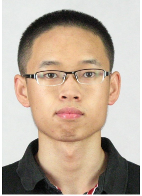
Tightening the Dependence on Horizon in the Sample Complexity of Q-Learning

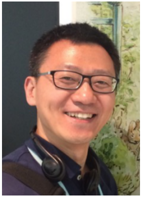
Changxiao Cai

ECE, Princeton University



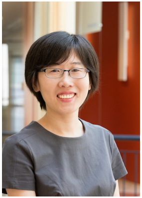
Gen Li
Tsinghua EE

Yuxin Chen
Princeton ECE



Yuantao Gu
Tsinghua EE

Yuting Wei
CMU Stats



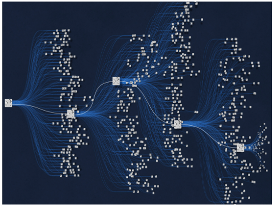
Yuejie Chi
CMU ECE

G. Li, C Cai, Y. Chen, Y. Gu, Y. Wei, and Y. Chi, “Tightening the Dependence on Horizon in the Sample Complexity of Q-Learning,” ICML2021

Sample-efficient reinforcement learning (RL)

In RL, an agent learns by interacting with an environment

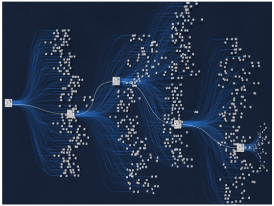
- Collecting data samples might be expensive or time-consuming



Sample-efficient reinforcement learning (RL)

In RL, an agent learns by interacting with an environment

- Collecting data samples might be expensive or time-consuming



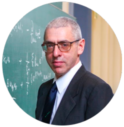
Calls for in-depth understanding about sample efficiency of RL algorithms

Q-learning: a classical model-free algorithm

γ-discounted infinite horizon MDP

- $Q^{\star}$: optimal action-value function
- $\mathcal{S}$: state space; $\mathcal{A}$: action space
- $r \in [0, 1]$: reward function

Chris Watkins



Peter Dayan

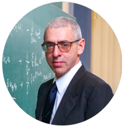
Stochastic approximation for solving Bellman equation $Q = \mathcal{T}(Q)$

Q-learning: a classical model-free algorithm

γ-discounted infinite horizon MDP

- $Q^\star$: optimal action-value function
- $\mathcal{S}$: state space; $\mathcal{A}$: action space
- $r \in [0,1]$: reward function

Chris Watkins



Peter Dayan

Stochastic approximation for solving Bellman equation $Q = \mathcal{T}(Q)$

$$Q_{t+1}(s,a) = (1-\eta_t)Q_t(s,a) + \eta_t \mathcal{T}_t(Q_t)(s,a), \quad t \geq 0$$

$$\mathcal{T}_t(Q)(s,a) := r(s,a) + \gamma \max_{a'} Q(s'_t, a')$$

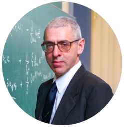
$$\mathcal{T}(Q)(s,a) := r(s,a) + \gamma \mathop{\mathbb{E}}_{s' \sim P(\cdot|s,a)} \Big[\max_{a'} Q(s', a')\Big]$$

Q-learning: a classical model-free algorithm

γ-discounted infinite horizon MDP

- $Q^{\star}$: optimal action-value function
- $\mathcal{S}$: state space; $\mathcal{A}$: action space
- $r \in [0, 1]$: reward function

Chris Watkins



Peter Dayan

Synchronous setting: in every iteration, draw a sample transition for each state-action pair, and update all state-action pairs at once

What is sample complexity of synchronous Q-learning?

A highly incomplete list of prior work

- Watkins, Dayan '92
- Tsitsiklis '94
- Jaakkola, Jordan, Singh '94
- Szepesvári '98
- Kearns, Singh '99
- Borkar, Meyn '00
- Even-Dar, Mansour '03
- Beck, Srikant '12
- Jin, Allen-Zhu, Bubeck, Jordan '18
- Shah, Xie '18
- Lee, He '18
- Wainwright '19
- Chen, Zhang, Doan, Maguluri, Clarke '19
- Yang, Wang '19
- Du, Lee, Mahajan, Wang '20
- Chen, Maguluri, Shakkottai, Shanmugam '20
- Qu, Wierman '20
- Devraj, Meyn '20
- Weng, Gupta, He, Ying, Srikant '20
- ...

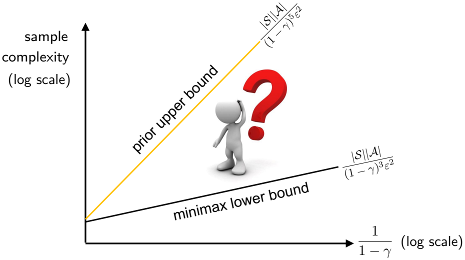
Prior art: achievability

Question: how many samples are needed to ensure $\|\widehat{Q} - Q^{\star}\|_{\infty} \leq \varepsilon$?

Prior art: achievability

Question: how many samples are needed to ensure $\|\widehat{Q} - Q^{\star}\|_{\infty} \leq \varepsilon$?

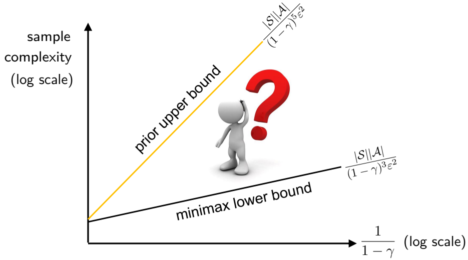
paper	sample complexity
Even-Dar & Mansour '03	$2^{\frac{1}{1-\gamma}} \frac{\|\mathcal{S}\|\|\mathcal{A}\|}{(1-\gamma)^4 \varepsilon^2}$
Beck & Srikant '12	$\frac{\|\mathcal{S}\|^2\|\mathcal{A}\|^2}{(1-\gamma)^5 \varepsilon^2}$
Wainwright '19	$\frac{\|\mathcal{S}\|\|\mathcal{A}\|}{(1-\gamma)^5 \varepsilon^2}$
Chen et al. '20	$\frac{\|\mathcal{S}\|\|\mathcal{A}\|}{(1-\gamma)^5 \varepsilon^2}$



Prior art: achievability

Question: how many samples are needed to ensure $\|\widehat{Q} - Q^{\star}\|_\infty \leq \varepsilon$?

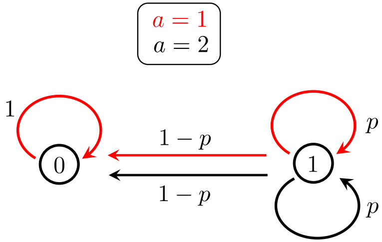
paper	sample complexity
Even-Dar & Mansour '03	$2^{\frac{1}{1-\gamma}} \frac{\|\mathcal{S}\|\|\mathcal{A}\|}{(1-\gamma)^4\varepsilon^2}$
Beck & Srikant '12	$\frac{\|\mathcal{S}\|^2\|\mathcal{A}\|^2}{(1-\gamma)^5\varepsilon^2}$
Wainwright '19	$\frac{\|\mathcal{S}\|\|\mathcal{A}\|}{(1-\gamma)^5\varepsilon^2}$
Chen et al. '20	$\frac{\|\mathcal{S}\|\|\mathcal{A}\|}{(1-\gamma)^5\varepsilon^2}$

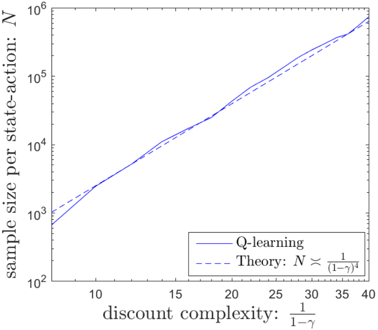


All prior results require sample size of at least $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^5\varepsilon^2}$!

Conjecture: Wainwright '19

Numerical evidence: $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}$ samples seem sufficient ...





Main result: sharpened upper bound

Theorem 1 (Li, Cai, Chen, Gu, Wei, Chi '21)

For any $0 < \varepsilon \leq 1$, sample complexity of sync Q-learning to yield $\|\widehat{Q} - Q^{\star}\|_{\infty} \leq \varepsilon$ is at most (up to log factor)

$$\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4 \varepsilon^2}$$

- Improves dependency on effective horizon $\frac{1}{1-\gamma}$

Main result: sharpened upper bound

Theorem 1 (Li, Cai, Chen, Gu, Wei, Chi '21)

For any $0 < \varepsilon \le 1$, *sample complexity of sync Q-learning to yield* $\|\widehat{Q} - Q^\star\|_\infty \le \varepsilon$ *is at most (up to log factor)*

$$\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4 \varepsilon^2}$$

- Improves dependency on effective horizon $\frac{1}{1-\gamma}$
- Holds for both constant and rescaled linear learning rates

Main result: matching lower bound

Theorem 2 (Li, Cai, Chen, Gu, Wei, Chi '21)

For any $0 < \varepsilon \leq 1$, there exist an MDP s.t. sample complexity of sync Q-learning to yield $\|\widehat{Q} - Q^{\star}\|_{\infty} \leq \varepsilon$ is at least (up to log factor)

$$\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}$$

- Tight algorithm-dependent lower bound

Main result: matching lower bound

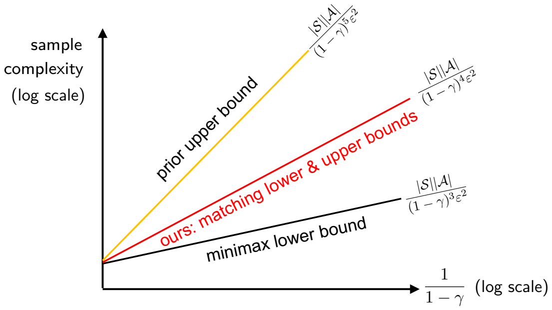
Theorem 2 (Li, Cai, Chen, Gu, Wei, Chi '21)

For any $0 < \varepsilon \leq 1$, there exist an MDP s.t. sample complexity of sync Q-learning to yield $\|\widehat{Q} - Q^{\star}\|_{\infty} \leq \varepsilon$ is at least (up to log factor)

$$\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4 \varepsilon^2}$$

- Tight algorithm-dependent lower bound
- Holds for both constant and rescaled linear learning rates

Takeaway message



- **Sharpens** sample complexity of sync Q-learning: $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^4\varepsilon^2}$
- Uncovers that vanilla Q-learning is **NOT minimax optimal**

  — minimax lower bound: $\frac{|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^3\varepsilon^2}$ (Azar et al '13)

Thanks for your attention!

G. Li, C Cai, Y. Chen, Y. Gu, Y. Wei, and Y. Chi, "Tightening the Dependence on Horizon in the Sample Complexity of Q-Learning," ICML2021

G. Li, C Cai, Y. Chen, Y. Gu, Y. Wei, and Y. Chi, "Is Q-learning minimax optimal? a tight sample complexity analysis," arXiv:2102.06548