Is posterior sampling reinforcement learning near-optimal?

Yingru Li[1,2], Tong Zhang[2,3] and Zhi-Quan Luo[1,2]

[1] The Chinese University of Hong Kong, Shenzhen [2] Shenzhen Research Institute of Big Data [3] The Hong Kong University of Science and Technology

Previews

1. For **linear function approximation**, we establish an $\tilde{\mathcal{O}}(dH\sqrt{T})$ Bayesian regret bound for PSRL, where d is the ambient dim, H is the planning horizon and T is the number of interactions. Our upper bound is the **first** one showing PSRL can **achieve the minimax lower bound** $\Omega(dH\sqrt{T})$ up to a logarithmic factor.
2. For **general function approximation**, we establish an $\tilde{\mathcal{O}}(H\sqrt{d_E d_K T})$ Bayesian regret bound which improves $\sqrt{H}$ factor compared with the bound in [OVR14]. Here d_E and d_K are complexity measures related to the function class.
3. As an implication from the general bound, it is the **first** time PSRL can be shown to achieve tight dependence in H for tabular setting.

Problem formulation

Learn to optimize a **random** *finite horizon MDP* M in repeated finite episodes of interaction.

Figure 1: Reinforcement Learning

- State space $\mathcal{S}$, action space $\mathcal{A}$
- Rewards $r_{h+1} \sim \mathcal{D}_R$ with mean $R(h, s_h, a_h)$
- Transitions $s_{h+1} \sim P(h, s_h, a_h) \in \mathcal{P}$
- Episode length H. Initial state $s_0 \sim \rho$.

- **Value function**: For an MDP $\hat{M}$ and a policy π,

$$V_{\pi,h}^{\hat{M}}(s) := \mathbb{E}\left[\sum_{j=h}^{H-1} r_{j+1} \,\middle|\, M = \hat{M}, s_h = s, \pi\right],$$

- **Expected value** of a policy π under an MDP $\hat{M}$ is

$$\overline{V}_{\pi}^{\hat{M}} = \mathbb{E}\left[V_{\pi,0}^{\hat{M}}(s_0) \mid \hat{M}, \pi\right].$$

- **Bayesian regret** of an alg over L episodes in the environment with underlying MDP M is

$$\mathfrak{BR}(\text{alg}, L) = \sum_{\ell=1}^{L} \mathbb{E}_{\text{alg}}\left[\overline{V}_{\pi^*}^{M} - \overline{V}_{\pi_\ell}^{M}\right],$$

where the expectation integrates over actions, state transitions, and any the algorithmic randomness used by alg, as well as the **prior distribution** $\mathbb{P}(M \in \cdot)$ over the random underlying MDP M.

Efficient Algorithm

Great challenge: data-efficiency in RL
Naive exploration such as Boltzman or ϵ-greedy can lead to exponential regret. Good performance requires balancing **exploration vs exploitation**.

Posterior sampling RL (PSRL):

Denote quantities for stage h of episode ℓ with subscript ℓ, h. Require prior $\mathbb{P}(M \in \cdot)$.

For each episode $\ell = 1, 2, \ldots$:

1. Sample an MDP from the posterior distribution for the true MDP:

$$\hat{M}_\ell \sim \mathbb{P}(M \in \cdot \mid \mathcal{H}_\ell)$$

2. Solve the optimal policy under MDP $\hat{M}_\ell$

$$\pi_\ell = \arg\max_{\pi} \overline{V}_{\pi}^{\hat{M}_\ell}$$

and execute the policy π_ℓ in episode ℓ.

3. Update history with new observations:

$$\mathcal{H}_{\ell+1} = \mathcal{H}_\ell \cup (s_{\ell,h}, a_{\ell,h}, r_{\ell,h+1}, s_{\ell,h+1})_{h=0}^{H-1}.$$

MDP Classes

- **Tabular MDP class**: Finite $\mathcal{S}$ and finite $\mathcal{A}$.
- **Linear Mixture MDPs class**: (ambient dim d)

$$P(s' \mid h, s, a) = \langle \theta_h^*, \phi(s' \mid s, a) \rangle.$$

Basis transition kernels (feature maps):

$$\phi(s' \mid s, a) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d.$$

Model parameters: $\Theta^* = (\theta_0^*, \ldots, \theta_{H-1}^*)$ with prior $\mathbb{P}(\Theta^* \in \cdot)$. The posterior sampling can implemented as first sampling the set $\hat{\Theta}_\ell = (\hat{\theta}_{\ell,0}, \ldots, \hat{\theta}_{\ell,H-1}) \sim \mathbb{P}(\Theta^* \in \cdot \mid \mathcal{H}_\ell)$ and then constructing virtual transition function

$$\hat{P}_\ell(\cdot \mid h, \cdot, \cdot) = \left\langle \hat{\theta}_{\ell,h}, \phi(\cdot \mid \cdot, \cdot) \right\rangle, \forall h \in [H]$$

as well as the MDP $\hat{M}_\ell = \left(\mathcal{S}, \mathcal{A}, \hat{P}_\ell, R, H, \rho\right)$.

- **General MDP class**: General transition function class $\mathcal{P}$. Denote d_K and d_E to be the Kolmogorov dimension and the eluder dimension of $\mathcal{P}$.

Main theorems

Theorem 1 (Linear analysis)

For any prior over models $\Theta^* = (\theta_0^*, \ldots, \theta_{H-1}^*)$ that are mutually independent and B-bounded in dimension d, PSRL have the Bayesian regret bound over L episodes interaction with the time-inhomogeneous linear mixture MDP with bounded known feature mapping,

$$\mathfrak{BR}(\text{PSRL}, L) \le \sqrt{3}d\sqrt{H^4 L/d + H^3 L + 64 d^2 H^6 \iota}\sqrt{\iota}$$

where $\iota = \log(1 + 2dLB^2)$ is a logarithmic factor.

Theorem 2 (General analysis)

Let $\beta_\ell = 8\log\left(\frac{2\mathcal{N}(\mathcal{F}, \alpha, \|\cdot\|_\infty)}{\delta}\right) + 4\ell\alpha\left(H + \sqrt{\log\left(\frac{4\ell(\ell+1)}{\delta}\right)}\right)$. For all $\alpha > 0$, $\delta \le 1/2T$ and $L \in \mathbb{N}$ and for any prior over models, the Bayesian regret of PSRL over L episode is upper bounded by

$$1 + \sqrt{\left(\frac{d_E H^2}{\beta_L} + LH^2 + 4H^2 + 16H^4(1 + d_E\beta_L H(1 + 4\log L)\right)(1 + d_E\beta_L H(1 + 4\log L)},$$

where $d_E = \dim_E(\mathcal{P}, \sqrt{1/T})$ is the eluder dimension of function class $\mathcal{P}$ at scale $\sqrt{1/T}$.

So What?

	Linear mixture MDPs	Tabular MDPs	General MDP class
Our linear analysis	$\tilde{\mathcal{O}}\left(dH\sqrt{T\log T}\right)$	$\tilde{\mathcal{O}}(S^2AH\sqrt{T})$	–
Our general Analysis	$\tilde{\mathcal{O}}(dH\sqrt{T}\log T)$ [a]	$\tilde{\mathcal{O}}(H\sqrt{S^2AT}\log T)$ [b]	$\tilde{\mathcal{O}}(H\sqrt{d_E d_K T\log T})$ [c]
[OVR17]	–	$\tilde{\mathcal{O}}(\sqrt{H^3SAT}\log T)$	–
[OVR14]	$\tilde{\mathcal{O}}(d\sqrt{H^3T}\log T)$ [d]	–	$\tilde{\mathcal{O}}(\sqrt{H^3 d_E d_K T\log T})$ [e]
Minimax Lower bound	$\Omega(dH\sqrt{T})$	$\Omega(H\sqrt{SAT})$	–

[a] This is translated from the general bound when $d_E = \tilde{\mathcal{O}}(d\log T)$ and $d_K = d$.
[b] This is translated from Theorem 2 when $d_E = S^2A$ and $\beta_L = \mathcal{O}(\log SAT)$.
[c] The Kolmogorov dimenstion d_K is derived from log-covering number in β_L.
[d] This result is not reported in [OVR14] but is a corollary derived from the bound in general MDP class.
[e] The original bound in [OVR14] is $\tilde{\mathcal{O}}(\mathbb{E}[K^*]\sqrt{d_K d_E T\log T})$ for time-homogeneous setting. Here $\mathbb{E}[K^*]$ is upper bounded by H. Translating this bound to time-inhomogeneous setting is simply multiplying with additional $\sqrt{H}$ factor.

References & More Information (Scan QR Code)

[OVR14] Ian Osband and Benjamin Van Roy. Model-based reinforcement learning and the eluder dimension. In *Advances in Neural Information Processing Systems*, pages 1466–1474, 2014.

[OVR17] Ian Osband and Benjamin Van Roy. Why is posterior sampling better than optimism for reinforcement learning? In *International Conference on Machine Learning*, pages 2701–2710, 2017.