Faster Non-asymptotic Convergence for Double Q-learning

Lin Zhao[1], Huaqing Xiong[2], Yingbin Liang[2]

NeurIPS 2021

1 National University of Singapore

2 The Ohio State University

2 Background

Vanilla Q-learning:

- Overestimation → volatile learning error & slow convergence
- max of sampled Q-function > max of expected Q-function

$$\max_{a' \in \mathcal{A}} Q(s', a') \qquad \boxed{\mathbb{E}_{s' \sim \mathbb{P}(\cdot|s,a)} \left[R_{sa}^{s'} + \gamma \max_{a' \in \mathcal{A}} Q(s', a') \right]}$$

Double Q-learning:

- Use two Q-estimators to reduce the overestimation

$$\begin{cases} Q_{t+1}^A(s,a) = (1 - \alpha\beta_t) Q_t^A(s,a) + \alpha\beta_t \left(R_t(s,a,s') + \gamma Q_t^B(s', a^*) \right) \\ Q_{t+1}^B(s,a) = (1 - \alpha(1-\beta_t)) \, Q_t^B(s,a) + \alpha(1-\beta_t) \left(R_t(s,a,s') + \gamma Q_t^A(s', b^*) \right) \end{cases}$$

At each iteration, randomly choose A or B to update.

$$\begin{cases} a^* = \arg\max_{a \in \mathcal{A}} Q^A(s', a) \\ b^* = \arg\max_{a \in \mathcal{A}} Q^B(s', a) \end{cases}$$

Problem Setup

- Discounted reward MDP: $\lambda \in (0,1)$, finite state-action space $D \coloneqq |\mathcal{S}| \times |\mathcal{A}|$

- Random reward: $R_t \in [0,1]$, constant step size/learning rate: $\alpha \in (0,1)$

- Sampling schemes:
  - Synchronous sampling (SynDQ): at each iteration, all state-action pairs are updated
  - Asynchronous sampling (AsynDQ): sample only one pair from a single Markovian trajectory to update

- Optimal Q-function Q^*: the unique solution of the Bellman equation

- Non-asymptotic convergence: how the learning error $\left\|Q_T^A - Q^*\right\|$ converges as a function of the iteration number T

4 SynDQ

Theorem (finite-time bound): with probability at least $1 - \delta$, the learning error $r_t := Q_t^A - Q^*$ satisfies

$$\|r_{t+1}\| \leq \boxed{h^t \|r_1\|} + \boxed{\frac{c}{(1-\gamma)^3}\sqrt{\alpha \ln \frac{2D}{\delta}},}$$

for all $t \geq 1$, where $h = 1 - \frac{1-\gamma}{2}\alpha$.

- Initialization error diminishes linearly; constant error scales as $\sqrt{\alpha}$. (Trade-off)

Corollary (sample complexity): $\forall \epsilon \in (0, \frac{1}{1-\gamma}]$, we have $\mathbb{P}(\|Q_T^A - Q^*\| \leq \epsilon) \geq 1 - \delta$, given

$$T(\epsilon, \gamma, \delta, D) = \tilde{\Omega}\left(\frac{\ln \frac{D}{\delta}}{(1-\gamma)^7 \epsilon^2}\right)$$

- Orders are tight in ϵ (up to logarithm factor), δ, and D, matching the lower bound Azar et al. (2013)

SynDQ

- Significantly improves Xiong et al. (2020) on major parameters $(\epsilon, 1-\gamma, D)$

SyncDQ	Stepsize	Time complexity†	
Xiong et al. (2020)	$\frac{1}{t^\omega}, \omega \in (\frac{1}{3}, 1)$	$\omega = 1-\eta \to 1$ $\tilde{\Omega}\left(\boxed{\frac{1}{\epsilon^{2+\eta}}} \vee \left(\frac{1}{1-\gamma}\right)^{\frac{1}{\eta}}\right)$	$\omega = 6/7$ $\tilde{\Omega}\left(\frac{1}{(1-\gamma)^7}\left(\boxed{\frac{1}{\epsilon^{3.5}}} \vee \left(\ln\frac{1}{1-\gamma}\right)^7\right)\right)$
This work	$\epsilon^2(1-\gamma)^6$	$\tilde{\Omega}\left(\frac{1}{\epsilon^2}\right)$	$\tilde{\Omega}\left(\frac{1}{(1-\gamma)^7\epsilon^2}\right)$

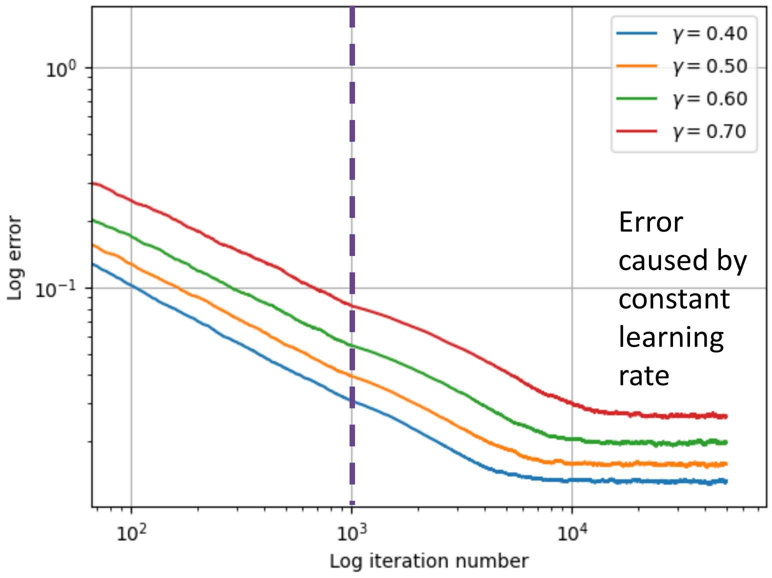
† The choices $\omega \to 1$ and $\omega = \frac{6}{7}$ optimize the dependence of time complexity on ϵ and $1-\gamma$ in Xiong et al. (2020) respectively. $a \vee b = \max\{a, b\}$.

SynDQ

Simulation Example

- Example adapted from Wainwright (2019b)
- Each curve is averaged over 1000 independent runs.
- Slope $\approx -\frac{1}{2}$, matches our analysis of $T = \mathcal{O}(\frac{1}{\epsilon^2})$
- Initially we use rescaled linear step size to reduce the initialization error. We switch to a constant step size of 0.001 after $T = 10^3$.

Error scaling versus iteration/discount



SynDQ

Proof sketch:

1. Reformulate double-Q as a pair of nested stochastic approximations (SA)
   - Inner SA: $\|Q_t^B - Q_t^A\|$ dynamics
   - Outer SA: $\|Q_t^A - Q^*\|$ dynamics, which takes the output of inner SA as an input
   - The two SAs have similar structures.
2. Derive a template finite-time bound applicable to both SAs
   - Per-iteration bound
   - Adapt the sandwich bound in Wainwright (2019b) and requires less assumptions
3. Construct martingales specific to each SA and apply Azuma-Hoeffding inequality to establish the finite-time bounds
4. Obtain the overall bound

AsynDQ

- $\forall \epsilon \in (0, \frac{1}{1-\gamma}]$, we have $\mathbb{P}(\|Q_T^A - Q^*\| \le \epsilon) \ge 1-\delta$, given

$$T = \tilde{\Omega}\left(\frac{L}{\epsilon^2(1-\gamma)^7} \ln \frac{1}{\epsilon(1-\gamma)^2}\right)$$

- Significantly improves Xiong et al. (2020) on major parameters $(\epsilon, 1-\gamma, L)$:

AsyncDQ	Stepsize	Time complexity †		
Xiong et al. (2020)	$\frac{1}{t^\omega}, \omega \in (\frac{1}{3}, 1)$	$\omega = 1-\eta \to 1$	$\omega = 6/7$	$\omega = 2/3$
		$\tilde{\Omega}\left(\boxed{\frac{1}{\epsilon^{2+\eta}}} \vee \left(\frac{1}{1-\gamma}\right)^{\frac{1}{\eta}}\right)$	$\tilde{\Omega}\left(\frac{1}{(1-\gamma)^7}\left(\boxed{\frac{1}{\epsilon^{3.5}}} \vee \left(\ln \frac{1}{1-\gamma}\right)^7\right)\right)$	$\tilde{\Omega}\left(\frac{L^6(\ln L)^{1.5}}{(1-\gamma)^9 \boxed{\epsilon^3}}\right)$
This work	$\epsilon^2(1-\gamma)^6$	$\tilde{\Omega}\left(\frac{1}{\epsilon^2}\right)$	$\tilde{\Omega}\left(\frac{1}{(1-\gamma)^7\epsilon^2}\right)$	$\tilde{\Omega}\left(\frac{L}{(1-\gamma)^7\epsilon^2}\right)$

† The choices $\omega \to 1, \omega = \frac{6}{7}$, and $\omega = \frac{2}{3}$ optimize the dependence of time complexity on $\epsilon, 1-\gamma$, and L in Xiong et al. (2020), respectively. In addition, we denote $a \vee b = \max\{a, b\}$.

AsynDQ

The analysis is more challenging:

- coupling between random switching of Q-estimators and Markovian sampling

Some of the Key steps include:

1. Capture the learning error in terms of key noise and error terms over all the preceding iterations
2. Construct an auxiliary Markov chain to derive a concentration inequality of the visitation probability
    - Enables a per-frame analysis adapted from Li et al. (2020) (the frame length determined by visitation probability)
3. Construct martingales for bounding learning errors using a conditional concentration analysis

Summary

This work:

- Tighter characterization of sample complexities for (a)synchronous double Q-learning: order-level better dependence on major parameters
- New proof techniques for nested SAs/double Q-learning

Future work:

- Further improve the bounds, possible match the vanilla Q-learning
- Analyze double Q-learning with function approximations

THANK YOU!

Correspondence:

Lin Zhao, Assistant Professor
Dept. of Electrical and Computer Engineering
National University of Singapore
Email: elezhli@nus.edu.sg
Homepage: https://sites.google.com/view/lzhao

References

- Azar, M. G., Munos, R., and Kappen, H. J. (2013). Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. Machine learning, 91(3):325–349.

- Xiong, H., Zhao, L., Liang, Y., and Zhang, W. (2020). Finite-time analysis for double Q-learning. In Advances in Neural Information Processing Systems (NeurIPS).

- Wainwright, M. J. (2019b). Stochastic approximation with cone-contractive operators: Sharp l_∞ bounds for Q-learning. arXiv:1905.06265

- Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. (2020). Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. In Advances in Neural Information Processing Systems (NeurIPS).