Variance-Dependent Regret Bounds of Model-Based and Model-Free RL

Runlong Zhou

Apr 26 2023

Qualification Presentation

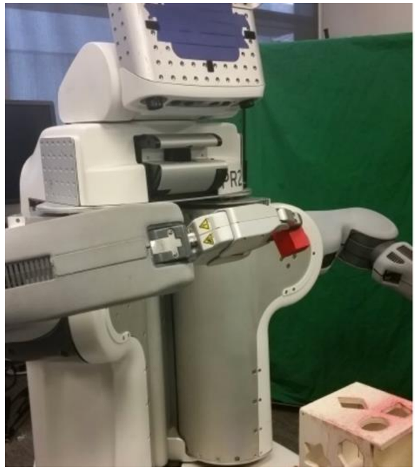
Acknowledgement

This is a joint work with

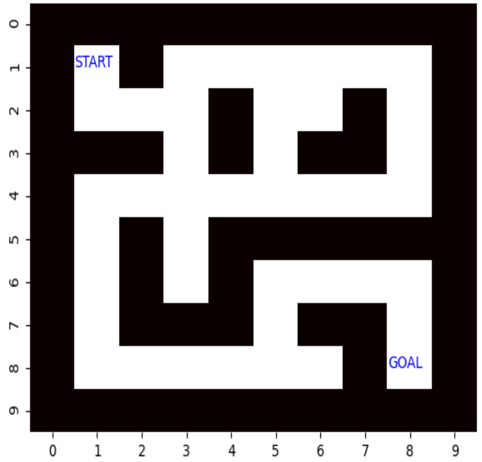
Zihan Zhang and **Simon Du**

Reinforcement Learning

Games with random environments



Robotics



Maze

The agent interacts wi… litioning on the observed sequence of…

Stochastic | **Deterministic**

Can we design an algorithm which **automatically** exploits determinism?

- Sometimes the **uncertainty** of “what is the **next state and reward**” is high
- Sometimes they are totally **determined**

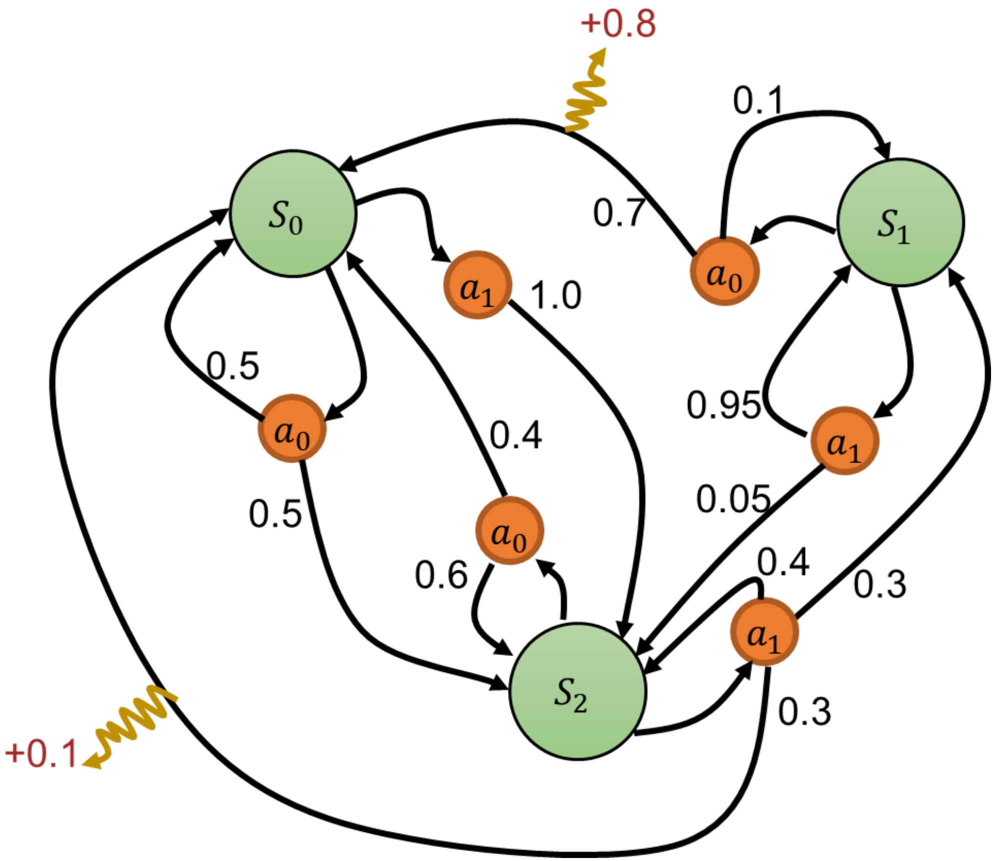
High variance | **Zero variance**

Markov Decision Processes (MDPs)

- **State** space $\mathcal{S}$, with size S
- **Action** space $\mathcal{A}$, with size A
- **Planning** horizon H
- **Reward** function $R_h(s,a) \in \Delta([0,1])$ with mean $r_h(s,a)$ for $h \in [H]$

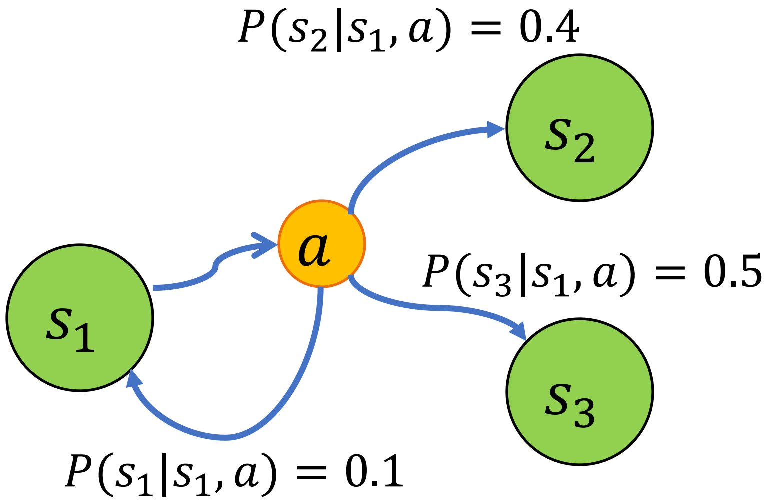
Probability simplex

- **Transition** model $P_h(s'|s,a)$



Maximum Transition Support

- $\Gamma = \max_{h,s,a} \|P_h(\cdot\,|s,a)\|_0$
- For **deterministic** MDPs, $\Gamma = 1$



An illustration for $\Gamma = 3$

Policy and Value Function

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

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

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

- **Optimal policy** denoted as $\pi^\star$, with $V_h^\star(s)$ and $Q_h^\star(s,a)$

Conditions for MDPs

Totally-bounded reward (TBR)

For any possible trajectory $\tau = \{(s_h, a_h, r_h)_{h=1}^{H}\} \cup \{s_{H+1}\}$,

$$\sum_{h=1}^{H} r_h \leq 1.$$

A fair comparison with **contextual bandits:** bandits have a single step reward $\in [0,1]$.

Time-homogeneous (TH)

For any $h, h' \leq H$, $P_h = P_{h'}$ and $R_h = R_{h'}$.

Previous Results

- **Model-based** algorithms can be **tight** under TBR and TH
  - Tight: upper bound matches lower bound

- No **model-free** algorithm is tight under TH
  - **Model-free**: space complexity $\leq O(SAH)$
    - Constructing P takes $O(S^2AH)$ space

Episodic RL for MDPs

- **Number of episodes** K
- Play a policy π^k in episode k
- **Regret**

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

$\tilde{O}$ hides poly log terms

- **Minimax (worst case)** regret, tight bounds
  - Upper bound (main order term, TBR & TH): $\tilde{O}(\sqrt{SAK})$ [Zanette and Brunskill, 2019, Zhang et al., 2021a]
  - Lower bound (TBR & TH): $\Omega(\sqrt{SAK})$ (a contextual bandit as a special MDP)

Problem-Dependent Regret

- Some MDPs are **easier** than the others
  - Applying an algorithm on them yields regrets better than **worst-case**
  - Example: **Deterministic** MDPs
    - Regret lower bound $\Omega(SA)$ (TBR & TH), no dependency on K
    - **Specially designed** algorithms can have regret upper bound $O(SA)$

> - Maintain a list of unexplored (s, a) pairs
> - In each episode, if anything in the list is reachable, visit it and remove from the list
> - After SA episodes the **accurate model** is established!

Main Contribution

- Define suitable **(problem-dependent)** quantities to **characterize the difficulty** of each MDP
- Design **generic** algorithms which
  - Preserve **minimax optimal regret**
  - **Automatically** adapt to structures of MDPs: regrets depend on the above quantities

Some Problem-Dependent Results

Gap-dependent regret [Even-Dar et al., 2006, Auer et al., 2008, Simchowitz and Jamieson, 2019, Xu et al., 2021, Yang et al., 2021, …]

- Proportional to the **inverse** of sub-optimality gap on optimal Q-functions
- May **not** recover **minimax** optimal regret, beyond scope of this work

Optimal value function

First-order regret

- Tabular MDP (TBR): $\tilde{O}\left(\sqrt{V_1^\star(s_0)HSAK}\right)$ [Jin et al., 2020]
- Linear function approximation: $\tilde{O}\left(\sqrt{V_1^\star(s_0)d^3H^3K}\right)$ [Wagenmaker et al., 2022]
- **Cannot** characterize **deterministic** MDPs!

Variances

Maximum per-step conditional variance [Zanette and Brunskill, 2019]

$$\mathbb{Q}^\star := \max_{h,s,a}\{\mathbb{V}(R_h(s,a)) + \mathbb{V}(P_{s,a,h}, V^\star_{h+1})\}.$$

Variance operator

- $\mathbb{V}(X) = \mathbb{E}[(X - \mathbb{E}X)^2], \mathbb{V}(p,x) = \sum_i \left(x_i - \sum_j p_j x_j\right)^2$

Example of **large** variances:

- $R_h(s,a) = \mathrm{Unif}_{\{0,1\}}$
- $P_{s,a,h} = (0.5, 0.5)$ and $V^\star_{h+1} = (0,1)$

Example of **small** variances:

- Deterministic MDPs
- $V^\star_{h+1}(s) = V^\star_{h+1}(s')$ for any s, s'

- $\mathbb{Q}^\star$ is not sufficient for our goal, need **more definitions**!

Variances (cont'd)

Total multi-step conditional variance

For any trajectory τ:

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

With $\mathrm{Var}_{K}^{\Sigma} \coloneqq \sum_{k=1}^{K} \mathrm{Var}_{\tau^k}^{\Sigma}$ as the total variance in episodic RL.

Variances (cont'd)

Maximum policy-value variance

For any policy π:

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

Further define $\mathrm{Var}^\pi := \max_s \mathrm{Var}_1^\pi(s)$.

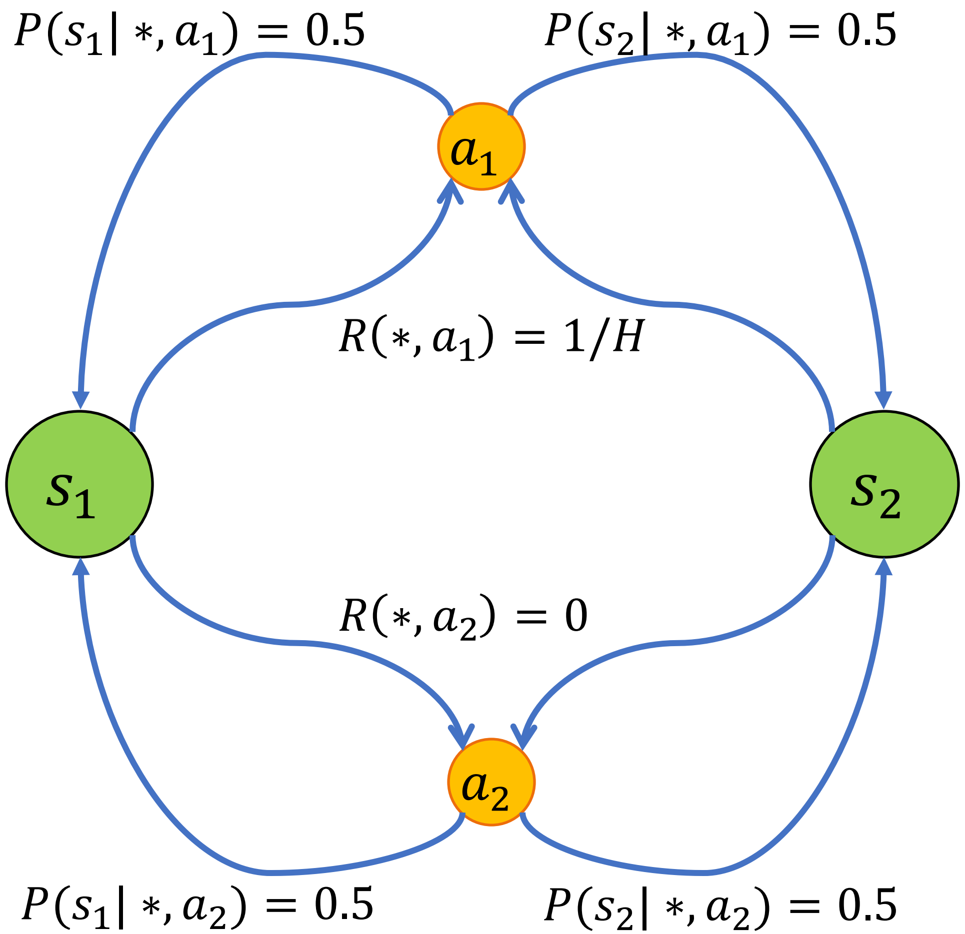
Maximum policy-value variance is defined as $\mathrm{Var}^\star := \max_\pi \mathrm{Var}^\pi$.

Comparing Variances

- $\mathrm{Var}_\tau^\Sigma \leq H\mathbb{Q}^\star$
  - They could all be $\Omega(H)$ (TBR) in the worst case
  - $\boldsymbol{\mathrm{Var}_\tau^\Sigma \leq \widetilde{O}(1)}$ **(TBR) with high probability** if τ is generated by a policy π
- $\mathrm{Var}^\star \leq V_1^\star \leq 1$ (TBR)
  - Better than first-order!
- Deterministic MDPs: $\mathrm{Var}_\tau^\Sigma = \mathrm{Var}^\star = 0$
- $\mathrm{Var}^\star = 0 \implies \mathrm{Var}_\tau^\Sigma = 0$
  - **Reverse is not true!**

$\mathrm{Var}_\tau^\Sigma = 0 < \mathrm{Var}^\star$

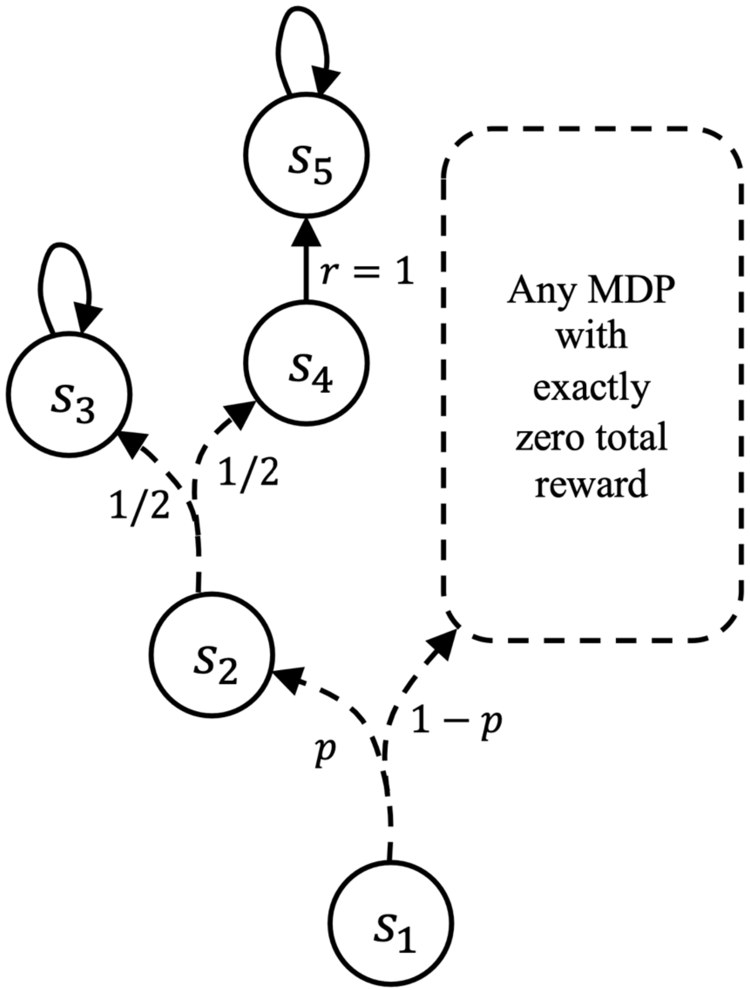
- $\pi_h^\star(s) = a_1$
  - $V_h^\star(s) = (H - h + 1)/H$, same across the same time step
  - $\mathrm{Var}_\tau^\Sigma = 0$
- $\pi_H(s_1) = a_2$ and any other action be a_1
  - Total reward $\in \{1, 1 - 1/H\}$
  - $\mathrm{Var}^\star > 0$



$\mathrm{Var}_\tau^\Sigma = 1/4 > \mathrm{Var}^\star \approx 0$

$$\mathsf{Var}_{(k)}^{\Sigma} \geqslant \mathbb{V}(R(s_2, a)) + \mathbb{V}(P_{s_2,a}, V_3^\star) = \frac{1}{4}.$$

- $\mathrm{Var}^\star \le \max_\pi V_1^\pi(s_1) \le p$
  - Take $p \to 0$



Model-Based Results (TBR & TH)

Totally-bounded reward

Time-homogeneous

Only log dependency on H

Algorithm	Regret	Variance-Dependent	Stochastic-Optimal	Deterministic-Optimal	Horizon-Free
Euler Zanette and Brunskill [2019]	$\tilde{O}(\sqrt{H\mathbb{Q}^\star \cdot SAK} + H^{5/2}S^2A)$	**Yes**	No	No	No
	$\tilde{O}(\sqrt{SAK} + H^{5/2}S^2A)$	No	**Yes**	No	No
MVP Zhang et al. [2021a]	$\tilde{O}(\sqrt{SAK} + S^2A)$	No	**Yes**	No	**Yes**
MVP-V This work	$\tilde{O}(\sqrt{\min\{\mathrm{Var}_K^\Sigma, \mathrm{Var}^\star K\}SA} + \Gamma SA)$	**Yes**	**Yes**	**Yes**	**Yes**

Model Estimation

- Transition: $\hat{P}^k_{s,a}(s') = \frac{n^k(s,a,s')}{n^k(s,a)}$
  - $n^k(s,a)$: total visitation to (s,a) before episode k
  - $n^k(s,a,s')$: total visitation to (s,a,s') before episode k
- Reward: $\hat{r}^k(s,a) = \frac{\theta^k(s,a)}{n^k(s,a)}$
  - $\theta^k(s,a)$: summation of observed reward on (s,a) before episode k
- **Variance** of reward: $\widehat{\mathrm{VarR}}^k(s,a) = \frac{\phi^k(s,a)}{n^k(s,a)} - \hat{r}^k(s,a)^2$
  - $\phi^k(s,a)$: summation of **squared** observed reward on (s,a) before episode k

Optimism

Value estimation

- In episode k, inductively calculate

$$Q_h^k(s,a) = \hat{r}^k(s,a) + \hat{P}_{s,a}^k V_{h+1}^k + b_h^k(s,a)$$

- $b_h^k(s,a)$: bonus function to account for empirical error in $\hat{r}^k(s,a)$ and $\hat{P}_{s,a}^k V_{h+1}^k$
- Policy $\pi_h^k(s) = \arg\max_a Q_h^k(s,a)$
- Value $V_h^k(s) = \max_a Q_h^k(s,a)$

Optimism

$$Q_h^k(s,a) \geq Q_h^\star(s,a)$$

- Large bonus ensures optimism, but incurs large regret

Bonus

Bennett's Inequality to introduce variances

$$\mathbb{P}\left[\left|\mathbb{E}[Z] - \frac{1}{n}\sum_{i=1}^{n} Z_i\right| > \sqrt{\frac{2\mathbb{V}[Z]\ln(2/\delta)}{n}} + \frac{b\ln(2/\delta)}{n}\right] \leqslant \delta.$$

Exploration bonus with empirical variances ($\iota = \tilde{O}(1)$)

$$b_h(s,a) \leftarrow 4\sqrt{\frac{\boxed{\mathbb{V}(\widehat{P}_{s,a}, V_{h+1})}\iota}{\max\{n(s,a),1\}}} + 2\sqrt{\frac{\boxed{\widehat{\mathsf{VarR}}(s,a)}\iota}{\max\{n(s,a),1\}}} + \frac{21\iota}{\max\{n(s,a),1\}};$$

- Change to MVP: using empirical variances of rewards

Analysis

Regret decomposition

- By optimism $\sum_k \left(V_1^\star(s_1^k) - V_1^{\pi^k}(s_1^k)\right) \le \sum_k \left(V_1^k(s_1^k) - V_1^{\pi^k}(s_1^k)\right) \lesssim \sum_{k,h} b_h^k(s_h^k, a_h^k)$

$$b_h(s,a) \leftarrow 4\sqrt{\frac{\mathbb{V}(\widehat{P}_{s,a}, V_{h+1})\iota}{\max\{n(s,a),1\}}} + 2\sqrt{\frac{\widehat{\mathsf{VarR}}(s,a)\iota}{\max\{n(s,a),1\}}} + \frac{21\iota}{\max\{n(s,a),1\}};$$

Cauchy-Schwarz: $\sum_{k,h} \sqrt{\frac{w_h^k}{n^k(s_h^k,a_h^k)}} \le \sqrt{\sum_{k,h} \frac{1}{n^k(s_h^k,a_h^k)}} \sqrt{\sum_{k,h} w_h^k} \lesssim \sqrt{SA \sum_{k,h} w_h^k}$

Regret main order term is $\tilde{O}\left(\sqrt{SAW}\right)$, where

$$W = \sum_{k=1}^{K} \sum_{h=1}^{H} \left(\mathbb{V}(R(s_h^k, a_h^k)) + \mathbb{V}(P_{s_h^k, a_h^k}, V_{h+1}^{\pi^k})\right)$$

Analysis: Var_K^Σ

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

$$W = \sum_{k=1}^{K} \sum_{h=1}^{H} (\mathbb{V}(R(s_h^k,a_h^k)) + \mathbb{V}(P_{s_h^k,a_h^k}, V_{h+1}^{\pi^k}))$$

By $\mathbb{V}(X+Y) \le 2\mathbb{V}(X) + 2\mathbb{V}(Y)$,

$$W \le 2\sum_{k=1}^{K}\sum_{h=1}^{H}\left(\mathbb{V}\left(R\left(s_h^k,a_h^k\right)\right) + \mathbb{V}\left(P_{s_h^k,a_h^k}, V_{h+1}^\star\right)\right)$$
$$+ 2\sum_{k=1}^{K}\sum_{h=1}^{H}\mathbb{V}\left(P_{s_h^k,a_h^k}, V_{h+1}^{\pi^k} - V_{h+1}^\star\right)$$

Var_K^Σ

$\le \tilde{O}(\sqrt{W})$, lower order term

Analysis: $\mathrm{Var}^\star$

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

$$W = \sum_{k=1}^{K} \boxed{\sum_{h=1}^{H} \left(\mathbb{V}(R(s_h^k, a_h^k)) + \mathbb{V}(P_{s_h^k, a_h^k}, V_{h+1}^{\pi^k}) \right)}$$

From estimation to expectation

- Each inner sum is **an unbiased estimation of $\mathrm{Var}_1^{\pi^k}(s_1^k)$**!
- Naïvely bounding W using a martingale concentration inequality yields an ***H*-dependency** on lower order terms

Solution: Truncation

- **Truncate** each inner sum to $\tilde{O}(1)$ before using martingale concentration inequality
- Effective truncation happens with low probability

Model-Free Results

Space $\leq O(SAH)$. No model estimation!

Algorithm	Regret	Variance-Dependent	Stochastic-Optimal
`Q-learning (UCB-B)` Jin et al. [2018]	$\tilde{O}(\sqrt{H^4SAK} + H^{9/2}S^{3/2}A^{3/2})$	No	No
`UCB-Advantage` Zhang et al. [2020]	$\tilde{O}(\sqrt{H^3SAK} + \sqrt[4]{H^{33}S^8A^6K})$	No	**Yes**
`Q-EarlySettled-Advantage` Li et al. [2021]	$\tilde{O}(\sqrt{H^3SAK} + H^6SA)$	No	**Yes**
`UCB-Advantage-V` This work	$\tilde{O}(\sqrt{\min\{\mathsf{Var}_K^{\Sigma}, \mathsf{Var}^{\star}K\}HSA} + \sqrt[4]{H^{15}S^5A^3K})$	**Yes**	**Yes**

Value Estimation

By some means of estimation

Naïve update rule: $Q_h(s,a) \leftarrow \widehat{P_{s,a,h}V_{h+1}} + \hat{r}_h(s,a) + b_h^k(s,a)$

- The **earlier** a sample is collected, the **more deviation** it is from $Q_h^\star$ because V changes

Reference value functions from UCB-Advantage [Zhang et al., 2020]

$$Q_h(s,a) \leftarrow \widehat{P_{s,a,h}V_{h+1}^{\text{ref}}} + P_{s,a,h}\widehat{(V_{h+1} - V_{h+1}^{\text{ref}})} + \hat{r}_h(s,a) + b_h^k(s,a),$$

- Assume V_h^{ref} is some **fixed** value and **approximates $V_h^\star$ well**
  - $\widehat{P_{s,a,h}V_{h+1}^{\text{ref}}} = \widehat{P_{s,a,h}}V_{h+1}^{\text{ref}}$, we can use **all the data to estimate** $P_{s,a,h}$ **(low deviation)**
  - Deviation in $P_{s,a,h}\widehat{(V_{h+1} - V_{h+1}^{\text{ref}})}$ is low order

Learn Reference Values

- Let V_h^{ref} be snapshots of V_h with a certain **frequency**
  - Only once in UCB-Advantage?
    - $V_h^{\text{ref}} - V_h^\star \leq$ either H (before snapshot) or β (the desired approximation error)
    - Less flexibility, making the main order term **variance-independent**
  - Whenever $n_h(s,a)$ doubles (uncapped-doubling)?
    - Gives arbitrary small $V_h^{\text{ref}} - V_h^\star$ (approximation error)
    - Each update **invalidates** previous V_h^{ref}, causing **another accumulated bias** which is a **variance-independent** main order term
  - **Capped-doubling!**
    - Balances between approximation error and data waste

Capped-Doubling

Capped to $i^\star$ updates!

Design

- Set $\mathcal{R} = \left\{N_i = \widetilde{\Theta}\left(2^{2i} H^3 SA\right) \mid i \leq i^\star\right\}$
- When $\sum_a n_h(s,a) = N_i \in \mathcal{R}$ for some i, assign $V_h^{\mathrm{ref}}(s) \leftarrow V_h(s)$
  - Ensures $V_h^{\mathrm{ref}}(s) - V_h^\star(s) \leq \beta_i = H/2^i$

Motivation

- $i^\star$ dictates the final precision we want
  - Setting $i^\star$ not too large helps reduce data waste on $P_{s,a,h}(\widehat{V_{h+1} - V_{h+1}^{\mathrm{ref}}})$
- Intermediate updates give a successively halving error sequence
  - It makes analysis more flexible than one-time update

Bonus

$$Q_h(s,a) \leftarrow \widehat{P_{s,a,h}V^{\text{ref}}_{h+1}} + P_{s,a,h}(\widehat{V_{h+1} - V^{\text{ref}}_{h+1}}) + \widehat{r}_h(s,a) + b^k_h(s,a),$$

Similar as MVP-V, we need variance terms for V^{ref}, $V - V^{\text{ref}}$ and R

$$b \leftarrow 4\sqrt{\frac{\nu^{\text{ref}}\iota}{n}} + 4\sqrt{\frac{\check{\nu}\iota}{\check{n}}} + 2\sqrt{\frac{\widehat{\text{VarR}}_h\iota}{n}} + \frac{90H\iota}{\check{n}};$$

- $\nu_h^{\text{ref},k}$: variance of V^{ref} at the beginning of episode k
- $\check{\nu}_h^k$: variance of $V - V^{\text{ref}}$ at the beginning of episode k
- n_h^k: number of visitation at the beginning of episode k
- $\check{n}_h^k$: number of visitation after the last V^{ref} update and before episode k

Limitation of Model-Free Algorithms

- UCB-Advantage-V cannot achieve $\tilde{O}\left(\sqrt{SAK}\right)$ regret under TBR & TH
- Is there a model-free algorithm achieving **tight regret** under TBR & TH?
- Hardness:
    - How to apply updates to all H layers when any (s_h, a_h) is collected?

Limitation of Model-Free Algorithms (cont'd)

- On deterministic MDPs, UCB-Advantage-V has regret $\propto K^{1/4}$, **not a constant** as MVP-V
- Is there a **generic** model-free algorithm achieving constant regret on deterministic MDPs?
- Hardness:
  - Using **all historical data** to estimate value function is biased
  - To converge in constant steps, must **discard initial data**

References

- Peter Auer, Thomas Jaksch, and Ronald Ortner. Near-optimal regret bounds for reinforcement learning. Advances in neural information processing systems, 21, 2008.
- Eyal Even-Dar, Shie Mannor, Yishay Mansour, and Sridhar Mahadevan. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. Journal of machine learning research, 7(6), 2006.
- Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? Advances in neural information processing systems, 31, 2018.
- Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforce- ment learning. In International Conference on Machine Learning, pages 4870–4879. PMLR, 2020.
- Gen Li, Laixi Shi, Yuxin Chen, Yuantao Gu, and Yuejie Chi. Breaking the sample complexity barrier to regret-optimal model-free reinforcement learning. Advances in Neural Information Processing Systems, 34:17762–17776, 2021.
- Max Simchowitz and Kevin G Jamieson. Non-asymptotic gap-dependent regret bounds for tabular mdps. Advances in Neural Information Processing Systems, 32, 2019.
- Andrew J Wagenmaker, Yifang Chen, Max Simchowitz, Simon Du, and Kevin Jamieson. First-order regret in reinforcement learning with linear function approximation: A robust estimation approach. In International Conference on Machine Learning, pages 22384–22429. PMLR, 2022.
- Haike Xu, Tengyu Ma, and Simon Du. Fine-grained gap-dependent bounds for tabular mdps via adaptive multi-step bootstrap. In Conference on Learning Theory, pages 4438–4472. PMLR, 2021.
- Kunhe Yang, Lin Yang, and Simon Du. Q-learning with logarithmic regret. In International Conference on Artificial Intelligence and Statistics, pages 1576–1584. PMLR, 2021.
- Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In ICML, 2019.
- Zihan Zhang, Yuan Zhou, and Xiangyang Ji. Almost optimal model-free reinforcement learning via reference- advantage decomposition. Advances in Neural Information Processing Systems, 33:15198–15207, 2020.
- Zihan Zhang, Xiangyang Ji, and Simon Shaolei Du. Is reinforcement learning more difficult than bandits? a near-optimal algorithm escaping the curse of horizon. In COLT, 2021a.

Thank You