: Best of Both Worlds in Stochastic and Deterministic Environments

Runlong Zhou[1] Zihan Zhang Simon S. Du[1]

[1]University of Washington

Preliminaries

Markov Decision Processes

- S states
- A actions
- Planning horizon H
- Reward function $R_h(s,a) \in \Delta([0,1])$
- Transition probability $P_h(s'|s,a)$
- Maximum transition support $\Gamma = \max_{h,s,a} \|P_h(\cdot|s,a)\|_0$

Policy and value functions

- $\pi = \{\pi_h\}_{h\in[H]}$ where $\pi_h : \mathcal{S} \to \mathcal{A}$, optimal $\pi^\star$
- Value and Q functions

$$V_h^\pi(s) := \mathbb{E}_\pi\left[\sum_{t=h}^{H} r_t \,\middle|\, s_h = s\right],$$

$$Q_h^\pi(s,a) := \mathbb{E}_\pi\left[\sum_{t=h}^{H} r_t \,\middle|\, (s_h,a_h) = (s,a)\right].$$

Episodic reinforcement learning

- K episodes, with policies $\pi^1, \pi^2, \ldots, \pi^k$
- Performance measure

$$\text{Regret}(K) := \sum_{k=1}^{K} (V_1^\star(s_1^k) - V_1^{\pi^k}(s_1^k)).$$

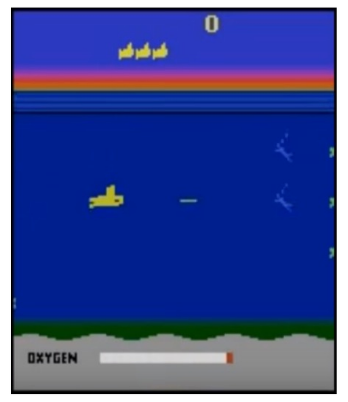
Conditions for MDPs

Condition 1: For any possible trajectory, its total reward in a single episode is upper-bounded by 1.

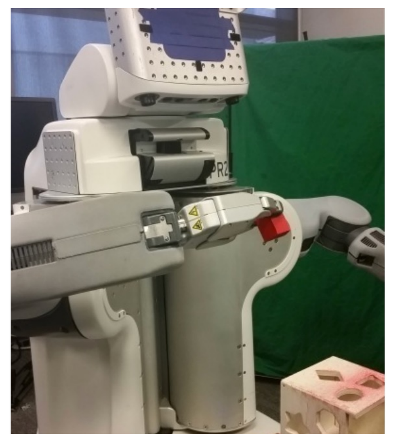
Condition 2: The MDP is time-homogeneous, i.e., the transition and reward are both independent on the timestep h.

Motivation

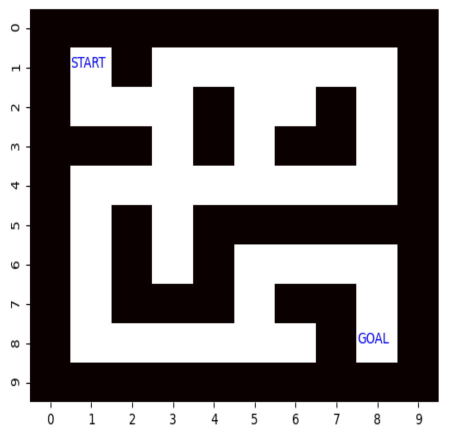
RL environments have different randomness:



Games with random environments



Robotics



Maze

Stochastic → **Deterministic**

High variance → **Zero variance**

Can we design an algorithm which automatically exploits randomness?

Total Multi-Step Conditional Variance

For trajectory $\tau = \{s_h, a_h\}_{h\in[H]}$, define

$$\text{Var}_\tau^\Sigma := \sum_{h=1}^{H} (\mathbb{V}(R_h(s_h,a_h)) + \mathbb{V}(P_{s_h,a_h,h}, V_{h+1}^\star)).$$

Let the trajectory of the k-th episode be τ^k, then we denote $\text{Var}_{(k)}^\Sigma := \text{Var}_{\tau^k}^\Sigma$, and $\text{Var}_K^\Sigma := \sum_{k=1}^K \text{Var}_{(k)}^\Sigma$.

Maximum Policy-Value Variance

For deterministic policy π, define

$$\text{Var}_1^\pi(s) := \mathbb{E}_\pi\left[\sum_{h=1}^{H} \left(\mathbb{V}(R_h(s_h,a_h)) + \mathbb{V}(P_{s_h,a_h,h}, V_{h+1}^\pi)\right) \,\middle|\, s_1 = s\right]$$

and $\text{Var}^\star := \max_{\pi\in\Pi, s\in\mathcal{S}} \text{Var}_1^\pi(s)$.

Discussion on Variances

- Under **Condition 1**, we have $\text{Var}_\tau^\Sigma \leqslant \widetilde{O}(1)$ with high probability if τ is sampled by any policy and $\text{Var}^\star \leqslant 1$. Without **Condition 1**, $\text{Var}_\tau^\Sigma \leqslant \widetilde{O}(H^2)$ and $\text{Var}^\star \leqslant H^2$.
- For deterministic MDPs, $\text{Var}_\tau^\Sigma = \text{Var}^\star = 0$.
- $\text{Var}^\star = 0 \implies \text{Var}_\tau^\Sigma = 0$, while reverse is not true.

Results of Model-Based Algorithms

The results are under **Condition 1** and **Condition 2**.

Algorithm	Regret	Variance-Dependent	Stochastic-Optimal	Deterministic-Optimal
`Euler`	$\widetilde{O}(\sqrt{H\mathbb{Q}^\star \cdot SAK} + H^{5/2}S^2A)$	Yes	No	No
	$\widetilde{O}(\sqrt{SAK} + H^{5/2}S^2A)$	No	Yes	No
`MVP`	$\widetilde{O}(\sqrt{SAK} + S^2A)$	No	Yes	No
`MVP-V` This work	$\widetilde{O}(\sqrt{\min\{\text{Var}_K^\Sigma, \text{Var}^\star K\}SA} + \Gamma SA)$	Yes	Yes	Yes
Lower bound	$\Omega(\sqrt{SAK})$ / $\Omega(SA)$	-	-	-

Remark:

- $\mathbb{Q}^\star$ could be as large as $\Omega(1)$ and $H\mathbb{Q}^\star \geqslant \text{Var}_\tau^\Sigma$.
- `MVP-V` recovers the optimal minimax rate because $\text{Var}^\star \leqslant O(1)$, and is optimal for deterministic MDPs because variances are 0 and $\Gamma = 1$.

Results of Model-Free Algorithms

Algorithm	Regret	Variance-Dependent	Stochastic-Optimal
`Q-learning (UCB-B)`	$\widetilde{O}(\sqrt{H^4SAK} + H^{9/2}S^{3/2}A^{3/2})$	No	No
`UCB-Advantage`	$\widetilde{O}(\sqrt{H^3SAK} + \sqrt[4]{H^{33}S^8A^6K})$	No	Yes
`Q-EarlySettled-Advantage`	$\widetilde{O}(\sqrt{H^3SAK} + H^6SA)$	No	Yes
`UCB-Advantage-V` This work	$\widetilde{O}(\sqrt{\min\{\text{Var}_K^\Sigma, \text{Var}^\star K\}HSA} + \sqrt[4]{H^{15}S^5A^3K})$	Yes	Yes
Lower bound	$\Omega(\sqrt{H^3SAK})$ / $\Omega(H^2SA)$	-	-

Remark: Currently no generic model-free algorithm achieves a constant regret for deterministic MDPs, while ours achieves a $K^{1/4}$ rate.

vectorzh@cs.washington.edu