Non-Asymptotic Guarantees for Average-Reward Q-Learning with Adaptive Stepsizes

Zaiwei Chen
Edwardson School of Industrial Engineering
Purdue University
West Lafayette, IN 47906
chen5252@purdue.edu

Abstract

This work presents the first finite-time analysis of average-reward Q-learning with an asynchronous implementation. A key feature of the algorithm we study is the use of adaptive stepsizes that act as local clocks for each state-action pair. We show that the mean-square error of this Q-learning algorithm, measured in the span seminorm, converges at a rate of $\tilde{\mathcal{O}}(1/k)$. Technically, the use of adaptive stepsizes causes each Q-learning update to depend on the full sample history, introducing strong correlations and making the algorithm a non-Markovian stochastic approximation (SA) scheme. Our approach to overcoming this challenge involves (1) a time-inhomogeneous Markovian reformulation of non-Markovian SA, and (2) a combination of almost-sure time-varying bounds, conditioning arguments, and Markov chain concentration inequalities to break the strong correlations between the adaptive stepsizes and the iterates.

1 Introduction

Reinforcement Learning (RL) has become as a powerful framework for solving sequential decision-making problems, as demonstrated by its growing impact across a range of real-world applications, including autonomous robotics [22], game-playing AI [21], and the development of large language models [6]. Given the promising potential of RL, establishing strong theoretical foundations to guide its practical implementation is of significant importance.

An RL problem is typically modeled as a Markov decision process (MDP) [25], but its objective can vary by application. Finite-horizon RL maximizes cumulative rewards over a fixed time, while infinite-horizon discounted RL introduces a discount factor $\gamma \in (0, 1)$ to prioritize immediate rewards over future rewards. However, selecting an appropriate discount factor can be challenging. For instance, it is often unclear how far into the future decisions should account for rewards. Moreover, in high-frequency decision-making tasks such as robotic control, queuing systems, and financial trading, introducing a discount factor may not be appropriate. The infinite-horizon average-reward setting addresses these challenges by optimizing the long-run average reward without requiring a discount factor. However, the absence of discounting introduces unique challenges for both algorithm design and theoretical analysis. For example, the associated Bellman operator is no longer a norm-contractive mapping [20], sample-based updates add additional complexities to the structure of the algorithm [1], and optimality depends on value differences rather than absolute values.

These challenges are particularly evident in Q-learning [30], one of the most classical and practically impactful RL algorithms. Due to its popularity and its role as a major milestone in RL [18], substantial efforts have been dedicated to providing theoretical guarantees, especially in terms of convergence rates, for Q-learning. In the discounted setting, the first finite-time analysis of Q-learning was conducted in the early 2000s [8], followed by a series of works over the past two decades that

Preprint. Under review.

eventually led to matching upper and lower bounds [3, 15]. In contrast, in the average-reward setting, due to the aforementioned challenges, existing results are largely limited to asymptotic convergence [1, 12, 24, 28, 29, 32, 33] and regret analysis [2, 9, 31, 35]. Regarding finite-time analysis, even for Q-learning with synchronous updates (which requires a generative model for i.i.d. sampling), results have only appeared very recently [5, 10, 34]. To the best of our knowledge, no existing work provides a finite-time analysis for the last-iterate convergence of Q-learning with asynchronous updates based on a single trajectory of Markovian samples, which is the contribution of this work.

2 From Average-Reward RL to the Seminorm Bellman Equation

In this section, we first provide background on average-reward RL and then introduce the seminorm Bellman equation, which plays a central role in motivating both the algorithm design and the analysis of Q-learning in subsequent sections.

Background on Average-Reward RL. Consider an infinite-horizon, undiscounted MDP defined by the tuple $(\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R})$ [20], where $\mathcal{S}$ and $\mathcal{A}$ denote the finite state and action spaces, respectively. The transition dynamics are given by $\mathcal{P} = \{p(s'|s,a)\}_{(s,a,s')\in\mathcal{S}\times\mathcal{A}\times\mathcal{S}}$, where $p(s'|s,a)$ denotes the probability of transitioning to state s' after taking action a in state s. The stage-wise reward function is denoted by $\mathcal{R} : \mathcal{S}\times\mathcal{A} \to [-1,1]$. The transition probabilities and the reward function are unknown to the agent, but the agent can interact with the environment by taking actions, observing transitions, and receiving rewards.

Given a policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ (where $\Delta(\mathcal{A})$ denotes the set of probability distributions supported on $\mathcal{A}$), the average reward $r^\pi \in \mathbb{R}^{|\mathcal{S}|}$ is defined as $r^\pi(s) = \lim_{K\to\infty} \mathbb{E}_\pi[\sum_{k=1}^K \mathcal{R}(S_k, A_k)/K \mid S_1 = s]$ for all $s \in \mathcal{S}$, where the expectation $\mathbb{E}_\pi[\,\cdot\,]$ is taken over the randomness in the trajectory generated by the policy π. It is shown in standard MDP theory [20] that if the Markov chain induced by the policy π has a single recurrent class, the limit exists and is independent of the initial state. In this work, we operate under this setting. Consequently, we slightly abuse notation and use r^π to denote the scalar average reward associated with policy π. The goal is to find an optimal policy π^* that maximizes the average reward. Define the optimal relative Q-function (also known as the Q-value bias function) $Q^* \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ as $Q^*(s,a) = \mathbb{E}_{\pi^*}[\sum_{k=1}^\infty (\mathcal{R}(S_k, A_k) - r^*) \mid S_1 = s, A_1 = a]$ for all $(s,a) \in \mathcal{S}\times\mathcal{A}$, where $r^* \in \mathbb{R}$ is the optimal average reward. It is known that any policy π that satisfies $\pi(s) \in \arg\max_{a\in\mathcal{A}} Q^*(s,a)$ for all state $s \in \mathcal{S}$ is an optimal policy [20]. Therefore, the problem reduces to finding the optimal relative Q-function Q^*.

To find Q^*, we next introduce the Bellman equation. Let $\mathcal{H} : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ denote the Bellman operator defined as $[\mathcal{H}(Q)](s,a) = \mathcal{R}(s,a) + \mathbb{E}[\max_{a'\in\mathcal{A}} Q(S_2, a') \mid S_1 = s, A_1 = a]$ for all (s,a) and $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$. The Bellman equation is then given by

$$\mathcal{H}(Q) - Q = r^* e, \tag{1}$$

where e denotes the all-ones vector. It is well known that Q^* is a solution to Eq. (1) [20]. However, such a solution is not unique. To see this, observe that for any $c \in \mathbb{R}$, it follows directly from the definition of $\mathcal{H}(\cdot)$ that $\mathcal{H}(Q^* + ce) - (Q^* + ce) = \mathcal{H}(Q^*) + ce - Q^* - ce = \mathcal{H}(Q^*) - Q^* = r^* e$, implying that $Q^* + ce$ is also a solution to Eq. (1). Fortunately, for the purpose of finding an optimal policy, errors in the direction of the all-ones vector have no impact on the result, because $\pi^*(s) \in \arg\max_{a\in\mathcal{A}} Q^*(s,a) = \arg\max_{a\in\mathcal{A}}(Q^*(s,a) + c)$ for all $c \in \mathbb{R}$. Therefore, it suffices to find any point in the set $\mathcal{Q} := \{Q^* + ce \mid c \in \mathbb{R}\}$.

The Seminorm Bellman Equation. Due to the lack of discounting, the Bellman operator $\mathcal{H}(\cdot)$ is not a contraction mapping under any norm. However, it has been shown in [20] that the operator $\mathcal{H}(\cdot)$ can be a contraction mapping with respect to the span seminorm under certain additional assumptions on the underlying stochastic model (to be discussed shortly). Since this property forms the foundation of the Q-learning algorithm we will present, we begin by formally introducing the span seminorm and discussing its properties.

For any $x \in \mathbb{R}^d$, the span seminorm of x, denoted by $p_{\text{span}}(x)$, is defined as $p_{\text{span}}(x) = (\max_i x_i - \min_j x_j)/2$. Similar to a norm, the span seminorm is non-negative and satisfies the triangle inequality: $p_{\text{span}}(x+y) \le p_{\text{span}}(x) + p_{\text{span}}(y)$ for all $x, y \in \mathbb{R}^d$; and absolute homogeneity: $p_{\text{span}}(\alpha x) = |\alpha|\, p_{\text{span}}(x)$ for all $\alpha \in \mathbb{R}$ and $x \in \mathbb{R}^d$ [19, Section 6.6.1]. However, unlike a norm, $p_{\text{span}}(x) = 0$ does

not imply $x = 0$. In fact, the set $\{x \in \mathbb{R}^d \mid p_{\text{span}}(x) = 0\}$ is called the kernel of the span seminorm and is denoted by $\ker(p_{\text{span}})$. Since $p_{\text{span}}(x) = 0$ if and only if $\max_i x_i = \min_j x_j$, in which case all entries of x must be identical, we have $\ker(p_{\text{span}}) = \{ce \mid c \in \mathbb{R}\}$. Another important property of the span seminorm is that $p_{\text{span}}(x)$ can be interpreted as the distance from x to the linear subspace $\{ce|c \in \mathbb{R}\}$ with respect to $\|\cdot\|_\infty$. Since this result will be used frequently throughout the paper, we formally state it in the following lemma.

Lemma 2.1. *For any $x \in \mathbb{R}^d$, we have $\arg\min_{c\in\mathbb{R}} \|x - ce\|_\infty = (\max_i x_i + \min_j x_j)/2$. As a result, the span seminorm of x can be equivalently written as $p_{span}(x) = \min_{c\in\mathbb{R}} \|x - ce\|_\infty$.*

With $p_{\text{span}}(\cdot)$ properly introduced, we next state our assumption on the Bellman operator $\mathcal{H}(\cdot)$.

Assumption 2.1. The Bellman operator $\mathcal{H}(\cdot)$ is a β-contraction mapping with respect to $p_{\text{span}}(\cdot)$.

A sufficient condition for Assumption 2.1 to hold is $\max_{(s,a),(s',a')} \|p(\cdot \mid s, a) - p(\cdot \mid s', a')\|_{\text{TV}} < 1$, where $\|\cdot\|_{\text{TV}}$ denotes the total variation distance. In this case, Assumption 2.1 is satisfied with $\beta = \max_{(s,a),(s',a')} \|p(\cdot \mid s, a) - p(\cdot \mid s', a')\|_{\text{TV}}$. This condition is adopted from standard MDP theory textbooks [20], where it was used to study the convergence of relative value iteration [19, Proposition 6.6.1 and Theorem 6.6.2]. Since the goal of this work is to establish the convergence rate for Q-learning in the span seminorm contraction setting, we do not attempt to relax Assumption 2.1.

Under Assumption 2.1, we have the following result.

Lemma 2.2. $\{Q^* + ce \mid c \in \mathbb{R}\} = \{Q \mid \mathcal{H}(Q) - Q = r^* e\} = \{Q \mid p_{span}(\mathcal{H}(Q) - Q) = 0\}$.

As a result of Lemma 2.2, the seminorm fixed-point equation

$$p_{\text{span}}(\mathcal{H}(Q) - Q) = 0 \tag{2}$$

and the Bellman equation (1) are equivalent in the sense that they have the same set of solutions: $\mathcal{Q} = \{Q^* + ce \mid c \in \mathbb{R}\}$. Therefore, it suffices to find a solution to Eq. (2) in order to compute an optimal policy. For this reason, we shall refer to Eq. (2) as the seminorm Bellman equation, or simply the Bellman equation.

3 Main Results

To solve the Bellman equation (2), we next introduce a Q-learning algorithm with finite-time convergence guarantees. Recall that in the RL setting, the parameters of the stochastic model (i.e., the transition dynamics and the reward function) are unknown to the agent, but the agent can interact with the environment by taking actions, receiving rewards, and observing environment transitions.

Algorithm. Let π denote the policy used by the agent to interact with the environment, commonly referred to as the behavior policy. Our Q-learning algorithm is represented in Algorithm 1.

Algorithm 1 Q-Learning

1: **Input:** Initializations $Q_1 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, $S_1 \in \mathcal{S}$, and a behavior policy π.
2: **for** $k = 1, 2, \cdots,$ **do**
3: Take $A_k \sim \pi(\cdot \mid S_k)$, observe $S_{k+1} \sim p(\cdot \mid S_k, A_k)$, and receive $\mathcal{R}(S_k, A_k)$.
4: Compute the temporal difference: $\delta_k = R(S_k, A_k) + \max_{a'} Q_k(S_{k+1}, a') - Q_k(S_k, A_k)$.
5: Update the Q-function: $Q_{k+1}(s, a) = Q_k(s, a) + \alpha_k(s, a)\mathbb{1}_{\{(s,a)=(S_k,A_k)\}}\delta_k$ for all (s, a), where $\alpha_k(s, a) = \alpha/(N_k(s, a)+h)$, and $N_k(s, a) := \sum_{i=1}^k \mathbb{1}_{\{(S_i,A_i)=(s,a)\}}$ denotes the total number of visits to the state-action pair (s, a) up to the k-th iteration.
6: **end for**
7: **Output:** $\{Q_k\}_{k\geq 1}$

Note that Algorithm 1 is surprisingly simple and represents the most natural extension of Q-learning in the discounted setting [30]. In fact, the only modification (aside from setting the discount factor to one) is using adaptive stepsizes of the form $\alpha_k(s, a) = \alpha/(N_k(s, a)+h)$, where $\alpha, h > 0$ are tunable parameters. Importantly, the stepsize $\alpha_k(s, a)$ depends on the specific state-action pair through the counter $N_k(s, a)$ and is therefore not universal. Throughout, we refer to such stepsizes as *adaptive stepsizes*. Although adaptive stepsizes of this form have been used in the existing asymptotic analysis of Q-learning and temporal-difference learning in both the discounted and average-reward settings [1, 26], they have been less explored in the context of finite-time analysis.

Finite-Time Analysis. We begin by stating our assumption regarding the behavior policy.

Assumption 3.1. The behavior policy satisfies $\pi(a|s) > 0$ for all (s, a), and the Markov chain $\{S_k\}$ induced by the behavior policy π is irreducible and aperiodic.

Assumption 3.1 is standard in the existing studies of both value-based and policy-based RL algorithms [16, 17, 23, 26, 27], and guarantees that all state-action pairs are visited infinitely often during learning. Specifically, under Assumption 3.1, the Markov chain $\{S_k\}$ induced by π has a unique stationary distribution, denoted by $\mu \in \Delta(\mathcal{S})$, which satisfies $\min_{s\in\mathcal{S}} \mu(s) > 0$ [14]. Moreover, there exist $C > 1$ and $\rho \in (0, 1)$ such that $\max_{s\in\mathcal{S}} \|p_\pi^k(S_k = \cdot|S_1 = s) - \mu(\cdot)\|_{\text{TV}} \leq C\rho^{k-1}$ for all $k \geq 1$ [14], where p_π denotes the transition kernel of the Markov chain $\{S_k\}$ induced by π.

To aid in the statement of our main theorem, we introduce the following notation. Let $\tau_k = \min\{t : C\rho^{t-1} \leq \alpha/(k + h)\}$, $b_k = \alpha|\mathcal{S}||\mathcal{A}| \log(\lceil \frac{k-1}{|\mathcal{S}||\mathcal{A}|} \rceil /h + 1)$, and $m_k = b_k + p_{\text{span}}(Q^*)$, where $\lceil x \rceil$ returns the smallest integer greater than or equal to x. Note that τ_k, b_k, and m_k all grow at most logarithmically in k. Let $D_{\min} = \min_{s,a} \mu(s)\pi(a|s)$, which is positive under Assumption 3.1.

Theorem 3.1. *Consider $\{Q_k\}$ generated by Algorithm 1. Suppose that Assumptions 2.1 and 3.1 are satisfied. Then, there exists $K > 0$ such that for any $k \in \{1, 2, \cdots, K\}$, we have $p_{span}(Q_k - Q^*) \leq b_k + p_{span}(Q^*)$ almost surely (a.s.), and for any $k \geq K + 1$, we have*

$$\mathbb{E}[p_{span}(Q_k - Q^*)^2] \leq \begin{cases} 3m_K^2 \left(\frac{K+h}{k+h}\right)^{\frac{\alpha(1-\beta)}{2}} + \frac{C_1\tau_k(m_k+1)^2}{(k+h)^{\frac{\alpha(1-\beta)}{2}}}, & \text{if } \alpha(1-\beta) < 2, \\ 3m_K^2 \left(\frac{K+h}{k+h}\right) + \frac{C_2\tau_k(m_k+1)^2\log(k+h)}{(k+h)}, & \text{if } \alpha(1-\beta) = 2, \\ 3m_K^2 \left(\frac{K+h}{k+h}\right)^{\frac{\alpha(1-\beta)}{2}} + \frac{C_1\tau_k(m_k+1)^2}{(k+h)}, & \text{if } \alpha(1-\beta) > 2. \end{cases}$$

Here, C_1 and C_2 are problem-dependent constants defined as

$$C_1 = \frac{C\alpha^2|\mathcal{S}||\mathcal{A}|\log(|\mathcal{S}||\mathcal{A}|)}{(1-\rho)(1-\beta)D_{\min}^2\min(1,\alpha)|2-(1-\beta)\alpha|}, \text{ and } C_2 = \frac{C|\mathcal{S}||\mathcal{A}|\log(|\mathcal{S}||\mathcal{A}|)}{(1-\rho)(1-\beta)^3D_{\min}^2},$$

where c_1 and c_2 are absolute constants.

Due to the presence of Markovian noise and the use of adaptive stepsizes, the convergence bound in Theorem 3.1 does not hold from the initial iteration. Specifically, prior to iteration K, we establish an almost-sure bound that grows at most logarithmically with k. After iteration K, the "averaging" effect becomes dominant, and the mean-square error begins to decay, with the rate of convergence depending critically on the choice of the constant α in Algorithm 1. In particular, if α is below the threshold $2/(1 - \beta)$, the convergence rate is $\mathcal{O}(k^{-\alpha(1-\beta)/2})$, which can be arbitrarily slow. Conversely, if α exceeds the threshold, the convergence rate improves to the optimal $\mathcal{O}(1/k)$ up to logarithmic factors. A qualitatively similar phenomenon has been observed in norm-contractive SA algorithms [7], linear SA algorithms [4], and stochastic gradient descent/ascent algorithms [13].

Based on Theorem 3.1, we have the following corollary for the sample complexity.

Corollary 3.1.1. *Given $\epsilon > 0$, to achieve $\mathbb{E}[p_{span}(Q_k - Q^*)] \leq \epsilon$ with Algorithm 1, the sample complexity is $\tilde{\mathcal{O}}\left(|\mathcal{S}|^3|\mathcal{A}|^3D_{\min}^{-2}(1-\beta)^{-5}\epsilon^{-2}\right)$.*

Notably, the dependence on the desired accuracy level is $\tilde{\mathcal{O}}(\epsilon^{-2})$, which is unimprovable in general. While we make the dependence on the size of the state–action space and the seminorm contraction factor explicit, these terms are by no means optimal in light of existing information-theoretic lower bounds [11]. It is worth noting, however, that the lower bound in [11] was derived under a generative model that provides i.i.d. samples, whereas our analysis considers the more challenging Markovian sampling setting. Despite this discrepancy, tightening the dependencies on $|\mathcal{S}||\mathcal{A}|$, $1/(1 - \beta)$, and $D_{\min}$, whether through refined analysis or improved algorithmic techniques such as Polyak averaging or variance reduction, remains an important direction for future research.

References

[1] Abounadi, J., Bertsekas, D., and Borkar, V. S. (2001). Learning algorithms for Markov decision processes with average cost. *SIAM Journal on Control and Optimization*, 40(3):681–698.

[2] Agrawal, P. and Agrawal, S. (2024). Optimistic Q-learning for average reward and episodic reinforcement learning. *arXiv preprint arXiv:2407.13743*.

[3] Azar, M. G., Gómez, V., and Kappen, H. J. (2012). Dynamic policy programming. *The Journal of Machine Learning Research*, 13(1):3207–3245.

[4] Bhandari, J., Russo, D., and Singal, R. (2018). A finite-time analysis of temporal difference learning with linear function approximation. In *Conference on learning theory*, pages 1691–1692. PMLR.

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

[6] Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

[7] Chen, Z., Maguluri, S. T., Shakkottai, S., and Shanmugam, K. (2024). A Lyapunov theory for finite-sample guarantees of Markovian stochastic approximation. *Operations Research*, 72(4):1352–1367.

[8] Even-Dar, E., Mansour, Y., and Bartlett, P. (2003). Learning rates for Q-learning. *Journal of machine learning Research*, 5(1).

[9] Jaksch, T., Ortner, R., and Auer, P. (2010). Near-optimal regret bounds for reinforcement learning. *Journal of Machine Learning Research*, 11(Apr):1563–1600.

[10] Jin, Y., Gummadi, R., Zhou, Z., and Blanchet, J. (2024). Feasible Q-learning for average reward reinforcement learning. In *International Conference on Artificial Intelligence and Statistics*, pages 1630–1638. PMLR.

[11] Jin, Y. and Sidford, A. (2021). Towards tight bounds on the sample complexity of average-reward MDPs. In *International Conference on Machine Learning*, pages 5055–5064. PMLR.

[12] Kara, A. D. and Yuksel, S. (2023). Q-learning for continuous state and action MDPs under average cost criteria. *Preprint arXiv:2308.07591*.

[13] Lan, G. (2020). *First-Order and Stochastic Optimization Methods for Machine Learning*. Springer.

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

[15] Li, G., Cai, C., Chen, Y., Wei, Y., and Chi, Y. (2024a). Is Q-learning minimax optimal? a tight sample complexity analysis. *Operations Research*, 72(1):222–236.

[16] Li, G., Wei, Y., Chi, Y., Gu, Y., and Chen, Y. (2020). Sample complexity of asynchronous Q-learning: Sharper analysis and variance reduction. *Advances in neural information processing systems*, 33:7031–7043.

[17] Li, T., Wu, F., and Lan, G. (2024b). Stochastic first-order methods for average-reward Markov decision processes. *Mathematics of Operations Research*.

[18] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533.

[19] Puterman, M. L. (1995). Markov Decision Processes: Discrete Stochastic Dynamic Programming. *Journal of the Operational Research Society*, 46(6):792–792.

[20] Puterman, M. L. (2014). *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons.

[21] Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., et al. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419):1140–1144.

[22] Singh, B., Kumar, R., and Singh, V. P. (2022). Reinforcement learning in robotic applications: a comprehensive survey. *Artificial Intelligence Review*, 55(2):945–990.

[23] Srikant, R. and Ying, L. (2019). Finite-time error bounds for linear stochastic approximation and TD-learning. In *Conference on Learning Theory*, pages 2803–2830.

[24] Suttle, W., Zhang, K., Yang, Z., Liu, J., and Kraemer, D. (2021). Reinforcement learning for cost-aware Markov decision processes. In *International Conference on Machine Learning*, pages 9989–9999. PMLR.

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

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

[27] Tsitsiklis, J. N. and Van Roy, B. (1999). Average cost temporal-difference learning. *Automatica*, 35(11):1799–1808.

[28] Wan, Y., Naik, A., and Sutton, R. S. (2021). Learning and planning in average-reward Markov decision processes. In *International Conference on Machine Learning*, pages 10653–10662. PMLR.

[29] Wan, Y., Yu, H., and Sutton, R. S. (2024). On convergence of average-reward Q-learning in weakly communicating Markov decision processes. *Preprint arXiv:2408.16262*.

[30] Watkins, C. J. and Dayan, P. (1992). Q-learning. *Machine learning*, 8(3-4):279–292.

[31] Wei, C.-Y., Jahromi, M. J., Luo, H., Sharma, H., and Jain, R. (2020). Model-free reinforcement learning in infinite-horizon average-reward Markov decision processes. In *International conference on machine learning*, pages 10170–10180. PMLR.

[32] Yang, X., Hu, J., and Hu, J.-Q. (2024). Relative Q-learning for average-reward Markov decision processes with continuous states. *IEEE Transactions on Automatic Control*.

[33] Yu, H., Wan, Y., and Sutton, R. S. (2024). Asynchronous stochastic approximation and average-reward reinforcement learning. *Preprint arXiv:2409.03915*.

[34] Zhang, S., Zhang, Z., and Maguluri, S. T. (2021). Finite sample analysis of average-reward TD-learning and Q-learning. *Advances in Neural Information Processing Systems*, 34:1230–1242.

[35] Zhang, Z. and Xie, Q. (2023). Sharper model-free reinforcement learning for average-reward Markov decision processes. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 5476–5477. PMLR.