Performance of Q-learning with Linear Function Approximation: Stability and Finite Time Analysis

Zaiwei Chen[1], **Sheng Zhang**[2], **Thinh T. Doan**[2], **Siva Theja Maguluri**[2], and **John-Paul Clarke**[2]

[1]Department of Aerospace Engineering, Georgia Institute of Technology
[2]Department of Industrial and Systems Engineering, Georgia Institute of Technology

Abstract

In this paper, we consider the model-free reinforcement learning problem and study the popular Q-learning algorithm with linear function approximation for finding the optimal policy. Despite its popularity, it is known that Q-learning with linear function approximation may diverge in general due to off-policy sampling. Our main contribution is to provide a finite-time bound on the performance of Q-learning with linear function approximation and constant step size under an assumption on the sampling policy. Unlike some prior work in the literature, we do not need to make the unnatural assumption that the samples are i.i.d. (since they are Markovian), and do not require an additional projection step in the algorithm. To show this result, we first consider a more general nonlinear stochastic approximation algorithm with Markovian noise, and derive a finite-time bound on the mean-square error, which we believe is of independent interest. Our proof is based on Lyapunov drift arguments and exploits the geometric mixing of the underlying Markov chain. We also provide numerical simulations to illustrate the effectiveness of our assumption on the sampling policy, and demonstrate the rate of convergence of Q-learning with linear function approximation.

1 Introduction

Reinforcement learning (RL) is a framework for solving sequential decision-making problems by repeatedly interacting with the environment [25]. This approach has demonstrated tremendous successes for solving many practical problems in several different areas, such as robotics [17], power management [28], autonomous driving [22], and board games [23].

An RL problem is often modeled as a Markov decision process (MDP) with unknown transition probabilities. The goal of the agent is to find an optimal policy to select actions so that the expected total future reward is maximized. Among potential methods, Q-learning, studied in [32], has been recognized as a promising solution for finding the optimal policy since it does not require any knowledge of the environment model. In particular, Q-learning iteratively estimates the optimal Q-function (state-action value function) based on a sequence of samples generated by applying a sampling policy to the unknown model. The optimal policy is then computed based on the optimal Q-function. This makes Q-learning an off-policy approach since it learns the optimal policy through data generated by a (possibly) non-optimal policy. This further has the advantage that learning can be decoupled from sampling and can be performed using data that is already collected.

Given the popularity and success of Q-learning, its performance has been studied in the literature. The asymptotic convergence of Q-learning has been studied in [32, 15, 29] using a martingale-based approach, while the rate of convergence has been characterized in [12, 27, 2, 13, 10]. However, since Q-learning requires to store the Q-function values for all state-action pairs, it has been limited to problems with small state and action spaces, and this challenge is often referred to as the curse

Optimization Foundations for Reinforcement Learning Workshop at NeurIPS 2019, Vancouver, Canada.

of dimensionality in RL. To overcome this drawback, we can approximate the Q-function by a parameterized function class with much smaller dimension. However, Q-learning with function approximation can in general diverge [1, 25]. The main reason is that Q-learning uses off-policy sampling to collect the data, making it potentially an expansion mapping [14, 1]. For this reason, the convergence of Q-learning with function approximation has been limited to special cases, such as, for optimal stopping problems [31], or when using state aggregation functions [4], or when using non-parametric regression method (*nearest neighbor Q-learning*) [21].

In this paper, we focus on studying Q-learning with linear function approximation [19, 33], where the Q-function is approximated by a linear combination of a given set of basis functions (or features). Q-learning with linear function approximation can diverge in general as illustrated by the counter examples in [1, 25]. Therefore, certain conditions need to be imposed on the sampling policy to guarantee the stability of this approach. One such condition was proposed in [19] to restrict the sampling policy to be close enough to the optimal policy. Later, based on [19], the work in [33] studies the finite-time analysis of Q-learning with linear function approximation. In particular, the approach in [33] is mainly motivated by the work in [5], where the convergence rate of the popular temporal-difference learning method for solving policy evaluation problems was studied. One drawback of the techniques used in [5, 33] is that: to stabilize the iterates, their algorithm requires an additional projection step onto a bounded set related to the unknown stationary distribution of the underlying Markov chain, which in some cases is impractical.

The contributions of this paper are threefold. Motivated by the work in [19], we first provide a new condition on the sampling policy in Q-learning to guarantee the stability of the algorithm. Second, we analyze a finite-time bound on the performance of Q-learning without requiring any projection steps. Our key technique is to view Q-learning as a stochastic approximation (SA) algorithm for finding the solution of a suitable nonlinear equation. We then study the convergence rate of such SA under general conditions, which we believe is of independent interest and may be applicable for other problems. Finally, we present some numerical experiments based on the example from [1] that shows divergence of Q-learning with linear function approximation. In particular, we illustrate the sufficiency of our proposed condition for the stability of the algorithm, and demonstrate the rate of convergence.

2 MDP and Q-learning

Consider an MDP denoted by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$, where $\mathcal{S}$ is a finite state space of size n, $\mathcal{A}$ is a finite action space of size m, and $\mathcal{P} = \{P_a \in \mathbb{R}^{n\times n} \mid a \in \mathcal{A}\}$ is a set of action dependent transition probability matrices. Moreover, $\mathcal{R} : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ is the reward function and $\gamma \in (0, 1)$ is the discount factor.

For each time step $k \geq 0$, the agent observes the current state $S_k = s$ of the environment and takes an action $A_k = a$ according to some policy π, which can be stochastic ($a \sim \pi(\cdot|S_k)$) or deterministic ($a = \pi(S_k)$). The system then moves to the next state $S_{k+1} = s'$ with probability $P_a(s, s')$. Moreover, as the transition occurs, the agent receives an instantaneous reward $\mathcal{R}(s, a)$. The goal of the agent is to find an optimal policy π^* such that its long term cumulative reward is maximized. Specifically, given a policy π, the value function at state s is defined to be the total expected reward starting from s, and then following policy π:

$$V_\pi(s) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k \mathcal{R}(S_k, A_k) \;\middle|\; S_0 = s\right], \; A_k \sim \pi(\cdot|S_k), \; \forall\, k \geq 0.$$

The goal is to find an optimal policy π^* such that $V_{\pi^*}(s) = \max_\pi V_\pi(s)$ for all state s. Denote the value function associated with π^* by V^*, it is well known that V^* satisfies the following Bellman equation:

$$V^*(s) = \max_{a\in\mathcal{A}} \underbrace{\{\mathcal{R}(s, a) + \gamma \mathbb{E}\left[V^*(s') \mid s, a\right]\}}_{Q^*(s,a)}, \tag{1}$$

where s' is the successor state after taking action a at state s. Let $Q^* : \mathcal{S} \times \mathcal{A} \mapsto \mathbb{R}$ be defined as in Eq. (1), we have $V^*(s) = \max_{a\in\mathcal{A}} Q^*(s, a)$, and therefore Q^* verifies another Bellman equation:

$$Q^*(s, a) = \mathcal{R}(s, a) + \gamma \mathbb{E}_{s'}\left[\max_{a'\in\mathcal{A}} Q^*(s', a') \;\middle|\; s, a\right]. \tag{2}$$

Once Q^* is obtained, an optimal policy π^* can be decided as $\pi^*(s) \in \arg\max_{a\in\mathcal{A}} Q^*(s,a)$, for all state s, which does not require any knowledge about the transition probabilities, and so is a model-free approach. In terms of finding Q^*, Q-learning can be viewed as an SA algorithm for finding the solution of the Bellman equation (2). In particular, given a sample trajectory $\{(S_k, A_k)\}$ generated by some "nice" policy π, Q-learning iteratively updates the estimate Q_k of Q^* as

$$Q_{k+1}(S_k,A_k) = Q_k(S_k,A_k) + \epsilon_k(S_k,A_k)\left[\mathcal{R}(S_k,A_k) + \gamma \max_{a\in\mathcal{A}} Q_k(S_{k+1},a) - Q_k(S_k,A_k)\right], \quad (3)$$

where $\{\epsilon_k(s,a)\}$ is the sequence of step sizes associated with the state-action pair (s,a). The sequence Q_k generated by Q-learning converges to Q^* w.p. 1 as long as every state-action pair is visited infinitely often under the sampling policy π, and the step size diminishes to zero at a proper rate [4, 29].

Observe that the estimates $\{Q_k\}$ are represented by a look-up table rather than an analytical expression. When the number of state-action pairs is very large, Q-learning can be intractable due to the curse of dimensionality. To overcome this difficulty, we use low-dimensional approximation $\tilde{Q}$ of Q^*, restricting $\tilde{Q}$ to a linear subspace $\mathcal{Q}$ with dimension $d \ll mn$. In particular, given a set of basis functions $\phi_\ell : \mathcal{S}\times\mathcal{A} \mapsto \mathbb{R}$, $\ell \in \{1,\dots,d\}$ called features, the approximation of Q^*, parameterized by a weight vector $\theta \in \mathbb{R}^d$, is given by $\tilde{Q}_\theta(s,a) = \phi(s,a)^\intercal\theta$, where $\phi(s,a) := (\phi_1(s,a),\cdots,\phi_d(s,a))^\intercal$. The Q-learning with linear function approximation for iteratively updating θ is then given by

$$\theta_{k+1} = \theta_k + \epsilon\phi(S_k,A_k)\left[\mathcal{R}(S_k,A_k) + \gamma\max_{a\in\mathcal{A}}\phi(S_{k+1},a)^\intercal\theta_k - \phi(S_k,A_k)^\intercal\theta_k\right], \quad (4)$$

where $\epsilon > 0$ is a constant step size. Our goal in this paper is to provide a finite-time error bound for Algo. (4). Note that unlike the work in [33], we do not assume a projection step to a predefined set related to the unknown transition probabilities of the underlying Markov chain. Finally, as mentioned above, Q-learning can be viewed as a nonlinear SA for solving the Bellman equation (2). Motivated by this observation, we first study the convergence rate of an SA algorithm for finding the solution of a general nonlinear equation. By utilizing this result, we then provide a finite-time bound for Q-learning with linear function approximation.

3 Main Results

In this section we present our main results. Specifically, the finite-time analysis of a general nonlinear SA is presented in Section 3.1. Following from this result, we provide the finite-time error bound of Q-learning with linear function approximation in Section 3.2. Further discussions about the condition on the sampling policy to guarantee the stability of Q-learning as well as numerical experiments are presented in Section 3.3. The proofs of all theorems can be found in the Appendix.

3.1 Finite time analysis of nonlinear stochastic approximation

In this section, we establish the finite-time error bound of a general nonlinear SA. Consider the problem of solving for θ^* in the equation

$$\bar{F}(\theta) := \mathbb{E}_\mu[F(X,\theta)] = 0,$$

where X is a random variable with finite state space $\mathcal{X}$ and distribution μ, which is assumed to be unknown. The function $F : \mathcal{X}\times\mathbb{R}^d \mapsto \mathbb{R}^d$ is a general nonlinear mapping. To solve this problem, we use the celebrated SA algorithm proposed in [20]. In particular, suppose that we can collect a sequence of samples $\{X_k\}$ of the random variable X. Then, with initialization θ_0, SA iteratively updates an estimate θ_k of θ^* as

$$\theta_{k+1} = \theta_k + \epsilon F(X_k,\theta_k). \quad (5)$$

A concrete example of this equation is the Q-learning update (4) given in the previous section. Under reasonable assumptions on the nonlinear mapping F, and i.i.d. assumption on the sequence $\{X_k\}$, the convergence properties of SA have been studied extensively in [26, 6]. In particular, in the context of optimization, [7] and the reference therein studied the same problem under the name of stochastic

gradient descend (SGD). On the other hand, when the samples $\{X_k\}$ are obtained from a Markov chain with stationary distribution μ, the asymptotic convergence of SA is provided in [3, 4, 6] using the ODE approach. That is, the sequence $\{\theta_k\}$ generated by Eq. (5) is shown to converge to the equilibrium point of the ODE

$$\dot{\theta} = \bar{F}(\theta) \tag{6}$$

under certain stability assumptions on the ODE (6) [16, 3]. As for the convergence rate, in the setting of linear SA, i.e., F is linear in terms of θ, finite-time analysis has been performed in [8] with i.i.d. noise, and in [24, 5] with Markovian noise. In optimization, [11] studied the SGD under Markovian noise, but a projection step is required there to maintain the stability.

Our goal here is to expand the frontier by providing a finite-time bound for general nonlinear SA (5) under Markovian noise without requiring a projection step in the algorithm. To do that, we start by presenting a sequence of standard assumptions. Throughout this paper, $\|\cdot\|$ stands for the Euclidean norm for vectors, and induced 2-norm for matrices. Our assumptions are given as follows.

Assumption 3.1. *The Markov chain $\{X_k\}$ is irreducible and aperiodic.*

Remark 3.1. *Assumption 3.1 is often assumed to study the asymptotic convergence of SA under Markovian noise; see for example [6, 4, 25].*

Assumption 3.2. *The function $F(x,\theta)$ is globally Lipschitz continuous with respect to θ uniformly in x, i.e., there exists $L > 0$ s.t. $\|F(x,\theta_1) - F(x,\theta_2)\| \leq L\|\theta_1 - \theta_2\|$, for all $\theta_1, \theta_2 \in \mathbb{R}^d$, $x \in \mathcal{X}$.*

Remark 3.2. *The Lipschitz continuity of F is sufficient to study a nonlinear system, e.g., it guarantees that ODE (6) has a unique solution. When $F(x,\theta)$ is linear in terms of θ as considered in [4, 24], i.e., $F(x,\theta) = A(x)\theta + b(x)$, Assumption 3.2 is automatically satisfied.*

Note that Lipschitz continuity implies at most linear growth rate in terms of $\|\theta\|$ for both $\|F(x,\theta)\|$ and $\|\bar{F}(\theta)\|$. In particular, assume without loss of generality that $L \geq \max_{x\in\mathcal{X}} \|F(x,0)\|$. By letting $\theta_1 = \theta$ and $\theta_2 = 0$ in Assumption 3.2, it is easy to show using triangular inequality and Jensen's inequality that

$$\|F(x,\theta)\| \leq L(\|\theta\| + 1), \text{ and } \|\bar{F}(\theta)\| \leq L(\|\theta\| + 1), \ \forall\, \theta \in \mathbb{R}^d, x \in \mathcal{X}.$$

Assumption 3.3. *The equation $\bar{F}(\theta) = 0$ has a solution θ^*, and there exists $\alpha > 0$ such that*

$$(\theta - \theta^*)^\intercal(\bar{F}(\theta) - \bar{F}(\theta^*)) \leq -\alpha\|\theta - \theta^*\|^2, \ \forall\, \theta \in \mathbb{R}^d. \tag{7}$$

Remark 3.3. *This assumption can be viewed as a strongly monotone property of the nonlinear mapping $-F$, or an exponential dissipativeness property of the dynamical system (6) with zero supply rate and a quadratic storage function.*

Before we present the finite-time error bound of SA (5), the following definition and lemma regarding the mixing time of the Markov chain $\{X_k\}$ are needed.

Definition 3.1. *For any $\delta > 0$, let τ_δ be defined as*

$$\tau_\delta = \min\{t \geq 1 : \|\mathbb{E}[F(X_k,\theta)|X_0 = x] - \bar{F}(\theta)\| \leq \delta(\|\theta\| + 1), \forall k \geq t, \forall\theta \in \mathbb{R}^d, \forall x \in \mathcal{X}\}.$$

Lemma 3.1. *Suppose Assumptions 3.1 and 3.2 hold, then there exists $L_1 > 0$ s.t. for any $\delta > 0$:*

$$\tau_\delta \leq L_1(\log(1/\delta) + 1)$$

(See Appendix A for the proof).

Remark 3.4. *Here τ_δ is referred to as the mixing time of the underlying Markov chain, which by Lemma 3.1 satisfies $\lim_{\delta\to 0} \delta\tau_\delta = 0$. Since we will always use $\delta = \epsilon$, where ϵ is the constant step size, for convenience, we drop the subscript ϵ in τ_ϵ in the following.*

We will now present our finite-time error bound for SA (5).

Theorem 3.1. *Consider iterates $\{\theta_k\}$ generated by SA (5). Suppose that Assumptions 3.1–3.3 hold, and $\epsilon\tau \leq \min(1/4L, \alpha/114L^2)$ (α is given in (7)), then we have for all $k \geq \tau$:*

$$\mathbb{E}[\|\theta_k - \theta^*\|^2] \leq \beta_1(1 - \alpha\epsilon)^{k-\tau} + \frac{\beta_2}{\alpha}\epsilon\tau, \tag{8}$$

where $\beta_1 = (\|\theta_0\| + \|\theta_0 - \theta^*\| + 1)^2$, and $\beta_2 = 114L^2(\|\theta^*\| + 1)^2$ (See Appendix B for the proof).

The r.h.s. of Eq. (8) can be viewed as a combination of the bias and variance. The first term shows the impact of the bias due to the initial estimate θ_0, which goes to zero geometrically fast. The second term corresponds to the variance of the Markovian noise. Since we use constant step size, the variance does not go to zero. However, note that we have $\lim_{\epsilon\to 0}\epsilon\tau = 0$ by Lemma 3.1, the variance vanishes when the step size diminishes to zero. We will next use Theorem 3.1 to establish the finite-time bound of Q-learning with linear function approximation.

3.2 Finite-time analysis of Q-learning with linear function approximation

First of all, we show that Q-learning can be reformulated as a variant of the nonlinear SA studied in Section 3.1. Recall the update of Q-learning from Eq. (4). Since π is predetermined and fixed, the MDP becomes a Markov chain $\{S_k\}$. Let $X_k = (S_k, A_k, S_{k+1})$, it is clear that $\{X_k\}$ is also a Markov chain with state space

$$\mathcal{X} = \{(s, a, s') \mid s \in \mathcal{S},\ \pi(a|s) > 0,\ P_a(s, s') > 0\}.$$

Thus, Eq. (4) can now be rewritten in the same form as the nonlinear SA given in Section 3.1:

$$\theta_{k+1} = \theta_k + \epsilon F(X_k, \theta_k), \tag{9}$$

where the nonlinear mapping F is defined as

$$F(x, \theta) = F((s, a, s'), \theta) = \phi(s, a)\left[\mathcal{R}(s, a) + \gamma \max_{a'\in\mathcal{A}} \phi(s', a')^\intercal\theta - \phi(s, a)^\intercal\theta\right]. \tag{10}$$

We next present the finite-time error bound of Algo. (9). Before that, we assume that the feature vectors $\{\phi_\ell\}_{1\le\ell\le d}$ are linearly independent and are normalized so that $\|\phi(s, a)\| \le 1$ for all (s, a). Note that this is w.o.l.g. since we can disregard dependent features. Let $r_{\max} = \max_{(s,a)} |\mathcal{R}(s, a)|$.

Theorem 3.2. *Consider iterates $\{\theta_k\}$ generated by Algo. (9). Suppose that:*
(a) The Markov chain $\{S_k\}$ induced by π is irreducible and aperiodic.
(b) The equation $\bar{F}(\theta) = 0$ has a solution θ^*, and

$$\gamma^2\mathbb{E}_\mu[\max_{a'\in\mathcal{A}}(\phi(s, a')^\intercal\theta)^2] - \mathbb{E}_\mu[(\phi(s, a)^\intercal\theta)^2] \le -\alpha\|\theta\|^2,\ \forall\ \theta \in \mathbb{R}^d. \tag{11}$$

(c) $\epsilon\tau \le \min(1/4L, \alpha/228L^2)$, where $L = \gamma + 1 + r_{\max}$.
Then we have for all $k \ge \tau$:

$$\mathbb{E}[\|\theta_k - \theta^*\|^2] \le \beta_1\left(1 - \frac{\alpha}{2}\epsilon\right)^{k-\tau} + \frac{2\beta_2}{\alpha}\epsilon\tau, \tag{12}$$

where $\beta_1 = (\|\theta_0\| + \|\theta_0 - \theta^*\| + 1)^2$, and $\beta_2 = 114L^2(\|\theta^*\| + 1)^2$ (See Appendix C for the proof).

Remark 3.5. *Condition (a) is standard in studying finite MDPs [4, 30]. As for condition (b), the solution θ^* to the Eq. $\bar{F}(\theta) = 0$ can be equivalently characterized as a solution to a projected Bellman equation [19]. However, we need to point out that the solution to such fixed point equation may not exist in general [9]. Eq. (11) is essentially a requirement on the sampling policy π to stabilize Q-learning, which will be analyzed in detail in the next section.*

Theorem 3.2 is qualitatively similar to Theorem 3.1 in that Q-learning achieves an exponential convergence rate in expectation to a ball centered at θ^*, and the size of this ball shrinks as a function of the step size ϵ. Eq. (11) is essentially to guarantee the stability of Q-learning with linear function approximation, i.e., to satisfy Assumption 3.3 in the case of nonlinear SA. A weaker form of (11) is

$$\gamma^2\mathbb{E}_\mu[\max_{a'\in\mathcal{A}}(\phi(s, a')^\intercal\theta)^2] < \mathbb{E}_\mu[(\phi(s, a)^\intercal\theta)^2],\ \forall\ \theta \ne 0, \tag{13}$$

which does not give exponential rate, but can be used to establish a.s. convergence [19]. In the next section, we present further discussions about when condition (13) may be satisfied and also present numerical simulations to verify its sufficiency.

3.3 Discussion and numerical experiments

We start by giving some illustration on condition (13). Consider Eq. (13). Since the action is chosen according to π on the r.h.s. while greedily on the l.h.s., clearly we have

$$\mathbb{E}_\mu[\max_{a\in\mathcal{A}}(\phi(s, a)^\intercal\theta)^2] \ge \mathbb{E}_\mu[(\phi(s, a)^\intercal\theta)^2],\ \forall\ \theta \in \mathbb{R}^d.$$

To meet condition (13), besides the presence of γ^2, there should be some requirements on the sampling policy π. To make it explicit, define

$$\delta(\pi) := \min_{\|\theta\|=1} \left\{ \mathbb{E}_\mu[(\phi(s,a)^\intercal \theta)^2] \Big/ \mathbb{E}_\mu[\max_{a \in \mathcal{A}} (\phi(s,a)^\intercal \theta)^2] \right\} \tag{14}$$

Note that $\delta(\pi) \in [0,1]$ for any policy π. Moreover, from Eq. (14) we see that the larger $\delta(\pi)$ is, the closer π is to the policy induced greedily from $\tilde{Q}_\theta^2$ for all θ. On the contrary, when $\delta(\pi)$ is small, convergence can happen only when the discount factor γ is sufficiently small. One way to compute $\delta(\pi)$ is presented in Appendix D for an MDP with a sampling policy π when the underlying model is known, which is used to perform numerical simulations.

In our simulation, we consider the divergent counter example of Q-learning introduced in [1], (see full description of this example in Appendix D). Since the reward function is identically zero, Q^* is zero, implying θ^* is zero. Because of this structure, it is possible for the Q-learning algorithm to converge even when constant step size is used. We choose the sampling policy π which takes each action with equal probability. It turns out that $\delta(\pi) \approx 0.5$, giving the threshold for γ being $\delta(\pi)^{1/2} \approx 0.7$. In our simulation, we choose $\epsilon = 0.01$, $\gamma \in \{0.7, 0.9, 0.97\}$, and plot $\|\theta_k\|$ as a function of the number of iterations k in Figure 1. Here, θ_k converges when $\gamma = 0.7$ and also when $\gamma = 0.9$ and diverges when $\gamma = 0.97$. This demonstrates that condition (13) is sufficient but not necessary for convergence. This also shows that by modifying the problem to ensure (13), the counter example from [1] can be made to converge. Finally, to show the convergence rate of Q-learning, we consider the convergence of θ_k when $\gamma = 0.7$ given in Figure 2, where we plot $\log \mathbb{E}[\|\theta_k\|^2]$ as a function on the number of iterations k. In this case, θ_k seems to converge exponentially to 0, which agrees with our theoretical result given in Theorem 3.2.

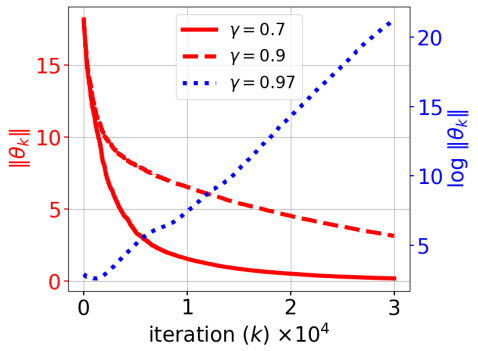

Figure 1: Convergence of Q-learning with linear function approximation for different discount factor γ

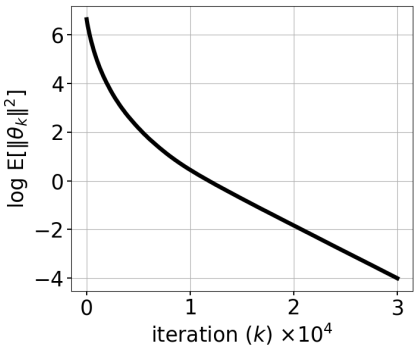

Figure 2: Exponentially fast convergence of Q-learning with linear function approximation for $\gamma = 0.7$

4 Conclusion

In this paper we establish a finite-time bound for Q-learning with linear function approximation and a constant step size, without either making an i.i.d. noise assumption, or requiring an additional projection step to bound the iterates. Our approach is to obtain finite-time bounds for a more general nonlinear SA algorithm with Markovian noise. We also provide sufficient conditions for the stability of Q-learning, and study the need of this condition numerically in the context of a well-known counter example. Future work includes obtaining finite-time error bounds under diminishing step sizes and optimizing the step sizes to achieve the best convergence rate. Since (11) is quite restrictive on the sampling policy π, another future direction is to relax this condition by considering using time-varying sampling policy. Studying finite-time error bounds for the on-policy variant of Q-learning called SARSA is probably a first step in this direction.

References

[1] L. Baird. Residual algorithms: Reinforcement learning with function approximation. In *Machine Learning Proceedings 1995*, pages 30–37. Elsevier, 1995.

[2] C. L. Beck and R. Srikant. Error bounds for constant step-size q-learning. *Systems & Control Letters*, 61:1203–1208, 2012.

[3] A. Benveniste, M. Métivier, and P. Priouret. *Adaptive algorithms and stochastic approximations*, volume 22. Springer Science & Business Media, 2012.

[4] D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-dynamic programming*, volume 5. Athena Scientific Belmont, MA, 1996.

[5] J. Bhandari, D. Russo, and R. Singal. A finite time analysis of temporal difference learning with linear function approximation. *arXiv preprint arXiv:1806.02450*, 2018.

[6] V. S. Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

[7] L. Bottou, F. E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning. *Siam Review*, 60(2):223–311, 2018.

[8] G. Dalal, B. Szörényi, G. Thoppe, and S. Mannor. Finite sample analyses for td (0) with function approximation. In *Thirty-Second AAAI Conference on Artificial Intelligence*, 2018.

[9] D. P. De Farias and B. Van Roy. On the existence of fixed points for approximate value iteration and temporal-difference learning. *Journal of Optimization theory and Applications*, 105(3): 589–608, 2000.

[10] A. M. Devraj and S. Meyn. Zap q-learning. In *Advances in Neural Information Processing Systems 30*. 2017.

[11] J. C. Duchi, A. Agarwal, M. Johansson, and M. I. Jordan. Ergodic mirror descent. *SIAM Journal on Optimization*, 22(4):1549–1578, 2012.

[12] E. Even-Dar and Y. Mansour. Learning rates for Q-learning. *Journal of Machine Learning Research*, 5(Dec):1–25, 2003.

[13] M. Ghavamzadeh, H. J. Kappen, M. G. Azar, and R. Munos. Speedy q-learning. In *Advances in Neural Information Processing Systems 24*. 2011.

[14] G. J. Gordon. Stable function approximation in dynamic programming. In *Machine Learning Proceedings 1995*, pages 261–268. Elsevier, 1995.

[15] T. Jaakkola, M. I. Jordan, and S. P. Singh. Convergence of stochastic iterative dynamic programming algorithms. In *Advances in neural information processing systems*, pages 703–710, 1994.

[16] H. K. Khalil and J. W. Grizzle. *Nonlinear systems*, volume 3. Prentice hall Upper Saddle River, NJ, 2002.

[17] J. Kober, J. A. Bagnell, and J. Peters. Reinforcement learning in robotics: A survey. *The International Journal of Robotics Research*, 32(11):1238–1274, 2013.

[18] D. A. Levin and Y. Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

[19] F. S. Melo, S. P. Meyn, and M. I. Ribeiro. An analysis of reinforcement learning with function approximation. In *Proceedings of the 25th international conference on Machine learning*, pages 664–671. ACM, 2008.

[20] H. Robbins and S. Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

[21] D. Shah and Q. Xie. Q-learning with nearest neighbors. *CoRR*, abs/1802.03900, 2018. URL `http://arxiv.org/abs/1802.03900`.

[22] S. Shalev-Shwartz, S. Shammah, and A. Shashua. Safe, multi-agent, reinforcement learning for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.

[23] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550 (7676):354, 2017.

[24] R. Srikant and L. Ying. Finite-time error bounds for linear stochastic approximation and TD learning. *arXiv preprint arXiv:1902.00923*, 2019.

[25] R. S. Sutton and A. G. Barto. *Introduction to Reinforcement Learning*. MIT Press Cambridge, 2 edition, 2018.

[26] R. S. Sutton, H. R. Maei, and C. Szepesvári. A convergent $o(n)$ temporal-difference algorithm for off-policy learning with linear function approximation. In *Advances in neural information processing systems*, pages 1609–1616, 2009.

[27] C. Szepesvári. The asymptotic convergence-rate of q-learning. In *Proceedings of the 10th International Conference on Neural Information Processing Systems*, NIPS'97, 1997.

[28] G. Tesauro, R. Das, H. Chan, J. Kephart, D. Levine, F. Rawson, and C. Lefurgy. Managing power consumption and performance of computing systems using reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 1497–1504, 2008.

[29] J. N. Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Machine learning*, 16 (3):185–202, 1994.

[30] J. N. Tsitsiklis and B. Van Roy. Average cost temporal-difference learning. *Automatica*, 35(11): 1799–1808, 1999.

[31] J. N. Tsitsiklis and B. Van Roy. Optimal stopping of markov processes: Hilbert space theory, approximation algorithms, and an application to pricing high-dimensional financial derivatives. *IEEE Transactions on Automatic Control*, 44(10):1840–1851, 1999.

[32] C. J. Watkins and P. Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

[33] S. Zou, T. Xu, and Y. Liang. Finite-sample analysis for sarsa and Q-learning with linear function approximation. *arXiv preprint arXiv:1902.02234*, 2019.

A Proof of Lemma 3.1

Since $\{X_k\}$ is irreducible and aperiodic, by Theorem 4.9 in [18], it has a unique stationary distribution μ, and there exist constants $C > 0$ and $\rho \in (0, 1)$ such that

$$d_{TV}(P(X_k = \cdot | X_0 = x), \mu) \leq C\rho^k, \quad \forall k \geq 0, x \in \mathcal{X},$$

where $d_{TV}(\cdot, \cdot)$ is the total-variation distance defined as

$$d_{TV}(P, Q) = \sup_{\{f: \|f\|_\infty \leq \frac{1}{2}\}} \left| \int f dP - \int f dQ \right|. \tag{15}$$

Let $F_i(x, \theta)$ be the i-th component of $F(x, \theta)$. Under Assumption 3.2 we have

$$|F_i(x, \theta)| \leq \|F(x, \theta)\| \leq L(\|\theta\| + 1), \quad \forall x \in \mathcal{X}, \theta \in \mathbb{R}^d.$$

Therefore, we obtain for any $\theta \in \mathbb{R}^d$ and $x \in \mathcal{X}$:

$$\begin{aligned}
&\|E[F(X_k, \theta)|X_0 = x] - E_\mu[F(X, \theta)]\| \\
\leq &\sum_{i=1}^{d} |E[F_i(X_k, \theta)|X_0 = x] - E_\mu[F_i(X, \theta)]| \\
= &2L(\|\theta\| + 1) \sum_{i=1}^{d} \left| E\left[\frac{F_i(X_k, \theta)}{2L(\|\theta\| + 1)} \,\middle|\, X_0 = x \right] - E_\mu\left[\frac{F_i(X, \theta)}{2L(\|\theta\| + 1)} \right] \right| \\
\leq &2L(\|\theta\| + 1) dC\rho^k,
\end{aligned}$$

Solve

$$2L(\|\theta\| + 1) dC\rho^t \leq \delta(\|\theta\| + 1)$$

and we get

$$t \geq \frac{\log \frac{1}{\delta} + \log(2LCd)}{\log \frac{1}{\rho}}.$$

Thus, by definition of τ_δ, we have

$$\tau_\delta \leq \frac{\log \frac{1}{\delta} + \log(2LCd)}{\log \frac{1}{\rho}} \leq L_1(\|\theta\| + 1),$$

where

$$L_1 = \frac{1 + |\log(2LCd)|}{\log\left(\frac{1}{\rho}\right)}.$$

B Proof of Theorem 3.1

The proof of Theorem 3.1 involves the following sequence of Lemmas.

Lemma B.1. *The following inequality holds for all $k \geq \tau$:*

$$E[\|\theta_{k+1} - \theta_k\|^2 | X_{k-\tau}, \theta_{k-\tau}] \leq 2L^2\epsilon^2 \left[E[\|\theta_k - \theta^*\|^2 | X_{k-\tau}, \theta_{k-\tau}] + (\|\theta^*\| + 1)^2 \right].$$

Proof of Lemma B.1. Assumption 3.2 implies $\|F(x, \theta)\| \leq L(\|\theta\| + 1)$ for all $x \in \mathcal{X}$ and $\theta \in \mathbb{R}^d$. Thus, using the update rule (5) and we have

$$\|\theta_{k+1} - \theta_k\| = \epsilon \|F(X_k, \theta_k)\| \leq L\epsilon(\|\theta_k\| + 1), \quad \forall k \geq 0. \tag{16}$$

The preceding inequality implies for all $k \geq \tau$:

$$\begin{aligned}
E[\|\theta_{k+1} - \theta_k\|^2 | X_{k-\tau}, \theta_{k-\tau}] &\leq L^2\epsilon^2 E[(\|\theta_k\| + 1)^2 | X_{k-\tau}, \theta_{k-\tau}] \\
&\leq L^2\epsilon^2 E[(\|\theta_k - \theta^*\| + \|\theta^*\| + 1)^2 | X_{k-\tau}, \theta_{k-\tau}] \\
&\leq 2L^2\epsilon^2 \left[E[\|\theta_k - \theta^*\|^2 | X_{k-\tau}, \theta_{k-\tau}] + (\|\theta^*\| + 1)^2 \right],
\end{aligned}$$

where the last line follows from $(x + y)^2 \leq 2(x^2 + y^2)$ for all $x, y \in \mathbb{R}$. □

Lemma B.2. *The following inequalities hold for all $k \geq \tau$:*

$$\|\theta_k - \theta_{k-\tau}\| \leq 2L\epsilon\tau(\|\theta_{k-\tau}\| + 1),$$
$$\|\theta_k - \theta_{k-\tau}\| \leq 4L\epsilon\tau(\|\theta_k\| + 1).$$

Proof of Lemma B.2. For any $k \geq \tau$, we first upper bound $\|\theta_t\|$ for $t \in [k-\tau, k]$. Indeed, by (16) we have

$$\|\theta_{t+1} - \theta_t\| \leq L\epsilon(\|\theta_t\| + 1), \tag{17}$$

which by the triangular inequality gives

$$(\|\theta_{t+1}\| + 1) \leq (L\epsilon + 1)(\|\theta_t\| + 1).$$

Recursively using the preceding inequality gives

$$\|\theta_t\| + 1 \leq (L\epsilon + 1)^{t-k+\tau}(\|\theta_{k-\tau}\| + 1), \quad \forall\, k - \tau \leq t \leq k.$$

Therefore,

$$\begin{aligned}
\|\theta_k - \theta_{k-\tau}\| &\leq \sum_{t=k-\tau}^{k-1} \|\theta_{t+1} - \theta_t\| \\
&\leq L\epsilon \sum_{t=k-\tau}^{k-1} (\|\theta_t\| + 1) \\
&\leq L\epsilon(\|\theta_{k-\tau}\| + 1) \sum_{t=k-\tau}^{k-1} (L\epsilon + 1)^{t-k+\tau} \\
&= \left[(L\epsilon + 1)^{\tau} - 1\right](\|\theta_{k-\tau}\| + 1) \\
&\leq (e^{L\epsilon\tau} - 1)(\|\theta_{k-\tau}\| + 1) \\
&\leq 2L\epsilon\tau(\|\theta_{k-\tau}\| + 1),
\end{aligned}$$

where in the last two inequalities we used $1 + x \leq e^x \leq 1 + 2x$ for all $x \in [0, 1/4]$. It follows from the preceding inequality that

$$\|\theta_k - \theta_{k-\tau}\| \leq 2L\epsilon\tau(\|\theta_{k-\tau}\| + 1) \leq 2L\epsilon\tau(\|\theta_k - \theta_{k-\tau}\| + \|\theta_k\| + 1),$$

which implies

$$(1 - 2L\epsilon\tau)\|\theta_k - \theta_{k-\tau}\| \leq 2L\epsilon\tau(\|\theta_k\| + 1).$$

Since $L\epsilon\tau \leq 1/4$, we have

$$\|\theta_k - \theta_{k-\tau}\| \leq 4L\epsilon\tau(\|\theta_k\| + 1).$$

□

Lemma B.3. *The following inequality holds for any $k \geq \tau$:*

$$\begin{aligned}
&E[(\theta_k - \theta^*)^\intercal(F(X_k, \theta_k) - \bar{F}(\theta_k))|X_{k-\tau}, \theta_{k-\tau}] \\
&\leq 56L^2\epsilon\tau\left[E[\|\theta_k - \theta^*\|^2|X_{k-\tau}, \theta_{k-\tau}] + (\|\theta^*\| + 1)^2\right].
\end{aligned}$$

Proof of Lemma B.3. We begin by considering the l.h.s. of the desired inequality as follows

$$\begin{aligned}
&E[(\theta_k - \theta^*)^\intercal(F(X_k, \theta_k) - \bar{F}(\theta_k))|X_{k-\tau}, \theta_{k-\tau}] \\
&= \underbrace{E[(\theta_k - \theta_{k-\tau})^\intercal(F(X_k, \theta_k) - \bar{F}(\theta_k))|X_{k-\tau}, \theta_{k-\tau}]}_{(T_1)} \\
&\quad + \underbrace{E[(\theta_{k-\tau} - \theta^*)^\intercal(F(X_k, \theta_k) - \bar{F}(\theta_k))|X_{k-\tau}, \theta_{k-\tau}]}_{(T_2)}.
\end{aligned}$$

First, we analyze the term (T_1) using Cauchy-Schwartz inequality and Lemma B.2:

$$\begin{aligned}(T_1) =&E[(\theta_k-\theta_{k-\tau})^\intercal(F(X_k,\theta_k)-\bar{F}(\theta_k))|X_{k-\tau},\theta_{k-\tau}]\\ \leq&E[\|\theta_k-\theta_{k-\tau}\|\|F(X_k,\theta_k)-\bar{F}(\theta_k)\||X_{k-\tau},\theta_{k-\tau}]\\ \leq&E[\|\theta_k-\theta_{k-\tau}\|(\|F(X_k,\theta_k)\|+\|\bar{F}(\theta_k)\|)|X_{k-\tau},\theta_{k-\tau}]\\ \leq&2LE[\|\theta_k-\theta_{k-\tau}\|(\|\theta_k\|+1)|X_{k-\tau},\theta_{k-\tau}]\\ \leq&8L^2\epsilon\tau E[(\|\theta_k\|+1)^2|X_{k-\tau},\theta_{k-\tau}]\\ \leq&8L^2\epsilon\tau E[(\|\theta_k-\theta^*\|+\|\theta^*\|+1)^2|X_{k-\tau},\theta_{k-\tau}]\\ \leq&16L^2\epsilon\tau\left[E[\|\theta_k-\theta^*\|^2|X_{k-\tau},\theta_{k-\tau}]+(\|\theta^*\|+1)^2\right]. \end{aligned}\tag{18}$$

Next we consider the term (T_2). Using the Lipschitz continuity of $F(x,\theta)$ and $\bar{F}(\theta)$, we have

$$\begin{aligned}&|(\theta_{k-\tau}-\theta^*)^\intercal(F(X_k,\theta_k)-\bar{F}(\theta_k))-(\theta_{k-\tau}-\theta^*)^\intercal(F(X_k,\theta_{k-\tau})-\bar{F}(\theta_{k-\tau}))|\\ \leq&|(\theta_{k-\tau}-\theta^*)^\intercal(F(X_k,\theta_k)-F(X_k,\theta_{k-\tau}))|+|(\theta_{k-\tau}-\theta^*)^\intercal(\bar{F}(\theta_k)-\bar{F}(\theta_{k-\tau}))|\\ \leq&\|\theta_{k-\tau}-\theta^*\|(\|F(X_k,\theta_k)-F(X_k,\theta_{k-\tau})\|+\|\bar{F}(\theta_k)-\bar{F}(\theta_{k-\tau})\|)\\ \leq&2L\|\theta_{k-\tau}-\theta^*\|\|\theta_k-\theta_{k-\tau}\|,\end{aligned}$$

which gives

$$\begin{aligned}(T_2) =&E[(\theta_{k-\tau}-\theta^*)^\intercal(F(X_k,\theta_k)-\bar{F}(\theta_k))|X_{k-\tau},\theta_{k-\tau}]\\ \leq&E[(\theta_{k-\tau}-\theta^*)^\intercal(F(X_k,\theta_{k-\tau})-\bar{F}(\theta_{k-\tau}))|X_{k-\tau},\theta_{k-\tau}]\\ &+2L\|\theta_{k-\tau}-\theta^*\|E[\|\theta_k-\theta_{k-\tau}\||X_{k-\tau},\theta_{k-\tau}]\\ =&(\theta_{k-\tau}-\theta^*)^\intercal(E[F(X_k,\theta_{k-\tau})|X_{k-\tau},\theta_{k-\tau}]-\bar{F}(\theta_{k-\tau}))\\ &+2L\|\theta_{k-\tau}-\theta^*\|E[\|\theta_k-\theta_{k-\tau}\||X_{k-\tau},\theta_{k-\tau}].\end{aligned}\tag{19}$$

On the one hand, Lemma 3.1 implies

$$\begin{aligned}&(\theta_{k-\tau}-\theta^*)^\intercal(E[F(X_k,\theta_{k-\tau})|X_{k-\tau},\theta_{k-\tau}]-\bar{F}(\theta_{k-\tau}))\\ \leq&\|\theta_{k-\tau}-\theta^*\|\|E[F(X_k,\theta_{k-\tau})|X_{k-\tau},\theta_{k-\tau}]-\bar{F}(\theta_{k-\tau})\|\\ \leq&\epsilon\|\theta_{k-\tau}-\theta^*\|(\|\theta_{k-\tau}\|+1).\end{aligned}$$

On the other hand, Lemma B.2 gives

$$\|\theta_{k-\tau}-\theta^*\|E[\|\theta_k-\theta_{k-\tau}\||X_{k-\tau},\theta_{k-\tau}]\leq 2L\epsilon\tau\|\theta_{k-\tau}-\theta^*\|(\|\theta_{k-\tau}\|+1).$$

Using the preceding two relations into Eq. (19) gives

$$\begin{aligned}(T_2)&\leq\epsilon\|\theta_{k-\tau}-\theta^*\|(\|\theta_{k-\tau}\|+1)+4L^2\epsilon\tau\|\theta_{k-\tau}-\theta^*\|(\|\theta_{k-\tau}\|+1)\\ &\leq 5L^2\epsilon\tau\|\theta_{k-\tau}-\theta^*\|(\|\theta_{k-\tau}\|+1)\\ &\leq 5L^2\epsilon\tau E[(\|\theta_{k-\tau}-\theta_k\|+\|\theta_k-\theta^*\|)\times\\ &\quad(\|\theta_{k-\tau}-\theta_k\|+\|\theta_k-\theta^*\|+\|\theta^*\|+1)|X_{k-\tau},\theta_{k-\tau}]\\ &\leq 5L^2\epsilon\tau E[(\|\theta_k\|+1+\|\theta_k-\theta^*\|)(\|\theta_k\|+\|\theta_k-\theta^*\|+\|\theta^*\|+2)|X_{k-\tau},\theta_{k-\tau}] \qquad (20)\\ &\leq 20L^2\epsilon\tau E[(\|\theta_k-\theta^*\|+\|\theta^*\|+1)^2|X_{k-\tau},\theta_{k-\tau}]\\ &\leq 40L^2\epsilon\tau\left[E[\|\theta_k-\theta^*\|^2|X_{k-\tau},\theta_{k-\tau}]+(\|\theta^*\|+1)^2\right], \qquad (21)\end{aligned}$$

where (20) follows from Lemma B.2 and our assumption that $L\epsilon\tau\leq\frac{1}{4}$. Combine the upper bounds for (T_1) and (T_2), and we obtain

$$\begin{aligned}&E[(\theta_k-\theta^*)^\intercal(F(X_k,\theta_k)-\bar{F}(\theta_k))|X_{k-\tau},\theta_{k-\tau}]\\ =&(T_1)+(T_2)\\ \leq&56L^2\epsilon\tau\left[E[\|\theta_k-\theta^*\|^2|X_{k-\tau},\theta_{k-\tau}]+(\|\theta^*\|+1)^2\right].\end{aligned}$$

□

Lemma B.4. *The following inequality holds for all $k\geq\tau$:*

$$E[\|\theta_{k+1}-\theta^*\|^2]\leq(1-\alpha\epsilon)E[\|\theta_k-\theta^*\|^2]+114L^2\epsilon^2\tau(\|\theta^*\|+1)^2.$$

Proof of Lemma B.4. For any $k \geq \tau$, we have

$$
\begin{aligned}
&E[\|\theta_{k+1} - \theta^*\|^2 - \|\theta_k - \theta^*\|^2 | X_{k-\tau}, \theta_{k-\tau}] \\
=&E[(\theta_{k+1} - \theta_k)^\intercal (\theta_{k+1} + \theta_k - 2\theta^*) | X_{k-\tau}, \theta_{k-\tau}] \\
=&E[(\theta_{k+1} - \theta_k)^\intercal (\theta_{k+1} - \theta_k + 2\theta_k - 2\theta^*) | X_{k-\tau}, \theta_{k-\tau}] \\
=&E[\|\theta_{k+1} - \theta_k\|^2 | X_{k-\tau}, \theta_{k-\tau}] + 2E[(\theta_k - \theta^*)^\intercal (\theta_{k+1} - \theta_k) | X_{k-\tau}, \theta_{k-\tau}] \\
=&E[\|\theta_{k+1} - \theta_k\|^2 | X_{k-\tau}, \theta_{k-\tau}] + 2\epsilon E[(\theta_k - \theta^*)^\intercal F(X_k, \theta_k) | X_{k-\tau}, \theta_{k-\tau}] \\
=&E[\|\theta_{k+1} - \theta_k\|^2 | X_{k-\tau}, \theta_{k-\tau}] + 2\epsilon E[(\theta_k - \theta^*)^\intercal (F(X_k, \theta_k) - \bar{F}(\theta_k)) | X_{k-\tau}, \theta_{k-\tau}] \\
&+ 2\epsilon E[(\theta_k - \theta^*)^\intercal (\bar{F}(\theta_k) - \bar{F}(\theta^*)) | X_{k-\tau}, \theta_{k-\tau}] \\
\leq& 2L^2\epsilon^2 \left[E[\|\theta_k - \theta^*\|^2 | X_{k-\tau}, \theta_{k-\tau}] + (\|\theta^*\| + 1)^2\right] \qquad (22) \\
&+ 112L^2\epsilon^2\tau \left[E[\|\theta_k - \theta^*\|^2 | X_{k-\tau}, \theta_{k-\tau}] + (\|\theta^*\| + 1)^2\right] \qquad (23) \\
&- 2\alpha\epsilon E[\|\theta_k - \theta^*\|^2 | X_{k-\tau}, \theta_{k-\tau}] \qquad (24) \\
\leq& (114L^2\epsilon\tau - 2\alpha)\epsilon E[\|\theta_k - \theta^*\|^2 | X_{k-\tau}, \theta_{k-\tau}] + 114L^2\epsilon^2\tau(\|\theta^*\| + 1)^2 \\
\leq& -\alpha\epsilon E[\|\theta_k - \theta^*\|^2 | X_{k-\tau}, \theta_{k-\tau}] + 114L^2\epsilon^2\tau(\|\theta^*\| + 1)^2, \qquad (25)
\end{aligned}
$$

where (22) follows from Lemma B.1, (23) follows from Lemma B.3, and (24) follows from Assumption 3.3. In the last inequality we used the assumption that $\epsilon\tau \leq \min(1/4L, \alpha/114L^2)$. Taking expectation with respect to $X_{k-\tau}$ and $\theta_{k-\tau}$ on both side of Eq. (25) gives the desired result. □

With the help of Lemmas B.1 – B.4, we next prove Theorem 3.1.

Proof of Theorem 3.1. Let $z_k = E[\|\theta_k - \theta^*\|^2]$, Lemma B.4 gives

$$z_{k+1} \leq az_k + b,$$

where $a = 1 - \alpha\epsilon \in (0, 1)$, and $b = 114L^2\epsilon^2\tau(\|\theta^*\| + 1)^2$. Note that the preceding Eq. can be equivalently written as

$$z_{k+1} - \frac{b}{1-a} \leq a\left(z_k - \frac{b}{1-a}\right),$$

which implies that $\{z_k - b/(1-a)\}$ is upper bounded by a geometric sequence with initial value $z_\tau - b/(1-a)$ and common ratio a. Therefore, we have for all $k \geq \tau$:

$$z_k \leq a^{k-\tau}\left(z_\tau - \frac{b}{1-a}\right) + \frac{b}{1-a} \leq a^{k-\tau}z_k + \frac{b}{1-a}.$$

Plugging in the definition of z_k, a, and b, we obtain

$$E[\|\theta_k - \theta^*\|^2] \leq (1 - \alpha\epsilon)^{k-\tau}E[\|\theta_\tau - \theta^*\|^2] + \frac{114L^2(\|\theta^*\| + 1)^2}{\alpha}\epsilon\tau.$$

The last thing to do is to control $E[\|\theta_\tau - \theta^*\|^2]$. By Lemma B.2 and our assumption that $\epsilon\tau \leq 1/4L$, we have

$$E[\|\theta_\tau - \theta^*\|^2] \leq E[(\|\theta_\tau - \theta_0\| + \|\theta_0 - \theta^*\|)^2] \leq (\|\theta_0\| + \|\theta_0 - \theta^*\| + 1)^2.$$

Combining the preceding two inequalities gives the desired finite-time error bound for SA (5):

$$E[\|\theta_k - \theta^*\|^2] \leq \beta_1(1 - \alpha\epsilon)^{k-\tau} + \frac{\beta_2}{\alpha}\epsilon\tau,$$

where $\beta_1 = (\|\theta_0\| + \|\theta_0 - \theta^*\| + 1)^2$, and $\beta_2 = 114L^2(\|\theta^*\| + 1)^2$. □

C Proof of Theorem 3.2

We will verify each Assumption in Section 3.1 as a lemma in the following, the resulting finite-time bound (12) then follows from Theorem 3.1.

Lemma C.1. *The function $F(x,\theta)$ defined in (10) is globally Lipschitz continuous w.r.t. θ uniformly in x, and $L = \gamma + 1 + r_{\max}$ is a valid Lipschitz constant.*

Proof of Lemma C.1. Since $\|\phi(s,a)\| \le 1$ for all state-action pairs, we have for any $\theta_1, \theta_2 \in \mathbb{R}^d$ and $x \in \mathcal{X}$:

$$\begin{aligned}
&\|F(x,\theta_1) - F(x,\theta_2)\| \\
=&\|\phi(s,a)(\mathcal{R}(s,a) + \gamma \max_{a_1\in\mathcal{A}} \phi(s',a_1)^\intercal\theta_1 - \phi(s,a)^\intercal\theta_1) \\
&- \phi(s,a)(\mathcal{R}(s,a) + \gamma \max_{a_2\in\mathcal{A}} \phi(s',a_2)^\intercal\theta_2 - \phi(s,a)^\intercal\theta_2)\| \\
\le&\gamma\|\phi(s,a)(\max_{a_1\in\mathcal{A}} \phi(s',a_1)^\intercal\theta_1 - \max_{a_2\in\mathcal{A}} \phi(s',a_2)^\intercal\theta_2)\| + \|\phi(s,a)\phi(s,a)^\intercal(\theta_1-\theta_2)\| \\
\le&\gamma|\max_{a_1\in\mathcal{A}} \phi(s',a_1)^\intercal\theta_1 - \max_{a_2\in\mathcal{A}} \phi(s',a_2)^\intercal\theta_2| + \|\theta_1-\theta_2\|.
\end{aligned}$$

To control $|\max_{a_1\in\mathcal{A}} \phi(s',a_1)^\intercal\theta_1 - \max_{a_2\in\mathcal{A}} \phi(s',a_2)^\intercal\theta_2|$, let $\bar{a} \in \arg\max_{a_2\in\mathcal{A}} \phi(s',a_2)^\intercal\theta_2$, we have

$$\begin{aligned}
\max_{a_1\in\mathcal{A}} \phi(s',a_1)^\intercal\theta_1 - \max_{a_2\in\mathcal{A}} \phi(s',a_2)^\intercal\theta_2 &= \max_{a_1\in\mathcal{A}} \phi(s',a_1)^\intercal\theta_1 - \phi(s',\bar{a})^\intercal\theta_2 \\
&\ge \phi(s',\bar{a})^\intercal(\theta_1-\theta_2) \\
&\ge \min_{a'\in\mathcal{A}} \phi(s',a')^\intercal(\theta_1-\theta_2).
\end{aligned}$$

Similarly, we also have

$$\max_{a_1\in\mathcal{A}} \phi(s',a_1)^\intercal\theta_1 - \max_{a_2\in\mathcal{A}} \phi(s',a_2)^\intercal\theta_2 \le \max_{a'\in\mathcal{A}} \phi(s',a')^\intercal(\theta_1-\theta_2).$$

Therefore,

$$\begin{aligned}
|\max_{a_1\in\mathcal{A}} \phi(s',a_1)^\intercal\theta_1 - \max_{a_2\in\mathcal{A}} \phi(s',a_1)^\intercal\theta_2| &\le \max(|\min_{a'\in\mathcal{A}} \phi(s',a')^\intercal(\theta_1-\theta_2)|, |\max_{a'\in\mathcal{A}} \phi(s',a')^\intercal(\theta_1-\theta_2)|) \\
&= \max_{a'\in\mathcal{A}} |\phi(s',a')^\intercal(\theta_1-\theta_2)| \\
&\le \max_{a'\in\mathcal{A}} \|\phi(s',a')\|\|\theta_1-\theta_2\| \\
&\le \|\theta_1-\theta_2\|.
\end{aligned}$$

It follows that

$$\begin{aligned}
\|F(x,\theta_1) - F(x,\theta_2)\| &\le \gamma|\max_{a_1\in\mathcal{A}} \phi(s',a_1)^\intercal\theta_1 - \max_{a_2\in\mathcal{A}} \phi(s',a_2)^\intercal\theta_2| + \|\theta_1-\theta_2\| \\
&\le (\gamma+1)\|\theta_1-\theta_2\|, \quad \forall\theta_1,\theta_2\in\mathbb{R}^d, \forall x\in\mathcal{X}.
\end{aligned}$$

Note that

$$\|F(x,0)\| = \|\phi(s,a)\mathcal{R}(s,a)\| \le r_{\max}, \quad \forall x \in \mathcal{X}.$$

It is clear that $L = \gamma + 1 + r_{\max}$ can be served as a Lipschitz constant for $F(x,\theta)$, and $L \ge \max_{x\in\mathcal{X}} \|F(x,0)\|$. □

Lemma C.2. *The Markov chain $\{X_k = (S_k, A_k, S_{k+1})\}$ is irreducible and aperiodic.*

Proof of Lemma C.2. Let $p^n(s,s')$ be the probability of the transition from s to s' in n steps following policy π. Consider two arbitrary states $x_1 = (s_1,a_1,s_1'), x_2 = (s_2,a_2,s_2') \in \mathcal{X}$. Since $\{S_k\}$ is irreducible, there exists $n > 0$ such that $p^n(s_1', s_2) > 0$. Hence we have

$$p^{n+1}(x_1,x_2) = p^n(s_1',s_2)\pi(a_2|s_2)P_{a_2}(s_2,s_2') > 0.$$

It follows that $\{X_k\}$ is irreducible. To show $\{X_k\}$ is aperiodic, assume for a contradiction that $\{X_k\}$ is periodic with period $T \ge 2$. Since $\{X_k\}$ is irreducible, every state in $\mathcal{X}$ has the same period. Therefore, for any $x = (s,a,s') \in \mathcal{X}$,

$$p^n(x,x) = 0 \text{ for all } n \text{ not divisible by } T.$$

However, notice that for any n not divisible by T, we have

$$\begin{aligned} p^n(s',s') &= \sum_{s\in\mathcal{S}} p^{n-1}(s',s)p(s,s') \\ &= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} p^{n-1}(s',s)\pi(a|s)P_a(s,s') \\ &= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} p^{n}((s,a,s'),(s,a,s')) \qquad (26) \\ &= \sum_{s\in\mathcal{S}}\sum_{a\in\mathcal{A}} 0 \\ &= 0. \qquad (27) \end{aligned}$$

To see (26), since $\{S_k\}$ is a Markov chain, we have

$$\begin{aligned} p^n((s,a,s'),(s,a,s')) &= P(S_n=s, A_n=a, S_{n+1}=s'|S_0=s, A_0=a, S_1=s') \\ &= P(S_n=s, A_n=a, S_{n+1}=s'|S_1=s') \\ &= p^{n-1}(s',s)\pi(a|s)P_a(s,s'). \end{aligned}$$

Therefore, (27) shows that the period of s' is at least T, which is a contradiction to the fact that $\{S_k\}$ being aperiodic. □

Lemma C.3. *The following inequality holds for all $\theta \in \mathbb{R}^d$:*

$$(\theta-\theta^*)^\intercal(\bar{F}(\theta)-\bar{F}(\theta^*)) \le -\frac{\alpha}{2}\|\theta-\theta^*\|^2,$$

where α is given in (11).

Proof of Lemma C.3. The existence of a solution to $\bar{F}(\theta)=0$ is also assumed in Theorem 3.2 condition (b), it is enough to show the drift.

$$\begin{aligned} &(\theta-\theta^*)^\intercal(\bar{F}(\theta)-\bar{F}(\theta^*)) \\ =&(\theta-\theta^*)^\intercal(\gamma E_\mu[\phi(s,a)(\max_{a_1\in\mathcal{A}}\phi(s',a_1)^\intercal\theta - \max_{a_2\in\mathcal{A}}\phi(s',a_2)^\intercal\theta^*)] - E_\mu[\phi(s,a)\phi(s,a)^\intercal](\theta-\theta^*)) \\ \le&\gamma\sqrt{E_\mu[(\phi(s,a)^\intercal(\theta-\theta^*))^2]}\sqrt{E_\mu[(\max_{a_1\in\mathcal{A}}\phi(s',a_1)^\intercal\theta - \max_{a_2\in\mathcal{A}}\phi(s',a_2)^\intercal\theta^*)^2]} \\ &- E_\mu[(\phi(s,a)^\intercal(\theta-\theta^*))^2]. \end{aligned}$$

Note that we have shown in Lemma B.4 that

$$|\max_{a_1\in\mathcal{A}}\phi(s',a_1)^\intercal\theta - \max_{a_2\in\mathcal{A}}\phi(s',a_2)^\intercal\theta^*| \le \max_{a'\in\mathcal{A}}|\phi(s',a')^\intercal(\theta-\theta^*)|,$$

which gives

$$\begin{aligned} (\max_{a_1\in\mathcal{A}}\phi(s',a_1)^\intercal\theta - \max_{a_2\in\mathcal{A}}\phi(s',a_2)^\intercal\theta^*)^2 &\le (\max_{a'\in\mathcal{A}}|\phi(s',a')^\intercal(\theta-\theta^*)|)^2 \\ &= \max_{a'\in\mathcal{A}}(\phi(s',a')^\intercal(\theta-\theta^*))^2. \end{aligned}$$

Therefore, we have

$$
\begin{aligned}
&(\theta-\theta^*)^\intercal(\bar{F}(\theta)-\bar{F}(\theta^*)) \\
\leq&\sqrt{E_\mu[(\phi(s,a)^\intercal(\theta-\theta^*))^2]}\sqrt{\gamma^2 E_\mu[\max_{a'\in\mathcal{A}}(\phi(s',a')^\intercal(\theta-\theta^*))^2]} \\
&-E_\mu[(\phi(s,a)^\intercal(\theta-\theta^*))^2] \\
=&\sqrt{E_\mu[(\phi(s,a)^\intercal(\theta-\theta^*))^2]}\times \\
&\times\left(\sqrt{\gamma^2 E_\mu[\max_{a'\in\mathcal{A}}(\phi(s',a')^\intercal(\theta-\theta^*))^2]}-\sqrt{E_\mu[(\phi(s,a)^\intercal(\theta-\theta^*))^2]}\right) \\
=&\sqrt{E_\mu[(\phi(s,a)^\intercal(\theta-\theta^*))^2]}\times \\
&\times\left(\frac{\gamma^2 E_\mu[\max_{a'\in\mathcal{A}}(\phi(s',a')^\intercal(\theta-\theta^*))^2]-E_\mu[(\phi(s,a)^\intercal(\theta-\theta^*))^2]}{\sqrt{\gamma^2 E_\mu[\max_{a'\in\mathcal{A}}(\phi(s',a')^\intercal(\theta-\theta^*))^2]}+\sqrt{E_\mu[(\phi(s,a)^\intercal(\theta-\theta^*))^2]}}\right) \\
\leq&\frac{-\alpha\sqrt{E_\mu[(\phi(s,a)^\intercal(\theta-\theta^*))^2]}\|\theta-\theta^*\|^2}{\sqrt{\gamma^2 E_\mu[\max_{a'\in\mathcal{A}}(\phi(s',a')^\intercal(\theta-\theta^*))^2]}+\sqrt{E_\mu[(\phi(s,a)^\intercal(\theta-\theta^*))^2]}} \\
=&-\alpha\|\theta-\theta^*\|^2\Bigg/\left(\frac{\gamma\sqrt{E_\mu[\max_{a'\in\mathcal{A}}(\phi(s',a')^\intercal(\theta-\theta^*))^2]}}{\sqrt{E_\mu[(\phi(s,a)^\intercal(\theta-\theta^*))^2]}}+1\right) \\
\leq&-\frac{\alpha}{2}\|\theta-\theta^*\|^2.
\end{aligned}
$$

□

D Numerical Experiments

We first present one way to compute $\delta(\pi)$ for an MDP with a sampling policy π when the underlying model is known. Before that, the following definitions are needed.

Definition D.1. *Define the feature matrix $\Phi\in\mathbb{R}^{nm\times d}$ as*

$$
\Phi=\begin{bmatrix} | & & | \\ \phi_1 & \cdots & \phi_d \\ | & & | \end{bmatrix}=\begin{bmatrix} — & \phi(s_1,a_1)^\intercal & — \\ \cdots & \cdots & \cdots \\ — & \phi(s_n,a_m)^\intercal & — \end{bmatrix}. \tag{28}
$$

Definition D.2. *Let $D_{\mu,\pi}\in\mathbb{R}^{mn\times mn}$ be a diagonal matrix with diagonal entries $\{\mu(s)\pi(a|s)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$, and let $\Sigma_{\mu,\pi}:=\Phi^\intercal D_{\mu,\pi}\Phi\in\mathbb{R}^{d\times d}$, where $\Phi\in\mathbb{R}^{mn\times d}$ is the feature matrix given in (28).*

Definition D.3. *Let $\mathcal{B}=\mathcal{A}^n\subseteq\mathbb{R}^n$ be the set of all deterministic policies.*

Definition D.4. *Let $D_\mu\in\mathbb{R}^{n\times n}$ be a diagonal matrix with diagonal entries $\{\mu(s)\}_{s\in\mathcal{S}}$, and let $\Sigma_{\mu,b}:=\Phi_b^\intercal D_\mu\Phi_b\in\mathbb{R}^{d\times d}$, where $\Phi_b=[\phi(s_1,b_1),\phi(s_2,b_2),\cdots,\phi(s_n,b_n)]^\intercal\in\mathbb{R}^{n\times d}$ for any $b\in\mathcal{B}$.*

Lemma D.1. *Suppose that $\mu(s)\pi(a|s)>0$ for all $(s,a)\in\mathcal{S}\times\mathcal{A}$. Let $\delta(\pi)$ be defined in (14) and let $\lambda_{\max}(M)$ be the largest eigenvalue of a positive semi-definite matrix M. Then we have*

$$
\delta(\pi)=\min_{b\in\mathcal{B}}\left[1/\lambda_{\max}(\Sigma_{\mu,\pi}^{-1/2}\Sigma_{\mu,b}\Sigma_{\mu,\pi}^{-1/2})\right].
$$

Proof of Lemma D.1. Recall our definition for $\delta(\pi)$:

$$
\delta(\pi)=\min_{\|\theta\|=1}\frac{\sum_{s\in\mathcal{S}}\mu(s)\sum_{a\in\mathcal{A}}\pi(a|s)(\phi(s,a)^\intercal\theta)^2}{\sum_{s\in\mathcal{S}}\mu(s)\max_{a\in\mathcal{A}}(\phi(s,a)^\intercal\theta)^2}. \tag{29}
$$

Let $f(\theta)$ be the numerator, we have

$$
f(\theta)=\sum_{s\in\mathcal{S}}\mu(s)\sum_{a\in\mathcal{A}}\pi(a|s)(\phi(s,a)^\intercal\theta)^2=\theta^\intercal\Phi^\intercal D_{\mu,\pi}\Phi\theta=\theta^\intercal\Sigma_{\mu,\pi}\theta.
$$

Since the diagonal entries of $D_{\mu,\pi}$ are all positive, and Φ is full column rank, $\Sigma_{\mu,\pi}$ is symmetric and positive definite. To represent the denominator of (29) in a similar form, let

$$g(\theta, b) = \sum_{i=1}^{n} \mu(s_i)(\phi(s_i, b_i)^\intercal \theta)^2 = \theta^\intercal \Phi_b^\intercal D_\mu \Phi_b \theta = \theta^\intercal \Sigma_{\mu,b} \theta, \quad \text{where } b \in \mathcal{B}.$$

Since the columns of Φ_b can be dependent, $\Sigma_{\mu,b}$ is in general only symmetric and positive semi-definite. With the definition of $f(\theta)$ and $g(\theta, b)$, $\delta(\pi)$ can be represented as

$$\delta(\pi) = \min_{\theta \neq 0} \frac{f(\theta)}{\max_{b \in \mathcal{B}} g(\theta, b)} = \min_{\theta \neq 0} \min_{b \in \mathcal{B}} \frac{f(\theta)}{g(\theta, b)} = \min_{b \in \mathcal{B}} \min_{\theta \neq 0} \frac{f(\theta)}{g(\theta, b)}.$$

Now since $\Sigma_{\mu,\pi}$ is positive definite, $\Sigma_{\mu,\pi}^{1/2}$ and $\Sigma_{\mu,\pi}^{-1/2}$ are both well-defined and positive definite, we have

$$\begin{aligned}
\min_{\theta \neq 0} \frac{f(\theta)}{g(\theta, b)} &= \left[\max_{\theta \neq 0} \frac{g(\theta, b)}{f(\theta)} \right]^{-1} = \left[\max_{\theta \neq 0} \frac{\theta^\intercal \Sigma_{\mu,b} \theta}{\theta^\intercal \Sigma_{\mu,\pi} \theta} \right]^{-1} = \left[\left(\max_{x \neq 0} \frac{\|\Sigma_{\mu,b}^{1/2} \Sigma_{\mu,\pi}^{-1/2} x\|}{\|x\|} \right)^2 \right]^{-1} \\
&= \frac{1}{\lambda_{\max}(\Sigma_{\mu,\pi}^{-1/2} \Sigma_{\mu,b} \Sigma_{\mu,\pi}^{-1/2})},
\end{aligned}$$

where the function $\lambda_{\max}(\cdot)$ returns the largest eigenvalue. It follows that

$$\delta(\pi) = \min_{b \in \mathcal{B}} \left[\frac{1}{\lambda_{\max}(\Sigma_{\mu,\pi}^{-1/2} \Sigma_{\mu,b} \Sigma_{\mu,\pi}^{-1/2})} \right].$$

□

Our numerical experiments in section 3.3 adopt the MDP model of the classical divergent example of Q-learning with linear function approximation introduced in [1]. Consider the infinite-horizon seven-state, two-action MDP shown in Figure 3. The *dashed* action takes the system to one of the six upper states with equal probability, whereas the *solid* action takes the system to the seventh state with probability one. The sampling policy π selects the *dashed* and *solid* actions with equal probability. The reward is zero on all transitions.

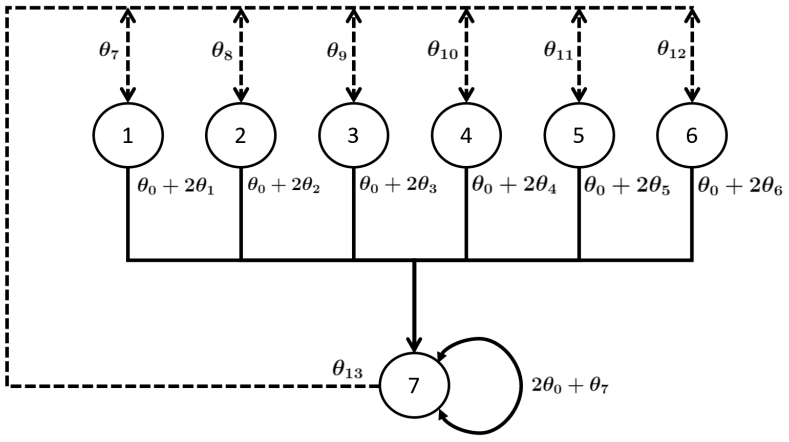

Figure 3: Baird's counterexample [1].

Consider estimating the Q-function under the linear parameterization indicated by the expression showing along each arrow in Figure 3. For example, the estimated value of state 1 taking the *solid* action is $\theta_0 + 2\theta_1$, where the subscript corresponds to the component of the overall weight vector $\theta \in \mathbb{R}^{14}$. It is easy to check that the feature matrix Φ is full column rank in this example.