Nearly Minimax Optimal Reinforcement Learning with Linear Function Approximation

Pihe Hu, Yu Chen, Longbo Huang

Tsinghua University

{*hph19,c-y19*}*@mails.tsinghua.edu.cn, longbohuang@tsinghua.edu.cn*

June 28, 2022

Markov Decision Process

An episodic finite horizon MDP is denoted as $\mathcal{M} = \{\mathcal{S}, \mathcal{A}, H, \{\mathbb{P}_h\}_h, \{r_h\}_h\}$

- Value function: $V_h^\pi(s) = \mathbb{E}[\sum_{h'=h}^{H} r_{h'}(s_{h'}, \pi_{h'}(s_{h'})) \mid s_h = s, \pi]$
- Learning goal - minimize cumulative regret

$$\text{Regret}(K) = \sum_{k=1}^{K} [V_1^\star(s_1^k) - V_1^{\pi_k}(s_1^k)], \tag{1}$$

where $V_1^*(\cdot)$ is the optimal value function.

For tabular MDPs:

- Minimax optimal regret $\widetilde{O}(\sqrt{H^2SAT})$ is achieved by UCBVI in [Azar et al., 2017]

RL with Linear Function Approximation

The curse-of-dimensionality → Function approximation

Open problem: *Does there exist a computation-efficient and minimax optimal algorithm for RL with linear function approximation?*

- Many problems can be linearly-parameterized structurally or linearly-combined with embedding.

Definition (Linear MDP)

Known feature mapping $\phi \in \mathbb{R}^d$, unknown measure $\boldsymbol{\mu}_h(s'), \boldsymbol{\theta}_h \in \mathbb{R}^d$:

$$\mathbb{P}_h(s' \mid s, a) = \langle \phi(s, a), \boldsymbol{\mu}_h(s') \rangle$$
$$r_h(s, a) = \langle \phi(s, a), \boldsymbol{\theta}_h \rangle$$

Significance

Table: Theoretical results on Linear MDPs

Algorithm	Setting	Regret
OPT-RLSVI [Zanette et al., 2020]	Linear MDP	$\tilde{O}(H^2 d^2 \sqrt{T})$
LSVI-UCB [Jin et al., 2020]	Linear MDP	$\tilde{O}(\sqrt{H^3 d^3 T})$
LSVI-UCB$^+$ (**this paper**)	Linear MDP	$\tilde{O}(Hd\sqrt{T})$
Lower Bound[Zhou et al., 2021]	Linear (Mixture) MDP	$\Omega(Hd\sqrt{T})$

- LSVI-UCB$^+$ is the first computationally-efficient and nearly minimax optimal algorithm.

Novelty: Eliminating Barriers to Minimax Optimality

- Overly Aggressive Exploration → $\sqrt{H}$ reduction
  - Hoeffding-type bonus → Bernstein-type bonus
- Extra Uniform Convergence Cost → $\sqrt{d}$ reduction
  - Bounding the deviation term[1] with the correction term

$$[(\widehat{\mathbb{P}}_{k,h} - \mathbb{P}_h)\widehat{V}_{k,h+1}](s_h^k, a_h^k) \simeq \underbrace{\| \sum_{i=1}^{k-1} \widehat{\sigma}_{i,h}^{-2} \phi(s_h^i, a_h^i) {\epsilon_h^i}^\top \widehat{\boldsymbol{V}}_{k,h+1} \|_{\widehat{\Lambda}_{k,h}^{-1}}}_{\text{Self-normalized Bound}}$$

$$\leqslant \underbrace{[(\widehat{\mathbb{P}}_{k,h} - \mathbb{P}_h)V_{h+1}^*](s_h^k, a_h^k)}_{\text{Dominant term with respect to } \widehat{V}_{h+1}^*} + \underbrace{[(\widehat{\mathbb{P}}_{k,h} - \mathbb{P}_h)(\widehat{V}_{k,h+1} - V_{h+1}^*)](s_h^k, a_h^k)}_{\text{Correction Term}}$$

Novel analytical tools:

- Bernstein self-normalized bound
- Conservatism of Elliptical Potentials

[1] $\epsilon_h^k := \mathbb{P}_h(\cdot \mid s_h^k, a_h^k) - \boldsymbol{\delta}(s_{h+1}^k)$, where $\boldsymbol{\delta}(s) \in \mathbb{R}^{|\mathcal{S}|}$ is a one-hot vector that is zero everywhere except the entry corresponding to state s is one.

Optimal Exploration for linear MDPs (LSVI-UCB$^+$)

- Linear Weighted Ridge Regression

$$\min_{\boldsymbol{\mu}\in\mathbb{R}^{d\times|\mathcal{S}|}} \sum_{i=1}^{k-1} \left\| \left[\boldsymbol{\mu}_h^\top \phi(s_h^k, a_h^k) - \boldsymbol{\delta}(s_{h+1}^i)\right] \widehat{\sigma}_{i,h}^{-1} \right\|_2^2 + \lambda\|\boldsymbol{\mu}\|_F^2,$$

where the weight $\widehat{\sigma}_{k,h}$ is the variance of value function $\rightarrow$ the Law of Total Variance (LTV) [Lattimore et al., 2012]

Theorem (Regret Upper Bound)

With high probability, the regret of LSVI-UCB$^+$ is upper bounded by

$$\text{Regret}(K) = \tilde{O}\left(Hd\sqrt{T} + H^3d^6 + \sqrt{H^7d^7}\right) \rightarrow \tilde{O}\left(Hd\sqrt{T}\right)$$

References

Azar, Mohammad Gheshlaghi, Ian Osband, and Rémi Munos. "Minimax regret bounds for reinforcement learning." International Conference on Machine Learning. PMLR, 2017.

Jin, Chi, Zhuoran Yang, Zhaoran Wang, and Michael I. Jordan. "Provably efficient reinforcement learning with linear function approximation." In Conference on Learning Theory, pp. 2137-2143. PMLR, 2020.

Zhou, Dongruo, Quanquan Gu, and Csaba Szepesvari. "Nearly minimax optimal reinforcement learning for linear mixture markov decision processes." In Conference on Learning Theory, pp. 4532-4576. PMLR, 2021.

Cai, Qi, Zhuoran Yang, Chi Jin, and Zhaoran Wang. "Provably efficient exploration in policy optimization." In International Conference on Machine Learning, pp. 1283-1294. PMLR, 2020.

Zanette, Andrea, David Brandfonbrener, Emma Brunskill, Matteo Pirotta, and Alessandro Lazaric. "Frequentist regret bounds for randomized least-squares value iteration." In International Conference on Artificial Intelligence and Statistics, pp. 1954-1964. PMLR, 2020.

Lattimore, Tor, and Marcus Hutter. "PAC bounds for discounted MDPs." In International Conference on Algorithmic Learning Theory, pp. 320-334. Springer, Berlin, Heidelberg, 2012.

The End