Finite-Time Error Bounds for Biased Stochastic Approximation with Application to Q-Learning

Gang Wang **Georgios B. Giannakis**
Department of Electrical and Computer Engineering, University of Minnesota, Minneapolis, MN 55455
E-mail: {gangwang, georgios}@umn.edu

Abstract

Inspired by the widespread use of Q-learning algorithms in reinforcement learning (RL), this present paper studies a class of biased stochastic approximation (SA) procedures under an 'ergodic-like' assumption on the underlying stochastic noise sequence. Leveraging a *multistep Lyapunov function* that looks ahead to several future updates to accommodate the gradient bias, we prove a general result on the convergence of the iterates, and use it to derive finite-time bounds on the mean-square error in the case of constant stepsizes. This novel viewpoint renders the finite-time analysis of *biased SA* algorithms under a broad family of stochastic perturbations possible. For direct comparison with past works, we also demonstrate these bounds by applying them to Q-learning with linear function approximation, under the realistic Markov chain observation model. The resultant finite-time error bound for Q-learning is *the first of its kind*, in the sense that it holds: i) for the unmodified version (i.e., without making any modifications to the updates), and ii), for Markov chains starting from any initial distribution, at least one of which has to be violated for existing results to be applicable.

1 INTRODUCTION

Stochastic approximation (SA) algorithms are widely used in a number of areas, including statistical signal processing, control, optimization, machine learning, and RL. Ever since the seminal contribution [Robbins and Monro, 1951], there have been a multitude of efforts on SA schemes, applications, and theoretical developments [Kushner and Yin, 2003], [Nemirovski et al., 2009]. On the theory side, conventional SA convergence analysis and error bounds are mostly asymptotic—that hold only in the limit as the number of iterations increases to infinity. Yet, recent efforts have focused on developing non-asymptotic performance guarantees—that hold even for finite iterations—for SA algorithms in different settings [Nemirovski et al., 2009], [Bach and Moulines, 2011], [Wainwright, 2019] mainly motivated by the emerging need for dealing with massive data examples in modern large-scale optimization and statistical learning tasks.

Many stochastic control tasks can be naturally formulated as Markov decision processes (MDPs), which provide a flexible framework for modeling decision making in scenarios where outcomes are partly random and partly under the control of a decision maker. Reinforcement learning is a collection of tools for solving MDPs, especially when the underlying transition mechanism is unknown [Watkins, 1989]. Originally introduced by [Watkins, 1989], Q-learning has become one of the most widely used RL algorithms nowadays, on which much of the modern artificial intelligence is built [Mnih et al., 2015]. The goal of Q-learning is to obtain a policy that informs an agent what action to taken under what circumstances. It is model-free, namely it does not require a model of the environment, and iteratively estimates the optimal state-action value function (a.k.a. Q-function) based on a sequence of samples generated by operating a fixed policy in the unknown environment. For any MDP with finite state and action spaces, Q-learning finds a policy that is optimal in the sense that it maximizes the expected value of the total reward from each state. Despite its popularity, convergence analysis of Q-learning (with function approximation) has proved challenging; see, e.g., [Tsitsiklis, 1994], [Szepesvári, 1998], [Melo et al., 2008], [Eryilmaz and Srikant, 2012], [Beck and Srikant, 2012]. Connections between Q-

Proceedings of the 23[rd]International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.

learning and SA were drawn in [Tsitsiklis, 1994], and [Bertsekas and Tsitsiklis, 1996]. Non-asymptotic guarantees of RL algorithms appeared only recently, and they remain limited [Bhandari et al., 2019], [Srikant and Ying, 2019], [Wainwright, 2019], [Chen et al., 2019].

Finite-time analysis of temporal-difference (TD) learning with linear approximation was studied in [Dalal et al., 2018], but their results require i.i.d. samples, which are difficult to obtain in practice. Dealing with the more realistic yet challenging Markov chain observation model, finite-time analysis of TD learning was studied in [Bhandari et al., 2019], [Srikant and Ying, 2019], and that of Q-learning appeared lately in [Chen et al., 2019]. However, the bound in [Chen et al., 2019] becomes applicable only after a certain mixing-time number of iterations, that is, after the Markov chain gets sufficiently "close" to its stationary distribution.

Targeting a deeper understanding for the statistical efficiency of Q-learning algorithms, the objective of this present paper is to derive finite-time guarantees for a certain class of biased SA procedures. In particular, we first characterize a set of easy-to-check conditions on the nonlinear operators used in SA updates, and introduce a mild assumption on the stochastic noise sequence satisfied by a broad family of discrete-time stochastic processes. We prove a general convergence result leveraging a novel multistep Lyapunov function, which relies on a number of future SA updates to gain control over the gradient bias arising from instantaneous stochastic perturbations. We further develop finite-time bounds on the mean-square error of the iterates. Finally, for direct comparison to past works, we specialize the results established for general SA algorithms to Q-learning with linear function approximation, from data samples gathered along a single trajectory of a Markov chain. We thereby obtain finite-time error bounds for Q-learning using (non-)linear function approximators in the case of constant stepsizes, under the most general assumptions to date. The merits of our bounds are that they directly apply to i) the unmodified Q-learning algorithm and, iii) Markov chains starting from any initial distribution, as well, as from the first iteration (meaning there is no need to wait until the Markov chain gets "close" to its unique stationary distribution as required by e.g., [Chen et al., 2019])).

2 PROBLEM SETUP

Consider the following nonlinear recursion with a constant stepsize $\epsilon > 0$, starting from $\Theta_0 \in \mathbb{R}^d$

$$\Theta_{k+1} = \Theta_k + \epsilon f(\Theta_k, X_k), \quad k = 0, 1, 2 \dots \tag{1}$$

where $\Theta_k \in \mathbb{R}^d$ denotes the k-th iterate, $\{X_k \in \mathbb{R}^m\}_k$ is a stochastic noise sequence defined on a complete probability space, and $f : \mathbb{R}^d \times \mathbb{R}^m \to \mathbb{R}^d$ is a continuous function of (θ, x). In the simplest setting, for example, $\{X_k\}$ is an i.i.d. random sequence of vectors, while $f(\Theta_k, X_k)$ is a conditionally unbiased estimate of the gradient $\bar{f}(\Theta_k) := \mathbb{E}[f(\Theta_k, X_k)|\mathcal{F}_k]$. Here, $(\mathcal{F}_k)_{k\geq 0}$ is an increasing family of σ-fields, with Θ_0 being $\mathcal{F}_0$-measurable, and $f(\theta, X_k)$ being $\mathcal{F}_k$-measurable. Depending on whether $\mathcal{F}_0$ is a trivial σ-field, the initial guess Θ_0 can be random or deterministic. For simplicity, the rest of this paper assumes a deterministic Θ_0, yet the obtained bounds hold true for a random Θ_0 after replacing $\|\Theta_0\|^2$ with $\mathbb{E}[\|\Theta_0\|^2]$. In a more complicated setting pertaining to MDPs, $\{X_k\}_k$ is a Markov chain assumed to have a unique stationary distribution, and $f(\Theta_k, X_k)$ can be viewed as a biased estimate of some gradient $\bar{f}(\Theta_k) = \lim_{k\to\infty} \mathbb{E}_{X_k}[f(\Theta_k, X_k)]$. In both cases, we are prompted to assume that the following limit exists for each $\theta \in \mathbb{R}^d$

$$\bar{f}(\theta) = \lim_{k\to\infty} \mathbb{E}[f(\theta, X_k)]. \tag{2}$$

Taking a dynamical system viewpoint [Borkar, 2008], the corresponding ODE for (1) is given by

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

Assume that this ODE admits an equilibrium point θ^* at the origin, i.e., $\bar{f}(0) = 0$. This assumption is made without loss of generality, as one can always shift a nonzero equilibrium point to zero through centering $\theta \leftarrow \theta - \theta^*$. Following the terminology in [Borkar, 2008], the recursion (1) is termed nonlinear SA. Our goal here is to provide a non-asymptotic convergence analysis of the iterate sequence $\{\Theta_k\}_{k\in\mathbb{N}^+}$ generated by a recursion of the form (1) to the equilibrium point θ^* of its corresponding ODE (3).

The motivating impetus for considering recursion (1) was to gain a deeper insight into the classical Q-learning algorithm [Watkins, 1989] from discounted MDPs and RL [Sutton and Barto, 2018], [Bertsekas and Tsitsiklis, 1996]. It is a biased SA procedure for solving a fixed point equation defined by the so-called Bellman's operator [Bertsekas and Tsitsiklis, 1996]. In fact, a large family of basic RL algorithms [Sutton and Barto, 2018], including TD(0), TD(λ), and GTD, as well as stochastic gradient descent for nonlinear least-squares estimation can also be described in this form (see e.g., [Srikant and Ying, 2019] for a detailed discussion).

Certainly, convergence guarantees of SA procedures as in (1) would not be possible without imposing assumptions on the operators $f(\theta, x)$ and $\bar{f}(\theta)$. In this work, motivated by the analysis of TD-learning and related

algorithms in RL, we consider a class of SA procedures that satisfy the following properties.

Assumption 1. *The function $f(\theta, x)$ satisfies the globally Lipschitz condition in θ, uniformly in x, i.e., there exists a constant $L_1 > 0$ such that for all θ, $\theta' \in \mathbb{R}^d$ and each $x \in \mathcal{X}$, it holds that*

$$\|f(\theta, x) - f(\theta', x)\| \leq L_1 \|\theta - \theta'\|. \tag{4}$$

Moreover, there exists a constant $L_2 > 0$ such that, for each $x \in \mathcal{X}$, it holds for all θ

$$\|f(\theta, x)\| \leq L_2(\|\theta\| + 1) \tag{5}$$

where $\mathcal{X} \subseteq \mathbb{R}^m$ denotes the living space of the stochastic process $\{X_k\}$.

It is worth pointing out that (5) is equivalent to assuming that $f(0, x)$ satisfying (4) is uniformly bounded for all $x \in \mathcal{X}$. To see this, suppose that $\|f(0, x)\| \leq \hat{f}$ holds for all $x \in \mathcal{X}$. Using (4), it follows readily that $\|f(\theta, x)\| \leq L_1\|\theta - \theta'\| + \|f(\theta', x)\|$, in which taking $\theta' = 0$ confirms that $\|f(\theta, x)\| \leq L_1\|\theta\| + \|f(0, x)\| \leq L_1\|\theta\| + \hat{f} \leq \max(L_1, \hat{f})(\|\theta\| + 1)$. By defining $L := \max\{L_1, L_2\}$, we will assume for simplicity that (4) and (5) hold with the same constant L.

Assumption 2. *Consider the ODE $\dot{\theta}(t) = \bar{f}(\theta(t))$ in (3). There exists a twice differentiable function $W(\theta)$ (a.k.a., Lyapunov function) that satisfies globally and uniformly the following conditions for all θ, $\theta' \in \mathbb{R}^d$*

$$c_1\|\theta\|^2 \leq W(\theta) \leq c_2\|\theta\|^2 \tag{6a}$$

$$\left(\left.\frac{\partial W}{\partial \theta}\right|_\theta\right)^\top \bar{f}(\theta) \leq -c_3 L\|\theta\|^2 \tag{6b}$$

$$\left\|\left.\frac{\partial W}{\partial \theta}\right|_\theta - \left.\frac{\partial W}{\partial \theta}\right|_{\theta'}\right\| \leq c_4\|\theta - \theta'\| \tag{6c}$$

for some constants $c_1, c_2, c_3, c_4 > 0$.

For an introduction to Lyapunov theory, see e.g., standard source [Khalil, 2002, Ch. 4]. Regarding these assumptions, two remarks come in order.

Remark 1. *As. 1 is standard and widely adopted in convergence analysis of SA algorithms; see e.g., [Borkar, 2008, Ch. 3], [Bach and Moulines, 2011], [Tsitsiklis, 1994] and [Srikant and Ying, 2019] in the case of linear SA (i.e., $f(\theta, x)$ is linear in θ).*

Remark 2. *By evaluating inequality* (6a) *at $\theta = 0$, one confirms that $W(\theta) > W(0) = 0$ for all $\theta \neq 0$. Since $W(\theta)$ is twice differentiable, it implies that $\frac{\partial W}{\partial \theta}|_{\theta=0} = 0$. From* (6b)*, it holds that both $\bar{f}(\theta) \neq 0$ and $\frac{\partial W}{\partial \theta}|_\theta \neq 0$ at any point $\theta \neq 0$. In words, Assumption 2 states that the equilibrium point $\theta = 0$ is unique, and globally, asymptotically stable for the ODE* (3)*. This also appeared in e.g., [Borkar, 2008, A5] and [Bach and Moulines, 2011] (strongly convex case). This is in the same spirit of requiring a Hurwitz matrix $\bar{A}$ (i.e., every eigenvalue has strictly negative real part) for the ODE $\dot{\theta} = \bar{A}\theta$ in linear SA by [Tsitsiklis and Van Roy, 1997, Thm. 2], [Dalal et al., 2018], [Srikant and Ying, 2019].*

In addition to As. 1 and 2, to leverage the ODE to study convergence of SA procedures, we make an assumption on the stochastic noise sequence $\{X_k\}_{k\in\mathbb{N}}$.

Assumption 3. *For each $\theta \in \mathbb{R}^d$, the random vector $f(\theta, X_k)$ is $\mathcal{F}_k$-measurable, and there exists a function $\sigma(T; T_0) : \mathbb{N}^+ \times \mathbb{N}^+ \to \mathbb{R}^+$ monotonically decreasing to zero as either $T \to \infty$ or $T_0 \to \infty$; i.e., $\lim_{T\to\infty} \sigma(T; T_0) = 0$ for any fixed $T_0 \in \mathbb{N}^+$, and $\lim_{T_0\to\infty} \sigma(T; T_0) = 0$ for any fixed $T \in \mathbb{N}^+$, such that*

$$\left\|\frac{1}{T}\sum_{k=T_0}^{T_0+T-1} \mathbb{E}\big[f(\theta, X_k)\big|\mathcal{F}_{T_0}\big] - \bar{f}(\theta)\right\| \leq \sigma(T; T_0)L(\|\theta\|+1) \tag{7}$$

where the expectation $\mathbb{E}$ is taken over $\{X_k\}_{k=T_0}^{T_0+T-1}$ conditioned on $\mathcal{F}_{T_0}$.

In fact, As. 3 requires that the bias of the 'ergodic' average of any T consecutive gradient estimates $\{f(\theta, X_k)\}_{k=T_0}^{T_0+T-1}$ from their limit $\bar{f}(\theta)$ vanishes (at least) sublinearly in T. Indeed, this is fairly mild and more general than those studied individually by e.g., [Bach and Moulines, 2011], [Bhandari et al., 2019], [Srikant and Ying, 2019], each of which imposes requirements on each gradient estimate $f(\theta, X_k)$. For example, [Bach and Moulines, 2011] entails an *unbiased* gradient estimate per iteration, [Dalal et al., 2018, Bhandari et al., 2019] has to incorporate a projection step for control of the instantaneous gradient bias, and [Srikant and Ying, 2019] requires the initial distribution of the Markov chain to be sufficiently close to the stationary distribution for the bound to be applicable.

In contrast, our condition (7) can allow for large instantaneous biased gradients $f(\Theta_k, X_k)$ of $\bar{f}(\Theta_k)$. Further, As. 3 is satisfied by a broad family of discrete-time stochastic processes, including i.i.d. random vector sequences [Bach and Moulines, 2011], finite-state irreducible and aperiodic Markov chains, and Ornstein-Uhlenbeck processes; whereas past works [Bach and Moulines, 2011], [Bhandari et al., 2019], [Srikant and Ying, 2019] deal only with one such type of those stochastic processes.

3 FINITE-TIME BOUNDS ON THE MEAN-SQUARE ERROR

In this paper, we seek to develop novel tools for proving non-asymptotic bounds on the mean-square error

of the iterates $\{\Theta_k\}_{k\geq 1}$ generated by a recursion of the form (1) (to the equilibrium point $\theta^* = 0$). Before presenting the main results, we start off by introducing an instrumental result which is the key to our novel approach to controlling possible bias present in the gradient estimate of SA procedures. Its proof is provided in Appendix A of the supplementary material.

Proposition 1. *Under As. 1 and 3, there exists a function $g'(k, T, \Theta_k)$ such that the next holds $\forall T \in \mathbb{N}^+$*

$$\Theta_{k+T} = \Theta_k + \epsilon T \bar{f}(\Theta_k) + g'(k, T, \Theta_k) \tag{8}$$

satisfying

$$\left\|\mathbb{E}\big[g'(k, T, \Theta_k)\big|\mathcal{F}_k\big]\right\| \leq \epsilon L T \beta(T, \epsilon)(\|\Theta_k\| + 1) \tag{9}$$

$$\beta(T, \epsilon) := \epsilon L T (1 + \epsilon L)^{T-2} + \sigma(T; k) \tag{10}$$

where the expectation is taken over $\{X_j\}_{j=k}^{k+T-1}$ conditioned on $\mathcal{F}_k$.

Evidently, Prop. 1 offers a bound on the average gradient bias over a number $T > 0$ of iterations, which is indeed motivated by our As. 3. Based on the results in Prop. 1, we present the following theorem, which establishes a general convergence result that applies to any stochastic sequence $\{X_k\}_{k\in\mathbb{N}}$ satisfying As. 3.

Theorem 1. *Under As. 1—3 and for any $\delta > 0$, there exist a function $W'(k, \Theta_k)$, and constants $(T^* \in \mathbb{N}^+, \epsilon_\delta)$ such that $\sigma(T^*; k) \leq \delta$ and the ensuing inequalities are globally and uniformly satisfied for all $\epsilon \in (0, \epsilon_\delta)$ and $k \in \mathbb{N}$*

$$c_1'\|\Theta_k\|^2 \leq W'(k, \Theta_k) \leq c_2'\|\Theta_k\|^2 + c_2''(\epsilon L)^2 \tag{11}$$

$$\begin{aligned}\mathbb{E}\big[W'(k+1, \Theta_{k+1}) - W'(k, \Theta_k)\big|\mathcal{F}_k\big] \\ \leq -\epsilon c_3'\|\Theta_k\|^2 + c_4'\epsilon^2 + c_5'\sigma(T^*; k)\epsilon\end{aligned} \tag{12}$$

where $c_1', c_2', c_2'', c_3', c_4', c_5' > 0$ are constants dependent on $c_1 \sim c_4$ of (6) but independent of $\epsilon > 0$.

Proof of Thm. 1 is relegated to Appendix B of the supplementary material. Our proof builds critically on the construction of function $W'(k, \Theta_k)$ from the Lyapunov function $W(\theta)$ of the ODE (3). To use the concentration bound (7), we are motivated to introduce a function candidate that necessarily looks ahead to a number of T future iterates, with parameter $T \geq 1$ to be designed such that the gradient bias can be made affordable, given by

$$W'(k, \Theta_k) = \sum_{j=k}^{k+T-1} W(\Theta_j(k, \Theta_k)) \tag{13}$$

where, to make dependence of $\Theta_{j\geq k}$ as a function of Θ_k explicit, we intentionally write $\Theta_j = \Theta_j(k, \Theta_k)$, understood as the iterate of the recursion (1) at time $j \geq k$ with initial condition Θ_k at time k. It is just this parameter $T \geq 1$ that allows us to exploit the monotonically decreasing function $\sigma(T; k) \to 0$ in (9) to gain control over large instantaneous gradient bias. This renders the *general* convergence bounds (11)–(12) possible, in the sense that they hold for any nonlinear SA procedure with underlying random sequence obeying As. 1–3. For instance, when the underlying noise sequence $\{X_k\}_k$ is i.i.d. [Bach and Moulines, 2011], or a Markov chain that has approximately arrived at its steady state (i.e., after a certain mixing time of recursions) [Srikant and Ying, 2019], they have shown that it suffices to choose $T = 1$, that is $W'(\Theta_k) = W(\Theta_k)$ to validate (11)–(12). For *general* Markov chains starting from any initial distribution however, functions like $W'(\Theta_k) = W(\Theta_k)$ may fail to yield finite-time bounds that hold for the entire sequence $\{\Theta_k\}_{k\geq 1}$. In a nutshell, our novel way of constructing this multistep Lyapunov function is indeed motivated by and well-suited for taking care of this kind of 'mixing' behavior. It goes beyond the Markov chain to be useful for finite-time analysis of general SA algorithms driven by a broad family of (discrete-time) stochastic processes.

We are now ready to study the drift of $W'(k, \Theta_k)$, which follows from Thm. 1, and whose proof is provided in Appendix C of the supplementary material.

Lemma 1. *Under As. 1—3, the following holds true for all $\epsilon \in (0, \epsilon_\delta)$ and $k \in \mathbb{N}$*

$$\begin{aligned}\mathbb{E}\big[W'(k+1, \Theta_{k+1})\big] \leq \left(1 - \frac{c_3'\epsilon}{c_2'}\right)\mathbb{E}\big[W'(k, \Theta_k)\big] + c_4''\epsilon^2 \\ + c_5'\sigma(T^*; k)\epsilon\end{aligned} \tag{14}$$

where $c_4'' > 0$ is an appropriate constant independent of ϵ, and $T^* \in \mathbb{N}^+$ is fixed in Thm. 1.

Theorem 2. *Let $k_\epsilon := \min\{k \in \mathbb{N}^+ | \sigma(T^*; k) \leq \epsilon\}$. Under As. 1–3, and choosing stepsize $\epsilon \in (0, \epsilon_\delta)$, the following finite-time error bounds hold for all $k \in \mathbb{N}$*

$$\begin{aligned}\mathbb{E}\big[\|\Theta_k\|^2\big] \leq \frac{c_2'}{c_1'}\left(1 - \frac{c_3'\epsilon}{c_2'}\right)^k \|\Theta_0\|^2 + \frac{c_2''L^2}{c_1'}\epsilon^2 + \frac{c_6}{c_1'}\epsilon \\ + \frac{c_6}{c_1'}\left(1 - \frac{c_3'\epsilon}{c_2'}\right)^{\max\{k-k_\epsilon, 0\}}\delta\end{aligned} \tag{15}$$

where $c_6 > 0$ is a constant, and δ is given in Thm. 1.

When a random initial estimate Θ_0 is considered, one just needs to replace the term $\|\Theta_0\|^2$ with its expectation $\mathbb{E}[\|\Theta_0\|^2]$ in (15), and the resulting bound holds. Proof of Thm. 2 is postponed to Appendix D of the supplemental document. At this point, some observations are worth making.

Remark 3. *Existing non-asymptotic results have focused on linear SA including e.g., [Dalal et al., 2018],*

[Srikant and Ying, 2019], [Bhandari et al., 2019], or nonlinear SA under i.i.d. noise e.g., [Nemirovski et al., 2009], [Bach and Moulines, 2011]. In contrast, our finite-time bound in Thm. 2 is applicable to a class of nonlinear SA procedures under a broad family of stochastic noise sequences.

Remark 4. *When the general recursion* (1) *is specialized to linear SA driven by Markovian noise* $\{X_k\}_{k\in\mathbb{N}}$*, i.e.,* $f(\Theta_k, X_k) = A(X_k)\Theta_k + b(X_k)$*, our established bound in* (15) *improves upon the state-of-the-art in [Srikant and Ying, 2019, Theorem 7]. In fact, the bound in [Srikant and Ying, 2019, Theorem 7] becomes applicable only after a mixing time of updates (i.e., for* $k \geq \tau$ *with* $\tau \gg 1$ *being the mixing time of the Markov chain* $\{X_k\}$*) till the Markov chain gets sufficiently 'close' to its stationary distribution; yet, in contrast, our bound* (15) *is effective from the first iteration for Markov chains starting with any initial distribution. Moreover, our stead-state value (the last term of* (15)*) scales only with the stepsize* $\epsilon > 0$ *(which has removed the independence on* τ *from the bound in [Srikant and Ying, 2019]), and it vanishes as* $\epsilon \to 0$*.*

Evidently, with the bound in (15), one can easily estimate the number of samples (e.g., the length of a Markov chain trajectory) required for the mean-square error to be of the same order as its steady-state value.

4 APPLICATIONS TO APPROXIMATE Q-LEARNING

We now turn to the consequences of our general results for the problem of Q-learning with linear function approximation. Toward this objective, we begin by providing a brief introduction to discounted MDPs and basic RL algorithms; interested readers can refer to standard sources (e.g., [Sutton and Barto, 2018], [Bertsekas and Tsitsiklis, 1996]) for more background.

4.1 Background and Problem Setup

Consider a discounted MDP, defined by the quintuple $(\mathcal{S}, \mathcal{U}, \mathcal{P}, R, \gamma)$, where $\mathcal{S}$ is a finite set of possible states (a.k.a. state space), $\mathcal{U}$ is a finite set of possible actions (a.k.a. action space), $\mathcal{P} := \{P^u \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|} | u \in \mathcal{U}\}$ is a collection of probability transition matrices, indexed by actions u, $R(s,u) : \mathcal{S}\times\mathcal{U} \to \mathbb{R}$ is a reward received upon executing action u while in state s, and $\gamma \in [0,1)$ is the discount factor. The results along with theoretical analysis developed in this paper may be generalized to deal with infinite and compact state and/or action spaces, but we restrict ourselves to finite spaces here for an ease of exposition.

An agent selects actions to interact with the MDP (the environment) by operating a policy. Specifically, at each time step $k \in \mathbb{N}$, the agent first observes the state $S_k = s \in \mathcal{S}$ of the environment, and takes an action $U_k = u \in \mathcal{U}$ by following a deterministic policy $\pi : \mathcal{S} \to \mathcal{U}$, or a stochastic one $U_k \sim \pi(\cdot|S_k)$, where $\pi(\cdot|s)$ is a probability distribution function supported on $\mathcal{U}$. The environment then moves to the next state $S_{k+1} = s' \in \mathcal{S}$ with probability $P^u_{ss'} = \Pr(S_{k+1} = s' | S_k = s, U_k = u)$, associated with which an instantaneous reward $R_k := R(S_k, U_k)$ is revealed to the agent. Repeating this procedure generates a single trajectory of states, actions, and rewards, namely, $S_0, U_0, R_0, S_1, \ldots, S_T, U_T, R_T, S_{T+1}, \ldots$ over $\mathcal{S}\times\mathcal{U}\times\mathbb{R}$.

We can define for control purpose the so-called action-value function (a.k.a., Q-function), which measures the quality of a given policy by the expected sum of discounted instantaneous rewards, conditioned on starting in a given state-action pair, and following the policy π to take subsequent actions; i.e.,

$$Q(s,u) = \mathbb{E}\Big[\sum_{k=0}^{\infty} \gamma^k R(S_k, U_k) \big| S_0 = s, U_0 = u\Big], \quad \text{where } U_k \sim \pi(\cdot|S_k) \text{ for all } k \in \mathbb{N}^+. \tag{16}$$

Naturally, one would like to choose the policy π such that the values of the Q-function are optimized. In fact, it has been established that the Q-function associated with the optimal policy π^*, yielding the optimal Q-function denoted by Q^*, satisfies the following Bellman equation [Bertsekas and Tsitsiklis, 1996, Tsitsiklis, 1994]

$$Q^*(s,u) = \mathbb{E}[R(s,u)] + \gamma \mathbb{E}\Big[\max_{u'\in\mathcal{U}} Q^*(s',u') \big| s, u\Big] \tag{17}$$

for all state-action pairs $(s,u) \in \mathcal{S}\times\mathcal{U}$. After assuming a canonical ordering on the elements of $\mathcal{S}\times\mathcal{U}$, the table Q can be treated as a matrix in $\mathbb{R}^{|\mathcal{S}|\times|\mathcal{U}|}$. Once $\{Q^*(s,u)\}_{s,u}$ becomes available, an optimal policy π^* can be recovered by setting $\pi^*(s) \in \arg\max_{u\in\mathcal{U}} Q^*(s,u)$ for all $s \in \mathcal{S}$, without any knowledge about the transition probabilities.

In the learning context of interest, the transition probabilities $\{P^u_{ss'}\}_{s,u,s'}$ are typically unknown and the dimensions $|\mathcal{S}|$ and/or $|\mathcal{U}|$ can be huge or even infinity in practice, so it is almost impossible to exactly evaluate the Bellman equation (17). As one of the most popular solutions for finding the optimal policy, Q-learning [Watkins, 1989] iteratively updates the estimate Q_k of Q^* using a single trajectory of samples $\{(S_k, U_k, S_{k+1})\}$ generated by following the policy π, according to the recursion

$$Q_{k+1}(S_k,U_k) = Q_k(S_k,U_k) + \epsilon_k\Big[R(S_k,U_k) + \gamma \max_{u'\in\mathcal{U}} Q_k(S_{k+1},u') - Q_k(S_k,U_k)\Big] \tag{18}$$

where $\{0 < \epsilon_k < 1\}$ is a sequence of stepsizes to be chosen by the user. Under standard conditions on the stepsizes, the sequence $\{Q_k\}$ converges to Q^* almost surely as long as every state-action pair $(s,u) \in \mathcal{S} \times \mathcal{U}$ is visited infinitely often; see, for instance, [Tsitsiklis, 1994, Bertsekas and Tsitsiklis, 1996].

It is known that for many important problems of interest, the computational requirements of exact function estimation are overwhelming, because of a large number of states and actions (i.e., Bellman's 'curse of dimensionality') [Bertsekas and Tsitsiklis, 1996]. Instead, a popular approach has been to leverage low-dimensional parametric approximants of the value function, or the Q-function. Although nonlinear approximators such as deep neural networks [Mnih et al., 2015], [Wang et al., 2019] could lead to more powerful approximation, the simplicity of RL with linear approximation [Sutton and Barto, 2018] allows us to analyze them in detail.

4.2 Q-learning with Linear Approximation

In this section, we provide a non-asymptotic analysis for the original Q-learning with linear function approximation. Specifically, we assume that the Q-function is parameterized by a linear function as follows

$$Q(s,u) \approx Q^\theta(s,u) = \psi^\top(s,a)\theta \tag{19}$$

where $\theta \in \mathbb{R}^d$ is a parameter vector to be learned, typically of size $d \ll |\mathcal{S}| \times |\mathcal{U}|$, the number of state-action pairs; and the feature vector $\psi(s,a) \in \mathbb{R}^d$ stacks up d features produced by pre-selected basis functions $\{\psi_\ell(s,u) : \mathcal{S} \times \mathcal{U} \to \mathbb{R}\}_{\ell=1}^d$. For future reference, we introduce the so-called feature matrix, given by

$$\Psi := \begin{bmatrix} \psi^\top(s_1,u_1) \\ \psi^\top(s_1,u_2) \\ \vdots \\ \psi^\top(s_{|\mathcal{S}|},u_{|\mathcal{U}|}) \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}||\mathcal{U}| \times d}$$

which is assumed to have full column rank (that is, linearly independent columns) and satisfy $\|\psi(s,u)\| \leq 1$ for all state-action pairs $(s,u) \in \mathcal{S} \times \mathcal{U}$.

The well-known Q-learning algorithm updates the parameter vector Θ, according to (e.g., [Watkins, 1989])

$$\Theta_{k+1} = \Theta_k + \epsilon\psi(S_k,U_k)\Big[R(S_k,U_k) + \gamma \max_{u\in\mathcal{U}} \psi^\top(S_{k+1},u)\Theta_k - \psi^\top(S_k,U_k)\Theta_k\Big] \tag{20}$$

for some constant stepsize $\epsilon \in (0,1)$. The goal here is to obtain finite-time error bounds for (20), when the observed data samples $\{(S_k,U_k,R(S_k,U_k),S_{k+1},U_{k+1})\}_{k\in\mathbb{N}}$ are collected along a single path of the Markov chain $\{S_k\}_{k\in\mathbb{N}}$ by following a deterministic policy π. With $X_k := (S_k,U_k,S_{k+1})$, considering

$$F(\theta,X_k) = \psi(S_k,U_k)\Big[R(S_k,U_k) + \gamma \max_{u\in\mathcal{U}} \psi^\top(S_{k+1},u)\theta - \psi^\top(S_k,U_k)\theta\Big] \tag{21}$$

it becomes obvious that (20) has the form of the SA update (1). For our non-asymptotic error guarantees established for nonlinear SA procedures in Theorem 2 to be applicable to Q-learning with linear function approximation, it suffices to show that As. 1–3 are satisfied by the Q-learning updates (20).

In general, Q-learning with even linear function approximation can diverge [Gordon, 1995]. This is mainly because Q-learning implements off-policy [1] sampling to collect data, which renders the expected Q-learning update possibly an expansive mapping [Gordon, 1995]. Under appropriate regularity conditions on the sampling policy, asymptotic convergence of Q-learning with linear function approximation was established in [Melo et al., 2008], and finite-time analysis was recently studied in [Chen et al., 2019]. In the following, we also impose a similar regularity condition on the sampling policy π [Chen et al., 2019].

Assumption 4. *Suppose that the Markov chain $\{S_k\}_{k\in\mathbb{N}}$ induced by policy π is irreducible and aperiodic, whose unique stationary distribution is denoted by μ. Assume that the equation $\bar{F}(\theta) := \mathbb{E}_\mu[F(\theta,X)] = 0$ has a unique solution θ^*, and the next inequality holds for all $\theta \in \mathbb{R}^d$*

$$\gamma^2\mathbb{E}_\mu\Big[\max_{u'\in\mathcal{U}}\big(\psi^\top(s',u')\,\theta\big)^2\Big] - \mathbb{E}_\mu\Big[\big(\psi^\top(s,u)\,\theta\big)^2\Big] \leq -c\|\theta\|^2$$

where $u \sim \pi(\cdot|s)$, for some constant $0 < c < 1$.

Now, let us turn to verify As. 1–3. To this end, we start by introducing $\tilde{\theta} := \theta - \theta^*$ and $X := (S,U,S')$. It then follows that

$$f(\tilde{\theta}) := F(\tilde{\theta}+\theta^*,X) = \psi(S,U)\Big[R(S,U) + \gamma \max_{u\in\mathcal{U}} \psi^\top(S',u)\theta - \psi^\top(S,U)\theta\Big]. \tag{22}$$

It is evident that $\bar{f}(\tilde{\theta}) := \mathbb{E}_\mu[F(\tilde{\theta}+\theta^*,X)] = 0$ has a unique solution $\tilde{\theta}^* = 0$. Now, we can rewrite (20) as

$$\tilde{\Theta}_{k+1} = \tilde{\Theta}_k + \epsilon f(\tilde{\Theta}_k, X_k). \tag{23}$$

[1] On-policy methods estimate the value of a policy while using it for control (namely, take actions); while in off-policy methods, the policy used to generate behavior, called the behavior/sampling policy, may be independent of the policy that is evaluated and improved, called the target/estimation policy [Sutton and Barto, 2018].

Verifying As. 1. For any $\tilde{\theta}_1$, $\tilde{\theta}_2$, and $x=(s,u,s')$, we have that

$$\begin{aligned}
&\|f(\tilde{\theta}_1,x)-\bar{f}(\tilde{\theta}_2,x)\| \\
&=\Big\|\psi(s,u)\Big[R(s,u)+\gamma\max_{u_1\in\mathcal{U}}\psi^\top(s',u_1)\big(\tilde{\theta}_1+\theta^*\big) \\
&\quad-\psi^\top(s,u)\big(\tilde{\theta}_1+\theta^*\big)\Big]-\psi(s,u)\Big[R(s,u) \\
&\quad+\gamma\max_{u_2\in\mathcal{U}}\psi^\top(s',u_2)\big(\tilde{\theta}_2+\theta^*\big)-\psi^\top(s,u)\big(\tilde{\theta}_2+\theta^*\big)\Big]\Big\| \\
&=\Big\|\gamma\psi(s,u)\Big[\max_{u_1\in\mathcal{U}}\psi^\top(s',u_1)\big(\tilde{\theta}_1+\theta^*\big) \\
&\quad-\max_{u_2\in\mathcal{U}}\psi^\top(s',u_2)(\tilde{\theta}_2+\theta^*)\Big]+\psi(s,u)\psi^\top(s,u)(\tilde{\theta}_1-\tilde{\theta}_2)\Big\| \\
&\leq\gamma\Big|\max_{u_1\in\mathcal{U}}\psi^\top(s',u_1)(\tilde{\theta}_1+\theta^*)-\max_{u_2\in\mathcal{U}}\psi^\top(s',u_2)(\tilde{\theta}_2+\theta^*)\Big| \\
&\quad+\big\|\tilde{\theta}_1-\tilde{\theta}_2\big\| \qquad (24)
\end{aligned}$$

where the last inequality follows from $\|\psi(s,u)\|\leq 1$ for all $(s,u)\in\mathcal{S}\times\mathcal{U}$.

Suppose that $u_1^*\in\max_{u_1\in\mathcal{U}}\psi^\top(s',u_1)\big(\tilde{\theta}_1+\theta^*\big)$, then

$$\begin{aligned}
&\max_{u_1\in\mathcal{U}}\psi^\top(s',u_1)\big(\tilde{\theta}_1+\theta^*\big)-\max_{u_2\in\mathcal{U}}\psi^\top(s',u_2)\big(\tilde{\theta}_2+\theta^*\big) \\
&=\psi^\top(s',u_1^*)\big(\tilde{\theta}_1+\theta^*\big)-\max_{u_2\in\mathcal{U}}\psi^\top(s',u_2)\big(\tilde{\theta}_2+\theta^*\big) \\
&\leq\psi^\top(s',u_1^*)\big(\tilde{\theta}_1+\theta^*\big)-\psi^\top(s',u_1^*)\big(\tilde{\theta}_2+\theta^*\big) \\
&=\psi^\top(s',u_1^*)\big(\tilde{\theta}_1-\tilde{\theta}_2\big) \\
&\leq\big\|\tilde{\theta}_1-\tilde{\theta}_2\big\| \qquad (25)
\end{aligned}$$

due again to $\|\psi(s',u_1^*)\|\leq 1$. On the other hand, if we let $u_2^*\in\max_{u_2\in\mathcal{U}}\psi^\top(s',u_2)\big(\tilde{\theta}_2+\theta^*\big)$, it follows similarly that

$$\begin{aligned}
&\max_{u_1\in\mathcal{U}}\psi^\top(s',u_1)\big(\tilde{\theta}_1+\theta^*\big)-\max_{u_2\in\mathcal{U}}\psi^\top(s',u_2)\big(\tilde{\theta}_2+\theta^*\big) \\
&=\max_{u_1\in\mathcal{U}}\psi^\top(s',u_1)\big(\tilde{\theta}_1+\theta^*\big)-\psi^\top(s',u_2^*)\big(\tilde{\theta}_2+\theta^*\big) \\
&\geq\psi^\top(s',u_2^*)\big(\tilde{\theta}_1+\theta^*\big)-\psi^\top(s',u_2^*)\big(\tilde{\theta}_2+\theta^*\big) \\
&=\psi^\top(s',u_2^*)\big(\tilde{\theta}_1-\tilde{\theta}_2\big) \\
&\geq-\big\|\tilde{\theta}_1-\tilde{\theta}_2\big\|. \qquad (26)
\end{aligned}$$

Combining (25) and (26) yields

$$\begin{aligned}
&\Big|\max_{u_1\in\mathcal{U}}\psi^\top(s',u_1)\big(\tilde{\theta}_1+\theta^*\big)-\max_{u_2\in\mathcal{U}}\psi^\top(s',u_2)\big(\tilde{\theta}_2+\theta^*\big)\Big| \\
&\leq\big\|\tilde{\theta}_1-\tilde{\theta}_2\big\| \qquad (27)
\end{aligned}$$

which, in conjunction with (24), proves that

$$\|f(\tilde{\theta}_1,x)-f(\tilde{\theta}_2,x)\|\leq(\gamma+1)\big\|\tilde{\theta}_1-\tilde{\theta}_2\big\|. \qquad (28)$$

In the meanwhile, it is easy to see that

$$\begin{aligned}
\|f(\tilde{\theta},x)\|&=\Big\|\psi(s,u)\big[R(s,u)+\gamma\max_{u_1\in\mathcal{U}}\psi^\top(s',u_1)\big(\tilde{\theta}+\theta^*\big) \\
&\quad-\psi^\top(s,u)\big(\tilde{\theta}+\theta^*\big)\big]\Big\| \\
&\leq|R(s,u)|+\big[\gamma\|\psi(s',u_1^*)\|+\|\psi(s,u)\|\big]\big\|\tilde{\theta}+\theta^*\big\| \\
&\leq\bar{r}+(\gamma+1)\big(\|\tilde{\theta}\|+\|\theta^*\|\big) \\
&=(\gamma+1)\|\tilde{\theta}\|+\big[\bar{r}+(\gamma+1)\|\theta^*\|\big] \qquad (29)
\end{aligned}$$

where we have used the fact that $|R(s,u)|\leq\bar{r}$ for all $(s,u)\in\mathcal{S}\times\mathcal{U}$. With (28) and (29), we have proved that As. 1 is met with $L:=\max\{\gamma+1,\ \bar{r}+(\gamma+1)\|\theta^*\|\}$.

Verifying As. 2. The ODE associated with the (centered) Q-learning update (23) is

$$\begin{aligned}
\dot{\tilde{\theta}}=\bar{f}(\tilde{\theta})&=\mathbb{E}_\mu\Big\{\psi(s,u)\big[R(s,u)+\gamma\max_{u'\in\mathcal{U}}\psi^\top(s',u')(\tilde{\theta}+\theta^*) \\
&\quad-\psi^\top(s,u)(\tilde{\theta}+\theta^*)\big]\Big\} \qquad (30)
\end{aligned}$$

for which we consider the Lyapunov candidate function $W(\tilde{\theta})=\|\tilde{\theta}\|^2/2$. Evidently, it follows that $W(\tilde{\theta})\geq 0$ for all $\tilde{\theta}\neq 0$, so (6a) holds with $c_1=c_2=1/2$. Secondly, using $\bar{f}(\tilde{\theta}^*)=0$, we have that

$$\begin{aligned}
&\big(\frac{\partial W(\tilde{\theta})}{\partial\tilde{\theta}}\big)^\top\bar{f}(\tilde{\theta}) \\
&=\big(\frac{\partial W(\tilde{\theta})}{\partial\tilde{\theta}}\big)^\top\big[\bar{f}(\tilde{\theta})-\bar{f}(\tilde{\theta}^*)\big] \\
&=\tilde{\theta}^\top\mathbb{E}_\mu\big\{\psi(s,u)\big[R(s,u)+\gamma\max_{u_1\in\mathcal{U}}\psi^\top(s',u_1)(\tilde{\theta}+\theta^*) \\
&\quad-\psi^\top(s,u)\big(\tilde{\theta}+\theta^*\big)\big]-\psi(s,u)\big[R(s,u) \\
&\quad+\gamma\max_{u_2\in\mathcal{U}}\psi^\top(s',u_2)\theta^*-\psi^\top(s,u)\theta^*\big]\Big\} \\
&=\gamma\mathbb{E}_\mu\Big\{\tilde{\theta}^\top\psi(s,u)\big[\max_{u_1\in\mathcal{U}}\psi^\top(s',u_1)(\tilde{\theta}+\theta^*) \\
&\quad-\max_{u_2\in\mathcal{U}}\psi^\top(s',u_2)\theta^*\big]\Big\}-\mathbb{E}_\mu\big[\psi^\top(s,u)\tilde{\theta}\big]^2 \\
&\leq-\mathbb{E}_\mu\big[\psi^\top(s,u)\tilde{\theta}\big]^2+\gamma\sqrt{\mathbb{E}_\mu\big[\psi^\top(s,u)\tilde{\theta}\big]^2} \\
&\quad\times\sqrt{\mathbb{E}_\mu\big[\max_{u_1\in\mathcal{U}}\psi^\top(s',u_1)(\tilde{\theta}+\theta^*)-\max_{u_2\in\mathcal{U}}\psi^\top(s',u_2)\theta^*\big]^2} \qquad (31) \\
&\leq\sqrt{\mathbb{E}_\mu[\psi^\top(s,u)\tilde{\theta}]^2}\Big\{\gamma\sqrt{\mathbb{E}_\mu\max_{u'\in\mathcal{U}}[\psi^\top(s',u')\tilde{\theta}]^2} \\
&\quad-\sqrt{\mathbb{E}_\mu\big[\psi^\top(s,u)\tilde{\theta}\big]^2}\Big\} \qquad (32) \\
&=\sqrt{\mathbb{E}_\mu\big[\psi^\top(s,u)\tilde{\theta}\big]^2} \\
&\quad\times\frac{\gamma^2\mathbb{E}_\mu\big[\max_{u'\in\mathcal{U}}\big(\psi^\top(s',u')\tilde{\theta}\big)^2\big]-\mathbb{E}_\mu\big[\psi^\top(s,u)\tilde{\theta}\big]^2}{\gamma\sqrt{\mathbb{E}_\mu\max_{u'\in\mathcal{U}}\big[\psi^\top(s',u')\tilde{\theta}\big]^2}+\sqrt{\mathbb{E}_\mu\big[\psi^\top(s,u)\tilde{\theta}\big]^2}} \qquad (33) \\
&\leq\frac{-c\|\tilde{\theta}\|^2}{\sqrt{\gamma^2\mathbb{E}_\mu\big[\max_{u'\in\mathcal{U}}\big(\psi^\top(s',u')\tilde{\theta}\big)^2\big]/\mathbb{E}_\mu\big[\psi^\top(s,u)\tilde{\theta}\big]^2}+1} \qquad (34)
\end{aligned}$$

$$\leq \frac{-c\|\tilde{\theta}\|^2}{2-c}$$

which suggests that (6b) holds with $c_3 := c/[(2-c)L]$.

On the other hand, it follows for any $\tilde{\theta}$, $\tilde{\theta}'$ that

$$\left\| \frac{\partial W}{\partial \tilde{\theta}}\Big|_{\tilde{\theta}} - \frac{\partial W}{\partial \tilde{\theta}}\Big|_{\tilde{\theta}'} \right\| = \|\tilde{\theta} - \tilde{\theta}'\|$$

validating (6c) with $c_4 = 1$.

Verifying As. 3. Let $P^u_{ss'}$ be the transition probability of the Markov chain $\{S_k\}_{k\in\mathbb{N}}$ from states s to s' after taking action u; and let $p^{(n)}_{ss'}$ be the n-step transition probability from states s to s' following policy π. Define $X_k := (S_k, U_k, S_{k+1})$. It can be verified that $\{X_k\}_{k\in\mathbb{N}}$ is a Markov chain with state space $\mathcal{X} := \{x = (s,u,s') : s \in \mathcal{S},\, \pi(u|s) > 0, P^u_{ss'} > 0\} \subseteq \mathcal{S}\times\mathcal{U}\times\mathcal{S}$. Next, we show that $\{X_k\}$ is aperiodic and irreducible.

Consider two arbitrary states $x_i = (s_i, u_i, s_i')$, $x_j = (s_j, u_j, s_j') \in \mathcal{X}$. Since $\{S_k\}_k$ is irreducible, there exists an integer $n > 0$ such that $p^{(n)}_{s_i', s_j} > 0$. Using the definition of $\{X_k\}_k$, it follows that

$$p^{(n+1)}_{x_i,x_j} = p^{(n)}_{s_i's_j}\pi(u_j|s_j)P^{u_j}_{s_js_j'} > 0 \tag{35}$$

which corroborates that the Markov chain $\{X_k\}_k$ is irreducible; see e.g., [Levin and Peres, 2017, Ch. 1.3].

To prove that $\{X_k\}_k$ is aperiodic, we assume, for the sake of contradiction, that $\{X_k\}_k$ is periodic with period $d \geq 2$. As $\{X_k\}_k$ has been shown irreducible, it follows readily that every state in $\mathcal{X}$ has the same period of d. Hence, for each state $x = (s,u,s') \in \mathcal{X}$, it holds that $p^{(n+1)}_{x,x} = 0$ for all integers $n+1 > 0$ not divisible by d. Further, we deduce for any positive integer $(n+1)$ not divisible by d that

$$\begin{aligned} p^{(n+1)}_{s's'} &= \sum_{s\in\mathcal{S}} p^{(n)}_{s's}p^{(1)}_{ss'} = \sum_{s\in\mathcal{S}} p^{(n)}_{s's}\sum_{u\in\mathcal{U}}\pi(u|s)P^u_{ss'} \\ &= \sum_{s\in\mathcal{S}}\sum_{u\in\mathcal{U}} p^{(n+1)}_{xx} = 0 \end{aligned} \tag{36}$$

where the last two equalities arise from (35) and the periodicity assumption of $\{X_k\}_k$, respectively. It becomes evident from (36) that $\{S_k\}_k$ is periodic too, and its period is at least d. This clearly contradicts with the assumption that $\{S_k\}_k$ is aperiodic. Therefore, we conclude that the Markov chain $\{X_k\}_k$ is irreducible and aperiodic provided that $\{S_k\}_k$ is irreducible and aperiodic.

Consider now two arbitrary states $x_{T_0} = (s_{T_0}, u_{T_0}, s'_{T_0})$ and $x = (s,u,s') \in \mathcal{X}$. It follows that

$$\begin{aligned} &\Big\| \frac{1}{T}\sum_{k=T_0+1}^{T_0+T} \mathbb{E}\big[f(\tilde{\theta}, X_k)\big|X_{T_0} = x_{T_0}\big] - \bar{f}(\tilde{\theta})\Big\| \\ &= \Big\| \frac{1}{T}\sum_{k=T_0+1}^{T_0+T}\sum_{x\in\mathcal{X}} \big[p^{(k)}_{x_{T_0}x} - \mu(x)\big] f(\tilde{\theta}, x)\Big\| \qquad (37) \\ &= \Big\| \frac{1}{T}\sum_{k=T_0+1}^{T}\sum_{s,s'\in\mathcal{S}}\sum_{u\in\mathcal{U}} \big[p^{(k-1)}_{s'_{T_0}s}\pi(u|s)P^u_{ss'} - \mu(x)\big] \\ &\quad \times \psi(s,u)\big[R(s,u) + \gamma\max_{u'\in\mathcal{U}}\psi^\top(s',u')(\tilde{\theta}+\theta^*) \\ &\quad - \psi^\top(s,u)(\tilde{\theta}+\theta^*)\big]\Big\| \qquad (38) \\ &\leq \max_{(s,u,s')\in\mathcal{X}} \Big\|\psi(s,u)\big[R(s,u) + \gamma\max_{u'\in\mathcal{U}}\psi^\top(s',u')(\tilde{\theta}+\theta^*) \\ &\quad - \psi^\top(s,u)(\tilde{\theta}+\theta^*)\big]\Big\| \\ &\quad \times \frac{1}{T}\sum_{k=T_0+1}^{T}\sum_{x\in\mathcal{X}} \big|p^{(k-1)}_{s'_{T_0}s}\pi(u|s)P^u_{ss'} - \mu(x)\big| \\ &\leq (\|\tilde{\theta}\|+1)\times\frac{1}{T}\sum_{k=T_0+1}^{T} 2c\eta^{k-T_0-1} \qquad (39) \\ &\leq \frac{2c/(1-\eta)}{T}(\|\tilde{\theta}\|+1) \end{aligned}$$

where (37) is due to $\bar{f}(\tilde{\theta}) = \mathbb{E}_{X\sim\mu}[f(\tilde{\theta}, X)] = \sum_{x\in\mathcal{X}}\mu(x)f(\tilde{\theta}, x)$; equality (38) uses (22) and (35); and, (39) arises from the geometric mixing property of irreducible, aperiodic Markov chain $\{X_k\}_k$ [Levin and Peres, 2017, Thm. 4.9] as well as (29).

We have proved that As. 1-3 are satisfied by Q-learning with linear function approximation, provided that certain conditions on the sampling policy and function approximators hold. Hence, our finite-time error bound in Thm. 2 also holds for Q-learning with linear approximation.

5 CONCLUSIONS

In this paper, we provided a non-asymptotic analysis for a class of biased SA algorithms driven by a broad family of stochastic perturbations, which include as special cases e.g., i.i.d. random sequences of vectors and ergodic Markov chains. Taking a dynamical system viewpoint, our novel approach has been to design a multistep Lyapunov function that involves future iterates to control the gradient bias. We proved a general convergence result based on this Lyapunov function, and developed non-asymptotic bounds on the mean-square error of the iterate generated by the SA procedure to the equilibrium point of the associated ODE. Subsequently, we illustrated this general result by applying it to obtain a finite-time error bound for Q-learning with linear function approximation from data gathered along a single trajectory of a Markov chain. Our bound holds for Markov chains with general mixing rates and from any initial distribution.

Acknowledgments

This work was supported in part by NSF grants 1711471 and 1901134.

References

[Bach and Moulines, 2011] Bach, F. and Moulines, E. (2011). Non-asymptotic analysis of stochastic approximation algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pages 451–459.

[Beck and Srikant, 2012] Beck, C. L. and Srikant, R. (2012). Error bounds for constant step-size Q-learning. *Syst. & Control Lett.*, 61(12):1203–1208.

[Bertsekas and Tsitsiklis, 1996] Bertsekas, D. P. and Tsitsiklis, J. N. (1996). *Neuro-Dynamic Programming*, volume 5. Athena Scientific Belmont, MA.

[Bhandari et al., 2019] Bhandari, J., Russo, D., and Singal, R. (2019). A finite time analysis of temporal difference learning with linear function approximation. In *Conference on Learning Theory*, pages 1691–1692.

[Borkar, 2008] Borkar, V. S. (2008). *Stochastic Approximation: A Dynamical Systems Viewpoint*, volume 48. Cambridge, New York, NY.

[Chen et al., 2019] Chen, Z., Zhang, S., Doan, T. T., Maguluri, S. T., and Clarke, J.-P. (2019). Finite-sample analysis for Q-Learning with linear function approximation. *arXiv:1905.11425*.

[Dalal et al., 2018] Dalal, G., Szörényi, B., Thoppe, G., and Mannor, S. (2018). Finite sample analyses for TD(0) with function approximation. In *AAAI Conf. on Artificial Intelligence*, pages 6144–6152.

[Eryilmaz and Srikant, 2012] Eryilmaz, A. and Srikant, R. (2012). Asymptotically tight steady-state queue length bounds implied by drift conditions. *Queueing Systems*, 72(3-4):311–359.

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

[Khalil, 2002] Khalil, H. K. (2002). Nonlinear Systems; 3rd ed.

[Kushner and Yin, 2003] Kushner, H. and Yin, G. G. (2003). *Stochastic Approximation and Recursive Algorithms and Applications*, volume 35. Springer Science & Business Media.

[Levin and Peres, 2017] Levin, D. A. and Peres, Y. (2017). *Markov Chains and Mixing Times*, volume 107. American Mathematical Society.

[Melo et al., 2008] Melo, F. S., Meyn, S. P., and Ribeiro, M. I. (2008). An analysis of reinforcement learning with function approximation. In *International Conf. on Machine Learning*, pages 664–671.

[Mnih et al., 2015] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529.

[Nemirovski et al., 2009] Nemirovski, A., Juditsky, A., Lan, G., and Shapiro, A. (2009). Robust stochastic approximation approach to stochastic programming. *SIAM Journal on Optimization*, 19(4):1574–1609.

[Robbins and Monro, 1951] Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407.

[Srikant and Ying, 2019] Srikant, R. and Ying, L. (2019). Finite-time error bounds for linear stochastic approximation and TD learning. In *Conference on Learning Theory*, pages 1–11.

[Sutton and Barto, 2018] Sutton, R. S. and Barto, A. G. (2018). *Reinforcement Learning: An Introduction*. MIT press.

[Szepesvári, 1998] Szepesvári, C. (1998). The asymptotic convergence-rate of Q-learning. In *Advances in Neural Information Processing Systems*, pages 1064–1070.

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

[Tsitsiklis and Van Roy, 1997] Tsitsiklis, J. N. and Van Roy, B. (1997). An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5).

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

[Wang et al., 2019] Wang, G., Giannakis, G. B., and Chen, J. (2019). Learning ReLU networks on linearly separable data: Algorithm, optimality, and generalization. *IEEE Transactions on Signal Processing*, 67(9):2357–2370.

[Watkins, 1989] Watkins, C. J. C. H. (1989). *Learning from delayed rewards*. PhD thesis, King's College, Cambridge.