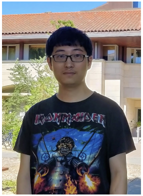
Provably Efficient Reinforcement Learning with Linear Function Approximation under Adaptivity Constraints

Tianhao Wang[1]



Dongruo Zhou[2]

Quanquan Gu[2]

[1]Department of Statistics and Data Science, Yale
[2]Department of Computer Science, UCLA

NeurIPS 2021

Outline

Motivation: adaptivity constraints in Reinforcement Learning

Problem setting

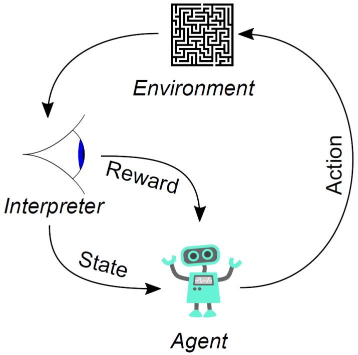
Main results: algorithm and analysis

Numerical experiment

Conclusion

(Online) Reinforcement Learning

In online Reinforcement Learning (RL), one of the most important tasks is to learn the optimal policy which maximizes the long-term cumulative rewards:



By Megajuice - Own work, CC0, https://commons.wikimedia.org/w/index.php?curid=57895741

Adaptivity constraints in RL

Typical online RL algorithm: execute policy $\rightleftharpoons$ update policy

Adaptivity constraints in RL

Typical online RL algorithm: execute policy $\rightleftharpoons$ update policy

- In practice, updating the policy might incur costs and there could be hard budget in this regard
- E.g., clinical trials

Adaptivity constraints in RL

Typical online RL algorithm: execute policy $\rightleftharpoons$ update policy

- In practice, updating the policy might incur costs and there could be hard budget in this regard
- E.g., clinical trials
- It is preferable to switch the policy less frequently instead of episodically

Adaptivity constraints in RL

Typical online RL algorithm: execute policy $\rightleftharpoons$ update policy

- In practice, updating the policy might incur costs and there could be hard budget in this regard
- E.g., clinical trials
- It is preferable to switch the policy less frequently instead of episodically
- The limited adaptivity setting has been studied for many online learning scenarios including PFE (Kalai and Vempala, 2005), MAB (Arora et al., 2012), etc.

Adaptivity constraints in RL

Typical online RL algorithm: execute policy $\rightleftharpoons$ update policy

- In practice, updating the policy might incur costs and there could be hard budget in this regard
- E.g., clinical trials
- It is preferable to switch the policy less frequently instead of episodically
- The limited adaptivity setting has been studied for many online learning scenarios including PFE (Kalai and Vempala, 2005), MAB (Arora et al., 2012), etc.
- A similar concept is known as *low switching cost* in RL (Bai et al., 2019), but the goal there is to achieve $\tilde{O}(\sqrt{K})$ regret with as few policy switches as possible

Our setting: limited number of policy updates

Given the number of episodes K, assume that there is a hard budget B on the number of policy switches:

$$\sum_{k=1}^{K-1} \mathbb{1}\{\pi^k \neq \pi^{k+1}\} \leq B$$

Our setting: limited number of policy updates

Given the number of episodes K, assume that there is a hard budget B on the number of policy switches:

$$\sum_{k=1}^{K-1} \mathbb{1}\{\pi^k \neq \pi^{k+1}\} \leq B$$

We consider two models of interest:

- Batch learning model: policy switches only happen at the prefixed grids $1 = t_1 < \cdots < t_B < t_{B+1} = K + 1$

Our setting: limited number of policy updates

Given the number of episodes K, assume that there is a hard budget B on the number of policy switches:

$$\sum_{k=1}^{K-1} \mathbb{1}\{\pi^k \neq \pi^{k+1}\} \leq B$$

We consider two models of interest:

- Batch learning model: policy switches only happen at the 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

Our setting: limited number of policy updates

Given the number of episodes K, assume that there is a hard budget B on the number of policy switches:

$$\sum_{k=1}^{K-1} \mathbb{1}\{\pi^k \neq \pi^{k+1}\} \leq B$$

We consider two models of interest:

- Batch learning model: policy switches only happen at the 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

We study the above two models in the context of linear MDPs, beyond tabular MDPs studied in Bai et al. (2019)

Linear Markov Decision Process (MDP)

We consider the setting of linear MDP (Yang and Wang, 2019; Jin et al., 2020) where both the transition probabilities and reward functions can be linearly parametrized as

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

Linear Markov Decision Process (MDP)

We consider the setting of linear MDP (Yang and Wang, 2019; Jin et al., 2020) where both the transition probabilities and reward functions can be linearly parametrized as

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

- Any tabular MDP is a linear MDP with one-hot features

Linear Markov Decision Process (MDP)

We consider the setting of linear MDP (Yang and Wang, 2019; Jin et al., 2020) where both the transition probabilities and reward functions can be linearly parametrized as

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

- Any tabular MDP is a linear MDP with one-hot features
- The action-value function $Q_h^\pi(s,a)$ is also linear in the feature mapping ϕ (Jin et al., 2020), i.e., $\exists \mathrm{w}_h^\pi$ s.t.

$$Q_h^\pi(s,a) = \langle \boldsymbol{\phi}(s,a), \mathrm{w}_h^\pi \rangle$$

Linear Markov Decision Process (MDP)

We consider the setting of linear MDP (Yang and Wang, 2019; Jin et al., 2020) where both the transition probabilities and reward functions can be linearly parametrized as

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

- Any tabular MDP is a linear MDP with one-hot features
- The action-value function $Q_h^\pi(s,a)$ is also linear in the feature mapping ϕ (Jin et al., 2020), i.e., $\exists \mathrm{w}_h^\pi$ s.t.

$$Q_h^\pi(s,a) = \langle \boldsymbol{\phi}(s,a), \mathrm{w}_h^\pi \rangle$$

- We adapt the original LSVI-UCB algorithm (Jin et al., 2020) to allow for adaptivity constraints

Batch learning model: `LSVI-UCB-Batch`

Algorithm 1 LSVI-UCB-Batch

Set $b \leftarrow 1,\ t_i \leftarrow (i-1)\lfloor \frac{K}{B} \rfloor + 1, i \in [B]$ (uniform batch grids)
for episode $k = 1, 2, \ldots, K$ **do**
    if $k = t_b$ (time to switch the policy) **then**
        $b \leftarrow b + 1,\ Q^k_{H+1}(\cdot, \cdot) \leftarrow 0$
        Compute optimistic estimates $\{Q^k_h\}$ by backward regression
        Update the greedy policy π^k induced by $\{Q^k_h\}_{h \in [H]}$
    else
        $\pi^k \leftarrow \pi^{k-1}$ (keep the current policy)
    end if
    Run policy π^k to obtain the trajectory $\{(s^k_h, a^k_h, r_h(s^k_h, a^k_h))\}$
end for

- A batched version of the original LSVI-UCB (Jin et al., 2020)

Regret of LSVI-UCB-Batch

Theorem (W., Zhou, Gu)

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

$$\mathit{Regret}(T) \leq \tilde{O}\left(dHT/B + \sqrt{d^3 H^3 T}\right).$$

Regret of `LSVI-UCB-Batch`

Theorem (W., Zhou, Gu)

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

$$Regret(T) \leq \tilde{O}\left(dHT/B + \sqrt{d^3H^3T}\right).$$

- $B = \Omega(\sqrt{\frac{T}{dH}})$ batches suffice to achieve a $\tilde{O}(\sqrt{d^3H^3T})$ regret, which is the same as that of the original LSVI-UCB
- Our algorithm requires much fewer policy switches ($\sqrt{\frac{T}{dH}}$ vs T)

Regret of `LSVI-UCB-Batch`

Theorem (W., Zhou, Gu)

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

$$Regret(T) \leq \tilde{O}\left(dHT/B + \sqrt{d^3H^3T}\right).$$

- $B = \Omega(\sqrt{\frac{T}{dH}})$ batches suffice to achieve a $\tilde{O}(\sqrt{d^3H^3T})$ regret, which is the same as that of the original LSVI-UCB
- Our algorithm requires much fewer policy switches ($\sqrt{\frac{T}{dH}}$ vs T)
- We also provide a $\Omega(dH\sqrt{T} + dHT/B)$ lower bound (for uniform grids), suggesting the above dependency on B is tight

Regret of LSVI-UCB-Batch

Theorem (W., Zhou, Gu)

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

$$\mathit{Regret}(T) \le \tilde{O}\left(dHT/B + \sqrt{d^3H^3T}\right).$$

- $B = \Omega(\sqrt{\frac{T}{dH}})$ batches suffice to achieve a $\tilde{O}(\sqrt{d^3H^3T})$ regret, which is the same as that of the original LSVI-UCB
- Our algorithm requires much fewer policy switches ($\sqrt{\frac{T}{dH}}$ vs T)
- We also provide a $\Omega(dH\sqrt{T} + dHT/B)$ lower bound (for uniform grids), suggesting the above dependency on B is tight
- Can we do better?

Regret of LSVI-UCB-Batch

Theorem (W., Zhou, Gu)

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

$$\mathit{Regret}(T) \leq \tilde{O}\left(dHT/B + \sqrt{d^3H^3T}\right).$$

- $B = \Omega(\sqrt{\frac{T}{dH}})$ batches suffice to achieve a $\tilde{O}(\sqrt{d^3H^3T})$ regret, which is the same as that of the original LSVI-UCB
- Our algorithm requires much fewer policy switches ($\sqrt{\frac{T}{dH}}$ vs T)
- We also provide a $\Omega(dH\sqrt{T} + dHT/B)$ lower bound (for uniform grids), suggesting the above dependency on B is tight
- Can we do better?
- Yes, by using adaptive batch size

Rare policy switch model: `LSVI-UCB-RareSwitch`

Algorithm 2 LSVI-UCB-RareSwitch

Initialize $\mathbf{\Lambda}_h = \mathbf{\Lambda}_h^0 = \lambda \mathbf{I}_d$ for all $h \in [H]$
for episode $k = 1, 2, \ldots, K$ **do**
    $\mathbf{\Lambda}_h^k \leftarrow \sum_{\tau=1}^{k-1} \phi(s_h^\tau, a_h^\tau)\phi(s_h^\tau, a_h^\tau)^\top + \lambda \mathbf{I}_d$ (covariance matrix)
    if $\exists h, \det(\mathbf{\Lambda}_h^k) > \eta \det(\mathbf{\Lambda}_h)$ (trigger policy switch) **then**
        $\{\mathbf{\Lambda}_h\} \leftarrow \{\mathbf{\Lambda}_h^k\}$ (maintain the last covariance matrix)
        Compute optimistic estimates $\{Q_h^k\}$ by backward regression, update the corresponding greedy policy π^k
    else
        $\pi^k \leftarrow \pi^{k-1}$ (keep the current policy)
    end if
    Run policy π^k to obtain the trajectory $\{(s_h^k, a_h^k, r_h(s_h^k, a_h^k))\}$
end for

Rare policy switch model: `LSVI-UCB-RareSwitch`

Algorithm 3 LSVI-UCB-RareSwitch

```
Initialize Λ_h = Λ_h^0 = λI_d for all h ∈ [H]
for episode k = 1, 2, ..., K do
    Λ_h^k ← Σ_{τ=1}^{k-1} φ(s_h^τ, a_h^τ) φ(s_h^τ, a_h^τ)^⊤ + λI_d (covariance matrix)
    if ∃h, det(Λ_h^k) > η det(Λ_h) (trigger policy switch) then
        {Λ_h} ← {Λ_h^k} (maintain the last covariance matrix)
        Compute optimistic estimates {Q_h^k} by backward regression, update the corresponding greedy policy π^k
    else
        π^k ← π^{k-1} (keep the current policy)
    end if
    Run policy π^k to obtain the trajectory {(s_h^k, a_h^k, r_h(s_h^k, a_h^k))}
end for
```

- Related to the doubling trick (Jaksch et al., 2010; Abbasi-Yadkori et al., 2011; Zhou et al., 2021)
- The policy switch slows down as k grows

Regret of LSVI-UCB-RareSwitch

Theorem (W., Zhou, Gu)

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

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

Regret of `LSVI-UCB-RareSwitch`

Theorem (W., Zhou, Gu)

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

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

- $B = \Omega(dH \log T)$ suffices to achieve a $\tilde{O}(\sqrt{d^3 H^3 T})$ regret
- This requires even fewer batches compared with `LSVI-UCB-Batch`, namely $\Omega(dH \log T)$

Regret of LSVI-UCB-RareSwitch

Theorem (W., Zhou, Gu)

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

$$\mathit{Regret}(T) \leq \tilde{O}\left(\sqrt{d^3 H^3 T[1 + T/(dH)]^{dH/B}}\right).$$

- $B = \Omega(dH \log T)$ suffices to achieve a $\tilde{O}(\sqrt{d^3 H^3 T})$ regret
- This requires even fewer batches compared with `LSVI-UCB-Batch`, namely $\Omega(dH \log T)$
- Trade-off between the total regret bound and the number of policy switches

Regret of `LSVI-UCB-RareSwitch`

Theorem (W., Zhou, Gu)

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

$$\mathit{Regret}(T) \leq \tilde{O}\left(\sqrt{d^3 H^3 T[1 + T/(dH)]^{dH/B}}\right).$$

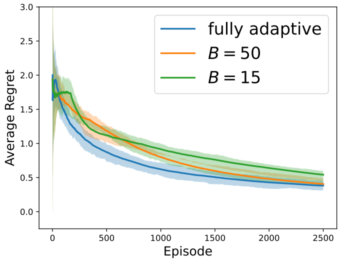
- $B = \Omega(dH \log T)$ suffices to achieve a $\tilde{O}(\sqrt{d^3 H^3 T})$ regret
- This requires even fewer batches compared with `LSVI-UCB-Batch`, namely $\Omega(dH \log T)$
- Trade-off between the total regret bound and the number of policy switches
- When choosing η to be a constant (or equivalently, $B = \Omega(\log T)$), `LSVI-UCB-RareSwitch` reduces to the algorithm studied in Gao et al. (2021)

Numerical experiments

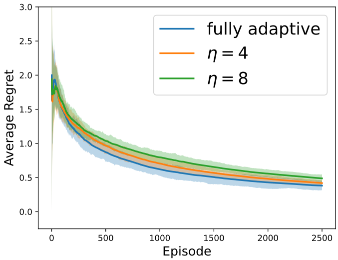
We examine the performance of our algorithms on a hard-to-learn linear MDP instance (Zhou et al., 2021) with $K = 2500$

Numerical experiments

We examine the performance of our algorithms on a hard-to-learn linear MDP instance (Zhou et al., 2021) with $K = 2500$



LSVI-UCB-Batch



LSVI-UCB-RareSwitch

Plot of average regret, Regret$(T)/K$, v.s. the number of episodes. The results are averaged over 50 rounds of each algorithm, and the error bars are the $[20\%, 80\%]$ empirical confidence intervals.

Conclusion

- We study episodic linear MDP under adaptivity constraints
- For the batch learning model, we propose `LSVI-UCB-Batch` which achieves a $\tilde{O}(\sqrt{d^3H^3T} + dHT/B)$ regret (the dependency on B is tight due to a complimentary lower bound)
- For the rare policy switch model, we propose `LSVI-UCB-RareSwitch` which achieves a $\tilde{O}(\sqrt{d^3H^3T[1 + T/(dH)]^{dH/B}})$ regret
- Compared with the fully adaptive LSVI-UCB algorithm (Jin et al., 2020), our algorithms can achieve the same order of regret with much fewer number of policy switches
- Synthetic numerical experiments corroborate our theory

Conclusion

- We study episodic linear MDP under adaptivity constraints
- For the batch learning model, we propose `LSVI-UCB-Batch` which achieves a $\tilde{O}(\sqrt{d^3H^3T} + dHT/B)$ regret (the dependency on B is tight due to a complimentary lower bound)
- For the rare policy switch model, we propose `LSVI-UCB-RareSwitch` which achieves a $\tilde{O}(\sqrt{d^3H^3T[1+T/(dH)]^{dH/B}})$ regret
- Compared with the fully adaptive LSVI-UCB algorithm (Jin et al., 2020), our algorithms can achieve the same order of regret with much fewer number of policy switches
- Synthetic numerical experiments corroborate our theory

Thank you!

ABBASI-YADKORI, Y., PÁL, D. and SZEPESVÁRI, C. (2011). Improved algorithms for linear stochastic bandits. In *Advances in Neural Information Processing Systems*, vol. 24.

ARORA, R., DEKEL, O. and TEWARI, A. (2012). Online bandit learning against an adaptive adversary: from regret to policy regret. *arXiv preprint arXiv:1206.6400* .

BAI, Y., XIE, T., JIANG, N. and WANG, Y.-X. (2019). Provably efficient q-learning with low switching cost. In *Advances in Neural Information Processing Systems*, vol. 32. URL https://proceedings.neurips.cc/paper/2019/file/473803f0f2ebd77d83ee60daaa61f381-Paper.pdf

GAO, M., XIE, T., DU, S. S. and YANG, L. F. (2021). A provably efficient algorithm for linear markov decision process with low switching cost. *arXiv preprint arXiv:2101.00494* .

JAKSCH, T., ORTNER, R. and AUER, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research* **11** 1563–1600.

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.

Kalai, A. and Vempala, S. (2005). Efficient algorithms for online decision problems. *Journal of Computer and System Sciences* **71** 291–307.

Yang, L. and Wang, M. (2019). Sample-optimal parametric q-learning using linearly additive features. In *International Conference on Machine Learning*.

Zhou, D., He, J. and Gu, Q. (2021). Provably efficient reinforcement learning for discounted mdps with feature mapping. In *International Conference on Machine Learning*. PMLR.