Proceedings of Machine Learning Research vol 125:1–21, 2020 33rd Annual Conference on Learning Theory

Finite-Time Analysis of Asynchronous Stochastic Approximation and Q-Learning

Guannan Qu GQU@CALTECH.EDU
Adam Wierman ADAMW@CALTECH.EDU
CMS, Caltech

Editors: Jacob Abernethy and Shivani Agarwal

Abstract

We consider a general asynchronous Stochastic Approximation (SA) scheme featuring a weighted infinity-norm contractive operator, and prove a bound on its finite-time convergence rate on a single trajectory. Additionally, we specialize the result to asynchronous Q-learning. The resulting bound matches the sharpest available bound for synchronous Q-learning, and improves over previous known bounds for asynchronous Q-learning.

Keywords: Stochastic approximation, Q-learning, finite time analysis.

1. Introduction

Reinforcement learning (RL) has received renewed interest recently due to its remarkable successes in diverse areas. Many RL algorithms can be viewed through the lens of Stochastic Approximation (SA) (Robbins and Monro, 1951). SA algorithms are widely used beyond RL in areas such as machine learning, stochastic control, signal processing, and communications and, as a result, there is a broad and deep literature focused on the analysis and applications of SA that has developed a rich class of ODE-based tools for proving convergence of SA schemes, e.g., see the books Borkar (2009); Benveniste et al. (2012). In the context of RL, it has been shown that linear SA captures TD-learning and that the ODE-based SA framework can be used to prove the convergence of TD-learning (Tsitsiklis and Van Roy, 1997). A similar connection can be found in the case of actor-critic methods (Konda and Tsitsiklis, 2000, 2003).

Most of the classical analysis in SA is asymptotic in nature; however this has changed recently. Driven by the interest in finite-time convergence of RL methods, the focus has shifted to non-asymptotic analysis of SA schemes. For example, in just the past year, a finite-time bound for linear SA is given in Srikant and Ying (2019), which leads to finite time error bounds for TD-learning, and a finite-time bound for a linear two time scale SA model is given in Gupta et al. (2019); Doan (2019); Xu et al. (2019), which leads to finite-time error bounds for the gradient TD method. These results can be viewed as extensions of the classical ODE-based SA framework, which requires the SA algorithm to admit a “limiting” ODE associated with a Lyapunov function that certifies stability.

While ODE-based approaches are powerful, there are popular classes of nonlinear SA schemes featuring a nonlinear operator with infinity-norm contraction that cannot be directly analyzed from the ODE-based SA framework (Tsitsiklis, 1994; Bertsekas and Tsitsiklis, 1996). This class of SA methods captures a particularly important class of RL methods, the Watkin’s Q-learning method (Watkins and Dayan, 1992), and so understanding the behavior of this class of SA schemes is important for understanding the finite-time behavior of Q-learning. Over the past year, progress

© 2020 G. Qu & A. Wierman.

has been made toward the finite-time analysis of these nonlinear SA schemes. In particular, Shah and Xie (2018) provides a finite-time convergence result for SA with an infinity-norm contractive operator, and Wainwright (2019a) provides sharp convergence rates for SA with a cone-contractive operator. However, both of these works consider the *synchronous* case, i.e., at each time all entries of the iterate are updated. This is a significant limitation since, in many applications, e.g., Q-learning on a single trajectory, the update is *asynchronous*, i.e., only one of the entries is updated at a time. This leads to the following question, which is the focus of this paper:

What is the finite-time convergence rate for asynchronous SA/Q-learning on a single trajectory?

Contribution. In this paper, we provide a finite-time analysis of asynchronous nonlinear SA schemes featuring a weighted infinity norm contraction. We prove an $O\left(\frac{1}{(1-\gamma)^{1.5}}\frac{1}{\sqrt{T}}\right)$ convergence rate in weighted infinity-norm for the SA scheme, where γ is the contraction coefficient (Theorem 4). Notably, our results are sharper than the result in the synchronous case in Shah and Xie (2018, Thm. 5).[1]

As a direct consequence, our result shows a $\tilde{O}(\frac{1}{(1-\gamma)^5}\frac{1}{\epsilon^2})$ convergence time to reach an ε-accurate (measured in infinity-norm) estimate of the Q-function for the asynchronous Q-learning method on a single trajectory in the infinite horizon γ-discounted MDP setting (Theorem 7). This result matches the sharpest known bound for synchronous Q-learning (Wainwright, 2019a), and to the best of our knowledge, improves over the best known finite-time bounds on asynchronous Q-learning (Even-Dar and Mansour, 2003) on a single trajectory in terms of its dependence on $\frac{1}{\varepsilon}$, $\frac{1}{1-\gamma}$, and the state-action space size (cf. Table 1 for comparison). Further, our results clarify a blow-up phenomenon in the asynchronous Q-learning literature where the error can blow up exponentially in $\frac{1}{1-\gamma}$. We show such a blow-up can be avoided by using a rescaled linear step size. This is consistent with related findings in other settings (Jin et al., 2018; Wainwright, 2019a).

Our proof technique is different from those in the literature, e.g., Even-Dar and Mansour (2003); Shah and Xie (2018); Wainwright (2019a). Specifically, we do not use an epoch-based analysis, as in Even-Dar and Mansour (2003); Shah and Xie (2018), where the error is controlled epoch-by-epoch. Instead, we decompose the error in a recursive manner, and this decomposition provides a more transparent approach for analyzing how the stochastic noise impacts the approximation error. This ultimately leads to a sharper bound. Further, our approach for handling asynchronicity is very different from Even-Dar and Mansour (2003) and is partially inspired by the "drift" analysis in the ODE-based SA literature Srikant and Ying (2019).

Related Work. Our results provide new insights about Q-learning and more generally, SA with an infinity-norm contractive operator. Q-learning was first proposed in Watkins and Dayan (1992). Its asymptotic convergence has been proven in Tsitsiklis (1994); Jaakkola et al. (1994), where its connection to SA with infinity-norm contractive operator was established. The first work on non-asymptotic analysis of Q-learning is Szepesvári (1998), which focused on an i.i.d. setting. A generalization beyond the i.i.d. setting was provided by Even-Dar and Mansour (2003), which proves finite-time bounds for synchronous and asynchronous Q-learning with polynomial and linear step sizes. Both Szepesvári (1998) and Even-Dar and Mansour (2003) discover that, when using a linear step size, there is an exponential blow-up in $\frac{1}{1-\gamma}$, where γ is the discounting factor; further, in the asynchronous setting, there is at least cubic dependence on the state-action space size

1. As another related work Wainwright (2019a) does not provide an explicit bound for the synchronous SA scheme, we can only compare with Wainwright (2019a) in the context of Q-learning.

Paper	Sample complexity	Assumption on trajectory
Szepesvári (1998)	$\left(\frac{1}{\varepsilon}\right)^{\frac{1}{R(1-\gamma)}}$	i.i.d.
Even-Dar and Mansour (2003)[2]	$\frac{(\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert)^5}{(1-\gamma)^5\epsilon^{2.5}}$, $\frac{(\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert)^{3.3}}{(1-\gamma)^{5.2}\varepsilon^{2.6}}$, or $\frac{(\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert)^{\frac{1}{1-\gamma}}}{(1-\gamma)^4\varepsilon^2}$	covering time
This work	$\frac{t_{\text{MIX}}(\lvert\mathcal{S}\rvert\lvert\mathcal{A}\rvert)^2}{(1-\gamma)^5\varepsilon^2}$	ergodic

Table 1: Sample complexity for asynchronous Q-learning to compute an ε-optimal Q-function in infinity norm.

(Even-Dar and Mansour, 2003, Thm. 4). Subsequently, Azar et al. (2011) proposes speedy Q-learning, a variant of synchronous Q-learning, by adding a momentum term, and shows it avoids the exponential blow-up with a finite time bound that scales in $\frac{1}{(1-\gamma)^4\epsilon^2}$. More recently, Shah and Xie (2018); Wainwright (2019a) provide finite time bounds for general synchronous SA, which indicates that even in the classical Q-learning setup, the exponential blow-up can be avoided by using a rescaled linear step size. Specifically, Wainwright (2019a) shows a finite time bound for synchronous Q-learning that scales in $\frac{1}{(1-\gamma)^5\epsilon^2}$. To the best of our knowledge, this is the sharpest known bound for synchronous Q-learning. Compared with the above papers, our result bridges the gap between the understanding of *synchronous* SA/Q-learning and *asynchronous* SA/Q-learning. Our finite time bounds for asynchronous Q-learning match the sharpest known scaling in $\frac{1}{(1-\gamma)}$ and $\frac{1}{\varepsilon}$ in synchronous Q-learning. Further, compared with the best known bounds for asynchronous Q-learning (Even-Dar and Mansour, 2003), our result improves the dependence on state-action space size from (at least) cubic to square. Additionally, our work presents a new analytic approach.

Other related work on SA and Q-learning include Lee and He (2019), which combines the ODE-based SA framework with the switch system theory to show the asymptotic convergence of asynchronous Q-learning in an i.i.d. setting; Beck and Srikant (2012), which studies the finite time error bound of constant step size Q-learning; and Melo et al. (2008); Chen et al. (2019), which analyze Q-learning with linear function approximation.

We also mention that there are other lines of work on Q-learning focusing on different models and performance measures. One line of work seeks to propose variants of Q-learning, e.g. recent work Wainwright (2019b) that achieves a minimax optimal rate. Earlier examples include Hasselt (2010); Azar et al. (2013); Sidford et al. (2018a,b); Devraj and Meyn (2017); Kearns and Singh (1999). Compared to these papers, our work focuses on general asynchronous SA and seeks to understand the convergence of the classical form of asynchronous SA/Q-learning. Another related line of work on Q-learning focuses on proving bounds on regret, e.g. Strehl et al. (2006); Jin et al. (2018); Dong et al. (2019); Wei et al. (2019). Regret is a fundamentally different goal than providing finite-time convergence bounds, and the results and techniques across the two communities are quite different. The reason is that regret bound results need to address the problem of exploration, and the performance metric focuses on the transient performance, without the need to approximate *every* entry of Q-function to the same accuracy. In contrast, infinity-norm finite-time error bound results typically assume a form of sufficient exploration (e.g. the i.i.d. assumption used in Szepesvári (1998); Lee and He (2019) and the covering time assumption used in Even-Dar and Mansour (2003)) and require every entry of the Q-function to be accurately estimated.

2. Even-Dar and Mansour (2003) studies different step sizes and obtains different sample complexity bounds.

2. Finite-Time Analysis of Stochastic Approximation

In this section, we present our results on the finite-time analysis of asynchronous SA with a (weighted) infinity-norm contractive operator. We apply the results in this section to Q-learning in Section 3.

To begin, we formally define the problem setting. Let $\mathcal{N} = \{1, \ldots, n\}$, $x \in \mathbb{R}^{\mathcal{N}}$, and $F : \mathbb{R}^{\mathcal{N}} \to \mathbb{R}^{\mathcal{N}}$ is an operator. We use F_i to denote the i'th entry of F. We consider the following stochastic approximation scheme that keeps updating $x(t) \in \mathbb{R}^{\mathcal{N}}$ starting from $x(0)$ being the all zero vector,

$$x_i(t+1) = x_i(t) + \alpha_t(F_i(x(t)) - x_i(t) + w(t)) \qquad \text{for } i = i_t, \tag{1}$$

$$x_i(t+1) = x_i(t) \qquad \text{for } i \neq i_t, \tag{2}$$

where $i_t \in \mathcal{N}$ is a stochastic process adapted to a filtration $\mathcal{F}_t$, and $w(t)$ is some noise that we will discuss later. As we show in Section 3, this stochastic approximation scheme captures the asynchronous Q-learning algorithm.

Given the setting described above, the following assumptions underlie our main result. Similar to Tsitsiklis (1994), the first assumption is concerned with the contraction of F in a weighted infinity norm, which we define in Definition 1. The reason that we consider the weighted infinity norm instead of the standard infinity norm is that its generality will capture not just the discounted case Q-learning, but also the undiscounted case, as shown by Tsitsiklis (1994, Sec. 7).

Definition 1 (Weighted Infinity Norm) *Given a positive vector $v = [v_1, \ldots, v_n]^\top \in \mathbb{R}^{\mathcal{N}}$, the weighted infinity norm $\|\cdot\|_v$ is given by $\|x\|_v = \sup_{i\in\mathcal{N}} \frac{|x_i|}{v_i}$.*

Throughout the rest of the section, we fix a positive vector $v \in \mathbb{R}^n$ and all the norms in the section are in $\|\cdot\|_v$. We also denote $\underline{v} = \inf_{i\in\mathcal{N}} v_i$, the smallest entry of v. We comment that when v is a all one vector, $\|\cdot\|_v$ becomes the standard infinity norm. We use the following result frequently on the induced matrix norm of $\|\cdot\|_v$, the proof of which can be found in Appendix A.1.

Proposition 2 *The induced matrix norm of $\|\cdot\|_v$ for a matrix $A = [a_{ij}]_{i,j\in\mathcal{N}}$ is given by $\|A\|_v = \sup_{i\in\mathcal{N}} \sum_{j\in\mathcal{N}} \frac{v_j}{v_i}|a_{ij}|$. When A is a diagonal matrix, $\|A\|_v = \sup_{i\in\mathcal{N}} |a_{ii}|$.*

With these preparations, we are now ready to state Assumption 1 on the contraction property of F. This assumption is standard in the literature, e.g., (Tsitsiklis, 1994; Wainwright, 2019a),[3] and is satisfied by the Q-learning algorithm as will be shown in Section 3. Note that, as a consequence of Assumption 1, F has a unique fixed point x^*. We also note that we do not require the monotonicity assumption needed in Wainwright (2019a).

Assumption 1 (Contraction) *(a) Operator F is γ contraction in $\|\cdot\|_v$, i.e. for any $x, y \in \mathbb{R}^{\mathcal{N}}$, $\|F(x) - F(y)\|_v \leq \gamma\|x - y\|_v$. (b) There exists some constant $C > 0$ s.t. $\|F(x)\|_v \leq \gamma\|x\|_v + C, \forall x \in \mathbb{R}^{\mathcal{N}}$.*

Assumption 1(a) directly implies Assumption 1(b) with $C = (1+\gamma)\|x^*\|_v$.[4] We write Assumption 1(b) as a separate assumption since, in some applications (e.g. Q-learning), the constant C can be better than $(1+\gamma)\|x^*\|_v$. Our next assumption concerns the noise sequence $w(t)$. It is also standard (Shah and Xie, 2018) and is satisfied by Q-learning.

3. Wainwright (2019a) considers contraction in a gauge norm associated with a cone, which is more general than the weighted infinity norm.
4. To see this, note $\|F(x)\|_v \leq \|F(x) - F(x^*)\|_v + \|F(x^*)\|_v \leq \gamma\|x - x^*\|_v + \|x^*\|_v \leq \gamma\|x\|_v + (1+\gamma)\|x^*\|_v$.

Assumption 2 (Martingale Difference Sequence) *$w(t)$ is $\mathcal{F}_{t+1}$measurable and satisfies $\mathbb{E}w(t)|\mathcal{F}_t = 0$. Further, $|w(t)| \leq \bar{w}$ almost surely for some constant $\bar{w}$.*

Lastly, we make an assumption regarding the stochastic process i_t.

Assumption 3 (Sufficient Exploration) *There exists a $\sigma \in (0,1)$ and positive integer, τ, such that, for any $i \in \mathcal{N}$ and $t \geq \tau$, $\mathbb{P}(i_t = i|\mathcal{F}_{t-\tau}) \geq \sigma$.*

Assumption 3 means that, given the history up to $t-\tau$, the distribution of i_t must have positive probability for every i. Its purpose is to ensure every i is visited by i_t sufficiently often. We note that Assumption 3 is more general than many typical ergodicity assumptions used in the SA literature, e.g., Srikant and Ying (2019). For example, the following proposition shows that if i_t is an ergodic Markov chain on state space $\mathcal{N}$, then Assumption 3 is automatically true with σ and τ depending on the stationary distribution and the mixing time of the Markov chain, where the mixing time refers to the minimum time it takes to reach within $1/4$ total variation distance of the stationary distribution regardless of the initial state (Levin and Peres, 2017, Sec. 4.5). The proof of Proposition 3 can be found in Appendix A.2.

Proposition 3 *If i_t is a ergodic Markov chain on state space $\mathcal{N}$ with stationary distribution μ and mixing time t_{MIX}, then Assumption 3 holds with $\sigma = \frac{1}{2}\mu_{\min}$, where $\mu_{\min} = \min_{i\in\mathcal{N}} \mu_i$, and $\tau = \lceil \log_2(\frac{2}{\mu_{\min}})\rceil t_{\mathrm{MIX}}$.*

With these assumptions, we are ready to state our main result,

Theorem 4 *Suppose Assumptions 1, 2 and 3 hold. Further, assume there exists constant $\bar{x} \geq \|x^*\|_v$ s.t. $\forall t, \|x(t)\|_v \leq \bar{x}$ almost surely. Let the step size be $\alpha_t = \frac{h}{t+t_0}$ with $t_0 \geq \max(4h, \tau)$, and $h \geq \frac{2}{\sigma(1-\gamma)}$. Then, with probability at least $1-\delta$,*

$$\|x(T) - x^*\|_v \leq \frac{12\bar{\epsilon}}{1-\gamma}\sqrt{\frac{(\tau+1)h}{\sigma}}\sqrt{\frac{\log(\frac{2(\tau+1)T^2 n}{\delta})}{T+t_0}} + \frac{4}{1-\gamma}\max(\frac{16\bar{\epsilon}h\tau}{\sigma}, 2\bar{x}(\tau+t_0))\frac{1}{T+t_0},$$

where $\bar{\epsilon} = 2\bar{x} + C + \frac{\bar{w}}{\underline{v}}$.

The assumption in Theorem 4 that $\|x(t)\|_v \leq \bar{x}$ is not necessary. In particular, it can be shown (see Proposition 5 below) that under Assumption 1 and Assumption 2, $\|x(t)\|_v$ can be bounded by some constant almost surely. The proof of Proposition 5 can be found in Appendix A.3. We treat the upper bound on $\|x(t)\|_v$ as a separate assumption because in the Q-learning case, the constant can be better than what is implied in Proposition 5.

Proposition 5 *Suppose Assumptions 1 and 2 hold. Then for all t, $\|x(t)\|_v \leq \frac{1}{1-\gamma}((1+\gamma)\|x^*\|_v + \frac{\bar{w}}{\underline{v}})$ almost surely.*

Theorem 4 shows that, when setting $h = \Theta(\frac{1}{\sigma(1-\gamma)})$ and $t_0 = \Theta(\max(h,\tau))$, $\|x(T) - x^*\|_v \leq \tilde{O}(\frac{\bar{\epsilon}\sqrt{\tau}}{(1-\gamma)^{1.5}\sigma}\frac{1}{\sqrt{T}}) + \tilde{O}(\frac{\bar{\epsilon}\tau}{\sigma^2(1-\gamma)^2}\frac{1}{T})$. This means that, to get an approximation error of ε, the number of time steps required is $T \gtrsim \frac{\bar{\epsilon}^2\tau}{\sigma^2(1-\gamma)^3}\frac{1}{\varepsilon^2}$. Compared to Shah and Xie (2018, Thm. 5), our result improves the dependence on $\frac{1}{\varepsilon}$. Note that Wainwright (2019a) does not provide an explicit approximation bound for the SA scheme, but state the bounds in the context of Q-learning instead. For this reason, we compare to Wainwright (2019a) in the context of Q-learning in Section 3.

We also comment that in the step size $\frac{h}{t+t_0}$ in Theorem 4, it is important for the h constant to scale with $\Theta(\frac{1}{(1-\gamma)\sigma})$ to avoid an exponential blow-up in $\frac{1}{1-\gamma}$. This fact is not apparent in the some of the earlier work like Even-Dar and Mansour (2003), but has been pointed out recently (Jin et al., 2018; Wainwright, 2019a). Specifically, Wainwright (2019a) shows that h needs to grow with $\frac{1}{1-\gamma}$ in the synchronous SA setting. Our result is consistent with Wainwright (2019a) and further shows that in the asynchronous setting, h also needs to scale with $\frac{1}{\sigma}$. If we interpret σ as the fraction of times that each state is visited, then such scaling in $\frac{1}{\sigma}$ will result in step size of $\Theta(\frac{1}{\sigma t})$, which is similar in spirit to a common practice in asynchronous Q-learning, where the step size is coordinate dependent, $\alpha_t = \Theta(\frac{1}{N_{i_t}^t})$ instead of $\Theta(\frac{1}{t})$, where $N_{i_t}^t$ means the number of times i_t has been visited up to time t.

3. Application to Q-learning

We now apply the results for SA to the important special case of Q-learning. The setting we study is defined as follows. We consider a γ-discounted infinite horizon Markov Decision Process (MDP) with finite state space $\mathcal{S}$ and finite action space $\mathcal{A}$. Our SA result applies to both the discounted ($\gamma < 1$) and undiscounted ($\gamma = 1$) case. For the undiscounted case ($\gamma = 1$), it is typically assumed the transition probability of the MDP satisfies some mixing rate assumption, which provides a contraction factor $\rho < 1$ for Assumption 1 to hold in a weighted infinity norm, see e.g. Tsitsiklis (1994) for more details. For ease of presentation, we focus on the discounted case ($\gamma < 1$), where we can let the norm be the standard infinity norm $\|\cdot\|_\infty$, i.e., v is the all-one vector.

Let the transition probability of the MDP be given by $\mathbb{P}(s_{t+1} = s'|s_t = s, a_t = a) = \mathbb{P}(s'|s,a)$. At time t, conditioned on the current state s_t and action a_t, the stage reward is a random variable r_t independently drawn from some fixed distribution depending on (s_t, a_t), with its expectation given by r_{s_t,a_t}, where $r \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ is a deterministic vector. A policy $\pi : \mathcal{S} \to \Delta(\mathcal{A}), s \mapsto \pi(\cdot|s)$ maps the state space to the probability simplex on the action space $\Delta(\mathcal{A})$, and under the policy, a_t is drawn from $\pi(\cdot|s_t)$. Given a policy π, the Q table $Q^\pi : \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ under this policy is,

$$Q^\pi_{s,a} = \mathbb{E}_\pi \big[\sum_{t=0}^{\infty} \gamma^t r_t | (s_0, a_0) = (s,a) \big],$$

where $\mathbb{E}_\pi$ means the expectation is taken with a_t drawn from $\pi(\cdot|s_t)$. The MDP problem seeks to find an optimal policy π^* such that $Q^\pi(s,a)$ is maximized simultaneously for all (s,a). Classical MDP theory (Bertsekas and Tsitsiklis, 1996) guarantees that such a π^* must exist and, further, the resulting Q-function, which we denote as Q^*, is the unique fixed point of the Bellman Operator $F : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ given by,

$$F_{s,a}(Q) = r_{s,a} + \gamma \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s,a)} \max_{a'\in\mathcal{A}} Q_{s',a'}. \tag{3}$$

Once Q^* is known, an optimal policy can be easily determined (Bertsekas and Tsitsiklis, 1996).

When the transition probabilities and the rewards are unknown, we cannot directly use (3) to calculate Q^*. The Q-learning algorithm is an off-policy learning algorithm that approximates Q^*.

In the asynchronous version of Q-learning, we sample a trajectory $\{(s_t, a_t, r_t)\}_{t=0}^{\infty}$ by taking a behavioral policy π. In this process, we maintain a Q table $Q(t)$, which is initialized with $Q(0)$ being the all-zero table, and is updated upon observing every new state action pair (s_{t+1}, a_{t+1}) using the following update rule,

$$Q_{s_t,a_t}(t+1) = (1-\alpha_t)Q_{s_t,a_t}(t) + \alpha_t[r_t + \gamma \max_{a\in\mathcal{A}} Q_{s_{t+1},a}(t)], \tag{4}$$

$$Q_{s,a}(t+1) = Q_{s,a}(t) \text{ for } (s,a) \neq (s_t, a_t). \tag{5}$$

Our results make the following standard assumptions regarding the MDP. Assumption 4(a) is an upper bound on the reward, and Assumption 4(b) is to ensure the sufficient exploration condition in Assumption 3 holds (cf. Proposition 3).[5] In the asynchronous Q-learning literature, it is common to require some type of sufficient exploration assumption. Assumption 4(b) is more general than the i.i.d. assumption in Szepesvári (1998); Lee and He (2019), and is similar in spirit to the covering time assumption in Even-Dar and Mansour (2003) and another related assumption in Beck and Srikant (2012).

Assumption 4 *The following conditions hold.*

(a) For all t, the stage reward r_t is upper bounded, $|r_t| \leq \bar{r}$ almost surely.

(b) Under the behavioral policy π, the induced Markov chain with state (s_t, a_t) is ergodic, has a stationary distribution μ and mixing time t_{MIX}. Further, define $\mu_{\min} = \inf_{s,a} \mu_{s,a} > 0$.

We now show that under this assumption, the Q-learning updates (4) and (5) can be written in the form of (1) and (2) and meet Assumptions 1, 2, 3. We first identify $\mathcal{N} = \mathcal{S} \times \mathcal{N}$, $i_t = (s_t, a_t)$, and $Q(t)$ with $x(t)$. We let $\mathcal{F}_t$ be the σ-algebra generated by $(s_0, a_0, r_0, \ldots, s_{t-1}, a_{t-1}, r_{t-1}, s_t, a_t)$. Then, clearly (s_t, a_t) is $\mathcal{F}_t$ measurable. We also define

$$\begin{aligned} w(t) &:= r_t + \gamma \max_{a\in\mathcal{A}} Q_{s_{t+1},a}(t) - F_{s_t,a_t}(Q(t)) \\ &= r_t - r_{s_t,a_t} + \gamma \max_{a\in\mathcal{A}} Q_{s_{t+1},a}(t) - \gamma \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)} \max_{a\in\mathcal{A}} Q_{s',a}(t). \end{aligned}$$

Then, (4) can be written as,

$$Q_{s_t,a_t}(t+1) = Q_{s_t,a_t}(t) + \alpha_t[F_{s_t,a_t}(Q(t)) + w(t) - Q_{s_t,a_t}(t)],$$

which shows the Q-learning algorithm (4) and (5) can be written in the form of (1) and (2). We then check Assumptions 1, 2, 3. For Assumption 1, it is known that the Bellman Operator F is a γ-contraction in infinity norm (Tsitsiklis, 1994); further, it easy to check $\|F(Q)\|_\infty \leq \bar{r} + \gamma\|Q\|_\infty$, and hence Assumption 1 is met with $C = \bar{r}$. For Assumption 2, clearly $w(t)$ is $\mathcal{F}_{t+1}$-measurable, and satisfies $\mathbb{E}w(t)|\mathcal{F}_t = 0$. For the boundedness of $w(t)$, we have the following proposition, which completes the verification of Assumption 2. The proof of Proposition 6 can be found in Appendix A.4.

5. Assumption 4(b) is a simple sufficient condition that leads to Assumption 3, but it is not necessary. For example, Assumption 3 does not even require the exploratory policy to be stationary.

Proposition 6 *Under Assumption 4, the Q-learning update satisfies the following. (a) For all t,* $\|Q(t)\|_\infty \le \bar{x} := \frac{\bar{r}}{1-\gamma}$ *almost surely; also,* $\|Q^*\|_\infty \le \bar{x}$. *(b) For all t,* $|w(t)| \le \bar{w} := \frac{2\bar{r}}{1-\gamma}$ *almost surely.*

Finally, using Assumption 4(b) and Proposition 3, we have that Assumption 3 holds with $\sigma = \frac{1}{2}\mu_{\min}$ and $\tau = \lceil \log_2 \frac{2}{\mu_{\min}} \rceil t_{\text{MIX}}$.

Combining the three assumptions together with the upper bound on $\|Q(t)\|_\infty$ in Proposition 6(a), we can directly apply Theorem 4 and obtain the following finite-time error bounds for Q-learning.

Theorem 7 *Suppose Assumption 4 holds and the step size is taken to be* $\alpha_t = \frac{h}{t+t_0}$ *with* $t_0 \ge \max(4h, \lceil \log_2 \frac{2}{\mu_{\min}} \rceil t_{\text{MIX}})$ *and* $h \ge \frac{4}{\mu_{\min}(1-\gamma)}$. *Then, with probability at least* $1-\delta$,

$$\|Q(T) - Q^*\|_\infty \le \frac{60\bar{r}}{(1-\gamma)^2}\sqrt{\frac{2(\lceil \log_2 \frac{2}{\mu_{\min}} \rceil t_{\text{MIX}} + 1)h}{\mu_{\min}}}\sqrt{\frac{\log(\frac{2(\lceil \log_2 \frac{2}{\mu_{\min}} \rceil t_{\text{MIX}}+1)T^2|\mathcal{S}||\mathcal{A}|}{\delta})}{T+t_0}}$$
$$+ \frac{4\bar{r}}{(1-\gamma)^2}\max\Big(\frac{160h\lceil \log_2 \frac{2}{\mu_{\min}} \rceil t_{\text{MIX}}}{\mu_{\min}}, 2(\lceil \log_2 \frac{2}{\mu_{\min}} \rceil t_{\text{MIX}} + t_0)\Big)\frac{1}{T+t_0}.$$

From the above theorem, if we take $h = \Theta(\frac{1}{\mu_{\min}(1-\gamma)})$, $t_0 = \tilde{\Theta}(\max(\frac{1}{\mu_{\min}(1-\gamma)}, t_{\text{MIX}}))$, the convergence rate becomes $\tilde{O}(\frac{\bar{r}\sqrt{t_{\text{MIX}}}}{(1-\gamma)^{5/2}\mu_{\min}}\frac{1}{\sqrt{T}} + \frac{\bar{r}t_{\text{MIX}}}{(1-\gamma)^3\mu_{\min}^2}\frac{1}{T})$. Therefore, to reach a ε accuracy in infinity norm, it takes $T \gtrsim \frac{\bar{r}^2 t_{\text{MIX}}}{(1-\gamma)^5\mu_{\min}^2}\frac{1}{\varepsilon^2}$ iterations. This bound matches the best known dependence on $\frac{1}{1-\gamma}$ and $\frac{1}{\varepsilon}$ in *synchronous* Q-learning (Wainwright, 2019a). The extra factor $\frac{t_{\text{MIX}}}{\mu_{\min}^2}$ is a result of the asynchronous updates. If we interpret $\frac{1}{\mu_{\min}}$ to scale with $|\mathcal{S}| \times |\mathcal{A}|$ (the state-action space size), the extra factor becomes $t_{\text{MIX}}(|\mathcal{S}| \times |\mathcal{A}|)^2$. We believe the scaling in t_{MIX} is inevitable. When compared with the results on *asynchronous* Q-learning, to the best of our knowledge, the best finite-time bound is that of Even-Dar and Mansour (2003, Thm. 4), where the scaling is $\frac{(|\mathcal{S}||\mathcal{A}|)^5}{(1-\gamma)^5\epsilon^{2.5}}$ when $\omega = 4/5$ (optimizing dependence on $\frac{1}{1-\gamma}$), or $\frac{(|\mathcal{S}||\mathcal{A}|)^{3.3}}{(1-\gamma)^{5.2}\epsilon^{2.6}}$ when $\omega = 0.77$ (optimizing dependence on $|\mathcal{S}||\mathcal{A}|$).[6] Here ω is a step size parameter in Even-Dar and Mansour (2003). While our result improves the dependence on $\frac{1}{\varepsilon}, \frac{1}{1-\gamma}, (|\mathcal{S}||\mathcal{A}|)$ over that of Even-Dar and Mansour (2003), we believe our square dependence on the state-action space size is not optimal. We leave it as future work to investigate whether this is an intrinsic property of the algorithm or it is an artifact of the proof.

4. Convergence Proof

In this section, we prove our main result, Theorem 4. The proof is divided into three steps. In the first step, we manipulate the update equation ((1) and (2)) and decompose the error in a recursive form, which provides a transparent view of how the stochastic noise affects the error. In the second step, we bound the contribution of the noise sequence to the error decomposition. In the third step, we use the error decomposition and the noise sequence bounds to prove the result.

Step 1: Decomposition of Error. Let $\mathbf{e}_i$ to be the unit vector (the i'th entry is 1 and others are zero). We let $D_t = \mathbb{E}\mathbf{e}_{i_t}\mathbf{e}_{i_t}^\top|\mathcal{F}_{t-\tau}$. Then, it is clear D_t is a $\mathcal{F}_{t-\tau}$-measurable n-by-n diagonal random matrix, with its i'th entry being $d_{t,i} = \mathbb{P}(i_t = i|\mathcal{F}_{t-\tau})$. By Assumption 3, we have

$$d_{t,i} \ge \sigma \text{ almost surely.} \tag{6}$$

6. Notably, Even-Dar and Mansour (2003) uses a different assumption on sufficient exploration.

With these definitions, we can rewrite the update equation (1) and (2) as follows,

$$\begin{aligned}
x(t+1) &= x(t) + \alpha_t[\mathbf{e}_{i_t}^\top F(x(t)) - \mathbf{e}_{i_t}^\top x(t) + w(t)]\mathbf{e}_{i_t} \\
&= x(t) + \alpha_t[\mathbf{e}_{i_t}\mathbf{e}_{i_t}^\top (F(x(t)) - x(t)) + w(t)\mathbf{e}_{i_t}] \\
&= x(t) + \alpha_t D_t(F(x(t)) - x(t)) + \alpha_t[(\mathbf{e}_{i_t}\mathbf{e}_{i_t}^\top - D_t)(F(x(t)) - x(t)) + w(t)\mathbf{e}_{i_t}] \\
&= x(t) + \alpha_t[D_t F(x(t)) - D_t x(t)] \\
&\quad + \alpha_t \underbrace{\left[(\mathbf{e}_{i_t}\mathbf{e}_{i_t}^\top - D_t)(F(x(t-\tau)) - x(t-\tau)) + w(t)\mathbf{e}_{i_t}\right]}_{:=\epsilon(t)} \\
&\quad + \alpha_t \underbrace{(\mathbf{e}_{i_t}\mathbf{e}_{i_t}^\top - D_t)[F(x(t)) - F(x(t-\tau)) - (x(t) - x(t-\tau))]}_{:=\phi(t)} \\
&= (I - \alpha_t D_t)x(t) + \alpha_t D_t F(x(t)) + \alpha_t(\epsilon(t) + \phi(t)).
\end{aligned} \tag{7}$$

Clearly, $x(t)$ is $\mathcal{F}_t$ measurable and $\epsilon(t)$ is $\mathcal{F}_{t+1}$ measurable (as $\epsilon(t)$ depends on $w(t)$, which is $\mathcal{F}_{t+1}$ measurable). Further,

$$\mathbb{E}\epsilon(t)|\mathcal{F}_{t-\tau} = \mathbb{E}[(\mathbf{e}_{i_t}\mathbf{e}_{i_t}^\top - D_t)|\mathcal{F}_{t-\tau}][F(x(t-\tau)) - x(t-\tau)] + \mathbb{E}[\mathbb{E}[w(t)|\mathcal{F}_t]\mathbf{e}_{i_t}|\mathcal{F}_{t-\tau}] = 0. \tag{8}$$

In other words, $\epsilon(t)$ is like a "shifted" martingale difference sequence, where here "shifted" means the conditioning in (8) is with respect to $\mathcal{F}_{t-\tau}$ instead of $\mathcal{F}_t$ as would be the case in a standard martingale difference sequence. Property (8) will be useful later in the proof. For now, we focus on (7) and expand it recursively, getting,

$$\begin{aligned}
x(t+1) &= \prod_{k=\tau}^{t}(I - \alpha_k D_k)x(\tau) + \sum_{k=\tau}^{t}\alpha_k D_k \prod_{\ell=k+1}^{t}(I - \alpha_\ell D_\ell)F(x(k)) + \sum_{k=\tau}^{t}\alpha_k \prod_{\ell=k+1}^{t}(I - \alpha_\ell D_\ell)(\epsilon(k) + \phi(k)) \\
&= \tilde{B}_{\tau-1,t}x(\tau) + \sum_{k=\tau}^{t} B_{k,t}F(x(k)) + \sum_{k=\tau}^{t}\alpha_k \tilde{B}_{k,t}(\epsilon(k) + \phi(k)),
\end{aligned} \tag{9}$$

where we have defined, $B_{k,t} = \alpha_k D_k \prod_{\ell=k+1}^{t}(I - \alpha_\ell D_\ell)$, $\tilde{B}_{k,t} = \prod_{\ell=k+1}^{t}(I - \alpha_\ell D_\ell)$. Clearly, $B_{k,t}$ and $\tilde{B}_{k,t}$ are n-by-n diagonal random matrices, with the i'th diagonal entry given by $b_{k,t,i}$ and $\tilde{b}_{k,t,i}$, where $b_{k,t,i} = \alpha_k d_{k,i}\prod_{\ell=k+1}^{t}(1 - \alpha_\ell d_{\ell,i})$ and $\tilde{b}_{k,t,i} = \prod_{\ell=k+1}^{t}(1 - \alpha_\ell d_{\ell,i})$. By (6), we have for any i, almost surely

$$b_{k,t,i} \le \beta_{k,t} := \alpha_k \prod_{\ell=k+1}^{t}(1 - \alpha_\ell \sigma), \quad \tilde{b}_{k,t,i} \le \tilde{\beta}_{k,t} = \prod_{\ell=k+1}^{t}(1 - \alpha_\ell \sigma). \tag{10}$$

With these preparations, we are ready to state the following Lemma, which decomposes the error $\|x(t) - x^*\|_v$ in a recursive form. The proof of Lemma 8 can be found in Appendix B.1.

Lemma 8 *Let $a_t = \|x(t) - x^*\|_v$, we have almost surely,*

$$a_{t+1} \le \tilde{\beta}_{\tau-1,t}a_\tau + \gamma \sup_{i\in\mathcal{N}} \sum_{k=\tau}^{t} b_{k,t,i}a_k + \Big\|\sum_{k=\tau}^{t}\alpha_k \tilde{B}_{k,t}\epsilon(k)\Big\|_v + \Big\|\sum_{k=\tau}^{t}\alpha_k \tilde{B}_{k,t}\phi(k)\Big\|_v.$$

From Lemma 8, it is clear that to control the error a_t, we need to bound $\|\sum_{k=\tau}^{t}\alpha_k \tilde{B}_{k,t}\epsilon(k)\|_v$ and $\|\sum_{k=\tau}^{t}\alpha_k \tilde{B}_{k,t}\phi(k)\|_v$, which will be the focus of the next step.

Step 2: Bounding $\|\sum_{k=\tau}^{t}\alpha_k \tilde{B}_{k,t}\epsilon(k)\|_v$ and $\|\sum_{k=\tau}^{t}\alpha_k \tilde{B}_{k,t}\phi(k)\|_v$. We start with a bound on each individual $\epsilon(k)$ and $\phi(k)$ in the following lemma, proven in Appendix B.2.

Lemma 9 *The following bounds hold almost surely. (a)* $\|\epsilon(t)\|_v \leq \bar{\epsilon} := 2\bar{x} + C + \frac{\bar{w}}{\underline{v}}$. *(b)* $\|\phi(t)\|_v \leq \sum_{k=t-\tau+1}^{t} 2\bar{\epsilon}\alpha_{k-1}$.

To bound $\|\sum_{k=\tau}^{t} \alpha_k \tilde{B}_{k,t}\epsilon(k)\|_v$ and $\|\sum_{k=\tau}^{t} \alpha_k \tilde{B}_{k,t}\phi(k)\|_v$, we also need to understand the behavior of α_k and $\tilde{B}_{k,t}$. Recall that, by (10), each entry of $B_{k,t}$ and $\tilde{B}_{k,t}$ are upper bounded by $\beta_{k,t}$ and $\tilde{\beta}_{k,t}$ respectively. We now provide the following results on the sequence $\beta_{k,t}$, $\tilde{\beta}_{k,t}$ which we will frequently use later to control $\alpha_k \tilde{B}_{k,t}$. The proof of Lemma 10 is provided in Appendix B.3.

Lemma 10 *If* $\alpha_t = \frac{h}{t+t_0}$, *where* $h > \frac{2}{\sigma}$ *and* $t_0 \geq \max(4h, \tau)$, *then* $\beta_{k,t}, \tilde{\beta}_{k,t}$ *satisfies the following.*

(a) $\beta_{k,t} \leq \frac{h}{k+t_0}\left(\frac{k+1+t_0}{t+1+t_0}\right)^{\sigma h}$, $\tilde{\beta}_{k,t} \leq \left(\frac{k+1+t_0}{t+1+t_0}\right)^{\sigma h}$.

(b) $\sum_{k=1}^{t} \beta_{k,t}^2 \leq \frac{2h}{\sigma}\frac{1}{(t+1+t_0)}$.

(c) $\sum_{k=\tau}^{t} \beta_{k,t} \sum_{\ell=k-\tau+1}^{k} \alpha_{\ell-1} \leq \frac{8h\tau}{\sigma}\frac{1}{t+1+t_0}$.

We are now ready to bound $\|\sum_{k=\tau}^{t} \alpha_k \tilde{B}_{k,t}\epsilon(k)\|_v$ and $\|\sum_{k=\tau}^{t} \alpha_k \tilde{B}_{k,t}\phi(k)\|_v$. Our bound on $\|\sum_{k=\tau}^{t} \alpha_k \tilde{B}_{k,t}\phi(k)\|_v$ is an immediate consequence of Lemma 9 (b) and Lemma 10 (c).

Lemma 11 *The following inequality holds almost surely,*

$$\Big\|\sum_{k=\tau}^{t} \alpha_k \tilde{B}_{k,t}\phi(k)\Big\|_v \leq \frac{16\bar{\epsilon}h\tau}{\sigma}\frac{1}{t+1+t_0} := C_\phi \frac{1}{t+1+t_0}.$$

Proof We have $\|\sum_{k=\tau}^{t} \alpha_k \tilde{B}_{k,t}\phi(k)\|_v \leq \sum_{k=\tau}^{t} \alpha_k \|\tilde{B}_{k,t}\|_v \|\phi(k)\|_v \leq \sum_{k=\tau}^{t} \beta_{k,t} \sum_{\ell=k-\tau+1}^{k} 2\bar{\epsilon}\alpha_{\ell-1} \leq \frac{16\bar{\epsilon}h\tau}{\sigma(t+t_0+1)}$. Here we have used by Proposition 2, $\|\tilde{B}_{k,t}\|_v = \sup_i |\tilde{b}_{k,t,i}| \leq \tilde{\beta}_{k,t}$. ■

Lemma 12 *For each t, with probability at least* $1-\delta$, *we have,*

$$\Big\|\sum_{k=\tau}^{t} \alpha_k \tilde{B}_{k,t}\epsilon(k)\Big\|_v \leq 6\bar{\epsilon}\sqrt{\frac{(\tau+1)h}{\sigma(t+1+t_0)}\log(\frac{2(\tau+1)tn}{\delta})}.$$

We now focus on proving Lemma 12. Recall $\epsilon(t)$ is $\mathcal{F}_{t+1}$ measurable is a "shifted" martingale difference sequence in the sense that $\mathbb{E}\epsilon(t)|\mathcal{F}_{t-\tau} = 0$ (cf. (8)). We will use a variant of the Azuma-Hoeffding bound in Lemma 13 that handles our "shifted" Martingale difference sequence. The proof of Lemma 13 is postponed to Appendix B.4.

Lemma 13 *Let* X_t *be a* $\mathcal{F}_t$*-adapted stochastic process, satisfying* $\mathbb{E}X_t|\mathcal{F}_{t-\tau} = 0$. *Further,* $|X_t| \leq \bar{X}_t$ *almost surely. Then with probability* $1-\delta$, *we have,* $|\sum_{k=0}^{t} X_k| \leq \sqrt{2\tau \sum_{k=0}^{t} \bar{X}_k^2 \log(\frac{2\tau}{\delta})}$.

To prove Lemma 12, recall that $\sum_{k=\tau}^{t} \alpha_k \tilde{B}_{k,t}\epsilon(k)$ is a random vector in $\mathbb{R}^{\mathcal{N}}$, with its i'th entry

$$\sum_{k=\tau}^{t} \alpha_k \epsilon_i(k) \prod_{\ell=k+1}^{t} (1 - \alpha_\ell d_{\ell,i}), \tag{11}$$

with $d_{\ell,i} \geq \sigma$ almost surely, cf. (6). Fixing i, as have been shown in (8), $\epsilon_i(k)$ is a $\mathcal{F}_{k+1}$ adapted stochastic process satisfying $\mathbb{E}\epsilon_i(k)|\mathcal{F}_{k-\tau} = 0$. However, $\prod_{\ell=k+1}^t (1-\alpha_\ell d_{\ell,i})$ is not $\mathcal{F}_{k-\tau}$-measurable, and as such we cannot directly apply the Azuma-Hoeffding bound in Lemma 13 to (11). To proceed, we need to get rid of the randomness of $\prod_{\ell=k+1}^t (1-\alpha_\ell d_{\ell,i})$ in the summation (11).This is done in Lemma 14 which shows that the absolute value of quantity (11) can be upper bounded by the sup of another quantity where the randomness caused by $\prod_{\ell=k+1}^t (1-\alpha_\ell d_{\ell,i})$ is removed through the use of $d_{\ell,i} \geq \sigma$, and to this new quantity we can directly apply Lemma 13. The proof of Lemma 14 is postponed to Appendix B.5.

Lemma 14 *For each i, we have almost surely,*

$$\Big|\sum_{k=\tau}^t \alpha_k \epsilon_i(k) \prod_{\ell=k+1}^t (1-\alpha_\ell d_{\ell,i})\Big| \leq \sup_{\tau \leq k_0 \leq t} \bigg(\Big|\sum_{k=k_0+1}^t \epsilon_i(k)\beta_{k,t}\Big| + 2\bar{\epsilon} v_i \beta_{k_0,t} \bigg).$$

With the help of Lemma 14, we use the Azuma-Hoeffding bound to prove Lemma 12.

Proof of Lemma 12. Fix i and $\tau \leq k_0 \leq t$. As have been shown in (8), $\frac{1}{v_i}\epsilon_i(k)\beta_{k,t}$ is a $\mathcal{F}_{k+1}$ adapted stochastic process satisfying $\mathbb{E}\frac{1}{v_i}\epsilon_i(k)\beta_{k,t}|\mathcal{F}_{k-\tau} = 0$. Also by Lemma 9(a), $|\frac{1}{v_i}\epsilon_i(k)\beta_{k,t}| \leq \bar{\epsilon}\beta_{k,t}$ almost surely. As a result, we can use the Azuma-Hoeffding bound in Lemma 13 to get with probability $1-\delta$,

$$\Big|\sum_{k=k_0+1}^t \frac{1}{v_i}\epsilon_i(k)\beta_{k,t}\Big| \leq \bar{\epsilon}\sqrt{2(\tau+1)\sum_{k=k_0+1}^t \beta_{k,t}^2 \log(\frac{2(\tau+1)}{\delta})}.$$

By a union bound on $\tau \leq k_0 \leq t$, we get with probability $1-\delta$,

$$\frac{1}{v_i}\sup_{\tau \leq k_0 \leq t}\Big|\sum_{k=k_0+1}^t \epsilon_i(k)\beta_{k,t}\Big| \leq \bar{\epsilon}\sqrt{2(\tau+1)\sum_{k=\tau+1}^t \beta_{k,t}^2 \log(\frac{2(\tau+1)t}{\delta})}.$$

Then, by Lemma 14, we have with probability $1-\delta$,

$$\begin{aligned}
\frac{1}{v_i}\Big|\sum_{k=\tau}^t \alpha_k \epsilon_i(k) \prod_{\ell=k+1}^t (1-\alpha_\ell d_{\ell,i})\Big| &\leq \sup_{\tau \leq k_0 \leq t}\bigg(\frac{1}{v_i}\Big|\sum_{k=k_0+1}^t \epsilon_i(k)\beta_{k,t}\Big| + 2\bar{\epsilon}\beta_{k_0,t}\bigg) \\
&\leq \bar{\epsilon}\sqrt{2(\tau+1)\sum_{k=\tau+1}^t \beta_{k,t}^2 \log(\frac{2(\tau+1)t}{\delta})} + \sup_{\tau \leq k_0 \leq t} 2\bar{\epsilon}\beta_{k_0,t} \\
&\leq 2\bar{\epsilon}\sqrt{\frac{(\tau+1)h}{\sigma(t+1+t_0)}\log(\frac{2(\tau+1)t}{\delta})} + \sup_{\tau \leq k_0 \leq t} 2\bar{\epsilon}\frac{h}{k_0+t_0}\Big(\frac{k_0+1+t_0}{t+1+t_0}\Big)^{\sigma h} \\
&\leq 2\bar{\epsilon}\sqrt{\frac{(\tau+1)h}{\sigma(t+1+t_0)}\log(\frac{2(\tau+1)t}{\delta})} + 2\bar{\epsilon}\frac{h}{t+t_0} \\
&\leq 6\bar{\epsilon}\sqrt{\frac{(\tau+1)h}{\sigma(t+1+t_0)}\log(\frac{2(\tau+1)t}{\delta})},
\end{aligned}$$

where in the third inequality, we have used the bounds on $\beta_{k,t}$ in Lemma 10; the last inequality is due to that $\frac{1}{t+t_0}$ is asymptotically smaller than $\sqrt{\frac{1}{t+1+t_0}}$. Finally, applying the union bound over $i \in \mathcal{N}$ will lead to the desired result. ■

Step 3: Bounding the error sequence. We are now ready to use the error decomposition in Lemma 8 and the bound on $\|\sum_{k=\tau}^{t}\alpha_k\tilde{B}_{k,t}\epsilon(k)\|_v$ and $\|\sum_{k=\tau}^{t}\alpha_k\tilde{B}_{k,t}\phi(k)\|_v$ in Lemma 12 and Lemma 11 to bound $a_t=\|x(t)-x^*\|_v$. Recall, we want to show that, with probability $1-\delta$,

$$a_T\le\frac{C_a}{\sqrt{T+t_0}}+\frac{C_a'}{T+t_0},\tag{12}$$

where $C_a=\frac{12\bar{\epsilon}}{1-\gamma}\sqrt{\frac{(\tau+1)h}{\sigma}\log(\frac{2(\tau+1)T^2n}{\delta})}$, $C_a'=\frac{4}{1-\gamma}\max(C_\phi,2\bar{x}(\tau+t_0))$. To prove (12), we start by applying Lemma 12 to $t\le T$ with δ replaced by δ/T. Then, using a union bound, we get with probability $1-\delta$, for any $t\le T$, $\|\sum_{k=\tau}^{t}\alpha_k\tilde{B}_{k,t}\epsilon(k)\|_v\le C_\epsilon\frac{1}{\sqrt{t+1+t_0}}$, where $C_\epsilon=6\bar{\epsilon}\sqrt{\frac{(\tau+1)h}{\sigma}\log(\frac{2(\tau+1)T^2n}{\delta})}$. Combine the above with Lemma 8 and use Lemma 11, we get with probability $1-\delta$, for all $\tau\le t\le T$,

$$\begin{aligned}a_{t+1}&\le\tilde{\beta}_{\tau-1,t}a_\tau+\gamma\sup_{i\in\mathcal{N}}\sum_{k=\tau}^{t}b_{k,t,i}a_k+\|\sum_{k=\tau}^{t}\alpha_k\tilde{B}_{k,t}\epsilon(k)\|_v+\|\sum_{k=\tau}^{t}\alpha_k\tilde{B}_{k,t}\phi(k)\|_v\\&\le\tilde{\beta}_{\tau-1,t}a_\tau+\gamma\sup_{i\in\mathcal{N}}\sum_{k=\tau}^{t}b_{k,t,i}a_k+\frac{C_\epsilon}{\sqrt{t+1+t_0}}+\frac{C_\phi}{t+1+t_0}.\end{aligned}\tag{13}$$

We now condition on (13) and use induction to show (12). Eq. (12) is true for $t=\tau$, as $\frac{C_a'}{\tau+t_0}\ge\frac{8}{1-\gamma}\bar{x}\ge a_\tau$, where we have used $a_\tau=\|x(\tau)-x^*\|_v\le\|x(\tau)\|_v+\|x^*\|_v\le 2\bar{x}$ by the definition of $\bar{x}$. Then, assuming (12) is true for up to $k\le t$, we have by (13),

$$\begin{aligned}a_{t+1}&\le\tilde{\beta}_{\tau-1,t}a_\tau+\gamma\sup_{i\in\mathcal{N}}\sum_{k=\tau}^{t}b_{k,t,i}[\frac{C_a}{\sqrt{k+t_0}}+\frac{C_a'}{k+t_0}]+C_\epsilon\frac{1}{\sqrt{t+1+t_0}}+C_\phi\frac{1}{t+1+t_0}\\&\le\tilde{\beta}_{\tau-1,t}a_\tau+\gamma C_a\sup_{i\in\mathcal{N}}\sum_{k=\tau}^{t}b_{k,t,i}\frac{1}{\sqrt{k+t_0}}+\gamma C_a'\sup_{i\in\mathcal{N}}\sum_{k=\tau}^{t}b_{k,t,i}\frac{1}{k+t_0}+C_\epsilon\frac{1}{\sqrt{t+1+t_0}}+C_\phi\frac{1}{t+1+t_0}.\end{aligned}$$

We use the following auxiliary Lemma, whose proof is provided in Appendix B.6.

Lemma 15 *Recall $\alpha_k=\frac{h}{k+t_0}$, and $b_{k,t,i}=\alpha_k d_{k,i}\prod_{\ell=k+1}^{t}(1-\alpha_\ell d_{\ell,i})$, here $d_{k,i}\ge\sigma$. If $\sigma h(1-\sqrt{\gamma})\ge 1$, $t_0\ge 1$, and $\alpha_0\le\frac{1}{2}$, then, for any $i\in\mathcal{N}$, and any $0<\omega\le 1$, we have $\sum_{k=\tau}^{t}b_{k,t,i}\frac{1}{(k+t_0)^\omega}\le\frac{1}{\sqrt{\gamma}(t+1+t_0)^\omega}$.*

With Lemma 15, and using the bound on $\tilde{\beta}_{\tau-1,t}$ in Lemma 10 (a), we have

$$\begin{aligned}a_{t+1}&\le\tilde{\beta}_{\tau-1,t}a_\tau+\sqrt{\gamma}C_a\frac{1}{\sqrt{t+1+t_0}}+\sqrt{\gamma}C_a'\frac{1}{t+1+t_0}+C_\epsilon\frac{1}{\sqrt{t+1+t_0}}+C_\phi\frac{1}{t+1+t_0}\\&\le\underbrace{\sqrt{\gamma}C_a\frac{1}{\sqrt{t+1+t_0}}+C_\epsilon\frac{1}{\sqrt{t+1+t_0}}}_{:=F_t}+\underbrace{\sqrt{\gamma}C_a'\frac{1}{t+1+t_0}+C_\phi\frac{1}{t+1+t_0}+\Big(\frac{\tau+t_0}{t+1+t_0}\Big)^{\sigma h}a_\tau}_{:=F_t'}.\end{aligned}$$

To finish the induction, it suffices to show $F_t\le\frac{C_a}{\sqrt{t+1+t_0}}$ and $F_t'\le\frac{C_a'}{t+1+t_0}$. To see this,

$$F_t\frac{\sqrt{t+1+t_0}}{C_a}=\sqrt{\gamma}+\frac{C_\epsilon}{C_a},\quad F_t'\frac{t+1+t_0}{C_a'}=\sqrt{\gamma}+\frac{C_\phi}{C_a'}+\frac{a_\tau(\tau+t_0)}{C_a'}\frac{(\tau+t_0)^{\sigma h-1}}{(t+1+t_0)^{\sigma h-1}}.$$

It suffices to show that, $\frac{C_\epsilon}{C_a}\le 1-\sqrt{\gamma}$, $\frac{C_\phi}{C_a'}\le\frac{1-\sqrt{\gamma}}{2}$, and $\frac{a_\tau(\tau+t_0)}{C_a'}\le\frac{1-\sqrt{\gamma}}{2}$. Using $a_\tau\le 2\bar{x}$, one can check that C_a and C_a' satisfy the above three inequalities, which concludes the proof. ■

Acknowledgments

The research was supported by Resnick Sustainability Institute Fellowship, ONR YIP: N00014-19-1-2217, the PIMCO Fellowship, and NSF grants AitF-1637598, CNS-1518941. We would also like to thank Prof. Na Li for her valuable suggestions during the planning and development of this research work.

References

Mohammad Gheshlaghi Azar, Rémi Munos, Mohammad Ghavamzadeh, and Hilbert Kappen. Speedy Q-learning. In *Advances in Neural Information Processing Systems*, 2011.

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

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

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

Dimitri P Bertsekas and John N Tsitsiklis. *Neuro-dynamic programming*, volume 5. Athena Scientific Belmont, MA, 1996.

Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer, 2009.

Zaiwei Chen, Sheng Zhang, Thinh T Doan, Siva Theja Maguluri, and John-Paul Clarke. Performance of Q-learning with linear function approximation: Stability and finite-time analysis. *arXiv preprint arXiv:1905.11425*, 2019.

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

Thinh T Doan. Finite-time analysis and restarting scheme for linear two-time-scale stochastic approximation. *arXiv preprint arXiv:1912.10583*, 2019.

Kefan Dong, Yuanhao Wang, Xiaoyu Chen, and Liwei Wang. Q-learning with UCB exploration is sample efficient for infinite-horizon MDP. *arXiv preprint arXiv:1901.09311*, 2019.

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

Harsh Gupta, R Srikant, and Lei Ying. Finite-time performance bounds and adaptive learning rate selection for two time-scale reinforcement learning. In *Advances in Neural Information Processing Systems*, pages 4706–4715, 2019.

Hado V Hasselt. Double Q-learning. In *Advances in neural information processing systems*, pages 2613–2621, 2010.

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

Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I Jordan. Is Q-learning provably efficient? In *Advances in Neural Information Processing Systems*, pages 4863–4873, 2018.

Michael J Kearns and Satinder P Singh. Finite-sample convergence rates for Q-learning and indirect algorithms. In *Advances in neural information processing systems*, pages 996–1002, 1999.

Vijay R Konda and John N Tsitsiklis. Actor-critic algorithms. In *Advances in neural information processing systems*, pages 1008–1014, 2000.

Vijay R Konda and John N Tsitsiklis. Linear stochastic approximation driven by slowly varying markov chains. *Systems & control letters*, 50(2):95–102, 2003.

Donghwan Lee and Niao He. A unified switching system perspective and ODE analysis of Q-learning algorithms. *arXiv preprint arXiv:1912.02270*, 2019.

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

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*, pages 664–671, 2008.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407, 1951.

Devavrat Shah and Qiaomin Xie. Q-learning with nearest neighbors. In *Advances in Neural Information Processing Systems*, pages 3111–3121, 2018.

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

Aaron Sidford, Mengdi Wang, Xian Wu, and Yinyu Ye. Variance reduced value iteration and faster algorithms for solving markov decision processes. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*, pages 770–787. SIAM, 2018b.

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

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

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

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

John N Tsitsiklis and Benjamin Van Roy. Analysis of temporal-diffference learning with function approximation. In *Advances in neural information processing systems*, pages 1075–1081, 1997.

Martin J Wainwright. Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-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 JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8(3-4):279–292, 1992.

Chen-Yu Wei, Mehdi Jafarnia-Jahromi, Haipeng Luo, Hiteshi Sharma, and Rahul Jain. Model-free reinforcement learning in infinite-horizon average-reward markov decision processes. *arXiv preprint arXiv:1910.07072*, 2019.

Tengyu Xu, Shaofeng Zou, and Yingbin Liang. Two time-scale off-policy TD learning: Non-asymptotic analysis over markovian samples. In *Advances in Neural Information Processing Systems*, pages 10633–10643, 2019.

Appendix A. Proofs of Auxiliary Propositions in Section 2 and Section 3

A.1. Proof of Proposition 2

Let $x \in \mathbb{R}^{\mathcal{N}}$ be any vector s.t. $\|x\|_v = 1$. Then,

$$\|Ax\|_v = \sup_{i\in\mathcal{N}} \frac{1}{v_i}\Big|\sum_{j\in\mathcal{N}} a_{ij}x_j\Big| \le \sup_{i\in\mathcal{N}} \sum_{j\in\mathcal{N}} |a_{ij}|\frac{v_j}{v_i}\frac{|x_j|}{v_j} \le \sup_{i\in\mathcal{N}} \sum_{j\in\mathcal{N}} |a_{ij}|\frac{v_j}{v_i}.$$

As a result, $\|A\|_v \le \sup_{i\in\mathcal{N}} \sum_{j\in\mathcal{N}} |a_{ij}|\frac{v_j}{v_i}$. On the other hand, let $i^* = \arg\max_{i\in\mathcal{N}} \sum_{j\in\mathcal{N}} |a_{ij}|\frac{v_j}{v_i}$ (ties broken arbitrarily). And we set $x = [x_1, \ldots, x_n]^\top$ with $x_j = v_j \mathrm{sign}(a_{i^*j})$, where $\mathrm{sign}(z) = 1$ when $z \ge 0$, and -1 otherwise. Then, clearly $\|x\|_v = 1$, and

$$\|Ax\|_v \ge \frac{1}{v_{i^*}}\Big|\sum_{j\in\mathcal{N}} a_{i^*j}x_j\Big| = \frac{1}{v_{i^*}}\Big|\sum_{j\in\mathcal{N}} a_{i^*j}\mathrm{sign}(a_{i^*j})v_j\Big| = \sum_{j\in\mathcal{N}} |a_{i^*j}|\frac{v_j}{v_i^*} = \sup_{i\in\mathcal{N}} \sum_{j\in\mathcal{N}} |a_{ij}|\frac{v_j}{v_i}.$$

This shows $\|A\|_v \ge \sup_{i\in\mathcal{N}} \sum_{j\in\mathcal{N}} |a_{ij}|\frac{v_j}{v_i}$ and finishes the proof. ■

A.2. Proof of Proposition 3

Let d be the distribution of i_t conditioned on $\mathcal{F}_{t-\tau}$. Then, by Levin and Peres (2017, eq. (4.33)),

$$\mathrm{TV}(d, \mu) \le 2^{-\lceil \log_2(\frac{2}{\mu_{\min}})\rceil} \le \frac{\mu_{\min}}{2},$$

where TV means the total-variation distance. As a result, for each $i \in \mathcal{N}$, $d_i \ge \mu_i - |\mu_i - d_i| \ge \mu_{\min} - \mathrm{TV}(d,\mu) \ge \frac{1}{2}\mu_{\min}$. This shows that for any i, $\mathbb{P}(i_t = i|\mathcal{F}_{t-\tau}) \ge \frac{1}{2}\mu_{\min}$ which verifies Assumption 3. ■

A.3. Proof of Proposition 5

Note that by Assumption 1(a), we have,

$$\|F(x)\|_v \leq \|F(x) - F(x^*)\|_v + \|F(x^*)\|_v \leq \gamma\|x - x^*\|_v + \|x^*\|_v \leq \gamma\|x\|_v + (1+\gamma)\|x^*\|_v.$$

In other words, Assumption 1(b) holds with $C = (1+\gamma)\|x^*\|_v$. Let $\bar{x} = \frac{1}{1-\gamma}((1+\gamma)\|x^*\|_v + \frac{\bar{w}}{\underline{v}})$. We prove $\|x(t)\|_v \leq \bar{x}$ by induction. The statement is obviously true for $t = 0$ as $x(0)$ is initialized to be the all-zero vector. Suppose it is true for t, then

$$\begin{aligned}
\|x(t+1)\|_v &\leq \max(\frac{1}{v_{i_t}}|x_{i_t}(t+1)|, \|x(t)\|_v) \\
&\leq \max(\frac{1}{v_{i_t}}|x_{i_t}(t+1)|, \bar{x}).
\end{aligned}$$

Then, notice that,

$$\begin{aligned}
\frac{1}{v_{i_t}}|x_{i_t}(t+1)| &\leq (1-\alpha_t)\frac{1}{v_{i_t}}|x_{i_t}(t)| + \alpha_t(\frac{1}{v_{i_t}}|F_{i_t}(x(t))| + \frac{1}{v_{i_t}}|w(t)|) \\
&\leq (1-\alpha_t)\|x(t)\|_v + \alpha_t(\|F(x(t))\|_v + \frac{1}{\underline{v}}\bar{w}) \\
&\leq (1-\alpha_t)\|x(t)\|_v + \alpha_t(\gamma\|x(t)\|_v + C + \frac{1}{\underline{v}}\bar{w}) \\
&\leq (1-\alpha_t)\bar{x} + \alpha_t(\gamma\bar{x} + C + \bar{w}) \\
&= \bar{x},
\end{aligned}$$

where in the second inequality, we have used $|w(t)| \leq \bar{w}$ almost surely (cf. Assumption 2), and in the last equality, we have used that $\gamma\bar{x} + C + \frac{\bar{w}}{\underline{v}} = \bar{x}$. This finishes the induction. ■

A.4. Proof of Proposition 6

We prove $\|Q(t)\|_\infty \leq \frac{\bar{r}}{1-\gamma}$ by induction. Firstly, the statement is true for $t = 0$ as $Q(0)$ is initialized to be the all zero table. Then, assume the statement is true for t. For $t+1$, clearly $\|Q(t+1)\|_\infty \leq \max(\|Q(t)\|_\infty, |Q_{s_t,a_t}(t+1)|)$. Further, notice,

$$\begin{aligned}
|Q_{s_t,a_t}(t+1)| &\leq (1-\alpha_t)|Q_{s_t,a_t}(t)| + \alpha_t(|r_t| + \gamma|\max_a Q_{s_{t+1},a}(t)|) \\
&\leq (1-\alpha_t)\|Q(t)\|_\infty + \alpha_t(\bar{r} + \gamma\|Q(t)\|_\infty) \\
&\leq (1-\alpha_t)\frac{\bar{r}}{1-\gamma} + \alpha_t(\bar{r} + \gamma\frac{\bar{r}}{1-\gamma}) \\
&= \frac{\bar{r}}{1-\gamma}.
\end{aligned}$$

This finishes the induction, and hence $\|Q(t)\|_\infty \leq \frac{\bar{r}}{1-\gamma}$ almost surely for all $t \geq 0$. As Q^* is the Q-function under an optimal policy π^*, we get for any $s \in \mathcal{S}, a \in \mathcal{A}$,

$$|Q^*_{s,a}| = \mathbb{E}_{\pi^*}[\sum_{t=0}^{\infty}\gamma^t r_t|(s_0,a_0) = (s,a)] \leq \sum_{t=0}^{\infty}\gamma^t\bar{r} = \frac{\bar{r}}{1-\gamma},$$

which concludes the proof of part (a). For part (b), notice,

$$\begin{aligned}|w(t)| &\le |r_t| + \gamma|\max_a Q_{s_{t+1},a}(t)| + |F_{s_t,a_t}(Q(t))| \\ &\le \bar{r} + \gamma\|Q(t)\|_\infty + \|F(Q(t))\|_\infty \\ &\le 2(\bar{r} + \gamma\|Q(t)\|_\infty) \\ &\le 2(\bar{r} + \gamma\frac{\bar{r}}{1-\gamma}) = \frac{2\bar{r}}{1-\gamma},\end{aligned}$$

which finishes the proof of part (b). ■

Appendix B. Proofs of Auxiliary Lemmas in Section 4

B.1. Proof of Lemma 8 (Error Decomposition)

By (9), we have,

$$\begin{aligned}&\|x(t+1) - x^*\|_v \\ &\le \sup_i \frac{1}{v_i}\Big|\tilde{b}_{\tau-1,t,i}x_i(\tau) + \sum_{k=\tau}^{t} b_{k,t,i}F_i(x(k)) - x_i^*\Big| + \|\sum_{k=\tau}^{t}\alpha_k\tilde{B}_{k,t}\epsilon(k)\|_v + \|\sum_{k=\tau}^{t}\alpha_k\tilde{B}_{k,t}\phi(k)\|_v.\end{aligned} \tag{14}$$

It easy to check that for each i, $\tilde{b}_{\tau-1,t,i} + \sum_{k=\tau}^{t} b_{k,t,i} = 1$. Then, for each i, we have

$$\begin{aligned}\frac{1}{v_i}\Big|\tilde{b}_{\tau-1,t,i}x_i(\tau) + \sum_{k=\tau}^{t} b_{k,t,i}F_i(x(k)) - x_i^*\Big| &\le \tilde{b}_{\tau-1,t,i}\frac{1}{v_i}|x_i(\tau) - x_i^*| + \sum_{k=\tau}^{t} b_{k,t,i}\frac{1}{v_i}\big|F_i(x(k)) - x_i^*\big| \\ &\le \tilde{b}_{\tau-1,t,i}\|x(\tau) - x^*\|_v + \sum_{k=\tau}^{t} b_{k,t,i}\|F(x(k)) - x^*\|_v \\ &\le \tilde{\beta}_{\tau-1,t}\|x(\tau) - x^*\|_v + \gamma\sum_{k=\tau}^{t} b_{k,t,i}\|x(k) - x^*\|_v,\end{aligned}$$

where in the last inequality, we have used that F is γ-contraction in $\|\cdot\|_v$ with fixed point x^*. Combining the above with (14), we have,

$$\begin{aligned}a_{t+1} &= \|x(t+1) - x^*\|_v \\ &\le \tilde{\beta}_{\tau-1,t}a_\tau + \gamma\sup_{i\in\mathcal{N}}\sum_{k=\tau}^{t} b_{k,t,i}a_k + \Big\|\sum_{k=\tau}^{t}\alpha_k\tilde{B}_{k,t}\epsilon(k)\Big\|_v + \Big\|\sum_{k=\tau}^{t}\alpha_k\tilde{B}_{k,t}\phi(k)\Big\|_v.\end{aligned}$$

■

B.2. Proof of Lemma 9 (Bounds on $\|\epsilon(t)\|_v$ and $\|\phi(t)\|_v$)

For part (a), we have,

$$\|\epsilon(t)\|_v = \|(\mathbf{e}_{i_t}\mathbf{e}_{i_t}^\top - D_t)[F(x(t-\tau)) - x(t-\tau)] + w(t)\mathbf{e}_{i_t}\|_v$$

$$
\begin{aligned}
&\leq \|\mathbf{e}_{i_t}\mathbf{e}_{i_t}^\top - D_t\|_v \|F(x(t-\tau)) - x(t-\tau))\|_v + |w(t)| \|\mathbf{e}_{i_t}\|_v \\
&\leq \|F(x(t-\tau))\|_v + \|x(t-\tau)\|_v + \frac{\bar{w}}{\underline{v}} \\
&\leq 2\bar{x} + C + \frac{\bar{w}}{\underline{v}} := \bar{\epsilon}.
\end{aligned}
$$

where we have used by Proposition 2, $\|\mathbf{e}_{i_t}\mathbf{e}_{i_t}^\top - D_t\|_v = \sup_i |\mathbf{1}(i_t = i) - d_{t,i}| \leq 1$ (here $\mathbf{1}$ is the indicator function); and $\|F(x(t-\tau))\|_v \leq \gamma \|x(t-\tau)\|_v + C \leq \bar{x} + C$.

For part (b), we have,

$$
\begin{aligned}
\|\phi(t)\|_v &= \|(\mathbf{e}_{i_t}\mathbf{e}_{i_t}^\top - D_t)(F(x(t)) - F(x(t-\tau))) - (\mathbf{e}_{i_t}\mathbf{e}_{i_t}^\top - D_t)(x(t) - x(t-\tau))\|_v \\
&\leq \|F(x(t)) - F(x(t-\tau))\|_v + \|x(t) - x(t-\tau)\|_v \\
&\leq 2\|x(t) - x(t-\tau)\|_v.
\end{aligned}
$$

Notice that $\|x(t) - x(t-1)\|_v \leq \alpha_{t-1}(\|F(x(t-1))\|_v + \|x(t-1)\|_v + \frac{1}{\underline{v}}\bar{w}) \leq \alpha_{t-1}(2\bar{x} + C + \frac{1}{\underline{v}}\bar{w}) = \alpha_{t-1}\bar{\epsilon}$. Summing up, we get

$$
\|\phi(t)\|_v \leq \sum_{k=t-\tau+1}^{t} 2\bar{\epsilon}\alpha_{k-1}
$$

■

B.3. Proof of Lemma 10 (Step Sizes)

For part (a), notice that $\log(1-x) \leq -x$ for all $x < 1$. Then,

$$
(1 - \sigma\alpha_t) = e^{\log(1 - \frac{\sigma h}{t+t_0})} \leq e^{-\frac{\sigma h}{t+t_0}}.
$$

Therefore,

$$
\begin{aligned}
\prod_{\ell=k+1}^{t} (1 - \sigma\alpha_\ell) &\leq e^{-\sum_{\ell=k+1}^{t} \frac{\sigma h}{\ell+t_0}} \\
&\leq e^{-\int_{k+1}^{t+1} \frac{\sigma h}{y+t_0} dy} \\
&= e^{-\sigma h \log(\frac{t+1+t_0}{k+1+t_0})} \\
&= \Big(\frac{k+1+t_0}{t+1+t_0}\Big)^{\sigma h},
\end{aligned}
$$

which leads to the bound on $\beta_{k,t}$ and $\tilde{\beta}_{k,t}$.

For part (b),

$$
\beta_{k,t}^2 \leq \frac{h^2}{(t+1+t_0)^{2\sigma h}} \frac{(k+1+t_0)^{2\sigma h}}{(k+t_0)^2} \leq \frac{2h^2}{(t+1+t_0)^{2\sigma h}} (k+t_0)^{2\sigma h - 2},
$$

where we have used $(k+1+t_0)^{2\sigma h} \leq 2(k+t_0)^{2\sigma h}$, which is true when $t_0 \geq 4h$. Then,

$$
\sum_{k=1}^{t} \beta_{k,t}^2 \leq \frac{2h^2}{(t+1+t_0)^{2\sigma h}} \sum_{k=1}^{t} (k+t_0)^{2\sigma h - 2} \leq \frac{2h^2}{(t+1+t_0)^{2\sigma h}} \int_1^{t+1} (y+t_0)^{2\sigma h - 2} dy
$$

$$< \frac{2h^2}{(t+1+t_0)^{2\sigma h}} \frac{1}{2\sigma h - 1}(t+1+t_0)^{2\sigma h-1} < \frac{2h}{\sigma}\frac{1}{t+1+t_0},$$

where in the last inequality we have used $2\sigma h - 1 > \sigma h$.

For part (c), notice that for $k-\tau \le \ell \le k-1$ where $k \ge \tau$, we have $\alpha_\ell \le \frac{h}{k-\tau+t_0} \le \frac{2h}{k+t_0}$ (using $t_0 \ge \tau$). Then,

$$\begin{aligned}
\sum_{k=\tau}^{t} \beta_{k,t} \sum_{\ell=k-\tau}^{k-1} \alpha_\ell &\le \sum_{k=\tau}^{t} \beta_{k,t} \frac{2h\tau}{k+t_0} \le \sum_{k=\tau}^{t} \frac{h}{k+t_0}\Big(\frac{k+1+t_0}{t+1+t_0}\Big)^{\sigma h} \frac{2h\tau}{k+t_0} \\
&\le \sum_{k=\tau}^{t} \frac{4h^2\tau}{(t+1+t_0)^{\sigma h}}(k+t_0)^{\sigma h-2} \\
&\le \frac{4h^2\tau}{(t+1+t_0)^{\sigma h}} \frac{(t+1+t_0)^{\sigma h-1}}{\sigma h-1} \\
&\le \frac{8h\tau}{\sigma}\frac{1}{t+1+t_0},
\end{aligned}$$

where we have used $(k+1+t_0)^{\sigma h} \le 2(k+t_0)^{\sigma h}$, and $\sigma h - 1 > \frac{1}{2}\sigma h$. ■

B.4. Proof of Lemma 13 (Azuma Hoeffding)

Let ℓ be an integer between 0 and $\tau-1$. For each ℓ, define process $Y_k^\ell = X_{\tau k+\ell}$, scalar $\bar{Y}_k^\ell = \bar{X}_{k\tau+\ell}$, and define Filtration $\tilde{\mathcal{F}}_k^\ell = \mathcal{F}_{\tau k+\ell}$. Then, Y_k^ℓ is $\tilde{\mathcal{F}}_k^\ell$-adapted, and satisfies

$$\mathbb{E} Y_k^\ell | \tilde{\mathcal{F}}_{k-1}^\ell = \mathbb{E} X_{k\tau+\ell} | \mathcal{F}_{k\tau+\ell-\tau} = 0.$$

Therefore, applying Azuma-Hoeffding bound on Y_k^ℓ, we have

$$P(|\sum_{k:k\tau+\ell\le t} Y_k^\ell| \ge t) \le 2\exp(-\frac{t^2}{2\sum_{k:k\tau+\ell\le t}(\bar{Y}_k^\ell)^2}),$$

i.e. with probability at least $1-\frac{\delta}{\tau}$,

$$|\sum_{k:k\tau+\ell\le t} X_{k\tau+\ell}| = |\sum_{k:k\tau+\ell\le t} Y_k^\ell| \le \sqrt{2\sum_{k:k\tau+\ell\le t} \bar{X}_{k\tau+\ell}^2 \log(\frac{2\tau}{\delta})}.$$

Using the union bound for $\ell = 0,\ldots,\tau-1$, we get that with probability at least $1-\delta$,

$$|\sum_{k=0}^{t} X_t| \le \sum_{\ell=0}^{\tau-1} |\sum_{k:k\tau+\ell\le t} X_{k\tau+\ell}| \le \sum_{\ell=0}^{\tau-1}\sqrt{2\sum_{k:k\tau+\ell\le t} \bar{X}_{k\tau+\ell}^2 \log(\frac{2\tau}{\delta})} \le \sqrt{2\tau \sum_{k=0}^{t} \bar{X}_k^2 \log(\frac{2\tau}{\delta})},$$

where the last inequality is due to Cauchy-Schwarz. ■

B.5. Proof of Lemma 14

Let p_k be a scalar sequence defined as follows. Set $p_\tau = 0$, and

$$p_k = (1 - \alpha_{k-1} d_{k-1,i}) p_{k-1} + \alpha_{k-1} \epsilon_i(k-1).$$

Then $p_{t+1} = \sum_{k=\tau}^{t} \alpha_k \epsilon_i(k) \prod_{\ell=k+1}^{t} (1 - \alpha_\ell d_{\ell,i})$, and to prove Lemma 14 we need to bound $|p_{t+1}|$. Let

$$k_0 = \sup\{k \le t : (1 - \alpha_k d_{k,i}) |p_k| \le \alpha_k |\epsilon_i(k)|\}.$$

We must have $k_0 \ge \tau$ since $|p_\tau| = 0$. With k_0 defined, we now define another scalar sequence $\tilde{p}$ s.t. $\tilde{p}_{k_0+1} = p_{k_0+1}$ and

$$\tilde{p}_k = (1 - \alpha_{k-1}\sigma) \tilde{p}_{k-1} + \alpha_{k-1} \epsilon_i(k-1).$$

We claim that for all $k \ge k_0 + 1$, p_k and $\tilde{p}_k$ have the same sign, and $|p_k| \le |\tilde{p}_k|$. This is obviously true for $k = k_0 + 1$. Suppose it is true for for $k - 1$. Without loss of generality, suppose both p_{k-1} and $\tilde{p}_{k-1}$ are non-negative. Since $k - 1 > k_0$ and by the definition of k_0, we must have

$$(1 - \alpha_{k-1} d_{k-1,i}) p_{k-1} > |\alpha_{k-1} \epsilon_i(k-1)|.$$

Therefore, $p_k > 0$. Further, since $d_{k-1,i} \ge \sigma$, we also have

$$(1 - \alpha_{k-1}\sigma) \tilde{p}_{k-1} \ge (1 - \alpha_{k-1} d_{k-1,i}) p_{k-1} > |\alpha_{k-1} \epsilon_i(k-1)|.$$

These imply $\tilde{p}_k \ge p_k > 0$. The case where both p_{k-1} and $\tilde{p}_{k-1}$ are negative is similar. This finishes the induction, and as a result, $|p_{t+1}| \le |\tilde{p}_{t+1}|$. Notice,

$$\tilde{p}_{t+1} = \sum_{k=k_0+1}^{t} \alpha_k \epsilon_i(k) \prod_{\ell=k+1}^{t} (1 - \alpha_\ell \sigma) + \tilde{p}_{k_0+1} \prod_{\ell=k_0+1}^{t} (1 - \alpha_\ell \sigma) = \sum_{k=k_0+1}^{t} \epsilon_i(k) \beta_{k,t} + \tilde{p}_{k_0+1} \tilde{\beta}_{k_0,t}.$$

By the definition of k_0, we have

$$|\tilde{p}_{k_0+1}| = |p_{k_0+1}| \le (1 - \alpha_{k_0} d_{k_0,i}) |p_{k_0}| + \alpha_{k_0} |\epsilon_i(k_0)| \le 2\alpha_{k_0} |\epsilon_i(k_0)| \le 2\alpha_{k_0} \bar{\epsilon} v_i,$$

where in the last step, we have used the upper bound on $\|\epsilon(k_0)\|_v$ in Lemma 9 (a). As a result,

$$\begin{aligned}
|p_{t+1}| \le |\tilde{p}_{t+1}| &\le \Big| \sum_{k=k_0+1}^{t} \epsilon_i(k) \beta_{k,t} \Big| + \big| \tilde{p}_{k_0+1} \tilde{\beta}_{k_0,t} \big| \\
&\le \Big| \sum_{k=k_0+1}^{t} \epsilon_i(k) \beta_{k,t} \Big| + \big| 2\alpha_{k_0} \bar{\epsilon} v_i \tilde{\beta}_{k_0,t} \big| \\
&= \Big| \sum_{k=k_0+1}^{t} \epsilon_i(k) \beta_{k,t} \Big| + 2\bar{\epsilon} v_i \beta_{k_0,t}.
\end{aligned}$$

■

B.6. Proof of Lemma 15

Throughout the proof, we fix i and will frequently use the property $d_{k,i} \geq \sigma$ which holds almost surely. Define the sequence

$$e_t = \sum_{k=\tau}^{t} b_{k,t,i} \frac{1}{(k+t_0)^\omega}.$$

We use induction to show that $e_t \leq \frac{1}{\sqrt{\gamma}(t+1+t_0)^\omega}$. The statement is clearly true for $t = \tau$, as $e_\tau = b_{\tau,\tau,i}\frac{1}{(\tau+t_0)^\omega} = \alpha_\tau d_{\tau,i}\frac{1}{(\tau+t_0)^\omega} \leq \frac{1}{\sqrt{\gamma}(\tau+1+t_0)^\omega}$ (the last step needs $\alpha_\tau \leq \frac{1}{2}$, $(1+\frac{1}{t_0})^\omega \leq \frac{2}{\sqrt{\gamma}}$, implied by $t_0 \geq 1$, $\omega \leq 1$). Let the statement be true for $t-1$. Then, notice that,

$$\begin{aligned}
e_t &= \sum_{k=\tau}^{t-1} b_{k,t,i}\frac{1}{(k+t_0)^\omega} + b_{t,t,i}\frac{1}{(t+t_0)^\omega} \\
&= (1-\alpha_t d_{t,i})\sum_{k=\tau}^{t-1} b_{k,t-1,i}\frac{1}{(k+t_0)^\omega} + \alpha_t d_{t,i}\frac{1}{(t+t_0)^\omega} \\
&= (1-\alpha_t d_{t,i})e_{t-1} + \alpha_t d_{t,i}\frac{1}{(t+t_0)^\omega} \\
&\leq (1-\alpha_t d_{t,i})\frac{1}{\sqrt{\gamma}(t+t_0)^\omega} + \alpha_t d_{t,i}\frac{1}{(t+t_0)^\omega} \\
&= \left[1-\alpha_t d_{t,i}(1-\sqrt{\gamma})\right]\frac{1}{\sqrt{\gamma}(t+t_0)^\omega},
\end{aligned}$$

where the inequality is based on induction assumption. Then, plug in $\alpha_t = \frac{h}{t+t_0}$ and use $d_{t,i} \geq \sigma$, we have,

$$\begin{aligned}
e_t &\leq \left[1-\frac{\sigma h}{t+t_0}(1-\sqrt{\gamma})\right]\frac{1}{\sqrt{\gamma}(t+t_0)^\omega} \\
&= \left[1-\frac{\sigma h}{t+t_0}(1-\sqrt{\gamma})\right]\left(\frac{t+1+t_0}{t+t_0}\right)^\omega\frac{1}{\sqrt{\gamma}(t+1+t_0)^\omega} \\
&= \left[1-\frac{\sigma h}{t+t_0}(1-\sqrt{\gamma})\right]\left(1+\frac{1}{t+t_0}\right)^\omega\frac{1}{\sqrt{\gamma}(t+1+t_0)^\omega}.
\end{aligned}$$

Now using the inequality that for any $x > -1$, $(1+x) \leq e^x$, we have,

$$\left[1-\frac{\sigma h}{t+t_0}(1-\sqrt{\gamma})\right]\left(1+\frac{1}{t+t_0}\right)^\omega \leq e^{-\frac{\sigma h}{t+t_0}(1-\sqrt{\gamma})+\omega\frac{1}{t+t_0}} \leq 1,$$

where in the last inequality, we have used $\omega \leq 1$ and the condition on h s.t. $\sigma h(1-\sqrt{\gamma}) \geq 1$. This shows $e_t \leq \frac{1}{\sqrt{\gamma}(t+1+t_0)^\omega}$ and finishes the induction. ∎