Automatica 146 (2022) 110623

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Finite-sample analysis of nonlinear stochastic approximation with applications in reinforcement learning☆

Zaiwei Chen [a,*], Sheng Zhang [a], Thinh T. Doan [b], John-Paul Clarke [c], Siva Theja Maguluri [a]

[a] Georgia Institute of Technology, Atlanta, GA 30332, USA
[b] Virginia Tech, Blacksburg, VA 24061, USA
[c] The University of Texas at Austin, Austin, TX 78712, USA

ARTICLE INFO

Article history:
Received 21 March 2021
Received in revised form 18 June 2022
Accepted 5 August 2022
Available online 28 September 2022

Keywords:
Markovian stochastic approximation
Finite-sample analysis
Reinforcement learning
Q-learning
Linear function approximation

ABSTRACT

Motivated by applications in reinforcement learning (RL), we study a nonlinear stochastic approximation (SA) algorithm under Markovian noise, and establish its finite-sample convergence bounds under various stepsizes. Specifically, we show that when using constant stepsize (i.e., $\alpha_k \equiv \alpha$), the algorithm achieves exponential fast convergence to a neighborhood (with radius $O(\alpha \log(1/\alpha))$) around the desired limit point. When using diminishing stepsizes with appropriate decay rate, the algorithm converges with rate $O(\log(k)/k)$. Our proof is based on Lyapunov drift arguments, and to handle the Markovian noise, we exploit the fast mixing of the underlying Markov chain. To demonstrate the generality of our theoretical results on Markovian SA, we use it to derive the finite-sample bounds of the popular Q-learning algorithm with linear function approximation, under a condition on the behavior policy. Importantly, we do not need to make the assumption that the samples are i.i.d., and do not require an artificial projection step in the algorithm. Numerical simulations corroborate our theoretical results.

1. Introduction

Large-scale optimization and machine learning problems are often solved using stochastic approximation (SA) methods (i.e., iterative algorithms in the presence of noise). For example, in optimization, the stochastic gradient descent (SGD) algorithm is commonly used to find an optimal solution of a target objective function (Bottou, Curtis, & Nocedal, 2018; Lan, 2020). In reinforcement learning (RL), Q-learning and TD-learning are popular algorithms used to solve the Bellman equations (Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 2018).

The behavior of SA algorithms is highly dependent on the nature of the associated noise (e.g., i.i.d., martingale difference, or Markovian). In robust optimization problems as considered in Duchi, Agarwal, Johansson, and Jordan (2012) where the data is generated by an auto-regressive process, the corresponding SGD algorithm naturally involves Markovian noise. In RL, algorithms such as Q-learning, TD-learning, and actor–critic use sample trajectories generated from a Markov decision process (MDP) to carry out the update, and hence can also be modeled as Markovian SA algorithms.

The asymptotic convergence of SA algorithms with Markovian noise has been studied extensively in the literature (Benveniste, Métivier, & Priouret, 2012; Bertsekas & Tsitsiklis, 1996; Borkar, 2009). Beyond asymptotic convergence, it is of more practical interest to study finite-sample guarantees, i.e., to provide performance guarantees on the output of SA algorithms after performing a finite number of iterations. More formally, suppose we perform k iterations of an SA algorithm and denote the output by θ_k. Then the goal of finite-sample analysis is to understand how the quantity $\mathbb{E}[\|\theta_k - \theta^*\|^2]$ decay as a function of k, where θ^* is the desired limit point, and $\|\cdot\|$ is a suitable norm. This leads to our main contributions in the following.

Finite-Sample Analysis for Nonlinear Markovian SA. We establish finite-sample convergence guarantees for nonlinear SA with Markovian noise for using various stepsizes, where we do not require an artificial projection step in the algorithm. The results state that Markovian SA algorithms enjoy exponential convergence rate to a neighborhood around the desired limit when using constant stepsize, and an $O(\log(k)/k)$ convergence rate when using appropriate diminishing stepsizes. We prove the results by applying a suitable Lyapunov function on the stochastic iterates, and show that in expectation it produces a negative drift.

☆ The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Mattia Zorzi under the direction of Editor Alessandro Chiuso.

* Corresponding author.
E-mail addresses: zchen458@gatech.edu (Z. Chen), shengzhang@gatech.edu (S. Zhang), thinhdoan@vt.edu (T.T. Doan), johnpaul@gatech.edu (J.-P. Clarke), siva.theja@gatech.edu (S.T. Maguluri).

https://doi.org/10.1016/j.automatica.2022.110623
0005-1098/

To handle the Markovian noise, we exploit the geometric mixing of the underlying Markov chain.

Finite-Sample Analysis of Q-Learning with Linear Function Approximation. To demonstrate the effectiveness of our SA results, we use them to establish for the first time finite-sample bounds for Q-learning under linear function approximation. Since the algorithm does not necessarily converge (Baird, 1995), we use our SA results to provide a sufficient condition which guarantees the convergence. In addition, we verify the sufficiency of our proposed condition and the resulting convergence rates via numerical experiments based on a well-known divergent counter-example of Q-learning from Baird (1995). Specifically, we demonstrate that if our condition is satisfied, the algorithm converges, and the rates match with our theoretical results.

1.1. Related literature

Stochastic Approximation. The SA method, originally proposed in Robbins and Monro (1951), is an iterative method for solving root-finding problems with incomplete information. The asymptotic behavior of SA algorithms is captured by its associated ordinary differential equation (ODE), which leads to the popular ODE approach for analyzing SA algorithms Benveniste et al. (2012) and Kushner and Clark (2012). Specifically, given certain assumptions, it was shown in Borkar (2009) and Ljung (1977) that the SA algorithm converges almost surely as long as the corresponding ODE is stable. The ODE approach was extended to more general cases in Benaim (1996), Karmakar and Bhatnagar (2021) and Yaji and Bhatnagar (2019), where the ODE lacks stability, or has multiple equilibrium points. The convergence of various SA algorithms such as SA with Markovian noise and multiple time-scale SA was studied in Bhatnagar and Borkar (1997, 1998), Karmakar and Bhatnagar (2021) and Ramaswamy and Bhatnagar (2018), respectively. While the results presented there were very general, they study SA algorithms in the asymptotic regime. In this paper, we perform finite-sample analysis, which is different in flavor and provides stronger finite-sample convergence guarantees.

For linear SA algorithms, finite-sample mean-square bounds were established under either i.i.d. sampling or Markovian sampling in Bhandari, Russo, and Singal (2018) and Srikant and Ying (2019). Concentration results were established in Dalal, Szörényi, Thoppe, and Mannor (2018) and Thoppe and Borkar (2019). For non-linear SA algorithms, finite-sample bounds in general are only derived in a special form of SA, namely SGD (Bottou et al., 2018; Lan, 2020; Moulines & Bach, 2011). Moreover, unlike i.i.d. sampling, in the case of Markovian sampling, an artificial projection (onto a ball) is introduced in the algorithm to ensure that the iterates are bounded (Duchi et al., 2012).

Q-Learning (with Linear Function Approximation). Q-learning (Watkins & Dayan, 1992) is perhaps one of the most popular algorithms for solving RL problems (Bertsekas & Tsitsiklis, 1996; Sutton & Barto, 2018). The asymptotic convergence and finite-sample guarantees of Q-learning were studied in Borkar and Meyn (2000), Jaakkola, Jordan, and Singh (1994), Tsitsiklis (1994) and Beck and Srikant (2012), Even-Dar and Mansour (2003) and Kearns and Singh (1998), respectively.

A major limitation with Q-learning is that it becomes computationally intractable when the size of the state–action space is large. One way to overcome this difficulty is to use function approximation. In this work, we consider Q-learning under linear function approximation, which can be modeled as a nonlinear Markovian SA algorithm (Melo, Meyn, & Ribeiro, 2008). However, as shown by the counter-example in Baird (1995), Q-learning with linear function approximation does not necessarily converge. Therefore, additional assumptions were imposed in Melo et al. (2008) to ensure the asymptotic convergence. Under a similar condition, we establish the finite-sample bounds by exploiting some natural properties of Q-learning (such as Lipschitz continuity), and the fast mixing of finite-state Markov chains. The mixing time argument for dealing with Markovian noise was inspired by Bertsekas and Tsitsiklis (1996, Section 4.4) and Srikant and Ying (2019), where linear SA algorithms were studied. Importantly, our approach does not require a projection step in the algorithm (Bhandari et al., 2018), which is impractical in RL since one needs to know the problem parameters to pick the projection set so that the desired limiting solution lies in it.

2. Nonlinear SA with Markovian noise

Consider the problem of solving for θ^* in the equation

$$\bar{F}(\theta) = \mathbb{E}_{\mu_X}[F(X,\theta)] = 0, \tag{1}$$

where $X \in \mathscr{X} \subseteq \mathbb{R}^{n_X}$ is a random vector with distribution μ_X, and the function $F : \mathscr{X} \times \mathbb{R}^d \mapsto \mathbb{R}^d$ is a general nonlinear operator. When the distribution μ_X is unknown, Eq. (1) cannot be solved analytically. Therefore, we consider solving the equation using the SA method. With initialization $\theta_0 \in \mathbb{R}^d$, the estimate θ_k of θ^* is updated according to

$$\theta_{k+1} = \theta_k + \alpha_k(F(X_k,\theta_k) + w_k), \tag{2}$$

where $\{X_k\}$ (taking values in $\mathscr{X}$) is a uniformly ergodic Markov chain with unique stationary distribution μ_X, $\{w_k\}$ represents the additive martingale difference noise that possibly depends on $\{\theta_k\}$, and $\{\alpha_k\}$ is the stepsize sequence. To better understand Algorithm (2), consider the special case where $F(x,\theta) = -\nabla J(\theta) + x$ for some cost function $J(\cdot)$, Algorithm (2) reduces to the popular SGD algorithm for minimizing $J(\cdot)$.

The behavior of Algorithm (2) is closely related to the trajectory of the ODE

$$\dot{\theta}(t) = \bar{F}(\theta(t)). \tag{3}$$

A popular approach to analyze an ODE is to construct a Lyapunov function and study the time-derivative of the Lyapunov function along the trajectory of the ODE. Inspired by the Lyapunov technique for ODE stability analysis, in this paper, we directly study SA algorithm (2) using a Lyapunov approach. See Fazlyab, Ribeiro, Morari, and Preciado (2017), Franca, Robinson, and Vidal (2018), Hu, Seiler, and Rantzer (2017), Hu and Syed (2019) and Romero and Benosman (2020) for more details on using Lyapunov functions to study the behavior of iterative algorithms. Since Algorithm (2) is a *discrete* and *stochastic* counterpart of ODE (3), a major challenge is to handle the error caused by the discretization and the noise. We begin by stating our assumptions to study Algorithm (2). Let $\|\cdot\|$ be the ℓ_2-norm for vectors and the induced 2-norm for matrices.

Assumption 1. There exists a constant $L_1 > 0$ s.t. (1) $\|F(x,\theta_1) - F(x,\theta_2)\| \le L_1\|\theta_1 - \theta_2\|$ for all θ_1, θ_2, and x, and (2) $\|F(x,\mathbf{0})\| \le L_1$ for all x.

Assumption 1 states that the operator $F(x,\theta)$ is L_1-Lipschitz continuous with respect to θ uniformly in x. In the special case where $F(x,\theta)$ is a linear function of θ as considered in Bhandari et al. (2018) and Srikant and Ying (2019), i.e., $F(x,\theta) = A(x)\theta + b(x)$, Assumption 1 is satisfied when $\sup_{x\in\mathscr{X}} \|A(x)\| < \infty$ and $\sup_{x\in\mathscr{X}} \|b(x)\| < \infty$. In our setting, although $F(x,\theta)$ is a nonlinear function of θ, Assumption 1 implies that the growth rate of both $\|F(x,\theta)\|$ and $\|\bar{F}(\theta)\|$ can at most be affine in terms of $\|\theta\|$. To

see this, under Assumption 1, we have by triangle inequality and Jensen's inequality that

$$\|F(x,\theta)\| \le L_1\|\theta\| + \|F(x,\mathbf{0})\| \le L_1(\|\theta\|+1), \tag{4}$$

$$\|\bar{F}(\theta)\| \le \mathbb{E}_{\mu_X}[\|F(X,\theta)\|] \le L_1(\|\theta\|+1). \tag{5}$$

These properties for $F(x,\theta)$ and $\bar{F}(\theta)$ essentially let us establish the finite-sample bounds akin to the case where $F(x,\theta)$ is a linear function of θ.

Assumption 2. The target equation $\bar{F}(\theta) = 0$ has a unique solution θ^*, and there exists $c_0 > 0$ s.t. $(\theta-\theta^*)^\top \bar{F}(\theta) \le -c_0\|\theta-\theta^*\|^2$ for all $\theta \in \mathbb{R}^d$.

In the SGD setting (i.e., $F(x,\theta) = -\nabla J(\theta) + x$), Assumption 2 is satisfied when the objective function $J(\cdot)$ is strongly convex. Moreover, Assumption 2 can be viewed as an exponential dissipativeness property of the ODE (3) with a quadratic storage function. In fact, this assumption guarantees that θ^* is the unique exponentially stable equilibrium point of ODE (3). To see this, let $W(\theta) = \|\theta-\theta^*\|^2$ be a candidate Lyapunov function. Then we have

$$\frac{d}{dt}W(\theta(t)) = 2(\theta(t)-\theta^*)^\top \dot{\theta}(t) \le -2c_0 W(\theta(t)), \tag{6}$$

which implies that $W(\theta(t)) \le W(\theta(0))e^{-2c_0 t}$ for all $t \ge 0$. The parameter c_0 is called the *negative drift*, and we see that the larger c_0 is, the faster $\theta(t)$ converges.

Our next assumption is about the noise sequences $\{X_k\}$ and $\{w_k\}$. Let $\mathscr{F}_k$ be the σ-algebra generated by $\{\theta_i, X_i, w_i\}_{0\le i\le k-1} \cup \{\theta_k, X_k\}$, and denote $\|\cdot\|_{\mathrm{TV}}$ as the total variation distance between probability distributions (Levin & Peres, 2017).

Assumption 3. (1) The Markov chain $\{X_k\}$ is uniformly geometrically ergodic with unique stationary distribution μ_X. (2) The sequence $\{w_k\}$ satisfies $\mathbb{E}[w_k \mid \mathscr{F}_k] = 0$ and $\|w_k\| \le L_2(\|\theta_k\|+1)$ for all $k \ge 0$, where $L_2 > 0$ is a constant.

Assumption 3 (1) is made to control the Markovian noise in Algorithm (2), and implies that there exist $C \ge 1$ and $\rho \in (0,1)$ s.t. $\sup_{x\in\mathscr{X}} \|p^k(x,\cdot) - \mu_X(\cdot)\|_{\mathrm{TV}} \le C\rho^k$ for all $k \ge 0$, where $p^k(x,\cdot)$ represents the distribution of X_k given $X_0 = x$. When compared to $\{X_k\}$ being i.i.d., the major difference for $\{X_k\}$ being Markovian is that there is a bias in the update, i.e., $\mathbb{E}[F(X_k,\theta) \mid X_0 = x] \ne \bar{F}(\theta)$. Since Assumption 3 (1) states that the Markov chain $\{X_k\}$ mixes geometrically fast, it enables us to control such bias and to show that it is not strong enough to cause major deviation from the desired direction of the update. In the special case where the state-space $\mathscr{X}$ of the Markov chain $\{X_k\}$ is finite, Assumption 3 (1) is satisfied when the Markov chain $\{X_k\}$ is irreducible and aperiodic (Levin & Peres, 2017, Theorem 4.9). Assumption 3 (2) states that $\{w_k\}$ is a martingale difference sequence, and w_k may depend on θ_k in the sense that $\|w_k\|$ is allowed to scale affinely with respect to $\|\theta_k\|$.

In addition to these assumptions, the choice of the stepsize sequence $\{\alpha_k\}$ is important. In order to state certain conditions on the stepsizes we pick, we need to use the mixing time of the Markov chain $\{X_k\}$ defined in the following.

Definition 1. For any $\delta > 0$, the mixing time of the Markov chain $\{X_k\}$ with precision δ is defined as $t_\delta = \min\{k \ge 0 : \sup_{x\in\mathscr{X}} \|p^k(x,\cdot) - \mu_X(\cdot)\|_{\mathrm{TV}} \le \delta\}$.

Under Assumption 3 (1), we have for any $\delta > 0$ that

$$t_\delta \le \frac{\log(1/\delta) + \log(C/\rho)}{\log(1/\rho)} \le L_3(\log(1/\delta)+1), \tag{7}$$

where $L_3 := \max(1, \frac{\log(C/\rho)}{\log(1/\rho)})$. As a result, we have $\lim_{\delta\to 0}\delta t_\delta = 0$. Analogous to Srikant and Ying (2019), we only require $t_\delta = o(1/\delta)$ to carry out our finite-sample analysis. We assume the stronger geometric mixing property merely for an ease of exposition. We next use t_δ to state our condition on the stepsize sequence $\{\alpha_k\}$. For simplicity of notation, denote $t_k = t_{\alpha_k}$ and $\alpha_{i,j} = \sum_{k=i}^{j}\alpha_k$. Let $L = L_1 + L_2$, and assume wolg. that $L \ge 1$.

Condition 1. *The stepsize sequence $\{\alpha_k\}$ is non-increasing and satisfies $\alpha_0 \in (0,1)$ and $\alpha_{k-t_k,k-1} < \frac{c_0}{130L^2}$ for all $k \ge t_k$.*

The reason we impose Condition 1 on the stepsize sequence is the following. Recall that a key step in deriving the convergence rate of ODE (3) is to establish the negative drift (cf. Eq. (6)). Similarly, when deriving finite-sample bounds for Algorithm (2), there will also be a negative drift term. In addition, there are error terms that arise because of the discretization and the stochastic noise. Using small stepsize helps suppressing these error terms and hence ensures that the negative drift is the dominant term in our analysis.

Suppose we use constant stepsize, i.e., $\alpha_k = \alpha$ for all $k \ge 0$. Since in this case we have $\alpha_{k-t_k,k-1} = \alpha t_\alpha$ and $\lim_{\alpha\to 0}\alpha t_\alpha = 0$, Condition 1 is satisfied when α is small enough. In addition to constant stepsize, consider using polynomially diminishing stepsizes of the form $\alpha_k = \alpha/(k+h)^\xi$. We show in Section 5.2 that Condition 1 is satisfied for any $\alpha > 0$ and $\xi \in (0,1]$, provided that h is appropriately chosen.

2.1. Finite-sample bounds for nonlinear SA

In this section, we present our main results. We begin with the finite-sample bound of Algorithm (2), the proof of which is presented in Section 2.2.

Theorem 1. *Consider $\{\theta_k\}$ of Algorithm (2). Suppose that Assumption 1–3 are satisfied, and $\{\alpha_k\}$ satisfies Condition 1. Let $K = \min\{k : k \ge t_k\}$. Then we have for all $k \ge K$:*

$$\mathbb{E}[\|\theta_k-\theta^*\|^2] \le \beta_1\prod_{j=K}^{k-1}(1-c_0\alpha_j) + \beta_2\sum_{i=K}^{k-1}\hat{\alpha}_i\prod_{j=i+1}^{k-1}(1-c_0\alpha_j),$$

where $\beta_1 = (\|\theta_0\| + \|\theta_0-\theta^*\|+1)^2$, $\beta_2 = 130L^2(\|\theta^*\|+1)^2$, and $\hat{\alpha}_i = \alpha_i\alpha_{i-t_i,i-1}$.

Remark 1. Although the parameter K is defined as $K = \min\{k : k \ge t_k\}$, we indeed have $K = t_K$. To see this, suppose that $K > t_K$. Since both K and t_K are integers, we must have $K-1 \ge t_K \ge t_{K-1}$, where the second inequality follows from the fact that $t_k = t_{\alpha_k}$ is an increasing function of k. This contradict to the definition of K and hence we have $K = t_K$.

On the RHS of the convergence bound, the first term represents the bias due to the initial guess θ_0, and the second term captures the variance due to the noise. Theorem 1 is one of our main contributions in that (1) the function $F(x,\theta)$ is allowed to be nonlinear, (2) it holds when $\{X_k\}$ is a Markov chain instead of being i.i.d., and (3) no modification on Algorithm (2) (e.g., adding a projection step) is needed to establish the results.

After establishing the finite-sample bounds of Algorithm (2) in its general form, we next consider several common choices of stepsizes, and derive the corresponding convergence rates. We begin by presenting the result when using constant stepsize, i.e., $\alpha_k \equiv \alpha$. The proof of the following corollary is presented in Section 5.1.

Corollary 1. *When α is chosen s.t. $\alpha t_\alpha \le \frac{c_0}{130L^2}$, we have $\mathbb{E}[\|\theta_k-\theta^*\|^2] \le \beta_1(1-c_0\alpha)^{k-t_\alpha} + \beta_2\frac{\alpha t_\alpha}{c_0}$ for all $k \ge t_\alpha$.*

We see from Corollary 1 that when using constant stepsize, the bias term converges to zero geometrically fast as the number of iterations increases, while the variance term remains as a constant of size $O(\alpha \log(1/\alpha))$. Since $t_\alpha \le L_3(\log(1/\alpha)+1)$ (cf. Eq. (7)), using constant stepsize efficiently eliminates the bias. However, since the noise is added to the iterates without being progressively suppressed, the variance does not converge to zero as k goes to infinity.

We next consider diminishing stepsizes. Let $\alpha_k = \alpha/(k+h)^\xi$ where $\alpha > 0$, $\xi \in (0,1]$, and h is chosen s.t. Condition 1 is satisfied. The requirement for choosing h and the proof of the following corollary are presented in Section 5.2.

Corollary 2. *Suppose $\alpha_k = \alpha/(k+h)^\xi$, then we have the following finite-sample bounds.*
(1) (a) When $\xi = 1$ and $\alpha < 1/c_0$, we have for all $k \ge K$:
$\mathbb{E}[\|\theta_k-\theta^*\|^2] \le \beta_1\left(\frac{K+h}{k+h}\right)^{c_0\alpha} + \frac{8\beta_2\alpha^2 L_3}{1-c_0\alpha}\frac{\log\left(\frac{k+h}{\alpha}\right)+1}{(k+h)^{c_0\alpha}}$.
(b) When $\xi = 1$ and $\alpha = 1/c_0$, we have for all $k \ge K$: $\mathbb{E}[\|\theta_k-\theta^*\|^2] \le \beta_1\left(\frac{K+h}{k+h}\right) + 8\beta_2\alpha^2 L_3\frac{\log(k+h)[\log\left(\frac{k+h}{\alpha}\right)+1]}{k+h}$.
(c) When $\xi = 1$ and $\alpha > 1/c_0$, we have for all $k \ge K$:
$\mathbb{E}[\|\theta_k-\theta^*\|^2] \le \beta_1\left(\frac{K+h}{k+h}\right)^{c_0\alpha} + \frac{8e\beta_2\alpha^2 L_3}{c_0\alpha-1}\frac{\log\left(\frac{k+h}{\alpha}\right)+1}{k+h}$.
(2) When $\xi \in (0,1)$ and $\alpha > 0$, suppose that $K \ge [2\xi/(c_0\alpha)]^{1/(1-\xi)}$, then we have for all $k \ge K$:
$\mathbb{E}[\|\theta_k-\theta^*\|^2] \le \beta_1 e^{-\frac{c_0\alpha}{1-\xi}\left((k+h)^{1-\xi}-(K+h)^{1-\xi}\right)} + \frac{4\beta_2\alpha L_3}{c_0}\frac{\log\left(\frac{k+h}{\alpha}\right)+1}{(k+h)^\xi}$.

Observe from Corollary 2 (1) that when using $\alpha_k = \alpha/(k+h)$, the constant α must be chosen carefully (i.e., $\alpha > 1/c_0$) to achieve the optimal $\tilde{O}(1/k)$ convergence rate, otherwise the convergence rate is $\tilde{O}(1/k^{c_0\alpha})$, which can be arbitrarily slow. From Corollary 2 (2), we see that when $\xi \in (0,1)$, the convergence rate is $\tilde{O}(1/k^\xi)$, which is sub-optimal, but more robust in the sense that it is independent of α. The above analysis indicates that our choice of stepsizes should depend on how precise our estimate of the negative drift parameter c_0 is. When our estimate of c_0 is accurate, we should use $\alpha_k = \alpha/(k+h)$ with $\alpha > 1/c_0$ so that the convergence rate is the optimal $\tilde{O}(1/k)$. When our understanding to the system model is poor (therefore inaccurate estimate of c_0), we should use $\alpha_k = \alpha/(k+h)^\xi$. In that case, we sacrifice the convergence rate for robustness.

Unlike almost sure convergence, where the usual requirements for stepsizes are $\sum_{k=0}^\infty \alpha_k = \infty$ and $\sum_{k=0}^\infty \alpha_k^2 < \infty$ (Robbins & Monro, 1951) (which correspond to $\xi \in (1/2, 1]$ in our case), we have convergence in the mean-square sense for all $\xi \in (0,1]$. The same phenomenon has been observed in Bhandari et al. (2018), where linear SA was studied.

2.2. Proof of Theorem 1

In this section, we present the proof of Theorem 1. Before going into the details, we first provide some intuition. Recall that the Lyapunov function $W(\theta) = \|\theta-\theta^*\|^2$ can be used to show the stability of ODE (3). To analyze the convergence rate of the iterates $\{\theta_k\}$ generated by Algorithm (2), naturally we want to use the Lyapunov function $W(\cdot)$ on $\{\theta_k\}$ to show something like

$$\mathbb{E}[W(\theta_{k+1})] - \mathbb{E}[W(\theta_k)] \le (-c_0\alpha_k + e_1)\mathbb{E}[W(\theta_k)] + e_2. \tag{8}$$

Note that on an aside, Eq. (8) is a discrete analog of Eq. (6), and so $W(\cdot)$ is a Lyapunov function (Haddad & Chellaboina, 2011). In continuous time, Eq. (6) enables one to determine the rate of convergence of ODE (3). Eq. (8) is the discrete-time equivalent for SA algorithm (2). To make connection to standard control literature, suppose we view e_2 as the input. Then when $e_2 = 0$, Eq. (8) is of the desired form used to prove the asymptotic stability (Sontag, 2008). In our case, due to a non-vanishing e_2, when using constant stepsize we do not have asymptotic convergence but have convergence to a neighborhood around θ^*.

We next proceed to elaborate our plan of proving Theorem 1. On the RHS of Eq. (8), the $-c_0\alpha_k$ term corresponds to the negative drift of the ODE, and the two terms e_1 and e_2 account for the discretization error and the stochastic error in Algorithm (2). The discretization error can be handled using the properties of the function $F(x,\theta)$ (cf. Assumption 1) and properly chosen stepsizes (cf. Condition 1). As for the stochastic error, since Markovian noise naturally produces bias in the update, we show that $\mathbb{E}[F(X_k,\theta) \mid X_0 = x]$ converges to $\bar{F}(\theta)$ (as k increases) fast enough for any θ, where we make use of Assumption 3 (1). Once we show that both error terms are dominated by the drift term, i.e., $e_1 = o(\alpha_k)$ and $e_2 = o(\alpha_k)$, Eq. (8) can be repeatedly used to establish a finite-sample bound of Algorithm (2).

Following from the high level idea stated above, we now prove Theorem 1. To begin, we apply $W(\theta) = \|\theta-\theta^*\|^2$ on the iterates θ_k of Algorithm (2). To utilize the mixing time of the Markov chain $\{X_k\}$, we take expectation conditioning on X_{k-t_k} and θ_{k-t_k}. For simplicity, we use $\mathbb{E}_k[\cdot]$ for $\mathbb{E}[\cdot \mid X_{k-t_k}, \theta_{k-t_k}]$ in the following. Then we have for all $k \ge t_k$:

$$\begin{aligned}
&\mathbb{E}_k[\|\theta_{k+1}-\theta^*\|^2] - \mathbb{E}_k[\|\theta_k-\theta^*\|^2] \\
&= 2\mathbb{E}_k[(\theta_k-\theta^*)^\top(\theta_{k+1}-\theta_k)] + \mathbb{E}_k[\|\theta_{k+1}-\theta_k\|^2] \\
&= \underbrace{2\alpha_k\mathbb{E}_k[(\theta_k-\theta^*)^\top\bar{F}(\theta_k)]}_{(a)} + \underbrace{2\alpha_k\mathbb{E}_k[(\theta_k-\theta^*)^\top w_k]}_{(b)} \\
&\quad + \underbrace{2\alpha_k\mathbb{E}_k[(\theta_k-\theta^*)^\top(F(X_k,\theta_k)-\bar{F}(\theta_k))]}_{(c)} \\
&\quad + \underbrace{\alpha_k^2\mathbb{E}_k[\|F(X_k,\theta_k)+w_k\|^2]}_{(d)},
\end{aligned} \tag{9}$$

where the last line follows by using the update Eq. (2) and by adding and subtracting $\bar{F}(\theta_k)$.

The term (a) corresponds to the negative drift of ODE (3), and we have $(a) \le -2c_0\alpha_k\mathbb{E}_k\left[\|\theta_k-\theta^*\|^2\right]$ under Assumption 2. The term (b) corresponds to the error due to martingale difference noise $\{w_k\}$. Using the tower property of conditional expectation and Assumption 3 (2), we have $(b) = 0$. The term (c) corresponds to the error due to the Markovian noise $\{X_k\}$, and the term (d) arises mainly because of the error due to discretization. What remains to show is that the terms (c) and (d) are dominated by the term (a). We begin by bounding the term (d) in the following lemma, the proof of which is presented in Section 5.3.

Lemma 1. *The following inequality holds for all $k \ge t_k$:*

$$(d) \le 2L^2\alpha_k^2\left[\mathbb{E}_k[\|\theta_k-\theta^*\|^2] + (\|\theta^*\|+1)^2\right].$$

Observe that Lemma 1 implies that $(d) = O(\alpha_k^2) = o(\alpha_k)$. We next consider the term (c). To control it, we need the following two results.

Lemma 2. *For any given $\delta > 0$, the following inequality holds for any x, θ, and $k \ge t_\delta$:*

$$\|\mathbb{E}[F(X_k,\theta) \mid X_0 = x] - \bar{F}(\theta)\| \le 2L_1\delta(\|\theta\|+1).$$

Lemma 2 uses the mixing time to bound the bias (due to Markovian noise) in Algorithm (2). See Section 5.4 for the proof. The next lemma enables us to control the difference between θ_{k_1} and θ_{k_2} when $k_2 - k_1$ is not too large.

Lemma 3. *For any $k_1 < k_2$ satisfying $\alpha_{k_1,k_2-1} \le \frac{1}{4L}$, the following two inequalities hold:*
(1) $\|\theta_{k_2}-\theta_{k_1}\| \le 2L\alpha_{k_1,k_2-1}(\|\theta_{k_1}\|+1)$,
(2) $\|\theta_{k_2}-\theta_{k_1}\| \le 4L\alpha_{k_1,k_2-1}(\|\theta_{k_2}\|+1)$.

The proof of Lemma 3 is presented in Section 5.5. With the help of Lemmas 2 and 3, we are now ready to bound the term (c) in the following lemma. See Section 5.6 for the proof.

Lemma 4. *The following inequality holds for all k s.t. $\alpha_{k-t_k,k-1} \leq \frac{1}{4L}$ (where we recall that $\alpha_{k-t_k,k-1} = \sum_{i=k-t_k}^{k-1} \alpha_i$):*

$$(c) \leq 128L^2\alpha_k\alpha_{k-t_k,k-1}\left[\mathbb{E}_k[\|\theta_k - \theta^*\|^2] + (\|\theta^*\|+1)^2\right].$$

Substituting the upper bounds we obtained for the terms $(a)-(d)$ into Eq. (9), we have the following result, the proof of which is presented in Section 5.7.

Lemma 5. *It holds for all k satisfying $\alpha_{k-t_k,k-1} \leq \frac{1}{4L}$ that:*

$$\begin{aligned}&\mathbb{E}[\|\theta_{k+1} - \theta^*\|^2]\\ &\leq (1 - 2c_0\alpha_k + 130L^2\alpha_k\alpha_{k-t_k,k-1})\mathbb{E}[\|\theta_k - \theta^*\|^2]\\ &\quad + 130L^2\alpha_k\alpha_{k-t_k,k-1}(\|\theta^*\|+1)^2. \end{aligned} \tag{10}$$

Eq. (10) is of the desired recursive form presented in Eq. (8). Therefore, as long as the drift term dominates the error terms, i.e., $2c_0\alpha_k > 130L^2\alpha_k\alpha_{k-t_k,k-1}$, we can repeatedly use Eq. (10) to derive finite-sample error bounds of Algorithm (2). When Condition 1 is satisfied and $k \geq K$ (see Theorem 1 for the definition of K), we have by Eq. (10) that

$$\mathbb{E}[\|\theta_{k+1} - \theta^*\|^2] \leq (1 - c_0\alpha_k)\mathbb{E}[\|\theta_k - \theta^*\|^2] + \beta_2\hat{\alpha}_k,$$

where $\hat{\alpha}_k$ and β_2 are defined in Theorem 1. Repeatedly using the previous inequality starting from K and we obtain

$$\begin{aligned}&\mathbb{E}[\|\theta_k - \theta^*\|^2]\\ &\leq \mathbb{E}[\|\theta_K - \theta^*\|^2]\prod_{j=K}^{k-1}(1 - c_0\alpha_j) + \beta_2\sum_{i=K}^{k-1}\hat{\alpha}_i\prod_{j=i+1}^{k-1}(1 - c_0\alpha_j).\end{aligned}$$

To bound $\mathbb{E}[\|\theta_K - \theta^*\|^2]$, we use Lemma 3 and $\alpha_{K-t_K,K-1} = \alpha_{0,K-1} \leq \frac{1}{4L}$ to obtain

$$\mathbb{E}[\|\theta_K - \theta^*\|^2] \leq \mathbb{E}[(\|\theta_K - \theta_0\| + \|\theta^* - \theta_0\|)^2] \leq \beta_1.$$

The proof is now complete.

3. Applications in reinforcement learning

We begin by describing the underlying model for RL. Consider an infinite horizon discounted MDP $\mathscr{M}$ comprised by a tuple $(\mathscr{S}, \mathscr{A}, p, \mathscr{R}, \gamma)$, where $\mathscr{S} \subseteq \mathbb{R}^{n_S}$ is a compact state-space, $\mathscr{A}$ is a finite action-space, $p : \mathscr{S} \times \mathscr{A} \times \mathscr{S} \mapsto \mathbb{R}_+$ is the transition function s.t. $\int_B p(s,a,s')ds' = \mathbb{P}(S_{k+1} \in B \mid S_k = s, A_k = a)$ where B is a (measurable) subset of $\mathscr{S}$, $\mathscr{R} : \mathscr{S} \times \mathscr{A} \mapsto [0, r_{\max}]$ is the reward function, and $\gamma \in (0,1)$ is the discount factor. The underlying model of the RL problem is essentially an MDP except that the transition function and reward function are unknown to the agent.

The goal of RL is to find a policy for choosing actions based on the state of the environment so that the expected long-term reward is maximized. Formally, define the state–action value function (aka. the Q-function) of a policy π at (s,a) by $Q_\pi(s,a) = \mathbb{E}_\pi[\sum_{k=0}^{\infty}\gamma^k\mathscr{R}(S_k,A_k) \mid S_0 = s, A_0 = a]$, where we use the notation $\mathbb{E}_\pi[\cdot]$ to mean that the actions are chosen according to policy π, i.e., $A_k \sim \pi(\cdot|S_k)$ for all $k \geq 1$. Our goal is to find an optimal policy π^* in the sense that its corresponding Q-function, denote by Q^*, satisfies $Q^*(s,a) \geq Q_\pi(s,a)$ for any (s,a) and π. A fundamental property of the function Q^* is that, if one simply selects actions greedy based on Q^*, then that is an optimal policy. More formally, we have $\{a \mid \pi^*(a|s) > 0\} \subseteq \arg\max_{a\in\mathscr{A}} Q^*(s,a)$ for all state $s \in \mathscr{S}$ (Bertsekas & Tsitsiklis, 1996). Therefore, solving the RL problem reduces to finding the optimal Q-function.

3.1. Q-learning with linear function approximation

The Q-learning algorithm proposed in Watkins and Dayan (1992) is a popular approach for estimating the function Q^*. However, a fundamental limitation of Q-learning is that the algorithm becomes intractable when the number of state–action pairs is large, or even infinite as considered in this work. Therefore, we consider approximating the optimal Q-function from a pre-specified function class parametrized by a finite number of parameters. We next describe the approximation model.

Let $\phi_i : \mathscr{S} \times \mathscr{A} \mapsto \mathbb{R}$, $1 \leq i \leq d$ be a set of basis functions. Denote $\phi(s,a) = [\phi_1(s,a), \ldots, \phi_d(s,a)]^\top$ (which is a column vector). We assume wolg. that the basis functions $\{\phi_i\}_{1\leq i\leq d}$ are linearly independent and are normalized so that $\|\phi(s,a)\| \leq 1$ for all (s,a). This is possible since we work with MDPs with compact state-spaces and finite action-spaces. The sub-space $\mathscr{W}$ spanned by the basis functions $\{\phi_i\}$ can be written as $\mathscr{W} = \{\tilde{Q}_\theta = \sum_{i=1}^d \phi_i\theta_i \mid \theta \in \mathbb{R}^d\}$. We will use $\mathscr{W}$ as our approximating function space, and the goal here is to find θ^* s.t. $\tilde{Q}_{\theta^*}$ best approximates Q^*.

Using the notation above, we now present Q-learning under linear function approximation (Bertsekas & Tsitsiklis, 1996). Let $\{(S_k, A_k)\}$ be a sample trajectory generated by applying some *behavior* policy π to the underlying MDP model. Note that $\{(S_k, A_k)\}$ forms a Markov chain. Then, the parameter θ of the approximation $\tilde{Q}_\theta$ is updated according to:

$$\theta_{k+1} = \theta_k + \alpha_k\phi(S_k,A_k)\Delta(\theta_k, S_k, A_k, S_{k+1}), \tag{11}$$

where $\Delta(\theta,s,a,s') = \mathscr{R}(s,a) + \gamma\max_{a'\in\mathscr{A}}\phi(s',a')^\top\theta - \phi(s,a)^\top\theta$ for all θ and (s,a,s'), and represents the temporal difference. Note that implementing Algorithm (11) requires computing $\max_{a'\in\mathscr{A}}\phi(s',a')^\top\theta_k$. Even when using linear parametrization, $\phi(s,a)^\top\theta$ as a function $a \in \mathscr{A}$ is not necessarily convex. This is the main reason for us to consider MDPs with finite action-spaces because it is in general hard to solve non-convex optimization problems.

Algorithm (11) can be viewed as an SA algorithm for solving the equation

$$\mathbb{E}_{S\sim\mu_S(\cdot),A\sim\pi(\cdot|S),S'\sim p(S,A,\cdot)}[\phi(S,A)\Delta(\theta,S,A,S')] = 0, \tag{12}$$

where μ_S stands for the stationary distribution of the Markov chain $\{S_k\}$ under policy π (provided that it exists and is unique). Under some mild conditions, Eq. (12) is equivalent to a so-called *projected Bellman equation* (Melo et al., 2008).

In general, Eq. (12) may not necessarily admit a solution, see Appendix A for such an example, and the iteration in Eq. (11) may diverge (Baird, 1995). However, it was shown in Melo et al. (2008) that under an assumption on the behavior policy π, θ_k converges to the solution of Eq. (12), denoted by θ^*, almost surely. In this paper, we work with a similar condition, and focus on establishing the finite-sample bounds of Algorithm (11). We next state our assumptions.

Assumption 4. The behavior policy π satisfies $\pi(a|s) > 0$ for all (s,a), and the Markov chain $\{S_k\}$ induced by π is uniformly geometrically ergodic.

Assumption 4 essentially requires that the behavior policy π has enough exploration, and is commonly used in studying value-based RL algorithms (Tsitsiklis & Van Roy, 1997, 1999). Under Assumption 4, the Markov chain $\{S_k\}$ has a unique stationary distribution, which we have denoted by μ_S. In addition, there exist $C' \geq 1$ and $\rho' \in (0,1)$ s.t. $\max_{s\in\mathscr{S}}\|p_\pi^k(s,\cdot) - \mu_S(\cdot)\|_{\text{TV}} \leq C'\rho'^k$ for all $k \geq 0$ (Levin & Peres, 2017), where $p_\pi(\cdot,\cdot)$ denotes the transition function of the Markov chain $\{S_k\}$ induced by π.

Assumption 5. The target equation (12) has a unique solution θ^*, and there exists $\kappa > 0$ s.t. the following inequality holds for all $\theta \in \mathbb{R}^d$:

$$\gamma^2 \mathbb{E}_{\mu_S}[\max_{a\in\mathscr{A}} \tilde{Q}_\theta(S,a)^2] - \mathbb{E}_{\mu_S,\pi}[\tilde{Q}_\theta(S,A)^2] \le -\kappa\|\theta\|^2. \tag{13}$$

We make Assumption 5 and especially Eq. (13) to ensure the stability of Algorithm (11), which is in the same spirit to the conditions proposed in Melo et al. (2008). A detailed discussion about this assumption and comparison to related conditions are presented in Section 3.3.

3.2. Finite-sample convergence guarantees

To apply our SA results, we begin by modeling Algorithm (11) in the form of Algorithm (2). Define $X_k = (S_k, A_k, S_{k+1})$ for all $k \ge 0$. It is clear that $\{X_k\}$ is also a Markov chain, with state-space $\mathscr{X} = \mathscr{S} \times \mathscr{A} \times \mathscr{S}$. Moreover, under Assumption 4, the Markov chain $\{X_k\}$ also has a unique stationary distribution, which we denote by μ_X and is given by $\mu_X(s,a,s') = \mu_S(s)\pi(a|s)p(s,a,s')$ for all $(s,a,s') \in \mathscr{X}$. Define an operator $F : \mathscr{S}\times\mathscr{A}\times\mathscr{S}\times\mathbb{R}^d \mapsto \mathbb{R}^d$ by

$$F(x,\theta) = F(s,a,s',\theta) = \phi(s,a)\Delta(\theta,s,a,s') \tag{14}$$

for all θ and $x = (s,a,s')$. Then Algorithm (11) can be written in the same form as SA algorithm (2) with the additive noise w_k being identically equal to zero. Let $\bar{F}(\theta) = \mathbb{E}_{\mu_X}[F(X,\theta)]$. We see that $\bar{F}(\theta) = 0$ is exactly the target equation (12).

To apply Theorem 1, we first show in the following proposition that Assumption 1, 2, and 3 are satisfied in the context of Q-learning. The proof is presented in Section 6.1.

Proposition 1. *Suppose that Assumptions 4 and 5 are satisfied, then we have the following results: (1) The Markov chain $\{X_k\}$ satisfies $\max_{x\in\mathscr{X}} \|p_\pi^{k+1}(x,\cdot) - \mu_X(\cdot)\|_{TV} \le C'\rho'^k$ for all $k \ge 0$. (2) Let $M = 1 + \gamma + r_{\max}$. Then we have (a) $\|F(x,\theta_1) - F(x,\theta_2)\| \le M\|\theta_1 - \theta_2\|$ for all x, θ_1, and θ_2, and (b) $\|F(x,\mathbf{0})\| \le M$ for all x. (3) The equation $\bar{F}(\theta) = 0$ has a unique solution θ^*, and we have $(\theta-\theta^*)^\top \bar{F}(\theta) \le -\frac{\kappa}{2}\|\theta-\theta^*\|^2$ for all $\theta \in \mathbb{R}^d$.*

Similarly as in Section 2, given $\delta > 0$, we define t_δ as the mixing time of the Markov chain $\{X_k\}$ with precision $\delta > 0$. Observe that Proposition 1 (1) implies that there exists a constant $M_1 = \max(1, \frac{\log(C'/\rho')}{\log(1/\rho')})$ s.t. $t_\delta \le M_1(\log(1/\delta)+1)$ for any $\delta > 0$. This is analogous to Eq. (7) in Section 2.

We next use Theorem 1 to establish the finite-sample bounds of the Q-learning algorithm (11). In the diminishing stepsize regime, we only present case (1) (c) of Corollary 2, which has the best convergence rate. Let $\eta_1 = (\|\theta_0\| + \|\theta_0 - \theta^*\| + 1)^2$ and $\eta_2 = 130M^2(\|\theta^*\|+1)^2$. The following theorem is a direct implication of Theorem 1, hence we omit its proof.

Theorem 2. *Consider $\{\theta_k\}$ of the Q-learning algorithm (11). Suppose that Assumptions 4 and 5 are satisfied, Then we have the following results.*
(1) When $\alpha_k \equiv \alpha$ with α chosen s.t. $\alpha t_\alpha \le \frac{\kappa}{260M^2}$, we have for all $k \ge t_\alpha$:
$\mathbb{E}[\|\theta_k - \theta^*\|^2] \le \eta_1(1-\kappa\alpha/2)^{k-t_\alpha} + 2\eta_2\alpha t_\alpha/\kappa$.
(2) When $\alpha_k = \alpha/(k+h)$, where $\alpha > 2/\kappa$ and h is large enough, there exists $K' > 0$ s.t. we have for all $k \ge K'$:
$\mathbb{E}[\|\theta_k - \theta^*\|^2] \le \eta_1\left(\frac{K'+h}{k+h}\right)^{\frac{\kappa\alpha}{2}} + \frac{16e\eta_2\alpha^2 M_1}{\kappa\alpha-2}\frac{\log\left(\frac{k+h}{\alpha}\right)+1}{k+h}$.

Theorem 2 (1) is qualitatively similar to Corollary 1 in that the iterates of Q-learning converge exponentially fast to a ball centered at θ^*, and the size of the ball is proportional to αt_α. This agrees with results in Bhandari et al. (2018) and Srikant and Ying (2019), where the popular TD-learning under linear function approximation was studied. Theorem 2 (2) suggests that for properly chosen diminishing stepsizes, the optimal convergence rate is roughly $O(\log(k)/k)$. The $\log(k)$ factor is a consequence of performing Markovian sampling of $\{(S_k, A_k)\}$.

3.3. Discussion about Assumption 5 on the Behavior Policy

In this section, we take a closer look at Assumption 5 and especially Eq. (13), which is made for the stability of Q-learning with linear function approximation. For ease of exposition, from now on, we assume that the state–action space of the MDP is finite, i.e., $n := |\mathscr{S}| < \infty$ and $m := |\mathscr{A}| < \infty$. Let $\Phi \in \mathbb{R}^{mn\times d}$ be the feature matrix defined as

$$\Phi = \begin{bmatrix} | & & | \\ \phi_1 & \dots & \phi_d \\ | & & | \end{bmatrix} = \begin{bmatrix} - & \phi(s_1,a_1)^\top & - \\ \dots & \dots & \dots \\ - & \phi(s_n,a_m)^\top & - \end{bmatrix}.$$

First note that Eq. (13) is equivalent to

$$\gamma^2 \mathbb{E}_{\mu_S}[\max_{a\in\mathscr{A}} \tilde{Q}_\theta(S,a)^2] < \mathbb{E}_{\mu_S,\pi}[\tilde{Q}_\theta(S,A)^2] \tag{15}$$

for all nonzero θ. The direction Eq. (13) implying Eq. (15) is trivial. As for the other direction, let

$$\kappa = -\max_{\theta:\|\theta\|=1} \{\gamma^2 \mathbb{E}_{\mu_S}[\max_{a\in\mathscr{A}} \tilde{Q}_\theta(S,a)^2] - \mathbb{E}_{\mu_S,\pi}[\tilde{Q}_\theta(S,A)^2]\}.$$

By Weierstrass extreme value theorem (Rudin et al., 1964), κ is well-defined and strictly positive because it is the maximum of a continuous function over a compact set. This immediately gives Eq. (13).

Similar assumptions on the behavior policy were also proposed in Lee and He (2019) and Melo et al. (2008). Although the exact form of the conditions are different, they all follow the same spirit. That is, with a chosen Lyapunov function, the condition should enable us to show that the corresponding ODE

$$\dot{\theta}(t) = \bar{F}(\theta(t)) \tag{16}$$

of the Q-learning algorithm (11) is globally asymptotically stable (GAS). We next briefly compare our condition to those proposed in Lee and He (2019) and Melo et al. (2008). The condition in Melo et al. (2008) (their Eq. (7)) implies

$$2\gamma^2 \mathbb{E}_{\mu_S}[(\max_{a\in\mathscr{A}} \tilde{Q}_\theta(S,a))^2] < \mathbb{E}_{\mu_S,\pi}[\tilde{Q}_\theta(S,A)^2] \tag{17}$$

for all nonzero θ[1]. The RHS is the same for both Eqs. (15) and (17). On the LHS, Eq. (17) has an additional factor of 2, and the square is outside the max operator. Although they are similar, our condition and the condition proposed in Melo et al. (2008) do not imply each other. As for the condition proposed in Lee and He (2019), while it is not clear if it is less restrictive than ours, it was shown that the condition in Lee and He (2019) implies the condition in Melo et al. (2008) under more restrictive assumptions. However, Lee and He (2019) assumes i.i.d. sampling, and studies only the asymptotic convergence rather than finite-sample error bounds.

We next analyze how the discount factor, the basis vectors $\{\phi_i\}$, and the behavior policy π impact condition (15). In terms of the dependence on the discount factor, it is clear that condition (15) is easier to satisfy for smaller discount factor. This agrees with our numerical simulations provided in Section 3.4. The use of smaller discount factors in RL was also noted in Jiang, Kulesza, Singh, and Lewis (2015), albeit in a completely different context

[1] The factor of 2 appears to be missing in Melo et al. (2008).

of generalization. To see the impact of the basis vectors and the behavior policy, consider the following two examples.

Uni-Dimension Case. Suppose that $d = 1$. That is, there is only one basis vector ϕ_1, and the weight θ is a scalar. Condition (15) reduces to

$$\gamma^2 \mathbb{E}_{\mu_S}[\max_{a\in\mathscr{A}} \phi(S,a)^2] < \mathbb{E}_{\mu_S,\pi}[\phi(S,A)^2]. \tag{18}$$

Define $H^+ = \mathbb{E}_{\mu_S,\pi}[\gamma\phi(S,A)\max_{a'\in\mathscr{A}}\phi(S',a') - \phi(S,A)^2]$, $H^- = \mathbb{E}_{\mu_S,\pi}[\gamma\phi(S,A)\min_{a'\in\mathscr{A}}\phi(S',a') - \phi(S,A)^2]$, and $r_\pi = \mathbb{E}_{\mu_S,\pi}[\phi(S,A)\mathscr{R}(S,A)]$. Then we have the following result. See Section 6.2 for the proof.

Proposition 2. *Eq. (18) implies $H^+ < 0$ and $H^- < 0$, and the following statements regarding the relation between the stability of ODE (16) and the sign of H^+ and H^- hold:*

$$\text{ODE (16) is GAS} \iff \begin{cases} H^+ < 0, H^- < 0, & \text{when } r_\pi = 0, \\ H^+ < 0, H^- \le 0, & \text{when } r_\pi > 0, \\ H^+ \le 0, H^- < 0, & \text{when } r_\pi < 0. \end{cases}$$

Proposition 2 states that Condition (18) implies $H^+, H^- < 0$, which is "almost necessary" for the GAS of ODE (16). Moreover, it is clear from Eq. (18) that when $d = 1$, there always exists a behavior policy π s.t. Eq. (18) is satisfied. For example, an ϵ-greedy policy (for a sufficiently small ϵ) with respect to $\phi(s,a)^2$ is a feasible behavior policy.

Full-Dimension Case. Suppose that $d = mn$, i.e., there is no dimension reduction at all. We want to emphasize that this is not equivalent to tabular Q-learning. Even when Φ is a full-rank square matrix, Q-learning with linear function approximation does not coincide with tabular Q-learning. In fact, the divergent counter-example provided in Baird (1995) belongs to this setting. We show in the following proposition that, in the full-dimension case, condition (15) is feasible in terms of the behavior policy π only when the discount factor γ is sufficiently small. See Section 6.3 for its proof.

Proposition 3. *When $d = mn$ and $\gamma^2 \ge 1/m$, condition (15) is infeasible for any behavior policy π.*

We now compare the results for the two extreme cases, i.e., $d = 1$ and $d = mn$. We see that in the uni-dimensional case, Eq. (15) implies a condition which is almost sufficient and necessary for the GAS of the equilibrium θ^* to ODE (16). Moreover, there always exists a behavior policy π satisfying (15). However, in the full-dimensional case, condition (15) is infeasible in terms of the behavior policy π when $\gamma^2 \ge 1/m$, which can usually happen in practice, especially when the number of actions is large.

3.4. Numerical simulations

In this section, we present numerical experiments to verify the sufficiency of Condition (13), and the convergence rates of Q-learning with linear function approximation. Let $\omega(\pi) = \min_{\{\theta:\|\theta\|=1\}} \mathbb{E}_{\mu_S,\pi}[\tilde{Q}_\theta(S,A)^2]/\mathbb{E}_{\mu_S}[\max_a \tilde{Q}_\theta(S,a)^2]$. Then Condition (13) is equivalent to $\omega(\pi) > \gamma^2$. One way to compute $\omega(\pi)$ is presented in Section 6.4.

In our simulation, we consider the divergent example of Q-learning with linear function approximation introduced in Baird (1995), which is an MDP with 7 states and 2 actions. To demonstrate the effectiveness of Condition (13) for the stability of Q-learning, in our first set of simulations, the reward function is set to zero. Since the reward function is identically zero, Q^* is zero, implying θ^* is zero. We choose the behavior policy π which takes each action with equal probability. In this case, we have $\omega(\pi) \approx 0.5$, giving the threshold for γ to satisfy Eq. (13) being $\omega(\pi)^{1/2} \approx 0.7$. In our simulation, we choose constant stepsize $\alpha = 0.01$, discount factor $\gamma \in \{0.7, 0.9, 0.97\}$, and plot $\|\theta_k\|$ as a function of the number of iterations k in Fig. 1. Here, θ_k converges when $\gamma = 0.7, 0.9$, but diverges when $\gamma = 0.97$. This demonstrates that Condition (13) is sufficient but not necessary for convergence. This also shows that when Eq. (13) is satisfied, the counter-example from Baird (1995) converges.

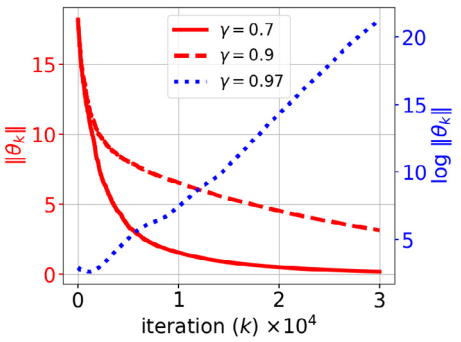

Fig. 1. Convergence of Q-learning with linear function approximation for different discount factors.

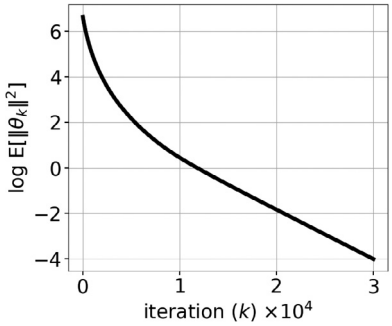

Fig. 2. Exponentially fast convergence of Q-learning with linear function approximation for $\gamma = 0.7$.

To show the exponential convergence rate for using constant stepsize, we consider the convergence of θ_k when $\gamma = 0.7$ given in Fig. 2, where we plot $\log \mathbb{E}[\|\theta_k\|^2]$ as a function of the number of iterations k. In this case, θ_k seems to converge geometrically, which agrees with Theorem 2 (1).

We next numerically verify the convergence rates of Q-learning with linear function approximation for using diminishing stepsizes $\alpha_k = \alpha/(k+h)^\xi$. We use the same MDP model and behavior policy. The only difference is that the reward is no longer set to zero, but is sampled independently from a uniform distribution on [0, 1] for all state–action pairs. The constant κ given in Eq. (13) is estimated by numerical optimization, and the discount factor γ is set to be 0.7 to ensure convergence. In Fig. 3, we plot $\mathbb{E}[\|\theta_k - \theta^*\|^2]$ as a function of k for $\xi \in \{0.4, 0.6, 0.8, 1\}$. In the case where $\xi = 1$, the constant coefficient α is chosen s.t. $\kappa\alpha \ge 2$ in order to achieve the optimal convergence rate. We see that the iterates converge for all $\xi \in (0, 1]$. Moreover, the larger the value of ξ is, the faster θ_k converges.

To further verify the convergence rates, we plot $\log \mathbb{E}[\|\theta_k - \theta^*\|^2]$ as a function of $\log(k)$ in Fig. 4 and look at its asymptotic behavior. We see that the slope is approximately $-\xi$, which agrees with Theorem 2 (2).

In addition to the MDP used in Baird's counter-example (Baird, 1995), numerical simulations corresponding to a larger MDP are presented in Appendix B, and the results are consistent with the theory as well as the outcomes of the simulations in this section.

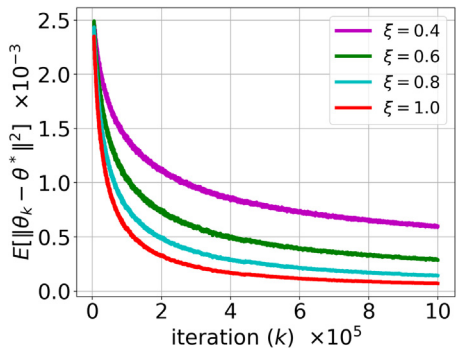

Fig. 3. Convergence for diminishing stepsizes.

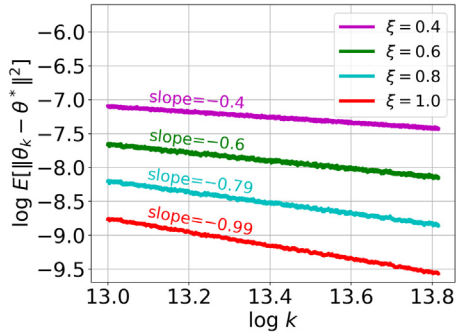

Fig. 4. Asymptotic convergence rate for diminishing stepsizes.

4. Conclusion

In this paper we establish finite-sample convergence guarantees for a general nonlinear SA algorithm with Markovian noise. We adopt a Lyapunov approach and control the error due to Markovian noise by exploiting the fast mixing of the underlying Markov chain. The result is used to derive, for the first time, finite-sample bounds for Q-learning with linear function approximation. Since such an algorithm is known to diverge in general, we study it under a condition on the basis functions, the behavior policy, and the discount factor that ensures stability. Sufficiency of this condition and the rate of convergence of Q-learning are verified numerically in the context of a well-known example.

5. Proof of all technical results in Section 2

5.1. Proof of Corollary 1

When $\alpha_k \equiv \alpha$, since $K = t_\alpha$, we have (1) $\prod_{j=K}^{k-1}(1 - c_0\alpha_j) = (1 - c_0\alpha)^{k-t_\alpha}$, and (2) $\sum_{i=K}^{k-1} \hat{\alpha}_i \prod_{j=i+1}^{k-1}(1 - c_0\alpha_j) = \alpha^2 t_\alpha \sum_{i=t_\alpha}^{k-1}(1 - c_0\alpha)^{k-i-1} \le \frac{\alpha t_\alpha}{c_0}$. This proves the result.

5.2. Proof of Corollary 2

We first verify Condition 1. When $\alpha_k = \alpha/(k+h)^\xi$, using Eq. (7) and we have

$$t_k \le L_3(\log(1/\alpha_k) + 1) = L_3(\xi \log(k+h) + \log(1/\alpha)).$$

It follows that

$$\begin{aligned}\alpha_{k-t_k,k-1} &\le t_k \alpha_{k-t_k} \\ &\le L_3(\log(1/\alpha_k) + 1)\frac{\alpha}{(k - t_k + h)^\xi} \\ &\le \frac{\alpha L_3(\log(1/\alpha_k) + 1)}{(k - L_3(\log(1/\alpha_k) + 1) + h)^\xi} \\ &= \frac{\alpha L_3(\log(1/\alpha_k) + 1)}{(k - L_3(\xi \log(k+h) + \log(1/\alpha)) + 1 + h)^\xi} \\ &= \frac{\alpha_k L_3(\log(1/\alpha_k) + 1)(k+h)^\xi}{(k - L_3(\xi \log(k+h) + \log(1/\alpha)) + 1 + h)^\xi},\end{aligned}$$

where the last line follows from multiplying α_k and dividing $\frac{\alpha}{(k+h)^\xi}$. Therefore, we have $\lim_{(k+h)\to\infty} \frac{\alpha_{k-t_k,k-1}}{\alpha_k L_3(\log(1/\alpha_k)+1)} = 1$, which implies that there exists $\bar{h}_1 = \bar{h}_1(\alpha, \xi) > 0$ such that $\alpha_{k-t_k,k-1} \le 2L_3\alpha_k(\log(1/\alpha_k) + 1)$ for all $k \ge 0$. Since we also have $\lim_{(k+h)\to\infty} \alpha_k(\log(1/\alpha_k) + 1) = 0$, there exists $\bar{h}_2 = \bar{h}_2(\alpha, \xi) > 0$ such that $\alpha_{k-t_k,k-1} \le \frac{c_0}{130L^2}$ for all $k \ge t_k$. Let $\bar{h} = \max(\bar{h}_1, \bar{h}_2)$. Then when $h \ge \bar{h}$, Condition 1 is satisfied for any $k \ge 0$. Moreover, we have in this case $\alpha_{k-t_k,k-1} \le 2L_3\alpha_k(\log(1/\alpha_k) + 1)$ for any $k \ge 0$. This is useful for us to derive the explicit convergence rate in the following.

To prove Corollary 2, we begin by simplifying the result of Theorem 1. For $k \ge K$, we have

$$\begin{aligned}&\mathbb{E}[\|\theta_k - \theta^*\|^2] \\ &\le \beta_1 \prod_{j=K}^{k-1}(1 - c_0\alpha_j) + \beta_2 \sum_{i=K}^{k-1} \alpha_i \alpha_{i-k_i,i-1} \prod_{j=i+1}^{k-1}(1 - c_0\alpha_j) \\ &\le \beta_1 \prod_{j=K}^{k-1}(1 - c_0\alpha_j) \\ &\quad + \beta_2 \sum_{i=K}^{k-1} 2L_3\alpha_i^2(\log(1/\alpha_i) + 1) \prod_{j=i+1}^{k-1}(1 - c_0\alpha_j) \\ &\le \beta_1 \prod_{j=K}^{k-1}(1 - c_0\alpha_j) \\ &\quad + 2L_3\beta_2(\log(1/\alpha_k) + 1) \sum_{i=K}^{k-1} \alpha_i^2 \prod_{j=i+1}^{k-1}(1 - c_0\alpha_j) \\ &\le \underbrace{\beta_1 \prod_{j=K}^{k-1}(1 - c_0\alpha_j)}_{A_1} + 2L_3\beta_2 \left(\log\left(\frac{k+h}{\alpha}\right) + 1\right) \times \\ &\quad \underbrace{\sum_{i=K}^{k-1} \alpha_i^2 \prod_{j=i+1}^{k-1}(1 - c_0\alpha_j)}_{A_2}. \end{aligned} \tag{19}$$

We next bound the terms A_1 and A_2. For A_1, we have

$$\begin{aligned}A_1 &= \prod_{j=K}^{k-1}\left(1 - \frac{c_0\alpha}{(j+h)^\xi}\right) \\ &\le \exp\left(-c_0\alpha \sum_{j=K}^{k-1} \frac{1}{(j+h)^\xi}\right) \\ &\le \exp\left(-c_0\alpha \int_K^k \frac{1}{(x+h)^\xi} dx\right) \\ &\le \begin{cases} \left(\frac{K+h}{k+h}\right)^{c_0\alpha}, & \xi = 1, \\ e^{-\frac{c_0\alpha}{1-\xi}\left((k+h)^{1-\xi} - (K+h)^{1-\xi}\right)}, & \xi \in (0,1). \end{cases}\end{aligned} \tag{20}$$

For A_2, when $\xi = 1$, we have

$$A_2 = \sum_{i=K}^{k-1} \frac{\alpha^2}{(i+h)^2} \prod_{j=i+1}^{k-1}\left(1 - \frac{c_0\alpha}{j+h}\right)$$

$$\leq \sum_{i=K}^{k-1} \frac{\alpha^2}{(i+h)^2}\left(\frac{i+1+h}{k+h}\right)^{c_0\alpha} \qquad \text{(Same to Eq. (20))}$$

$$= \frac{\alpha^2}{(k+h)^{c_0\alpha}} \sum_{i=K}^{k-1} \left(\frac{i+1+h}{i+h}\right)^2 \frac{1}{(i+1+h)^{2-c_0\alpha}}$$

$$= \frac{\alpha^2}{(k+h)^{c_0\alpha}} \sum_{i=K}^{k-1} \left(1+\frac{1}{i+h}\right)^2 \frac{1}{(i+1+h)^{2-c_0\alpha}}$$

$$\leq \frac{4\alpha^2}{(k+h)^{c_0\alpha}} \sum_{i=K}^{k-1} \frac{1}{(i+1+h)^{2-c_0\alpha}}$$

$$\leq \begin{cases} \frac{4\alpha^2}{(1-c_0\alpha)(k+h)^{c_0\alpha}}, & \alpha \in (0, 1/c_0), \\ \frac{4\alpha^2 \log(k+h)}{k+h}, & \alpha = 1/c_0, \\ \frac{4e\alpha^2}{(c_0\alpha-1)(k+h)}, & \alpha \in (1/c_0, \infty). \end{cases} \tag{21}$$

Substituting the upper bounds for the terms A_1 (cf. Eq. (20)) and A_2 (cf. Eq.) into Eq. (19) proves Corollary 2 (1).

Now consider the case where $\xi \in (0, 1)$. Let $\{u_k\}_{k\geq K}$ be a sequence defined as

$$u_{k+1} = \left(1 - \frac{c_0\alpha}{(k+h)^\xi}\right) u_k + \frac{\alpha^2}{(k+h)^{2\xi}}, \quad u_K = 0.$$

It is easy to verify that $u_k = A_2$. We next use induction on u_k to show that

$$u_k \leq \frac{2\alpha}{c_0} \frac{1}{(k+h)^\xi}. \tag{22}$$

Since $u_K = 0 \leq \frac{2\alpha}{c_0}\frac{1}{(K+h)^\xi}$, we have the base case. Now suppose $u_k \leq \frac{2\alpha}{c_0}\frac{1}{(k+h)^\xi}$ for some $k \geq K$. Consider the difference between $\frac{2\alpha}{c_0}\frac{1}{(k+1+h)^\xi}$ and u_{k+1}. We have

$$\frac{2\alpha}{c_0}\frac{1}{(k+1+h)^\xi} - u_{k+1}$$

$$\geq \frac{2\alpha}{c_0}\frac{1}{(k+1+h)^\xi} - \left(1 - \frac{c_0\alpha}{(k+h)^\xi}\right)\frac{2\alpha}{c_0}\frac{1}{(k+h)^\xi} - \frac{\alpha^2}{(k+h)^{2\xi}}$$

$$= \frac{2\alpha}{c_0}\frac{1}{(k+h)^{2\xi}}\left[\frac{c_0\alpha}{2} - (k+h)^\xi\left(1 - \left(\frac{k+h}{k+1+h}\right)^\xi\right)\right]$$

$$\geq \frac{2}{c_0\alpha}\frac{1}{(k+h)^{2\xi}}\left[\frac{c_0\alpha}{2} - \frac{\xi}{(k+h)^{1-\xi}}\right] \tag{23}$$

$$\geq 0, \tag{24}$$

where Eq. (23) follows from

$$\left(\frac{k+h}{k+1+h}\right)^\xi = \left[\left(1+\frac{1}{k+h}\right)^{k+h}\right]^{-\xi/(k+h)}$$
$$\geq e^{-\xi/(k+h)}$$
$$\geq 1 - \frac{\xi}{k+h},$$

and Eq. (24) follows from $k \geq K \geq [2\xi/(c_0\alpha)]^{1/(1-\xi)}$. The induction is now complete.

Substituting the upper bounds for the terms A_1 (cf. Eq. (20)) and A_2 (cf. Eq. (22)) into Eq. (19) proves Corollary 2 (2).

5.3. Proof of Lemma 1

For all $k \geq t_k$, we have

$$\mathbb{E}_k[\|F(X_k,\theta_k) + w_k\|^2]$$

$$\leq \mathbb{E}_k[(\|F(X_k,\theta_k)\| + \|w_k\|)^2] \qquad \text{(triangle inequality)}$$

$$\leq \mathbb{E}_k[(L_1+L_2)^2(\|\theta_k\|+1)^2] \qquad \text{(Eq. (4) and Assumption 3 (2))}$$

$$\leq L^2\mathbb{E}_k[(\|\theta_k - \theta^*\| + \|\theta^*\| + 1)^2]$$
$$(L = L_1 + L_2 \text{ and triangle inequality})$$

$$\leq 2L^2\mathbb{E}_k[\|\theta_k - \theta^*\|^2 + (\|\theta^*\|+1)^2],$$

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

5.4. Proof of Lemma 2

We first recall an equivalent formula of computing the total variation distance (Charalambous, Tzortzis, Loyka, & Charalambous, 2014).

Lemma 6. *Let μ_1 and μ_2 be two probability measures on a probability space $(\Omega, \mathscr{F})$. Let p_1, p_2 be the Radon–Nikodym derivatives of μ_1 and μ_2 w.r.t. some base probability measure ν. Then we have $\|\mu_1 - \mu_2\|_{TV} = \frac{1}{2}\int_\Omega |p_1 - p_2| d\nu$.*

We next proceed to prove Lemma 2. For any given state $x \in \mathscr{X}$, let p_x^k and q_X be the Radon–Nikodym derivatives of the probability measures $p^k(x,\cdot)$ and $\mu_X(\cdot)$ with respect to some base probability measure ν. Use the definition of mixing time (cf. Definition 1) and Lemma 6, and we have for all x, θ, and $k \geq t_\delta$ that

$$\|\mathbb{E}[F(X_k,\theta) \mid X_0 = x] - \bar{F}(\theta)\|$$

$$= \left\| \int_{\mathscr{X}} F(y,\theta) p^k(x, d(y)) - \int_{\mathscr{X}} F(y,\theta)\mu_X(dy) \right\|$$

$$= \left\| \int_{\mathscr{X}} F(y,\theta)(p_x^k - q_X) d\nu \right\|$$

$$\leq \int_{\mathscr{X}} \| F(y,\theta) \| |p_x^k - q_X| d\nu$$

$$\leq L_1(\|\theta\|+1)\int_{\mathscr{X}} |p_x^k - q_X| d\nu \qquad \text{(Eq. (4))}$$

$$= 2L_1(\|\theta\|+1)\|p^k(x,\cdot) - \mu_X(\cdot)\|_{TV} \qquad \text{(Lemma 6)}$$

$$\leq 2L_1(\|\theta\|+1)C\rho^k \qquad \text{(Assumption 3)}$$

$$\leq 2L_1(\|\theta\|+1)\delta \qquad (k \geq t_\delta \text{ and Definition 1})$$

5.5. Proof of Lemma 3

Given $k_1 < k_2$, we first upper bound $\|\theta_t\|$ for any $t \in [k_1, k_2]$. Using Eq. (4) and Assumption 3 (2), we have

$$\|\theta_{t+1}\| - \|\theta_t\| \leq \|\theta_{t+1} - \theta_t\|$$
$$= \alpha_k\|F(X_k,\theta_k) + w_k\|$$
$$\leq (L_1+L_2)\alpha_k(\|\theta\|+1)$$
$$\leq L\alpha_t(\|\theta_t\|+1), \tag{25}$$

which gives $(\|\theta_{t+1}\|+1) \leq (L\alpha_t+1)(\|\theta_t\|+1)$. Recursively applying the previous inequality, then using the fact that $1+x \leq e^x$ for all $x \in \mathbb{R}$, we have for all $t \in [k_1, k_2]$:

$$\|\theta_t\| + 1 \leq \prod_{j=k_1}^{t-1}(L\alpha_j + 1)(\|\theta_{k_1}\| + 1)$$
$$\leq \exp(L\alpha_{k_1,k_2-1})(\|\theta_{k_1}\| + 1).$$

Since $e^x \le 1+2x$ for all $x \in [0, 1/2]$ and $\alpha_{k_1,k_2-1} \le \frac{1}{4L}$, we further obtain

$$\|\theta_t\| + 1 \le (1 + 2L\alpha_{k_1,k_2-1})(\|\theta_{k_1}\| + 1) \le 2(\|\theta_{k_1}\| + 1).$$

It follows from the previous inequality and Eq. (25) that

$$\|\theta_{k_2} - \theta_{k_1}\| \le \sum_{t=k_1}^{k_2-1} \|\theta_{t+1} - \theta_t\| \le 2L\alpha_{k_1,k_2-1}(\|\theta_{k_1}\| + 1).$$

Since $\alpha_{k_1,k_2-1} \le \frac{1}{4L}$ and

$$\begin{aligned}\|\theta_{k_2} - \theta_{k_1}\| &\le 2L\alpha_{k_1,k_2-1}(\|\theta_{k_1}\| + 1)\\ &\le 2L\alpha_{k_1,k_2-1}(\|\theta_{k_2} - \theta_{k_1}\| + \|\theta_{k_2}\| + 1)\\ &\le \frac{1}{2}\|\theta_{k_2} - \theta_{k_1}\| + 2L\alpha_{k_1,k_2-1}(\|\theta_{k_2}\| + 1),\end{aligned}$$

we have by rearranging terms that

$$\|\theta_{k_2} - \theta_{k_1}\| \le 4L\alpha_{k_1,k_2-1}(\|\theta_{k_2}\| + 1).$$

5.6. Proof of Lemma 4

We begin by decomposing the following term on the LHS of the target inequality:

$$\begin{aligned}&\mathbb{E}_k[(\theta_k - \theta^*)^\top(F(X_k,\theta_k) - \bar{F}(\theta_k))]\\ &= \underbrace{\mathbb{E}_k[(\theta_k - \theta_{k-t_k})^\top(F(X_k,\theta_k) - \bar{F}(\theta_k))]}_{(T_1)}\\ &\quad + \underbrace{(T_2}(\theta_{k-t_k} - \theta^*)^\top(\mathbb{E}_k[F(X_k,\theta_{k-t_k})] - \bar{F}(\theta_{k-t_k}))\\ &\quad + \underbrace{\mathbb{E}_k(\theta_{k-t_k} - \theta^*)^\top(F(X_k,\theta_k) - F(X_k,\theta_{k-t_k}))}_{T_3}\\ &\quad + \underbrace{\mathbb{E}_k[(\theta_{k-t_k} - \theta^*)^\top(\bar{F}(\theta_{k-t_k}) - \bar{F}(\theta_k))]}_{(T_4)}.\end{aligned}$$

Consider the term (T_1). Since $\alpha_{k-t_k,k-1} \le \frac{1}{4L}$, Lemma 3 is applicable for $k_1 = k - t_k$ and $k_2 = k$. Therefore, we have:

$$\begin{aligned}(T_1) &= \mathbb{E}_k[(\theta_k - \theta_{k-t_k})^\top(F(X_k,\theta_k) - \bar{F}(\theta_k))]\\ &\le \mathbb{E}_k[\|\theta_k - \theta_{k-t_k}\|\|F(X_k,\theta_k) - \bar{F}(\theta_k)\|]\\ &\le \mathbb{E}_k[\|\theta_k - \theta_{k-t_k}\|(\|F(X_k,\theta_k)\| + \|\bar{F}(\theta_k)\|)]\\ &\le 2L\mathbb{E}_k[\|\theta_k - \theta_{k-t_k}\|(\|\theta_k\| + 1)]\\ &\le 8L^2\alpha_{k-t_k,k-1}\mathbb{E}_k\left[(\|\theta_k\| + 1)^2\right] \qquad \text{(Lemma 3)}\\ &\le 8L^2\alpha_{k-t_k,k-1}\mathbb{E}_k\left[(\|\theta_k - \theta^*\| + \|\theta^*\| + 1)^2\right]\\ &\le 16L^2\alpha_{k-t_k,k-1}(\mathbb{E}_k\left[\|\theta_k - \theta^*\|^2\right] + (\|\theta^*\| + 1)^2). \qquad (26)\end{aligned}$$

Now consider the term (T_2). Since Lemma 2 implies that

$$\|\mathbb{E}_k[F(X_k,\theta_{k-t_k})] - \bar{F}(\theta_{k-t_k})\| \le 2\alpha_k L(\|\theta_{k-t_k}\| + 1),$$

we have by Cauchy–Schwarz inequality that

$$(T_2) \le 2\alpha_k L\mathbb{E}_k[(\|\theta_{k-t_k}\| + 1)\|\theta_{k-t_k} - \theta^*\|].$$

To further control (T_2), using Lemma 3 together with our assumption that $\alpha_{k-t_k,k-1} \le \frac{1}{4L}$, we have

$$\|\theta_k - \theta_{k-t_k}\| \le 4L\alpha_{k-t_k,k-1}(\|\theta_k\| + 1) \le \|\theta_k\| + 1. \qquad (27)$$

Therefore, we have

$$\begin{aligned}&\|\theta_{k-t_k} - \theta^*\|(\|\theta_{k-t_k}\| + 1)\\ &\le (\|\theta_k - \theta_{k-t_k}\| + \|\theta_k - \theta^*\|)\\ &\quad \times (\|\theta_k - \theta_{k-t_k}\| + \|\theta_k - \theta^*\| + \|\theta^*\| + 1)\\ &\le (\|\theta_k\| + \|\theta_k - \theta^*\| + 1)(\|\theta_k\| + \|\theta_k - \theta^*\| + \|\theta^*\| + 2)\\ &\le (2\|\theta_k - \theta^*\| + \|\theta^*\| + 1)(2\|\theta_k - \theta^*\| + 2\|\theta^*\| + 2)\\ &\le 4(\|\theta_k - \theta^*\| + \|\theta^*\| + 1)^2\\ &\le 8[\|\theta_k - \theta^*\|^2 + (\|\theta^*\| + 1)^2].\end{aligned}$$

It follows that

$$(T_2) \le 16\alpha_k L(\mathbb{E}_k[\|\theta_k - \theta^*\|^2] + (\|\theta^*\| + 1)^2). \qquad (28)$$

We next bound the terms (T_3) and (T_4). Using Assumption 1, we have

$$\begin{aligned}&(T_3) + (T_4)\\ &\le 2L\|\theta_{k-t_k} - \theta^*\|\mathbb{E}_k[\|\theta_k - \theta_{k-t_k}\|]\\ &\le 8L^2\alpha_{k-t_k,k-1}\mathbb{E}_k[\|\theta_{k-t_k} - \theta^*\|(\|\theta_k\| + 1)] \qquad \text{(Lemma 3)}\\ &\le 8L^2\alpha_{k-t_k,k-1}\mathbb{E}_k[(\|\theta_k - \theta_{k-t_k}\| + \|\theta_k - \theta^*\|)\times\\ &\quad (\|\theta_k\| + 1)]\\ &\le 8L^2\alpha_{k-t_k,k-1}\mathbb{E}_k[(\|\theta_k\| + \|\theta_k - \theta^*\| + 1)\times\\ &\quad (\|\theta_k - \theta^*\| + \|\theta^*\| + 1)] \qquad \text{(Eq. (27))}\\ &\le 8L^2\alpha_{k-t_k,k-1}\mathbb{E}_k[(2\|\theta_k - \theta^*\| + \|\theta^*\| + 1)\times\\ &\quad (\|\theta_k - \theta^*\| + \|\theta^*\| + 1)]\\ &\le 16L^2\alpha_{k-t_k,k-1}\mathbb{E}_k[(\|\theta_k - \theta^*\| + \|\theta^*\| + 1)^2]\\ &\le 32L^2\alpha_{k-t_k,k-1}\mathbb{E}_k[\|\theta_k - \theta^*\|^2 + (\|\theta^*\| + 1)^2]. \qquad (29)\end{aligned}$$

Finally, combining the upper bounds we derived for the terms $\{(T_i)\}_{1\le i\le 4}$ in Eqs. (26), (28), and (29), we have

$$\begin{aligned}(c) &= 2\alpha_k((T_1) + (T_2) + (T_3) + (T_4))\\ &\le 128L^2\alpha_k\alpha_{k-t_k,k-1}\left[\mathbb{E}_k[\|\theta_k - \theta^*\|^2] + (\|\theta^*\| + 1)^2\right],\end{aligned}$$

where the last line follows from $L \ge 1$ and $\alpha_k \le \alpha_{k-t_k,k-1}$.

5.7. Proof of Lemma 5

Substituting the upper bounds we obtained for the terms (a)–(d) into Eq. (9) and we have

$$\begin{aligned}&\mathbb{E}_k[\|\theta_{k+1} - \theta^*\|^2] - \mathbb{E}_k[\|\theta_k - \theta^*\|^2]\\ &\le (-2c_0\alpha_k + 128L^2\alpha_k\alpha_{k-t_k,k-1} + 2L_1^2\alpha_k^2)\mathbb{E}_k[\|\theta_k - \theta^*\|^2]\\ &\quad + (128L^2\alpha_k\alpha_{k-t_k,k-1} + 2L_1^2\alpha_k^2)(\|\theta^*\| + 1)^2\\ &\le (-2c_0\alpha_k + 130L^2\alpha_k\alpha_{k-t_k,k-1})\mathbb{E}_k[\|\theta_k - \theta^*\|^2]\\ &\quad + 130L^2\alpha_k\alpha_{k-t_k,k-1}(\|\theta^*\| + 1)^2.\end{aligned}$$

The result then follows by taking the total expectation on both sides of the previous inequality.

6. Proof of all technical results in Section 3

6.1. Proof of Proposition 1

(1) For any $x_0 = (s_0, a_0, s_0') \in \mathscr{X}$, where the state space $\mathscr{X}$ of the Markov chain $\{X_k\}$ is given by $= \{(s, a, s') \mid s \in \mathscr{S}, \pi(a|s) > 0, p(s, a, s') > 0\}$, we have

$$\begin{aligned}&\|p_\pi^{k+1}(x_0,\cdot) - \mu_X(\cdot)\|_{\mathrm{TV}}\\ &= \frac{1}{2}\sup_{u:\|u\|_\infty\le 1}\left|\int_{\mathscr{X}} u(x)p_\pi^{k+1}(x_0,dx) - \int_{\mathscr{X}} u(x)\mu_X(dx)\right|.\end{aligned}$$

For simplicity, for any $x = (s, a, s') \in \mathscr{X}$, we denote $v_u(s) = \sum_{a\in\mathscr{A}}\pi(a|s)\left(\int_{\mathscr{S}} u(s, a, s')p(s, a, s')ds'\right)$. Note that we have $\|v_u\|_\infty \le 1$ for all $u(\cdot)$ such that $\|u\|_\infty \le 1$. Using the definition of $v_u(\cdot)$, we have for any $x_0 = (s_0, a_0, s_0') \in \mathscr{X}$ that

$$\|p_\pi^{k+1}(x_0,\cdot) - \mu_X(\cdot)\|_{\mathrm{TV}}$$

$$
\begin{aligned}
&= \frac{1}{2}\sup_{u:\|u\|_\infty\le 1}\left|\int_{\mathscr{X}} u(x)p_\pi^{k+1}(x_0,dx)-\int_{\mathscr{X}} u(x)\mu_X(dx)\right|\\
&= \frac{1}{2}\sup_{u:\|u\|_\infty\le 1}\left|\int_{\mathscr{S}} v_u(s)p_\pi^{k}(s_0',ds)-\int_{\mathscr{S}} v_u(s)\mu_S(ds)\right|\\
&\le \frac{1}{2}\sup_{v:\|v\|_\infty\le 1}\left|\int_{\mathscr{S}} v(s)p_\pi^{k}(s_0',ds)-\int_{\mathscr{S}} v(s)\mu_S(ds)\right|\\
&= \|p_\pi^k(s_0',\cdot)-\mu_S(\cdot)\|_{\mathrm{TV}}\\
&\le C'\rho'^k
\end{aligned}
$$

for all $k \ge 0$, where the last line follows from Assumption 4. Since the previous inequality holds for all $x_0 \in \mathscr{X}$, we have $\max_{x\in\mathscr{X}} \|p_\pi^{k+1}(x,\cdot)-\mu_X(\cdot)\|_{\mathrm{TV}} \le C'\rho'^k$ for all $k \ge 0$.

(2) Using Cauchy–Schwarz inequality, and our assumption that $\|\phi(s,a)\| = 1$ for all state–action pairs, we have for any θ_1, θ_2 and $x = (s,a,s')$:

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

Since

$$
\begin{aligned}
&|\max_{a_1\in\mathscr{A}}\phi(s',a_1)^\top\theta_1-\max_{a_2\in\mathscr{A}}\phi(s',a_1)^\top\theta_2|\\
&\le \max_{a'\in\mathscr{A}}|\phi(s',a')^\top(\theta_1-\theta_2)| \qquad (30)\\
&\le \max_{a'\in\mathscr{A}}\|\phi(s',a')\|\|\theta_1-\theta_2\|\\
&\le \|\theta_1-\theta_2\|,
\end{aligned}
$$

we have for any θ_1, θ_2 and x:

$$\|F(x,\theta_1)-F(x,\theta_2)\| \le (\gamma+1)\|\theta_1-\theta_2\| \le M\|\theta_1-\theta_2\|.$$

Moreover, we have $\|F(x,\mathbf{0})\| = \|\phi(s,a)\mathscr{R}(s,a)\| \le r_{\max}$ for any $x \in \mathscr{X}$.

(3) Using the fact that $\bar{F}(\theta^*) = 0$, we have

$$
\begin{aligned}
&(\theta-\theta^*)^\top(\bar{F}(\theta)-\bar{F}(\theta^*))\\
&= \gamma(\theta-\theta^*)^\top\times\\
&\quad \mathbb{E}_{\mu_S}[\phi(S,A)(\max_{a_1\in\mathscr{A}}\phi(S',a_1)^\top\theta-\max_{a_2\in\mathscr{A}}\phi(S',a_2)^\top\theta^*)]\\
&\quad -\mathbb{E}_{\mu_S}[(\phi(S,A)^\top(\theta-\theta^*))^2]\\
&\le \gamma\mathbb{E}_{\mu_S}[|\phi(S,A)^\top(\theta-\theta^*)|\max_{a'\in\mathscr{A}}|\phi(S',a')^\top(\theta-\theta^*)|]\\
&\quad -\mathbb{E}_{\mu_S}[(\phi(S,A)^\top(\theta-\theta^*))^2] \qquad (31)\\
&\le \gamma\sqrt{\mathbb{E}_{\mu_S}[(\phi(S,A)^\top(\theta-\theta^*))^2]}\\
&\quad \times\sqrt{\mathbb{E}_{\mu_S}[\max_{a\in\mathscr{A}}(\phi(S,a)^\top(\theta-\theta^*))^2]}\\
&\quad -\mathbb{E}_{\mu_S}[(\phi(S,A)^\top(\theta-\theta^*))^2]. \qquad (32)
\end{aligned}
$$

Eq. (31) follows from Eq. (30). Eq. (32) follows from the fact that when $S \sim \mu_S$, we have $S' \sim \mu_S$. For simplicity of notation, denote $A = \sqrt{\mathbb{E}_{\mu_S}[(\phi(S,A)^\top(\theta-\theta^*))^2]}$ and $B = \sqrt{\mathbb{E}_{\mu_S}[\max_{a\in\mathscr{A}}(\phi(S,a)^\top(\theta-\theta^*))^2]}$. Since Assumption 5 gives $\gamma^2B^2-A^2 \le -\kappa\|\theta-\theta^*\|^2$, we have

$$(\theta-\theta^*)^\top\bar{F}(\theta) \le \frac{\gamma^2B^2-A^2}{\gamma B/A+1} \le -\frac{\kappa}{2}\|\theta-\theta^*\|^2.$$

6.2. Proof of Proposition 2

We first show that Eq. (18) implies $H^+ < 0$, and $H^- < 0$. Note that Jensen's inequality implies

$$
\begin{aligned}
&\mathbb{E}_{\mu_S}[\max_{a'\in\mathscr{A}}\phi(S,a')^2]\\
&= \mathbb{E}_{\mu_S}\left\{\max\left[(\max_{a'\in\mathscr{A}}\phi(S,a'))^2,(\min_{a'\in\mathscr{A}}\phi(S,a'))^2\right]\right\}\\
&\ge \max\left\{\mathbb{E}_{\mu_S}[(\max_{a'\in\mathscr{A}}\phi(S,a'))^2],\mathbb{E}_{\mu_S}[(\min_{a'\in\mathscr{A}}\phi(S,a'))^2]\right\}. \qquad (33)
\end{aligned}
$$

Thus, using Eq. (18) and we have

$$
\begin{aligned}
H^+ &= \mathbb{E}_{\mu_S}[\gamma\phi(S,A)\max_{a'\in\mathscr{A}}\phi(S',a')]-\mathbb{E}_{\mu_S}[\phi(S,A)^2]\\
&= \mathbb{E}_{\mu_S}[\gamma\phi(S,A)\max_{a'\in\mathscr{A}}\phi(S',a')]\\
&\quad -\sqrt{\mathbb{E}_{\mu_S}[\phi(S,A)^2]\mathbb{E}_{\mu_S}[\phi(S,A)^2]}\\
&< \mathbb{E}_{\mu_S}[\gamma\phi(S,A)\max_{a'\in\mathscr{A}}\phi(S',a')]\\
&\quad -\gamma\sqrt{\mathbb{E}_{\mu_S}[\max_{a'\in\mathscr{A}}\phi(S,a')^2]\mathbb{E}_{\mu_S}[\phi(S,A)^2]}\\
&\le 0,
\end{aligned}
$$

where the last inequality follows from Cauchy–Schwarz inequality and the fact that S' and S are equal in distribution if $S \sim \mu_S$. Similarly, we also have $H^- < 0$.

We next prove the equivalence stated in Proposition 2. By definition of H^+ and H^-, in uni-dimensional case, ODE (16) can be equivalently written as

$$\dot{\theta}(t) = \begin{cases} H^+\theta(t)+r_\pi, & \theta(t)\ge 0,\\ H^-\theta(t)+r_\pi, & \theta(t)<0.\end{cases}$$

In the case where $r_\pi = 0$, it is easy to see that ODE (16) is globally asymptotically stable if and only if $H^+, H^- < 0$. Now we assume without loss of generality that $r_\pi > 0$. The proof for the other case is entirely similar.

Sufficiency: We first note that $\theta^* = -r_\pi/H^+ > 0$. Let $W(\theta) = \frac{1}{2}(\theta-\theta^*)^2$ be a candidate Lyapunov function. It is clear that $W(\theta) \ge 0$ for all $\theta \in \mathbb{R}$, and $W(\theta) = 0$ if and only if $\theta = \theta^*$. Moreover, we have

$$
\begin{aligned}
\dot{W}(\theta(t)) &= (\theta(t)-\theta^*)\dot{\theta}(t)\\
&= \begin{cases} H^+(\theta(t)-\theta^*)^2, & \theta(t)\ge 0\\ (\theta(t)-\theta^*)(H^-\theta(t)-H^+\theta^*), & \theta(t)<0.\end{cases}
\end{aligned}
$$

It is clear that $\dot{W}(\theta(t)) < 0$ when $\theta(t) \in [0,\theta^*)\cup(\theta^*,\infty)$. For $\theta(t) < 0$, since $\theta(t)-\theta^* < 0$, $H^+\theta^* = -r_\pi < 0$, and $H^-\theta(t) \ge 0$, we must also have $\dot{W}(\theta(t)) < 0$. Therefore, the time derivative of the Lyapunov function $W(\theta)$ along the trajectory of ODE (16) is strictly negative when $\theta(t) \ne \theta^*$. It then follows from the Lyapunov stability theorem (Haddad & Chellaboina, 2011; Khalil & Grizzle, 2002) that θ^* is globally asymptotically stable.

Necessity: We prove by contradiction. Suppose that the equilibrium point θ^* is globally asymptotically stable, but $H^+ \ge 0$ or $H^- > 0$. Suppose that $H^+ \ge 0$. When $\theta(0) > \max(0,\theta^*)$, we have $\dot{\theta}(t) = H^+\theta(t)+r_\pi \ge r_\pi > 0$. It follows that $\theta(t) > \theta(0) > \theta^*$ for all $t \ge 0$, which contradict to the fact that θ^* is a globally asymptotically stable equilibrium point. Suppose that $H^- > 0$. When $\theta(0) < \min(\theta^*, -(1+r_\pi)/H^-)$, we have $\dot{\theta}(t) = H^-\theta(t)+r_\pi \le -1 < 0$. It follows that $\theta(t) < \theta(0) < \theta^*$ for all $t \ge 0$, which also contradict to the fact θ^* being globally asymptotically stable.

6.3. Proof of Proposition 3

When $d = mn$, the feature matrix Φ is a square matrix. Define

$$\Theta_{s,a} = \text{span}\left(\{\phi(s',a') | (s',a') \in \mathscr{S}\times\mathscr{A},\ (s',a') \neq (s,a)\}\right)^{\perp}.$$

Note that $\Theta_{s,a}$ exists for all state–action pairs. Now for a given state–action pair (s,a), let $\theta \neq 0$ be in $\Theta_{s,a}$, Eq. (15) implies $\gamma^2\mu_S(s)(\phi(s,a)^\top\theta)^2 < \mu_S(s)\pi(a|s)(\phi(s,a)^\top\theta)^2$, which further gives $\gamma^2 < \pi(a|s)$. Therefore, by running (s,a) though all state–action pairs, we have $\gamma^2 < \min_{(s,a)\in\mathscr{S}\times\mathscr{A}}\pi(a|s) \leq \frac{1}{m}$. Thus, if $\gamma^2 \geq 1/m$, there is no behavior policy π that satisfies Condition (15).

6.4. Computing $\omega(\pi)$

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

Definition 2. Let $D \in \mathbb{R}^{mn\times mn}$ be a diagonal matrix with diagonal entries $\{\mu_S(s)\pi(a|s)\}_{(s,a)\in\mathscr{S}\times\mathscr{A}}$, and let $\Sigma = \Phi^\top D\Phi \in \mathbb{R}^{d\times d}$, where $\Phi \in \mathbb{R}^{mn\times d}$ is the feature matrix.

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

Definition 4. Let $D_S \in \mathbb{R}^{n\times n}$ be a diagonal matrix with diagonal entries $\{\mu_S(s)\}_{s\in\mathscr{S}}$, and let $\Sigma_b = \Phi_b^\top D_S\Phi_b \in \mathbb{R}^{d\times d}$, where $\Phi_b \in \mathbb{R}^{n\times d}$ $(b \in \mathscr{B})$ is defined by:

$$\Phi_b = \begin{bmatrix} - & \phi(s_1,b)^\top & - \\ \cdots & \cdots & \cdots \\ - & \phi(s_n,b)^\top & - \end{bmatrix}.$$

We now compute $\omega(\pi)$ given in the following lemma. Let $\lambda_{\max}(\cdot)$ return the largest eigenvalue of a positive semi-definite matrix

Lemma 7. $\omega(\pi) = \min_{b\in\mathscr{B}}\left[1/\lambda_{\max}(\Sigma^{-1/2}\Sigma_b\Sigma^{-1/2})\right]$.

Proof of Lemma 7. Recall our definition for $\omega(\pi)$:

$$\omega(\pi) = \min_{\theta\neq 0}\frac{\sum_{s\in\mathscr{S}}\mu_S(s)\sum_{a\in\mathscr{A}}\pi(a|s)(\phi(s,a)^\top\theta)^2}{\sum_{s\in\mathscr{S}}\mu_S(s)\max_{a\in\mathscr{A}}(\phi(s,a)^\top\theta)^2}. \tag{34}$$

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

$$\begin{aligned} f(\theta) &= \sum_{s\in\mathscr{S}}\mu_S(s)\sum_{a\in\mathscr{A}}\pi(a|s)(\phi(s,a)^\top\theta)^2 \\ &= \theta^\top\Phi^\top D\Phi\theta = \theta^\top\Sigma\theta. \end{aligned}$$

Since the diagonal entries of D are all positive, and Φ is full column rank, the matrix Σ is symmetric and positive definite. To represent the denominator of (34) in a similar form, let

$$\begin{aligned} g(\theta,b) &= \sum_s \mu_S(s)(\phi(s,b)^\top\theta)^2 \\ &= \theta^\top\Phi_b^\top D_S\Phi_b\theta = \theta^\top\Sigma_b\theta, \end{aligned}$$

where $b \in \mathscr{B}$. Since the columns of Φ_b can be dependent, the matrix Σ_b is in general only symmetric and positive semi-definite. Using the definition of $f(\theta)$ and $g(\theta,b)$, we can rewrite $\omega(\pi)$ as

$$\begin{aligned} \omega(\pi) &= \min_{\theta\neq 0}\frac{f(\theta)}{\max_{b\in\mathscr{B}} g(\theta,b)} \\ &= \min_{\theta\neq 0}\min_{b\in\mathscr{B}}\frac{f(\theta)}{g(\theta,b)} \\ &= \min_{b\in\mathscr{B}}\min_{\theta\neq 0}\frac{f(\theta)}{g(\theta,b)}. \end{aligned}$$

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

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

It follows that

$$\omega(\pi) = \min_{b\in\mathscr{B}}[1/\lambda_{\max}(\Sigma^{-1/2}\Sigma_b\Sigma^{-1/2})].$$

Acknowledgments

This work was partially supported by NSF Grants EPCN-2144316, CCF-1740776, CMMI-2112533, and awards from Raytheon Technologies and Delta Airlines.

Appendix A. On the existence of solutions to Eq. (12)

In this section, we construct an example to show that Eq. (12) for Q-learning with linear function approximation may not admit a solution. Consider an MDP with states-space $\mathscr{S} = \{1,2\}$, action-space $\mathscr{A} = \{1,2\}$, transition probability matrices $P_1 = [1,0\,;\,1,0]$, $P_2 = [0,1\,;\,0,1]$, reward function

$$R = \begin{bmatrix} R(1,1) \\ R(1,2) \\ R(2,1) \\ R(2,2) \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 2 \\ 4 \end{bmatrix},$$

and a tunable discount factor $\gamma \in (0,1)$. Let the feature matrix be defined by

$$\Phi = \begin{bmatrix} \phi(1,1) \\ \phi(1,2) \\ \phi(2,1) \\ \phi(2,2) \end{bmatrix} = \begin{bmatrix} 1 \\ 2 \\ 2 \\ 4 \end{bmatrix}.$$

We use a uniform behavior policy, i.e., $\pi(1|1) = \pi(2|1) = \pi(1|2) = \pi(2|2) = 0.5$. Then the transition probability matrix P_π under policy π is given by

$$P_\pi = 0.5P_1 + 0.5P_2 = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix},$$

and the unique stationary distribution μ_S of the Markov chain $\{S_k\}$ under policy π is given by $\mu_S(1) = \mu_S(2) = 0.5$.

Consider the target Eq. (12). In this example, after straightforward calculation, Eq. (12) reduces to

$$\theta = \begin{cases} 1 + \gamma\frac{6}{5}\theta, & \theta \geq 0, \\ 1 + \gamma\frac{3}{5}\theta, & \theta < 0, \end{cases}$$

which has no solution when $\gamma \in (5/6, 1)$.

Appendix B. More numerical simulations

To complement the numerical experiments presented in Section 3.4, here we implement the Q-learning with linear function approximation algorithm on a larger MDP. We first introduce our experimental setup and then state our results.

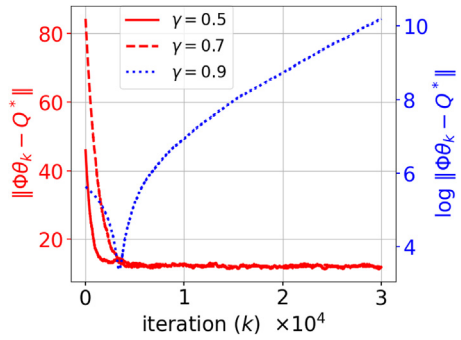

Fig. B.1. Convergence of Q-learning with linear function approximation for discount factor $\gamma \in \{0.5, 0.7, 0.9\}$.

B.1. Setup

We consider an MDP with 100 states and 10 actions, where rewards and transition probabilities are generated as follows:

Rewards. The reward $\mathscr{R}(s, a)$ for each state–action pair (s, a) is drawn from the uniform distribution on $[0, 1]$.

Transition probabilities. For each state–action pair $(s, a) \in \mathscr{S} \times \mathscr{A}$, the probabilities $p(s, a, s')$ of each successor state $s' \in \mathscr{S}$ are chosen as random partitions of the unit interval. That is, 99 numbers are chosen uniformly randomly between 0 and 1, dividing that interval into 100 numbers that sum to one — the probabilities of the 100 successor states.

Moreover, we consider a feature matrix Φ with 100 features (recall that there are total 1000 state–action pairs) for each state–action pair $(s, a) \in \mathscr{S} \times \mathscr{A}$, where each element is drawn from the Bernoulli distribution with success probability $p = 0.5$. We repeat this process until we obtain a full column rank feature matrix Φ. We further normalize the features to ensure $\|\phi(s, a)\| \leq 1$ for all $(s, a) \in \mathscr{S} \times \mathscr{A}$. Furthermore, the behavior policy π is chosen to take each action with equal probability in each state $s \in \mathscr{S}$.

B.2. Results

In our first set of experiments, we choose constant stepsize $\alpha = 0.01$ and discount factor $\gamma \in \{0.5, 0.7, 0.9\}$. In Fig. B.1, we plot $\|\Phi\theta_k - Q^*\|$ as a function of the iteration k, where Q^* associated with each γ is the optimal Q value function computed by the value iteration algorithm. Here, $\Phi\theta_k$ converges when $\gamma \in \{0.5, 0.7\}$, but diverges when $\gamma = 0.9$. This again shows that the algorithm is likely to diverge when γ is close to 1 and that the Condition (13) is sufficient but not necessary for convergence. To demonstrate the exponential convergence rate for constant stepsize, we plot $\log \mathbb{E}\left[\|\theta_k - \theta^*\|^2\right]$ as the function of the iteration k when $\gamma = 0.5$, where θ^* is the solution of the projected Bellman equation (12), estimated by the projected value iteration algorithm. Note that, we repeat running the algorithm for 1000 times and use the average as an approximation to the expectation. In Fig. B.2, we observe that the graph is nearly a straight line when k is large enough, meaning that θ_k converges to θ^* geometrically fast, which agrees with Theorem 2 (1).

In our second set of experiments, we consider diminishing stepsizes $\alpha_k = \frac{\alpha}{k^\xi}$, where $\xi \in \{0.4, 0.6, 0.8, 1.0\}$. In the case where $\xi = 1$, the constant α is chosen s.t. $\kappa\alpha > 2$ to achieve the optimal convergence rate. In addition, the discount factor γ is set to be 0.5. Fig. B.3 shows that the algorithm converges for all $\xi \in \{0.4, 0.6, 0.8, 1.0\}$ and the algorithm converges faster with larger ξ. To further illustrate the rate of convergence for each choice of ξ, we plot $\log \mathbb{E}\left[\|\theta_k - \theta^*\|^2\right]$ as a function of $\log k$ in Fig. B.4 and focus on its asymptotic behavior. We can observe that the slope is approximately $-\xi$, which agrees with Corollary 2.

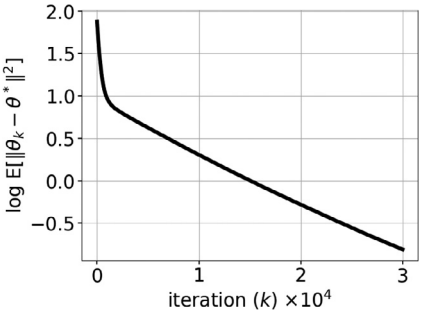

Fig. B.2. Exponential convergence rate of Q-learning with linear function approximation for $\gamma = 0.5$.

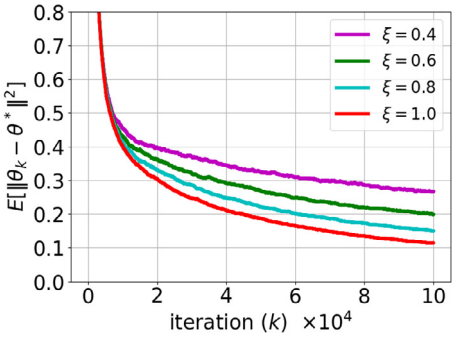

Fig. B.3. Convergence for diminishing stepsizes.

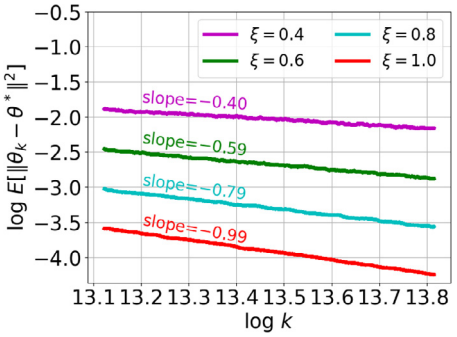

Fig. B.4. Asymptotic convergence rate.

References

Baird, Leemon (1995). Residual algorithms: Reinforcement learning with function approximation. In *Machine learning proceedings 1995* (pp. 30–37). Elsevier.

Beck, Carolyn L., & Srikant, Rayadurgam (2012). Error bounds for constant step-size Q-learning. *Systems & Control Letters*, *61*(12), 1203–1208.

Benaim, Michel (1996). A dynamical system approach to stochastic approximations. *SIAM Journal on Control and Optimization*, *34*(2), 437–472.

Benveniste, Albert, Métivier, Michel, & Priouret, Pierre (2012). *vol. 22, Adaptive algorithms and stochastic approximations*. Springer Science & Business Media.

Bertsekas, Dimitri P., & Tsitsiklis, John N. (1996). *Neuro-dynamic programming*. Athena Scientific.

Bhandari, Jalaj, Russo, Daniel, & Singal, Raghav (2018). A finite time analysis of temporal difference learning with linear function approximation. In *Conference on learning theory* (pp. 1691–1692).

Bhatnagar, Shalabh, & Borkar, Vivek S. (1997). Multiscale stochastic approximation for parametric optimization of hidden Markov models. *Probability in the Engineering and Informational Sciences*, *11*(4), 509–522.

Bhatnagar, Shalabh, & Borkar, Vivek S. (1998). A two timescale stochastic approximation scheme for simulation-based parametric optimization. *Probability in the Engineering and Informational Sciences*, *12*(4), 519–531.

Borkar, Vivek S. (2009). *vol. 48, Stochastic approximation: A dynamical systems viewpoint*. Springer.

Borkar, Vivek S., & Meyn, Sean P. (2000). The ODE method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, *38*(2), 447–469.

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

Charalambous, Charalambos D, Tzortzis, Ioannis, Loyka, Sergey, & Charalambous, Themistoklis (2014). Extremum problems with total variation distance and their applications. *IEEE Transactions on Automatic Control*, *59*(9), 2353–2368.

Dalal, Gal, Szörényi, Balázs, Thoppe, Gugan, & Mannor, Shie (2018). Finite sample analysis for TD(0) with function approximation. In *Thirty-second AAAI conference on artificial intelligence*.

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

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

Fazlyab, Mahyar, Ribeiro, Alejandro, Morari, Manfred, & Preciado, Victor M (2017). A dynamical systems perspective to convergence rate analysis of proximal algorithms. In *2017 55th annual allerton conference on communication, control, and computing (Allerton)* (pp. 354–360). IEEE.

Franca, Guilherme, Robinson, Daniel, & Vidal, Rene (2018). ADMM and accelerated ADMM as continuous dynamical systems. In *International conference on machine learning* (pp. 1559–1567). PMLR.

Haddad, Wassim M., & Chellaboina, VijaySekhar (2011). *Nonlinear dynamical systems and control: A Lyapunov-based approach*. Princeton University Press.

Hu, Bin, Seiler, Peter, & Rantzer, Anders (2017). A unified analysis of stochastic optimization methods using jump system theory and quadratic constraints. In *Conference on learning theory* (pp. 1157–1189). PMLR.

Hu, Bin, & Syed, Usman Ahmed (2019). Characterizing the exact behaviors of temporal difference learning algorithms using Markov jump linear system theory. In *Proceedings of the 33rd international conference on neural information processing systems* (pp. 8479–8490).

Jaakkola, Tommi, Jordan, Michael I., & Singh, Satinder P. (1994). Convergence of stochastic iterative dynamic programming algorithms. In *Advances in neural information processing systems* (pp. 703–710).

Jiang, Nan, Kulesza, Alex, Singh, Satinder, & Lewis, Richard (2015). The dependence of effective planning horizon on model accuracy. In *Proceedings of the 2015 international conference on autonomous agents and multiagent systems* (pp. 1181–1189). Citeseer.

Karmakar, Prasenjit, & Bhatnagar, Shalabh (2021). Stochastic approximation with iterate-dependent Markov noise under verifiable conditions in compact state space with the stability of iterates not ensured. *IEEE Transactions on Automatic Control*.

Kearns, Michael, & Singh, Satinder (1998). Finite-sample convergence rates for Q-learning and indirect algorithms. In *Proceedings of the 11th international conference on neural information processing systems* (pp. 996–1002).

Khalil, Hassan K., & Grizzle, Jessy W. (2002). *vol. 3*, *Nonlinear systems*. Prentice hall Upper Saddle River, NJ.

Kushner, Harold Joseph, & Clark, Dean S. (2012). *vol. 26*, *Stochastic approximation methods for constrained and unconstrained systems*. Springer Science & Business Media.

Lan, Guanghui (2020). *First-order and stochastic optimization methods for machine learning*. Springer.

Lee, Donghwan, & He, Niao (2019). A unified switching system perspective and ODE analysis of Q-learning algorithms. Preprint arXiv:1912.02270.

Levin, David A., & Peres, Yuval (2017). *vol. 107*, *Markov chains and mixing times*. American Mathematical Soc..

Ljung, Lennart (1977). Analysis of recursive stochastic algorithms. *IEEE Transactions on Automatic Control*, *22*(4), 551–575.

Melo, Francisco S., Meyn, Sean P., & Ribeiro, M. Isabel (2008). An analysis of reinforcement learning with function approximation. In *Proceedings of the 25th international conference on machine learning* (pp. 664–671).

Moulines, Eric, & Bach, Francis (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. *Advances in Neural Information Processing Systems*, *24*, 451–459.

Ramaswamy, Arunselvan, & Bhatnagar, Shalabh (2018). Stability of stochastic approximations with "controlled markov" noise and temporal difference learning. *IEEE Transactions on Automatic Control*, *64*(6), 2614–2620.

Robbins, Herbert, & Monro, Sutton (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 400–407.

Romero, Orlando, & Benosman, Mouhacine (2020). Finite-time convergence in continuous-time optimization. In *International conference on machine learning* (pp. 8200–8209). PMLR.

Rudin, Walter, et al. (1964). *vol. 3*, *Principles of mathematical analysis*. McGraw-hill New York.

Sontag, Eduardo D. (2008). Input to state stability: Basic concepts and results. In *Nonlinear and optimal control theory* (pp. 163–220). Springer.

Srikant, R., & Ying, Lei (2019). Finite-time error bounds for linear stochastic approximation and TD learning. In *Conference on learning theory* (pp. 2803–2830).

Sutton, Richard S., & Barto, Andrew G. (2018). *Reinforcement learning: an introduction*. MIT Press.

Thoppe, Gugan, & Borkar, Vivek (2019). A concentration bound for stochastic approximation via Alekseev's formula. *Stochastic Systems*, *9*(1), 1–26.

Tsitsiklis, John N. (1994). Asynchronous stochastic approximation and Q-learning. *Machine Learning*, *16*(3), 185–202.

Tsitsiklis, John N., & Van Roy, Benjamin (1997). Analysis of temporal-difference learning with function approximation. In *Advances in neural information processing systems* (pp. 1075–1081).

Tsitsiklis, John N., & Van Roy, Benjamin (1999). Average cost temporal-difference learning. *Automatica*, *35*(11), 1799–1808.

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

Yaji, Vinayaka G., & Bhatnagar, Shalabh (2019). Analysis of stochastic approximation schemes with set-valued maps in the absence of a stability guarantee and their stabilization. *IEEE Transactions on Automatic Control*, *65*(3), 1100–1115.

Zaiwei Chen is currently a CMI postdoctoral fellow at Caltech CMS department, hosted by Prof. Adam Wierman and Prof. Eric Mazumdar. He received his Ph.D. degree in Machine Learning from the School of Industrial & Systems Engineering at Georgia Tech, advised by Prof. Siva Theja Maguluri and Prof. John-Paul Clarke. He also obtained two M.S. degrees from Georgia Tech, one in Mathematics, and the other in Operations Research. Before that, he received his B.S. degree in Electrical Engineering from Chu Kochen Honors College, Zhejiang University, China. He is interested in applied probability with applications in reinforcement learning, optimization, and control theory.

Sheng Zhang is currently a Ph.D. student in the H. Milton Stewart School of Industrial and Systems Engineering at Georgia Tech. He received his B.S. in Mathematics and Applied Mathematics from Wuhan University and M.S. in Applied Mathematics from Columbia University. His primary research interests are in sequential decision making under uncertainty, reinforcement learning, bandit algorithms, decentralized and distributed optimization, statistical machine learning and their various applications.

Thinh T. Doan is an Assistant Professor in the Electrical and Computer Engineering Department at Virginia Tech. He obtained his Ph.D. degree at the University of Illinois, Urbana-Champaign, his M.S. at the University of Oklahoma, and his B.S. at Hanoi University of Science and Technology, Vietnam, all in Electrical Engineering. His research interests span the intersection of control theory, optimization, machine learning, reinforcement learning, game theory, and applied probability theory.

John-Paul Clarke is a professor of Aerospace Engineering and Engineering Mechanics at The University of Texas at Austin, where he holds the Ernest Cockrell Jr. Memorial Chair in Engineering. Previously, he was a faculty member at the Georgia Institute of Technology and the Massachusetts Institute of Technology (MIT); Vice President of Strategic Technologies at United Technologies Corporation (now Raytheon); and a researcher at Boeing and NASA JPL.

Siva Theja Maguluri is Fouts Family Early Career Professor and Assistant Professor in the H. Milton Stewart School of Industrial and Systems Engineering at Georgia Tech. He received his B.Tech in Electrical Engineering from IIT Madras in 2008, M.S in ECE, M.S. in Applied Math and a Ph.D. in ECE all from University of Illinois at Urbana Champaign. His research interests span the areas of Networks, Control, Optimization, Algorithms, Applied Probability and Reinforcement Learning. He is a recipient of the biennial "Best Publication in Applied Probability" award, NSF CAREER award, "CTL/BP Junior Faculty Teaching Excellence Award", and "Student Recognition of Excellence in Teaching: Class of 1934 ClOS Award".