Provably Efficient Reinforcement Learning with Linear Function Approximation under Adaptivity Constraints

Tianhao Wang[1,*] and **Dongruo Zhou**[2,*] and **Quanquan Gu**[2]

[1]Department of Statistics and Data Science, Yale University

[2]Department of Computer Science, University of California, Los Angeles

Problem Setting

- **Episodic Markov Decision Processes:** $\mathcal{M}(\mathcal{S}, \mathcal{A}, H, \{r_h\}_{h=1}^H, \{\mathbb{P}_h\}_{h=1}^H)$
  - State space $\mathcal{S}$, action space $\mathcal{A}$
  - Reward function $r_h : \mathcal{S} \times \mathcal{A} \to [0, 1]$
  - Transition probability function $\mathbb{P}_h(s' \mid s, a)$
  - Episode length H
- **Policy:** A policy π consists of H mappings, $\{\pi_h\}_{h=1}^H$, from $\mathcal{S}$ to $\mathcal{A}$
- **Goal:** Find a policy to maximize the return
- **Value function:** Expected accumulative reward for policy π: $V_1^\pi(s) = \mathbb{E}\big[\sum_{h=1}^H r_h(s_h, \pi_h(s_h)) | s_1 = s\big]$
- **Regret:** The sum of sub-optimality over K episodes

$$\text{Regret}(T) = \sum_{k=1}^{K} V_1^*(s_1^k) - V_1^{\pi^k}(s_1^k),$$

where $T = KH$ and $V_1^*(s_t) = \sup_\pi V_1^\pi(s_t)$
- **Adaptivity constraint:** Given the number of episodes K, there is a hard budget B on the number of policy switches: $\sum_{k=1}^{K-1} \mathbb{1}\{\pi^k \neq \pi^{k+1}\} \le B$
- **Batch learning model:** policy switches only happen at prefixed grids $1 = t_1 < \cdots < t_B < t_{B+1} = K + 1$
- **Rare policy switch model:** the agent can adaptively choose when to switch the policy

Assumptions

- **Linear MDPs:** Assume there exist unknown measures $\{\boldsymbol{\mu}_h = (\boldsymbol{\mu}_h^{(1)}, \dots, \boldsymbol{\mu}_h^{(d)})\}_{h=1}^H$, unknown vectors $\{\boldsymbol{\theta}_h\}_{h=1}^H$, and a known feature mapping $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$, s.t.
  - $\mathbb{P}_h(s'|s, a) = \langle \boldsymbol{\phi}(s, a), \boldsymbol{\mu}_h(s') \rangle$
  - $r_h(s, a) = \langle \boldsymbol{\phi}(s, a), \boldsymbol{\theta}_h \rangle$

  for each $h \in [H]$.

Main Results: Batch Learning Model

- **Algorithm** `LSVI-UCB-Batch`

```
Set b ← 1, t_i ← (i − 1)⌊K/B⌋ + 1 (uniform grid)
for episode k = 1, 2, ..., K do
  if k = t_b (time to switch the policy) then
    b ← b + 1, Q^k_{H+1}(·,·) ← 0
    Compute optimistic estimates {Q^k_h}_{h=1}^H by backward regression (Jin et al., 2020)
    Compute greedy policy π^k induced by {Q^k_h}_{h=1}^H
  else
    π^k ← π^{k−1} (keep the current policy)
  Run policy π^k to obtain {(s^k_h, a^k_h, r_h(s^k_h, a^k_h))}_{h=1}^H
```

- **Regret upper bound**

Under technical assumptions and with appropriate choice of parameters, the total regret of `LSVI-UCB-Batch` is bounded by

$$\text{Regret}(T) = \widetilde{O}\left(dHT/B + \sqrt{d^3H^3T}\right)$$

Main Results: Rare Policy Switch Model

- **Algorithm** `LSVI-UCB-RareSwitch`

```
Initialize Λ_h = Λ^0_h = λI_d for all h ∈ [H]
for episode k = 1, 2, ..., K do
  Λ^k_h ← Σ_{τ=1}^{k−1} φ(s^τ_h, a^τ_h) φ(s^τ_h, a^τ_h)^T + λI_d
  if ∃h, det(Λ^k_h) > η det(Λ_h) (criterion) then
    {Λ_h}_{h=1}^H ← {Λ^k_h}_{h=1}^H
    Compute optimistic estimates {Q^k_h}_{h=1}^H by backward regression, update greedy policy π^k
  else
    π^k ← π^{k−1} (keep the current policy)
  Run policy π^k to obtain {(s^k_h, a^k_h, r_h(s^k_h, a^k_h))}_{h=1}^H
```

Main Results: Rare Policy Switch Model (cont.)

- **Regret upper bound**

Under technical assumptions and with appropriate choice of parameters, the total regret of `LSVI-UCB-RareSwitch` satisfies

$$\text{Regret}(T) \le \widetilde{O}\left(\sqrt{d^3H^3T[1 + T/(dH)]^{dH/B}}\right)$$

Discussion

- **Comparison with LSVI-UCB:** To achieve a $\widetilde{O}(\sqrt{d^3H^3T})$ regret which is attained by the original LSVI-UCB algorithm (Jin et al., 2020), the proposed algorithms require a much smaller number of policy switches (K for LSVI-UCB):
  - For `LSVI-UCB-Batch`, $B = \Omega(\sqrt{T/(dH)})$
  - For `LSVI-UCB-RareSwitch`, $B = \Omega(dH \log T)$
- **Regret lower bound** For batch learning model, a complimentary lower bound is proved:

Suppose $B \ge (d - 1)H/2$. Then for any batch learning algorithm with B batches, there exists a linear MDP such that the regret satisfies

$$\text{Regret}(T) = \Omega(dH\sqrt{T} + dHT/B)$$

  - This suggests that the dependency on B in the upper bound for `LSVI-UCB-Batch` is tight
  - It remains an open problem to establish a similar lower bound for the rare policy switch model.

Reference

Jin, C., Yang, Z., Wang, Z. and Jordan, M. I. (2020). Provably efficient reinforcement learning with linear function approximation. In *Conference on Learning Theory*. PMLR.