Proceedings of Machine Learning Research vol 120:1–17, 2020 2nd Annual Conference on Learning for Dynamics and Control

Periodic Q-Learning

Donghwan Lee DONGHWAN@KAIST.AC.KR
Korea Advanced Institute of Science and Technology

Niao He NIAOHE@ILLINOIS.EDU
University of Illinois at Urbana-Champaign

Editors: A. Bayen, A. Jadbabaie, G. J. Pappas, P. Parrilo, B. Recht, C. Tomlin, M. Zeilinger

Abstract

The use of target networks is a common practice in deep reinforcement learning for stabilizing the training; however, theoretical understanding of this technique is still limited. In this paper, we study the so-called *periodic Q-learning* algorithm (PQ-learning for short), which resembles the technique used in deep Q-learning for solving infinite-horizon discounted Markov decision processes (DMDP) in the tabular setting. PQ-learning maintains two separate Q-value estimates – the online estimate and target estimate. The online estimate follows the standard Q-learning update, while the target estimate is updated *periodically*. In contrast to the standard Q-learning, PQ-learning enjoys a simple finite time analysis and achieves better sample complexity for finding an ε-optimal policy. Our result provides a preliminary justification of the effectiveness of utilizing target estimates or networks in Q-learning algorithms.

Keywords: Reinforcement learning; Q-learning; stochastic optimization; sample complexity

1. Introduction

Reinforcement learning (RL) addresses the optimal control problem for unknown systems through experiences (Sutton and Barto, 1998). Among many others, Q-learning (Watkins and Dayan, 1992) is one of the most popular RL algorithms. Recent deep Q-learning (Mnih et al., 2015) has captured significant attentions in the RL community for outperforming humans in several challenging tasks. Besides the effective use of deep neural networks as function approximators, the success of deep Q-learning is also indispensable to the utilization of target networks when calculating target values at each iteration. Specifically, deep Q-learning maintains two separate networks, the Q-network that approximates the state-action value function, and the target network that is synchronized with the Q-network periodically. In practice, using target networks is proven to substantially improve the performance of Q-learning algorithms (Mnih et al., 2015). However, theoretical understanding of this technique remains rather limited. Lee and He (2019b) recently explores a family of target-based temporal-difference learning algorithms for policy evaluation and develops their convergence analyses. Yang et al. (2019) provides theoretical analysis of the neural fitted Q-iteration under some simplification.

In this paper, we investigate a simple algorithm, called the *periodic Q-learning* (PQ-learning), for finding the optimal policy of infinite-horizon discounted Markov decision process (DMDP) in the tabular setting (i.e., discrete finite state-space). The algorithm mimics deep Q-learning by maintaining two separate Q-value estimates – the online estimate and target estimate: the online estimate takes the standard Q-learning updating rule while freezing the target estimate, and the target esti-

© 2020 D. Lee & N. He.

Initialize $Q_0(s,a) \in [-\frac{1}{1-\gamma}, \frac{1}{1-\gamma}], \forall (s,a) \in \mathcal{S} \times \mathcal{A}$; iteration T; steps $\{N_k\}_{k=0}^{T-1}$; stepsizes $\{\beta_t\}_{t=0}^{\infty}$;
for *iteration* $k = 0, 1, \ldots, T-1$ **do**
  for *step* $t = 0, 1, \ldots, N_k - 1$ **do**
    Obtain sample $(s,a) \sim d_a(s)$, $s' \sim P_a(s,\cdot)$ and $r(s,a)$
    Update $Q_{k,t+1}(s,a) = Q_{k,t}(s,a) - \beta_t(r(s,a) + \gamma \max_{a' \in \mathcal{A}} Q_k(s',a') - Q_{k,t}(s,a))$
  end
  Update $Q_{k+1} = Q_{k,N_k}$
end
Return Q_T

Algorithm 1: Periodic Q-Learning (PQ-learning)

mate is updated through a periodic fashion. Formally speaking, consider the DMDP represented by the tuple $(\mathcal{S}, \mathcal{A}, P, r, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ are the finite state and action spaces, $r(s,a) \in [-1,1]$ is the state-action reward, $P_a(s,s')$ is the probability of transiting from s to s' when taking action a, and $\gamma \in (0,1)$ is a discount factor. The objective is to find a deterministic optimal policy, $\pi^* : \mathcal{S} \to \mathcal{A}$, such that the cumulative discounted rewards over infinite time horizons is maximized, i.e., $\pi^* := \arg\max_\pi \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k r(s_k, \pi(s_k)) \middle| \pi\right]$. For simplicity, throughout, we assume that we have access to a sequence of *i.i.d.* random variables $\{(s_k, a_k)\}_{k=0}^{\infty}$ from a fixed underlying probability distribution, $d_a(s), s \in \mathcal{S}, a \in \mathcal{A}$, of the state and action pair (s,a).[1] The PQ-learning algorithm is formally presented in Algorithm 1.

Notice that when setting $N_k = 1$ for all k, PQ-learning reduces to the standard Q-learning. To the authors' knowledge, finite-time convergence analysis of the standard Q-learning for DMDPs can be challenging, and only few results are reported in the literature. In particular, Even-Dar and Mansour (2003) shows that to achieve an ε-optimal Q-function such that $\|Q_T - Q^*\|_2 \leq \varepsilon$ with probability at least $1-\delta$, the number of samples required by the standard Q-learning algorithm with linear learning rate is $\mathcal{O}\left(\frac{(|\mathcal{S}||\mathcal{A}|)^{\frac{2\ln(1/\varepsilon)}{1-\gamma}}}{(1-\gamma)^2\varepsilon^2} \ln\left(\frac{|\mathcal{S}||\mathcal{A}|}{\delta(1-\gamma)\varepsilon}\right)\right)$. However, for PQ-learning, we show that the finite sample analysis can be easily established based on standard tools of analyzing the stochastic gradient descent and the contraction property of Bellman equation. Our analysis hinges on a key observation that PQ-learning can be viewed as solving a sequence of mean-squared Bellman error minimization subproblems through the stochastic gradient descent routine.

As a main result, we prove that to find an ε-optimal Q function such that $\mathbb{E}[\|Q_T - Q^*\|_\infty] \leq \varepsilon$, the number of samples needed for PQ-learning is at most

$$\mathcal{O}\left(\frac{|\mathcal{S}||\mathcal{A}|}{\varepsilon^2(1-\gamma)^4}\frac{L}{c^3}\ln\left(\frac{1}{(1-\gamma)^2\varepsilon}\right)\right),$$

where c and L corresponds to the minimal and maximal probabilities of the state-action distribution $d_a(s)$. The most efficient sample complexity can be achieved when the state-action distribution is uniform. In this case, the sample complexity becomes

$$\mathcal{O}\left(\frac{|\mathcal{S}|^3|\mathcal{A}|^3}{\varepsilon^2(1-\gamma)^4}\ln\left(\frac{1}{(1-\gamma)^2\varepsilon}\right)\right).$$

1. This assumption can be relaxed to ergodicity conditions. But for simplicity, we focus only on the *i.i.d.* case.

Table 1: Comparison of complexities in terms of $|\mathcal{S}|$, $|\mathcal{A}|$, and ε up to logarithmic factors. The notation "$-$" means that the run-time complexity is the same order as the sample complexity.

Algorithm	Sample Complexity	Run-time Complexity
PQ-learning (this paper)	$\tilde{\mathcal{O}}\left(\frac{\lvert\mathcal{S}\rvert^3\lvert\mathcal{A}\rvert^3}{\varepsilon^2}\right)$	$-$
Q-learning (Even-Dar and Mansour, 2003)	$\tilde{\mathcal{O}}\left(\frac{(\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert)^{\frac{2\ln(1/\varepsilon)}{1-\gamma}}}{\varepsilon^2}\right)$	$-$
Delayed Q-learning (Strehl et al., 2006)	$\tilde{\mathcal{O}}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{\varepsilon^4}\right)$	$-$
Phased Q-learning (Kearns and Singh, 1999)	$\tilde{\mathcal{O}}\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{\varepsilon^2}\right)$	N/A
SPD Q-learning (Lee and He, 2019a)	$\tilde{\mathcal{O}}\left(\frac{\lvert\mathcal{S}\rvert^8\lvert\mathcal{A}\rvert^8}{\varepsilon^2}\right)$	$-$
Lower bound (Azar et al., 2013)	$\Omega\left(\frac{\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert}{\varepsilon^2}\right)$	N/A

Compared to the standard Q-learning, PQ-learning is also fairly simple to implement, with $\mathcal{O}(1)$ per-iteration complexity. On the theory side, the sample complexity (also run-time complexity) of PQ-learning greatly improves that of the standard Q-learning, in terms of the dependence on the sizes of state and action spaces. Our result, in some sense, sheds light on the effectiveness of using target estimates or networks commonly observed in practice. Finally, while the current paper only focuses on the tabular setting, PQ-learning can be extended to incorporate linear function approximation or neural network approximation (i.e., similar to deep Q-learning). We hope this work would open the door to further investigation of the theory of deep Q-learning and the design of more efficient reinforcement learning algorithms with target networks.

Related works. There exists a significant body of Q-learning variations for tabular DMDP in the literature. Some representative model-free algorithms include phased Q-learning (Kearns and Singh, 1999), delayed Q-learning (Strehl et al., 2006), fitted Q-learning (Ernst et al., 2005), double Q-learning (Hasselt, 2010), Zap Q-learning (Devraj and Meyn, 2017), stochastic primal-dual Q-learning (SPD Q-learning) (Lee and He, 2019a), etc. Besides, there are other sampling algorithms for tabular DMDP, such as R-MAX (Strehl et al., 2009), empirical QVI (Azar et al., 2013), sublinear randomized QVI (Sidford et al., 2018), etc. We briefly summarize the computational complexities of some representative works in Table 1, ignoring the dependence on the logarithmic factors and the discount factor.[2] We can see that PQ-learning is reasonably efficient comparing to several existing Q-learning variations. For instance, the phased Q-learning (Kearns and Singh, 1999) has lower sample complexity, but its run-time complexity can be very large. The delayed Q-learning (Strehl et al., 2006) is efficient in that the complexity scales linearly in $|\mathcal{S}|$ and $|\mathcal{A}|$, but the complexity in terms of ε is much worse than PQ-learning. Note that the sample complexity achieved by PQ-learning, is by no means optimal, in terms of the dependence on $|\mathcal{S}|, |\mathcal{A}|$ and the factor $1-\gamma$, comparing to the lower bound established in the literature for DMDP under a generative sampling model (Strehl et al., 2009; Azar et al., 2013). However, we emphasize that the goal of this work is

2. The complexities in Table 1 are obtained after simplifications for convenience of presentations, to provide a rough overview rather than a detailed comparative analysis since many algorithms operate under different assumptions.

to understand the convergence and efficiency of target-based Q-learning, rather than establishing an optimal algorithm that matches the sample complexity lower bound for tabular DMDP.

2. Periodic Q-learning

Recall that the Q-function under policy π is defined as $Q^\pi(s,a) = \mathbb{E}\big[\sum_{k=0}^{\infty}\gamma^k r(s_k,\pi(s_k))\big|\, s_0 = s, a_0 = a, \pi\big]$, $s \in \mathcal{S}, a \in \mathcal{A}$, and the optimal Q-function is defined as $Q^*(s,a) = Q^{\pi^*}(s,a)$ for all $s \in \mathcal{S}, a \in \mathcal{A}$. Consider the Bellman operator $\mathbf{T} : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$,

$$(\mathbf{T}Q)(s,a) := \sum_{s'\in\mathcal{S}} P_a(s,s')\left(r(s,a) + \gamma \max_{a'\in A} Q(s',a')\right).$$

The Bellman operator is known to be a contraction with respect to the max-norm (Puterman, 2014), and the optimal Q-function, Q^*, is the unique fixed point of this operator. Once Q^* is known, then an optimal policy can be retrieved by $\pi^*(s) = \arg\max_{a\in\mathcal{A}} Q^*(s,a)$. Therefore, the MDP can be solved by finding the optimal Q-function.

For the PQ-learning, at each iteration k, the algorithm can be viewed as approximately computing the Bellman operator $\mathbf{T}Q_k$ through minimizing the mean-squared loss function

$$\min_{Q\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}} l(Q;Q_k) := \frac{1}{2}\mathbb{E}_{s,a}\left[\left(\mathbb{E}_{s'}\left[r(s,a) + \gamma\max_{a'\in\mathcal{A}} Q_k(s',a')\right] - Q(s,a)\right)^2\right]. \tag{1}$$

Here $\mathbb{E}_{s,a}$ is the expectation taken with respect to the current state-action pair which has the distribution $d_a(s)$, i.e., $\mathbb{P}[s_k = s, a_k = a] = d_a(s)$ for any time step $k \geq 0$, and $\mathbb{E}_{s'}$ is the expectation taken with respect to the next state $s' \sim P_a(s,\cdot)$. In particular, PQ-learning approximately solves the subproblem (1) through N_k steps of stochastic gradient descent:

$$Q_{k,t+1} = Q_{k,t} - \beta_t\, \tilde{\nabla}_Q l(Q;Q_k)\Big|_{Q=Q_{k,t}},\ t = 0,\ldots,N_k - 1$$

where $\tilde{\nabla}_Q l(Q;Q_k)$ is a stochastic estimator of the gradient $\nabla_Q l(Q;Q_k)$. Here, Q_k can be treated as the target estimate of the optimal Q-values, and $Q_{k,t}$ is the online estimate. The target estimate is synchronized with the online estimate after N_k step, i.e., $Q_{k+1} = Q_{k,N_k}$. Here, for generality, we allow the inner steps N_k to vary at every iteration. PQ-learning resembles the original deep Q-learning (Mnih et al., 2015) when N_k is set to a constant. If $N_k = 0$ for all $k = 0,1,\ldots,T-1$, then PQ-learning corresponds to the standard Q-learning.

As we only apply a finite number N_k steps of SGD to solve (1), the SGD subroutine will return an approximate solution with a certain error bound. Throughout, we denote $\{\varepsilon_k\}_{k=1}^T$ as the approximation errors such that

$$\mathbb{E}[\|Q_{k+1} - \mathbf{T}Q_k\|_2^2] \leq \varepsilon_{k+1}.$$

As a result, the convergence of PQ-learning depends on the cumulative approximation errors $\{\varepsilon_k\}_{k=1}^T$ induced from SGD subroutine, whereas these error terms are determined by the number of inner steps, the learning rate, and the variance of stochastic gradient. In what follows, we derive the convergences of the outer and inner iterations, and further establish the overall sample complexity for achieving ε-optimal policy.

Remark. It is worth pointing out that PQ-learning shares some similarity with fitted Q-iteration (FQI) (Ernst et al., 2005) algorithms in that both approaches can be viewed as approximating the Bellman operator at every iteration, but there exist notable differences. FQI usually solves the resulting linear regression problem through matrix inversion computation or batch algorithms, requiring expensive memory and computation cost. On the other hand, PQ-learning solves the subproblems through the online SGD subroutine, which is much more efficient. By using the SGD steps, PQ-learning can be naturally interpreted as Q-learning with periodic target update (Algorithm 1), thus much simpler than the FQI algorithms. As will be shown in the next section, PQ-learning enjoys a much simpler finite-time convergence and complexity analysis, in stark contrast to many other online Q-learning algorithms.

3. Main Results

Throughout, we assume that the sampling distribution satisfies that $d_a(s) > 0$ for all $s \in \mathcal{S}, a \in \mathcal{A}$. Throughout the paper, we define the constants $c := \min_{s\in\mathcal{S},a\in\mathcal{A}} d_a(s)$ and $L := \max_{s\in\mathcal{S},a\in\mathcal{A}} d_a(s)$, the nonsingular diagonal matrix $D := \text{diag}[D_1; \ldots, D_{|\mathcal{A}|}] \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$, where each $D_a, a \in \{1, 2, \cdots, |\mathcal{A}|\}$ is a diagonal matrix whose diagonal entries consist of the distribution $d_a(s), s \in \mathcal{S}$. We start by characterizing the outer and inner iteration convergence, respectively.

Proposition 1 (Outer iteration convergence) *We have*

$$\mathbb{E}[\|Q_T - Q^*\|_\infty] \leq \sum_{k=1}^{T} \gamma^{T-k}\sqrt{\varepsilon_k} + \gamma^T \mathbb{E}[\|Q_0 - Q^*\|_\infty].$$

Particularly, if $\varepsilon_k = \varepsilon$ for all $k \geq 0$, then $\mathbb{E}[\|Q_T - Q^*\|_\infty] \leq \frac{\sqrt{\varepsilon}}{1-\gamma} + \gamma^T \mathbb{E}[\|Q_0 - Q^*\|_\infty]$.

One can see that the error is essentially decomposed into two terms, one from the approximation errors induced from SGD procedures and one from the contraction property of solving the subproblems, which can also be viewed as approximately computing the Bellman operators. To further analyze the approximation error from the SGD procedure, existing convergence results for SGD can be applied with some modifications.

Proposition 2 (Inner iteration convergence) *Suppose that $\mathbb{E}[\|Q_i - \mathbf{T}Q_{i-1}\|_2^2] \leq \varepsilon, i \in \{1, 2, \ldots, k\}$ and $\varepsilon \leq (1-\gamma)^2$ hold. Suppose the PQ-learning algorithm is run with a step-size rule $\beta_t = \beta/(\lambda + t)$ with $\beta = 2/c$ and $\lambda = 16L/(c^2)$. Then,*

$$\mathbb{E}[\|\mathbf{T}Q_k - Q_{k,t}\|_D^2] \leq \frac{512|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}\frac{L}{c^3} \cdot \frac{1}{\lambda + t}, \quad \forall t \geq 0.$$

Proposition 2 ensures that the inner iterate, $Q_{k,t}$, converges to the solution of the subproblem at the rate of $\mathcal{O}(1/t)$. Combining Proposition 2 with Proposition 1, the overall sample complexity can be easily derived.

Theorem 3 (Sample Complexity I) *Let $\beta_t = \beta/(\lambda + t)$ with $\beta = 2/c$ and $\lambda = 16L/(c^2)$ and the maximum number of steps*

$$N_k = N \geq \frac{2048|\mathcal{S}||\mathcal{A}|}{\varepsilon^2(1-\gamma)^4}\frac{L}{c^3}.$$

PQ-learning achieves an ε-optimal solution, $\mathbb{E}[\|Q_T - Q^*\|_\infty] \leq \varepsilon$, with the number of samples at most

$$\frac{2048|\mathcal{S}||\mathcal{A}|}{\varepsilon^2(1-\gamma)^4 \ln \gamma^{-1}} \frac{L}{c^3} \ln\left(\frac{4}{(1-\gamma)\varepsilon}\right).$$

Note that in Theorem 3, the constant c and L depend on $|\mathcal{S}|$ and $|\mathcal{A}|$. In particular, c is upper bounded by $1/(|\mathcal{S}||\mathcal{A}|)$, while L is lower bounded by $1/(|\mathcal{S}||\mathcal{A}|)$. The upper and lower bounds are achieved when the state-action distribution is uniform. In other words, the best sample complexity bound in Theorem 3 is obtained when the state-action distribution is uniform because in this case, $c = L = 1/(|\mathcal{S}||\mathcal{A}|)$ according to the definitions of c and L, and the quantity, $\frac{L}{c^3}$, is minimized. Accordingly, the sample complexity becomes $\tilde{\mathcal{O}}\left(\frac{|\mathcal{S}|^3|\mathcal{A}|^3}{\varepsilon^2(1-\gamma)^4}\right)$. Since the per-iteration (inner iteration) complexity is $\mathcal{O}(1)$, the run-time complexity of PQ-learning has the same order as the sample complexity.

Although Theorem 3 provides a finite-time convergence analysis of PQ-learning in terms of the Q-function, it does not reflect convergence of the corresponding policy recovered from Q_T, namely, $\pi_{Q_T}(s) := \arg\max_{a\in\mathcal{A}} Q_T(s,a)$. In the sequel, we focus on the convergence of the policy π_{Q_T}. A policy, π_{Q_T}, is called the ε-optimal policy if $\mathbb{E}[\|V^{\pi_{Q_T}} - V^*\|_\infty] \leq \varepsilon$ holds. Before proceeding, we present a simple lemma that characterizes the relationship between the $\mathbb{E}[\|Q_T - Q^*\|_\infty]$ with $\mathbb{E}[\|V^{\pi_{Q_T}} - V^*\|_\infty]$.

Lemma 4 *If $\mathbb{E}[\|Q_T - Q^*\|_\infty] \leq \varepsilon$, then $\mathbb{E}[\|V^{\pi_{Q_T}} - V^*\|_\infty] \leq \frac{2\varepsilon}{1-\gamma}$.*

Invoking Lemma 4 and Theorem 3, we immediately arrive at the following result.

Theorem 5 (Sample Complexity II) *Under the same setting as above, PQ-learning returns an ε-optimal policy π_{Q_T} such that $\mathbb{E}[\|V^{\pi_{Q_T}} - V^*\|_\infty] \leq \varepsilon$, with $\varepsilon \leq 1$, with the number of samples at most*

$$\frac{8192|\mathcal{S}||\mathcal{A}|}{\varepsilon^2(1-\gamma)^6 \ln \gamma^{-1}} \frac{L}{c^3} \ln\left(\frac{8}{(1-\gamma)^2\varepsilon}\right),$$

where c and L are constants as defined before.

Similarly, when the state-action distribution is uniform, the sample complexity bound for PQ-learning to obtain an ε-optimal policy becomes $\tilde{\mathcal{O}}\left(\frac{|\mathcal{S}|^3|\mathcal{A}|^3}{\varepsilon^2(1-\gamma)^6}\right)$. It is worth pointing out that the sample complexity lower bound for solving DMDP under a generative sampling model established in the literature is $\Omega\left(\frac{|\mathcal{S}||\mathcal{A}|}{\varepsilon^2(1-\gamma)^3}\right)$ (Azar et al., 2013). Hence, PQ-learning is *optimal* in terms of dependence on ϵ, but not in the dependence on $|\mathcal{S}|, |\mathcal{A}|$ and the factor $1-\gamma$. Nonetheless, the sample complexity greatly improves over the existing results on standard Q-learning as reported in Table 1, e.g., $\tilde{\mathcal{O}}((|\mathcal{S}||\mathcal{A}|)^{\frac{2\ln(1/\varepsilon)}{1-\gamma}}/(\varepsilon^2))$ in (Even-Dar and Mansour, 2003).

4. Technical Proofs

In this section, we provide detailed analysis of the main results described in the previous section. The characterizations of inner and outer iteration convergences, as stated in Propositions 1 and 2 form the backbone of the main results. Below we present the proofs for these two propositions. As can be seen, the analysis of PQ-learning is fairly simple and elegant based on the contraction property and standard arguments of SGD convergence.

4.1. Proof of Proposition 1

We have

$$\begin{aligned}\mathbb{E}[\|Q_{k+1}-Q^*\|_\infty] \leq& \mathbb{E}[\|Q_{k+1}-\mathbf{T}Q_k\|_\infty] + \mathbb{E}[\|(\mathbf{T}Q_k - Q^*\|_\infty] \\ \leq& \sqrt{\mathbb{E}[\|Q_{k+1}-\mathbf{T}Q_k\|_2^2]} + \mathbb{E}[\|\mathbf{T}Q_k - \mathbf{T}Q^*\|_\infty] \\ \leq& \sqrt{\varepsilon_{k+1}} + \gamma\mathbb{E}[\|Q_k - Q^*\|_\infty],\end{aligned}$$

where the last inequality comes from the contraction property of the Bellman operator $\mathbf{T}$. Therefore, we have

$$\mathbb{E}[\|Q_{k+1}-Q^*\|_\infty] \leq \sqrt{\varepsilon_{k+1}} + \gamma\mathbb{E}[\|Q_k - Q^*\|_\infty].$$

Combining the last inequality over $k = 0, 1, \ldots, T-1$, the desired result is obtained. ■

4.2. Proof of Proposition 2

Proposition 2 describes the convergence of the SGD subroutine when applied to solving the sub-problem, $\min_{Q\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}} l(Q;Q_k)$. Our proof follows the standard analysis in Bottou et al. (2018, Theorem 4.7), adapted to the specific problem. Notice that the objective $l(Q;Q_k)$ is c-strongly convex and L-Lipschitz smooth.

Lemma 6 (Strong convexity and Lipschitz continuity) *The objective function $l(Q;Q_k)$ is c-strongly convex with $c = \min_{s\in\mathcal{S},a\in\mathcal{A}} d_a(s)$ and is Lipschitz continuous with parameter $L = \max_{s\in\mathcal{S},a\in\mathcal{A}} d_a(s)$.*

Moreover, we introduce two key lemmas showing that the variance of the stochastic gradient and target estimates can be properly bounded. Proofs are deferred in Appendix.

Lemma 7 (Boundedness of variance) *We have for any $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$,*

$$\mathbb{E}[\|\tilde{\nabla}_Q l(Q;Q_k)\|_2^2|Q_k, Q] \leq 12\gamma^2|\mathcal{S}||\mathcal{A}|\|Q^*-Q_k\|_\infty^2 + 8\|\nabla_Q l(Q;Q_k)\|_{D^{-1}}^2 + \frac{18|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}.$$

Lemma 8 (Boundedness of estimate) *Suppose $\mathbb{E}[\|Q_i - \mathbf{T}Q_{i-1}\|_2^2] \leq \varepsilon, \forall i \leq k$ and $\epsilon \leq (1-\gamma)^2$. Then*

$$\mathbb{E}[\|Q_k - Q^*\|_\infty^2] \leq \frac{8}{(1-\gamma)^2}.$$

Invoking the smoothness and strong convexity conditions, one can easily show that

$$\begin{aligned}&\|\mathbf{T}Q_k - Q_{k,t+1}\|_D^2 - \|\mathbf{T}Q_k - Q_{k,t}\|_D^2 \\ \leq& -2\beta_t\|D\mathbf{T}Q_k - DQ_{k,t}\|_2^2 + L\beta_t^2\mathbb{E}[\|\tilde{\nabla}_Q l(Q_{k,t};Q_k)\|_2^2|Q_{k,t},Q_k] \\ \leq& (8L\beta_t^2 - 2c\beta_t)\|\mathbf{T}Q_k - Q_{k,t}\|_D^2 + 12|\mathcal{S}||\mathcal{A}|\gamma^2 L\beta_t^2\|Q^*-Q_k\|_\infty^2 + \frac{18|\mathcal{S}||\mathcal{A}|L\beta_t^2}{(1-\gamma)^2}.\end{aligned}$$

Recall that the step-size satisfies $0 \leq \beta_t \leq \frac{c}{8L}$, which implies $8L\beta_t^2 - 2\beta_t c \leq -c\beta_t$. Thus,

$$\mathbb{E}[\|\mathbf{T}Q_k - Q_{k,t+1}\|_D^2] \leq (1-c\beta_t)\mathbb{E}[\|\mathbf{T}Q_k - Q_{k,t}\|_D^2] + \beta_t^2\left(\frac{114|\mathcal{S}||\mathcal{A}|L}{(1-\gamma)^2}\right).$$

By induction, it follows from the above recursion and $\beta_t = \frac{\beta}{\lambda+t}$ that

$$\mathbb{E}[\|\mathbf{T}Q_k - Q_{k,t}\|_D^2] \leq \frac{v}{\lambda+t}, \quad \forall t \geq 0, \forall k \geq 1$$

where $v = \max\{\lambda\mathbb{E}[\|\mathbf{T}Q_k - Q_{k,0}\|_D^2], \beta^2 C\}$, and $C = \frac{114|\mathcal{S}||\mathcal{A}|L}{(1-\gamma)^2}$. Note that the term $\mathbb{E}[\|\mathbf{T}Q_k - Q_{k,0}\|_D^2]$ can be further bounded as

$$\begin{aligned}
\mathbb{E}[\|\mathbf{T}Q_k - Q_{k,0}\|_D^2] \leq & \max_{s,a} d_a(s)\mathbb{E}[\|\mathbf{T}Q_k - Q_{k,0}\|_2^2] \\
\leq & 2L|\mathcal{S}||\mathcal{A}|\mathbb{E}[\|\mathbf{T}Q_k - \mathbf{T}Q^*\|_\infty^2] + 2L|\mathcal{S}||\mathcal{A}|\mathbb{E}[\|\mathbf{T}Q^* - Q_k\|_\infty^2] \\
\leq & 2L|\mathcal{S}||\mathcal{A}|\gamma^2\mathbb{E}[\|Q_k - Q^*\|_\infty^2] + 2L|\mathcal{S}||\mathcal{A}|\mathbb{E}[\|Q^* - Q_k\|_\infty^2] \\
\leq & \frac{32L|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2},
\end{aligned}$$

Combing these facts then leads to the desired result in Proposition 2. ■

5. Discussions

In this paper, we introduce PQ-learning and provide its finite-time convergence and complexity analysis. The algorithm is relevantly simple, intuitive, and efficient. While this paper only focuses on the tabular case, the algorithm and analysis might be extended to PQ-learning with function approximations. When linear function approximation is used, a potential challenge arises from the mismatch between the ∞-norm used for the contraction property of the Bellman equation and the 2-norm used for the projection of the Q-function onto the ranges of the feature vectors. Therefore, the composition of the Bellman operator and the projection operator is no longer a contraction. However, under certain conditions on the feature matrix, the composite mapping can be proven to be a contraction and our analysis would still apply. We leave this for future investigation.

Another important issue is the *i.i.d.* assumption of samples. This assumption is rather restrictive as in practice the samples are often acquired from past trajectories or experiences. Our current result can be easily extended to Markovian sampling by applying standard mixing time arguments in stochastic optimization; see e.g., Duchi et al. (2012); Sun et al. (2018); Dalal et al. (2018). Note that our analysis of PQ-learning is mainly based on characterizing the cumulative errors from the SGD subroutines. Combining the existing result of SGD under Markovian samples with our analysis immediately leads to the finite-time convergence of PQ-learning under Markovian samples, which we leave for future investigation. Lastly, building on the stochastic optimization framework, various methods such as the variance reduction techniques and acceleration schemes (Kingma and Ba, 2015; Allen-Zhu, 2017; Johnson and Zhang, 2013; Defazio et al., 2014a,b) can be applied to replace the SGD subroutine used in the algorithm, to further improve the convergence or sample efficiency.

References

Zeyuan Allen-Zhu. Katyusha: the first direct acceleration of stochastic gradient methods. *The Journal of Machine Learning Research*, 18(1):8194–8244, 2017.

Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J Kappen. Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine Learning*, 91(3): 325–349, 2013.

Léon Bottou, Frank E Curtis, and Jorge Nocedal. Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311, 2018.

Gal Dalal, Balázs Szörényi, Gugan Thoppe, and Shie Mannor. Finite sample analyses for TD(0) with function approximation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

Aaron Defazio, Francis Bach, and Simon Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in Neural Information Processing Systems*, pages 1646–1654, 2014a.

Aaron Defazio, Justin Domke, et al. Finito: A faster, permutable incremental gradient method for big data problems. In *International Conference on Machine Learning*, pages 1125–1133, 2014b.

Adithya M Devraj and Sean Meyn. Zap Q-learning. In *Advances in Neural Information Processing Systems*, pages 2235–2244, 2017.

John C Duchi, Alekh Agarwal, Mikael Johansson, and Michael I Jordan. Ergodic mirror descent. *SIAM Journal on Optimization*, 22(4):1549–1578, 2012.

Damien Ernst, Pierre Geurts, and Louis Wehenkel. Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6(Apr):503–556, 2005.

Eyal Even-Dar and Yishay Mansour. Learning rates for Q-learning. *Journal of Machine Learning Research*, 5(Dec):1–25, 2003.

Hado V Hasselt. Double Q-learning. In *Advances in Neural Information Processing Systems*, pages 2613–2621, 2010.

Rie Johnson and Tong Zhang. Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in Neural Information Processing Systems*, pages 315–323, 2013.

Michael J Kearns and Satinder P Singh. Finite-sample convergence rates for q-learning and indirect algorithms. In *Advances in Neural Information Processing Systems*, pages 996–1002, 1999.

Diederik P Kingma and Jimmy Ba. Adam: a method for stochastic optimization. *Proceedings of the 3rd International Conference on Learning Representations*, 2015.

Donghwan Lee and Niao He. Stochastic primal-dual Q-learning algorithm for discounted MDPs. In *2019 American Control Conference (ACC)*, pages 4897–4902. IEEE, 2019a.

Donghwan Lee and Niao He. Target-based temporal-difference learning. In *International Conference on Machine Learning*, pages 3713–3722, 2019b.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

Martin L. Puterman. *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

Aaron Sidford, Mengdi Wang, Xian Wu, Lin Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving markov decision processes with a generative model. In *Advances in Neural Information Processing Systems*, pages 5186–5196, 2018.

Alexander L Strehl, Lihong Li, Eric Wiewiora, John Langford, and Michael L Littman. PAC model-free reinforcement learning. In *Proceedings of the 23rd International Conference on Machine learning*, pages 881–888, 2006.

Alexander L Strehl, Lihong Li, and Michael L Littman. Reinforcement learning in finite MDPs: PAC analysis. *Journal of Machine Learning Research*, 10:2413–2444, 2009.

Tao Sun, Yuejiao Sun, and Wotao Yin. On markov chain gradient descent. In *Advances in Neural Information Processing Systems*, pages 9896–9905, 2018.

Richard S. Sutton and Andrew G. Barto. *Reinforcement learning: An introduction*. MIT Press, 1998.

Christopher J. C. H. Watkins and Peter Dayan. Q-learning. *Machine Learning*, 8(3-4):279–292, 1992.

Zhuora Yang, Yuchen Xie, and Zhaoran Wang. A theoretical analysis of deep Q-learning. *arXiv preprint arXiv:1901.00137*, 2019.

Appendix

1. Matrix representation

In this section, we introduce vector and matrix notations for the matrix representation of the Bellman operator. To this end, we introduce the following compact notations:

$$P := \begin{bmatrix} P_1 \\ \vdots \\ P_{|\mathcal{A}|} \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}||\mathcal{A}|}, \quad R := \begin{bmatrix} R_1 \\ \vdots \\ R_{|\mathcal{A}|} \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}, \quad Q := \begin{bmatrix} Q_1 \\ \vdots \\ Q_{|\mathcal{A}|} \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|},$$

$$D_a := \begin{bmatrix} d_a(1) & & \\ & \ddots & \\ & & d_a(|\mathcal{S}|) \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}, \quad D := \begin{bmatrix} D_1 & & \\ & \ddots & \\ & & D_{|\mathcal{A}|} \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|},$$

where $Q_a = Q(\cdot, a) \in \mathbb{R}^{|\mathcal{S}|}, a \in \mathcal{A}$ and $R_a(s) := \mathbb{E}[r(s,a)|s,a]$. Note that D is a nonsingular diagonal matrix with strictly positive diagonal elements. Moreover, for any deterministic policy, $\pi : \mathcal{S} \to \mathcal{A}$, we define the corresponding distribution vector

$$\boldsymbol{\pi}(s) := e_{\pi(s)} \in \Delta_{|\mathcal{S}|},$$

where $\Delta_{|\mathcal{S}|}$ is the set of all discrete probability distributions over $\mathcal{S}$, and define the matrix

$$\Pi_\pi := \begin{bmatrix} \boldsymbol{\pi}(1)^T \otimes e_1^T \\ \boldsymbol{\pi}(2)^T \otimes e_2^T \\ \vdots \\ \boldsymbol{\pi}(|\mathcal{S}|)^T \otimes e_{|\mathcal{S}|}^T \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}||\mathcal{A}|},$$

where $e_j, j \in \{1, 2, \ldots, |\mathcal{A}|\}$ is used to denote the j-th basis vector of dimension $|\mathcal{A}|$. Note that Π_π is a matrix function which depends on the policy $\pi \in \Theta$. With these notations, the Bellman operator can be compactly written by

$$\mathbf{T}Q = R + \gamma P \Pi_{\pi_Q} Q,$$

where $\pi_Q(s) := \arg\max_{a\in\mathcal{A}} e_s^T Q_a \in \mathcal{A}$. We note that, for any $\pi \in \Theta$, $P\Pi_\pi$ is the state-action pair transition probability matrix under the deterministic policy π, and Π_{π_Q} is a nonlinear matrix function which depends on Q. This matrix representation of the Bellman operator plays an important role in the subsequent developments.

2. Proofs of Technical Lemmas

Throughout the remaining part, we denote by $F(Q) := l(Q; Q_k) := \frac{1}{2}\|R + P\Pi_{\pi_{Q_k}} Q_k - Q\|_D^2$. Given a sample (s, a, s'), define the corresponding stochastic Bellman operator

$$\hat{\mathbf{T}}Q_k := \gamma \sum_{j\in\mathcal{S}, i\in\mathcal{A}} (e_i \otimes e_j)(e_{s'})^T \Pi_{\pi_{Q_k}} Q_k + R$$

and the matrix $\hat{D} := (e_a \otimes e_s)(e_a \otimes e_s)^T$. Then, the stochastic gradient is written by

$$g(Q; Q_k) = -(\hat{D}\hat{\mathbf{T}}Q_k - \hat{D}Q) = -((e_a \otimes e_s)((e_{s'})^T \gamma \Pi_{\pi_{Q_k}} Q_k + (e_a \otimes e_s)^T R - (e_a \otimes e_s)^T Q)).$$

2.1. Proof of Lemma 4

We first show that if $\mathbb{E}[\|Q_T - Q^*\|_\infty] \leq \varepsilon$, then

$$\mathbb{E}[\|Q^{\pi_{Q_T}} - Q^*\|_\infty] \leq \frac{2\gamma\varepsilon}{1-\gamma}.$$

This is because

$$\begin{aligned}
\mathbb{E}[\|Q^{\pi_{Q_T}} - Q^*\|_\infty] =& \mathbb{E}[\|Q^{\pi_{Q_T}} - \mathbf{T}Q_T + \mathbf{T}Q_T - Q^*\|_\infty] \\
\leq& \mathbb{E}[\|Q^{\pi_T} - \mathbf{T}Q_T\|_\infty] + \mathbb{E}[\|\mathbf{T}Q_T - Q^*\|_\infty] \\
=& \mathbb{E}[\|R + \gamma P\Pi_{\pi_{Q_T}} Q^{\pi_{Q_T}} - (R + \gamma P\Pi_{\pi_{Q_T}} Q_T)\|_\infty] + \gamma\varepsilon \\
\leq& \mathbb{E}[\gamma\|P\Pi_{\pi_{Q_T}}\|_\infty \|Q^{\pi_{Q_T}} - Q_T\|_\infty] + \gamma\varepsilon \\
\leq& \gamma\mathbb{E}[\|Q^{\pi_{Q_T}} - Q_T\|_\infty] + \gamma\varepsilon \\
\leq& \gamma\mathbb{E}[\|Q^{\pi_{Q_T}} - Q^*\|_\infty] + \gamma\mathbb{E}[\|Q^* - Q_T\|_\infty] + \gamma\varepsilon \\
\leq& \gamma\mathbb{E}[\|Q^{\pi_{Q_T}} - Q^*\|_\infty] + 2\gamma\varepsilon,
\end{aligned}$$

where the second line is due to the triangle inequality, the second line uses the optimal Bellman equation, $Q^{\pi_{Q_T}} = R + \gamma P\Pi_{\pi_{Q_T}} Q^{\pi_{Q_T}}$ and $Q^* = \mathbf{T}Q^*$ and the contraction of the Bellman operator, the forth line is due to the hypothesis $\mathbb{E}[\|Q_T - Q^*\|_\infty] \leq \varepsilon$, the fifth line is due to the Cauchy-Schwarz inequality, the sixth line is due to the fact that $P\Pi_{\pi_{Q_T}}$ is a stochastic matrix whose row sum is one, the seventh line is due to the triangle inequality, and the last line uses the hypothesis $\mathbb{E}[\|Q_T - Q^*\|_\infty] \leq \varepsilon$. The desired result is obtained by rearranging the last inequality. Therefore, we have

$$\begin{aligned}
\|V^{\pi_{Q_T}} - V^*\|_\infty =& \max_{s\in\mathcal{S}} |V^{\pi_{Q_T}}(s) - V^*(s)| \\
=& \max_{s\in\mathcal{S}} |Q^{\pi_{Q_T}}(s, \pi_{Q_T}(s)) - \max_{a\in\mathcal{A}} Q^*(s,a)| \\
\leq& \max_{s\in\mathcal{S}} |Q^{\pi_{Q_T}}(s, \pi_T(s)) - \max_{a\in\mathcal{A}} Q_T(s,a)| + \max_{s\in\mathcal{S}} |\max_{a\in\mathcal{A}} Q_T(s,a) - \max_{a\in\mathcal{A}} Q^*(s,a)| \\
=& \max_{s\in\mathcal{S}} |Q^{\pi_{Q_T}}(s, \pi_{Q_T}(s)) - \max_{a\in\mathcal{A}} Q_T(s,a)| + \|Q_T - Q^*\|_\infty. \qquad (2)
\end{aligned}$$

For any fixed s, we have $|Q^{\pi_{Q_T}}(s, \pi_{Q_T}(s)) - \max_{a\in\mathcal{A}} Q_T(s,a)| \leq \max_{a\in\mathcal{A}} |Q_T(s,a) - Q^{\pi_{Q_T}}(s,a)|$. By taking the max over s on the right-hand side first and then on the left-hand side, we obtain $\max_{s\in\mathcal{S}} |Q^{\pi_{Q_T}}(s, \pi_{Q_T}(s)) - \max_{a\in\mathcal{A}} Q_T(s,a)| \leq \|Q_T - Q^{\pi_{Q_T}}\|_\infty$. Using this inequality and taking the expectation on both sides of the inequality (2), it is further bounded as

$$\begin{aligned}
\mathbb{E}[\|V^{\pi_{Q_T}} - V^*\|_\infty] \leq& \mathbb{E}[\|Q_T - Q^{\pi_{Q_T}}\|_\infty] + \mathbb{E}[\|Q_T - Q^*\|_\infty] \\
\leq& \mathbb{E}[\|Q_T - Q^{\pi_{Q_T}}\|_\infty] + \varepsilon \\
\leq& \mathbb{E}[\|Q_T - Q^*\|_\infty] + \mathbb{E}[\|Q^* - Q^{\pi_{Q_T}}\|_\infty] + \varepsilon \\
\leq& \mathbb{E}[\|Q^* - Q^{\pi_{Q_T}}\|_\infty] + 2\varepsilon \\
\leq& \frac{2\gamma\varepsilon}{1-\gamma} + 2\varepsilon \\
\leq& \frac{2\varepsilon}{1-\gamma}.
\end{aligned}$$

This completes the proof. ■

2.2. Proof of Lemma 6

Noting that

$$F(Q) = \frac{1}{2}(R + P\Pi_{\pi_{Q_k}} Q_k)^T D(R + P\Pi_{\pi_{Q_k}} Q_k) + \frac{1}{2}Q^T DQ - (R + P\Pi_{\pi_{Q_k}} Q_k)^T DQ,$$

which is a convex quadratic function. Since $F(Q)$ is twice differentiable, $F(Q)$ is c-strongly convex if and only if there exists a constant $c > 0$ such that $\nabla_Q^2 F(Q) \geq cI$. Moreover, since the Hessian is $\nabla_Q^2 F(Q) = D$, and $D \geq \min_{s\in\mathcal{S},a\in\mathcal{A}} d_a(s)I$, $F(Q)$ is c-strongly convex with $c = \min_{s\in\mathcal{S},a\in\mathcal{A}} d_a(s)$.

Moreover, noting that $\nabla F(Q) = -D(R + P\Pi_{\pi_{Q_k}} Q_k - Q)$, we have

$$\begin{aligned}\|\nabla F(Q) - \nabla F(Q')\|_2 =& \|D(R + P\Pi_{\pi_{Q_k}} Q_k - Q) - D(R + P\Pi_{\pi_{Q_k}} Q_k - Q')\|_2 \\ =& \|D(Q - Q')\|_2 \\ \leq& \|D\|_2 \|Q - Q'\|_2,\end{aligned}$$

where $\|D\|_2 = \sqrt{\lambda_{\max}(DD)} = \max_{s\in\mathcal{S},a\in\mathcal{A}} d_a(s)$, which proves the desired result. ■

2.3. Proof of Lemma 7

We want to show that

$$\mathbb{E}[\|g(Q; Q_k)\|_2^2 | Q_k, Q] \leq 12\gamma^2 |\mathcal{S}||\mathcal{A}| \|Q^* - Q_k\|_\infty^2 + 8\|\nabla F(Q)\|_{D^{-1}}^2 + \frac{18|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}.$$

We first have

$$\begin{aligned}\mathbb{E}[\|g(Q; Q_k)\|_2^2 | Q_k, Q] =& \mathbb{E}[\|\hat{D}\hat{\mathbf{T}}Q_k - \hat{D}Q | Q_k, Q] \\ =& \mathbb{E}[\|\hat{D}\hat{\mathbf{T}}Q_k - \hat{D}Q - (\hat{D}\hat{\mathbf{T}}Q^* - \hat{D}Q^*) + \hat{D}\hat{\mathbf{T}}Q^* - \hat{D}Q^*\|_2^2 | Q_k, Q] \\ \leq& 2\mathbb{E}[\|(e_a \otimes e_s)((e_{s'})^T \gamma \Pi_{\pi_{Q_k}} Q_k - (e_{s'})^T \gamma \Pi_{\pi_{Q^*}} Q^*) \\ & - (e_a \otimes e_s)((e_a \otimes e_s)^T Q - (e_a \otimes e_s)^T Q^*)\|_2^2 | Q_k] + 2\sigma^2,\end{aligned}$$

where $\sigma^2 := \mathbb{E}[\|\hat{D}\hat{\mathbf{T}}Q^* - \hat{D}Q^*\|_2^2]$ and the last inequality is due to $\|a + b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$. Now, we further have

$$\begin{aligned}& \mathbb{E}[\|g(Q; Q_k)\|_2^2 | Q_k, Q] \\ \leq& 2\mathbb{E}[\|e_a \otimes e_s\|_2^2]\mathbb{E}[\|(e_{s'})^T(\gamma\Pi_{\pi_{Q_k}} Q_k - \gamma\Pi_{\pi_{Q^*}} Q^*) - (e_a \otimes e_s)^T(Q - Q^*)\|_2^2 | Q_k, Q] + 2\sigma^2 \\ \leq& 2\mathbb{E}[\|(e_{s'})^T(\gamma\Pi_{\pi_{Q_k}} Q_k - \gamma\Pi_{\pi_{Q^*}} Q^*) - (e_a \otimes e_s)^T(Q - Q^*)\|_2^2 | Q_k, Q] + 2\sigma^2 \\ \leq& 4\gamma^2\mathbb{E}[\|(e_{s'})^T(\Pi_{\pi_{Q_k}} Q_k - \Pi_{\pi_{Q^*}} Q^*)\|_2^2 | Q_k] + 4\mathbb{E}[\|(e_a \otimes e_s)^T(Q - Q^*)\|_2^2 | Q] + 2\sigma^2 \\ =& 4\gamma^2\mathbb{E}[(\Pi_{\pi_{Q_k}} Q_k - \Pi_{\pi_{Q^*}} Q^*)^T e_{s'}(e_{s'})^T(\Pi_{\pi_{Q_k}} Q_k - \Pi_{\pi_{Q^*}} Q^*) | Q_k, Q] \\ & + 4\mathbb{E}[(Q - Q^*)^T(e_a \otimes e_s)(e_a \otimes e_s)^T(Q - Q^*) | Q] + 2\sigma^2 \\ =& 4\gamma^2\|\Pi_{\pi_{Q_k}} Q_k - \Pi_{\pi_{Q^*}} Q^*\|_{\mathbb{E}[e_{s'}(e_{s'})^T]}^2 + 4\|Q - Q^*\|_D^2 + 2\sigma^2 \\ \leq& 4\gamma^2|\mathcal{S}|\|\Pi_{\pi_{Q_k}} Q_k - \Pi_{\pi_{Q^*}} Q^*\|_\infty^2 + 4\|Q - Q^*\|_D^2 + 2\sigma^2,\end{aligned}$$

where the first inequality follows from the CauchySchwarz inequality and Holder's inequality, the second inequality is due to $\mathbb{E}[\|e_a \otimes e_s\|_2^2] \leq 1$, and the third inequality is due to $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$. Invoking the following inequality

$$\|\Pi_{\pi_{Q_k}} Q_k - \Pi_{\pi_{Q^*}} Q^*\|_\infty^2 \leq \|Q_k - Q^*\|_\infty^2,$$

we further bound $\mathbb{E}[\|g(Q;Q_k)\|_2^2 | Q_k, Q]$ as

$$\mathbb{E}[\|g(Q;Q_k)\|_2^2 | Q_k, Q] \leq 4\gamma^2 |\mathcal{S}| \|Q_k - Q^*\|_\infty^2 + 4\|Q - Q^*\|_D^2 + 2\sigma^2.$$

For later uses, we will bound $\|Q - Q^*\|_D^2$ in terms of $\nabla F(Q) = -(D\mathbf{T}Q_k - DQ)$. To this end, we first observe

$$\begin{aligned}
\mathbb{E}[\|g(Q;Q_k)\|_2^2 | Q_k, Q] \leq& 4\gamma^2 |\mathcal{S}| \|Q_k - Q^*\|_\infty^2 + 4\|Q - Q^*\|_D^2 + 2\sigma^2 \\
=& 4\gamma^2 |\mathcal{S}| \|Q_k - Q^*\|_\infty^2 + 4\|DQ - DQ^*\|_{D^{-1}}^2 + 2\sigma^2 \\
=& 4\gamma^2 |\mathcal{S}| \|Q_k - Q^*\|_\infty^2 + 4\|DQ^* - DQ - \nabla F(Q) + \nabla F(Q)\|_{D^{-1}}^2 + 2\sigma^2 \\
\leq& 4\gamma^2 |\mathcal{S}| \|Q_k - Q^*\|_\infty^2 + 8\|DQ^* - DQ - \nabla F(Q)\|_{D^{-1}}^2 + 8\|\nabla F(Q)\|_{D^{-1}}^2 + 2\sigma^2 \\
=& 4\gamma^2 |\mathcal{S}| \|Q_k - Q^*\|_\infty^2 + 8\|DQ^* - D\mathbf{T}Q_k\|_{D^{-1}}^2 + 8\|\nabla F(Q)\|_{D^{-1}}^2 + 2\sigma^2 \\
\leq& 4\gamma^2 |\mathcal{S}| \|Q_k - Q^*\|_\infty^2 + 8L\|Q^* - \mathbf{T}Q_k\|_2^2 + 8\|\nabla F(Q)\|_{D^{-1}}^2 + 2\sigma^2 \\
\leq& 4\gamma^2 |\mathcal{S}| \|Q_k - Q^*\|_\infty^2 + 8L|\mathcal{S}||\mathcal{A}|\|\mathbf{T}Q^* - \mathbf{T}Q_k\|_\infty^2 + 8\|\nabla F(Q)\|_{D^{-1}}^2 + 2\sigma^2 \\
\leq& 4\gamma^2 |\mathcal{S}| \|Q_k - Q^*\|_\infty^2 + 8|\mathcal{S}||\mathcal{A}|\gamma^2\|Q^* - Q_k\|_\infty^2 + 8\|\nabla F(Q)\|_{D^{-1}}^2 + 2\sigma^2 \\
\leq& 12\gamma^2 |\mathcal{S}||\mathcal{A}| \|Q^* - Q_k\|_\infty^2 + 8\|\nabla F(Q)\|_{D^{-1}}^2 + 2\sigma^2,
\end{aligned}$$

where the fourth line follows from $\|a+b\|^2 \leq 2\|a\|^2 + 2\|b\|^2$, the eight line comes from the contraction property of the Bellman operator, and the last line follows after simplifications.

Finally, we will find a bound on σ^2. By the definition, we have

$$\begin{aligned}
\sigma^2 =& \mathbb{E}[\|\hat{D}\hat{\mathbf{T}}Q^* - \hat{D}Q^*\|_2^2] \\
=& \mathbb{E}[\|(e_a \otimes e_s)((e_{s'})^T \gamma \Pi_{\pi_{Q^*}} Q^* - (e_a \otimes e_s)^T R - (e_a \otimes e_s)^T Q^*)\|_2^2] \\
\leq& \mathbb{E}[\|(e_a \otimes e_s)\|_2^2] \mathbb{E}[\|(e_{s'})^T \gamma \Pi_{\pi_{Q^*}} Q^* - (e_a \otimes e_s)^T R - (e_a \otimes e_s)^T Q^*\|_2^2] \\
\leq& \mathbb{E}[\|(e_{s'})^T \gamma \Pi_{\pi_{Q^*}} Q^* - (e_a \otimes e_s)^T R - (e_a \otimes e_s)^T Q^*\|_2^2] \\
\leq& 3\mathbb{E}[\|(e_{s'})^T \gamma \Pi_{\pi_{Q^*}} Q^*\|_2^2] + 3\mathbb{E}[\|(e_a \otimes e_s)^T R\|_2^2] + 3\mathbb{E}[\|(e_a \otimes e_s)^T Q^*\|_2^2] \\
\leq& 3\mathbb{E}[\|(e_{s'})^T \gamma \Pi_{\pi_{Q^*}}\|_2^2 \|Q^*\|_2^2] + 3\mathbb{E}[\|(e_a \otimes e_s)^T R\|_2^2] + 3\mathbb{E}[\|e_a \otimes e_s\|_2^2 \|Q^*\|_2^2] \\
\leq& 3|\mathcal{S}||\mathcal{A}|\mathbb{E}[\|(e_{s'})^T \gamma \Pi_{\pi_{Q^*}}\|_2^2 \|Q^*\|_\infty^2] + 3\mathbb{E}[\|(e_a \otimes e_s)^T R\|_2^2] + 3|\mathcal{S}||\mathcal{A}|\mathbb{E}[\|e_a \otimes e_s\|_2^2 \|Q^*\|_\infty^2] \\
\leq& 3|\mathcal{S}||\mathcal{A}|\|Q^*\|_\infty^2 + 3 + 3|\mathcal{S}||\mathcal{A}|\|Q^*\|_\infty^2 \\
\leq& 3|\mathcal{S}|\mathcal{A}|\frac{1}{(1-\gamma)^2} + 3 + 3|\mathcal{S}||\mathcal{A}|\frac{1}{(1-\gamma)^2} \\
\leq& \frac{9|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2},
\end{aligned}$$

where third line is due to the CauchySchwarz inequality and Holder's inequality, the forth line is due to $\mathbb{E}[\|e_a \otimes e_s\|_2^2] \leq 1$, the fifth line uses $\|a+b+c\|^2 \leq 3\|a\|^2 + 3\|b\|^2 + 3\|c\|^2$, the sixth line

follows from the CauchySchwarz inequality again, the eighth line comes from $\|(e_{s'})^T\gamma\Pi_{\pi_{Q^*}}\|_2^2 \leq 1$, $\|(e_a \otimes e_s)^T R\|_2^2 \leq 1$, $\|e_a \otimes e_s\|_2^2 \leq 1$, the ninth line is due to $\|Q^*\|_\infty \leq 1/(1-\gamma)$, and tenth line follows after simplifications. Combining the bound $\sigma^2 \leq 9|\mathcal{S}||\mathcal{A}|/(1-\gamma)^2$ with the previous result, the desired result follows. This completes the proof. ■

2.4. Proof of Lemma 8

We want to show that if $\mathbb{E}[\|Q_i - \mathbf{T}Q_{i-1}\|_2^2] \leq \varepsilon, \forall i \leq k$ and $\epsilon \leq (1-\gamma)^2$, then $\mathbb{E}[\|Q_k - Q^*\|_\infty^2] \leq \frac{8}{(1-\gamma)^2}$. We start with the following claim.

Claim: If $\mathbb{E}[\|Q_i - \mathbf{T}Q_{i-1}\|_2^2] \leq \varepsilon, i \in \{1, \ldots, T\}$ with $T \geq 1$, then we have

$$\mathbb{E}[\|Q_T - Q^*\|_\infty^2] \leq \left(\frac{1+\gamma^2}{1-\gamma^2}\right)\varepsilon\frac{2}{1-\gamma^2} + \left(\frac{\gamma^2+1}{2}\right)^T \mathbb{E}[\|Q_0 - Q^*\|_\infty^2].$$

This is because

$$\begin{aligned}
\mathbb{E}[\|Q_T - Q^*\|_\infty^2] =& \mathbb{E}[\|Q_T - \mathbf{T}Q_{T-1} + \mathbf{T}Q_{T-1} - Q^*\|_\infty^2] \\
\leq& \mathbb{E}[(\|Q_T - \mathbf{T}Q_{T-1}\|_\infty + \|\mathbf{T}Q_{T-1} - Q^*\|_\infty)^2] \\
\leq& \mathbb{E}[(1+\delta^{-1})\|Q_T - \mathbf{T}Q_{T-1}\|_\infty^2 + (1+\delta)\|\mathbf{T}Q_{T-1} - Q^*\|_\infty^2] \\
=& (1+\delta^{-1})\mathbb{E}[\|Q_T - \mathbf{T}Q_{T-1}\|_2^2] + (1+\delta)\mathbb{E}[\|\mathbf{T}Q_{T-1} - \mathbf{T}Q^*\|_\infty^2] \\
\leq& (1+\delta^{-1})\varepsilon + (1+\delta)\gamma^2\mathbb{E}[\|Q_{T-1} - Q^*\|_\infty^2],
\end{aligned}$$

where the second line is due to the triangle inequality, the third line is due to the fact that $\|a+b\|_2^2 \leq (1+\delta)\|a\|_2^2 + (1+\delta^{-1})\|b\|_2^2$ for any $\delta > 0$, and the last line is due to the hypothesis, $\mathbb{E}[\|Q_i - \mathbf{T}Q_{i-1}\|_2^2] \leq \varepsilon, i \in \{1, \ldots, T\}$. Note that we can choose $\delta > 0$ such that $(1+\delta)\gamma^2 < 1$ or equivalently, $\delta < \frac{1-\gamma^2}{\gamma^2}$. Simply choosing $\delta = \frac{1-\gamma^2}{2\gamma^2}$ yields

$$\mathbb{E}[\|Q_T - Q^*\|_\infty^2] \leq \left(1 + \frac{2\gamma^2}{1-\gamma^2}\right)\varepsilon + \frac{\gamma^2+1}{2}\mathbb{E}[\|Q_{T-1} - Q^*\|_\infty^2].$$

By the induction argument in T, we get

$$\begin{aligned}
\mathbb{E}[\|Q_T - Q^*\|_\infty^2] &\leq \left(1 + \frac{2\gamma^2}{1-\gamma^2}\right)\varepsilon\sum_{k=1}^{T}\left(\frac{\gamma^2+1}{2}\right)^k + \left(\frac{\gamma^2+1}{2}\right)^T \mathbb{E}[\|Q_0 - Q^*\|_\infty^2] \\
&\leq \left(1 + \frac{2\gamma^2}{1-\gamma^2}\right)\varepsilon\sum_{k=1}^{\infty}\left(\frac{\gamma^2+1}{2}\right)^k + \left(\frac{\gamma^2+1}{2}\right)^T \mathbb{E}[\|Q_0 - Q^*\|_\infty^2] \\
&= \left(\frac{1+\gamma^2}{1-\gamma^2}\right)\varepsilon\frac{2}{1-\gamma^2} + \left(\frac{\gamma^2+1}{2}\right)^T \mathbb{E}[\|Q_0 - Q^*\|_\infty^2],
\end{aligned}$$

which proves the claim. As an immediate result,

$$\mathbb{E}[\|Q_k - Q^*\|_\infty^2] \leq \left(\frac{1+\gamma^2}{1-\gamma^2}\right)\varepsilon\frac{2}{1-\gamma^2} + \left(\frac{\gamma^2+1}{2}\right)^k \mathbb{E}[\|Q_0 - Q^*\|_\infty^2]$$

$$
\begin{aligned}
&\leq \left(\frac{1+\gamma^2}{1-\gamma^2}\right)\varepsilon\frac{2}{1-\gamma^2}+\frac{4}{(1-\gamma)^2}\\
&\leq \frac{2}{(1-\gamma)^2}\varepsilon\frac{2}{1-\gamma^2}+\frac{4}{(1-\gamma)^2}\\
&\leq \frac{4}{1-\gamma^2}+\frac{4}{(1-\gamma)^2},\\
&\leq \frac{8}{(1-\gamma)^2},
\end{aligned}
\tag{3}
$$

where the first line comes from the claim, the second inequality is due to $\mathbb{E}[\|Q_0 - Q^*\|_\infty^2] \leq 1/((1-\gamma)^2)$ and $(\gamma^2+1)/2 < 1$, the third line comes from $\frac{1+\gamma^2}{1-\gamma^2} \leq \frac{2}{(1-\gamma)^2}$, the fourth line follows from the hypothesis $\varepsilon \leq (1-\gamma)^2$, and the last line follows after simplifications. For the case $k=0$, the bound is also valid since $\mathbb{E}[\|Q_0 - Q^*\|_\infty^2] \leq \frac{4}{(1-\gamma)^2}$. Taking the total expectation on both sides and combining it with (3), we arrive at the desired conclusion. ■

3. Proof of Theorem 3

The main line of the proof is to balance the inner and outer iteration numbers to achieve $\mathbb{E}[\|Q_T - Q^*\|_\infty] \leq \varepsilon$.

First of all, we define numbers $\xi_i > 0, i \in \{1,2,\ldots,T\}$ as those satisfying $\mathbb{E}[\|Q_i - \mathbf{T}Q_{i-1}\|_\infty] \leq \sqrt{\xi_i}, i \in \{1,2,\ldots,T\}$ through the previous inner iterations. By Proposition 1, we first have

$$
\mathbb{E}[\|Q_T - Q^*\|_\infty] \leq \sum_{k=1}^{T}\gamma^{T-k}\sqrt{\xi_k} + \gamma^T\mathbb{E}[\|Q_0 - Q^*\|_\infty] \leq \sum_{k=1}^{T}\gamma^{T-k}\sqrt{\xi_k} + \frac{2\gamma^T}{1-\gamma},
$$

where the last inequality is due to $\mathbb{E}[\|Q_0 - Q^*\|_\infty] \leq 2/(1-\gamma)$ and $\mathbb{E}[\|Q_i - \mathbf{T}Q_{i-1}\|_\infty] \leq \sqrt{\xi_i}, i \in \{1,2,\ldots,T\}$. Moreover, by Proposition 2, a constant $N_k = N$ implies that ξ_i can be uniformly bounded by a constant ξ. Letting $\xi_i = \xi$, the last inequality can be bounded as

$$
\mathbb{E}[\|Q_T - Q^*\|_\infty] \leq \frac{\sqrt{\xi}}{1-\gamma} + \frac{2\gamma^T}{1-\gamma}.
$$

To achieve $\mathbb{E}[\|Q_T - Q^*\|_\infty] \leq \varepsilon$, it suffices to ensure

$$
\begin{aligned}
\text{Inner iteration}: \frac{\sqrt{\xi}}{1-\gamma} &\leq \frac{1}{2}\varepsilon,\\
\text{Outer iteration}: \frac{2\gamma^T}{1-\gamma} &\leq \frac{1}{2}\varepsilon.
\end{aligned}
$$

By rearranging terms and taking the logarithm on both sides of the second inequality, it follows that the second inequality is equivalent to

$$
T \geq \ln\left(\frac{1-\gamma}{4}\varepsilon\right)/\ln\gamma = \ln\left(\frac{4}{(1-\gamma)\varepsilon}\right)/\ln\gamma^{-1}.
$$

The first inequality is equivalent to $\xi \leq \frac{\varepsilon^2(1-\gamma)^2}{4}$, which is ensured if

$$
\mathbb{E}[\|Q_i - \mathbf{T}Q_{i-1}\|_\infty] \leq \sqrt{\frac{\varepsilon^2(1-\gamma)^2}{4}}, \quad \forall i \in \{1,2,\ldots,T\}. \tag{4}
$$

The square of the left-hand side of the above inequality is bounded by

$$\begin{aligned}\mathbb{E}[\|Q_i - \mathbf{T}Q_{i-1}\|_\infty]^2 \leq& \mathbb{E}[\|Q_i - \mathbf{T}Q_{i-1}\|_2]^2 \\ \leq& \mathbb{E}[\|Q_i - \mathbf{T}Q_{i-1}\|_2^2] \\ \leq& c^{-1}\mathbb{E}[\|Q_i - \mathbf{T}Q_{i-1}\|_D^2] \\ \leq& \frac{1}{\lambda + N}\frac{512|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}\frac{L}{c^3},\end{aligned}$$

where the second line is due to the Jensen's inequality and the last line is due to Theorem 3. Therefore, a sufficient condition to ensure (4) is $\frac{1}{\lambda+N}\frac{512|\mathcal{S}||\mathcal{A}|}{(1-\gamma)^2}\frac{L}{c^3} \leq \frac{\varepsilon^2(1-\gamma)^2}{4}$. The last inequality is implied by

$$\frac{2048|\mathcal{S}||\mathcal{A}|}{\varepsilon^2(1-\gamma)^4}\frac{L}{c^3} \leq N.$$

In summary, if the total number of inner iteration N is greater than or equal to $\frac{2048|\mathcal{S}||\mathcal{A}|}{\varepsilon^2(1-\gamma)^4}\frac{L}{c^3}$, then the first inequality $\frac{\sqrt{\xi}}{1-\gamma} \leq \frac{1}{2}\varepsilon$ holds. By combining the total number of inner iterations and outer iterations, the total number of samples that are required to achieve $\mathbb{E}[\|Q_T - Q^*\|_\infty] \leq \varepsilon$ is $\frac{2048|\mathcal{S}||\mathcal{A}|}{\varepsilon^2(1-\gamma)^4 \ln\gamma^{-1}}\frac{L}{c^3}\ln\left(\frac{4}{(1-\gamma)\varepsilon}\right)$. This completes the proof. ■