Achieving Constant Regret in Linear Markov Decision Processes

Weitong Zhang* [1] Zhiyuan Fan* [2] Jiafan He [3] Quanquan Gu [3]

*Equal Contribution [1]UNC-Chapel Hill [2]MIT [3]UCLA

Main question

Can we design RL algorithm that incurs only constant regret?

What is the minimal assumptions to achieve constant regret?

Key Results

- First gap-dependent constant regret bound for linear MDP without structural assumptions.

Algorithm	Allow Misspecification?	Result
`LSVI-UCB` [He et al., 2021]	×	$\widetilde{\mathcal{O}}(d^3H^5\Delta^{-1}\log(K))$
`LSVI-UCB` [Papini et al., 2021]	×	$\widetilde{\mathcal{O}}(d^3H^5\Delta^{-1}\log(1/\lambda))$
`Cert-LSVI-UCB` **(this paper)**	✓	$\widetilde{\mathcal{O}}(d^3H^5\Delta^{-1})$

Table 1. Comparison of instance-dependent regret bounds. [He et al., 2021] achieves a regret bound logarithmic in the number of episodes K, while [Papini et al., 2021] achieves a regret bound logarithmic in the minimal eigenvalue λ. Both previous algorithms do not work in the presence of misspecification ($\zeta > 0$).

Problem Setting

Misspecified Linear Markov Decision Process with **state space** $\mathcal{S}$, **action space** $\mathcal{A}$:

- In each stage $h \in [\![H]\!]$, the algorithm picks an action $a \in \mathcal{A}$ based on the current state $s \in \mathcal{S}$.
- Receives immediate reward $r_h(s,a)$ and observes next state $s' \sim \mathbb{P}_h(s,a)$
- **Misspecification level** ζ of known feature map $\boldsymbol{\phi} : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}^d$ with unknown parameters $\boldsymbol{\theta}_h, \boldsymbol{\mu}_h$.

$$\|\mathbb{P}_h(\cdot|s,a) - \langle \boldsymbol{\phi}(s,a), \boldsymbol{\mu}_h(\cdot)\rangle\|_{\mathrm{TV}} \le \zeta, \qquad |h(s,a) - \langle \boldsymbol{\phi}(s,a), \boldsymbol{\theta}_h\rangle| \le \zeta$$

- **Minimal sub-Optimality gap** Δ of the optimal policy π^*:

$$\Delta := \min\{\Delta_h(s,a) : \Delta_h(s,a) > 0\}, \qquad \Delta_h(s,a) := V^{\pi^*}(s) - Q_h^{\pi^*}(s,a)$$

Goal: Find policy $\pi : \mathcal{S} \times [\![H]\!] \to \mathcal{A}$ maximizing expected reward $V^{\pi}(s_1)$. Performance measured via **regret:**

$$\mathrm{Regret}(K) := \sum_{k=1}^{K} \left(V^{\pi^*}(s_1) - V^{\pi^k}(s_1)\right)$$

Main Theorem

[Theorem 5.1] When $\zeta/\Delta < \widetilde{\mathcal{O}}(1/(\sqrt{d}H^2))$, `Cert-LSVI-UCB` achieves a worst-case regret bound of

$$\mathrm{Regret}(K) \le \widetilde{\mathcal{O}}(d^3H^5\Delta^{-1})$$

with high probability, even when run indefinitely as $K \to \infty$.

Proposed Algorithm: `Cert-LSVI-UCB`

Set $V_{H+1}^k(s) = 0$ for all $(s,k) \in \mathcal{S} \times [K]$, $\mathcal{C}_{h,l}^k = \varnothing$ for all $(h,l) \in [H] \times \mathbb{N}^+$, $\lambda = 16$
for episode $k = 1, \cdots, K$ **do**
  Set $L_k = \max\{\lceil \log_4(k/d) \rceil, 0\}$
  for step $h = H, \cdots, 1$ **do**
    for phase $l = 1, \cdots, L_k + 1$ **do**
      $\mathbf{U}_{h,l}^k = \lambda \mathbf{I} + \sum_{\tau \in \mathcal{C}_{h,l}^{k-1}} \boldsymbol{\phi}_h^\tau (\boldsymbol{\phi}_h^\tau)^\top$
      $\mathbf{w}_{h,l}^k = (\mathbf{U}_{h,l}^k)^{-1} \sum_{\tau \in \mathcal{C}_{h,l}^{k-1}} \boldsymbol{\phi}_h^\tau \left(r_h^\tau + \widehat{V}_{h+1}^k(s_{h+1}^\tau)\right)$
      $\widetilde{\mathbf{U}}_{h,l}^{k,-1} = \kappa_l \lceil (\mathbf{U}_{h,l}^k)^{-1}/\kappa_l \rfloor$, $\widetilde{\mathbf{w}}_{h,l}^k = \kappa_l \lceil \mathbf{w}_{h,l}^k/\kappa_l \rfloor$ where $\kappa_l = 0.01 \cdot 2^{-4l} d^{-1}$
    end for
    $\widehat{V}_h^k(s_h^\tau), \cdot, \cdot, \cdot = \texttt{Cert-LinUCB}(s_h^\tau; \{\widetilde{\mathbf{w}}_{h,l}^k\}_l, \{\widetilde{\mathbf{U}}_{h,l}^{k,-1}\}_l, L_k)$ for all $\tau \in [k-1]$
  end for
  Observe $s_1^k \in \mathcal{S}$
  for step $h = 1, \cdots, H$ **do**
    $\cdot, \pi_h^k(s_h^k), l_h^k(s_h^k), f_h^k(s_h^k) = \texttt{Cert-LinUCB}(s_h^k; \{\widetilde{\mathbf{w}}_{h,l}^k\}_l, \{\widetilde{\mathbf{U}}_{h,l}^{k,-1}\}_l, L_k)$
    $\mathcal{C}_{h,l_h^k(s_h^k)}^k = \mathcal{C}_{h,l_h^k(s_h^k)}^{k-1} \cup \{k\}$ **if** $f_h^k(s_h^k) = 1$ **else** $\mathcal{C}_{h,l_h^k(s_h^k)}^{k-1}$
    $\mathcal{C}_{h,l}^k = \mathcal{C}_{h,l}^{k-1}$ for all $l \ne l_h^k(s_h^k)$
    Play $\pi_h^k(s_h^k)$, set $\boldsymbol{\phi}_h^k = \boldsymbol{\phi}(s_h^k, \pi_h^k(s_h^k))$, receive r_h^k and observe $s_{h+1}^k \in \mathcal{S}$
  end for
end for

Auxiliary Algorithm: `Cert-LinUCB`

input: $s \in \mathcal{S}, \forall l : \widetilde{\mathbf{w}}_{h,l}^k \in \mathbb{R}^d, \widetilde{\mathbf{U}}_{h,l}^{k,-1} \in \mathbb{R}^{d \times d}, L \in \mathbb{N}^+$
output: $\widehat{V}_h^k(s) \in \mathbb{R}, \pi_h^k(s) \in \mathcal{A}, l_h^k(s) \in \mathbb{N}^+, f_h^k(s) \in \{0,1\}$
$\mathcal{A}_{h,1}^k(s) = \mathcal{A}, \widetilde{V}_{h,0}^k(s) = 0, \widehat{V}_{h,0}^k(s) = H$
for phase $l = 1, \cdots, L+1$ **do**
  Set $Q_{h,l}^k(s,a) = \langle \boldsymbol{\phi}(s,a), \widetilde{\mathbf{w}}_{h,l}^k \rangle$
  Set $\pi_{h,l}^k(s) = \mathrm{argmax}_{a \in \mathcal{A}_{h,l}^k} Q_{h,l}^k(s,a)$, $V_{h,l}^k(s) = Q_{h,l}^k(s, \pi_{h,l}^k(s))$
  if $l > L$ **then**
    return $(\widehat{V}_h^k(s), \pi_h^k(s), l_h^k(s), f_h^k(s)) = (\widehat{V}_{h,l-1}^k(s), \pi_{h,l-1}^k(s), l, 1)$
  else if $\gamma_l \cdot \max_{a \in \mathcal{A}_{h,l}^k(s)} \|\boldsymbol{\phi}(s,a)\|_{\widetilde{\mathbf{U}}_{h,l}^{k,-1}} \ge 2^{-l}$ **then**
    return $(\widehat{V}_h^k(s), \pi_h^k(s), l_h^k(s), f_h^k(s)) = (\widehat{V}_{h,l-1}^k(s), \mathrm{argmax}_{a \in \mathcal{A}_{h,l}^k(s)} \|\boldsymbol{\phi}(s,a)\|_{\widetilde{\mathbf{U}}_{h,l}^{k,-1}}, l, 1)$
  else if $\max\{V_{h,l}^k(s) - 3 \cdot 2^{-l}, \widetilde{V}_{h,l-1}^k(s)\} > \min\{V_{h,l}^k(s) + 3 \cdot 2^{-l}, \widetilde{V}_{h,l-1}^k(s)\}$ **then**
    return $(\widehat{V}_h^k(s), \pi_h^k(s), l_h^k(s), f_h^k(s)) = (\widehat{V}_{h,l-1}^k(s), \pi_{h,l-1}^k(s), l, 0)$
  else
    $\widehat{V}_{h,l}^k(s) = \min\{V_{h,l}^k(s) + 3 \cdot 2^{-l}, \widehat{V}_{h,l-1}^k(s)\}$
    $\widetilde{V}_{h,l}^k(s) = \max\{V_{h,l}^k(s) - 3 \cdot 2^{-l}, \widetilde{V}_{h,l-1}^k(s)\}$
    $\mathcal{A}_{h,l+1}^k(s) = \left\{a \in \mathcal{A}_{h,l}^k(s) : Q_{h,l}^k(s,a) \ge V_{h,l}^k(s) - 4 \cdot 2^{-l}\right\}$
  end if
end for

Technical highlight

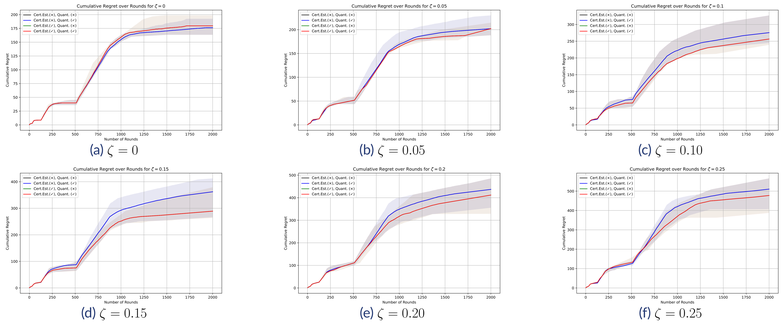
Multi-Phase Arm-Elimination

- The estimation precision in phase l is $\mathcal{O}(2^{-l} + \sqrt{l}\zeta)$.
- Arm-elimination ensures that the selected actions are guaranteed by the optimal phase l.

Certified Estimator

- Arm-elimination introduces uncontrolled error due to the misspecification ζ in each phase.
- Certified estimator prevents the error from accumulating across phases.

Numerical Verification



(a) $\zeta = 0$ (b) $\zeta = 0.05$ (c) $\zeta = 0.10$ (d) $\zeta = 0.15$ (e) $\zeta = 0.20$ (f) $\zeta = 0.25$

Figure 1. Cumulative regret with respect to the number of episodes under different misspecification level ζ. **Certified Estimator** plays an important role when misspecification ζ is large.

Conclusion

- A high probability constant regret bound for linear MDP
- An interplay between misspefication level and sub-optimality gap
- A **Certified Estimator** for achieving these robust results.

Checkout our paper on OpenReview! ⟹

Jiafan He, Dongruo Zhou, and Quanquan Gu. Logarithmic regret for reinforcement learning with linear function approximation. In *International Conference on Machine Learning*, pages 4171–4180. PMLR, 2021.

Matteo Papini, Andrea Tirinzoni, Aldo Pacchiano, Marcello Restelli, Alessandro Lazaric, and Matteo Pirotta. Reinforcement learning in linear mdps: Constant regret and representation selection. *Advances in Neural Information Processing Systems*, 34:16371–16383, 2021.

weitongz@unc.edu, fanzy@mit.edu