Central Limit Theorems for Asynchronous Averaged Q-Learning

Xingtu Liu XINGTU_LIU@SFU.CA
School of Computing Science, Simon Fraser University, Burnaby, Canada

Editors: G. Sukhatme, L. Lindemann, S. Tu, A. Wierman, N. Atanasov

Abstract

This paper establishes central limit theorems for Polyak–Ruppert averaged Q-learning under asynchronous updates. We present a non-asymptotic central limit theorem, where the convergence rate in Wasserstein distance explicitly reflects the dependence on the number of iterations, state–action space size, the discount factor, and the quality of exploration. In addition, we derive a functional central limit theorem, showing that the partial-sum process converges weakly to a Brownian motion.

Keywords: Reinforcement Learning, Central Limit Theorem, Stochastic Approximation, Q-Learning

1. Introduction

Reinforcement Learning (RL) has emerged as a powerful paradigm in artificial intelligence, achieving successes in various applications such as Atari [23], Go [33], robot manipulation [37, 41], and aligning large language models to human preferences [25, 32]. Q-learning [39], which directly learns the optimal action-value function (Q-function) from experience trajectories, is one of the most widely used RL algorithms.

Stochastic approximation (SA) [2, 5] is a general iterative framework to solve fixed-point equation problems. Since the Bellman operator in RL is a contraction map with a unique fixed point, many RL algorithms can be interpreted as instances of SA. For example, TD learning [36] can be viewed as an instance of linear SA. Synchronous Q-learning, by contrast, is a special case of nonlinear SA with martingale noise. The asynchronous Q-learning algorithm studied in this work, however, is a nonlinear SA problem with Markovian noise. There is a growing line of work on finite-sample analysis of SA with applications to RL algorithms [3, 6, 7, 9, 11, 18, 29, 35, 38].

Polyak-Ruppert averaging is a classical variance-reduction technique to stabilize and accelerate SA algorithms. A key motivation for focusing on Polyak-Ruppert averaging is its dual advantage of practical robustness and statistical efficiency. Standard stochastic approximation algorithms are often sensitive to the specific choice of decaying stepsize, requiring hyperparameter tuning to achieve stable convergence. By contrast, averaging the iterates makes the algorithm significantly more robust to the underlying stepsize schedule. Furthermore, Polyak-Ruppert averaging is known to achieve the optimal rate of convergence by minimizing the asymptotic covariance matrix of the estimates. In this paper, we are interested in establishing central limit theorems (CLTs) for Polyak-Ruppert averaged Q-learning under asynchronous updates. Building CLTs provides a foundational understanding of the algorithm's statistical properties. This asymptotic normality is crucial for uncertainty quantification and statistical inference in RL. Building on the seminal work by Polyak and Juditsky [27], a non-asymptotic CLT for Polyak-Ruppert averaged SGD was established [1]. Mou et al. [24], Samsonov et al. [31] derive non-asymptotic CLTs for linear SA with Polyak–Ruppert av-

eraged iterates. Similar results for two-time-scale SA are also studied [16, 17, 19]. Recently, CLTs for SA with applications to RL algorithms are studied [4]. As a special case linear SA, Samsonov et al. [30], Srikant [34] derive non-asymptotic CLTs for TD-learning with averaging. However, non-asymptotic CLTs for Q-learning remain unexplored.

As a special case of nonlinear SA, Q-learning is substantially more challenging to analyze than linear SA and TD learning. Functional CLTs for Polyak–Ruppert averaged synchronous Q-learning was established in Li et al. [21], Panda et al. [26], Xie and Zhang [40]. Synchronous Q-learning only considers martingale noises. By contrast, asynchronous Q-learning updates a single state–action pair based on one transition sample at each iteration, which involves Markovian noises that are non-IID. Moreover, the empirical Bellman operator in asynchronous Q-learning is non-smooth. Thus, the challenges in analyzing asynchronous Q-learning come from nonlinearity, Markovian samples, and a non-smooth operator. Recently, Zhang and Xie [43] established a functional CLT for asynchronous Q-learning with a constant stepsize. Constant stepsize does not satisfy the necessary conditions for establishing a non-asymptotic CLT, which we detail in Section 3. To the best of our knowledge, no non-asymptotic CLT is currently known for Q-learning, even in the synchronous setting. In this work, we close this gap and prove both a non-asymptotic CLT and a functional CLT for asynchronous averaged Q-learning with decaying stepsizes.

2. Preliminaries

An infinite-horizon discounted Markov decision process (MDP) is denoted by $\mathcal{M}$, and is defined by the tuple $\langle \mathcal{S}, \mathcal{A}, P, r, \gamma \rangle$ where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the action set, $P : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}$ is the transition probability function, and $\gamma \in [0, 1)$ is the discount factor. Let $\Delta_{\mathcal{A}}$ denotes the simplex over the action space. The action-value function (Q-function) of a stationary and stochastic policy $\pi : \mathcal{S} \to \Delta_{\mathcal{A}}$ is defined as $Q^\pi(s, a) = \mathbb{E}\Big[\sum_{t=0}^{\infty} \gamma^t r(s_t, a_t) | s_0 = s, a_0 = a\Big]$, where $a_t \sim \pi(\cdot|s_t)$ and $s_{t+1} \sim P(\cdot|s_t, a_t)$. The optimal Q-function is defined as $Q^* := \max_\pi Q^\pi$. The value function is defined as $V^\pi = \pi Q^\pi$, where $(\pi Q)(s) := \langle \pi(\cdot|s), Q(s, \cdot) \rangle$. We also define $P^\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}$ such that $P^\pi Q = P(\pi Q)$. We make the following assumption over a specific optimal policy.

Assumption 1 *There exists an optimal policy π^* such that for $Q \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$ we have $\|(P^\pi - P^{\pi^*})(Q - Q^*)\|_\infty \le L\|Q - Q^*\|_\infty^2$ where $\pi(s) := \operatorname{argmax}_{a \in \mathcal{A}} Q(s, a)$.*

Adopted from Li et al. [21], this assumption provides a localized smoothness condition required to establish asymptotic normality for Polyak-Ruppert averaging. Unlike linear stochastic approximation, the Bellman optimality operator is non-smooth due to the $\max$ operation, meaning the greedy policy π can shift abruptly near Q^*. Assumption 1 addresses this by imposing a margin condition, where the difference in transition dynamics between the current and optimal policies must shrink at a quadratic rate relative to the estimation error, $||Q - Q^*||_\infty$. This quadratic decay ensures that non-linear residual terms vanish asymptotically, allowing the leading-order martingale noise to dictate the covariance of the averaged iterates.

The asynchronous Q-learning algorithm maintains a Q-function estimator Q_k and the update rule is the following:

$$Q_{k+1} = Q_k + \alpha_k(F_k - Q_k) \tag{1}$$

where we let $F_k = F(Q_k, y_k)$, $y_k = (s_k, a_k, s_{k+1})$,

$$[F(Q_k, s_k, a_k, s_{k+1})](s, a) = \mathbb{1}_{\{(s_k, a_k) = (s, a)\}} \Gamma(Q_k, s_k, a_k, s_{k+1}) + Q_k(s, a), \tag{2}$$

and

$$\Gamma(Q_k, s_k, a_k, s_{k+1}) = r_k(s_k, a_k) + \gamma \max_a Q_k(s_{k+1}, a) - Q_k(s_k, a_k).$$

Γ is the temporal difference in the Q-function iterate. The sample trajectory $\{(s_k, a_k)\}$ is collected by the MDP under a behavior policy π_b. We define $V_k(s) := \max_a Q_k(s, a)$. Now we make the following assumption on the Markov chain, which is standard in the literature [10, 20, 29, 42, 43].

Assumption 2 *$\{y_k\}_{k\geq 0}$ is an irreducible and aperiodic finite state Markov chain $\mathcal{M}$.*

Under Assumption 2, the Markov chain $\mathcal{M}$ admits a unique stationary distribution $\tilde{\mu}$. We denote $\tilde{S}$ as the state-space and $\tilde{P}$ as the transition kernel. Next, we define the Bellman operator for the Q-function:

$$[\mathcal{T}(Q)](s, a) = r(s, a) + \gamma \mathbb{E}_{s' \sim P(\cdot|s,a)} \max_{a' \in \mathcal{A}} Q(s', a').$$

Define π_k such that $\mathcal{T}(Q_k) = r + \gamma \pi_k Q_k$. Denote by $\bar{F}_k$ the expected value of $F(Q_k, y_k)$, i.e. $\bar{F}_k := \bar{F}(Q_k) := \mathbb{E}_{y_k \sim \tilde{\mu}}[F(Q_k, y_k)]$. Further, denote by $D \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}$ the diagonal matrix with $\{p(s, a)\}_{(s,a)\in\mathcal{S}\times\mathcal{A}}$ on its diagonal, where $p(s, a)$ is the stationary visitation probability of the state-action pair (s, a). We denote $\rho := \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} p(s, a)$, which captures the quality of exploration. The following lemmas are consequences of Assumption 2.

Lemma 3 (Proposition 3.1 in [10]) *Suppose that Assumption 2 holds, we have*

$$\bar{F}(Q) = D\mathcal{T}(Q) + (I - D)Q.$$

Lemma 4 (Proposition 3.1 in [10]) *For the operator $F(\cdot, \cdot)$ defined in eq. (2), under Assumption 2, we have $\|F(Q_1, s) - F(Q_2, s)\|_\infty \leq 2\|Q_1 - Q_2\|_\infty$ for any $s \in \tilde{S}$.*

We denote the Markov chain mixing time at iteration k as t_k. Formally, the mixing time t_k of the Markov chain $\mathcal{M}$ is defined as $t_k := \min\{i \geq 0 : \max_{s\in\tilde{S}} \|\tilde{P}^i(s, \cdot) - \tilde{\mu}(\cdot)\|_{\text{TV}} \leq \alpha_k\}$.

3. Main Results

In this section, we present our main results. We first establish a non-asymptotic CLT for the averaged Q-learning iterates, providing an explicit rate at which their distribution approaches a normal distribution. The deviation is measured by using the 1-Wasserstein distance. We then derive a functional central limit theorem (FCLT), showing that the partial-sum process converges weakly to a Brownian motion.

3.1. Non-Asymptotic Central Limit Theorem

Let $\Delta_k = Q_k - Q^*$. Our goal is to study the rate at which $\frac{1}{\sqrt{K}} \sum_{k=1}^K \Delta_k$ converges in distribution to normality. We present the main result as follows, where we use big O notation to hide all constants.

Theorem 5 *Let $\alpha_k = \alpha(k+b)^{-\beta}$ for some constants $\alpha, b > 0$ and $\beta \in (0.5, 1)$. Under Assumption 1 and 2, we have the following rate of convergence*

$$\mathcal{W}_1\left(K^{-\frac{1}{2}}\sum_{k=1}^{K}\Delta_k, \tilde{\mathcal{N}}\right) \leq \frac{(|\mathcal{S}||\mathcal{A}|)^{\frac{1}{2}}}{\rho(1-\gamma)^2 K^{\frac{1}{2}}} \cdot \tilde{O}\left((\rho(1-\gamma))^{\frac{\beta-2}{1-\beta}} + K^{\beta/2}\rho^{-1}(1-\gamma)^{-1} + K^{1-\beta} + K^{\frac{1-\beta}{2}}\rho^{-1-\beta}(1-\gamma)^{-\beta}\right)$$

where $\tilde{\mathcal{N}} = (A^{-1}\Sigma A^{-\top})^{1/2}\mathcal{N}(0, I)$, $A = D - \gamma D P^{\pi^*}$, $\Sigma := \sum_{i,j\in\tilde{S}} \tilde{\mu}(i)\tilde{P}(i,j)(X(j) - \mathbb{E}[X(Y_1)|Y_0 = i])(X(j) - \mathbb{E}[X(Y_1)|Y_0 = i])^\top$ and X is the solution to a Poisson's equation.

We now derived a non-asymptotic CLT showing that the distribution of the algorithm's average error converges towards a normal distribution. The asymptotic covariance matrix Σ describes the variance in the learning process that comes from sampling transitions from the environment. The asynchronous Q-learning updates are noisy because they are based on single transition samples, which is not IID. The matrix Σ quantifies the long-term structure of this randomness. The parameter ρ quantifies the quality of exploration. Recall that $\tilde{u}, \tilde{P}$, and $\tilde{S}$ are the stationary distribution, state-space, and transition kernel of the Markov chain $\mathcal{M}$. We define $X : \tilde{S} \to \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ to be the solution of the Poisson's equation: $F(Q^*, i) - \mathbb{E}[F(Q^*, i)] = X(i) - \mathbb{E}[X(Y_1)|Y_0 = i] \ \forall \ i \in \tilde{S}$.

The stepsize in the Q-learning update chosen in this work is $\alpha_k = \alpha(k+b)^{-\beta}$ for two reasons. First, convergence of stochastic approximation with averaging schemes relies on several key conditions [21, 27]: (i) $0 \leq \sup_k \alpha_k \leq 1$, $\alpha_k \downarrow 0$ and $k\alpha_k \uparrow \infty$; (ii) $\frac{\alpha_{k-1}-\alpha_k}{\alpha_{k-1}} = o(\alpha_{k-1})$; (iii) $\frac{1}{\sqrt{K}}\sum_{k=0}^{K}\alpha_k \to 0$; (iv) $\frac{\sum_{k=0}^{K}\alpha_k}{K\alpha_K} \leq C$. Constant stepsizes violate conditions (i) and (iii), while linear stepsizes violate condition (ii). By contrast, polynomial stepsizes satisfy all of the above. Second, the problem-dependent constants α and b are crucial for establishing a finite-sample convergence guarantee [10], which we leverage in our analysis. The parameter α acts as a scaling factor for balancing the trade-off between the speed of convergence and the final error of the algorithm. The parameter b is used to control the magnitude of the initial stepsizes and ensure the stability of the algorithm during the early stages. Setting $\beta = 2/3$, the rate of convergence can be simplified as follows.

Corollary 6 *Under Assumption 1, 2, and with $\alpha_k = \alpha(k+b)^{-\frac{2}{3}}$, $K \geq (\rho(1-\gamma))^{-12}$, we have*

$$\mathcal{W}_1\left(K^{-\frac{1}{2}}\sum_{k=1}^{K}\Delta_k, (A^{-1}\Sigma A^{-\top})^{1/2}\mathcal{N}(0, I)\right) \leq \tilde{O}\left(\frac{(|\mathcal{S}||\mathcal{A}|)^{\frac{1}{2}}}{K^{\frac{1}{6}}\rho^2(1-\gamma)^3}\right).$$

3.2. Proof Sketch of Theorem 5

The proof of the non-asymptotic central limit theorem for asynchronous averaged Q-learning addresses the core challenges of nonlinearity, the non-smoothness of the empirical Bellman operator, and non-IID Markovian noises. The proof contains five main steps.

Step 1: Constructing Bounding Processes Note that the empirical Bellman operator in asynchronous Q-learning is non-smooth due to the $\max$ operator and the indicator function. The proof introduces upper and lower bounding processes, denoted as $\Delta_k^{\uparrow}$ and $\Delta_k^{\downarrow}$, to track the error $\Delta_k = Q_k - Q^*$. By carefully bounding the update steps using the properties of the greedy and optimal policies, it is established by induction that $\Delta_k^{\downarrow} \leq \Delta_k \leq \Delta_k^{\uparrow}$ holds for all $k \in [K]$.

Step 2: Recursive Error Decomposition We first analyze the upper bounding process, where the accumulated error sum $\sum_{k=1}^{K} \Delta_k^{\uparrow}$ is expanded recursively. We decompose the sum into five primary terms. These terms isolate distinct sources of error: the initialization bias, the smooth temporal difference errors, and the stochastic noise term.

Step 3: Handling Markovian Noise via Poisson Equation Asynchronous updates rely on single transition samples, which introduces Markovian noise into the sequence. To address this lack of independence, the proof leverages the Poisson equation technique [8, 13, 14, 22]. By applying the solution to the Poisson equation, the non-stationary Markovian noise is rewritten as the sum of a bounded martingale difference sequence and a telescoping-like correction term. This cleanly isolates the true martingale noise required for the central limit theorem.

Step 4: Bounding the Residual Components With the error decomposed, the non-martingale residual terms, such as the initialization decay and the Poisson equation correction terms, are bounded in the infinity norm. The analysis demonstrates that when these residual terms are scaled by $1/\sqrt{K}$, they vanish asymptotically at a fast enough rate and do not affect the limiting distribution.

Step 5: Applying the Martingale CLT and Sandwiching After bounding all remainder terms, a non-asymptotic martingale central limit theorem is applied directly to the isolated martingale difference sequence. This establishes that the normalized sum of the upper bounding process converges to the normal distribution in the 1-Wasserstein distance. Finally, since the lower bounding process $\Delta_k^{\downarrow}$ follows an analogous decomposition and converges to the exact same distribution, a sandwich argument concludes that the true averaged Q-learning error $\frac{1}{\sqrt{K}} \sum_{k=1}^{K} \Delta_k$ inherits this identical convergence rate.

3.3. Functional Central Limit Theorem

The FCLT is an important extension to the conventional CLT. Donsker's FCLT [12] states that the normalized partial sum process of i.i.d. random variables converges weakly to a Brownian motion in the Skorokhod space. In this section, we establish an FCLT for asynchronous Q-learning iterates, showing that the partial-sum process converges in distribution to a rescaled Brownian motion. Let $\mathcal{D}[0,1]$ denote the Skorokhod space. For $\zeta \in [0,1]$, we define the standardized partial sum processes associated with $\{Q_k\}_{k\geq 1}$ as

$$\Phi_K(\zeta) = \frac{1}{\sqrt{K}} \sum_{k=1}^{\lfloor \zeta K \rfloor} \Delta_k = \frac{1}{\sqrt{K}} \sum_{k=1}^{\lfloor \zeta K \rfloor} (Q_k - Q^*).$$

Theorem 7 *Under the setting of Theorem 5, the partial sum process $\Phi_K(\cdot)$ converges weakly to $(A^{-1}\Sigma A^{-\top})^{1/2}\boldsymbol{B}(\cdot)$ on $\mathcal{D}[0,1]$, where $\boldsymbol{B}(\cdot)$ is the standard Brownian motion on $[0,1]$.*

We can see that the conventional CLT is a special case of the FCLT when $\zeta = 1$. As the FCLT provides a basis for the asymptotic normality of certain functionals of stochastic processes, it is important for uncertainty quantification and statistical inference in Q-learning. Previous works have established the FCLT for synchronous Q-learning [21, 26, 40]. A recent work by Zhang and Xie [43] established a FCLT for asynchronous Q-learning with a constant step size. In contrast, our result concerns diminishing step-sizes.

4. Conclusion

We present a non-asymptotic central limit theorem for asynchronous averaged Q-learning, showing that the averaged iterate converges to a normal distribution in the Wasserstein distance at a rate of $\tilde{O}\left((|\mathcal{S}||\mathcal{A}|)^{\frac{1}{2}}K^{-\frac{1}{6}}\rho^{-2}(1-\gamma)^{-3}\right)$. We also derive a functional CLT, showing weak convergence of the partial-sum process to a Brownian motion. Compared with linear stochastic approximation and TD learning, the analysis of Q-learning poses additional challenges due to its nonlinearity and the non-stationarity of the process. Asynchronous updates further complicate the problem by introducing Markovian noise. This work identifies and addresses all of these challenges to provide the first non-asymptotic CLT for Q-learning. An important future direction is to strengthen the convergence rate and to extend the results to other metrics beyond the 1-Wasserstein distance.

References

[1] Andreas Anastasiou, Krishnakumar Balasubramanian, and Murat A Erdogdu. Normal approximation for stochastic gradient descent via non-asymptotic rates of martingale clt. In *Conference on Learning Theory*, pages 115–137. PMLR, 2019.

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

[3] Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference on learning theory*, pages 1691–1692. PMLR, 2018.

[4] Vivek Borkar, Shuhang Chen, Adithya Devraj, Ioannis Kontoyiannis, and Sean Meyn. The ode method for asymptotic statistics in stochastic approximation and reinforcement learning. *The Annals of Applied Probability*, 35(2):936–982, 2025.

[5] Vivek S Borkar and Vivek S Borkar. *Stochastic approximation: a dynamical systems viewpoint*, volume 9. Springer, 2008.

[6] Siddharth Chandak. $o(1/k)$ finite-time bound for non-linear two-time-scale stochastic approximation. *arXiv preprint arXiv:2504.19375*, 2025.

[7] Siddharth Chandak, Shaan Ul Haque, and Nicholas Bambos. Finite-time bounds for two-time-scale stochastic approximation with arbitrary norm contractions and markovian noise. *arXiv preprint arXiv:2503.18391*, 2025.

[8] Shuhang Chen, Adithya Devraj, Ana Busic, and Sean Meyn. Explicit mean-square error bounds for monte-carlo and linear stochastic approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 4173–4183. PMLR, 2020.

[9] 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, 2020.

[10] 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.

[11] 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, 2022.

[12] Monroe David Donsker. *An invariance principle for certain probability limit theorems*. 1951.

[13] Randal Douc, Eric Moulines, Pierre Priouret, Philippe Soulier, Randal Douc, Eric Moulines, Pierre Priouret, and Philippe Soulier. *Markov chains: Basic definitions*. Springer, 2018.

[14] Peter W Glynn and Sean P Meyn. A liapounov bound for solutions of the poisson equation. *The Annals of Probability*, pages 916–931, 1996.

[15] Peter Hall and Christopher C Heyde. *Martingale limit theory and its application*. Academic press, 2014.

[16] Yuze Han, Xiang Li, Jiadong Liang, and Zhihua Zhang. Decoupled functional central limit theorems for two-time-scale stochastic approximation. *arXiv preprint arXiv:2412.17070*, 2024.

[17] Jie Hu, Vishwaraj Doshi, et al. Central limit theorem for two-timescale stochastic approximation with markovian noise: Theory and applications. In *International Conference on Artificial Intelligence and Statistics*, pages 1477–1485. PMLR, 2024.

[18] Sajad Khodadadian and Martin Zubeldia. A general-purpose theorem for high-probability bounds of stochastic approximation with polyak averaging. *arXiv preprint arXiv:2505.21796*, 2025.

[19] Seo Taek Kong, Sihan Zeng, Thinh T Doan, and R Srikant. Nonasymptotic clt and error bounds for two-time-scale stochastic approximation. *arXiv preprint arXiv:2502.09884*, 2025.

[20] 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.

[21] 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*, pages 2207–2261. PMLR, 2023.

[22] Armand M Makowski and Adam Shwartz. The poisson equation for countable markov chains: probabilistic methods and interpretations. In *Handbook of Markov Decision Processes: Methods and Applications*, pages 269–303. Springer, 2002.

[23] 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.

[24] 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 *Conference on Learning Theory*, pages 2947–2997. PMLR, 2020.

[25] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[26] Saunak Kumar Panda, Ruiqi Liu, and Yisha Xiang. Asymptotic analysis of sample-averaged q-learning. *IEEE Transactions on Information Theory*, 2025.

[27] Boris T Polyak and Anatoli B Juditsky. Acceleration of stochastic approximation by averaging. *SIAM journal on control and optimization*, 30(4):838–855, 1992.

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

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

[30] Sergey Samsonov, Eric Moulines, Qi-Man Shao, Zhuo-Song Zhang, and Alexey Naumov. Gaussian approximation and multiplier bootstrap for polyak-ruppert averaged linear stochastic approximation with applications to td learning. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

[31] Sergey Samsonov, Marina Sheshukova, Eric Moulines, and Alexey Naumov. Statistical inference for linear stochastic approximation with markovian noise. *arXiv preprint arXiv:2505.19102*, 2025.

[32] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024.

[33] David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529 (7587):484–489, 2016.

[34] R Srikant. Rates of convergence in the central limit theorem for markov chains, with an application to td learning. *arXiv preprint arXiv:2401.15719*, 2024.

[35] Rayadurgam Srikant and Lei Ying. Finite-time error bounds for linear stochastic approximation andtd learning. In *Conference on Learning Theory*, pages 2803–2830. PMLR, 2019.

[36] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*, volume 1. MIT press Cambridge, 1998.

[37] Jie Tan, Tingnan Zhang, Erwin Coumans, Atil Iscen, Yunfei Bai, Danijar Hafner, Steven Bohez, and Vincent Vanhoucke. Sim-to-real: Learning agile locomotion for quadruped robots. *arXiv preprint arXiv:1804.10332*, 2018.

[38] M. J. Wainwright. Stochastic approximation with cone-contractive operators: Sharp ℓ_∞-bounds for q-learning. Technical Report arxiv:1905.06265, UC Berkeley, May 2019.

[39] Christopher JCH Watkins and Peter Dayan. Q-learning. *Machine learning*, 8:279–292, 1992.

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

[41] Andy Zeng, Shuran Song, Johnny Lee, Alberto Rodriguez, and Thomas Funkhouser. Tossingbot: Learning to throw arbitrary objects with residual physics. *IEEE Transactions on Robotics*, 36(4):1307–1319, 2020.

[42] Shangtong Zhang, Remi Tachet Des Combes, and Romain Laroche. Global optimality and finite sample analysis of softmax off-policy actor critic under state distribution mismatch. *Journal of Machine Learning Research*, 23(343):1–91, 2022.

[43] Yixuan Zhang and Qiaomin Xie. Constant stepsize q-learning: Distributional convergence, bias and extrapolation. *arXiv preprint arXiv:2401.13884*, 2024.

Appendix A. Proof of Theorem 5

We begin by establishing the following lemma.

A.1. Proof of Lemma 8

Lemma 8 *Denote $\Delta_k = Q_k - Q^*$. For all $k \in [K]$, if $\alpha_k \leq 1$, then Δ_k is bounded as follows:*

$$\Delta_k^{\downarrow} \leq \Delta_k \leq \Delta_k^{\uparrow},$$

where $\Delta_0^{\downarrow} = \Delta_0 = \Delta_0^{\uparrow}$ and the upper and lower bounds evolve according to

$$\Delta_{k+1}^{\uparrow} = (I - \alpha_k D + \alpha_k \gamma D P^{\pi^*})\Delta_k^{\uparrow} + \alpha_k \gamma D(P^{\pi_k} - P^{\pi^*})\Delta_k + \alpha_k(F_k - \bar{F}_k),$$

and

$$\Delta_{k+1}^{\downarrow} = (I - \alpha_k D + \alpha_k \gamma D P^{\pi^*})\Delta_k^{\downarrow} + \alpha_k(F_k - \bar{F}_k).$$

Proof We first show that

$$\Delta_{k+1} = (I - \alpha_k D + \alpha_k \gamma D P^{\pi^*})\Delta_k + \alpha_k \gamma D(P^{\pi_k} - P^{\pi^*})Q_k + \alpha_k(F_k - \bar{F}_k). \tag{3}$$

By the asynchronous Q-learning update rule, we have

$$\begin{aligned}
Q_{k+1} &= Q_k + \alpha_k(F_k - Q_k) \\
&= Q_k + \alpha_k(\bar{F}_k - Q_k) + \alpha_k(F_k - \bar{F}_k) \\
&= Q_k + \alpha_k(D\mathcal{T}(Q_k) + (I - D)Q_k - Q_k) + \alpha_k(F_k - \bar{F}_k) \\
&= Q_k + \alpha_k(D\mathcal{T}(Q_k) - DQ_k) + \alpha_k(F_k - \bar{F}_k) \\
&= Q_k + \alpha_k D(\mathcal{T}(Q_k) - Q_k) + \alpha_k(F_k - \bar{F}_k).
\end{aligned}$$

Subtracting Q^* from both sides yields

$$\begin{aligned}
Q_{k+1} - Q^* &= Q_k + \alpha_k D(\mathcal{T}(Q_k) - Q_k) + \alpha_k(F_k - \bar{F}_k) - Q^* \\
&= (I - \alpha_k D)Q_k + \alpha_k D\mathcal{T}(Q_k) + \alpha_k(F_k - \bar{F}_k) - Q^* \\
&= (I - \alpha_k D)(Q_k - Q^*) + \alpha_k D(\mathcal{T}(Q_k) - Q^*) + \alpha_k(F_k - \bar{F}_k).
\end{aligned}$$

Therefore, using the definition of Δ_k, we obtain

$$\Delta_{k+1} = (I - \alpha_k D)\Delta_k + \alpha_k D(\mathcal{T}(Q_k) - Q^*) + \alpha_k(F_k - \bar{F}_k). \tag{4}$$

Let $V_k(s) := \max_a Q_k(s,a) = Q_k(s, \pi_k(s))$ and define $P^\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}$ such that $P^\pi Q = P(\pi Q)$, we observe

$$\begin{aligned}
\alpha_k D(\mathcal{T}(Q_k) - Q^*) &= \alpha_k D((r + \gamma P V_k) - (r + \gamma P V^*)) \\
&= \alpha_k \gamma D(P V_k - P V^*) \\
&= \alpha_k \gamma D(P^{\pi_k} Q_k - P^{\pi^*} Q^*) \\
&= \alpha_k \gamma D(P^{\pi_k} Q_k - P^{\pi^*} Q_k + P^{\pi^*} Q_k - P^{\pi^*} Q^*)
\end{aligned}$$

$$= \alpha_k \gamma D(P^{\pi_k} - P^{\pi^*})Q_k + \alpha_k \gamma D P^{\pi^*}(Q_k - Q^*).$$

Thus, eq. (3) holds by substituting the above expression into eq. (4).

Next, we prove $\Delta_k^{\downarrow} \leq \Delta_k \leq \Delta_k^{\uparrow}$ by induction. The base case $k = 0$ holds by initialization. Suppose the statement holds at k. We observe that, since α_k and the entries in matrices D and P^{π^*} are all bounded between 0 and 1, the entries in matrix $I - \alpha_k D + \alpha_k \gamma D P^{\pi^*}$ are nonnegative. Consequently,

$$(I - \alpha_k D + \alpha_k \gamma D P^{\pi^*})\Delta_k^{\downarrow} \leq (I - \alpha_k D + \alpha_k \gamma D P^{\pi^*})\Delta_k \leq (I - \alpha_k D + \alpha_k \gamma D P^{\pi^*})\Delta_k^{\uparrow}.$$

We now have

$$\begin{aligned}
\Delta_{k+1}^{\downarrow} &= (I - \alpha_k D + \alpha_k \gamma D P^{\pi^*})\Delta_k^{\downarrow} + \alpha_k(F_k - \bar{F}_k) \\
&\leq (I - \alpha_k D + \alpha_k \gamma D P^{\pi^*})\Delta_k + \alpha_k(F_k - \bar{F}_k) \\
&\leq (I - \alpha_k D + \alpha_k \gamma D P^{\pi^*})\Delta_k + \alpha_k \gamma D(P^{\pi_k} - P^{\pi^*})Q_k + \alpha_k(F_k - \bar{F}_k) \\
&= \Delta_{k+1},
\end{aligned}$$

where the last inequality holds because $(P^{\pi_k} - P^{\pi^*})Q_k \geq 0$, as π_k is greedy w.r.t. Q_k. We remark that π_k is the greedy policy w.r.t. Q_k over all states, as implied by the definition of the Bellman optimality operator $\mathcal{T}$. Next, we have

$$\begin{aligned}
\Delta_{k+1} &= (I - \alpha_k D + \alpha_k \gamma D P^{\pi^*})\Delta_k + \alpha_k \gamma D(P^{\pi_k} - P^{\pi^*})Q_k + \alpha_k(F_k - \bar{F}_k) \\
&\leq (I - \alpha_k D + \alpha_k \gamma D P^{\pi^*})\Delta_k^{\uparrow} + \alpha_k \gamma D(P^{\pi_k} - P^{\pi^*})Q_k + \alpha_k(F_k - \bar{F}_k) \\
&= (I - \alpha_k D + \alpha_k \gamma D P^{\pi^*})\Delta_k^{\uparrow} + \alpha_k \gamma D(P^{\pi_k} - P^{\pi^*})\Delta_k + \alpha_k \gamma D(P^{\pi_k} - P^{\pi^*})Q^* + \alpha_k(F_k - \bar{F}_k) \\
&\leq (I - \alpha_k D + \alpha_k \gamma D P^{\pi^*})\Delta_k^{\uparrow} + \alpha_k \gamma D(P^{\pi_k} - P^{\pi^*})\Delta_k + \alpha_k(F_k - \bar{F}_k) \\
&= \Delta_{k+1}^{\uparrow},
\end{aligned}$$

where the last inequality holds because $(P^{\pi_k} - P^{\pi^*})Q^* \leq 0$, as π^* is greedy w.r.t. Q^*. Thus, the statement holds at $k + 1$, which completes the proof. ■

A.2. Proof of Theorem 5

Proof We first recall

$$\Delta_{k+1}^{\uparrow} = (I - \alpha_k D + \alpha_k \gamma D P^{\pi^*})\Delta_k^{\uparrow} + \alpha_k \gamma D(P^{\pi_k} - P^{\pi^*})\Delta_k + \alpha_k(F_k - \bar{F}_k).$$

Denoting $A = D - \gamma D P^{\pi^*}$, $Z_k = \gamma D(P^{\pi_k} - P^{\pi^*})\Delta_k$, and $Z_k' = F_k - \bar{F}_k$, by recursion we have

$$\Delta_{k+1}^{\uparrow} = \prod_{i=0}^{k}(I - \alpha_i A)\Delta_0 + \sum_{i=0}^{k}\left(\prod_{j=i+1}^{k}(I - \alpha_j A)\right)\alpha_i(Z_i + Z_i').$$

Thus,

$$\sum_{k=1}^{K}\Delta_k^{\uparrow} = \sum_{k=1}^{K}\prod_{i=0}^{k-1}(I - \alpha_i A)\Delta_0 + \sum_{k=1}^{K}\sum_{i=0}^{k-1}\left(\prod_{j=i+1}^{k-1}(I - \alpha_j A)\right)\alpha_i(Z_i + Z_i')$$

$$= \sum_{k=1}^{K} \prod_{i=0}^{k-1} (I - \alpha_i A)\Delta_0 + \sum_{i=0}^{K-1} \alpha_i \sum_{k=i+1}^{K} \left(\prod_{j=i+1}^{k-1} (I - \alpha_j A) \right) (Z_i + Z_i').$$

Denote $\Psi_i^K = \alpha_i \sum_{k=i+1}^{K} \left(\prod_{j=i+1}^{k-1} (I - \alpha_j A) \right)$. We further expand:

$$\begin{aligned}
&\sum_{k=1}^{K} \Delta_k^{\uparrow} \\
&= \sum_{k=1}^{K} \prod_{i=0}^{k-1} (I - \alpha_i A)\Delta_0 + \sum_{i=0}^{K-1} \Psi_i^K (Z_i + Z_i') \\
&= \sum_{k=1}^{K} \prod_{i=0}^{k-1} (I - \alpha_i A)\Delta_0 + \sum_{i=0}^{K-1} A^{-1}(Z_i + Z_i') + \sum_{i=0}^{K-1} (\Psi_i^K - A^{-1})(Z_i + Z_i') \\
&= \underbrace{\sum_{k=1}^{K} \prod_{i=0}^{k-1} (I - \alpha_i A)\Delta_0}_{\text{Term (1)}} + \underbrace{\sum_{i=0}^{K-1} A^{-1} Z_i}_{\text{Term (2)}} + \underbrace{\sum_{i=0}^{K-1} A^{-1} Z_i'}_{\text{Term (3)}} + \underbrace{\sum_{i=0}^{K-1} (\Psi_i^K - A^{-1}) Z_i}_{\text{Term (4)}} + \underbrace{\sum_{i=0}^{K-1} (\Psi_i^K - A^{-1}) Z_i'}_{\text{Term (5)}}.
\end{aligned} \tag{5}$$

Bounding Term (1). By applying Lemma 13 and using the bound $\|\Delta_0\|_\infty \leq \frac{1}{1-\gamma}$, we have $\|K^{-\frac{1}{2}} \text{Term (1)}\|_\infty \leq O\left(K^{-\frac{1}{2}} \rho^{\frac{-1}{1-\beta}} (1-\gamma)^{\frac{\beta-2}{1-\beta}}\right)$.

Bounding Term (2). We first expand the expression

$$A^{-1} Z_i = (D(I - \gamma P^{\pi^*}))^{-1} \gamma D (P^{\pi_i} - P^{\pi^*})(Q_i - Q^*).$$

Denoting $\rho := \min_{(s,a) \in \mathcal{S} \times \mathcal{A}} p(s,a)$, we observe

$$\|A^{-1}\|_\infty = \|(D(I - \gamma P^{\pi^*}))^{-1}\|_\infty \leq \frac{1}{(1-\gamma)\rho} \tag{6}$$

and by Assumption 1 and Lemma 16,

$$\|Z_i\|_\infty \leq \|(P^{\pi_i} - P^{\pi^*})(Q_i - Q^*)\|_\infty = L\|Q_i - Q^*\|_\infty^2 \leq O\left(\frac{t_i L}{\rho(1-\gamma)^2 i}\right) \tag{7}$$

where t_i is the mixing time. Thus,

$$\begin{aligned}
\left\| \frac{1}{\sqrt{K}} \sum_{i=0}^{K-1} A^{-1} Z_i \right\|_\infty &\leq \frac{1}{\sqrt{K}} \sum_{i=1}^{K} O\left(\frac{t_i L}{i(1-\gamma)^2 \rho}\right) \leq \frac{1}{\sqrt{K}} \cdot O\left(\frac{t_{\max} L}{(1-\gamma)^2 \rho}\right) \cdot \sum_{i=1}^{K} \frac{1}{i} \\
&\leq \tilde{O}\left(\frac{L}{\sqrt{K}(1-\gamma)^2 \rho}\right).
\end{aligned}$$

Decomposing Term (3). We now analyze the Markovian noise term

$$\sum_{i=0}^{K-1} A^{-1} Z_i' = \sum_{i=0}^{K-1} A^{-1}(F_i - \bar{F}_i) = \sum_{i=0}^{K-1} A^{-1}(F(Q_i, Y_i) - \mathbb{E}[F(Q_i, Y_i)]).$$

We decompose this term into two parts, where the first part has a bounded norm and the second part is a bounded martingale difference sequence. To this end, we use the Poisson equation technique [8, 13, 14, 22] to transform the Markovian noise into a martingale difference sequence. By a standard use of the technique [13], we know there exists a solution $X_k : \tilde{\mathcal{S}} \to \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ to the following Poisson's equation for all $k \in [K]$,

$$F(Q_k, Y_k) - \mathbb{E}[F(Q_k, Y_k)] = X_k(y_k) - \mathbb{E}[X_k(Y_{k+1})|Y_k = y_k].$$

For $i \in \tilde{\mathcal{S}}$, the closed form of the solution $X_k(i)$ is given by

$$X_k(i) = \sum_{j\in\tilde{\mathcal{S}}} [I - \tilde{P} - \mathbf{1}\tilde{\mu}^\top]^{-1}(i,j)(F(Q_k, i) - \bar{F}_k).$$

Under Assumption 2, there exists a constant $c_0 > 0$ and $\kappa \in (0,1)$ such that

$$\max_{i\in\tilde{S}} \|\tilde{P}^t(i,\cdot) - \tilde{\mu}(\cdot)\|_{\mathrm{TV}} \le c_0\kappa^t.$$

We now state two important properties for X_k. The first is a boundedness property that $\|X_k\|_\infty \le O(\frac{1}{(1-\gamma)(1-\kappa)})$, which follows directly from the above results. Next, we prove Lipschitzness by showing that

$$\begin{aligned}
\|X_k(i) - X_{k'}(i)\|_\infty &= \|\sum_{j\in\tilde{\mathcal{S}}} [I - \tilde{P} - \mathbf{1}\tilde{\mu}^\top]^{-1}(i,j)(F(Q_k, i) - F(Q_{k'}, i))\|_\infty \\
&\le \frac{c_0}{1-\kappa}\|F(Q_k, i) - F(Q_{k'}, i)\|_\infty \\
&\le \frac{2c_0}{1-\kappa}\|Q_k - Q_{k'}\|_\infty. \qquad \text{(By Lemma 17)}
\end{aligned}$$

We now decompose Term (3),

$$\begin{aligned}
&\sum_{k=0}^{K-1} A^{-1}(F(Q_k, Y_k) - \mathbb{E}[F(Q_k, Y_k)]) = \sum_{k=0}^{K-1} A^{-1}(X_k(Y_k) - \mathbb{E}[X_k(Y_{k+1})|Y_k]) \\
&= \sum_{k=0}^{K-1} A^{-1}(X_k(Y_k) - X_{k+1}(Y_{k+1}) + X_{k+1}(Y_{k+1}) - X_k(Y_{k+1}) + X_k(Y_{k+1}) - \mathbb{E}[X_k(Y_{k+1})|Y_k]) \\
&= \underbrace{A^{-1}(X_0(Y_0) - X_K(Y_K))}_{\text{Term (3a)}} + \underbrace{\sum_{k=0}^{K-1} A^{-1}(X_{k+1}(Y_{k+1}) - X_k(Y_{k+1}))}_{\text{Term (3b)}} \\
&\quad + \underbrace{\sum_{k=0}^{K-1} A^{-1}(X_k(Y_{k+1}) - \mathbb{E}[X_k(Y_{k+1})|Y_k])}_{\text{Term (3c)}}.
\end{aligned}$$

For Term (3a), note that $\|A^{-1}(X_0(Y_0) - X_K(Y_K))\|_\infty \le O(\frac{1}{(1-\gamma)^2(1-\kappa)\rho})$ by boundedness property. By the Lipschitzness property, for Term (3b) we obtain

$$\left\|\sum_{k=0}^{K-1} A^{-1}(X_{k+1}(Y_{k+1}) - X_k(Y_{k+1}))\right\|_\infty \le \frac{1}{(1-\gamma)\rho}\sum_{k=0}^{K-1}\|X_{k+1}(Y_{k+1}) - X_k(Y_{k+1})\|_\infty$$

$$\leq \frac{2c_0}{(1-\gamma)(1-\kappa)\rho}\sum_{k=0}^{K-1}\|Q_{k+1}-Q_k\|_\infty = \frac{2c_0}{(1-\gamma)(1-\kappa)\rho}\sum_{k=0}^{K-1}\|\alpha_k(F_k-Q_k)\|_\infty$$
$$\leq O\left(\frac{1}{(1-\gamma)^2(1-\kappa)\rho}\right)\sum_{k=0}^{K-1}\frac{1}{(k+b)^\beta} = O\left(\frac{K^{1-\beta}}{(1-\gamma)^2(1-\kappa)\rho}\right).$$

We have analyzed the first two terms. We defer the analysis of Term (3c) to the end of the proof.

Bounding Term (4). By combining eq. (12) and eq. (7), we have

$$\left\|\frac{1}{\sqrt{K}}\sum_{i=0}^{K-1}(\Psi_i^K - A^{-1})Z_i\right\|_\infty$$
$$\leq \frac{1}{\sqrt{K}}\sum_{i=1}^{K} O\left(\frac{1}{i(\rho(1-\gamma))^{\frac{2-\beta}{1-\beta}}} + \frac{(i-1)^\beta}{i\rho^2(1-\gamma)^2} + \frac{(1-\rho(1-\gamma)\alpha_K)^{K-i+1}}{\rho(1-\gamma)}\right)\cdot\frac{1}{\rho(1-\gamma)^2 i}$$
$$\leq \tilde{O}\left(\frac{1}{\sqrt{K}\rho^{\frac{3-2\beta}{1-\beta}}(1-\gamma)^{\frac{4-3\beta}{1-\beta}}}\right).$$

Bounding Term (5). Similarly to Term (3), we have

$$\sum_{k=0}^{K-1}(\Psi_k^K - A^{-1})Z_k' = \sum_{k=0}^{K-1}(\Psi_k^K - A^{-1})(X_k(Y_k) - \mathbb{E}[X_k(Y_{k+1})|Y_k])$$
$$= \sum_{k=0}^{K-1}(\Psi_k^K - A^{-1})(X_k(Y_k) - X_{k+1}(Y_{k+1}) + X_{k+1}(Y_{k+1}) - X_k(Y_{k+1}) + X_k(Y_{k+1}) - \mathbb{E}[X_k(Y_{k+1})|Y_k])$$
$$= \underbrace{\sum_{k=0}^{K-1}(\Psi_k^K - A^{-1})(X_k(Y_k) - X_{k+1}(Y_{k+1}))}_{\text{Term (5a)}} + \underbrace{\sum_{k=0}^{K-1}(\Psi_k^K - A^{-1})(X_{k+1}(Y_{k+1}) - X_k(Y_{k+1}))}_{\text{Term (5b)}}$$
$$+ \underbrace{\sum_{k=0}^{K-1}(\Psi_k^K - A^{-1})(X_k(Y_{k+1}) - \mathbb{E}[X_k(Y_{k+1})|Y_k])}_{\text{Term (5c)}}$$

Now we analyze each term individually. For Term (5a),

$$\sum_{k=0}^{K-1}(\Psi_k^K - A^{-1})(X_k(Y_k) - X_{k+1}(Y_{k+1}))$$
$$= \sum_{k=0}^{K-1}\left[(\Psi_k^K - A^{-1})X_k(Y_k) - (\Psi_{k+1}^K - A^{-1})X_{k+1}(Y_{k+1})\right.$$
$$\left.+(\Psi_{k+1}^K - A^{-1})X_{k+1}(Y_{k+1}) - (\Psi_k^K - A^{-1})X_{k+1}(Y_{k+1})\right]$$
$$= (\Psi_0^K - A^{-1})X_0(Y_0) - (\Psi_K^K - A^{-1})X_K(Y_K) + \sum_{k=0}^{K-1}\left[(\Psi_{k+1}^K - \Psi_k^K)X_{k+1}(Y_{k+1})\right].$$

By the boundedness of X_k, eq. (12), and Lemma 15, we obtain

$$\begin{aligned}\|\text{Term (5a)}\|_\infty &\le O\left(\frac{1}{(1-\gamma)(\rho(1-\gamma))^{\frac{2-\beta}{1-\beta}}}+\frac{1}{1-\gamma}\sum_{k=1}^{K}\frac{1}{k^\beta}\right)\\ &\le O\left(\frac{1}{(1-\gamma)(\rho(1-\gamma))^{\frac{2-\beta}{1-\beta}}}+\frac{K^{1-\beta}}{1-\gamma}\right).\end{aligned}$$

By eq. (12) and the Lipschitzness property of X_k, we have

$$\begin{aligned}\|\text{Term (5b)}\|_\infty &\le \sum_{k=1}^{K} O\left(\frac{1}{k(\rho(1-\gamma))^{\frac{2-\beta}{1-\beta}}}+\frac{(k-1)^\beta}{k\rho^2(1-\gamma)^2}+\frac{(1-\rho(1-\gamma)\alpha_K)^{K-k+1}}{\rho(1-\gamma)}\right)\cdot\frac{1}{k^\beta}\\ &\le \tilde{O}\left(\frac{1}{(\rho(1-\gamma))^{\frac{2-\beta}{1-\beta}}}+\frac{K^{1-\beta}}{\rho(1-\gamma)}\right).\end{aligned}$$

Next, we bound $\left\|\frac{1}{\sqrt{K}}\mathbb{E}[\text{Term (5c)}]\right\|_\infty$. We first note that $\{M_k,\mathcal{F}_k\}_{k\in[K]}$ is a martingale difference sequence where $\{M_k\}_{k\in[K]} := \{X_k(Y_{k+1})-\mathbb{E}[X_k(Y_{k+1})|Y_k]\}_{k\in[K]}$ and $\mathcal{F}_k$ is σ-field generated by all randomness until iteration k. Thus, by the martingale difference property we have $\mathbb{E}[M_k|\mathcal{F}_{k-1}]=0$ and $\mathbb{E}[\langle M_i,M_j\rangle]=\mathbb{E}[\langle M_i,\mathbb{E}[M_j|\mathcal{F}_{j-1}]\rangle]=0$ for $i\neq j$. This leads to

$$\begin{aligned}&\left\|\frac{1}{\sqrt{K}}\mathbb{E}[\text{Term (5c)}]\right\|_\infty^2\\ &=\left\|\frac{1}{\sqrt{K}}\mathbb{E}\sum_{k=0}^{K-1}(\Psi_k^K-A^{-1})(X_k(Y_{k+1})-\mathbb{E}[X_k(Y_{k+1})|Y_k])\right\|_\infty^2\\ &\le\frac{1}{(1-\gamma)^2K}\sum_{k=0}^{K-1}\|\Psi_k^K-A^{-1}\|_\infty^2\\ &\le\frac{1}{(1-\gamma)^2K}\sum_{i=1}^{K}O\left(\frac{1}{i(\rho(1-\gamma))^{\frac{2-\beta}{1-\beta}}}+\frac{(i-1)^\beta}{i\rho^2(1-\gamma)^2}+\frac{(1-\rho(1-\gamma)\alpha_K)^{K-i+1}}{\rho(1-\gamma)}\right)^2\\ &\qquad\qquad\text{(by eq. (12))}\\ &\le\frac{1}{(1-\gamma)^2}\cdot\tilde{O}\left(\frac{1}{K(\rho(1-\gamma))^{\frac{4-2\beta}{1-\beta}}}+\frac{1}{K^{1-\beta}\rho^4(1-\gamma)^4}\right).\end{aligned}$$

Thus,

$$\left\|\frac{1}{\sqrt{K}}\mathbb{E}[\text{Term (5c)}]\right\|_\infty\le\tilde{O}\left(\frac{1}{K^{1/2}(1-\gamma)(\rho(1-\gamma))^{\frac{2-\beta}{1-\beta}}}+\frac{1}{K^{1/2-\beta/2}\rho^2(1-\gamma)^3}\right).$$

Combining three terms, we have

$$\left\|\frac{1}{\sqrt{K}}\mathbb{E}[\text{Term (5)}]\right\|_\infty\le\tilde{O}\left(\frac{1}{K^{1/2}(1-\gamma)(\rho(1-\gamma))^{\frac{2-\beta}{1-\beta}}}+\frac{1}{K^{1/2-\beta/2}\rho^2(1-\gamma)^3}+\frac{1}{K^{\beta-1/2}\rho(1-\gamma)}\right).$$

Putting Everything Together. At this stage, we have decomposed $\sum_{k=1}^{K}\Delta_k^{\uparrow}$ into six components $\{\phi_i\}_{i=1}^{6}$, where ϕ_i corresponds to Term (i) for $i = 1,2,4,5$ and Term (3) is further split into $\phi_3 = A^{-1}(X_0(Y_0) - X_K(Y_K))$ and $\phi_6 = \sum_{i=0}^{K-1} A^{-1}(X_k(Y_{k+1}) - \mathbb{E}[X_k(Y_{k+1})|Y_k])$. Accordingly,

$$\sum_{k=1}^{K}\Delta_k^{\uparrow} = \sum_{i=1}^{6}\phi_i = \sum_{i=1}^{5}\phi_i + \sum_{k=0}^{K-1} A^{-1}(X_k(Y_{k+1}) - \mathbb{E}[X_k(Y_{k+1})|Y_k]) \tag{8}$$

where ϕ_6 is a bounded martingale difference sequence. Note we have also established bounds for $\{\phi_i\}_{i=1}^{5}$. Therefore, to establish CLTs for the averaged Q-learning iterates, we can apply any suitable known martingale CLTs. To proceed, we choose the non-asymptotic martingale CLT given in Srikant [34]. We prove in Lemma 10 that $\mathcal{W}_1\left(\frac{1}{\sqrt{K}}\phi_6, (A^{-1}\Sigma A^{-\top})^{1/2}\mathcal{N}(0,I))\right) \leq O\left(\frac{1}{(1-\gamma)\rho K^{\beta/2}}\right)$. Note that

$$\begin{aligned}
&\mathcal{W}_1\left(\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\Delta_k^{\uparrow}, (A^{-1}\Sigma A^{-\top})^{1/2}\mathcal{N}(0,I)\right) \\
&= \sup_{h\in \mathrm{Lip}_1} \mathbb{E}\left[h\left(\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\Delta_k^{\uparrow}\right) - h((A^{-1}\Sigma A^{-\top})^{1/2}\mathcal{N}(0,I))\right].
\end{aligned}$$

For any $h \in \mathrm{Lip}_1$, we have

$$\begin{aligned}
&\mathbb{E}\left[h\left(\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\Delta_k^{\uparrow}\right) - h((A^{-1}\Sigma A^{-\top})^{1/2}\mathcal{N}(0,I))\right] \\
&= \mathbb{E}\left[h\left(\frac{1}{\sqrt{K}}\sum_{i=1}^{6}\phi_i\right) - h((A^{-1}\Sigma A^{-\top})^{1/2}\mathcal{N}(0,I))\right] \\
&= \underbrace{\sum_{i=1}^{5}\mathbb{E}\left[h\left(\frac{1}{\sqrt{K}}\sum_{k=i}^{6}\phi_k\right) - h\left(\frac{1}{\sqrt{K}}\sum_{j=i+1}^{6}\phi_j\right)\right]}_{Ta} + \underbrace{\mathbb{E}\left[h\left(\frac{1}{\sqrt{K}}\phi_6\right)\right] - h((A^{-1}\Sigma A^{-\top})^{1/2}\mathcal{N}(0,I))}_{T_b}.
\end{aligned}$$

By Lemma 10, we have $T_b \leq O\left(((1-\gamma)\rho)^{-2-\beta}K^{-\beta/2}\right)$. To bound T_a, by combining all bounds analyzed above, merging alike terms, and ignoring constants, we have

$$\begin{aligned}
T_a &\leq \sum_{i=1}^{5}\mathbb{E}\left\|\frac{1}{\sqrt{K}}\phi_i\right\|_2 \leq \sum_{i=1}^{5}\sqrt{|\mathcal{S}||\mathcal{A}|}\mathbb{E}\left\|\frac{1}{\sqrt{K}}\phi_i\right\|_\infty \\
&\leq \frac{\sqrt{|\mathcal{S}||\mathcal{A}|}}{\rho(1-\gamma)^2}\cdot\tilde{O}\left(\frac{1}{K^{1/2}(\rho(1-\gamma))^{\frac{2-\beta}{1-\beta}}} + \frac{1}{K^{1/2-\beta/2}\rho(1-\gamma)} + \frac{1}{K^{\beta-1/2}}\right).
\end{aligned}$$

Thus, we have shown that

$$\mathcal{W}_1\left(\frac{1}{\sqrt{K}}\sum_{k=1}^{K}\Delta_k^{\uparrow}, \tilde{\mathcal{N}}\right) \leq R(K,\rho,1-\gamma,|\mathcal{S}|,|\mathcal{A}|)$$

where $\tilde{\mathcal{N}} := (A^{-1}\Sigma A^{-\top})^{1/2}\mathcal{N}(0, I)$ and

$$R(K, \rho, 1-\gamma, |\mathcal{S}|, |\mathcal{A}|) := \frac{\sqrt{|\mathcal{S}||\mathcal{A}|}}{\rho(1-\gamma)^2} \cdot \tilde{O}\left(\frac{1}{K^{1/2}(\rho(1-\gamma))^{\frac{2-\beta}{1-\beta}}} + \frac{1}{K^{1/2-\beta/2}\rho(1-\gamma)} + \frac{1}{K^{\beta-1/2}} + \frac{1}{K^{\beta/2}\rho^{1+\beta}(1-\gamma)^{\beta}} \right).$$

Next, we show that a similar convergence also holds for $\Delta_k^{\downarrow}$. By Lemma 8, we know

$$\Delta_{k+1}^{\downarrow} = (I - \alpha_k D + \alpha_k \gamma D P^{\pi^*})\Delta_k^{\downarrow} + \alpha_k(F_k - \bar{F}_k).$$

By a similar decomposition as in eq. (5), we obtain

$$\sum_{k=1}^{K} \Delta_k^{\downarrow} = \sum_{k=1}^{K} \prod_{i=0}^{k-1} (I - \alpha_i A)\Delta_0 + \sum_{i=0}^{K-1} A^{-1} Z_i' + \sum_{i=0}^{K-1} (\Psi_i^K - A^{-1}) Z_i'$$

which matches Term (1), Term (3), and Term (5) in eq. (5). Thus, following the same steps as before,

$$\mathcal{W}_1\left(\frac{1}{\sqrt{K}} \sum_{k=1}^{K} \Delta_k^{\downarrow}, \tilde{\mathcal{N}} \right) = R(K, \rho, 1-\gamma, |\mathcal{S}|, |\mathcal{A}|).$$

By Lemma 8, we have $\Delta_k^{\downarrow} \le \Delta_k \le \Delta_k^{\uparrow}$ for all $k \in [K]$. Therefore, we conclude that

$$\begin{aligned} \mathcal{W}_1\left(\frac{1}{\sqrt{K}} \sum_{k=1}^{K} \Delta_k, \tilde{\mathcal{N}} \right) &\le \mathcal{W}_1\left(\frac{1}{\sqrt{K}} \sum_{k=1}^{K} \Delta_k^{\uparrow}, \tilde{\mathcal{N}} \right) + \mathcal{W}_1\left(\frac{1}{\sqrt{K}} \sum_{k=1}^{K} \Delta_k^{\downarrow}, \tilde{\mathcal{N}} \right) \\ &= R(K, \rho, 1-\gamma, |\mathcal{S}|, |\mathcal{A}|). \end{aligned}$$

■

A.3. Martingale CLT

Theorem 9 (Restatement of Theorem 1 in Srikant [34]) *Let $\{m_k\}_{k\ge 1}$ be a d-dimensional martingale difference sequence with respect to a filtration $\{\mathcal{F}_k\}_{k\ge 0}$. Assume (i) $\mathbb{E}[\|m_k\|_2] \le \infty$ and $\mathbb{E}[m_k|\mathcal{F}_{k-1}] = 0$ for all $k \ge 1$; (ii) $\mathbb{E}[\|m_k\|_2^{2+\beta}]$ exists almost surely for all $k \ge 1$ and some $\beta \in (0,1)$ and (iii) $\Sigma_k = \mathbb{E}[m_k m_k^{\top}|\mathcal{F}_{k-1}]$ exists and further assume that $\lim_{n\to\infty}(\Sigma_1 + \cdots + \Sigma_n)/n = \Sigma_\infty$ almost surely for some positive definite Σ_∞. It follows that*

$$\mathcal{W}_1\left(\sum_{k=1}^{n} \frac{m_k}{\sqrt{n}}, \Sigma_\infty^{1/2}\mathcal{N}(0, I) \right) \le \frac{1}{\sqrt{n}} \sum_{k=1}^{n} O\left(\frac{\|\Sigma_\infty^{1/2}\|_{\mathrm{op}} \mathbb{E}[\|\Sigma_\infty^{-1/2} m_k\|_2^{\beta+2} + \|\Sigma_\infty^{-1/2} m_k\|_2^{\beta}]}{(n-k+1)^{(1+\beta)/2}} - \frac{1}{n-k+1} \mathrm{Tr}(M_k(\Sigma_\infty^{-1/2}\mathbb{E}[\Sigma_k]\Sigma_\infty^{-1/2} - I)) \right)$$

where M_k is a matrix with the property $\|M_k\|_{\mathrm{op}} \le O(\sqrt{n-k+1}\|\Sigma_\infty^{1/2}\|_{\mathrm{op}})$.

Lemma 10 *Under Assumption 2,*

$$\mathcal{W}_1\left(\frac{1}{\sqrt{K}}\sum_{k=1}^{K} A^{-1}(X_{k-1}(Y_k)-\mathbb{E}[X_{k-1}(Y_k)|Y_{k-1}]),\tilde{\mathcal{N}}\right)\leq O\left(((1-\gamma)\rho)^{-2-\beta}K^{-\beta/2}\right)$$

where $\tilde{\mathcal{N}}=A^{-1}\Sigma A^{-\top}\mathcal{N}(0,I)$ *and* $\Sigma:=\sum_{i,j\in\tilde{S}}\tilde{\mu}(i)\tilde{P}(i,j)(X(j)-\mathbb{E}[X(Y_1)|Y_0=i])(X(j)-\mathbb{E}[X(Y_1)|Y_0=i])^\top$.

Proof We first define $X:\tilde{S}\to\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ to be the solution of the following Poisson's equation,

$$F(Q^*,i)-\mathbb{E}[F(Q^*,i)]=X(i)-\mathbb{E}[X(Y_1)|Y_0=i]\ \text{ for all }\ i\in\tilde{S}.$$

Denote $\tilde{p}_t(i):=\mathbb{P}(Y_t=i)$. We further define the covariance matrix of the martingale noise characterized via the solution of Poisson's equation and its asymptotic matrix by

$$\tilde{\Sigma}_k=\sum_{i,j\in\tilde{S}}\tilde{p}_k(i)\tilde{P}(i,j)(X_k(j)-\mathbb{E}[X_k(Y_1)|Y_0=i])(X_k(j)-\mathbb{E}[X_k(Y_1)|Y_0=i])^\top$$

and

$$\Sigma:=\sum_{i,j\in\tilde{S}}\tilde{\mu}(i)\tilde{P}(i,j)(X(j)-\mathbb{E}[X(Y_1)|Y_0=i])(X(j)-\mathbb{E}[X(Y_1)|Y_0=i])^\top.$$

Now we can substitute $n=K$, $m_k=A^{-1}(X_{k-1}(Y_k)-\mathbb{E}[X_{k-1}(Y_k)|Y_{k-1}])$, $\Sigma_k=A^{-1}\tilde{\Sigma}_kA^{-\top}$, and $\Sigma_\infty=A^{-1}\Sigma A^{-\top}$ into Theorem 9. Note that under Assumption 2, the three conditions in Theorem 9 are satisfied. The rest of the proof follows from the proof of Theorem 2 in Srikant [34], with necessary modifications to accommodate our setting. To conclude, with the substitutions, we have

$$\sum_{k=1}^{n}\frac{\|\Sigma_\infty^{1/2}\|_{\text{op}}\mathbb{E}[\|\Sigma_\infty^{-1/2}m_k\|_2^{\beta+2}+\|\Sigma_\infty^{-1/2}m_k\|_2^{\beta}]}{(n-k+1)^{(1+\beta)/2}}\leq O\left(n^{(1-\beta)/2}/((1-\gamma)\rho)^{2+\beta}\right)$$

and

$$\sum_{k=1}^{n}\frac{1}{n-k+1}\text{Tr}(M_k(\Sigma_\infty^{-1/2}\mathbb{E}[\Sigma_k]\Sigma_\infty^{-1/2}-I))\leq O\left(1/(1-\gamma)^2\rho^2\right)$$

which completes the proof. ■

Appendix B. Proof of Theorem 7

Polish space is a separable and complete function space. It is a crucial structure for applying convergence in distribution results such as FCLT. Recall that we denote $\mathcal{D}[0,1]$ as the Skorokhod space. Equipped with the Skorokhod J_1 topology with a particular metric [28], $\mathcal{D}[0,1]$ is a Polish space. We use $\xrightarrow{\text{w}}$ to denote weak convergence for some sequence of random elements. To prove the theorem, we need the following result.

Proposition 11 *For two random sequences* $\{X_t\}_{t\geq 0}, \{Y_t\}_{t\geq 0} \subseteq \mathcal{D}[0,1]$ *satisfying* $\mathbb{E}[\sup_{\kappa\in[0,1]}\|Y_T(\kappa)\|] \to 0$ *and* $X_T \xrightarrow{\text{w}} X$*, we have* $X_T + Y_T \xrightarrow{\text{w}} X$.

Now we prove Theorem 7.

Proof For $\zeta \in [0,1]$, by a similar decomposition as in eq. (8), we have

$$\begin{aligned}\Phi_K^{\uparrow}(\zeta) &:= \frac{1}{\sqrt{K}}\sum_{k=1}^{\lfloor \zeta K\rfloor}\Delta_k^{\uparrow} = \frac{1}{\sqrt{K}}\sum_{i=1}^{6}\phi_i(\zeta)\\ &= \frac{1}{\sqrt{K}}\sum_{i=1}^{5}\phi_i(\zeta) + \frac{1}{\sqrt{K}}\sum_{k=1}^{\lfloor \zeta K\rfloor}A^{-1}(X_{k-1}(Y_k) - \mathbb{E}[X_{k-1}(Y_k)|Y_{k-1}]).\end{aligned}$$

From the proof of Theorem 5, we know $\sup_{\zeta\in[0,1]}\left\|\frac{1}{\sqrt{K}}\phi_i(\zeta)\right\|_\infty = o(1)$ for $i \in \{1,2,3,4,5\}$. Let X and Σ as defined in the proof of Lemma 10. The following lemma, which establishes the FCLT for $\frac{1}{\sqrt{K}}\phi_6(\zeta)$, is a direct consequence of Theorem 4.2 in Hall and Heyde [15].

Lemma 12 *For any* $\zeta \in [0,1]$,

$$\frac{1}{\sqrt{K}}\sum_{k=1}^{\lfloor \zeta K\rfloor}A^{-1}(X_{k-1}(Y_k) - \mathbb{E}[X_{k-1}(Y_k)|Y_{k-1}]) \xrightarrow{\text{w}} (A^{-1}\Sigma A^{-\top})^{1/2}\boldsymbol{B}(\zeta)$$

where $\boldsymbol{B}$ *is the standard Brownian motion and* $\Sigma := \sum_{i,j\in\tilde{S}}\tilde{\mu}(i)\tilde{P}(i,j)(X(j) - \mathbb{E}[X(Y_1)|Y_0 = i])(X(j) - \mathbb{E}[X(Y_1)|Y_0 = i])^\top$.

Thus, we have $\frac{1}{\sqrt{K}}\phi_6(\cdot) \xrightarrow{\text{w}} (A^{-1}\Sigma A^{-\top})^{1/2}\mathbf{B}(\cdot)$. Besides, we observe that

$$\sup_{\zeta\in[0,1]}\left\|\Phi_K^{\uparrow}(\zeta) - \frac{1}{\sqrt{K}}\phi_6(\zeta)\right\|_\infty \leq \sum_{i=1}^{5}\sup_{\zeta\in[0,1]}\left\|\frac{1}{\sqrt{K}}\phi_i(\zeta)\right\|_\infty = o(1),$$

which implies $\Phi_K^{\uparrow}(\cdot) \xrightarrow{\text{w}} (A^{-1}\Sigma A^{-\top})^{1/2}\mathbf{B}(\cdot)$ by Proposition 11. The FCLT for $\Phi_K^{\downarrow}(\cdot) := \frac{1}{\sqrt{K}}\sum_{k=1}^{\lfloor \cdot K\rfloor}\Delta_k^{\downarrow}$ can be established in the same way. Therefore, by the sandwich inequality, we have

$$\begin{aligned}&\sup_{\zeta\in[0,1]}\left\|\Phi_K(\zeta) - (A^{-1}\Sigma A^{-\top})^{1/2}\mathbf{B}(\zeta)\right\|_\infty\\ &\leq \sup_{\zeta\in[0,1]}\left\|\Phi_K^{\uparrow}(\zeta) - (A^{-1}\Sigma A^{-\top})^{1/2}\mathbf{B}(\zeta)\right\|_\infty + \sup_{\zeta\in[0,1]}\left\|\Phi_K^{\downarrow}(\zeta) - (A^{-1}\Sigma A^{-\top})^{1/2}\mathbf{B}(\zeta)\right\|_\infty = o(1),\end{aligned}$$

which implies that $\Phi_K(\cdot) \xrightarrow{\text{w}} (A^{-1}\Sigma A^{-\top})^{1/2}\mathbf{B}(\cdot)$. This completes the proof. ■

Appendix C. Supporting Lemmas

In this section we present several supporting lemmas. Lemma 13 and 15 analyze Term (1) and Ψ_i^K appeared in the proof of Theorem 5. Next, by leveraging the results in Chen et al. [10], Lemma 16 gives a non-asymptotic convergence rate for $\Delta_k = Q_k - Q^*$ under asynchronous updates. Lastly, Lemma 17 provides a Lipschitz property for the operator $F(\cdot, s)$ defined in eq. (2).

Lemma 13 *Let $\alpha_i = \alpha(i+b)^{-\beta}$ for some problem-dependent constants $\alpha, b > 0$ and $\beta \in (0,1)$. Then the following bounds hold:*

$$\left\|\sum_{k=1}^{K}\prod_{i=0}^{k-1}(I-\alpha_i A)\right\|_\infty \le O\left((\rho(1-\gamma))^{\frac{-1}{1-\beta}}\right),$$
$$\sum_{i=1}^{K}\|\Psi_i^K - A^{-1}\|_\infty \le \tilde{O}\left((\rho(1-\gamma))^{\frac{\beta-2}{1-\beta}} + \frac{K^\beta}{\rho^2(1-\gamma)^2}\right).$$

Proof The analysis of polynomial step sizes has been well studied in prior work (see, e.g., Li et al. [21], Polyak and Juditsky [27], Srikant [34]). However, due to a slightly modified choice of the step-size and the different update rule in the asynchronous setting, we provide a complete proof for the sake of completeness. Recall that $\alpha_i = \alpha(i+b)^{-\beta}$ and $\rho := \min_{(s,a)\in\mathcal{S}\times\mathcal{A}} p(s,a)$. We now have

$$\begin{aligned}
\left\|\sum_{k=1}^{K}\prod_{i=0}^{k-1}(I-\alpha_i A)\right\|_\infty &= \left\|\sum_{k=1}^{K}\prod_{i=0}^{k-1}(I-\alpha_i(D-\gamma DP^{\pi^*}))\right\|_\infty \\
&\le \sum_{k=1}^{K}\prod_{i=0}^{k-1}(1-\alpha_i\rho(1-\gamma)) \\
&= \sum_{k=1}^{K}\prod_{i=0}^{k-1}\left(1-\frac{\alpha\rho(1-\gamma)}{(i+b)^\beta}\right) \\
&\le \sum_{k=1}^{K}\exp\left(-\alpha\rho(1-\gamma)\sum_{i=0}^{k-1}(i+b)^{-\beta}\right). \qquad (1-x\le\exp(-x))
\end{aligned}$$

For $\beta \in (0,1)$, we have $\sum_{i=0}^{k-1}(i+b)^{-\beta} \ge \int_0^{k-1}(x+b)^{-\beta}dx = \frac{(k-1+b)^{1-\beta}+b^{1-\beta}}{1-\alpha}$,

$$\begin{aligned}
\left\|\sum_{k=1}^{K}\prod_{i=0}^{k-1}(I-\alpha_i A)\right\|_\infty &\le \sum_{k=1}^{K}\exp\left(-\alpha\rho(1-\gamma)\frac{(k-1+b)^{1-\beta}+b^{1-\beta}}{1-\alpha}\right) \\
&\le \int_1^\infty \exp\left(-\alpha\rho(1-\gamma)\frac{(k-1+b)^{1-\beta}+b^{1-\beta}}{1-\alpha}\right)dk
\end{aligned}$$

by the change of variable $u = -\alpha\rho(1-\gamma)\frac{(k-1+b)^{1-\beta}+b^{1-\beta}}{1-\alpha}$,

$$\le \frac{1}{\alpha\rho(1-\gamma)}\int_0^\infty\left(\frac{(1-\beta)u}{\alpha\rho(1-\gamma)}+b^{1-\beta}\right)^{\frac{\beta}{1-\beta}}\exp(-u)du$$

$$\leq \frac{\max\{2^{\frac{\beta}{1-\beta}},1\}}{\alpha\rho(1-\gamma)}\int_0^\infty \left(\left(\frac{(1-\beta)u}{\alpha\rho(1-\gamma)}\right)^{\frac{\beta}{1-\beta}}+b^\beta\right)\exp(-u)du.$$

Since $\int_0^\infty \exp(-u)du = 1$ and $\int_0^\infty u^{\frac{\beta}{1-\beta}}\exp(-u)du = \Gamma(\frac{1}{1-\beta}) \leq \frac{\sqrt{2\pi e}}{\sqrt{1-\beta}}(\frac{1}{1-\beta})^{\frac{\beta}{1-\beta}}$,

$$\begin{aligned}\left\|\sum_{k=1}^K\prod_{i=0}^{k-1}(I-\alpha_i A)\right\|_\infty &\leq \frac{\max\{2^{\frac{\beta}{1-\beta}},1\}}{\alpha\rho(1-\gamma)}\left(\left(\frac{1}{\alpha\rho(1-\gamma)}\right)^{\frac{\beta}{1-\beta}}\frac{\sqrt{2\pi e}}{\sqrt{1-\beta}}+b^\beta\right)\\ &\leq O\left(\frac{1}{(\alpha\rho(1-\gamma))^{\frac{1}{1-\beta}}(1-\beta)^{\frac{1}{2}}}\right).\end{aligned}$$

Next, we prove the second part. Recall $\Psi_i^K = \alpha_i\sum_{k=i+1}^K\left(\prod_{j=i+1}^{k-1}(I-\alpha_j A)\right)$. Since $A^{-1} = \alpha_i^{-1}(I-(I-\alpha_i A))$, we have

$$\begin{aligned}\Psi_i^K - A^{-1} &= (\Psi_i^K A - I)A^{-1}\\ &= \left(\sum_{t=i+1}^K\left(\prod_{j=i+1}^{t-1}(I-\alpha_j A)-\prod_{j=i}^{t-1}(I-\alpha_j A)\right)A^{-1}-A^{-1}\right)\\ &= \sum_{t=i+1}^K\left(\left(\prod_{j=i+1}^{t-1}(I-\alpha_j A)-\prod_{j=i}^{t-2}(I-\alpha_j A)\right)A^{-1}\right)-\left(\prod_{j=i}^K(I-\alpha_j A)\right)A^{-1}\\ &= \underbrace{\sum_{t=i+1}^K(\alpha_i-\alpha_t)\prod_{j=i+1}^{t-2}(I-\alpha_j A)}_{T_1}-\underbrace{\left(\prod_{j=i}^K(I-\alpha_j A)\right)A^{-1}}_{T_2}. \qquad (9)\end{aligned}$$

For T_1, since $A = D-\gamma DP^{\pi^*}$ and $1-x\leq\exp(-x)$, we have

$$\begin{aligned}\|T_1\|_\infty &= \left\|\sum_{t=i+1}^K(\alpha_i-\alpha_t)\prod_{j=i+1}^{t-2}(I-\alpha_j A)\right\|_\infty\\ &\leq \sum_{t=i+1}^K|\alpha_i-\alpha_t|\exp\left(-\sum_{j=i+1}^{t-2}\rho(1-\gamma)\alpha_j\right)\\ &\leq \sum_{t=i+1}^K\sum_{k=i}^{t-1}|\alpha_{k+1}-\alpha_k|\exp\left(-\sum_{j=i+1}^{t-2}\rho(1-\gamma)\alpha_j\right).\end{aligned}$$

Note that $\frac{\alpha_k-\alpha_{k+1}}{\alpha_k} = 1-\left(1-\frac{1}{k+1+b}\right)^\beta \leq 1-\exp(-\frac{\beta}{k+1+b}) \leq \frac{\beta}{k}$,

$$\|T_1\|_\infty \leq \sum_{t=i+1}^K\sum_{k=i}^{t-1}\frac{\beta\alpha_k}{k}\exp\left(-\sum_{j=i+1}^{t-2}\rho(1-\gamma)\alpha_j\right)$$

$$\leq \frac{\beta}{\rho(1-\gamma)i}\sum_{t=i+1}^{K}\sum_{k=i}^{t-1}\rho(1-\gamma)\alpha_k\exp\left(-\sum_{j=i+1}^{t-2}\rho(1-\gamma)\alpha_j\right)$$
$$\leq O\left(\frac{1}{i(\rho(1-\gamma))^{\frac{2-\beta}{1-\beta}}}+\frac{(i-1)^{\beta}}{i\rho^2(1-\gamma)^2}\right). \tag{10}$$

For T_2, we obtain

$$\|T_2\|_\infty = \left\|\left(\prod_{j=i}^{K}(I-\alpha_j A)\right)A^{-1}\right\|_\infty \leq \|A^{-1}\|_\infty\prod_{j=i}^{K}\|I-\alpha_j A\|_\infty$$
$$\leq \frac{\prod_{j=i}^{K}(1-\rho(1-\gamma)\alpha_j)}{\rho(1-\gamma)} \leq \frac{(1-\rho(1-\gamma)\alpha_K)^{K-i+1}}{\rho(1-\gamma)}. \tag{11}$$

Combining eqs. (9) to (11), we have

$$\|\Psi_i^K - A^{-1}\|_\infty = O\left(\frac{1}{i(\rho(1-\gamma))^{\frac{2-\beta}{1-\beta}}}+\frac{(i-1)^{\beta}}{i\rho^2(1-\gamma)^2}+\frac{(1-\rho(1-\gamma)\alpha_K)^{K-i+1}}{\rho(1-\gamma)}\right). \tag{12}$$

Therefore,

$$\sum_{i=1}^{K}\|\Psi_i^K - A^{-1}\|_\infty \leq \tilde{O}\left(\frac{1}{(\rho(1-\gamma))^{\frac{2-\beta}{1-\beta}}}+\frac{K^{\beta}}{\rho^2(1-\gamma)^2}\right).$$

■

Lemma 14 *Let $\alpha_i = \alpha(i+b)^{-\beta}$ for some problem-dependent constants $\alpha, b > 0$ and $\beta \in (0,1)$. It follows that*

$$\sum_{t=i+1}^{K}\sum_{k=i}^{t-1}\rho(1-\gamma)\alpha_k\exp\left(-\sum_{j=i+1}^{t-2}\rho(1-\gamma)\alpha_j\right) \leq O\left(\frac{1}{(\rho(1-\gamma))^{\frac{1}{1-\beta}}}+\frac{(i-1)^{\beta}}{\rho(1-\gamma)}\right).$$

Proof Since

$$\sum_{k=i}^{t-1}\rho(1-\gamma)\alpha_k \leq \frac{\rho\alpha(1-\gamma)}{1-\beta}((t-1)^{1-\beta}-(i-1)^{1-\beta}) \leq \sum_{k=i-1}^{t-2}\rho(1-\gamma)\alpha_k, \tag{13}$$

we have

$$\sum_{t=i+1}^{K}\sum_{k=i}^{t-1}\rho(1-\gamma)\alpha_k\exp\left(-\sum_{j=i+1}^{t-2}\rho(1-\gamma)\alpha_j\right)$$
$$= \sum_{t=i+1}^{K}\sum_{k=i}^{t-1}\rho(1-\gamma)\alpha_k\exp\left(-\sum_{j=i-1}^{t-2}\rho(1-\gamma)\alpha_j\right)\exp\left(\rho(1-\gamma)(\alpha_i+\alpha_{i-1})\right)$$

$$
\begin{aligned}
&\leq e \sum_{t=i+1}^{K} \sum_{k=i}^{t-1} \rho(1-\gamma)\alpha_k \exp\left(-\sum_{j=i-1}^{t-2} \rho(1-\gamma)\alpha_j\right) \\
&\leq e \sum_{t=i+1}^{K} u \exp(-u) \qquad \text{(let } u = \tfrac{\rho\alpha(1-\gamma)}{1-\beta}((t-1)^{1-\beta} - (i-1)^{1-\beta}) \text{ and by eq. (13))} \\
&\leq e \int_0^{\infty} u \exp(-u) \frac{1}{\rho\alpha(1-\gamma)} \left(\frac{1-\beta}{\rho\alpha(1-\gamma)} u + (i-1)^{1-\beta}\right)^{\frac{\beta}{1-\beta}} dt \\
&\leq \frac{e \max\{2^{\frac{\beta}{1-\beta}}, 2\}}{\rho\alpha(1-\gamma)} \int_0^{\infty} u \exp(-u) \left(\left(\frac{1-\beta}{\rho\alpha(1-\gamma)} u\right)^{\frac{\beta}{1-\beta}} + (i-1)^{\beta}\right) dt \\
&\leq \frac{e 2^{\frac{1}{1-\beta}}}{\rho\alpha(1-\gamma)} \left(\left(\frac{1-\beta}{\rho\alpha(1-\gamma)}\right)^{\frac{\beta}{1-\beta}} \Gamma\left(1 + \frac{1}{1-\beta}\right) + (i-1)^{\beta}\right) \\
&\leq O\left(\frac{1}{(\rho(1-\gamma))^{\frac{1}{1-\beta}}} + \frac{(i-1)^{\beta}}{\rho(1-\gamma)}\right).
\end{aligned}
$$

■

Lemma 15 *Let* $\Psi_k^K = \alpha_k \sum_{i=k+1}^{K} \left(\prod_{j=k+1}^{i-1}(I - \alpha_j A)\right)$. *For* $k \geq (\rho(1-\gamma))^{-1/\beta(1-\beta)}$, *we have*

$$
\left\|\Psi_{k+1}^K - \Psi_k^K\right\|_{\infty} \leq O\left(\frac{1}{k^{\beta}}\right).
$$

Proof First, we have

$$
\begin{aligned}
\Psi_{k+1}^K - \Psi_k^K &= \alpha_{k+1} \sum_{i=k+1}^{K} \left(\prod_{j=k+1}^{i-1} (I - \alpha_j A)\right) - \alpha_k \sum_{i=k}^{K} \left(\prod_{j=k}^{i-1} (I - \alpha_j A)\right) \\
&= (\alpha_{k+1} - \alpha_k) \sum_{i=k+1}^{K} \left(\prod_{j=k+1}^{i-1} (I - \alpha_j A)\right) \\
&\quad + \alpha_k \left[\sum_{i=k+1}^{K} \left(\prod_{j=k+1}^{i-1} (I - \alpha_j A)\right) - \sum_{i=k}^{K} \left(\prod_{j=k}^{i-1} (I - \alpha_j A)\right)\right].
\end{aligned}
$$

For the first term above, by a similar analysis as in the proof of Lemma 13 we have

$$
\begin{aligned}
\left\|(\alpha_{k+1} - \alpha_k) \sum_{i=k+1}^{K} \left(\prod_{j=k+1}^{i-1} (I - \alpha_j A)\right)\right\|_{\infty} &\leq \left(\frac{1}{k^{\beta}} - \frac{1}{(k+1)^{\beta}}\right) \cdot O\left(\frac{1}{(\rho(1-\gamma))^{\frac{1}{1-\beta}}}\right) \\
&\leq O\left(\frac{1}{k^{1+\beta}(\rho(1-\gamma))^{\frac{1}{1-\beta}}}\right) \leq O\left(\frac{1}{k}\right)
\end{aligned}
$$

for $k \geq (\rho(1-\gamma))^{-1/\beta(1-\beta)}$. For the second term, we observe

$$\begin{aligned}\sum_{i=k+1}^{K}\left(\prod_{j=k+1}^{i-1}(I-\alpha_j A)\right)-\sum_{i=k}^{K}\left(\prod_{j=k}^{i-1}(I-\alpha_j A)\right) &= I+\sum_{i=k}^{K}\left(\prod_{j=k+1}^{i-1}(I-\alpha_j A)-\prod_{j=k}^{i-1}(I-\alpha_j A)\right)\\ &= I+\sum_{i=k}^{K}\left(\alpha_k A\prod_{j=k+1}^{i-1}(I-\alpha_j A)\right)\\ &= O(I)\end{aligned}$$

for $k \geq (\rho(1-\gamma))^{-1/\beta(1-\beta)}$. Thus, the second term is of order α_k. Putting them together,

$$\left\|\Psi_{k+1}^K-\Psi_k^K\right\|_\infty \leq O\left(\frac{1}{k^\beta}\right).$$

■

The following lemmas provide finite-sample convergence guarantees of asynchronous Q-learning and Lipschitzness of the operator $F(\cdot, s)$ [10].

Lemma 16 (Theorem B.1 in Chen et al. [10]) *Let $\alpha_i = \alpha(i+b)^{-\beta}$ for some problem-dependent constants $\alpha, b > 0$ and $\beta \in (0,1)$. For the Q-learning updates in eq. (1), under Assumption 2, we have $\mathbb{E}\|Q_k - Q^*\|_\infty \leq O\left(\sqrt{\frac{t_k}{(1-\gamma)^2\rho^2 k}}\right)$, where $t_k = O(\log(1/\alpha_k))$ denotes the mixing time.*

Lemma 17 (Proposition 3.1 in Chen et al. [10]) *For the operator $F(\cdot,\cdot)$ defined in eq. (2), under Assumption 2, we have $\|F(Q_1,s) - F(Q_2,s)\|_\infty \leq 2\|Q_1 - Q_2\|_\infty$ for any $s \in \tilde{S}$.*