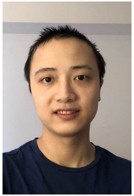
On the Sample Complexity of Learning Infinite-horizon Discounted Linear Kernel MDPs


Yuanzhou Chen [1] Jiafan He [1] Quanquan Gu [1]

[1]Department of Computer Science, UCLA

July 17, 2022

Reinforcement Learning on Discounted MDPs

Discounted Markov Decision Processes (MDP),

$$M(\underbrace{s \in \mathcal{S}}_{\text{state space}}, \underbrace{a \in \mathcal{A}}_{\text{action space}}, \underbrace{\gamma}_{\text{discount factor}}, \underbrace{r(s,a)}_{\text{reward function}}, \underbrace{\mathbb{P}(s'|s,a)}_{\text{transition dynamic}})$$

Starting from s_1, at round t,

- Select action $a_t \leftarrow \pi_t(s_t)$
- Observe reward $r(s_t, a_t)$ and next-state s_{t+1}

Reinforcement Learning on Discounted MDPs

Discounted Markov Decision Processes (MDP),

$$M(\underbrace{s \in \mathcal{S}}_{\text{state space}}, \underbrace{a \in \mathcal{A}}_{\text{action space}}, \underbrace{\gamma}_{\text{discount factor}}, \underbrace{r(s,a)}_{\text{reward function}}, \underbrace{\mathbb{P}(s'|s,a)}_{\text{transition dynamic}})$$

Starting from s_1, at round t,

- Select action $a_t \leftarrow \pi_t(s_t)$
- Observe reward $r(s_t, a_t)$ and next-state s_{t+1}

Goal: to find (non-stationary) policy $\pi = (\pi_t)_t$ to maximize the value function $V_t^\pi(s_t)$, where $a_i \sim \pi_i(s_i)$,

$$V_t^\pi(s_t) = Q_t^\pi(s_t, \pi_t(s_t)),\ Q_t^\pi(s,a) = \mathbb{E}\left[\sum_{i=0}^{\infty} \gamma^i r(s_{t+i}, a_{t+i}) \middle| s_t = s, a_t = a\right]$$

RL with Linear Function Approximation

- Tradition tabular reinforcement algorithms
  - Value function $Q(s, a)$ can be represented as a table
- Limitation: Inefficient when $|\mathcal{S}|$ or $|\mathcal{A}|$ is large (e.g. $|\mathcal{S}| = \Omega(2^{100})$)

RL with Linear Function Approximation

- Tradition tabular reinforcement algorithms
  - Value function $Q(s, a)$ can be represented as a table
- Limitation: Inefficient when $|\mathcal{S}|$ or $|\mathcal{A}|$ is large (e.g. $|\mathcal{S}| = \Omega(2^{100})$)
- Solution: Use **linear function to approximate** the underlining discounted MDPs

Definition (Linear Kernel MDPs Zhou et al. 2021; Ayoub et al. 2020)

MDP $\mathcal{M}$ is linear kernel MDP if there exists a *known* feature mapping $\boldsymbol{\phi}(\cdot|\cdot,\cdot) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$ and an *unknown* vector $\boldsymbol{\theta} \in \mathbb{R}^d$, such that

$$\mathbb{P}(s'|s, a) = \langle \boldsymbol{\phi}(s'|s, a), \boldsymbol{\theta}^* \rangle, \boldsymbol{\phi}(s'|s, a) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to \mathbb{R}^d$$

PAC-bound Guarantee for discounted MDPs

Definition ((ϵ, δ)-PAC-bounds)

For an RL algorithm **Alg** and a fixed ϵ, let $\pi_t (t \in \mathbb{N})$ be the policies generated by **alg** at round t. Let $N_\epsilon = \sum_{t=1}^{\infty} \mathbb{1}\{V_t^*(s_t) - V_t^{\pi_t}(s_t) > \epsilon\}$ be the number of rounds whose suboptimality gap is greater than ϵ. Then we say **alg** is (ϵ, δ)-PAC with sample complexity $f(\epsilon, \delta)$ if

$$\mathbb{P}(N_\epsilon > f(\epsilon, \delta)) \leq \delta.$$

- Widely used performance measure for tabular discounted MDPs
- Only have regret guarantee for discounted MDPs with linear function approximation

PAC-bound Guarantee for discounted MDPs

Definition ((ϵ, δ)-PAC-bounds)

For an RL algorithm **Alg** and a fixed ϵ, let $\pi_t (t \in \mathbb{N})$ be the policies generated by **alg** at round t. Let $N_\epsilon = \sum_{t=1}^{\infty} \mathbb{1}\{V_t^*(s_t) - V_t^{\pi_t}(s_t) > \epsilon\}$ be the number of rounds whose suboptimality gap is greater than ϵ. Then we say **alg** is (ϵ, δ)-PAC with sample complexity $f(\epsilon, \delta)$ if

$$\mathbb{P}(N_\epsilon > f(\epsilon, \delta)) \leq \delta.$$

- Widely used performance measure for tabular discounted MDPs
- Only have regret guarantee for discounted MDPs with linear function approximation

Efficient algorithms for discounted MDPs with linear function approximation to provide sample complexity guarantee?

UPAC-UCLK Algorithm

Uniform-PAC UCLK algorithm needs ...

Multi-level partition scheme

Confidence sets of $\boldsymbol{\theta}^*$

UPAC-UCLK Algorithm

Uniform-PAC UCLK algorithm needs ...

Multi-level partition scheme

Confidence sets of $\boldsymbol{\theta}^*$

In round t, UPAC-UCLK maintains confidence sets $\{\mathcal{C}_l\}_{l=1}^L \ni \boldsymbol{\theta}^*$ and ...

- Run Multi-level extended value iteration (ML-EVI) over $\{\mathcal{C}_l\}_{l=1}^L$, set optimistic estimations $\{Q_l\}_{l=1}^L \leftarrow$ ML-EVI$(\{\mathcal{C}_l\}_{l=1}^L)$
- Select action $a_t \leftarrow \mathrm{argmax}_a \min_{1 \le l} Q_l(s_t, a)$, observe reward and next-state s_{t+1}
- Find the minimum level l_t that $\|\boldsymbol{\phi}_{V_t}(s_t, a_t)\|_{(\boldsymbol{\Sigma}^{l_t})^{-1}} \ge 2^{-l_t}\sqrt{d}/(1-\gamma)$ update the covaraince matrix for level $l \ge l_t$ with discounted Data Inheritance.

$$\boldsymbol{\Sigma}^l \leftarrow \boldsymbol{\Sigma}^l + 2^{l_t - l}\boldsymbol{\phi}_{V_t}(s_t, a_t)\boldsymbol{\phi}_{V_t}(s_t, a_t)^\top.$$

- Update the confidence sets $\{\mathcal{C}_l\}_{l=1}^L \ni \boldsymbol{\theta}^*$ with updated covariance matrix.

Our Results for Linear Kernel MDPs

Theorem (Regret upper bound for linear kernel MDPs)

With high probability, the number of rounds in Algorithm UPAC-UCLK which have sub-optimality no less than ϵ is bounded by

$$\Gamma\big(1/\epsilon, \log(1/\delta); \gamma, d\big) = \widetilde{O}\bigg(\frac{1}{(1-\gamma)^6\epsilon^2} + \frac{d^2 + d\log(1/\delta)}{(1-\gamma)^4\epsilon^2}\bigg),$$

where d is the dimension of feature mapping and γ is the discount factor.

Our Results for Linear Kernel MDPs

Theorem (Regret upper bound for linear kernel MDPs)

With high probability, the number of rounds in Algorithm UPAC-UCLK which have sub-optimality no less than ϵ is bounded by

$$\Gamma\big(1/\epsilon, \log(1/\delta); \gamma, d\big) = \widetilde{O}\bigg(\frac{1}{(1-\gamma)^6\epsilon^2} + \frac{d^2 + d\log(1/\delta)}{(1-\gamma)^4\epsilon^2}\bigg),$$

where d is the dimension of feature mapping and γ is the discount factor.

- First sample complexity guarantee for discounted MDPs with linear function approximation
- Can further provide uniform-PAC guarantee for discounted MDPs
  - Strictly **stronger** than PAC-bound and regert
  - Guarantees the **convergence to the optimal policy**

Thank you!

Reference I

Ayoub, Alex et al. (2020). “Model-Based Reinforcement Learning with Value-Targeted Regression”. In: *arXiv preprint arXiv:2006.01107*.

Zhou, Dongruo, Jiafan He, and Quanquan Gu (2021). “Provably Efficient Reinforcement Learning for Discounted MDPs with Feature Mapping”. In: *International Conference on Machine Learning*. PMLR.