Constant Stepsize Q-learning: Distributional Convergence, Bias and Extrapolation

Yixuan Zhang
yzhang2554@wisc.edu
Department of Industrial and Systems Engineering
University of Wisconsin-Madison

Qiaomin Xie
qiaomin.xie@wisc.edu
Department of Industrial and Systems Engineering
University of Wisconsin-Madison

Abstract

Stochastic Approximation (SA) is a widely used algorithmic approach in various fields, including optimization and reinforcement learning (RL). Among RL algorithms, Q-learning is particularly popular due to its empirical success. In this paper, we study asynchronous Q-learning with constant stepsize, which is commonly used in practice for its fast convergence. By connecting the constant stepsize Q-learning to a time-homogeneous Markov chain, we show the distributional convergence of the iterates in Wasserstein distance and establish its exponential convergence rate. We also establish a Central Limit Theory for Q-learning iterates, demonstrating the asymptotic normality of the averaged iterates. Moreover, we provide an explicit expansion of the asymptotic bias of the averaged iterate in stepsize. Specifically, the bias is proportional to the stepsize up to higher-order terms, and we provide an explicit expression for the linear coefficient. This precise characterization of the bias allows the application of Richardson-Romberg (RR) extrapolation technique to construct a new estimate that is provably closer to the optimal Q function. Numerical results corroborate our theoretical finding on the improvement of the RR extrapolation method.

1 Introduction

Stochastic Approximation (SA) is a fundamental algorithmic paradigm in various fields, including machine learning, stochastic control and reinforcement learning (RL). SA uses recursive stochastic updates to solve fixed-point equations. One prominent example is the stochastic gradient descent (SGD) algorithm for optimizing an objective function (Lan, 2020). In RL, well-known algorithms such as Q-learning and TD-learning can be viewed as SA algorithms for solving Bellman equations (Bertsekas & Tsitsiklis, 1996). Classical SA theory suggests using diminishing stepsize, ensuring asymptotic convergence to the desired solution (Borkar, 2008). However, SA with constant stepsize is commonly used in practice due to its simplicity and faster convergence. In this case, SA iterates can be viewed as a *time-homogeneous Markov chain*. Adopting this perspective, a growing line of recent work establishes weak convergence of constant stepsize SA and characterizes the stationary distribution (Durmus et al., 2021a; Huo et al., 2023; Dieuleveut et al., 2020; Yu et al., 2021).

In this paper, we investigate constant-stepsize Q-learning, which is an important instance of *nonsmooth* SA with *Markovian* noise. Q-learning is a popular algorithm that has played a significant role in the empirical success of RL (Mnih et al., 2015). It aims to learn the optimal action-value function q^* by iteratively updating the estimator q_k from sample trajectories. Consequently, the iterations inherently involve *Markovian* noise resulting from the sampling process of a Markov chain under the behavior policy. Finite-time guarantees of Q-learning variants have been extensively studied (Tsitsiklis, 1994; Szepesvári, 1997; Even-Dar et al., 2003; Chen et al., 2021; Li et al., 2020). These non-asymptotic results provide *upper bounds* on either the mean squared error (MSE) $\mathbb{E}[\|q_k - q^*\|_\infty^2]$ or high probability ℓ_∞ error $\|q_k - q^*\|_\infty$ of the estimated Q-function q_k.

The main goal of this paper is to gain a more comprehensive understanding of the behavior of constant-stepsize Q-learning and its error decomposition. In the discounted setting, Q-learning aims to solve the fixed-point equation involving the Bellman operator, which is contractive in the *nonsmooth* ℓ_∞ norm. Hence Q-learning is an instance of SA with a *nonsmooth* operator and *Markovian* noise. Recently, non-asymptotic analysis of Markovian SA has been gaining attention (Bhandari et al., 2021; Srikant & Ying, 2019; Chen et al., 2020a; 2021; Huo et al., 2023). However, these results either concern linear SA or provide upper bounds on the error.

In this work, we study Q-learning through Markov chain theory, which allows us to quantify the fluctuations and bias of the iterates. Our results lead to a more precise characterization of the error $\|q_k - q^*\|$: the error is composed of a stochastic part $q_k - \mathbb{E}q_\infty^{(\alpha)}$ and a deterministic part (bias) $\mathbb{E}q_\infty^{(\alpha)} - q^*$, where $q_\infty^{(\alpha)}$ denotes the limit random vector of the Q-learning iterate $\{q_k\}$ with stepsize α. Our contributions are summarized as follows.

- **(Weak Convergence)** Viewing the joint process of the iterates $\{q_k\}_{k\geq 0}$ and data trajectory as a time-homogeneous Markov chain, we establish its distributional convergence in W_2, Wasserstein distance of order 2. Moreover, $\{q_k\}_{k\geq 0}$ converges to a limit random vector $q_\infty^{(\alpha)}$ *exponentially* fast due to the use of a constant stepsize α. We further prove a central limit theorem (CLT) for the iterates $\{q_k\}_{k\geq 0}$, thus proving the asymptotic normality of the averaging iterates.
- **(Bias Characterization)** We provide an explicit expansion of the deterministic bias $\mathbb{E}q_\infty^{(\alpha)} - q^*$ with respect to the stepsize α:

$$\mathbb{E}q_\infty^{(\alpha)} - q^* = \alpha B + \widetilde{\mathcal{O}}(\alpha^2)^1,$$

  where B is a vector *independent* of the stepsize α. Importantly, the leading term in bias scales linearly with α. Consequently, one can use the Richardson-Romberg (RR) extrapolation technique to reduce the bias and obtain an estimate closer to q^* with order-wise smaller bias $\widetilde{\mathcal{O}}(\alpha^2)$.
- For the stochastic part, $\mathbb{E}\|q_k - \mathbb{E}q_\infty^{(\alpha)}\|^2 \asymp \left\|\mathbb{E}q_k - \mathbb{E}q_\infty^{(\alpha)}\right\|^2 + \mathrm{Var}\left(q_k\right)$, we show that the optimization error $\|\mathbb{E}q_k - \mathbb{E}q_\infty^{(\alpha)}\|$ decays *exponentially* in k. The convergence rate cannot be obtained from the existing upper bound on $\mathbb{E}\|q_k - q^*\|^2$ or $\|q_k - q^*\|_\infty$, which does not vanish as $k \to \infty$. We further show that the variance $\mathrm{Var}\left(q_k\right)$ is of order $\mathcal{O}(1)$. By law of large numbers, one can use Polyak-Ruppert averaging to achieve a variance of order $\mathcal{O}(1/k)$. Consequently, for large k, the deviation between the averaged iterate and q^* for large k is dominated by the deterministic bias.

Compared with prior work focusing on MSE guarantee, we establish the distributional convergence, CLT and bias expansion of asynchronous Q-learning, which are completely new in this setting. On the technical side, we emphasize that Markovian noise and nonsmoothness of Q-learning operator bring additional challenges in showing weak convergence and bias characterization. The recent work of Huo et al. (2023) establishes the weak convergence of *linear* SA with Markovian noise, by analyzing the difference of two coupled iterates, which reduces to a special instance of linear SA. However, this observation does not apply to Q-learning due to the nonsmooth/nonlinear dynamic. We note that the very recent work by Lauand & Meyn (2023) studies nonlinear SA with Markovian noise, and a similar challenge arises. To this end, we develop a novel technique to analyze the difference of two coupled iterates; See Section 4.1 for a detailed discussion. For bias characterization, to deal with the nonsmooth operator, we employ a *local linearization* of the operator in the neighborhood of the optimal solution q^*. While local linearization has been explored in nonlinear SA literature, they mainly consider the asymptotic regime with diminishing stepsizes (Lee & He, 2020; Li et al., 2023b; Melo et al., 2008; Gopalan & Thoppe, 2023). We generalize this approach to characterize the dependence on the constant stepsize. It is worth noting that while the linear approximation component resembles similar behavior as linear SA (Huo et al., 2023), a precise characterization of the bias requires a careful analysis of the linear approximation error to show a proper higher order of α. We establish this result by analyzing the fourth moment of the iterates. Our techniques may be of independent interest and have the potential to be applied to the analysis of other nonsmooth/nonlinear SA algorithms.

[1] In this paper, $\widetilde{\mathcal{O}}$ denotes the variant of big $\mathcal{O}$ that ignores logarithmic order.

1.1 Related Work

We discuss closely related work and defer other related work to Section A in supplementary materials.

Q-learning. Recent work has been dedicated to understanding finite-time guarantees of Q-learning variants, with two main types of results: high probability bounds and mean (square) error bounds. For asynchronous Q-learning, as considered in this paper, Beck & Srikant (2012) provide the first result on MSE with constant stepsize and Chen et al. (2021) improve the result by a $|\mathcal{S}||\mathcal{A}|$ factor. The work by Li et al. (2023a) presents the best known high probability sample complexity. It is worth noting that these two types of bounds are not directly comparable, as discussed in Chen et al. (2021). Importantly, existing results are achieved either by rescaled linear stepsize $\alpha_k = a/(b+k)$ (Qu & Wierman, 2020; Chen et al., 2021) or by a carefully chosen constant stepsize based on the target accuracy (Chen et al., 2021; Li et al., 2023a). In contrast, we precisely characterize the convergence rate and the bias induced by any constant-stepsize α in a given range. Our explicit characterization enables the application of RR technique, leading to an estimate with reduced bias, while simultaneously enjoying the exponential convergence of the optimization error.

Some recent work also studies Polyak-Ruppert averaged Q-learning. Xie & Zhang (2022) and Li et al. (2023b) prove a functional CLT for *synchronous* Q-learning with constant stepsize and diminishing stepsize, respectively. In this work, we focus on asynchronous Q-learning involving Markovian data.

Stochastic approximation. There is a growing interest in investigating general SA with constant stepsize. Most work considers i.i.d. or martingale difference noise, and establishes finite-time guarantees for contractive/linear SA (Chen et al., 2020a; Mou et al., 2020; Durmus et al., 2021b) or SGD (Dieuleveut et al., 2020; Yu et al., 2021). Recent work investigates SA with Markovian noise, motivated by applications in RL (Srikant & Ying, 2019; Mou et al., 2021; Chen et al., 2022b).

Our results have some similarities to Dieuleveut et al. (2020, Proposition 2), Durmus et al. (2021b, Theorem 3) and Huo et al. (2023), in that we also study instances of SA with constant stepsizes through Markov chain theory. However, our setting is different from Dieuleveut et al. (2020); Durmus et al. (2021b), where they assume i.i.d. data. While the work (Huo et al., 2023) also considers Markovian noise, their focus is linear SA. In contrast, Q-learning involves nonsmooth update, which brings additional challenges on the analysis of convergence and bias as discussed earlier.

2 Preliminaries

Consider a discounted Markov decision process (MDP) defined by the tuple $(\mathcal{S}, \mathcal{A}, T, r, \gamma)$, where $\mathcal{S}$ and $\mathcal{A}$ are the (finite) state space and action space, $T : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel, $r : \mathcal{S} \times \mathcal{A} \to [0, r_{\max}]$ is the reward function, and $\gamma \in (0,1)$ is the discounted factor. At time $t \in \{0, 1, \dots\}$, the system is in state $s_t \in \mathcal{S}$; upon taking action $a_t \in \mathcal{A}$, the system transits to $s_{t+1} \in \mathcal{S}$ with probability $T(s_{t+1}|s_t, a_t)$ and generates a reward $r_t = r(s_t, a_t)$.

A stationary policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ maps each state to a distribution over the actions. For each policy π, the Q-function is defined as follows: $\forall s \in \mathcal{S}, \forall a \in \mathcal{A}$, $q^\pi(s,a) = \mathbb{E}\big[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a\big]$, where $a_k \sim \pi(\cdot|s_k)$ for all $k > 0$. An optimal policy π^* is the policy that maximizes $q^\pi(s,a)$ for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$ simultaneously (Bertsekas, 2017). We denote the associated Q functions as $q^* \equiv q^{\pi^*}$. Notably, given q^*, one can obtain the optimal policy $\pi^*(s) \in \arg\max_{a \in \mathcal{A}} Q^*(s,a)$.

Behavior policy. The goal of RL is to learn the optimal policy based on transition data from the system with unknown model (T, r). In this paper, we consider the off-policy setting, where we have access to a sample trajectory $\{s_k, a_k, r_k\}_{k \geq 0}$ generated by the MDP under a fixed *behavior policy* $\widetilde{\pi}$. Define $\mathcal{X} := \mathcal{S} \times \mathcal{A} \times \mathcal{S}$, and let $x_k = (s_k, a_k, s_{k+1})$. Note that when $\widetilde{\pi}$ is stationary, $\{x_k\}_{k \geq 0}$ forms a time-homogeneous Markov chain. We use $P = (p_{ij})$ to denote the corresponding transition matrix.

Assumption 1. *$\{x_k\}_{k \geq 0}$ is an irreducible and aperiodic Markov chain on a finite state $\mathcal{X}$ with stationary distribution $\mu_{\mathcal{X}}$. Also, the distribution of the initial state x_0 is $\mu_{\mathcal{X}}$.*

Assumption 1 is equivalent to assuming that the Markov chain $\{s_k, a_k\}_{k\geq 0}$ induced by the behavior policy $\widetilde{\pi}$ is uniformly ergodic with a unique stationary distribution $\mu_{\mathcal{S}}$ (Chen et al., 2021). This assumption is standard for analyzing off-policy Q-learning (Li et al., 2020; Qu & Wierman, 2020). It implies that $\{x_k\}_{k\geq 0}$ mixes geometrically fast to the stationary distribution $\mu_{\mathcal{X}}$ (Levin & Peres, 2017), and there exist $c \geq 0$ and $\rho \in (0,1)$ s.t. $\max_{x\in\mathcal{X}} \|p^k(x,\cdot) - \mu_{\mathcal{X}}(\cdot)\|_{TV} \leq c\rho^k$, where $p^k(x,\cdot)$ denotes the distribution of x_k at time k given $x_0 = x$.

To quantify how fast $\{x_k\}_{k\geq 0}$ mixes to a specified precision, we define the mixing time below.

Definition 1. $\forall \delta > 0$, *define the mixing time* $t_\delta := \min\{k \geq 0 : \max_{x\in\mathcal{X}} \|p^k(x,\cdot) - \mu_{\mathcal{X}}(\cdot)\|_{TV} \leq \delta\}$.

Under Assumption 1, we have $t_\alpha \leq \frac{\log(c/\rho)+\log(1/\alpha)}{\log(1/\rho)}$, which implies $\lim_{\alpha\to 0} \alpha^{m_1} t_{\alpha^{m_2}} = 0, \forall m_1, m_2 > 0$. We assume that $x_0 \sim \mu_{\mathcal{X}}$ to simplify some presentation. This assumption can be relaxed by adapting our result after the Markov chain $\{x_k\}_{k\geq 0}$ has almost mixed. The same assumption is considered in many previous works (Bhandari et al., 2021; Huo et al., 2023; Mou et al., 2021).

Q-learning. The Q-learning algorithm (Watkins & Dayan, 1992) is an iterative method for estimating the function Q^* based on the sample trajectory $\{s_k, a_k, r_k\}_{k\geq 0}$. It generates a sequence of Q-function estimate $\{q_k : \mathcal{S} \times \mathcal{A} \to \mathbb{R}\}_k$, according to the following recursion:

$$q_{k+1} = q_k + \alpha_k F(x_k, q_k), \tag{1}$$

where α_k is the stepsize. Here the operator $F : \mathcal{X} \times \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \mapsto \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, known as empirical Bellman operator, is defined as: $\forall (s,a) \neq (s_k, a_k), [F(x,q)](s,a) = 0$; and

$$[F(x,q)](s_k, a_k) = r(s_k, a_k) + \gamma \max_v q_k(s_{k+1}, v) - q_k(s_k, a_k).$$

In this paper, we focus on constant stepsize $\alpha_k \equiv \alpha > 0$. We use superscript $q_k^{(\alpha)}$ to emphasize the dependence on the stepsize α, but omit it when it is clear from the context.

We state some properties of Q-learning. (1) By the boundedness of reward, there exists a constant $q_{\max}$ such that $\|q_k\|_\infty \leq q_{\max}, \forall k$. (2) Denote the expected operator of F by $\bar{F}(q) := \mathbb{E}_{x\sim\mu_{\mathcal{X}}}[F(x,q)]$. It has been shown that $\bar{F}(q) + q$ is a β-contraction mapping w.r.t. $\|\cdot\|_\infty$ (Chen et al., 2021), where $\beta = 1 - (1-\gamma)\min_{(s,a)} \mu_{\mathcal{S}}(s,a)$. Recall that $\mu_{\mathcal{S}}$ is the stationary distribution of Markov chain $\{s_k, a_k\}_{k\geq 0}$. By Assumption 1, $\min_{(s,a)} \mu_{\mathcal{S}}(s,a) > 0$, thus $\beta < 1$. (3) Crucially, the iterates $\{q_k\}$ generated by Q-learning is not a Markov chain. On the other hand, we can see that the joint process $\{x_k, q_k\}_{k\geq 0}$ is a Markov chain on the state space $\mathcal{X} \times \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$.

Part of our results on Q-learning (cf. Theorem 3) requires the following assumption.

Assumption 2. *The optimal policy π is unique. That is, $\exists \Delta > 0$ such that for $\forall s \in \mathcal{S}$, $q^*(s, a_s^*) - q^*(s,a) \geq 2\Delta, \forall a \neq a_s^*$, where $a_s^* := \arg\max_a q^*(s,a)$ denotes the optimal action for each state s.*

Similar conditions have been considered in prior work on the analysis of Q-learning variants (Devraj & Meyn, 2017; Li et al., 2023b). Assumption 2 implies that the operator in (1) can be approximated by local linearization around q^* and high-order approximation error, which leads to our precise characterization of the bias induced by constant stepsize.

3 Main Results

In this section, we present our main results. In Section 3.1, we show that joint data-iterates $\{x_k, q_k\}_{k\geq 0}$ converges to a unique limit distribution exponentially fast. We show a central limit theorem (CLT) for the iterates $\{q_k\}_{k\geq 0}$ in Section 3.2. We then precisely characterize the relationship between the limit and the stepsize in Section 3.3. Furthermore, we investigate the implications of these results for Polyak-Ruppert averaging and Richardson-Romberg extrapolation in Section 3.4.

3.1 Stationary Distribution and Convergence Rate

Note that the Q-learning iterate $\{q_k\}_{k\geq 0}$ is not a Markov chain by itself, as its dynamic depends on the Markovian data $\{x_k\}_{k\geq 0}$. To show the distributional convergence of $\{q_k\}_{k\geq 0}$, we consider the joint process $\{x_k, q_k\}_{k\geq 0}$, which can be cast as a time-homogeneous Markov chain. We will analyze the convergence of this Markov chain using the Wasserstein 2-distance, which is defined as follows for any distributions μ and ν in $\mathcal{P}_2(\mathbb{R}^d)$, the space of square-integrable distributions on $\mathbb{R}^d$:

$$W_2(\mu,\nu) = \inf_{\xi\in\Pi(\mu,\nu)} \Big(\int_{\mathbb{R}^d} \|u-v\|_\infty^2 \,\mathrm{d}\xi(u,v) \Big)^{1/2} = \inf \Big\{ \big(\mathbb{E}\big[\, \|\theta-\theta'\|_\infty^2 \,\big] \big)^{\frac{1}{2}} : \mathcal{L}(\theta) = \mu, \mathcal{L}\left(\theta'\right) = \nu \Big\},$$

where $\mathcal{L}(\theta)$ denote the distribution of θ and $\Pi(\mu,\nu)$ is the set of all joint distributions in $\mathcal{P}_2(\mathbb{R}^d\times\mathbb{R}^d)$ with marginal distributions μ and ν. To analyze the Markov chain $\{x_k, q_k\}_{k\geq 0}$, we define the extended Wasserstein 2-distance. Let $\bar{d}\left((x,\theta),(x',\theta')\right) := \sqrt{\mathbb{1}\left\{x \neq x'\right\} + \|\theta-\theta'\|_\infty^2}$, which defines a metric on $\mathcal{X}\times\mathbb{R}^d$. The extended Wasserstein 2-distance w.r.t. the metric $\bar{d}$ is defined as follows:

$$\bar{W}_2(\bar{\mu},\bar{\nu}) = \inf \Big\{ \big(\mathbb{E}[\bar{d}(z,z')^2] \big)^{1/2} : \mathcal{L}(z) = \bar{\mu}, \mathcal{L}\left(z'\right) = \bar{\nu} \Big\}, \quad \forall \bar{\mu},\bar{\nu} \in \mathcal{P}_2(\mathcal{X}\times\mathbb{R}^d). \tag{2}$$

We show that the Markov chain $\{x_k, q_k\}_{k\geq 0}$ converges in $\bar{W}_2$ to a unique stationary distribution, *geometrically* fast, as stated in the following Theorem.

Theorem 1 (Weak Convergence)**.** *Suppose that Assumption 1 holds, and the stepsize α for Q-learning (1) satisfies $\alpha t_\alpha \leq c_0 \frac{(1-\beta)^2}{\log(|\mathcal{S}||\mathcal{A}|)}$ for some constant c_0.*

1. *Under all initial distribution of q_0, the chain $\{x_k, q_k\}_{k\geq 0}$ converges in $\bar{W}_2$ to a a unique limit $(x_\infty, q_\infty) \sim \bar{\mu}$. Moreover, we have $\mathrm{Var}(q_\infty) \leq c_Q \frac{\log(|\mathcal{S}||\mathcal{A}|)}{(1-\beta)^2} \alpha t_\alpha$, where $c_Q = 912e\left(3\|q^*\|_\infty + r_{\max}\right)$.*
2. *$\bar{\mu}$ is the unique stationary distribution of the Markov chain $\{x_k, q_k\}_{k\geq 0}$.*
3. *Let $\mu := \mathcal{L}(q_\infty)$ be the second marginal of $\bar{\mu}$. Let $\eta = 1-(1-\beta)\,\alpha/2$. For all $k \geq t_\alpha$, we have*
$$W_2^2\left(\mathcal{L}(q_k),\mu\right) \leq 24\eta^{k-t_\alpha}\left(\mathbb{E}\left[\|q_0\|_\infty^2\right] + \mathbb{E}\left[\|q_\infty\|_\infty^2\right]\right). \tag{3}$$

Theorem 1 states that the Markov chain $\{x_k, q_k\}_{k\geq 0}$ admits a unique stationary distribution. Recall that under Assumption 1, for all $m_1, m_2 > 0$, we have $\lim_{\alpha\to 0}\alpha^{m_1} t_{\alpha^{m_2}} = 0$. Therefore, there always exists a sufficiently small stepsize α such that the condition in Theorem 1 holds.

We remark that the convergence results of Theorem 1 cannot be obtained from the existing error bounds on Q-learning. For example, the sharpest high probability bound on ℓ_∞ error scales as $\|q_k - q^*\|_\infty \lesssim (1-\rho)^k\|q_0 - q^*\|_\infty + \mathcal{O}(\sqrt{\alpha})$, where $\rho\in(0,1)$ (Li et al., 2020). Another type of upper bound is on the MSE that scales as $\mathbb{E}[\|q_k - q^*\|_\infty^2] \lesssim (1-(1-\beta)\alpha/2)^{k-t_\alpha}\|q_0 - q^*\|_\infty^2 + \mathcal{O}(\alpha t_\alpha)$ (Chen et al., 2021). Both upper bounds imply that the sequence eventually falls in a neighbor of the optimal solution q^* and the initial condition is forgotten exponentially fast. However, these results do not imply the distributional convergence of the sequence $\{q_k\}_{k\geq 0}$ or its convergence rate in the W_2 metric.

We would like to highlight the techniques employed to prove Theorem 1. A standard method to prove the convergence of a Markov chain is to verify the irreducibility and the Lyapunov drift condition (Meyn & Tweedie, 2009), as used in prior work on SA (Borkar et al., 2021) and SGD (Yu et al., 2021). However, this method requires a strong condition on the randomness of the Markov chain dynamics, which typically does not hold in Q-learning. Instead, we draw inspiration from recent work on constant-stepsize SA (Dieuleveut et al., 2020; Huo et al., 2023), and prove weak convergence by showing the convergence in W_2 distance through coupling arguments. We remark that the coupling argument in our proof is more involved due to the nonsmoothness of the update operator F. We sketch the proof outline in Section 4.1 and defer the complete proof to Section B.

A direct consequence of the convergence in W_2 metric is the convergence of the first two moments. We can also obtain explicit convergence rates from Theorem 1, as detailed in the following corollary.

Corollary 1. *Under the setting of Theorem 1, for all $k \geq t_\alpha$,*

$$\|\mathbb{E}[q_k - q_\infty]\|_\infty^2 \leq C \cdot \left(1 - (1-\beta)\,\alpha/2\right)^{k-t_\alpha}, \quad \left\|\mathbb{E}\left[q_k q_k^\top\right] - \mathbb{E}\left[q_\infty q_\infty^\top\right]\right\|_\infty \leq C' \cdot \left(1 - (1-\beta)\,\alpha/2\right)^{\frac{k-t_\alpha}{2}},$$

where C and C' are constants independent of α and k.

3.2 Central Limit Theorem

Building on the convergence result, we establish a CLT for $\{q_k\}_{k\geq 0}$. Here we define $S_n = \sum_{k=0}^{n-1}\left(q_k - \mathbb{E}[q_\infty]\right)$ and $Y_n(t) = n^{-\frac{1}{2}} S_{\lfloor nt \rfloor}$. Let $\mathcal{D} = \mathcal{D}[0,1]$ denote the Skorokhod space, which is a separable and complete function space under some proper metrics (Prokhorov, 1956).

Theorem 2 (CLT)**.** *Under the setting of Theorem 1, $\boldsymbol{\Sigma} := \lim_{n\to\infty} n^{-1}\mathbb{E}_\pi\left(S_n S_n^\top\right)$ exists, and for $\bar{\mu}$-almost every point (x_0, q_0), the sequence $\{S_n/\sqrt{n}\}_{n\geq 0}$ converge in distribution to the Gaussian distribution $\mathcal{N}(\mathbf{0}, \boldsymbol{\Sigma})$. Furthermore, the process $(Y_n(t))_{0\leq t\leq 1}$ converges weakly to $\left(\boldsymbol{\Sigma}^{\frac{1}{2}}\boldsymbol{B}(t)\right)_{0\leq t\leq 1}$ on the Skorokhod space $D[0,1]$, where $\boldsymbol{B} = (\boldsymbol{B}(t))_{t\geq 0}$ is the standard Brownian motion.*

Theorem 2 states that the average of Q-learning iterates is asymptotically normally distributed around the expected value of the unique stationary distribution. Establishing such a CLT is important for uncertainty quantification and statistical inference (Li et al., 2023b). A similar result has been established for synchronous Q-learning with constant stepsize (Xie & Zhang, 2022), where the data used in each iteration is *independently* generated. It is worth highlighting that one key step in Xie & Zhang (2022) uses the Kantorovich–Rubinstein theorem (Edwards, 2011) defined on a Wasserstein distance with single-step contraction. However, such result does not hold in our setting due to Markovian data. To this end, we use the result in Theorem 1 and ergodicity of $\{x_k\}_{k\geq 0}$ to establish CLT. The detailed proof is provided in Section C.

3.3 Bias Characterization

Under constant stepsize α, Theorem 1 asserts that the convergence of $q_k^{(\alpha)}$ to $q_\infty^{(\alpha)}$, which is of distribution μ. Therefore, the estimates $q_k^{(\alpha)}$ of Q-learning with constant stepsize do not converge to a point, but oscillate around the mean $\mathbb{E}[q_\infty^{(\alpha)}]$. Here we would like to quantify the *bias*, i.e., the deviation of the mean $\mathbb{E}[q_\infty^{(\alpha)}]$ from the optimal solution q^*. One of our main contributions is to provide an *explicit* expansion of the bias $\mathbb{E}[q_\infty^{(\alpha)}] - q^*$ in the step-size α.

Theorem 3 (Bias Characterization)**.** *Suppose that Assumptions 1 and 2 hold and $\alpha \leq \alpha_0$ for some α_0. Then the following holds for a vector $B = B(r, \gamma, P)$ independent of α:*

$$\mathbb{E}\left[q_\infty\right] = q^* + \alpha B + \mathcal{O}(\alpha^2 + \alpha^2 t_{\alpha^2}^2). \tag{4}$$

Theorem 3 states that the asymptotic bias of Q-learning can be decomposed into a linear term and a higher order term of α. We emphasize that our bias characterization of the linear dependence on α is *exact*. As discussed in the previous subsection, existing results are typically in the form of an upper bound on the bias. Specifically, the high probability upper bound on ℓ_∞ error (Li et al., 2020) implies a bias of $\mathcal{O}(\sqrt{\alpha})$. In contrast, our analysis reveals a refined result with $\alpha B + \widetilde{\mathcal{O}}(\alpha^2)$ bias.

One key step in the proof of Theorem 3 is to calculate $\mathbb{E}[F(x_\infty, q_\infty) \mid x_\infty = i], \forall i \in \mathcal{X}$. For linear SA, this step is straightforward. However, for asynchronous Q learning, the operator F is nonlinear and not even smooth, making the analysis more complicated. In our proof, we develop a local linearization method which can bridge the gap between nonlinear SA and linear SA. We outline the proof of Theorem 3 in Section 4.2. The complete proof is provided in Section D.

We remark that the coefficient B of the linear term is independent of α. It depends only on the underlying MDP and the behavior policy. One can find an explicit expression of B in the proof (cf. Equation (31)). Importantly, for the special case where the associated data sequence $\{x_k\}_{k\geq 0}$ is i.i.d., we have $B = \mathbf{0}$. However, the bias term $\mathcal{O}(\alpha^2 + \alpha^2 t_{\alpha^2}^2)$ still remains, due to the nonlinearity of the Q-learning operator. This should be contrasted with the LSA where the bias vanishes under i.i.d. data (Huo et al., 2023). In general, the existence of bias implies that the mean of the sequence $\{q_k\}_{k\geq 0}$ limit deviates from the optimal solution q^*. Therefore, averaging the iterates q_k does not eliminate the bias. However, thanks to the independence of B on α, we can leverage an extrapolation technique to reduce the bias, as detailed in Section 3.4.

3.4 Polyak-Ruppert Tail Average and Richardson-Romberg Extrapolation

We now utilize the bias expansion result Theorem 3 to study the behavior of Q-learning when combined with Polyak-Ruppert (PR) average and Richardson-Romberg extrapolation.

Polyak-Ruppert Averaging. The celebrated PR averaging procedure (Ruppert, 1988; Polyak & Juditsky, 1992) can reduce the estimator variance and accelerate the convergence rate. Here we consider the PR tail averaging (Jain et al., 2018), defined as follows with a burn-in period k_0:

$$\bar{q}_{k_0,k} := \frac{1}{k-k_0}\sum_{t=k_0}^{k-1} q_t, \quad \text{for } k \geq k_0+1. \tag{5}$$

The following corollary provides non-asymptotic results for the first and second moments of $\bar{q}_{k_0,k}$.

Corollary 2. *Under the setting of Theorem 3, the tail-averaged iterates (5) satisfy the following:* $\forall k > k_0 \geq t_{\alpha^2}$*:*

$$\mathbb{E}\left[\bar{q}_{k_0,k}\right] - q^* = \alpha B + \mathcal{O}(\alpha^2 + \alpha^2 t_{\alpha^2}^2) + \mathcal{O}\Big(\frac{1}{\alpha(k-k_0)}\exp\big(-\frac{\alpha(1-\beta)k_0}{4}\big)\Big), \tag{6}$$

$$\mathbb{E}\big[\|\bar{q}_{k_0,k}-q^*\|^2\big] = \underbrace{\alpha^2 B' + \mathcal{O}(\alpha^3 + \alpha^3 t_{\alpha^2}^2)}_{\text{asymptotic squared bias}} + \underbrace{\mathcal{O}\Big(\frac{1}{(k-k_0)\alpha}\Big)}_{\text{variance}} + \underbrace{\mathcal{O}\Big(\frac{1}{(k-k_0)^2\alpha^2}\exp\big(-\frac{\alpha(1-\beta)k_0}{4}\big)\Big)}_{\text{optimization error}}, \tag{7}$$

where B and B' are independent of α.

The proof is provide in Section E. For simplicity, let us consider the case $k_0 = k/2$ and discuss the mean squared distance between the averaged-iterate $\bar{q}_{k/2,k}$ and q^*. The MSE can be decomposed into three parts: (1) the asymptotic squared bias term $\|\mathbb{E}[\bar{q}_{\infty/2,\infty} - q^*]\|^2$ is independent of k and averaging; (2) the variance of $\bar{q}_{k/2,k}$ scales as $1/k$; (3) and the optimization error $\|\mathbb{E}[\bar{q}_{\infty/2,\infty} - \bar{q}_{k/2,k}]\|^2$ decays to 0 geometrically fast. Importantly, the larger the stepsize α is, the faster the variance and optimization error decay.

Richardson-Romberg Extrapolation. Given the explicit expansion of the bias in stepsize α (cf. Theorem 3), we can leverage the RR extrapolation technique from numerical analysis (Gautschi, 2011) to reduce the bias. Specifically, consider running two Q-learning recursions using the *same* data stream $\{x_k\}_{k\geq 0}$, but with different stepsizes α and 2α. Denote by $\bar{q}_{k_0,k}^{(\alpha)}$ and $\bar{q}_{k_0,k}^{(2\alpha)}$ the corresponding tail-averaged iterates. The corresponding RR extrapolated iterates are given by

$$\widetilde{q}_{k_0,k}^{(\alpha)} = 2\bar{q}_{k_0,k}^{(\alpha)} - \bar{q}_{k_0,k}^{(2\alpha)}. \tag{8}$$

With $k_0, k \to \infty$, Theorems 1 and 3 imply that $\widetilde{q}_{k_0,k}^{(\alpha)}$ converges to $2q_\infty^{(\alpha)} - q_\infty^{(2\alpha)}$, which has a bias

$$2\mathbb{E}q_\infty^{(\alpha)} - \mathbb{E}q_\infty^{(2\alpha)} - q^* = 2\left(\alpha B + \mathcal{O}(\alpha^2 + \alpha^2 t_{\alpha^2}^2)\right) - \left(2\alpha B + \mathcal{O}(\alpha^2 + \alpha^2 t_{\alpha^2}^2)\right) = \mathcal{O}(\alpha^2 + \alpha^2 t_{\alpha^2}^2).$$

Compared with $q_\infty^{(\alpha)}$ and $q_\infty^{(2\alpha)}$, the extrapolated sequence reduces the bias by a factor of α. We formally state the result in the following corollary, which quantifies the non-asymptotic behavior of the first two moments of extrapolated sequence $\{\widetilde{q}_{k_0,k}^{(\alpha)}\}_{k\geq 0}$. The proof is provided in Section F.

Corollary 3. *Under the setting of Theorem 3, the RR extrapolated iterates (8) with stepsizes α and 2α satisfy the following for all $k > k_0 \geq t_{\alpha^2}$:*

$$\mathbb{E}\left[\widetilde{q}_{k_0,k}^{(\alpha)}\right] - q^* = \mathcal{O}(\alpha^2 + \alpha^2 t_{\alpha^2}^2) + \mathcal{O}\left(\frac{1}{\alpha(k-k_0)}\exp\left(-\frac{\alpha(1-\beta)k_0}{4}\right)\right), \tag{9}$$

$$\mathbb{E}\left[\|\widetilde{q}_{k_0,k}^{(\alpha)} - q^*\|^2\right] \in \mathcal{O}\left(\alpha^4 + \alpha^4 t_{\alpha^2}^4\right) + \mathcal{O}\left(\frac{1}{(k-k_0)\alpha}\right) + \mathcal{O}\left(\frac{1}{(k-k_0)^2\alpha^2}\exp\left(-\frac{\alpha(1-\beta)k_0}{4}\right)\right). \tag{10}$$

Let us compare the MSE bounds (7) on the PR-averaged iterates and the extrapolated iterates (10). Note that the asymptotic squared bias is reduced from $\mathcal{O}(\alpha^2)$ to $\mathcal{O}(\alpha^4)$ by RR extrapolation! Meanwhile, RR extrapolation still enjoys similar decaying rates of variance and optimization error. We remark that the RR procedure involves the computation of two parallel Q-learning iterates, using either the same or different data sequences. This makes the RR procedure inherently parallelizable, offering potential performance improvements when implemented on parallel computing architectures.

4 Proof Outlines

4.1 Proof Outline for Theorem 1 on Weak Convergence

Here we outline the proof of the existence of the limit distribution, which is the most challenging part. Note that the space $\mathcal{P}(\mathcal{X}\times\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|})$ endowed with our extended Wasserstein 2-distance $\bar{W}_2$ is a Polish space (Villani et al., 2009, Theorem 6.18). We will show that $\sum_{k=0}^{\infty}\bar{W}_2^2\left(\mathcal{L}\left(x_k,q_k\right),\mathcal{L}\left(x_{k+1},q_{k+1}\right)\right) < \infty$, thus the sequence $\{x_k,q_k\}_{k\geq 0}$ forms a Cauchy sequence. This result implies the existence of the limit distribution, by the fact that all Cauchy sequences converge in a Polish space.

The key step involves coupling through the construction of two Markov chains, $\{x_k^{[1]},q_k^{[1]}\}_{k\geq 0}$ and $\{x_k^{[2]},q_k^{[2]}\}_{k\geq 0}$, which share the same underlying data stream $\{x_k^{[1]}\}_{k\geq 0} = \{x_k^{[2]}\}_{k\geq 0} = \{x_k\}_{k\geq 0}$. We observe that the iterates difference $w_k := q_k^{[1]} - q_k^{[2]}$ exhibits the following *double-recursion* that involves both w_k and $q_k^{[1]}$: $w_{k+1}(s_k,a_k) = (1-\alpha)w_k(s_k,a_k)+\alpha\gamma\big(\max_a q_k^{[1]}(s_{k+1},a)-\max_a q_k^{[2]}(s_{k+1},a)\big)$.

Proposition 1. *Under the setting of Theorem 1, the following bound holds with $\eta = 1-(1-\beta)\alpha/2$:*

$$\mathbb{E}\left[\,\|w_k\|_\infty^2\,\right] \leq 12\mathbb{E}\left[\|w_0\|_\infty^2\right]\eta^{k-t_\alpha}, \qquad \forall k \geq t_\alpha.$$

For linear SA (Huo et al., 2023), the iterates difference w_k is a *single-recursion* that only involves w_k, which reduces to a special case of linear SA. In constrast, the nonsmoothness of Q-learning leads to the double recursion of w_t, which brings an additional challenge in analyzing the convergence of w_t. Our key idea for proving Proposition 1 is to exploit the fact that the difference between two max operators can be lower bounded by the minimum of the difference, and upper bounded by the maximum of the difference. We thus construct two new sequences that serve as lower and upper bounds on $\{w_k\}_{k\geq 0}$, and prove that both sequences decay geometrically fast to 0, which immediately implies a geometric decay of $\{w_k\}_{k\geq 0}$. Next, by carefully choosing the initial distribution of $q_0^{[2]}$, we can ensure that $(x_k,q_k^{[2]}) \overset{d}{=} (x_{k+1},q_{k+1}^{[1]})$. Consequently, $\bar{W}_2^2\big(\mathcal{L}\big(x_k,q_k\big),\mathcal{L}\big(x_{k+1},q_{k+1}\big)\big) \to 0$ geometrically fast, which allows us to show $\sum_{k=0}^{\infty}\bar{W}_2^2\big(\mathcal{L}\big(x_k,q_k\big),\mathcal{L}\big(x_{k+1},q_{k+1}\big)\big) < \infty$.

4.2 Proof Outline for Theorem 3 on Bias Expansion

A crucial technique employed in the proof of Theorem 3 is the linearization of the non-smooth operator $F(x,q)$. Specifically, for a fixed x, we linearize $F(x,q)$ around the optimal solution q^* according to the following proposition.

Proposition 2. *There exists a function $G_{q^*}: \mathcal{X} \mapsto \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$ s.t. for any $(x,q) \in \mathcal{X}\times\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$,*

$$F(x,q) = F(x,q^*) + (G_{q^*}(x) - I_d)(q-q^*) + R(x,q), \tag{11}$$

where $d = |\mathcal{S}||\mathcal{A}|$, $\|R(x,q)\|_\infty = \mathcal{O}\left(\|q-q^*\|_\infty^4\right)$, *and* $\mathbb{E}_{x\sim\mu_\mathcal{X}}[G_{q^*}(x)]$ *does not have eigenvalue of 1.*

We next provide a finite-time upper bound on the fourth moment of the error, which shows that the remaining term $R(x,q)$ in Proposition 2 is of a higher order of α. We remark that existing non-asymptotic results for Q-learning are limited to the first moment and second moment of the error.

Proposition 3. *Suppose that Assumption 1 holds and* $\alpha \in (0, \alpha_0)$ *for some* α_0*. Then*

$$\mathbb{E}[\|q_k - q^*\|_\infty^4] \le b_1(1-\alpha(1-\gamma)^2)^{k-t_{\alpha^2}} + b_2\alpha^2 + b_3\alpha^2 t_{\alpha^2}^2, \qquad \forall k \ge t_{\alpha^2}, \tag{12}$$

where b_1*,* b_2 *and* b_3 *are constants independent of* α*.*

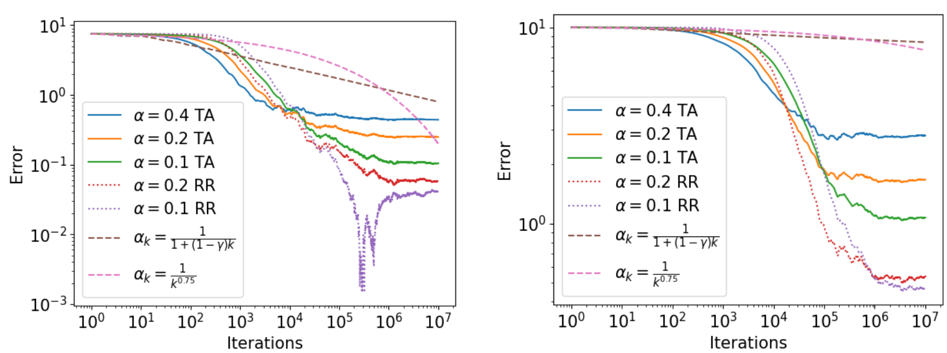
To deal with the nonsmooth ℓ_∞ norm, we consider a Generalized Moreau Envelope $M(\cdot)$, which has been used to analyze MSE $\|\cdot\|_\infty^2$ (Chen et al., 2021). We derive the bound for $M(\cdot)^2$, which provides a bound for $\|\cdot\|_\infty^4$. We defer the complete proof of Proposition 2 and 3 to Section D.1.

Note that the first term on RHS of (12) decays geometrically in k, whereas the remaining two terms are independent of k. Consequently, as $k \to \infty$, the upper bound is of order $\mathcal{O}\left(\alpha^2 + \alpha^2 t_{\alpha^2}^2\right)$. Therefore, the RHS of equation (11) can be viewed as a combination of a linear operator and a high-order remaining term $R(x,q)$ of order $\mathcal{O}\left(\alpha^2 + \alpha^2 t_{\alpha^2}^2\right)$. We then can analyze the dynamic of $\{x_k, q_k\}_{k\ge 0}$ as a combination of linear SA with a remaining term.

5 Numerical Experiments

We consider two MDPs: the first example is a 1×3 Gridword with two actions (left/right); the second one is a classical 4×4 Gridworld with the slippery mechanism in Frozen-Lake, and four actions (left/up/right/down). For both MDPs, the discounted factor is $\gamma = 0.9$ and the Markovian data $\{x_k\}_{k\ge 0}$ is generated from a uniformly random behavior policy. We defer details of the reward function and the transition kernel for the MDPs to Section G.

We run Q-learning with constant stepsize $\alpha \in \{0.1, 0.2, 0.4\}$. We also consider two commonly used diminishing stepsizes: a rescaled linear stepsize $\alpha_k = 1/\big(1+(1-\gamma)k\big)$ (Qu & Wierman, 2020; Chen et al., 2020b) and a polynomial stepsize $\alpha_k = 1/k^{0.75}$. The results are illustrated in Figure 1(a) and 1(b). We plot the ℓ_1-norm error $\|\bar{q}_{k/2,k}^{(\alpha)} - q^*\|_1$ for the tail-averaged (TA) iterates $\bar{q}_{k/2,k}^{(\alpha)}$, the RR extrapolated iterates $\widetilde{q}_k^{(\alpha)}$ with stepsizes α and 2α, and iterates with diminishing stepsizes.



(a) 1×3 Gridworld.

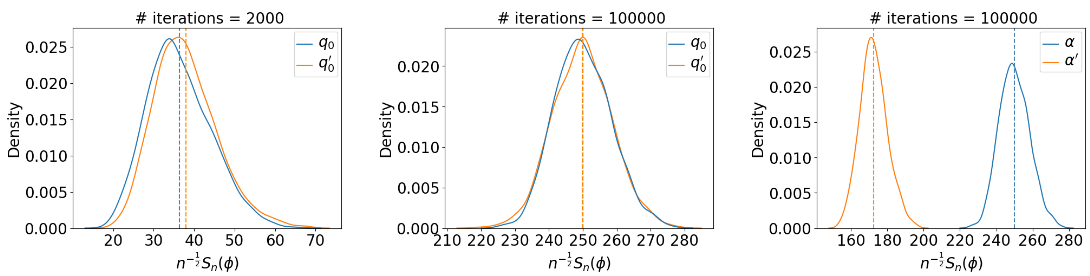
(b) 4×4 Gridworld.

Figure 1: Errors of tail-averaged (TA) iterates and RR extrapolated iterates with different stepsizes.

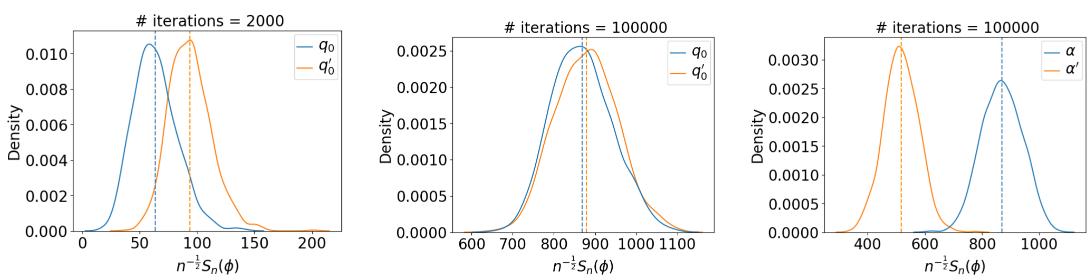
We first observe that the larger the stepsize α, the faster it converges, as implied by Corollary 2. We note that the final TA error, which corresponds to the asymptotic bias, is approximately proportional to the stepsize, as indicated by the roughly equal space between three TA lines in the log-scale plots. Moreover, RR extrapolated iterates reduce the bias, which can be observed by comparing, e.g, the solid orange line (TA with $\alpha = 0.2$) and the dotted red line (RR with $\alpha = 0.2$ and 0.4). These results are consistent with Corollary 3. Furthermore, the TA and RR-extrapolated iterates with constant stepsizes enjoy significantly faster initial convergence than those with diminishing stepsizes. A general choice of diminishing stepsize is of the form $\alpha_k = a/(b + k^c)$, where a, b and c are hyperparameters. Tuning the best hyperparameters for diminishing stepsize is generally more challenging than a single parameter for constant stepsize.

We also perform experiments on MDPs with linear function approximation. We observe similar behaviors of TA iterates and RR extrapolated iterates as in the tabular case; see Section G for details.

Our next set of experiments demonstrates the asympotitic normality of Q-learning averaging iterates. We consider different initializations q_0, different number of iterations n and different stepsizes $\alpha = 0.4$ and $\alpha' = 0.2$. We plot the density of $n^{-1/2}S_n(\phi) = n^{-1/2}\sum_{k=1}^{n}\phi(q_k)$ with the test function $\phi(q_k) = \|q_k - q^*\|_\infty$ for 1000 Monte Carlo runs. Figures (2(a),2(d)) show the effect of different initializations (blue, orange) on the normality after a moderate number of iterations $n = 2\times 10^3$. We observe that the impact of initialization becomes negligible in the long run from Figures (2(b),2(e)), and the distribution is approximately Gaussian. Lastly, Figures (2(c),2(f)) show the impact of stepsize on the normality. In particular, a larger stepsize α (blue) induces a larger mean. These observations are consistent with our Theorems 2 and 3.



(a) (b) (c)



(d) (e) (f)

Figure 2: First and second rows correspond to 1×3 Gridword and 4×4 Gridworld, respectively. Figures (2(a), 2(d)) and (2(b), 2(e)) show the density of $n^{-\frac{1}{2}}S_n(\phi)$ with different initializations for different number of iterations. Figure (2(c), 2(f)) show the density with different stepsizes.

6 Conclusions

In this work, we provide a comprehensive study of asynchronous Q-learning with constant stepsizes, through the framework of Markov chain theory. We establish the distributional convergence of the iterates, characterize the convergence rate, and prove a central limit theorem for the averaged iterates. Our convergence results lead to a refined characterization of the error. In particular, the explicit expansion of the asymptotic bias w.r.t. stepsize α allows one to use the RR extrapolation for bias reduction. There are several interesting directions one can take to extend our work. First, our CLT, together with our bias characterization and the Richardson-Romberg de-biasing scheme, allow one to create confidence intervals for the output of the Q-learning algorithms. Second, our current results require the assumption of local linearity in the neighborhood of the optimal solution. Extending our analysis without this assumption is a direction worth pursuing.

Acknowledgement

Q. Xie and Y. Zhang are supported in part by National Science Foundation (NSF) grants CNS-1955997 and EPCN-2339794. Y. Zhang is also supported in part by NSF Award DMS-2023239.

References

Mohammad Gheshlaghi Azar, Remi Munos, Mohammad Ghavamzadeh, and Hilbert Kappen. Speedy q-learning. In *Advances in neural information processing systems*, 2011.

Yu Bai, Tengyang Xie, Nan Jiang, and Yu-Xiang Wang. Provably efficient q-learning with low switching cost. *Advances in Neural Information Processing Systems*, 32, 2019.

Carolyn L Beck and Rayadurgam Srikant. Error bounds for constant step-size q-learning. *Systems & control letters*, 61(12):1203–1208, 2012.

D. Bertsekas and J.N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996. ISBN 9781886529106. URL https://books.google.com/books?id=WxCCQgAACAAJ.

D.P. Bertsekas. *Dynamic Programming and Optimal Control*. Athena Scientific, 2017.

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. *Operations Research*, 69(3):950–973, May 2021. ISSN 0030-364X. doi: 10.1287/opre.2020.2024. URL https://doi.org/10.1287/opre.2020.2024.

Pascal Bianchi, Walid Hachem, and Sholom Schechtman. Convergence of constant step stochastic gradient descent for non-smooth non-convex functions. *Set-Valued and Variational Analysis*, 30 (3):1117–1147, 2022.

Julius R. Blum. Approximation methods which converge with probability one. *The Annals of Mathematical Statistics*, 25(2):382 – 386, 1954. doi: 10.1214/aoms/1177728794. URL https://doi.org/10.1214/aoms/1177728794.

Vivek Borkar, Shuhang Chen, Adithya Devraj, Ioannis Kontoyiannis, and Sean Meyn. The ode method for asymptotic statistics in stochastic approximation and reinforcement learning. *arXiv preprint arXiv:2110.14427*, 2021.

Vivek S. Borkar. *Stochastic Approximation: A Dynamical Systems Viewpoint*. Hindustan Book Agency Gurgaon, 2008. ISBN 978-93-86279-38-5. doi: 10.1007/978-93-86279-38-5. URL https://link.springer.com/book/10.1007/978-93-86279-38-5.

Vivek S. Borkar and Sean P. Meyn. The O.D.E. method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2):447–469, Jan 2000. ISSN 0363-0129.

Mario Bravo and Roberto Cominetti. Stochastic fixed-point iterations for nonexpansive maps: Convergence and error bounds. *SIAM Journal on Control and Optimization*, 62(1):191–219, 2024.

Qi Cai, Zhuoran Yang, Jason D Lee, and Zhaoran Wang. Neural temporal-difference learning converges to global optima. *Advances in Neural Information Processing Systems*, 32, 2019.

Zaiwei Chen, Siva Theja Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. Finite-sample analysis of contractive stochastic approximation using smooth convex envelopes. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 8223–8234. Curran Associates, Inc., 2020a. URL https://proceedings.neurips.cc/paper/2020/file/5d44ee6f2c3f71b73125876103c8f6c4-Paper.pdf.

Zaiwei Chen, Siva Theja Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. Finite-sample analysis of contractive stochastic approximation using smooth convex envelopes. *Advances in Neural Information Processing Systems*, 33:8223–8234, 2020b.

Zaiwei Chen, Siva Theja Maguluri, Sanjay Shakkottai, and Karthikeyan Shanmugam. A lyapunov theory for finite-sample guarantees of asynchronous q-learning and td-learning variants. *arXiv preprint arXiv:2102.01567*, 2021.

Zaiwei Chen, Shancong Mou, and Siva Theja Maguluri. Stationary behavior of constant stepsize sgd type algorithms: An asymptotic characterization. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 6(1), 02 2022a. doi: 10.1145/3508039. URL https://doi.org/10.1145/3508039.

Zaiwei Chen, Sheng Zhang, Thinh T Doan, John-Paul Clarke, and Siva Theja Maguluri. Finite-sample analysis of nonlinear stochastic approximation with applications in reinforcement learning. *Automatica*, 146:110623, 2022b.

Adithya M Devraj and Sean P Meyn. Fastest convergence for q-learning. *arXiv preprint arXiv:1707.03770*, 2017.

Aymeric Dieuleveut, Alain Durmus, and Francis Bach. Bridging the gap between constant step size stochastic gradient descent and Markov chains. *The Annals of Statistics*, 48(3):1348 – 1382, 2020. doi: 10.1214/19-AOS1850. URL https://doi.org/10.1214/19-AOS1850.

Simon S Du, Jason D Lee, Gaurav Mahajan, and Ruosong Wang. Agnostic *q*-learning with function approximation in deterministic systems: Near-optimal bounds on approximation error and sample complexity. *Advances in Neural Information Processing Systems*, 33:22327–22337, 2020.

Alain Durmus, Pablo Jiménez, Éric Moulines, and SAID Salem. On riemannian stochastic approximation schemes with fixed step-size. In *International Conference on Artificial Intelligence and Statistics*, pp. 1018–1026. PMLR, 2021a.

Alain Durmus, Eric Moulines, Alexey Naumov, Sergey Samsonov, Kevin Scaman, and Hoi-To Wai. Tight high probability bounds for linear stochastic approximation with fixed step-size. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 30063–30074. Curran Associates, Inc., 2021b. URL https://proceedings.neurips.cc/paper/2021/file/fc95fa5740ba01a870cfa52f671fe1e4-Paper.pdf.

Alain Durmus, Eric Moulines, Alexey Naumov, Sergey Samsonov, and Hoi-To Wai. On the stability of random matrix product with markovian noise: Application to linear stochastic approximation and td learning. In *Conference on Learning Theory*, pp. 1711–1752. PMLR, 2021c.

Alain Durmus, Eric Moulines, Alexey Naumov, and Sergey Samsonov. Finite-time high-probability bounds for Polyak-Ruppert averaged iterates of linear stochastic approximation, 2022. URL https://arxiv.org/abs/2207.04475.

David A Edwards. On the kantorovich–rubinstein theorem. *Expositiones Mathematicae*, 29(4): 387–398, 2011.

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

Walter Gautschi. *Numerical analysis*. Springer Science & Business Media, 2011.

Aditya Gopalan and Gugan Thoppe. Demystifying approximate value-based rl with ϵ-greedy exploration: A differential inclusion view, 2023.

Dongyan Huo, Yudong Chen, and Qiaomin Xie. Bias and extrapolation in markovian linear stochastic approximation with constant stepsizes. In *Abstract Proceedings of the 2023 ACM SIGMETRICS International Conference on Measurement and Modeling of Computer Systems*, pp. 81–82, 2023.

Prateek Jain, Sham M. Kakade, Rahul Kidambi, Praneeth Netrapalli, and Aaron Sidford. Parallelizing stochastic gradient descent for least squares regression: Mini-batching, averaging, and model misspecification. *The Journal of Machine Learning Research*, 18(223):1–42, 2018. URL http://jmlr.org/papers/v18/16-595.html.

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is q-learning provably efficient? *Advances in neural information processing systems*, 31, 2018.

Harold J. Kushner and G. George Yin. *Stochastic Approximation and Recursive Algorithms and Applications*. Stochastic Modelling and Applied Probability. Springer, New York, NY, USA, 2nd edition, 2003. ISBN 9780387008943. doi: 10.1007/b97441. URL https://link.springer.com/book/10.1007/b97441.

Chandrashekar Lakshminarayanan and Csaba Szepesvári. Linear stochastic approximation: How far does constant step-size and iterate averaging go? In Amos Storkey and Fernando Perez-Cruz (eds.), *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pp. 1347–1355. PMLR, 09–11 Apr 2018. URL https://proceedings.mlr.press/v84/lakshminarayanan18a.html.

G. Lan. *First-order and Stochastic Optimization Methods for Machine Learning*. Springer Series in the Data Sciences. Springer International Publishing, 2020. ISBN 9783030395681. URL https://books.google.com/books?id=7dTkDwAAQBAJ.

Caio Kalil Lauand and Sean Meyn. Bias in stochastic approximation cannot be eliminated with averaging. In *2022 58th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1–4. IEEE, 2022.

Caio Kalil Lauand and Sean Meyn. The curse of memory in stochastic approximation. In *2023 62nd IEEE Conference on Decision and Control (CDC)*, pp. 7803–7809, 2023. URL https://doi.org/10.1109/CDC49753.2023.10383986.

Donghwan Lee and Niao He. A unified switching system perspective and convergence analysis of q-learning algorithms. *Advances in Neural Information Processing Systems*, 33:15556–15567, 2020.

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

Gen Li, Yuting Wei, Yuejie Chi, Yuantao Gu, and Yuxin Chen. Sample complexity of asynchronous q-learning: Sharper analysis and variance reduction. *Advances in neural information processing systems*, 33:7031–7043, 2020.

Gen Li, Changxiao Cai, Yuxin Chen, Yuantao Gu, Yuting Wei, and Yuejie Chi. Tightening the dependence on horizon in the sample complexity of q-learning. In *International Conference on Machine Learning*, pp. 6296–6306. PMLR, 2021.

Gen Li, Changxiao Cai, Yuxin Chen, Yuting Wei, and Yuejie Chi. Is q-learning minimax optimal? a tight sample complexity analysis. *Operations Research*, 2023a.

Xiang Li, Wenhao Yang, Jiadong Liang, Zhihua Zhang, and Michael I Jordan. A statistical analysis of polyak-ruppert averaged q-learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 2207–2261. PMLR, 2023b.

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

Sean P. Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge Mathematical Library. Cambridge University Press, Cambridge, 2nd edition, 2009. ISBN 9780521731829. doi: 10.1017/CBO9780511626630. URL https://doi.org/10.1017/CBO9780511626630.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533, 2015.

Wenlong Mou, Chris Junchi Li, Martin J. Wainwright, Peter L. Bartlett, and Michael I. Jordan. On linear stochastic approximation: Fine-grained Polyak-Ruppert and non-asymptotic concentration. In Jacob Abernethy and Shivani Agarwal (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 2947–2997. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/mou20a.html.

Wenlong Mou, Ashwin Pananjady, Martin J. Wainwright, and Peter L. Bartlett. Optimal and instance-dependent guarantees for Markovian linear stochastic approximation, 2021. URL https://arxiv.org/abs/2112.12770.

James R Norris. *Markov chains*. Number 2. Cambridge university press, 1998.

Boris T. Polyak and Anatoli B. Juditsky. Acceleration of stochastic approximation by averaging. *SIAM Journal on Control and Optimization*, 30(4):838–855, Jul 1992. ISSN 0363-0129. doi: 10.1137/0330046. URL https://doi.org/10.1137/0330046.

Yu V Prokhorov. Convergence of random processes and limit theorems in probability theory. *Theory of Probability & Its Applications*, 1(2):157–214, 1956.

Guannan Qu and Adam Wierman. Finite-time analysis of asynchronous stochastic approximation and q-learning. In *Conference on Learning Theory*, pp. 3185–3205. PMLR, 2020.

David Ruppert. Efficient estimations from a slowly convergent Robbins-Monro process. Technical report, Cornell University, February 1988.

Aaron Sidford, Mengdi Wang, Xian Wu, Lin Yang, and Yinyu Ye. Near-optimal time and sample complexities for solving markov decision processes with a generative model. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018. URL https://proceedings.neurips.cc/paper_files/paper/2018/file/bb03e43ffe34eeb242a2ee4a4f125e56-Paper.pdf.

Rayadurgam Srikant and Lei Ying. Finite-time error bounds for linear stochastic approximation and TD learning. In Alina Beygelzimer and Daniel Hsu (eds.), *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pp. 2803–2830. PMLR, 25–28 Jun 2019. URL https://proceedings.mlr.press/v99/srikant19a.html.

Csaba Szepesvári. The asymptotic convergence-rate of Q-Learning. In M. Jordan, M. Kearns, and S. Solla (eds.), *Advances in Neural Information Processing Systems*, volume 10. MIT Press, 1997. URL https://proceedings.neurips.cc/paper/1997/file/cd0dce8fca267bf1fb86cf43e18d5598-Paper.pdf.

John N. Tsitsiklis. Asynchronous stochastic approximation and Q-Learning. *Machine Learning*, 16 (3):185–202, Sep 1994. ISSN 1573-0565. doi: 10.1023/A:1022689125041. URL https://doi.org/10.1023/A:1022689125041.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Martin J Wainwright. Stochastic approximation with cone-contractive operators: Sharp ell_∞-bounds for q-learning. *arXiv preprint arXiv:1905.06265*, 2019a.

Martin J Wainwright. Variance-reduced q-learning is minimax optimal. *arXiv preprint arXiv:1906.04697*, 2019b.

Christopher J. C. H. Watkins and Peter Dayan. Q-Learning. *Machine Learning*, 8(3):279–292, May 1992. ISSN 1573-0565. doi: 10.1007/BF00992698. URL https://doi.org/10.1007/BF00992698.

Wentao Weng, Harsh Gupta, Niao He, Lei Ying, and R Srikant. The mean-squared error of double q-learning. *Advances in Neural Information Processing Systems*, 33:6815–6826, 2020.

Chuhan Xie and Zhihua Zhang. A statistical online inference approach in averaged stochastic approximation. *Advances in Neural Information Processing Systems*, 35:8998–9009, 2022.

Pan Xu and Quanquan Gu. A finite-time analysis of q-learning with neural network function approximation. In *International Conference on Machine Learning*, pp. 10555–10565. PMLR, 2020.

Lu Yu, Krishnakumar Balasubramanian, Stanislav Volgushev, and Murat A. Erdogdu. An analysis of constant step size SGD in the non-convex regime: Asymptotic normality and bias. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 4234–4248. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper/2021/file/21ce689121e39821d07d04faab328370-Paper.pdf.

A Additional Related Work

Q-learning. An increasing volume of recent work has been dedicated to understanding finite-time guarantees of Q-learning variants. There are two types of results on the error of the estimate q_k: high probability bounds and mean (square) error bounds. For classical asynchronous Q-learning, as considered in this paper, Beck & Srikant (2012) provide the first result on MSE with constant stepsize and Chen et al. (2021) improve the result by at least a $|\mathcal{S}||\mathcal{A}|$ factor. The work by Li et al. (2023a) presents the best known high probability sample complexity. It is worth noting that these two types of bounds are not directly comparable, as discussed in Chen et al. (2021). Importantly, these results are achieved either by rescaled linear stepsize $\alpha_k = a/(b+k)$ (Qu & Wierman, 2020; Chen et al., 2021) or by a carefully chosen constant stepsize based on the target accuracy (Chen et al., 2021; Li et al., 2023a; Bravo & Cominetti, 2024). Contrasting with these findings, our results provide a precise characterization of the convergence rate as well as the bias induced by any constant-stepsize α in a given range. Our explicit characterization enables the application of RR technique, leading to an estimate with reduced bias, while simultaneously enjoying the exponential convergence of the optimization error.

Some recent work also studies Polyak-Ruppert averaged Q-learning. Xie & Zhang (2022) and Li et al. (2023b) prove a functional central limit theorem for the averaged iterates of *synchronous* Q-learning with constant stepsize and diminishing stepsize, respectively. In contrast, we focus on asynchronous Q-learning involving Markovian data.

Stochastic approximation. There is a growing interest in investigating general SA with constant stepsize. Most work along this line considers i.i.d. or martingale difference noise, and establishes finite-time guarantees for contractive/linear SA (Chen et al., 2020a; Mou et al., 2020; Durmus et al., 2021b) or SGD (Dieuleveut et al., 2020; Yu et al., 2021). Recent work investigates constant-stepsize SA with Markovian noise, motivated by applications in RL. For linear SA, the work by Srikant & Ying (2019) provides finite-time upper bounds on the MSE. Mou et al. (2021) study LSA with PR averaging and presents instance-dependent MSE upper bounds with tight dimension dependence. The work by Durmus et al. (2021c) shows a finite-time upper bound for the p-th of LSA iterate on general state space. The paper Lauand & Meyn (2022) shows that LSA with Markovian noise admits a biass that can not be eliminated by averaging. The work Huo et al. (2023) establishes the distributional convergence of LSA iterates, and provides an explicit asymptotic expansion of the bias in stepsize. Going beyond LSA, the work Chen et al. (2022b) considers contractive SA under a strong monotone condition and provides finite-time upper bound on the MSE.

Our results have some similarities to Dieuleveut et al. (2020, Proposition 2), Durmus et al. (2021b, Theorem 3) and Huo et al. (2023), in that we also study instances of SA with constant stepsizes through Markov chain theory. However, our setting is different from Durmus et al. (2021b, Theorem 3) as the sampling process in RL naturally induces Markovian noise, whereas they consider i.i.d. data. While the work Huo et al. (2023) also considers Markovian noise, their focus is on linear SA. In contrast, Q-learning involves nonsmooth update, which brings additional challenges on the analysis of convergence and bias. In particular, for convergence proof, the difference between two coupled LSA iterates can be reformulated as an LSA; however, this is not the case for Q-learning, which requires a novel analysis for the coupled iterates. For the bias analysis, we employ a local linearization method to decompose the Q-learning operator into a linear term and a remaining approximation term. While the technique for LSA (Huo et al., 2023) can be used to analyse the linear part, it is highly nontrivial to show the remaining term is of higher order dependence on α. We establish this result by analyzing the fourth moment of the iterates. Our techniques may be of independent interest and have the potential to be applied to the analysis of other nonsmooth/nonlinear SA algorithms.

Q-learning Earlier work established the asymptotic convergence of Q-learning algorithm with diminishing stepsize (Tsitsiklis, 1994; Szepesvári, 1997). Over the past few years, an increasing volume of work has been dedicated to understanding finite-time guarantees of Q-learning in various scenarios. from tabular setting (Beck & Srikant, 2012; Chen et al., 2021; Qu & Wierman, 2020;

Wainwright, 2019a; Li et al., 2023a) to function approximation (Chen et al., 2022b; Xu & Gu, 2020; Du et al., 2020; Cai et al., 2019). In this paper we focus on the classical asynchrounous Q-learning. There is another variant of Q-learning that concerns an *synchronous* setting, where all state-action pairs are updated simultaneously at each step. This setting requires access to a simulator, which generates independent samples for each state-action pair. For synchronous Q-learning, the best-known sample complexity for mean error bound is $\widetilde{O}(SA(1-\gamma)^{-5}\epsilon^{-2})$ (Wainwright, 2019a; Chen et al., 2020b). The paper Li et al. (2021) provides the state-of-art high probability sample complexity $\widetilde{O}(\frac{SA}{(1-\gamma)^4\epsilon^2})$. In this paper, we focus on the classical asynchronous Q-learning which updates only a single state-action pair upon each observation. The Markovian noise inherited in the asynchronous model makes it considerably more challenging to analyze than the synchronous case.

We also note that there are other lines of work focusing on Q-learning variants that aim to accelerate convergence and improve sample complexity, such as variance-reduced Q-learning (Li et al., 2020; Wainwright, 2019b; Sidford et al., 2018), speedy Q-learning (Azar et al., 2011) and double Q-learning (Weng et al., 2020). Another direction considers Q-learning with sophisticated exploration strategies, with an emphasis on regret bound (Jin et al., 2018; Bai et al., 2019). Regret is a metric fundamentally different from finite-sample bounds, and techniques for these two types of guarantees are quite different. A comparison with these results is beyond the scope of this paper.

Stochastic approximation. There is a rich literature on the study of SA. Classical SA theory mainly focuses on the asymptotic convergence (Kushner & Yin, 2003; Borkar, 2008; Borkar & Meyn, 2000; Blum, 1954), typically assuming a diminishing stepsize sequence. More recent studies have shifted the focus to non-asymptotic results. In particular, there is a growing interest in investigating general SA and SGD algorithms with constant stepsize. Most work along this line considers SA or SGD with i.i.d. or martingale difference noise, and establishes finite-time bounds. The paper Chen et al. (2020a) considers contractive SA and presents an upper bounds on the MSE. Lakshminarayanan & Szepesvári (2018) analyzes linear SA (LSA) and establishes finite-time upper and lower bounds on the MSE. The work Mou et al. (2020) refines these results, providing tight bounds with the optimal dependence on problem-specific constants as well as a central limit theorem (CLT) for the averaged iterates. There are also some recent studies developing new bounds on random matrix products to analyze LSA: Durmus et al. (2021b) establishes tight concentration bounds of LSA, and Durmus et al. (2022) extends these bounds to LSA with iterate averaging. In the context of SGD, the work in Dieuleveut et al. (2020) considers strongly convex and smooth functions. They prove that the iterates converge to a unique stationary distribution by Markov chain theory. Subsequent work generalizes this result to non-convex and non-smooth functions with quadratic growth (Yu et al., 2021), and proves asymptotic normality of the averaged SGD iterates. The work Chen et al. (2022a) exams the limit of the stationary distribution as stepsize goes to zero. All these results are established under the i.i.d. noise setting. Additionally, Bianchi et al. (2022) explores SGD for non-smooth non-convex functions with martingale difference noise, and establishes the weak convergence of the iterates to the set of critical points of the objective function.

B Proof of Theorem 1

In this section, we provide the proof of Theorem 1. The first part of the proof, Section B.1, involves coupling through the construction of two iterates of Q-learning. Using the result of this step, we then establish the existence and uniqueness of the stationary distribution for the joint Markov chain $(x_k, q_k)_{k\geq 0}$ (part 1 and 2 of Theorem 1) in Section B.2. We prove the convergence rate (part 3 of Theorem 1) in Section B.3.

B.1 Coupling and Geometric Convergence

We construct a pair of coupled Markov chains, $(x_k, q_k^{[1]})_{k\geq 0}$ and $(x_k, q_k^{[2]})_{k\geq 0}$, defined as

$$\begin{aligned} q_{k+1}^{[1]}(s_k,a_k) &= q_k^{[1]}(s_k,a_k) + \alpha\left(r(s_k,a_k) + \gamma \max_a q_k^{[1]}(s_{k+1},a) - q_k^{[1]}(s_k,a_k)\right), \\ q_{k+1}^{[2]}(s_k,a_k) &= q_k^{[2]}(s_k,a_k) + \alpha\left(r(s_k,a_k) + \gamma \max_a q_k^{[2]}(s_{k+1},a) - q_k^{[2]}(s_k,a_k)\right). \end{aligned} \tag{13}$$

Here $(q_k^{[1]})_{k\geq 0}$ and $(q_k^{[2]})_{k\geq 0}$ are two iterates generated by the Q-learning algorithm, coupled by sharing the underlying data stream $(x_k)_{k\geq 0}$. We assume that the initial iterates $q_0^{[1]}$ and $q_0^{[2]}$ may depend on each other and on x_0, but are independent of $(x_k)_{k\geq 1}$ given x_0.

Define the iterates difference as $w_k := q_k^{[1]} - q_k^{[2]}$. Note that the dynamic for $\{w_k\}_{k\geq 0}$ can be formulated as follows:

$$w_{k+1}(s_k,a_k) = (1-\alpha)w_k(s_k,a_k) + \alpha\gamma\left(\max_a q_k^{[1]}(s_{k+1},a) - \max_a q_k^{[2]}(s_{k+1},a)\right).$$

We can exploit the dynamic of $\{w_k\}_{k\geq 0}$ to establish its convergence rate, as stated in Proposition 1. The proof of Proposition 1 is deferred to Section B.4.

When $\alpha t_\alpha \leq c_0 \frac{(1-\beta)^2}{\log(|\mathcal{S}||\mathcal{A}|)}$, we can apply Proposition 1 to bound the square of W_2 distance between $q_k^{[1]}$ and $q_k^{[2]}$ as follows: for all $k \geq t_\alpha$,

$$\begin{aligned} W_2^2\left(\mathcal{L}\left(q_k^{[1]}\right), \mathcal{L}\left(q_k^{[2]}\right)\right) &\overset{\text{(i)}}{\leq} \bar{W}_2^2\left(\mathcal{L}\left(x_k, q_k^{[1]}\right), \mathcal{L}\left(x_k, q_k^{[2]}\right)\right) \\ &\overset{\text{(ii)}}{\leq} \mathbb{E}\left[\left\|q_k^{[1]} - q_k^{[2]}\right\|_\infty^2\right] \\ &= \mathbb{E}\left[\|w_k\|_\infty^2\right] \\ &\overset{\text{(iii)}}{\leq} 12\mathbb{E}\left[\|w_0\|_\infty^2\right]\left(1 - \frac{(1-\beta)\,\alpha}{2}\right)^{k-t_\alpha}, \end{aligned} \tag{14}$$

where the inequality (i) follows from the definition of W_2 and $\bar{W}_2$; the inequality (ii) holds as the $\bar{W}_2$ is defined by an infimum as in equation (2); the inequality (iii) follows from applying Proposition 1.

Therefore, $W_2^2\left(\mathcal{L}\left(q_k^{[1]}\right), \mathcal{L}\left(q_k^{[2]}\right)\right)$ decays geometrically. We will use this result in the next subsection to prove that $(x_k, q_k)_{k\geq 0}$ converges to a unique stationary distribution.

B.2 Existence and Uniqueness of Stationary Distribution

Additional Notations. Throughout the proof, we denote the discrete metric $d_0(x_0', x_0) := \mathbb{1}\{x_0' \neq x_0\}$, which is used in the definition of extended Wasserstein distance (2). Part of our analysis uses the reversed Markov chains. An implication of Assumption 1 is that the chain $\{x_k\}_{k\geq 0}$ running backward in time is also a Markov chain (Norris, 1998), with transition kernel $\hat{P} = (\hat{p}_{ij})$ given by $\mu_{\mathcal{X}}(j)\hat{p}_{ji} = \mu_{\mathcal{X}}(i)p_{ij}$.

Note that equation (14) always holds for any joint distribution of initial iterates $(x_0, q_0^{[1]}, q_0^{[2]})$. After fixing an arbitrarily chosen distribution of $(x_0, q_0^{[1]})$, we need to carefully choose the conditional distribution of $q_0^{[2]}$ to ensure that $(x_k, q_k^{[2]}) \overset{d}{=} (x_{k+1}, q_{k+1}^{[1]})$ holds for all $k \geq 0$, where $\overset{d}{=}$ denotes equality in distribution. Recall that $\hat{P}$ represents the transition kernel for the time-reversed Markov chain of $(x_k)_{k\geq 0}$, and the initial distribution of x_0 is assumed to be mixed already. Given a specific x_0, we sample x_{-1} from $\hat{P}(\cdot \mid x_0)$. Additionally, we use $q_{-1}^{[2]}$ to denote a random variable that satisfies

$q_{-1}^{[2]} \overset{d}{=} q_0^{[1]}$ and is independent of $(x_k)_{k\geq 0}$. Finally, we set $q_0^{[2]}$ as

$$q_0^{[2]} = q_{-1}^{[2]} + \alpha F(x_{-1}, q_{-1}^{[2]}). \tag{15}$$

By the property of time-reversed Markov chains, we have $(x_k)_{k\geq -1} \overset{d}{=} (x_k)_{k\geq 0}$. Given that $q_{-1}^{[2]} \overset{d}{=} q_0^{[1]}$ and $q_{-1}^{[2]}$ is independent with $(x_k)_{k\geq -1}$, we can prove $(x_k, q_k^{[2]}) \overset{d}{=} (x_{k+1}, q_{k+1}^{[1]})$ for all $k \geq 0$ by comparing the dynamic of $(q_k^{[1]})_{k\geq 0}$ and $(q_k^{[2]})_{k\geq 0}$ as given in equations (13) and (15).

We thus have for all $k \geq t_\alpha$:

$$\begin{aligned}\bar{W}_2^2\left(\mathcal{L}\left(x_k, q_k^{[1]}\right), \mathcal{L}\left(x_{k+1}, q_{k+1}^{[1]}\right)\right) &= \bar{W}_2^2\left(\mathcal{L}\left(x_k, q_k^{[1]}\right), \mathcal{L}\left(x_k, q_k^{[2]}\right)\right) \\ &\leq 12\mathbb{E}\left[\|w_0\|_\infty^2\right]\left(1 - \frac{(1-\beta)\,\alpha}{2}\right)^{k-t_\alpha},\end{aligned}$$

where the second inequality follows from equation (14). It follows that

$$\begin{aligned}&\sum_{k=0}^{\infty} \bar{W}_2^2\left(\mathcal{L}\left(x_k, q_k^{[1]}\right), \mathcal{L}\left(x_{k+1}, q_{k+1}^{[1]}\right)\right) \\ &\leq \sum_{k=0}^{t_\alpha - 1} \bar{W}_2^2\left(\mathcal{L}\left(x_k, q_k^{[1]}\right), \mathcal{L}\left(x_{k+1}, q_{k+1}^{[1]}\right)\right) + 12\mathbb{E}\left[\|w_0\|_\infty^2\right]\sum_{k=0}^{\infty}\left(1 - \frac{(1-\beta)\,\alpha}{2}\right)^k \\ &< \infty,\end{aligned}$$

where the last step holds since $\frac{(1-\beta)\alpha}{2} \in (0,1)$. Consequently, $(\mathcal{L}(x_k, q_k^{[1]}))_{k\geq 0}$ forms a Cauchy sequence with respect to the metric $\bar{W}_2$. Since the space $\mathcal{P}_2(\mathcal{X} \times \mathbb{R}^d)$ endowed with $\bar{W}_2$ is a Polish space, every Cauchy sequence converges (Villani et al., 2009, Theorem 6.18). Furthermore, convergence in Wasserstein 2-distance also implies weak convergence (Villani et al., 2009, Theorem 6.9). Therefore, we conclude that the sequence $(\mathcal{L}(x_k, q_k^{[1]}))_{k\geq 0}$ converges weakly to a limit distribution $\bar{\mu} \in \mathcal{P}_2(\mathcal{X} \times \mathbb{R}^d)$.

Next, we show that $\bar{\mu}$ is independent of the initial iterate distribution of $q_0^{[1]}$, when x_0 is initialized from its unique stationary distribution $\mu_{\mathcal{X}}$. Suppose there exists another sequence $(x_k, \widetilde{q}_k^{[1]})_{k\geq 0}$ with a different initial distribution that converges to a limit $\widetilde{\mu}$. By triangle inequality, we have

$$\bar{W}_2(\bar{\mu}, \tilde{\mu}) \leq \bar{W}_2\left(\bar{\mu}, \mathcal{L}\left(x_k, q_k^{[1]}\right)\right) + \bar{W}_2\left(\mathcal{L}\left(x_k, q_k^{[1]}\right), \mathcal{L}\left(x_k, \widetilde{q}_k^{[1]}\right)\right) + \bar{W}_2\left(\mathcal{L}\left(x_k, \widetilde{q}_k^{[1]}\right), \tilde{\mu}\right) \overset{k\to\infty}{\longrightarrow} 0.$$

Note that the last step holds since $\bar{W}_2\left(\mathcal{L}\left(x_k, q_k^{[1]}\right), \mathcal{L}\left(x_k, \widetilde{q}_k^{[1]}\right)\right) \overset{k\to\infty}{\longrightarrow} 0$ by equation (14). We thus have $\bar{W}_2(\bar{\mu}, \widetilde{\mu}) = 0$, which implies the uniqueness of the limit $\bar{\mu}$.

Moreover, we will show that the unique limit distribution μ is also a stationary distribution for the Markov chain $(x_k, q_k)_{k\geq 0}$, as stated in the following lemma.

Lemma 1. *Let $(x_k, q_k)_{k\geq 0}$ and $(x'_k, q'_k)_{k\geq 0}$ be two trajectories of Q-learning iterates, where $\mathcal{L}(x_0, q_0) = \bar{\mu}$ and $\mathcal{L}(x'_0, q'_0) \in \mathcal{P}_2(\mathcal{X} \times \mathbb{R}^d)$ is arbitrary. Under Assumption 1 we have*

$$\bar{W}_2^2\left(\mathcal{L}\left(x_1, q_1\right), \mathcal{L}(x'_1, q'_1)\right) \leq \rho \bar{W}_2^2\left(\mathcal{L}\left(x_0, q_0\right), \mathcal{L}(x'_0, q'_0)\right),$$

where the quantity $\rho := \max\left(1 + 2(\alpha R_{\max} + \alpha\gamma q_{\max})^2, 2(1+\alpha\gamma)^2\right)$ is independent of $\mathcal{L}(x'_0, q'_0)$. In particular, for any $k \geq 0$, if we set $\mathcal{L}(x'_0, q'_0) = \mathcal{L}(x_k, q_k)$, then

$$\bar{W}_2^2\left(\mathcal{L}\left(x_1, q_1\right), \mathcal{L}(x_{k+1}, q_{k+1})\right) \leq \rho \bar{W}_2^2\left(\bar{\mu}, \mathcal{L}(x_k, q_k)\right).$$

Proof of Lemma 1. We prove this lemma by coupling the two processes $(x_k, q_k)_{k\geq 0}$ and $(x'_k, q'_k)_{k\geq 0}$ such that

$$\begin{aligned}\bar{W}_2^2\left(\mathcal{L}\left(x_0, q_0\right), \mathcal{L}(x'_0, q'_0)\right) &= \mathbb{E}\left[d_0(x_0, x'_0) + \|q_0 - q'_0\|_\infty^2\right] \text{ and} \\ x_{k+1} &= x'_{k+1} \quad \text{if } x_k = x'_k, \quad \forall k \geq 0.\end{aligned}$$

Since $\bar{W}_2$ is defined by infimum over all couplings, we have

$$\bar{W}_2^2\left(\mathcal{L}\left(x_1, q_1\right), \mathcal{L}(x_1', q_1')\right) \leq \mathbb{E}\left[d_0(x_1, x_1') + \|q_1 - q_1'\|_\infty^2\right].$$

We denote by $e_{(s,a)} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ the one-hot vector with only one "1" in the location of (s, a). We have

$$\begin{aligned}
\|q_1 - q_1'\|_\infty &= \|q_0 - q_0' - \alpha e_{(s_0,a_0)} q_0(s_0, a_0) + \alpha e_{(s_0',a_0')} q_0'(s_0', a_0') \\
&\quad + \alpha e_{(s_0,a_0)} r(s_0, a_0) - \alpha e_{(s_0',a_0')} r(s_0', a_0') \\
&\quad + \alpha\gamma e_{(s_0,a_0)} \max_a q_0(s_1, a) - \alpha\gamma e_{(s_0',a_0')} \max_a q_0'(s_1', a)\|_\infty \\
&\leq \|q_0 - q_0' - \alpha e_{(s_0,a_0)} q_0(s_0, a_0) + \alpha e_{(s_0',a_0')} q_0'(s_0', a_0')\|_\infty \\
&\quad + \alpha\|e_{(s_0,a_0)} r(s_0, a_0) - e_{(s_0',a_0')} r(s_0', a_0')\|_\infty \\
&\quad + \alpha\gamma\|e_{(s_0,a_0)} \max_a q_0(s_1, a) - e_{(s_0',a_0')} \max_a q_0'(s_1', a)\|_\infty \\
&\leq \|q_0 - q_0'\|_\infty + \alpha d_0(x_0', x_0) q_{\max} + \alpha r_{\max} d_0(x_0', x_0) + \alpha\gamma\|q_0 - q_0'\|_\infty + \alpha\gamma q_{\max} d_0(x_0', x_0) \\
&= (1 + \alpha\gamma)\|q_0 - q_0'\|_\infty + (\alpha r_{\max} + \alpha(\gamma + 1) q_{\max}) d_0(x_0', x_0).
\end{aligned}$$

Therefore, we obtain

$$\begin{aligned}
&\mathbb{E}\left[d_0(x_1, x_1') + \|q_1 - q_1'\|_\infty^2\right] = \mathbb{E}\left[d_0(x_1, x_1')\right] + \mathbb{E}\left[\|q_1 - q_1'\|_\infty^2\right] \\
&\quad \leq \mathbb{E}\left[d_0(x_0, x_0')\right] + 2(1 + \alpha\gamma)^2 \mathbb{E}\left[\|q_0 - q_0'\|_\infty^2\right] \\
&\qquad + 2(\alpha r_{\max} + \alpha(\gamma + 1) q_{\max})^2 \mathbb{E}\left[d_0(x_0, x_0')\right] \\
&\quad \leq \rho \bar{W}_2^2\left(\mathcal{L}\left(x_0, q_0\right), \mathcal{L}(x_0', q_0')\right),
\end{aligned}$$

with $\rho = \max\left(1 + 2(\alpha r_{\max} + \alpha(\gamma + 1) q_{\max})^2, 2(1 + \alpha\gamma)^2\right)$. □

By the triangle inequality of extended Wasserstein 2-distance, we obtain

$$\begin{aligned}
\bar{W}_2\left(\mathcal{L}\left(x_1, q_1\right), \bar{\mu}\right) &\leq \bar{W}_2\left(\mathcal{L}\left(x_1, q_1\right), \mathcal{L}\left(x_{k+1}, q_{k+1}\right)\right) + \bar{W}_2\left(\mathcal{L}\left(x_{k+1}, q_{k+1}\right), \bar{\mu}\right) \\
&\leq \rho \bar{W}_2^2\left(\bar{\mu}, \mathcal{L}(x_k, q_k)\right) + \bar{W}_2\left(\mathcal{L}\left(x_{k+1}, q_{k+1}\right), \bar{\mu}\right) \\
&\overset{k\to\infty}{\longrightarrow} 0,
\end{aligned} \tag{16}$$

where the second inequality holds by Lemma 1 and last step comes from the weak convergence result. Therefore, we have proved that $(x_k, q_k)_{k\geq 0}$ converge to a unique stationary distribution $\bar{\mu}$.

Next, we provide the mean squared error (MSE) bound for Q-learning algorithm by restating a variant of Theorem 3.1 in Chen et al. (2021) as follows without the assumption that $r_{\max} \leq 1$, which can be proved by Theorem 2.1 and 3.1 in Chen et al. (2021).

Proposition 4. *Under Assumption 1, and $\alpha t_\alpha \leq c_0 \frac{(1-\beta)^2}{\log(|\mathcal{S}||\mathcal{A}|)}$ (c_0 is a constant), for all $k \geq t_\alpha$, we obtain*

$$\mathbb{E}\left[\|q_k - q^*\|_\infty^2\right] \leq c_{Q,1}\left(1 - \frac{(1-\beta)\,\alpha}{2}\right)^{k - t_\alpha} + c_Q \frac{\log(|\mathcal{S}||\mathcal{A}|)}{(1-\beta)^2} \alpha t_\alpha, \tag{17}$$

where $c_{Q,1} = 3\left(\|q_0 - q^*\|_\infty + \|q_0\|_\infty + \frac{r_{\max}}{3}\right)^2$ and $c_Q = 912e\left(3\|q^*\|_\infty + r_{\max}\right)$.

Here, c_0 is the same constant as $c_{Q,0}$ appearing in Theorem 3.1 in Chen et al. (2021). We remark that under a constant stepsize, the MSE can be upper bounded by one geometrically decaying term and one bias term that cannot be eliminated as $k \to \infty$; in contrast, using diminishing stepsize $\alpha_k \propto \frac{1}{k}$ can ensure that the MSE decays to zero, but the decaying rate is linear (Chen et al., 2021).

Finally, we establish the following lemma to bound the variance of the limit random vector q_∞, $\text{Var}\left(q_\infty\right)$.

Lemma 2. *Under Assumption 1, and* $\alpha t_\alpha \leq c_0 \frac{(1-\beta)^2}{\log(|\mathcal{S}||\mathcal{A}|)}$ *(c_0 is a constant), we obtain*

$$\mathrm{Var}\left(q_\infty\right) \leq c_Q \frac{\log(|\mathcal{S}||\mathcal{A}|)}{\left(1-\beta\right)^2} \alpha t_\alpha$$

and

$$(\mathbb{E}[\|q_\infty\|_\infty])^2 \leq \mathbb{E}[\|q_\infty\|_\infty^2] \leq 2c_Q c_0 + 2\|q^*\|^2,$$

where $c_Q = 912e\left(3\|q^*\|_\infty + r_{\max}\right).$

Proof for Lemma 2. We have shown that the sequence $(q_k)_{k\geq 0}$ converges weakly to q_∞ in $\mathcal{P}_2(\mathbb{R}^d)$. It is well known that weak convergence in $\mathcal{P}_2(\mathbb{R}^d)$ is equivalent to convergence in distribution and the convergence of the first two moments. As a result, we have

$$\mathbb{E}\left[\|q_\infty - q^*\|_\infty^2\right] = \lim_{k\to\infty} \mathbb{E}\left[\|q_k - q^*\|_\infty^2\right]. \tag{18}$$

Taking $k \to \infty$ on the both sides of equation (17) and combining with equation 18 yields

$$\mathbb{E}[\|q_\infty - q^*\|_\infty^2] \leq c_Q \frac{\log(|\mathcal{S}||\mathcal{A}|)}{\left(1-\beta\right)^2} \alpha t_\alpha.$$

Note that q^* is a deterministic quantity. We thus have

$$\mathrm{Var}\left(q_\infty\right) \overset{(i)}{\leq} \max_{s,a} \mathrm{Var}\left(q_\infty(s,a)\right) \leq \mathbb{E}[\|q_\infty - q^*\|_\infty^2] \leq c_Q \frac{\log(|\mathcal{S}||\mathcal{A}|)}{\left(1-\beta\right)^2} \alpha t_\alpha,$$

where the inequality (i) means an upper bound on elementwise ℓ_∞ norm for the covariance matrix $\mathrm{Var}\left(q_\infty\right)$.

In addition, we have

$$\begin{aligned}
(\mathbb{E}[\|q_\infty\|_\infty])^2 &\leq \mathbb{E}[\|q_\infty\|_\infty^2] \\
&\leq \mathbb{E}[(\|q_\infty - q^*\|_\infty + \|q^*\|_\infty)^2] \\
&\leq 2\mathbb{E}(\|q_\infty - q^*\|_\infty^2) + 2\|q^*\|_\infty^2 \\
&\leq 2c_Q \frac{\log(|\mathcal{S}||\mathcal{A}|)}{\left(1-\beta\right)^2} \alpha t_\alpha + 2\|q^*\|_\infty^2 \\
&\leq 2c_Q c_0 + 2\|q^*\|_\infty^2.
\end{aligned}$$

□

Therefore, we have proved parts 1 and 2 of Theorem 1.

B.3 Convergence Rate

So far we have established that the Markov chain $(x_k, q_k)_{k\geq 0}$ converges to a unique stationary distribution $\bar{\mu} \in \mathcal{P}_2(\mathcal{X} \times \mathbb{R}^{|\mathcal{S}||\mathcal{A}|})$. As a result, $(q_k)_{k\geq 0}$ converges weakly to $\mu \in \mathcal{P}_2(\mathbb{R}^{|\mathcal{S}||\mathcal{A}|})$, where μ is the second marginal of $\bar{\mu}$ over $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. We next focus on the convergence rate of $(q_k)_{k\geq 0}$.

Let us consider the coupled processes defined as equation (13) in Section B.1. Suppose that the initial iterate $(x_0, q_0^{[2]})$ follows the stationary distribution $\bar{\mu}$, thus $\mathcal{L}(x_k, q_k^{[2]}) = \bar{\mu}$ and $\mathcal{L}(q_k^{[2]}) = \mu$ for all $k \geq 0$. By equation (14), we have for all $k \geq 0$:

$$\begin{aligned} W_2^2\left(\mathcal{L}(q_k^{[1]}), \mu\right) &= W_2^2\left(\mathcal{L}(q_k^{[1]}), \mathcal{L}(q_k^{[2]})\right) \\ &\leq \bar{W}_2^2\left(\mathcal{L}(x_k, q_k^{[1]}), \mathcal{L}(x_k, q_k^{[2]})\right) \\ &\leq 12\mathbb{E}\left[\|q_0^{[1]} - q_0^{[2]}\|_\infty^2\right]\left(1 - \frac{(1-\beta)\,\alpha}{2}\right)^{k-t_\alpha} \\ &\leq 24\left(1 - \frac{(1-\beta)\,\alpha}{2}\right)^{k-t_\alpha} \cdot \\ &\quad \left(\mathbb{E}\left[\|q_0^{[1]}\|_\infty^2\right] + \mathbb{E}\left[\|q_\infty^{[1]}\|_\infty^2\right]\right). \end{aligned} \tag{19}$$

Here the last step follows from the fact that $(x_0, q_0^{[2]})$ follows the stationary distribution, and thus $\mathbb{E}\left[\|q_0^{[2]}\|_\infty^2\right] = \mathbb{E}\left[\|q_\infty^{[2]}\|_\infty^2\right] = \mathbb{E}\left[\|q_\infty^{[1]}\|_\infty^2\right]$.

We have completed the proof of Theorem 1.

B.4 Proof of Proposition 1

To analyze the convergence rate of w_k, we construct two new sequences $\{\underline{w}_k\}_{k\geq 0}$ and $\{\bar{w}_k\}_{k\geq 0}$ that satisfy the following recursion:

$$\begin{aligned} \underline{w}_{k+1}(s_k, a_k) &= (1-\alpha)\underline{w}_k(s_k, a_k) + \alpha\gamma\left(\min_{a'} \underline{w}_k(s_{k+1}, a')\right), \\ \bar{w}_{k+1}(s_k, a_k) &= (1-\alpha)\bar{w}_k(s_k, a_k) + \alpha\gamma\left(\max_{a'} \bar{w}_k(s_{k+1}, a')\right). \end{aligned}$$

Let $\underline{w}_0 = w_0 = \bar{w}_0$. We then prove that $\underline{w}_k$ and $\bar{w}_k$ provide a lower bound and upper bound for w_k, respectively.

Lemma 3. *For all $k \geq 0$ and all $(s,a) \in \mathcal{S} \times \mathcal{A}$, $\underline{w}_k(s,a) \leq w_k(s,a) \leq \bar{w}_k(s,a)$.*

Proof of Lemma 3. We use an inductive argument to prove this lemma.

For $k = 0$, $\underline{w}_0 = w_0 = \bar{w}_0$ by definition.

Now assume for $k = k_0$, $\underline{w}_{k_0} \leq w_{k_0} \leq \bar{w}_{k_0}$. For $k = k_0 + 1$, we consider the following two cases:

For $(s,a) \neq (s_{k_0}, a_{k_0})$, we have

$$\underline{w}_{k_0+1}(s,a) = \underline{w}_{k_0}(s,a) \leq w_{k_0}(s,a) = w_{k_0+1}(s,a) \leq \bar{w}_{k_0}(s,a) = \bar{w}_{k_0+1}(s,a).$$

For $(s,a) = (s_{k_0}, a_{k_0})$, we have

$$\begin{aligned} w_{k_0+1}(s,a) &= (1-\alpha)w_{k_0}(s,a) + \alpha\gamma\left(\max_{a'} q_{k_0}^{[1]}(s_{k_0+1}, a') - \max_{a'} q_{k_0}^{[2]}(s_{k_0+1}, a')\right) \\ &\leq (1-\alpha)w_{k_0}(s,a) + \alpha\gamma \max_{a'}\left(q_{k_0}^{[1]}(s_{k_0+1}, a') - q_{k_0}^{[2]}(s_{k_0+1}, a')\right) \\ &= (1-\alpha)w_{k_0}(s,a) + \alpha\gamma \max_{a'}\left(w_{k_0}(s_{k_0+1}, a')\right) \\ &\leq (1-\alpha)\bar{w}_{k_0}(s,a) + \alpha\gamma \max_{a'}\left(\bar{w}_{k_0}(s_{k_0+1}, a')\right) = \bar{w}_{k_0+1}(s,a). \\ w_{k_0+1}(s,a) &= (1-\alpha)w_{k_0}(s,a) + \alpha\gamma\left(\max_{a'} q_{k_0}^{[1]}(x_{k_0+1}, a') - \max_{a'} q_{k_0}^{[2]}(s_{k_0+1}, a')\right) \\ &\geq (1-\alpha)w_{k_0}(s,a) + \alpha\gamma \min_{a'}\left(q_{k_0}^{[1]}(s_{k_0+1}, a') - q_{k_0}^{[2]}(s_{k_0+1}, a')\right) \\ &= (1-\alpha)w_{k_0}(s,a) + \alpha\gamma \min_{a'}\left(w_{k_0}(s_{k_0+1}, a')\right) \\ &\geq (1-\alpha)\underline{w}_{k_0}(s,a) + \alpha\gamma \min_{a'}\left(\underline{w}_{k_0}(s_{k_0+1}, a')\right) = \underline{w}_{k_0+1}(s,a). \end{aligned}$$

By induction, we complete the proof of Lemma 3.

□

Notice that $\{-\underline{w}_k\}$ and $\{\bar{w}_k\}$ can be viewed as the iterates generated by the Q-learning algorithm with $r(s,a)=0$ for all (s,a). Then, for both $\{-\underline{w}_k\}$ and $\{\bar{w}_k\}$, we obtain the following bound for the second moment of $\underline{w}_k$ and $\bar{w}_k$ by Proposition 4 with the special case of $q^*=0$ and $r_{\max}=0$.

$$\mathbb{E}\left[\|\underline{w}_k\|_\infty^2\right] \leq 12\mathbb{E}\left[\|w_0\|_\infty^2\right]\left(1-\frac{(1-\beta)\,\alpha}{2}\right)^{k-t_\alpha},$$
$$\mathbb{E}\left[\|\bar{w}_k\|_\infty^2\right] \leq 12\mathbb{E}\left[\|w_0\|_\infty^2\right]\left(1-\frac{(1-\beta)\,\alpha}{2}\right)^{k-t_\alpha}.$$

By Lemma 3, the same bound can also be applied to $\mathbb{E}\left[\|w_k\|_\infty^2\right]$. We thus have

$$\mathbb{E}\left[\|w_k\|_\infty^2\right] \leq 12\mathbb{E}\left[\|w_0\|_\infty^2\right]\left(1-\frac{(1-\beta)\,\alpha}{2}\right)^{k-t_\alpha}.$$

B.5 Proof of Corollary 1

Lemma 2 states that the second moment of q_∞ is bounded by a constant, which is $\mathbb{E}\left[\|q_\infty\|_\infty^2\right]=\mathcal{O}(1)$. Combining this bound with equation (3) in Theorem 1, we obtain

$$W_2^2(\mathcal{L}(q_k),\mu) \leq C(r,\gamma,P)\cdot\left(1-\frac{(1-\beta)\,\alpha}{2}\right)^{k-t_\alpha},$$

where $C(r,\gamma,P)$ is a numerical constant that only depends on the reward function r, discounted factor γ, and stationary distribution for Markov chain $(x_k)_{k\geq 0}$.

By (Villani et al., 2009, Theorem 4.1), there exists a coupling between q_k and q_∞ such that

$$W_2^2(\mathcal{L}(q_k),\mu)=\mathbb{E}[\|q_k-q_\infty\|_\infty^2].$$

By the above bounds and applying Jensen's inequality twice, we obtain that

$$\begin{aligned}\|\mathbb{E}[q_k-q_\infty]\|_\infty^2 &\leq \left(\mathbb{E}[\|q_k-q_\infty\|_\infty]\right)^2\\ &\leq \mathbb{E}[\|q_k-q_\infty\|_\infty^2]\\ &\leq C(r,\gamma,P)\left(1-\frac{(1-\beta)\,\alpha}{2}\right)^{k-t_\alpha}.\end{aligned}$$

We thus have for all $k\geq t_\alpha$,

$$\|\mathbb{E}[q_k]-\mathbb{E}[q_\infty]\|_\infty \leq \mathbb{E}[\|q_k-q_\infty\|_\infty] \leq C(r,\gamma,P)\left(1-\frac{(1-\beta)\,\alpha}{2}\right)^{\frac{k-t_\alpha}{2}}.$$

For the second moment, we notice that

$$\begin{aligned}
&\left\|\mathbb{E}\left[q_k q_k^\top\right] - \mathbb{E}\left[q_\infty q_\infty^\top\right]\right\|_\infty \\
&= \left\|\mathbb{E}\left[\left(q_k - q_\infty + q_\infty\right)\left(q_k - q_\infty + q_\infty\right)^\top\right] - \mathbb{E}\left[q_\infty q_\infty^\top\right]\right\|_\infty \\
&= \left\|\mathbb{E}\left[\left(q_k - q_\infty\right)\left(q_k - q_\infty\right)^\top\right] + \mathbb{E}\left[q_\infty\left(q_k - q_\infty\right)^\top\right] + \mathbb{E}\left[\left(q_k - q_\infty\right) q_\infty^\top\right]\right\|_\infty \\
&\leq \left\|\mathbb{E}\left[\left(q_k - q_\infty\right)\left(q_k - q_\infty\right)^\top\right]\right\|_\infty + \left\|\mathbb{E}\left[q_\infty\left(q_k - q_\infty\right)^\top\right]\right\|_\infty + \left\|\mathbb{E}\left[\left(q_k - q_\infty\right) q_\infty^\top\right]\right\|_\infty \qquad (20)\\
&\leq \mathbb{E}\left[\left\|\left(q_k - q_\infty\right)\left(q_k - q_\infty\right)^\top\right\|_\infty\right] + \mathbb{E}\left[\left\|q_\infty\left(q_k - q_\infty\right)^\top\right\|_\infty\right] + \mathbb{E}\left[\left\|\left(q_k - q_\infty\right) q_\infty^\top\right\|_\infty\right] \\
&\leq \mathbb{E}\left[\left\|q_k - q_\infty\right\|_\infty^2\right] + 2\mathbb{E}\left[\left\|q_\infty^\top\left(q_k - q_\infty\right)\right\|_\infty\right] \\
&\leq \mathbb{E}\left[\left\|q_k - q_\infty\right\|_\infty^2\right] + 2\left(\mathbb{E}\left[\left\|q_k - q_\infty\right\|_\infty^2\right]\mathbb{E}\left[\left\|q_\infty\right\|_\infty^2\right]\right)^{1/2}.
\end{aligned}$$

Meanwhile, we have

$$\mathbb{E}\left[\left\|q_k - q_\infty\right\|_\infty^2\right] \leq C(r,\gamma,P)\left(1 - \frac{(1-\beta)\,\alpha}{2}\right)^{k-t_\alpha} \quad \text{and} \quad \mathbb{E}\left[\left\|q_\infty\right\|_\infty^2\right] = \mathcal{O}(1).$$

Substituting the above bounds into the right-hand side of inequality (20) yields

$$\left\|\mathbb{E}\left[q_k q_k^\top\right] - \mathbb{E}\left[q_\infty q_\infty^\top\right]\right\|_\infty \leq C'(r,\gamma,P)\left(1 - \frac{(1-\beta)\,\alpha}{2}\right)^{\frac{k-t_\alpha}{2}},$$

thereby completing the proof for Corollary 1.

C Proof of Theorem 2

Define $f : \mathcal{X} \times \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, such that $f(x,q) := q - \mathbb{E}(q_\infty)$. Consider $\{(x_k, q_k)\}_{k\geq 0}$ with $x_0 \sim \mu_{\mathcal{X}}$ and $q_0 \sim \bar{\mu}(\cdot \mid x_0)$.

$$\begin{aligned}
\left\|\sum_{k=0}^{n-1} P^k f\right\|_{\infty, L^2_{\bar{\mu}}} &= \sqrt{\mathbb{E}_{(x_0,q_0)\sim\bar{\mu}}\|\sum_{k=0}^{n-1}\mathbb{E}[f(x_k,q_k) \mid x_0,q_0]\|_\infty^2} \\
&= \sqrt{\mathbb{E}_{(x_0,q_0)\sim\bar{\mu}}\|\sum_{k=0}^{n-1}\mathbb{E}[q_k \mid x_0,q_0] - n\mathbb{E}(q_\infty)\|_\infty^2} \\
&\leq \sqrt{\mathbb{E}_{(x_0,q_0)\sim\bar{\mu}}\|\sum_{k=0}^{n-1}\mathbb{E}[q_k \mid x_0,q_0] - n\mathbb{E}(q_\infty)\|_2^2} \\
&= \sqrt{\mathbb{E}_{(x_0,q_0)\sim\bar{\mu}}\sum_{i,j=0}^{n-1}\mathbb{E}[q_i - \mathbb{E}(q_\infty) \mid x_0,q_0]^T\mathbb{E}[q_j - \mathbb{E}(q_\infty) \mid x_0,q_0]}.
\end{aligned}$$

Define $g_k(x,q) := \mathbb{E}[q_k - \mathbb{E}(q_\infty) \mid (x_0,q_0) = (x,q)]$, we then give the following Lemma 4 to uniformly bound $g_k(x,q)$ for all $(x,q) \in \mathcal{X} \times \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. The proof of Lemma 4 is given at section C.1.

Lemma 4. *For all $(x,q) \in \mathcal{X} \times \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, when $k \geq t_\alpha$, there exist two constant λ_0, λ_1 such that*

$$\|g_k(x,q)\|_2 \leq \lambda_0 \cdot \lambda_1^k,$$

where $\lambda_0 > 0$ and $0 < \lambda_1 < 1$.

By Lemma 4, we obtain

$$\begin{aligned}\left\|\sum_{k=0}^{n-1} P^k f\right\|_{\infty, L^2_{\bar{\mu}}} &\leq \sqrt{\mathbb{E}_{(x_0,q_0)\sim\bar{\mu}} \sum_{i,j=0}^{n-1} \mathbb{E}[q_i - \mathbb{E}(q_\infty) \mid x_0, q_0]^T \mathbb{E}[q_j - \mathbb{E}(q_\infty) \mid x_0, q_0]} \\ &\leq \sqrt{\mathbb{E}_{(x_0,q_0)\sim\bar{\mu}} \sum_{i,j=0}^{n-1} \|g_i(x_0,q_0)\|_2 \|g_j(x_0,q_0)\|_2} \\ &\leq \sqrt{\frac{\lambda_0^2}{(1-\lambda_1)^2}} = \mathcal{O}(1).\end{aligned}$$

By Lemma 2, we can observe that $\int \|f(x,q)\|_\infty^2 \bar{\mu}(\mathrm{d}(x,q)) < \infty$ and $\int f(x)\bar{\mu}(\mathrm{d}(x,q)) = \mathbf{0}$. Therefore, by Theorem 2.1 in Xie & Zhang (2022), we complete the proof for Theorem 2.

C.1 Proof of Lemma 4

Recall that the Markov chain $\{x_k\}_{k\geq 0}$ mixes geometrically fast to the stationary distribution $\mu_\mathcal{X}$, and there exist $c \geq 0$ and $\rho \in (0,1)$ s.t.

$$\max_{x\in\mathcal{X}} \|p^k(x,\cdot) - \mu_\mathcal{X}(\cdot)\|_{TV} \leq c\rho^k,$$

When $k \geq t_\alpha$, we have

$$\begin{aligned}g_k(x,q) &= \sum_{x'\in\mathcal{X}} \int_{q'\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}} \mathbb{E}[q_k - \mathbb{E}(q_\infty) \mid (x_{\lfloor\frac{k}{2}\rfloor}, q_{\lfloor\frac{k}{2}\rfloor}) = (x,q)] \mathrm{d}\mathbb{P}\left((x_{\lfloor\frac{k}{2}\rfloor}, q_{\lfloor\frac{k}{2}\rfloor}) = (x',q') \mid (x_0,q_0) = (x,q)\right) \\ &= \sum_{x'\in\mathcal{X}} \int_{q'\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}} g_{k-\lfloor\frac{k}{2}\rfloor}(x',q') \mathrm{d}\mathbb{P}\left((x_{\lfloor\frac{k}{2}\rfloor}, q_{\lfloor\frac{k}{2}\rfloor}) = (x',q') \mid (x_0,q_0) = (x,q)\right) \\ &= \sum_{x'\in\mathcal{X}} \int_{q'\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}} g_{k-\lfloor\frac{k}{2}\rfloor}(x',q') \underbrace{\mathbb{P}\left(x_{\lfloor\frac{k}{2}\rfloor} = x' \mid (x_0,q_0) = (x,q)\right)}_{p(x')} \mathrm{d}\underbrace{\mathbb{P}\left(q_{\lfloor\frac{k}{2}\rfloor} = q' \mid x_{\lfloor\frac{k}{2}\rfloor} = x', (x_0,q_0) = (x,q)\right)}_{\eta(q'|x')} \\ &= \underbrace{\sum_{x'\in\mathcal{X}} \int_{q'\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}} g_{k-\lfloor\frac{k}{2}\rfloor}(x',q') \mu_\mathcal{X}(x') \mathrm{d}\eta(q' \mid x')}_{T_1} + \underbrace{\sum_{x'\in\mathcal{X}} \int_{q'\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}} g_{k-\lfloor\frac{k}{2}\rfloor}(x',q')(p(x') - \mu_\mathcal{X}(x')) \mathrm{d}\eta(q' \mid x')}_{T_2}.\end{aligned}$$

By Corollary 1, when $x_0 \sim \mu_\mathcal{X}$, for all $k \geq t_\alpha$ and arbitrary q_0 we have

$$\|\mathbb{E}[q_k] - \mathbb{E}[q_\infty]\|_\infty \leq C(r,\gamma,P)\left(1 - \frac{(1-\beta)\alpha}{2}\right)^{\frac{k-t_\alpha}{2}}.$$

Therefore, we obtain

$$\begin{aligned}\|T_1\|_2 &\leq \sqrt{|\mathcal{S}||\mathcal{A}|}\|T_1\|_\infty \\ &= \sqrt{|\mathcal{S}||\mathcal{A}|}\|\mathbb{E}_{x'\sim\mu_\mathcal{X}, q'\sim\eta(q'|x')} g_{k-\lfloor\frac{k}{2}\rfloor}\|_\infty \\ &\leq \sqrt{|\mathcal{S}||\mathcal{A}|} C(r,\gamma,P)\left(1 - \frac{(1-\beta)\alpha}{2}\right)^{\frac{k-\lfloor\frac{k}{2}\rfloor - t_\alpha}{2}} \\ &\leq \left(\sqrt{|\mathcal{S}||\mathcal{A}|} C(r,\gamma,P)\left(1 - \frac{(1-\beta)\alpha}{2}\right)^{\frac{-t_\alpha}{2}}\right)\left(1 - \frac{(1-\beta)\alpha}{2}\right)^{\frac{k}{2}}.\end{aligned}$$

Note that $\|q_k\|_\infty \le q_{\max}$, $\|g_k(x,q)\|_\infty \le 2q_{\max}$, we have

$$\|T_2\|_2 \le \sqrt{|\mathcal{S}||\mathcal{A}|}\|T_2\|_\infty \le \sqrt{|\mathcal{S}||\mathcal{A}|}c\rho^{\lfloor \frac{k}{2} \rfloor}|\mathcal{S}|^2|\mathcal{A}| = |\mathcal{S}|^{\frac{5}{2}}|\mathcal{A}|^{\frac{3}{2}}c\rho^{\lfloor \frac{k}{2} \rfloor} \le |\mathcal{S}|^{\frac{5}{2}}|\mathcal{A}|^{\frac{3}{2}}c\rho^{-1}\rho^{\frac{k}{2}}.$$

Therefore, we have

$$\begin{aligned}\|g_k(x,q)\|_2 &= \|T_1 + T_2\|_2 \\ &\le \|T_1\|_2 + \|T_2\|_2 \\ &\le \left(\left(\sqrt{|\mathcal{S}||\mathcal{A}|}C(r,\gamma,P)\left(1 - \frac{(1-\beta)\,\alpha}{2}\right)^{\frac{-t_\alpha}{2}}\right) + |\mathcal{S}|^{\frac{5}{2}}|\mathcal{A}|^{\frac{3}{2}}c\rho^{-1}\right)\left(\max\left\{\sqrt{\left(1 - \frac{(1-\beta)\,\alpha}{2}\right)}, \sqrt{\rho}\right\}\right)^k.\end{aligned}$$

Let $\lambda_0 = \left(\left(\sqrt{|\mathcal{S}||\mathcal{A}|}C(r,\gamma,P)\left(1 - \frac{(1-\beta)\alpha}{2}\right)^{\frac{-t_\alpha}{2}}\right) + |\mathcal{S}|^{\frac{5}{2}}|\mathcal{A}|^{\frac{3}{2}}c\rho^{-1}\right)$ and $\lambda_1 = \max\left\{\sqrt{\left(1 - \frac{(1-\beta)\alpha}{2}\right)}, \sqrt{\rho}\right\}$, we complete the proof of Lemma 4.

D Proof of Theorem 3

In this section, we prove Theorem 3 on the characterization of the bias $\mathbb{E}(q_\infty) - q^*$. The proof consists of five steps, which are given in the following five sub-sections.

D.1 Step 1: Local linearization of Operator F

Unlike linear SA, the operator F in the update rule of Q-learning (cf. equation (1)) is nonlinear and nonsmooth, which makes the analysis considerably more challenging. To address this issue, we employ the local linearization of the operator F around the optimal solution q^*, with a higher order remaining term as stated in Proposition 2 and 3. We provide complete proof here.

Proof of Proposition 2. Recall that we define the unique optimal action with respect to the optimal Q-function q^* as

$$a_s^* := \arg\max_a q^*(s,a).$$

We define a function $G_{q^*} : \mathcal{X} \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times|\mathcal{S}||\mathcal{A}|}$ as follows: for each $x = (s_0, a_0, s_1) \in \mathcal{X}$,

$$[G_{q^*}(x)]\,[(s,a),(\bar{s},\bar{a})] = \begin{cases} 1, & (s,a) = (\bar{s},\bar{a}) \ne (s_0,a_0) \\ \gamma, & (s,a) = (s_0,a_0), (\bar{s},\bar{a}) = (s_1, a^*_{s_1}) \\ 0, & \text{otherwise.}\end{cases}$$

Note that the operator $F(x,\cdot)$ is nonsmooth and does not admit any gradient. On the other hand, by the uniqueness of the optimal policy π^*, we can locally linearize $F(x,\cdot)$ around q^*. In particular, $G_{q^*}(x) - I_d$ serves as an approximate "gradient" of the operator $F(x,\cdot)$ around q^*. Define

$$R(x,q) = F(x,q) - F(x,q^*) - (G_{q^*}(x) - I_d)(q - q^*).$$

We can observe that for $\forall (s,a) \ne (s_0,a_0)$, $[R(x,q)]\,(s,a) = 0$. For $(s,a) = (s_0,a_0)$, we have

$$[R(x,q)]\,(s_0,a_0) = \gamma\left(\max_a q(s_1,a) - q(s_1,a^*_{s_1})\right) \ge 0.$$

If $\|q - q^*\|_\infty < \Delta$, by Assumption 2, for any action $a \ne a^*_{s_1}$, we have

$$\begin{aligned} q(s_1,a^*_{s_1}) &> q^*(s_1,a^*_{s_1}) - \delta \\ &\ge q^*(s_1,a) + \delta \\ &> q(s_1,a).\end{aligned}$$

Thus,

$$[R(x,q)](s_0,a_0) = \gamma \left(\max_a q(s_1,a) - q(s_1,a^*_{s_1}) \right) = 0.$$

If $\|q-q^*\|_\infty \geq \Delta$, we have

$$\begin{aligned} |[R(x,q)](s_0,a_0)| &= \gamma | \max_a q(s_1,a) - q(s_1,a^*_{s_1})| \\ &= \gamma | \max_a q(s_1,a) - \max_a q^*(s_1,a) + q^*(s_1,a^*_{s_1}) - q(s_1,a^*_{s_1})| \\ &\leq 2\gamma \|q-q^*\|_\infty \\ &\leq \frac{2\gamma}{\Delta^3}\|q-q^*\|^4_\infty. \end{aligned}$$

Combining the two situations considered above, we finally obtain that

$$\|R(x,q)\|_\infty \leq \frac{2\gamma}{\Delta^3}\|q-q^*\|^4_\infty.$$

which proves the first part of Proposition 2.

For the second part, we can multiply the $G_{q^*}(x)$ by an arbitrary nonzero vector $H \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. Let $(s_h,a_h) = \arg\max_{(s,a)\in\mathcal{S}\times\mathcal{A}} H(s,a)$ and $p_h = \mu_{\mathcal{S}}(s_h,a_h)$. By Assumption 1, $p_h > 0$. Without loss of generality, we can assume $H(s_h,a_h) > 0$, otherwise we can replace H with $-H$. We then have

$$\begin{aligned} \mathbb{E}\left[G'_{q^*}(x)H\right](s_h,a_h) &= \gamma p_h \mathbb{E}\left(H(s_1,a^*_{s_1}) \mid s_0 = s_h, a_0 = a_h\right) + (1-p_h)H(s_h,a_h) \\ &\leq \gamma p_h H(s_h,a_h) + (1-p_h)H(s_h,a_h) \\ &< H(s_h,a_h), \end{aligned}$$

where the second step hold as the definition of (s_h,a_h) uses the maximum.

We thus have

$$\mathbb{E}[G_{q^*}(x)]H = \mathbb{E}\left[G_{q^*}(x)H\right] \neq H.$$

As H is an arbitrary vector, we conclude that $\mathbb{E}(G_{q^*}(x))$ does not have an eigenvalue of 1, thereby completing the proof for Proposition 2. □

Proof of Proposition 3: Let $f(z) = \frac{1}{2}\|z\|^2_\infty$ and $g(z) = \frac{1}{2}\|z\|^2_2$. Note that $g(\cdot)$ is a convex, differentiable, and 1-smooth function. In Proposition 3, we work with a finite demensional space $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. By Cauchy-Schwarz Inequality, $\frac{1}{\sqrt{|\mathcal{S}||\mathcal{A}|}}\|\cdot\|_2 \leq \|\cdot\|_\infty \leq \|\cdot\|_2$. We construct the Generalized Moreau Envelope of $f(\cdot)$ with respect to $g(\cdot)$ as follows:

$$M^{\eta,g}_f(z) = \min_{u\in\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}} \left\{ f(u) + \frac{1}{\eta} g(z-u) \right\},$$

where $\eta > 0$. For the ease of exposition, we use $M(\cdot)$ to denote $M^{\eta,g}_f(\cdot)$. We restate Lemma 2.1 in Chen et al. (2020b) below on the properties of $M(\cdot)$.

Lemma 5 (Lemma 2.1 in Chen et al. (2020b))**.** *For given $\eta > 0$. Then $M(\cdot)$ constructed above has the following properties:*

1. *(Smoothness) $M(\cdot)$ is convex, $\frac{1}{\eta}$-smooth with respect to $\|\cdot\|_2$.*
2. *There exists a norm $\|\cdot\|_m$ such that $M(z) = \frac{1}{2}\|z\|^2_m$. Furthermore, there exist $l_m, u_m > 0$, such that $l_m\|\cdot\|_m \leq \|\cdot\|_\infty \leq u_m\|\cdot\|_m$. Specifically, we can let $l_m = (1 + \frac{\eta}{|\mathcal{S}||\mathcal{A}|})^{1/2}$, $u_m = (1+\eta)^{1/2}$.*

Therefore, $M(\cdot)$ serves as a smooth approximation of the non-smooth function $f(\cdot)$. Then, we have for $\forall k \geq 0$:

$$
\begin{aligned}
&M^2(q_{k+1}-q^*)\\
&\overset{(a)}{\leq} \left(M(q_k-q^*)+\langle\nabla M(q_k-q^*), q_{k+1}-q_k\rangle+\frac{1}{2\eta}\|q_{k+1}-q_k\|_2^2\right)^2\\
&\overset{(b)}{=} \left(M(q_k-q^*)+\alpha\langle\nabla M(q_k-q^*), F(x_k,q_k)\rangle+\frac{\alpha^2}{2\eta}\|F(x_k,q_k)\|_2^2\right)^2\\
&= (M(q_k-q^*)+\alpha\langle\nabla M(q_k-q^*), \bar{F}(q_k)\rangle\\
&\quad +\alpha\langle\nabla M(q_k-q^*), F(x_k,q_k)-\bar{F}(q_k)\rangle+\frac{\alpha^2}{2\eta}\|F(x_k,q_k)\|_2^2)^2\\
&\overset{(c)}{\leq} M^2(q_k-q^*)+\underbrace{2\alpha M(q_k-q^*)\langle\nabla M(q_k-q^*), \bar{F}(q_k)\rangle}_{T_1} \qquad (21)\\
&\quad +\underbrace{2\alpha M(q_k-q^*)\langle\nabla M(q_k-q^*), F(x_k,q_k)-\bar{F}(q_k)\rangle}_{T_2}+\underbrace{\frac{\alpha^2}{\eta}M(q_k-q^*)\|F(x_k,q_k)\|_2^2}_{T_3}\\
&\quad +\underbrace{3\alpha^2\langle\nabla M(q_k-q^*), \bar{F}(q_k)\rangle^2}_{T_4}+3\alpha^2\underbrace{\langle\nabla M(q_k-q^*), F(x_k,q_k)-\bar{F}(q_k)\rangle^2}_{T_5}\\
&\quad +\underbrace{\frac{3\alpha^4}{4\eta^2}\|F(x_k,q_k)\|_2^4}_{T_6},
\end{aligned}
$$

where (a) follows from the smoothness of $M(\cdot)$ in Lemma 5, (b) follows from the update rule of q_k in (1), and (c) holds by the inequality $(x+y+z)^2 \leq 3(x^2+y^2+z^2)$.

Next we derive an upper bound on $M(\cdot)^2$ by bounding each term of $T_1 - T_6$.

Lemma 6. *For all $k \geq 0$, $T_1 \leq -4\alpha(1-\gamma)^2 M^2(q_k-q^*)$.*

Proof of Lemma 6. By Proposition 2.1 in Chen et al. (2020b), we have that

$$
\langle\nabla M(q_k-q^*), \bar{F}(q_k)\rangle \leq -2\left(1-\gamma\left(\frac{1+\eta\sqrt{|\mathcal{S}||\mathcal{A}|}}{1+\eta}\right)^{\frac{1}{2}}\right)M(q_k-q^*).
$$

We can always choose a sufficiently small η such that $\left(\frac{1+\eta\sqrt{|\mathcal{S}||\mathcal{A}|}}{1+\eta}\right)^{\frac{1}{2}} \leq 2-\gamma$ because $\gamma < 1$, which is equivalent to

$$
\eta \leq \frac{(2-\gamma)^2-1}{\sqrt{|\mathcal{S}||\mathcal{A}|}-1}. \qquad (22)
$$

Since $M(\cdot)$ is non-negative, we complete the proof by multiplying $2\alpha M(q_k-q^*)$ on both sides. □

By Lemma 6, T_1 can give us a desired negative drift term of order $-\mathcal{O}(\alpha)$.

By Cauchy-Schwarz Inequality, we can bound T_2 by two terms. One term is proportional to $M^2(q_k-q^*)$ but still keep the negative drift generated by T_1 and the other term is proportional to T_5:

$$
\begin{aligned}
T_2 &\leq \alpha(1-\gamma)^2M^2(q_k-q^*)+\alpha(1-\gamma)^{-2}\langle\nabla M(q_k-q^*), F(x_k,q_k)-\bar{F}(q_k)\rangle^2\\
&= \alpha(1-\gamma)^2M^2(q_k-q^*)+\alpha(1-\gamma)^{-2}T_5.
\end{aligned}
$$

Then, we can simplify equation (21) as follows when $3\alpha \leq (1-\gamma)^{-2}$:

$$M^2(q_{k+1}-q^*) \leq (1-3\alpha(1-\gamma)^2)M^2(q_k-q^*) + T_3 + T_4 + 2\alpha(1-\gamma)^{-2}T_5 + T_6.$$

By Cauchy-Schwarz Inequality and Lemma A.5 in Chen et al. (2021), T_3 can be bounded as follows:

$$\begin{aligned}
T_3 &= \frac{\alpha^2}{\eta}M(q_k-q^*)\|F(x_k,q_k)\|_2^2 \\
&\leq \frac{\alpha^2}{\eta}M(q_k-q^*)\left(36u_m^2|\mathcal{S}||\mathcal{A}|M(q_k-q^*) + 2|\mathcal{S}||\mathcal{A}|(3\|q^*\|_\infty + r_{\max})^2\right) \\
&= \frac{36u_m^2|\mathcal{S}||\mathcal{A}|\alpha^2}{\eta}M^2(q_k-q^*) + \frac{2|\mathcal{S}||\mathcal{A}|\alpha^2}{\eta}M(q_k-q^*)(3\|q^*\|_\infty + r_{\max})^2 \\
&\leq \frac{36u_m^2|\mathcal{S}||\mathcal{A}|\alpha^2}{\eta}M^2(q_k-q^*) + \alpha(1-\gamma)^2M^2(q_k-q^*) + \frac{|\mathcal{S}|^2|\mathcal{A}|^2\alpha^3}{(1-\gamma)^2\eta^2}(3\|q^*\|_\infty + r_{\max})^4.
\end{aligned}$$

The term T_4 can be directly bounded as follows:

$$\begin{aligned}
T_4 &= 3\alpha^2\langle\nabla M(q_k-q^*), \bar{F}(q_k)\rangle^2 \\
&\leq 3\alpha^2\left(\|\nabla M(q_k-q^*)\|_2\|\bar{F}(q_k)\|_2\right)^2 \\
&= 3\alpha^2\left(\|\nabla M(q_k-q^*) - \nabla M(q^*-q^*)\|_2\|\bar{F}(q_k)-\bar{F}(q^*)\|_2\right)^2 \\
&\leq 3\alpha^2\left(\frac{\sqrt{|\mathcal{S}||\mathcal{A}|}}{\eta}\|q_k-q^*\|_2\|\bar{F}(q_k)-\bar{F}(q^*)\|_\infty\right)^2 \\
&\leq 3\alpha^2\left(\frac{2}{\eta}|\mathcal{S}||\mathcal{A}|\|q_k-q^*\|_\infty^2\right)^2 \\
&\leq \frac{12u_m^4|\mathcal{S}|^2|\mathcal{A}|^2\alpha^2}{\eta^2}\|q_k-q^*\|_m^4 \\
&= \frac{48u_m^4|\mathcal{S}|^2|\mathcal{A}|^2\alpha^2}{\eta^2}M^2(q_k-q^*).
\end{aligned}$$

By Cauchy-Schwarz Inequality, we bound T_5 by the following three parts:

$$\begin{aligned}
T_5 \leq\ & \underbrace{3\langle\nabla M(q_k-q^*) - \nabla M(q_{k-t_{\alpha^2}}-q^*), F(x_k,q_k)-\bar{F}(q_k)\rangle^2}_{T_{51}} \\
&+ \underbrace{3\langle\nabla M(q_{k-t_{\alpha^2}}-q^*), F(x_k,q_k) - F(x_k,q_{k-t_{\alpha^2}}) + \bar{F}(q_{k-t_{\alpha^2}}) - \bar{F}(q_k)\rangle^2}_{T_{52}} \\
&+ \underbrace{3\langle\nabla M(q_{k-t_{\alpha^2}}-q^*), F(x_k,q_{k-t_{\alpha^2}}) - \bar{F}(q_{k-t_{\alpha^2}})\rangle^2}_{T_{53}}.
\end{aligned}$$

By Lemma A.3 in Chen et al. (2021), for all $k \geq t_{\alpha^2}$ with α satisfying $\alpha t_{\alpha^2} \leq \frac{1}{12}$:

$$\begin{aligned}
T_{51} &\leq 3\left(\frac{144u_m^2|\mathcal{S}||\mathcal{A}|\alpha t_{\alpha^2}}{\eta}M(q_k-q^*) + \frac{8|\mathcal{S}||\mathcal{A}|\alpha t_{\alpha^2}}{\eta}(3\|q^*\|_\infty + r_{\max})^2\right)^2 \\
&\leq \frac{124416u_m^4|\mathcal{S}|^2|\mathcal{A}|^2\alpha^2t_{\alpha^2}^2}{\eta^2}M^2(q_k-q^*) + \frac{384|\mathcal{S}|^2|\mathcal{A}|^2\alpha^2t_{\alpha^2}^2}{\eta^2}(3\|q^*\|_\infty + r_{\max})^4, \\
T_{52} &\leq 3\left(\frac{576u_m^2|\mathcal{S}||\mathcal{A}|\alpha t_{\alpha^2}}{\eta}M(q_k-q^*) + \frac{32|\mathcal{S}||\mathcal{A}|\alpha t_{\alpha^2}}{\eta}(3\|q^*\|_\infty + r_{\max})^2\right)^2 \\
&\leq \frac{1990656u_m^4|\mathcal{S}|^2|\mathcal{A}|^2\alpha^2t_{\alpha^2}^2}{\eta^2}M^2(q_k-q^*) + \frac{6144|\mathcal{S}|^2|\mathcal{A}|^2\alpha^2t_{\alpha^2}^2}{\eta^2}(3\|q^*\|_\infty + r_{\max})^4.
\end{aligned}$$

For term T_{53}, we use the conditional expectation as follows:

$$\begin{aligned}
&\mathbb{E}[T_{53}|x_{k-t_{\alpha^2}}, q_{k-t_{\alpha^2}}] \\
=&3\mathbb{E}\left[\langle \nabla M(q_{k-t_{\alpha^2}} - q^*), F(x_k, q_{k-t_{\alpha^2}}) - \bar{F}(q_{k-t_{\alpha^2}})\rangle^2 | x_{k-t_{\alpha^2}}, q_{k-t_{\alpha^2}}\right]
\end{aligned} \tag{23}$$

Let $H = \nabla M(q_{k-t_{\alpha^2}} - q^*) \cdot \nabla M(q_{k-t_{\alpha^2}} - q^*)^\top$. Equation (23) can be reformulated as follows:

$$\begin{aligned}
&\mathbb{E}[T_{53}|x_{k-t_{\alpha^2}}, q_{k-t_{\alpha^2}}] \\
&= 3\mathbb{E}\left[(F(x_k, q_{k-t_{\alpha^2}}) - \bar{F}(q_{k-t_{\alpha^2}}))^\top H (F(x_k, q_{k-t_{\alpha^2}}) - \bar{F}(q_{k-t_{\alpha^2}})) | x_{k-t_{\alpha^2}}, q_{k-t_{\alpha^2}}\right] \\
&= 3\mathbb{E}\left[F(x_k, q_{k-t_{\alpha^2}})^\top H F(x_k, q_{k-t_{\alpha^2}}) - \bar{F}(q_{k-t_{\alpha^2}})^\top H \bar{F}(q_{k-t_{\alpha^2}}) | x_{k-t_{\alpha^2}}, q_{k-t_{\alpha^2}}\right] \\
&\quad - 6\mathbb{E}\left[(F(x_k, q_{k-t_{\alpha^2}}) - \bar{F}(q_{k-t_{\alpha^2}}))^\top H \bar{F}(q_{k-t_{\alpha^2}}) | x_{k-t_{\alpha^2}}, q_{k-t_{\alpha^2}}\right] \\
&= 3\left(\sum_{x\in\mathcal{X}} \left(P^{t_{\alpha^2}}\left(x_{k-t_{\alpha^2}}, x\right) - \mu_{\mathcal{X}}(x)\right) F\left(x, q_{k-t_{\alpha^2}}\right)^\top H F\left(x, q_{k-t_{\alpha^2}}\right)\right) \\
&\quad - 6\left(\sum_{x\in\mathcal{X}} \left(P^{t_{\alpha^2}}\left(x_{k-t_{\alpha^2}}, x\right) - \mu_{\mathcal{X}}(x)\right) F\left(x, q_{k-t_{\alpha^2}}\right)^\top H \bar{F}\left(q_{k-t_{\alpha^2}}\right)\right) \\
&\overset{(a)}{\leq} 6\alpha^2 \|F(\widetilde{x_0}, q_{k-t_{\alpha^2}})\|_2^2 \|H\|_2 + 12\alpha^2 \|F(\widetilde{x_1}, q_{k-t_{\alpha^2}})\|_2 \|H\|_2 \|\bar{F}(q_{k-t_{\alpha^2}})\|_2 \\
&\leq \frac{18\alpha^2|\mathcal{S}|^2|\mathcal{A}|^2}{\eta^2}(2\|q_{k-t_{\alpha^2}}\|_\infty + r_{\max})^2 \|q_{k-t_{\alpha^2}} - q^*\|_\infty^2 \\
&\leq \frac{18\alpha^2|\mathcal{S}|^2|\mathcal{A}|^2}{\eta^2}(2\|q_{k-t_{\alpha^2}} - q^*\|_\infty + 2\|q^*\|_\infty + r_{\max})^2 \|q_{k-t_{\alpha^2}} - q^*\|_\infty^2 \\
&\leq \frac{18\alpha^2|\mathcal{S}|^2|\mathcal{A}|^2}{\eta^2}(2\|q_{k-t_{\alpha^2}} - q_k\|_\infty + 2\|q_k - q^*\|_\infty + 2\|q^*\|_\infty + r_{\max})^2 \\
&\quad \cdot (\|q_{k-t_{\alpha^2}} - q_k\|_\infty + \|q_k - q^*\|_\infty)^2 \\
&\leq \frac{18\alpha^2|\mathcal{S}|^2|\mathcal{A}|^2}{\eta^2}(2(\|q_k\|_\infty + \frac{r_{\max}}{3}) + 2\|q_k - q^*\|_\infty + 2\|q^*\|_\infty + r_{\max})^2 \\
&\quad \cdot ((\|q_k\|_\infty + \frac{r_{\max}}{3}) + \|q_k - q^*\|_\infty)^2 \\
&\leq \frac{18\alpha^2|\mathcal{S}|^2|\mathcal{A}|^2}{\eta^2}(6\|q_k - q^*\|_\infty + 6\|q^*\|_\infty + 2r_{\max})^2 (3\|q_k - q^*\|_\infty + 3\|q^*\|_\infty + r_{\max})^2 \\
&= \frac{72\alpha^2|\mathcal{S}|^2|\mathcal{A}|^2}{\eta^2}(3\|q_k - q^*\|_\infty + 3\|q^*\|_\infty + r_{\max})^4 \\
&\leq \frac{186624\alpha^2|\mathcal{S}|^2|\mathcal{A}|^2}{\eta^2} M^2(q_k - q^*) + \frac{576\alpha^2|\mathcal{S}|^2|\mathcal{A}|^2}{\eta^2}(3\|q^*\|_\infty + r_{\max})^4.
\end{aligned}$$

where (a) follows with some $\widetilde{x_0}, \widetilde{x_1} \in \mathcal{X}$. Here we use the facts that $\sum_{x\in\mathcal{X}} |P^{t_{\alpha^2}}\left(x_{k-t_{\alpha^2}}, x\right) - \mu_{\mathcal{X}}(x)| \leq 2\alpha^2$ (by Definition 1 of mixing time) and $\|q_{k-t_{\alpha^2}} - q_k\|_\infty \leq \|q_k\|_\infty + \frac{r_{\max}}{3}$, which has been proved in Lemma A.2 in Chen et al. (2021).

By putting these three terms together, we obtain the following bound for $\mathbb{E}[T_5]$:

$$\begin{aligned}
\mathbb{E}(T_5) \leq& \frac{|\mathcal{S}|^2|\mathcal{A}|^2(2115072u_m^4\alpha^2 t_{\alpha^2}^2 + 186624\alpha^2)}{\eta^2} M^2(q_k - q^*) \\
&+ \frac{|\mathcal{S}|^2|\mathcal{A}|^2(6528\alpha^2 t_{\alpha^2}^2 + 576\alpha^2)}{\eta^2}(3\|q^*\|_\infty + r_{\max})^4.
\end{aligned}$$

By Lemma A.5 in Chen et al. (2021), we have

$$
\begin{aligned}
T_6 &\leq \frac{3\alpha^4}{4\eta^2}\left(36u_m^2|\mathcal{S}||\mathcal{A}|M(q_k-q^*)+2|\mathcal{S}||\mathcal{A}|(3\|q^*\|_\infty+r_{\max})^2\right)^2\\
&\leq \frac{1944u_m^4|\mathcal{S}|^2|\mathcal{A}|^2\alpha^4}{\eta^2}M^2(q_k-q^*)+\frac{6|\mathcal{S}|^2|\mathcal{A}|^2\alpha^4}{\eta^2}(3\|q^*\|_\infty+r_{\max})^4.
\end{aligned}
$$

Using the above bounds for $T_1 - T_6$, we can finally bound $\mathbb{E}[M^2(q_{k+1}-q^*)]$ by following:

$$
\begin{aligned}
\mathbb{E}[M^2(q_{k+1}-q^*)] &\leq (1-3\alpha(1-\gamma)^2)\mathbb{E}[M^2(q_k-q^*)]\\
&+\frac{36u_m^2|\mathcal{S}||\mathcal{A}|\alpha^2}{\eta}\mathbb{E}[M^2(q_k-q^*)]+\alpha(1-\gamma)^2\mathbb{E}[M^2(q_k-q^*)]\\
&+\frac{|\mathcal{S}|^2|\mathcal{A}|^2\alpha^3}{(1-\gamma)^2\eta^2}(3\|q^*\|_\infty+r_{\max})^2+\frac{48u_m^4|\mathcal{S}|^2|\mathcal{A}|^2\alpha^2}{\eta^2}\mathbb{E}[M^2(q_k-q^*)]\\
&+\frac{|\mathcal{S}|^2|\mathcal{A}|^2(4230144u_m^4\alpha^3t_{\alpha^2}^2+373248\alpha^3)}{\eta^2(1-\gamma)^2}\mathbb{E}[M^2(q_k-q^*)]\\
&+\frac{|\mathcal{S}|^2|\mathcal{A}|^2(13056\alpha^3t_{\alpha^2}^2+1152\alpha^3)}{\eta^2(1-\gamma)^2}(3\|q^*\|_\infty+r_{\max})^4\\
&+\frac{1944u_m^4|\mathcal{S}|^2|\mathcal{A}|^2\alpha^4}{\eta^2}\mathbb{E}[M^2(q_k-q^*)]+\frac{6|\mathcal{S}|^2|\mathcal{A}|^2\alpha^4}{\eta^2}(3\|q^*\|_\infty+r_{\max})^4\\
&\leq (1-\alpha(1-\gamma)^2)\mathbb{E}[M^2(q_k-q^*)]\\
&+\frac{|\mathcal{S}|^2|\mathcal{A}|^2(374007\alpha^3+13056\alpha^3t_{\alpha^2}^2)}{\eta^2(1-\gamma)^2}(3\|q^*\|_\infty+r_{\max})^4,
\end{aligned}
$$

where there exists a $\alpha_0 > 0$ such that the last step always hold for all $\alpha \leq \alpha_0$.
Then, we obtain that for all $k \geq t_{\alpha^2}$:

$$
\begin{aligned}
\mathbb{E}[M^2(q_k-q^*)] &\leq \mathbb{E}[M^2(q_{t_{\alpha^2}}-q^*)](1-\alpha(1-\gamma)^2)^{k-t_{\alpha^2}}\\
&+\frac{|\mathcal{S}|^2|\mathcal{A}|^2(374007\alpha^2+13056\alpha^2t_{\alpha^2}^2)}{\eta^2(1-\gamma)^4}(3\|q^*\|_\infty+r_{\max})^4.
\end{aligned}
$$

We can choose $\eta = \frac{(1-\gamma)^2}{\sqrt{|\mathcal{S}||\mathcal{A}|}}$ satisfying equation (22) and by (Chen et al., 2021, Theorem A.1), we obtain the following bound for $\mathbb{E}[\|q_k - q^*\|_\infty^4]$:

$$\begin{aligned}
\mathbb{E}[\|q_k - q^*\|_\infty^4] &\le 4u_m^4 \mathbb{E}[M^2(q_k - q^*)] \\
&\le 4u_m^4 \mathbb{E}[M^2(q_{t_{\alpha^2}} - q^*)](1-\alpha(1-\gamma)^2)^{k-t_{\alpha^2}} \\
&+ 4u_m^4 \frac{|\mathcal{S}|^3|\mathcal{A}|^3(374007\alpha^2 + 13056\alpha^2 t_{\alpha^2}^2)}{(1-\gamma)^8}(3\|q^*\|_\infty + r_{\max})^4 \\
&\le \frac{u_m^4}{l_m^4}\mathbb{E}[\|q_{t_{\alpha^2}} - q^*\|_\infty^4](1-\alpha(1-\gamma)^2)^{k-t_{\alpha^2}} \\
&+ \frac{4u_m^4|\mathcal{S}|^3|\mathcal{A}|^3(374007\alpha^2 + 13056\alpha^2 t_{\alpha^2}^2)}{(1-\gamma)^8}(3\|q^*\|_\infty + r_{\max})^4 \\
&\le \frac{u_m^4}{l_m^4}\mathbb{E}((\|q_{t_{\alpha^2}} - q_0\|_\infty + \|q_0 - q^*\|_\infty)^4)(1-\alpha(1-\gamma)^2)^{k-t_{\alpha^2}} \\
&+ \frac{4u_m^4|\mathcal{S}|^3|\mathcal{A}|^3(374007\alpha^2 + 13056\alpha^2 t_{\alpha^2}^2)}{(1-\gamma)^8}(3\|q^*\|_\infty + r_{\max})^4 \\
&\le \frac{u_m^4}{l_m^4}(\|q_0\|_\infty + \|q_0 - q^*\|_\infty + \frac{r_{\max}}{3})^4(1-\alpha(1-\gamma)^2)^{k-t_{\alpha^2}} \\
&+ \frac{4u_m^4|\mathcal{S}|^3|\mathcal{A}|^3(374007\alpha^2 + 13056\alpha^2 t_{\alpha^2}^2)}{(1-\gamma)^8}(3\|q^*\|_\infty + r_{\max})^4.
\end{aligned}$$

By Lemma 5, we can let $l_m = (1 + \frac{\eta}{|\mathcal{S}||\mathcal{A}|})^{1/2}$, $u_m = (1+\eta)^{1/2}$. Define

$$\begin{aligned}
b_1 &= \frac{(1 + \frac{(1-\gamma)^2}{\sqrt{|\mathcal{S}||\mathcal{A}|}})^2}{(1 + \frac{(1-\gamma)^2}{(|\mathcal{S}||\mathcal{A}|)^{\frac{3}{2}}})^2}(\|q_0\|_\infty + \|q_0 - q^*\|_\infty + \frac{r_{\max}}{3})^4, \\
b_2 &= \frac{374007 \times 4(1 + \frac{(1-\gamma)^2}{\sqrt{(|\mathcal{S}||\mathcal{A}|)}})^2|\mathcal{S}|^3|\mathcal{A}|^3}{(1-\gamma)^8}(3\|q^*\|_\infty + r_{\max})^4, \\
b_3 &= \frac{13056 \times 4(1 + \frac{(1-\gamma)^2}{\sqrt{(|\mathcal{S}||\mathcal{A}|)}})^2|\mathcal{S}|^3|\mathcal{A}|^3}{(1-\gamma)^8}(3\|q^*\|_\infty + r_{\max})^4.
\end{aligned}$$

We have for all $k \ge t_{\alpha^2}$,

$$\mathbb{E}[\|q_k - q^*\|_\infty^4] \le b_1(1-\alpha(1-\gamma)^2)^{k-t_{\alpha^2}} + b_2\alpha^2 + b_3\alpha^2 t_{\alpha^2}^2.$$

This completes the proof of Proposition 3. □

D.2 Step 2: Basic Adjoint Relationship

We first derive a recursive relationship for the following quantities

$$z(i) := \mathbb{E}\left[q_\infty \mid x_\infty = i\right], \quad i \in \mathcal{X}.$$

Recall that $(x_k)_{k\ge0}$ is a time-homogeneous Markov chain with transition probability matrix $P = (p_{ij})$ and a unique stationary distribution $\mu_\mathcal{X}$. Theorem 1 shows that $(x_k, q_k)_{k\ge0}$ converges in distribution to a limit $(x_\infty, q_\infty) \sim \bar{\mu}$, with marginal $q_\infty \sim \mu$ and $x_\infty \sim \mu_\mathcal{X}$. Given (x_∞, q_∞), let $x_{\infty+1}$ be a random variable with conditional distribution $\mathbb{P}(x_{\infty+1} = j \mid x_\infty = i) = p_{ij}$, and $q_{\infty+1} = q_\infty + \alpha F(x_\infty, q_\infty)$.

Since (x_∞, q_∞) is in the stationary, $(x_{\infty+1}, q_{\infty+1})$ also follows the stationary distribution $\bar{\mu}$. Let $d = |\mathcal{S}||\mathcal{A}|$. Therefore, for any test function $f : \mathcal{X} \times \mathbb{R}^d \mapsto \mathbb{R}^d$ that satisfies $\|f(x,q)\|_\infty \le C(1+\|q\|_\infty^2)$

for some $C \in \mathbb{R}$, the following relationship holds (Villani et al., 2009, Theorem 6.9)

$$\mathbb{E}[f(x_\infty, q_\infty)] = \mathbb{E}[f(x_{\infty+1}, q_{\infty+1})],$$

which is called Basic Adjoint Relationship (BAR).

Consider the test function $f^{(i)}$, $i \in \mathcal{X}$, defined as

$$f^{(i)}(x, q) = q \cdot \mathbb{1}\{x = i\}.$$

Substituting $f = f^{(i)}$ into BAR gives

$$\mathbb{E}[q_\infty \cdot \mathbb{1}\{x_\infty = i\}] = \mathbb{E}[q_{\infty+1} \cdot \mathbb{1}\{x_{\infty+1} = i\}]. \tag{24}$$

To simplify the presentation, we denote by $\nu(i) := \mu_{\mathcal{X}}(i)$ the probability of the Markov chain $(x_k)_{k\geq 0}$ being in state $i \in \mathcal{X}$ when in stationary. The LHS of equation (24) can be written as follows

$$\mathbb{E}\left[q_\infty \cdot \mathbb{1}\{x_\infty = i\}\right] = \nu(i) \cdot \mathbb{E}\left[q_\infty \mid x_\infty = i\right] = \nu(i) z(i).$$

Recall that $\hat{P} = (\hat{p}_{ij})$ is the transition kernel of the time-reversal of the Markov chain $(x_k)_{k\geq 0}$. The RHS of equation (24) can be reformulated as

$$\begin{aligned}
\mathbb{E}\left[q_{\infty+1} \cdot \mathbb{1}\{x_{\infty+1} = i\}\right] &= \nu(i)\mathbb{E}\left[q_{\infty+1} \mid x_{\infty+1} = i\right] \\
&= \nu(i)\mathbb{E}\left[q_\infty + \alpha F(x_\infty, q_\infty) \mid x_{\infty+1} = i\right] \\
&= \nu(i) \sum_{j\in\mathcal{X}} \hat{p}_{ij} \mathbb{E}\left[q_\infty + \alpha F(x_\infty, q_\infty) \mid x_\infty = j, x_{\infty+1} = i\right] \\
&= \nu(i) \sum_{j\in\mathcal{X}} \hat{p}_{ij} \mathbb{E}\left[q_\infty + \alpha F(j, q_\infty) \mid x_\infty = j\right].
\end{aligned}$$

The last step follows from the fact that condition on x_k, q_k is conditionally independent of x_{k+1} for all $k \geq 1$.

By Proposition 2, we can further rewrite the above equation as

$$\begin{aligned}
&\mathbb{E}\left[q_{\infty+1} \cdot \mathbb{1}\{x_{\infty+1} = i\}\right] \\
&= \nu(i) \sum_{j\in\mathcal{X}} \hat{p}_{ij} \mathbb{E}\left[q_\infty + \alpha \left(F(j, q^*) + (G_{q^*}(j) - I_d)(q_\infty - q^*) + R(j, q_\infty)\right) \mid x_\infty = j\right] \\
&= \nu(i) \sum_{j\in\mathcal{X}} \hat{p}_{ij} \left[z(j) + \alpha \left(F(j, q^*) + (G_{q^*}(j) - I_d)(z(j) - q^*) + \mathbb{E}(R(j, q_\infty) \mid x_\infty = j)\right)\right].
\end{aligned}$$

We thus obtain the following recursive relationship for $\{z(i)\}_{i\in\mathcal{X}}$:

$$\begin{aligned}
z(i) &= \sum_{j\in\mathcal{X}} \hat{p}_{ij} \left[z(j) + \alpha \left(F(j, q^*) + (G_{q^*}(j) - I_d)(z(j) - q^*) + \mathbb{E}(R(j, q_\infty) \mid x_\infty = j)\right)\right] \\
&= \sum_{j\in\mathcal{X}} \hat{p}_{ij} \left[z(j) + \alpha \left(F(j, q^*) + (G_{q^*}(j) - I_d)(z(j) - q^*)\right)\right] + \alpha \mathbb{E}\left[R(x_\infty, q_\infty) \mid x_{\infty+1} = i\right].
\end{aligned} \tag{25}$$

Note that the second term of the RHS of equation (25) can be bounded as

$$\begin{aligned}
\mathbb{E}\left[R(x_\infty, q_\infty) \mid x_{\infty+1} = i\right] &= \frac{1}{\nu(i)} \mathbb{E}\left[R(x_\infty, q_\infty) \mathbb{1}\{x_{\infty+1} = i\}\right] \\
&\overset{(i)}{\leq} \frac{1}{\nu(i)} \mathbb{E}\left[R(x_\infty, q_\infty)\right] \\
&\overset{(ii)}{=} \mathcal{O}(\alpha^2 + \alpha^2 t_{\alpha^2}^2),
\end{aligned}$$

where (i) holds because $R(x,q)$ is always positive, as shown in the proof of Proposition 2; (ii) follows from Proposition 2, Proposition 3 and Hölder's inequality.

Let $A(x) = G_{q^*}(x) - I_d$ and $b(x) = F(x,q^*) - (G_{q^*}(x) - I_d)q^*$. Let D denote the operator given by $(Df)(x) = A(x)f(x)$ for each $x \in \mathcal{X}$. We thus can simplify equation (25) by

$$z = \hat{P}\left(z + \alpha(Dz + b)\right) + \mathcal{O}(\alpha^3 + \alpha^3 t_{\alpha^2}^2). \tag{26}$$

D.3 Step 3: Setting up System of δ

Define the difference

$$\delta(i) := z(i) - \mu_{\mathcal{X}} z \text{ for each } i \in \mathcal{X},$$

where $\mu_{\mathcal{X}} z := \sum_{i\in\mathcal{X}} \nu(i) z(i)$. Let $\Pi = 1 \otimes \mu_{\mathcal{X}}$. Then, by applying the operator $(\hat{P} - \Pi)$ to both side of above equation we obtain

$$\left(P^* - \Pi\right) z = \left(P^* - \Pi\right) \delta.$$

Subtracting Πz from both sides of equation (26), we obtain

$$\begin{aligned} \delta &= (\hat{P} - \Pi)z + \alpha\hat{P}(Dz + b) + \mathcal{O}(\alpha^3 + \alpha^3 t_{\alpha^2}^2) \\ &= (\hat{P} - \Pi)\delta + \alpha\hat{P}(Dz + b) + \mathcal{O}(\alpha^3 + \alpha^3 t_{\alpha^2}^2). \end{aligned} \tag{27}$$

Applying $\mu_{\mathcal{X}}$ to both sides of equation (26), we obtain

$$\mu_{\mathcal{X}}(Dz + b) = \mathcal{O}(\alpha^2 + \alpha^2 t_{\alpha^2}^2). \tag{28}$$

Subtracting equation (28) from equation (27), we obtain

$$\delta = (\hat{P} - \Pi)\delta + \alpha(\hat{P} - \Pi)(Dz + b) + \mathcal{O}(\alpha^3 + \alpha^3 t_{\alpha^2}^2).$$

Then, we have

$$(I - \hat{P} + \Pi)\delta = \alpha(\hat{P} - \Pi)(Dz + b) + \mathcal{O}(\alpha^3 + \alpha^3 t_{\alpha^2}^2).$$

It is well-known that $(I - \hat{P} + \Pi)^{-1}$ exists by Huo et al. (2023). Therefore, we obtain

$$\delta = \alpha(I - \hat{P} + \Pi)^{-1}(\hat{P} - \Pi)(Dz + b) + \mathcal{O}(\alpha^3 + \alpha^3 t_{\alpha^2}^2). \tag{29}$$

D.4 Step 4: Establishing $\delta = \mathcal{O}(\alpha)$

In this sub-section, we show that $\vec{\delta} = \mathcal{O}(\alpha)$, as stated in the following Lemma.

Lemma 7. *Under Assumption 1, and* $\alpha t_\alpha \le c_0 \frac{(1-\beta)^2}{\log(|\mathcal{S}||\mathcal{A}|)}$, *we have*

$$\|\vec{\delta}\|_\infty \le \alpha \cdot B''(r,\gamma,P)$$

for some number $B''(r,\gamma,P) \in \mathbb{R}$ *that is independent of* α.

Proof of Lemma 7. Recalling the definition for $z(i)$, we have

$$z(i) = \mathbb{E}\left[q_\infty \mid x_\infty = i\right] = \frac{\mathbb{E}\left[q_\infty \mathbb{1}\left\{x_\infty = i\right\}\right]}{\nu(i)}.$$

Then by Lemma 2 and the fact that $\nu(i) > 0$, we have

$$\|z(i)\|_\infty \le \frac{\mathbb{E}\left[\|q_\infty\|_\infty\right]}{\nu(i)} \le \frac{1}{\nu_{\min}} \cdot \sqrt{2c_Q c_0 + 2\|q^*\|_\infty^2},$$

where $\nu_{\min} := \min_i \nu(i) > 0$.

By equation (29), we conclude that

$$\|\vec{\delta}\|_\infty \le \alpha \cdot B''(r, \gamma, P)$$

for some number $B''(r, \gamma, P)$ that is independent of α.

□

D.5 Step 5: Expansion of the bias

By definition, $\bar{F}(q^*) = 0$ and $R(x, q^*) \equiv 0$. Define $\bar{A} = \mathbb{E}_{\mu_\mathcal{X}} A(x)$ and $\bar{b} = \mathbb{E}_{\mu_\mathcal{X}} b(x)$. Then, we have $\bar{A}q^* + \bar{b} = 0$. From Proposition 2, $\bar{A}$ is a non-singular matrix. Define $\bar{D}$ be the normalized D such that $(\bar{D}f)(x) = \bar{A}^{-1}A(x)f(x)$. Therefore, we obtain

$$\begin{aligned} q^* &= -\bar{A}^{-1}\bar{b} \\ &= -\bar{A}^{-1}\mu_\mathcal{X} b \\ &= \mu_\mathcal{X}\bar{D}z + \mathcal{O}(\alpha^2 + \alpha^2 t_{\alpha^2}^2), \end{aligned}$$

where the last inequality holds by equation (28).

Because $\delta = z - \Pi z$, we can further obtain

$$q^* = \mu_\mathcal{X}\bar{D}\delta + \mu_\mathcal{X} z + \mathcal{O}(\alpha^2 + \alpha^2 t_{\alpha^2}^2).$$

Then,

$$\mu_\mathcal{X} z = q^* - \mu_\mathcal{X}\bar{D}\delta + \mathcal{O}(\alpha^2 + \alpha^2 t_{\alpha^2}^2).$$

$$z(i) = \delta(i) + \mu_\mathcal{X} z = \delta(i) + q^* - \mu_\mathcal{X}\bar{D}\delta + \mathcal{O}(\alpha^2 + \alpha^2 t_{\alpha^2}^2).$$

Therefore, we obtain

$$z = q^* + (I - \Pi\bar{D})\delta + \mathcal{O}(\alpha^2 + \alpha^2 t_{\alpha^2}^2). \tag{30}$$

Substituting equation (30) into equation (29), we obtain

$$\begin{aligned} \delta &= \alpha(I - \hat{P} + \Pi)^{-1}(\hat{P} - \Pi)(Dz + b) + \mathcal{O}(\alpha^3 + \alpha^3 t_{\alpha^2}^2) \\ &= \alpha\,\underbrace{(I - \hat{P} + \Pi)^{-1}(\hat{P} - \Pi)(Aq^* + b)}_{v} \\ &\quad + \alpha\,\underbrace{(I - \hat{P} + \Pi)^{-1}(\hat{P} - \Pi)D(I - \Pi\bar{D})}_{\Xi}\,\delta \\ &\quad + \mathcal{O}(\alpha^3 + \alpha^3 t_{\alpha^2}^2) \\ &= \alpha v + \alpha\Xi\delta + \mathcal{O}(\alpha^3 + \alpha^3 t_{\alpha^2}^2). \end{aligned}$$

Therefore, we can finally obtain

$$\begin{aligned}\mathbb{E}(q_\infty) &= \mu_{\mathcal{X}} z \\ &= q^* - \mu_{\mathcal{X}} \bar{D}\delta + \mathcal{O}(\alpha^2 + \alpha^2 t_{\alpha^2}^2) \\ &= q^* - \alpha\mu_{\mathcal{X}} \bar{D} v - \alpha\mu_{\mathcal{X}} \bar{D}\Xi\delta + \mathcal{O}(\alpha^2 + \alpha^2 t_{\alpha^2}^2)\end{aligned}$$

Let $B = -\mu_{\mathcal{X}} \bar{D} v$. By Lemma 7, we have $\mu_{\mathcal{X}} \bar{D}\Xi\delta = \mathcal{O}(\alpha)$

Therefore, we have

$$\mathbb{E}(q_\infty) = q^* + \alpha B + \mathcal{O}(\alpha^2 + \alpha^2 t_{\alpha^2}^2)$$

with

$$B = -\mu_{\mathcal{X}} \bar{D}(I - \hat{P} + \Pi)^{-1}(\hat{P} - \Pi)(Aq^* + b). \tag{31}$$

We complete the proof of Theorem 3.

E Proof of Corollary 2

In this section, we provide the proof of the first and second moment bounds in Corollary 2.

E.1 First Moment

First, we have

$$\mathbb{E}\left[\bar{q}_{k_0,k}\right] - q^* = \left(\mathbb{E}\left[q_\infty\right] - q^*\right) + \frac{1}{k-k_0} \underbrace{\sum_{t=k_0}^{k-1} \mathbb{E}\left[q_t - q_\infty\right]}_{T_1}.$$

By Corollary 1, we have that for $k \geq t_\alpha$,

$$\|\mathbb{E}[q_k] - \mathbb{E}[q_\infty]\|_\infty \leq C(r,\gamma,P) \cdot \left(1 - \frac{(1-\beta)\,\alpha}{2}\right)^{\frac{k-t_\alpha}{2}}.$$

Then, when $\alpha t_\alpha \leq 1$, we have the following bound for T_1,

$$\begin{aligned}\|T_1\|_\infty &= \left\|\sum_{t=k_0}^{k-1} \mathbb{E}\left[q_t - q_\infty\right]\right\|_\infty \leq \sum_{t=k_0}^{k-1} \left\|\mathbb{E}\left[q_t\right] - \mathbb{E}\left[q_\infty\right]\right\|_\infty \\ &\leq C(r,\gamma,P)\left(1 - \frac{(1-\beta)\,\alpha}{2}\right)^{\frac{k_0-t_\alpha}{2}} \frac{1}{1 - \sqrt{1 - \frac{(1-\beta)\alpha}{2}}} \\ &\leq C(r,\gamma,P)\left(1 - \frac{(1-\beta)\,\alpha}{2}\right)^{\frac{k_0-t_\alpha}{2}} \frac{4}{(1-\beta)\,\alpha} \\ &\overset{(i)}{\leq} C(r,\gamma,P)\exp\left(-\frac{(1-\beta)\alpha(k_0 - t_\alpha)}{4}\right)\frac{4}{(1-\beta)\,\alpha} \\ &\leq C''(r,\gamma,P) \cdot \frac{1}{\alpha} \cdot \exp\left(-\frac{\alpha(1-\beta)k_0}{4}\right),\end{aligned}$$

where (i) follows from the inequality that $(1-u)^m \leq \exp(-um)$ for $0 < u < 1$.

Together with Theorem 3, we have

$$\mathbb{E}\left[\bar{q}_{k_0,k}\right] - q^* = \alpha B(r,\gamma,P) + \mathcal{O}(\alpha^2 + \alpha^2 t_{\alpha^2}^2) + \mathcal{O}\left(\frac{1}{\alpha(k-k_0)}\exp\left(-\frac{\alpha(1-\beta)k_0}{4}\right)\right),$$

thereby finishing the proof of equation (6) for the first moment.

E.2 Second Moment

We first derive the bound for the second moment of the tail-averaged iterate. Note that

$$\begin{aligned}
&\mathbb{E}\Big[\left(\bar{q}_{k_0,k} - \mathbb{E}\left[q_\infty\right]\right)\left(\bar{q}_{k_0,k} - \mathbb{E}\left[q_\infty\right]\right)^\top\Big] \\
=&\frac{1}{(k-k_0)^2}\mathbb{E}\left[\left(\sum_{t=k_0}^{k-1}(q_t - \mathbb{E}\left[q_\infty\right])\right)\left(\sum_{t=k_0}^{k-1}(q_t - \mathbb{E}\left[q_\infty\right])\right)^\top\right] \\
=&\underbrace{\frac{1}{(k-k_0)^2}\sum_{t=k_0}^{k-1}\mathbb{E}\left[\left(q_t - \mathbb{E}\left[q_\infty\right]\right)\left(q_t - \mathbb{E}\left[q_\infty\right]\right)^\top\right]}_{T_1} \\
&+\underbrace{\frac{1}{(k-k_0)^2}\sum_{t=k_0}^{k-1}\sum_{l=t+1}^{k-1}\left(\mathbb{E}\left[\left(q_t - \mathbb{E}\left[q_\infty\right]\right)\left(q_l - \mathbb{E}\left[q_\infty\right]\right)^\top\right] + \mathbb{E}\left[\left(q_l - \mathbb{E}\left[q_\infty\right]\right)\left(q_t - \mathbb{E}\left[q_\infty\right]\right)^\top\right]\right)}_{T_2}.
\end{aligned}$$

For the term T_1, we have the following decomposition,

$$\begin{aligned}
&\mathbb{E}\left[\left(q_t - \mathbb{E}\left[q_\infty\right]\right)\left(q_t - \mathbb{E}\left[q_\infty\right]\right)^\top\right] \\
=&\,\mathbb{E}\left[q_t q_t^\top - q_t\mathbb{E}\left[q_\infty^\top\right] - \mathbb{E}\left[q_\infty\right]q_t^\top + \mathbb{E}\left[q_\infty\right]\mathbb{E}\left[q_\infty^\top\right]\right] \\
=&\,\mathbb{E}\left[q_t q_t^\top\right] - \mathbb{E}\left[q_t\right]\mathbb{E}\left[q_\infty^\top\right] - \mathbb{E}\left[q_\infty\right]\mathbb{E}\left[q_t^\top\right] + \mathbb{E}\left[q_\infty\right]\mathbb{E}\left[q_\infty^\top\right] \\
=&\,\left(\mathbb{E}\left[q_t q_t^\top\right] - \mathbb{E}\left[q_\infty q_\infty^\top\right]\right) + \left(\mathbb{E}\left[q_\infty q_\infty^\top\right] - \mathbb{E}\left[q_\infty\right]\mathbb{E}\left[q_\infty^\top\right]\right) \\
&- \left(\mathbb{E}\left[q_t\right]\mathbb{E}\left[q_\infty^\top\right] + \mathbb{E}\left[q_\infty\right]\mathbb{E}\left[q_t^\top\right] - 2\mathbb{E}\left[q_\infty\right]\mathbb{E}\left[q_\infty^\top\right]\right) \\
=&\,\left(\mathbb{E}\left[q_t q_t^\top\right] - \mathbb{E}\left[q_\infty q_\infty^\top\right]\right) + \mathrm{Var}\left(q_\infty\right) - \mathbb{E}\left[q_t - q_\infty\right]\mathbb{E}\left[q_\infty^\top\right] - \mathbb{E}\left[q_\infty\right]\mathbb{E}\left[\left(q_t - q_\infty\right)^\top\right]
\end{aligned} \tag{32}$$

Corollary 1 and Lemma 2 imply the following bounds for $k \geq t_\alpha$,

$$\mathbb{E}\left[\|q_t - q_\infty\|_\infty\right] \leq C(r,\gamma,P)\cdot\left(1 - \frac{(1-\beta)\,\alpha}{2}\right)^{\frac{t-t_\alpha}{2}} \tag{33}$$

$$\begin{aligned}
\left\|\mathbb{E}\left[q_t q_t^\top\right] - \mathbb{E}\left[q_\infty q_\infty^\top\right]\right\|_\infty &\leq C'(r,\gamma,P)\cdot\left(1 - \frac{(1-\beta)\,\alpha}{2}\right)^{\frac{t-t_\alpha}{2}} \\
\mathbb{E}\left[\|q_\infty\|_\infty\right] &\leq C''(r,\gamma,P), \\
\mathrm{Var}\left(q_\infty\right) &\leq C'''(r,\gamma,P)\cdot\alpha t_\alpha.
\end{aligned} \tag{34}$$

Substituting these bounds into equation (32), we have

$$\mathbb{E}\left[\left(q_t - \mathbb{E}\left[q_\infty\right]\right)\left(q_t - \mathbb{E}\left[q_\infty\right]\right)^\top\right] = \mathcal{O}\left(\left(1 - \frac{(1-\beta)\,\alpha}{2}\right)^{\frac{t-t_\alpha}{2}} + \alpha t_\alpha\right).$$

Therefore, we can bound T_1 as follows,

$$\begin{aligned}
T_1 &= \frac{1}{(k-k_0)^2}\sum_{t=k_0}^{k-1}\mathbb{E}\left[(q_t-\mathbb{E}[q_\infty])(q_t-\mathbb{E}[q_\infty])^\top\right] \\
&= \frac{1}{(k-k_0)^2}\sum_{t=k_0}^{k-1}\mathcal{O}\left(\left(1-\frac{(1-\beta)\alpha}{2}\right)^{\frac{t-t_\alpha}{2}}+\alpha t_\alpha\right) \\
&= \mathcal{O}\left(\frac{1}{\alpha(k-k_0)^2}\exp\left(-\frac{\alpha(1-\beta)k_0}{4}\right)\right)+\mathcal{O}\left(\frac{\alpha t_\alpha}{k-k_0}\right) \\
&= \mathcal{O}\left(\frac{1}{\alpha(k-k_0)^2}\exp\left(-\frac{\alpha(1-\beta)k_0}{4}\right)+\frac{\alpha t_\alpha}{k-k_0}\right).
\end{aligned}$$

Regarding the term T_2, notice that for $l>t$, we have

$$\begin{aligned}
\mathbb{E}\left[(q_t-\mathbb{E}[q_\infty])(q_l-\mathbb{E}[q_\infty])^\top\right] &= \mathbb{E}\left[\mathbb{E}\left[(q_t-\mathbb{E}[q_\infty])(q_l-\mathbb{E}[q_\infty])^\top \mid q_t\right]\right] \\
&= \mathbb{E}\left[(q_t-\mathbb{E}[q_\infty])\,\mathbb{E}[q_l-\mathbb{E}[q_\infty]\mid q_t]^\top\right] \\
&= \mathbb{E}\left[(q_t-\mathbb{E}[q_\infty])(\mathbb{E}[q_l\mid q_t]-\mathbb{E}[q_\infty])^\top\right].
\end{aligned}$$

Note that for any $y\in\mathbb{R}^d$, it holds that

$$\|\mathbb{E}[q_l\mid q_t=y]-\mathbb{E}[q_\infty]\| = \|\mathbb{E}[q_{l-t}\mid q_0=y]-\mathbb{E}[q_\infty]\| \le C(r,\gamma,P)\cdot\left(1-\frac{(1-\beta)\alpha}{2}\right)^{\frac{l-t-t_\alpha}{2}},$$

where the second inequality holds since Corollary 1 holds for all initial value of q_0.

Therefore, when $l>t$, we have

$$\begin{aligned}
&\mathbb{E}\left[\left\|(q_t-\mathbb{E}[q_\infty])(\mathbb{E}[q_l\mid q_t]-\mathbb{E}[q_\infty])^\top\right\|_\infty\right] \\
\le&\mathbb{E}\left[\|q_t-\mathbb{E}[q_\infty]\|_\infty\|\mathbb{E}[q_l\mid q_t]-\mathbb{E}[q_\infty]\|_\infty\right] \\
\le&\mathbb{E}\left[\|q_t-\mathbb{E}[q_\infty]\|_\infty\right]\cdot\left(C(r,\gamma,P)\cdot\left(1-\frac{(1-\beta)\alpha}{2}\right)^{\frac{l-t-t_\alpha}{2}}\right) \\
\le&\left(\mathbb{E}\left[\|q_t-q_\infty\|_\infty\right]+\mathbb{E}\left[\|q_\infty-\mathbb{E}[q_\infty]\|_\infty\right]\right)\cdot\left(C(r,\gamma,P)\cdot\left(1-\frac{(1-\beta)\alpha}{2}\right)^{\frac{l-t-t_\alpha}{2}}\right) \\
\overset{(i)}{\le}&\left(\mathbb{E}\left[\|q_t-q_\infty\|_\infty\right]+(\mathrm{Tr}(\mathrm{Var}(q_\infty)))^{1/2}\right)\cdot\left(C(r,\gamma,P)\cdot\left(1-\frac{(1-\beta)\alpha}{2}\right)^{\frac{l-t-t_\alpha}{2}}\right) \\
\overset{(ii)}{\le}&\left(C(r,\gamma,P)\cdot\left(1-\frac{(1-\beta)\alpha}{2}\right)^{\frac{t-t_\alpha}{2}}+C'(r,\gamma,P)\sqrt{\alpha t_\alpha}\right)\cdot\left(C(r,\gamma,P)\cdot\left(1-\frac{(1-\beta)\alpha}{2}\right)^{\frac{l-t-t_\alpha}{2}}\right) \\
=&C^2(r,\gamma,P)\cdot\left(1-\frac{(1-\beta)\alpha}{2}\right)^{\frac{l-2t_\alpha}{2}}+C''''(r,\gamma,P)\cdot\sqrt{\alpha t_\alpha}\cdot\left(1-\frac{(1-\beta)\alpha}{2}\right)^{\frac{l-t-t_\alpha}{2}},
\end{aligned}$$

where in (i) $\mathrm{Tr}(\cdot)$ denotes the trace operator and we use the fact that $\mathbb{E}\left[\|q_\infty-\mathbb{E}[q_\infty]\|_\infty\right]\le\sqrt{\mathbb{E}\left[\|q_\infty-\mathbb{E}[q_\infty]\|_\infty^2\right]}=\mathrm{Tr}(\mathrm{Var}(q_\infty))^{1/2}$; in (ii) we use the bounds in equations (33) and (34).

In addition, note that

$$\begin{aligned}
&\frac{1}{(k-k_0)^2}\sum_{t=k_0}^{k-1}\sum_{l=t+1}^{k-1}\mathcal{O}\left(\left(1-\frac{(1-\beta)\,\alpha}{2}\right)^{\frac{l-2t_\alpha}{2}}\right)\\
\leq&\frac{1}{(k-k_0)^2}\sum_{t=k_0}^{\infty}\sum_{l=t+1}^{\infty}\mathcal{O}\left(\left(1-\frac{(1-\beta)\,\alpha}{2}\right)^{\frac{l-2t_\alpha}{2}}\right)\\
\leq&\frac{1}{(k-k_0)^2}\left(\frac{4}{(1-\beta)\alpha}\right)^2\mathcal{O}\left(\left(1-\frac{(1-\beta)\,\alpha}{2}\right)^{\frac{k_0-2t_\alpha}{2}}\right)\\
=&\mathcal{O}\left(\frac{1}{(k-k_0)^2\alpha^2}\exp\left(-\frac{\alpha(1-\beta)k_0}{4}\right)\right),
\end{aligned}$$

and

$$\begin{aligned}
&\frac{1}{(k-k_0)^2}\sum_{t=k_0}^{k-1}\sum_{l=t+1}^{k-1}\mathcal{O}\left(\left(1-\frac{(1-\beta)\,\alpha}{2}\right)^{\frac{l-t-t_\alpha}{2}}\right)\\
\leq&\frac{1}{(k-k_0)^2}\sum_{t=k_0}^{k-1}\sum_{l=t+1}^{\infty}\mathcal{O}\left(\left(1-\frac{(1-\beta)\,\alpha}{2}\right)^{\frac{l-t-t_\alpha}{2}}\right)\\
=&\mathcal{O}\left(\frac{1}{(k-k_0)\alpha}\right).
\end{aligned}$$

Putting together, we obtain the following upper bound for T_2,

$$\begin{aligned}
T_2=&\frac{1}{(k-k_0)^2}\sum_{t=k_0}^{k-1}\sum_{l=t+1}^{k-1}\mathcal{O}\left(\left(1-\frac{(1-\beta)\,\alpha}{2}\right)^{\frac{l-2t_\alpha}{2}}+\sqrt{\alpha t_\alpha}\left(1-\frac{(1-\beta)\,\alpha}{2}\right)^{\frac{l-t-t_\alpha}{2}}\right)\\
=&\mathcal{O}\left(\frac{1}{(k-k_0)^2\alpha^2}\exp\left(-\frac{\alpha(1-\beta)k_0}{4}\right)+\frac{\sqrt{\alpha t_\alpha}}{(k-k_0)\alpha}\right).
\end{aligned}$$

Combining the above bounds for T_1 and T_2, we obtain

$$\begin{aligned}
&\mathbb{E}\Big[\left(\bar{q}_{k_0,k}-\mathbb{E}\left[q_\infty\right]\right)\left(\bar{q}_{k_0,k}-\mathbb{E}\left[q_\infty\right]\right)^\top\Big]\\
=&\mathcal{O}\left(\frac{1}{\alpha(k-k_0)^2}\exp\left(-\frac{\alpha(1-\beta)k_0}{4}\right)+\frac{\alpha t_\alpha}{k-k_0}\right)\\
&+\mathcal{O}\left(\frac{1}{(k-k_0)^2\alpha^2}\exp\left(-\frac{\alpha(1-\beta)k_0}{4}\right)+\frac{\sqrt{t_\alpha/\alpha}}{(k-k_0)}\right)\\
=&\mathcal{O}\left(\frac{\sqrt{t_\alpha/\alpha}}{(k-k_0)}+\frac{1}{(k-k_0)^2\alpha^2}\exp\left(-\frac{\alpha(1-\beta)k_0}{4}\right)\right).
\end{aligned} \tag{35}$$

Now we are ready to bound the LHS of equation (7). First, we have the following decomposition

$$\begin{aligned}
&\mathbb{E}\left[\left(\bar{q}_{k_0,k}-q^*\right)\left(\bar{q}_{k_0,k}-q^*\right)^\top\right]\\
=&\mathbb{E}\left[\left(\bar{q}_{k_0,k}-\mathbb{E}\left[q_\infty\right]+\mathbb{E}\left[q_\infty\right]-q^*\right)\left(\bar{q}_{k_0,k}-\mathbb{E}\left[q_\infty\right]+\mathbb{E}\left[q_\infty\right]-q^*\right)^\top\right]\\
=&\mathbb{E}\left[\left(\bar{q}_{k_0,k}-\mathbb{E}\left[q_\infty\right]\right)\left(\bar{q}_{k_0,k}-\mathbb{E}\left[q_\infty\right]\right)^\top\right]+\mathbb{E}\left[\left(\mathbb{E}\left[q_\infty\right]-q^*\right)\left(\bar{q}_{k_0,k}-\mathbb{E}\left[q_\infty\right]\right)^\top\right]\\
&+\mathbb{E}\left[\left(\bar{q}_{k_0,k}-\mathbb{E}\left[q_\infty\right]\right)\left(\mathbb{E}\left[q_\infty\right]-q^*\right)^\top\right]+\mathbb{E}\left[\left(\mathbb{E}\left[q_\infty\right]-q^*\right)\left(\mathbb{E}\left[q_\infty\right]-q^*\right)^\top\right].
\end{aligned} \tag{36}$$

For the second term of RHS of equation 36, we have

$$
\begin{aligned}
&\mathbb{E}\left[\left(\bar{q}_{k_0,k}-\mathbb{E}\left[q_\infty\right]\right)\left(\mathbb{E}\left[q_\infty\right]-q^*\right)^\top\right]\\
=&\frac{1}{k-k_0}\left(\sum_{t=k_0}^{k-1}\mathbb{E}\left[q_t-q_\infty\right]\right)\left(\mathbb{E}\left[q_\infty\right]-q^*\right)^\top\\
=&\mathcal{O}\left(\frac{1}{\alpha(k-k_0)}\exp\left(-\frac{\alpha(1-\beta)k_0}{4}\right)\right)\left(\alpha B(r,\gamma,P)+\mathcal{O}(\alpha^2+\alpha^2 t_{\alpha^2}^2)\right)\\
=&\mathcal{O}\left(\frac{1}{k-k_0}\exp\left(-\frac{\alpha(1-\beta)k_0}{4}\right)\right).
\end{aligned}
$$

Similarly, we have the same bound for the third term of equation (36). For the last term of RHS of equation (36), we have

$$
\begin{aligned}
\mathbb{E}\left[\left(\mathbb{E}\left[q_\infty\right]-q^*\right)\left(\mathbb{E}\left[q_\infty\right]-q^*\right)^\top\right]&=\left(\mathbb{E}\left[q_\infty\right]-q^*\right)\left(\mathbb{E}\left[q_\infty\right]-q^*\right)^\top\\
&=\left(\alpha B(r,\gamma,P)+\mathcal{O}(\alpha^2+\alpha^2 t_{\alpha^2}^2)\right)\left(\alpha B(r,\gamma,P)+\mathcal{O}(\alpha^2+\alpha^2 t_{\alpha^2}^2)\right)^\top\\
&=\alpha^2 B'(r,\gamma,P)+\mathcal{O}(\alpha^3+\alpha^3 t_{\alpha^2}^2).
\end{aligned}
$$

Combining all these bounds, we obtain

$$
\begin{aligned}
&\mathbb{E}\left[\left(\bar{q}_{k_0,k}-q^*\right)\left(\bar{q}_{k_0,k}-q^*\right)^\top\right]\\
&=\alpha^2 B'(r,\gamma,P)+\mathcal{O}(\alpha^3+\alpha^3 t_{\alpha^2}^2)\\
&\quad+\mathcal{O}\left(\frac{\sqrt{t_\alpha/\alpha}}{(k-k_0)}+\frac{1}{(k-k_0)^2\alpha^2}\exp\left(-\frac{\alpha(1-\beta)k_0}{4}\right)\right)\\
&=\alpha^2 B'+\mathcal{O}\left(\alpha^3+\alpha^3 t_{\alpha^2}^2+\frac{\sqrt{t_\alpha/\alpha}}{(k-k_0)}++\frac{1}{(k-k_0)^2\alpha^2}\exp\left(-\frac{\alpha(1-\beta)k_0}{4}\right)\right).
\end{aligned}
$$

thereby completing the proof of Corollary 2.

F Proof of Corollary 3

In this section, we give the proof of the first and second moment bounds in Corollary 3.

F.1 First Moment

We have

$$
\begin{aligned}
\mathbb{E}\left[\widetilde{q}_{k_0,k}^{(\alpha)}\right]-q^*=&\left(2\bar{q}_{k_0,k}^{(\alpha)}-\bar{q}_{k_0,k}^{(2\alpha)}\right)-q^*\\
=&2\left(\bar{q}_{k_0,k}^{(\alpha)}-q^*\right)-\left(\bar{q}_{k_0,k}^{(2\alpha)}-q^*\right)\\
\overset{(i)}{=}&2\left(\alpha B(r,\gamma,P)+\mathcal{O}(\alpha^2+\alpha^2 t_{\alpha^2}^2)+\mathcal{O}\left(\frac{1}{\alpha(k-k_0)}\exp\left(-\frac{\alpha(1-\beta)k_0}{4}\right)\right)\right)\\
&-\left(2\alpha B(r,\gamma,P)+\mathcal{O}(\alpha^2+\alpha^2 t_{\alpha^2}^2)+\mathcal{O}\left(\frac{1}{\alpha(k-k_0)}\exp\left(-\frac{\alpha(1-\beta)k_0}{2}\right)\right)\right)\\
=&\mathcal{O}(\alpha^2+\alpha^2 t_{\alpha^2}^2)+\mathcal{O}\left(\frac{1}{\alpha(k-k_0)}\exp\left(-\frac{\alpha(1-\beta)k_0}{4}\right)\right)
\end{aligned}
$$

where (i) follows from Corollary 2.

F.2 Second Moment

We first introduce the following short-hands:

$$u_1 := \bar{q}_{k_0,k}^{(\alpha)} - \mathbb{E}\left[q_\infty^{(\alpha)}\right], \quad u_2 := \bar{q}_{k_0,k}^{(2\alpha)} - \mathbb{E}\left[q_\infty^{(2\alpha)}\right]$$
$$\text{and} \quad v := 2\mathbb{E}\left[q_\infty^{(\alpha)}\right] - \mathbb{E}\left[q_\infty^{(2\alpha)}\right] + q^*.$$

With these notations, $\tilde{q}_{k_0,k} - q^* = 2u_1 - u_2 + v$. We then have the following bound

$$\begin{aligned}
\left\|\mathbb{E}\left[\left(\widetilde{q}_{k_0,k}^{(\alpha)} - q^*\right)\left(\widetilde{q}_{k_0,k}^{(\alpha)} - q^*\right)^\top\right]\right\|_\infty &\le \left\|\mathbb{E}\left[\left(\widetilde{q}_{k_0,k}^{(\alpha)} - q^*\right)\left(\widetilde{q}_{k_0,k}^{(\alpha)} - q^*\right)^\top\right]\right\|_2 \\
&= \left\|\mathbb{E}\left[(2u_1 - u_2 + v)(2u_1 - u_2 + v)^\top\right]\right\|_2 \\
&\le \mathbb{E}\left[\|2u_1 - u_2 + v\|_2^2\right] \\
&\le 3\mathbb{E}\,\|2u_1\|_2^2 + 3\mathbb{E}\,\|u_2\|_2^2 + 3\|v\|_2^2.
\end{aligned}$$

By equation (35), we have

$$\mathbb{E}\,\|u_1\|_2^2 = \operatorname{Tr}\left(\mathbb{E}\left[u_1 u_1^\top\right]\right) = \mathcal{O}\left(\frac{\sqrt{t_\alpha/\alpha}}{(k-k_0)} + \frac{1}{(k-k_0)^2\alpha^2}\exp\left(-\frac{\alpha(1-\beta)k_0}{4}\right)\right).$$

Similarly, we have

$$\mathbb{E}\,\|u_2\|_2^2 = \mathcal{O}\left(\frac{\sqrt{t_\alpha/\alpha}}{(k-k_0)} + \frac{1}{(k-k_0)^2\alpha^2}\exp\left(-\frac{\alpha(1-\beta)k_0}{2}\right)\right).$$

By Theorem 3, we have $\|v\|_2^2 = \mathcal{O}\left(\alpha^4 + \alpha^4 t_{\alpha^2}^4\right)$.

Combining these bounds together, we have

$$\begin{aligned}
&\mathbb{E}\left[(\tilde{q}_{k-k_0} - q^*)(\tilde{q}_{k-k_0} - q^*)^\top\right] \\
&= \mathcal{O}\left(\alpha^4 + \alpha^4 t_{\alpha^2}^4\right) + \mathcal{O}\left(\frac{\sqrt{t_\alpha/\alpha}}{(k-k_0)} + \frac{1}{(k-k_0)^2\alpha^2}\exp\left(-\frac{\alpha(1-\beta)k_0}{4}\right)\right).
\end{aligned}$$

G Experiment Details

Tabular case. We consider two MDPs for our numerical experiments.

The first example is a 1×3 Gridword with $\mathcal{S} = \{0, 1, 2\}$ and $\mathcal{A} = \{-1, 1\}$. For each step, the agent can walk in two directions: left or right. If the agent walks out of the space, the agent would get a reward of -4 and stay at the same state. Otherwise, the agent can walk to the next state with probability of 0.95 or still stay at the same state with probability of 0.05. For the case that the agent does not exceed the space, the reward function is determined by the current state $r(s, a) = r(s)$ with $r(0) = 0, r(1) = 10$ and $r(2) = 0.5$. The discounted factor is set as $\gamma = 0.9$.

The second example is a classical 4×4 Gridworld combined with the slippery mechanism in Frozen-Lake. For each step, the agent can walk in four directions: left, up, right or down. Specially, there are two state A and B in which the agent can only intend to move to A' and B'. After the action is selected by the behavior policy, the agent will walk in the intended direction with probability of 0.9 else will move in either perpendicular direction with equal probability of 0.05 in both directions. If the agent walks out of the space, the agent would get a reward of -1 and stay in the same state. Otherwise, the reward function is also determined by the current state with $r(A) = 10$, $r(B) = 5$ and $r(s) = 0$ for $s \neq A, B$. The discounted factor is set as $\gamma = 0.9$.

Linear function approximation. Our second set of experiments consider Q-learning with linear function approximation. More specifically, we consider approximating the Q-function by a linear subspace spanned by basis vectors $\phi = (\phi_1, \ldots, \phi_d)^\top : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^d$. The goal is to find θ^* such that $\tilde{q}_{\theta^*} := \Phi\theta^*$ best approximates the optimal Q function q^*, where Φ denotes the feature matrix $\Phi = \left[\phi(s_1, a_1) \quad \cdots \quad \phi(s_{|\mathcal{S}|}, a_{|\mathcal{A}|})\right]^\top \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times d}$. We assume that Φ has a full column rank, which is standard in literature (Bertsekas & Tsitsiklis, 1996; Chen et al., 2022b; Melo et al., 2008). Note that θ^* can be calculated by projected value iteration algorithm.

In this case, the Q-earning algorithm reduces to updating the parameter $\theta \in \mathbb{R}^d$ as follows Bertsekas & Tsitsiklis (1996):

$$\theta_{k+1} = \theta_k + \alpha\phi(s_k, a_k)\left(r_k + \gamma \max_{a'} \phi(s_{k+1}, a')^\top \theta_k - \phi(s_k, a_k)^\top \theta_k\right), \tag{37}$$

where (s_k, a_k, r_k, s_{k+1}) is the sample generated by the behavior policy at time step k.

For the MDP and feature vectors, we consider a similar setup as the work (Chen et al., 2022b, Appendix D.1). We provide the detail description here for completeness. We consider an MDP with $|\mathcal{S}| = 20$ states and $|\mathcal{A}| = 5$ actions. We generate the rewards and transition probabilities as follows: for each $(s, a) \in \mathcal{S} \times \mathcal{A}$,

- The reward $r(s, a)$ is drawn uniformly in $[0, 1]$.
- For the transition probability $T(\cdot|s, a)$, we first obtain $|\mathcal{S}|$ numbers by uniformly sampling of $[0, 1]$, and then normalize these $|\mathcal{S}|$ numbers by their sum to make it a valid probability distribution.

As for the feature matrix, we consider $d = 10$. For each (s, a), each element of $\phi(s, a)$ is drawn from Bernoulli distribution with parameter $p = 0.5$, and then we normalize the features to ensure $\|\phi(s, a)\| \leq 1$. We repeat this process until the matrix Φ has a full column rank.

We set the discounted factor to be $\gamma = 0.5$ and the Markovian data $\{x_k\}_{k \geq 0}$ is generated from a uniformly random behavior policy.

We run Q-learning with linear function approximation (37) with initialization $\theta_0^{(\alpha)} = \theta^* + 10$ and stepsize $\alpha \in \{0.1, 0.2, 0.4\}$. We also consider two diminishing stepsizes: $\alpha_k = 1/\big(1 + (1-\gamma)k\big)$ and $\alpha_k = 1/k^{0.75}$ as we used in tabular Q-learning. The simulation results for the Q-learning with linear function approximation are illustrated in Figure 3. We plot the ℓ_1-norm error $\|\bar{\theta}_{k/2,k}^{(\alpha)} - \theta^*\|_1$ for the tail-averaged (TA) iterates $\bar{\theta}_{k/2,k}^{(\alpha)}$, the RR extrapolated iterates $\widetilde{\theta}_k^{(\alpha)}$ with stepsizes α and 2α, and iterates with diminishing stepsizes.

We can observe some similar results as tabular Q-learning's:

- The larger the stepsize α, the faster it converges.
- The final TA error, which corresponds to the asymptotic bias, is approximately proportional to the stepsize.
- RR extrapolated iterates reduce the bias.
- The TA and RR-extrapolated iterates with constant stepsizes enjoy significantly faster initial convergence than those with diminishing stepsizes.

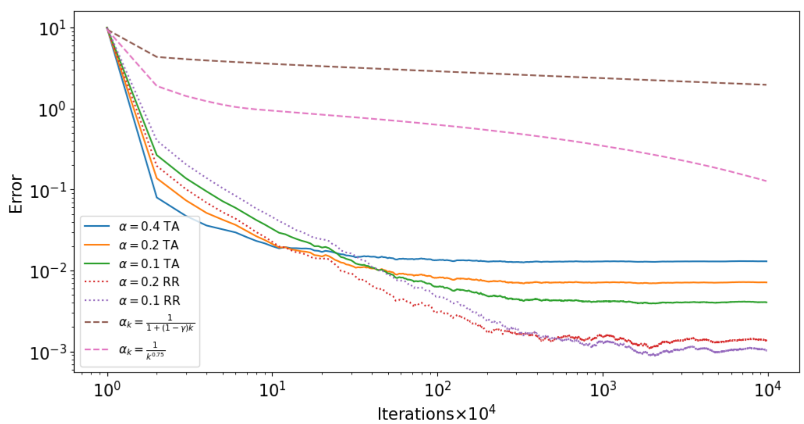


Figure 3: The Q-learning with linear function approximation errors of tail-averaged (TA) iterates and RR extrapolated iterates with different stepsizes.