清華大學 交叉信息研究院
Institute for Interdisciplinary Information Sciences, Tsinghua University

Sharp Gap-Dependent Variance-Aware Regret Bounds for Tabular MDPs

Shulun Chen[1*] Runlong Zhou[2] Zihan Zhang[3] Maryam Fazal[2] Simon S. Du[2]

Preliminaries

Markov Decision Processes

- S states, A actions, planning horizon H
- Reward distribution $R_{s,a,h} \in \Delta([0, H])$
- Transition probability $P_{s,a,h}$

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

$$V_h^\pi(s) := \mathbb{E}^\pi\left[\sum_{t=h}^{H} r_t \,\middle|\, s_h = s\right], V_0^\pi := \mathbb{E}^\pi\left[\sum_{t=1}^{H} r_t\right],$$

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

- Optimal and Suboptimal Actions

$$\mathcal{Z}_{\text{opt}} = \{(s,a,h) \in \mathcal{S}\times\mathcal{A}\times[H] : \Delta_h(s,a) = 0\},$$
$$\mathcal{Z}_{\text{sub}} = \mathcal{S}\times\mathcal{A}\times[H]\backslash\mathcal{Z}_{\text{opt}}.$$

- Gap quantities

$$\Delta_h(s,a) := V_h^*(s) - Q_h^*(s,a), \Delta_{\min} := \min_{(s,a,h)\in\mathcal{Z}_{\text{sub}}} \Delta_h(s,a).$$

Episodic reinforcement learning

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

$$\text{Regret}(K) := \sum_{k=1}^{K}(V_0^\star - V_0^{\pi^k}).$$

Maximum Conditional Total Variance

We define $\texttt{Var}^{\text{c}}_{\max}$, the maximum conditional total variance.

$$\texttt{Var}_h^\star(s,a) := \mathbb{V}^{r\sim R_{s,a,h}, s'\sim P_{s,a,h}}[r + V_{h+1}^\star(s')],$$
$$\texttt{Var}^{\text{c}}_{\max} := \max_{\pi,s,h} \mathbb{E}^\pi\left[\sum_{h'=1}^{h} \texttt{Var}_{h'}^\star(s_{h'}, a_{h'}) \,\middle|\, s_h = s\right].$$

We call

$$\texttt{Var}_{\max} := \max_{\pi} \mathbb{E}^\pi\left[\sum_{h=1}^{H} \texttt{Var}_h^\star(s_h, a_h)\right]$$

the maximum unconditional total variance.

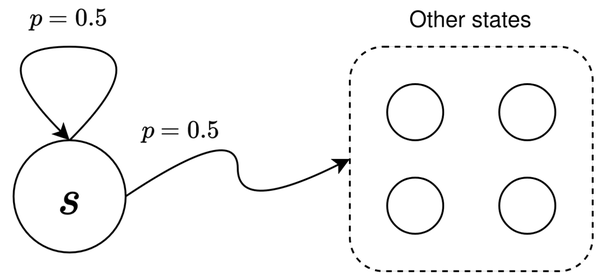

Figure 1. Conditioning on $s_h = s$, $\texttt{Var}^{\text{c}}_{\max}$ can be as large as $\Theta(H^3)$

Conditions for MDPs

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

Question under Research

We know

$$\mathbb{E}^\pi\left[\sum_{h'=h}^{H} \texttt{Var}_h^*(s_{h'}, a_{h'}) \,\middle|\, s_h = s\right] \leqslant H^2,$$

but previous gap-dependent analysis does not use this fact.

Can we find the tightest problem-dependent regret while considering both variance and gap?

Regret Bound Comparison

Algorithm	Regret	Variance-Dependent
StrongEuler	$\widetilde{O}(\sum_{h,s,a} H\mathbb{Q}^\star/(\Delta_h(s,a)\vee\Delta_{\min}))$	**Yes**
	$\widetilde{O}(\sum_{(h,s,a)\in\mathcal{Z}_{\text{sub}}} H^3/\Delta_h(s,a) + \lvert\mathcal{Z}_{\text{opt}}\rvert H^3/\Delta_{\min})$	No
Q-Learning (UCB-H)	$\widetilde{O}(H^6SA/\Delta_{\min})$	No
AMB	$\widetilde{O}(\sum_{(h,s,a)\in\mathcal{Z}_{\text{sub}}} H^5/\Delta_h(s,a) + \lvert\mathcal{Z}_{\text{mul}}\rvert H^5/\Delta_{\min})$	No
UCB-Advantage	$\widetilde{O}((H\mathbb{Q}^\star + H)H^2SA/\Delta_{\min})$	No
Q-EarlySettled-Advantage	$\widetilde{O}((H\mathbb{Q}^\star + H)H^2SA/\Delta_{\min})$	No
MVP This work	$\widetilde{O}(\sum_{(h,s,a)\in\mathcal{Z}_{\text{sub}}} W/\Delta_h(s,a) + \lvert\mathcal{Z}_{\text{opt}}\rvert W/\Delta_{\min})$	**Yes**

Remark:

- $\tilde{O}$ hides all poly $\log(SAHK)$ terms, and burn-in terms are not presented in the table. See our paper for more details.
- $W = H^2 \wedge \texttt{Var}^{\text{c}}_{\max} \leqslant H\mathbb{Q}^\star \leqslant H^3$, so our regret bound recovers all previous results except for AMB.

Lower Bound Result

Given (with some mild conditions):

- Target gap set $\{\Delta_i\}$
- Target maximum conditional total variance L,

there exists an MDP instance with previous targets so that

$$\text{Regret}(K) \geqslant \Omega\left(\sum_{i:\Delta_i>0} \frac{L}{\Delta_i}\cdot\log K\right).$$

Remark:

- Our lower bound matches with the first term of our regret bound.
- The MDP instance can be taken with small maximum unconditional total variance $\texttt{Var}_{\max}$.

NeurIPS 2025, San Diego
chensl22@mails.tsinghua.edu.cn
[1]Tsinghua University [2]University of Washington [3]HKUST
*Work done while Shulun Chen was visiting University of Washington