

# A FINITE-TIME ANALYSIS OF DISTRIBUTED Q-LEARNING

**Anonymous authors**

Paper under double-blind review

## ABSTRACT

Multi-agent reinforcement learning (MARL) has witnessed a remarkable surge in interest, fueled by the empirical success achieved in applications of single-agent reinforcement learning (RL). In this study, we consider a distributed Q-learning scenario, wherein a number of agents cooperatively solve a sequential decision making problem without access to the central reward function which is an average of the local rewards. In particular, we study finite-time analysis of a distributed Q-learning algorithm, and provide a new sample complexity result of  $\tilde{O}\left(\max\left\{\frac{1}{\epsilon^2} \frac{t_{\text{mix}}}{(1-\gamma)^6 d_{\text{min}}^4}, \frac{1}{\epsilon} \frac{\sqrt{|S||A|}}{(1-\sigma_2(\mathbf{W}))(1-\gamma)^4 d_{\text{min}}^3}\right\}\right)$  under tabular lookup setting for Markovian observation model.

## 1 INTRODUCTION

Multi-agent reinforcement learning (MARL) aims to solve a sequential decision making problem, where a number of agents sharing an environment collaborates. Accompanied by advancements in algorithms (Sunehag et al., 2017; Rashid et al., 2020), MARL has shown impressive success in various fields such as robotics (de Witt et al., 2020) and autonomous driving (Shalev-Shwartz et al., 2016). Beyond its empirical success, there has also been notable interest in theoretical investigations (Zhang et al., 2018b; Dou et al., 2022).

MARL has been studied under various scenarios including an access to central reward function (Tan, 1993; Claus and Boutilier, 1998; Littman, 2001). In particular, our interest lies in the the distributed learning paradigm where agents collaborate to solve a shared problem, constrained to communicate solely with their neighboring agents and does not have access to central reward function. Such setting has come of interest due to its wide applications (Blumenkamp et al., 2022; Prabuchandran et al., 2014; Zhao et al., 2021). Compared to scenarios where a centralized coordinate exists, the distributed paradigm has advantage in terms of privacy-preservation and scalability. One notable example is the distributed adaptation of temporal-difference (TD) learning, as demonstrated in studies by Doan et al. (2019); Wang et al. (2020); Lim and Lee (2023), to name a few.

Meanwhile, in the literature of single-agent RL, Q-learning (Watkins and Dayan, 1992) is one of the most important algorithms in RL. The non-linear max-operator in Q-learning algorithm imposes difficulty in the analysis, and its non-asymptotic analysis has been an active research area recently (Even-Dar et al., 2003; Chen et al., 2021; Lee et al., 2023; Li et al., 2024). However, distributed learning framework for Q-learning has not been studied in detail. In particular, distributed Q-learning has been studied in an asymptotic sense (Kar et al., 2013), i.e., the algorithm converges over time as it approaches infinity, or in a non-asymptotic sense under additional assumptions on the problem (Heredia et al., 2020; Zeng et al., 2022b). **Wang et al. (2022) studied a version of distributed Q-learning in tabular setting but differs from the one in Kar et al. (2013).** This motivates our study to understand its non-asymptotic behavior under tabular setup, i.e., all the state-action values are stored in a table. Our contribution can be summarized as follows:

1. For Markovian observation model, we provide the sample complexity  $\tilde{O}\left(\max\left\{\frac{1}{\epsilon^2} \frac{t_{\text{mix}}}{(1-\gamma)^6 d_{\text{min}}^4}, \frac{1}{\epsilon} \frac{\sqrt{|S||A|}}{(1-\sigma_2(\mathbf{W}))(1-\gamma)^4 d_{\text{min}}^3}\right\}\right)$  in terms of the infinity norm under

054 tabular setting. We derive, for the first time, the finite-time analysis of QD-learning (Kar  
 055 et al., 2013) in its original form, which is one of the most fundamental and widely  
 056 used distributed Q-learning methods. While several works have addressed other types  
 057 of distributed Q-learning, the analysis of QD-learning has remained unexplored until  
 058 now. Furthermore, we also provide a sample complexity result for the independent and  
 059 identically distributed (i.i.d.) observation model.

- 060 2. Our analysis relies on switched system modeling of Q-learning, providing new insights  
 061 for interpretation of distributed Q-learning algorithms. We show that the distributed  
 062 Q-learning also allows switched system interpretation as in the single-agent case.

063 **Related Works:**

064  
 065 The non-asymptotic behavior of distributed TD-learning was studied in Doan et al. (2019);  
 066 Sun et al. (2020); Wang et al. (2020); Lim and Lee (2023), which were motivated from the  
 067 distributed optimization and control literature (Nedic and Ozdaglar, 2009; Wang and Elia,  
 068 2010; Pu and Nedić, 2021). Distributed versions of various TD-learning algorithms were  
 069 investigated in Macua et al. (2014); Lee et al. (2018). As for actor-critic algorithm (Konda  
 070 and Tsitsiklis, 1999), its extension to distributed setting was studied in Zhang et al. (2018a;b);  
 071 Zhang and Zavlanos (2019); Zeng et al. (2022a). Meanwhile, Yang et al. (2023) considered  
 072 a distributed policy gradient approach. Moreover, Zhang et al. (2021) investigated distributed  
 073 algorithm for fitted Q-iteration, which is similar to solving a least squares problem. Fur-  
 074 thermore, a line of research has focused on dealing with exponential scaling in the action  
 075 space Lin et al. (2021); Qu et al. (2022); Zhang et al. (2023); Gu et al. (2024).

076 The distributed Q-learning algorithm under the setting when only the local reward is observ-  
 077 able, was first studied by Kar et al. (2013). They proposed the so-called QD-learning proving  
 078 asymptotic convergence using two-time scale stochastic approximation approaches. Zeng  
 079 et al. (2022b); Heredia et al. (2020) proved finite-time bounds of distributed Q-learning with  
 080 linear function approximation. However, the works require additional strong assumptions,  
 081 which may not hold even in the tabular setup. In particular, Zeng et al. (2022b) considered a  
 082 strongly monotone condition to hold, and Heredia et al. (2020) posed a particular assumption  
 083 on the state-action distribution. Wang et al. (2022) studied a distributed Q-learning model  
 084 motivated from the adapt-then-combine scheme (Chen and Sayed, 2012) in the distributed  
 085 optimization literature and provided a sample complexity bound in terms of high-probability.

086 Considering a single-agent case, the non-asymptotic analysis of Q-learning has made great  
 087 success. An incomplete list is provided in the following: An early result by Even-Dar et al.  
 088 (2003) studied the sample complexity under i.i.d. observation model. Lee et al. (2023)  
 089 developed a switched system method to analyze the behavior of Q-learning. Qu and Wierman  
 090 (2020) considered a shifted Martingale approach to deal with the Markovian observation  
 091 model. Li et al. (2024) proved the sample complexity using refined analysis under the  
 092 Markovian observation model.

093 Meanwhile, a separate line of research focusing on multi-agent problems is the federated  
 094 reinforcement learning literature (Khodadadian et al., 2022; Woo et al., 2023; Zheng et al.,  
 095 2023). This approach differs from the distributed learning scenario in two key aspects: it  
 096 employs a centralized controller, and all agents share a common reward function.

097 The paper is organized as follows: Section 2 provides background for the MARL setting.  
 098 Section 3 provides result under i.i.d. observation model and sketch of the proof. The result  
 099 for Markovian observation model is provided in Section 4.

100 **2 PRELIMINARIES**

101 **2.1 MULTI AGENT MDP**

102  
 103 A multi-agent Markov decision process (MAMDP) consists of the tuple  
 104  $(\mathcal{S}, \{\mathcal{A}_i\}_{i=1}^N, \mathcal{P}, \{r^i\}_{i=1}^N, \gamma)$ , where  $\mathcal{S} := \{1, 2, \dots, |\mathcal{S}|\}$  is the finite set of states,  $\mathcal{A}_i :=$   
 105  $\{1, 2, \dots, |\mathcal{A}_i|\}$  is the finite set of actions for each agent  $i \in \mathcal{V}$ ,  $\mathcal{P} : \mathcal{S} \times \prod_{i=1}^N \mathcal{A}_i \times \mathcal{S} \rightarrow [0, 1]$   
 106 is the transition probability, and  $r^i : \mathcal{S} \times \prod_{i=1}^N \mathcal{A}_i \times \mathcal{S} \rightarrow \mathbb{R}$  is the reward function of agent  
 107

108  $i \in \mathcal{V}$ . We will use the notation  $\mathcal{A} := \prod_{i=1}^N \mathcal{A}_i = \{1, 2, \dots, |\mathcal{A}|\}$  where tuple of actions are  
 109 mapped to unique integer.  $\gamma \in (0, 1)$  is the discount factor.

110  
 111 At time  $k \in \mathbb{N}$ , the agents share the state  $s \in \mathcal{S}$ , and each agent  $i \in \mathcal{V}$  selects an action  
 112  $a_i \in \mathcal{A}_i$  following its own policy  $\pi^i : \mathcal{S} \rightarrow \Delta^{|\mathcal{A}_i|}$ . The collection of the actions selected by  
 113 each agents are denoted as  $\mathbf{a} = (a_1, a_2, \dots, a_N)$ , and transition occurs to  $s' \sim \mathcal{P}(s, \mathbf{a}, \cdot)$ .  
 114 Each agents receives local reward  $r^i(s, \mathbf{a}, s')$ , which is not shared with other agents.

115 The main goal of MAMDP is to find a deterministic optimal policy,  $\pi^* := (\pi^1, \pi^2, \dots, \pi^N) : \mathcal{S} \rightarrow \mathcal{A}$   
 116 such that the average of cumulative discounted rewards of each agents is  
 117 maximized:  $\pi^* := \arg \max_{\pi \in \Omega} \mathbb{E} \left[ \sum_{k=0}^{\infty} \sum_{i=1}^N \frac{\gamma^k}{N} r^i(s_k, \mathbf{a}_k, s_{k+1}) \mid \pi \right]$ , where  $\Omega$  is the set  
 118 of possible deterministic policies, and  $\{(s_k, \mathbf{a}_k)\}_{k \geq 0}$  is a state-action trajectory gener-  
 119 ated by Markov chain under policy  $\pi$ . The Q-function for a policy  $\pi : \mathcal{S} \rightarrow \mathcal{A}$ , de-  
 120 notes the average of cumulative discounted rewards of each agents following the policy  
 121  $\pi$ , i.e.,  $Q^\pi(s, \mathbf{a}) := \mathbb{E} \left[ \sum_{k=0}^{\infty} \sum_{i=1}^N \frac{\gamma^k}{N} r_{k+1}^i \mid \pi, (s_0, a_0) = (s, \mathbf{a}) \right]$  for  $s \in \mathcal{S}, \mathbf{a} \in \mathcal{A}$ , where  
 122  $r_{k+1}^i := r^i(s_k, \mathbf{a}_k, s'_k)$ . The optimal Q-function,  $Q^{\pi^*}$ , which is the Q-function induced by the  
 123 optimal policy  $\pi^*$ , is denoted as  $Q^*$ . The optimal policy can be recovered via a greedy policy  
 124 over  $Q^*$ , i.e.,  $\pi^*(s) = \arg \max_{\mathbf{a} \in \mathcal{A}} Q^*(s, \mathbf{a})$  for  $s \in \mathcal{S}$ . The optimal Q-function,  $Q^*$  satisfies  
 125 the following so-called optimal Bellman equation (Bellman, 1966):

$$126 \quad Q^*(s, \mathbf{a}) = \mathbb{E} \left[ \frac{1}{N} \sum_{i=1}^N r^i(s, \mathbf{a}, s') + \gamma \max_{\mathbf{u} \in \mathcal{A}} Q^*(s', \mathbf{u}) \right], \quad \forall s \in \mathcal{S}, \mathbf{a} \in \mathcal{A}. \quad (1)$$

130 Since each agent only has an access to its local reward  $r^i$ , it is impossible to learn the central  
 131 optimal Q-function without sharing additional information among the agents. However,  
 132 we assume that there is no central coordinator that can communicate with all the agents.  
 133 Instead, we will consider a more restricted communication scenario where each agent can  
 134 share its learning parameter only with a subset of the agents. This communication constraint  
 135 can be caused by several reasons such as infrastructures, privacy, and spacial topology.  
 136 The communication structure among the agents can be described by an undirected simple  
 137 connected graph  $\mathcal{G} := (\mathcal{V}, \mathcal{E})$ , where  $\mathcal{V}$  denotes the set of vertices and  $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$  is the  
 138 set of edges. Each agent will be described by a vertex  $v \in \mathcal{V} := \{1, 2, \dots, N\}$ , where  $N$  is  
 139 the number of agents. Moreover, each agent  $i \in \mathcal{V}$  only communicates with its neighbours,  
 140 denoted as  $\mathcal{N}_i := \{j \in \mathcal{V} \mid (i, j) \in \mathcal{E}\}$ .

141 To further proceed, we will use the following matrix and vector notations:  $\mathbf{P} :=$   
 142  $[\mathbf{P}_{1,1} \quad \mathbf{P}_{1,2} \quad \dots \quad \mathbf{P}_{|\mathcal{S}|,|\mathcal{A}|}]^\top$ ,  $\mathbf{R}^i := [\mathbf{R}_1^{i\top} \quad \dots \quad \mathbf{R}_{|\mathcal{S}|}^{i\top}]^\top$  where  $\mathbf{P}_{s,\mathbf{a}} \in \mathbb{R}^{|\mathcal{S}|}$  and  $\mathbf{R}_s^i \in$   
 143  $\mathbb{R}^{|\mathcal{A}|}$  are column vectors such that  $[\mathbf{P}_{s,\mathbf{a}}]_{s'} = \mathcal{P}(s, \mathbf{a}, s')$  for  $s' \in \mathcal{S}$ , and  $[\mathbf{R}_s^i]_{\mathbf{a}} =$   
 144  $\mathbb{E} [r^i(s, \mathbf{a}, s') \mid s, \mathbf{a}]$ , respectively. We assume that  $\|\mathbf{R}^i\|_\infty \leq R_{\max}$  for some positive real num-  
 145 ber  $R_{\max}$ . Throughout the paper, we will represent a policy in a matrix form. A greedy policy  
 146 over  $\mathbf{Q} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$ , which is denoted as  $\pi_{\mathbf{Q}} : \mathcal{S} \rightarrow \mathcal{A}$ , i.e.,  $\pi_{\mathbf{Q}}(s) = \arg \max_{\mathbf{a} \in \mathcal{A}} (\mathbf{e}_s \otimes \mathbf{e}_{\mathbf{a}})^\top \mathbf{Q}$ ,  
 147 can be represented as a matrix as follows:

$$148 \quad \mathbf{\Pi}^{\mathbf{Q}} := [\mathbf{e}_1 \otimes \mathbf{e}_{\pi(1)} \quad \mathbf{e}_2 \otimes \mathbf{e}_{\pi(2)} \quad \dots \quad \mathbf{e}_{|\mathcal{S}|} \otimes \mathbf{e}_{\pi(|\mathcal{S}|)}]^\top \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}| \times |\mathcal{A}|},$$

150 where  $\mathbf{e}_s$  and  $\mathbf{e}_{\mathbf{a}}$  represent the canonical basis vector whose  $s$ -th and  $\mathbf{a}$ -th element is  
 151 only one and others are all zero in  $\mathbb{R}^{|\mathcal{S}|}$  and  $\mathbb{R}^{|\mathcal{A}|}$ , respectively, and  $\otimes$  denotes the  
 152 Kronecker product. We can prove that  $\mathbf{P} \mathbf{\Pi}^{\mathbf{Q}}$  for  $\mathbf{Q} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$  represents a transition  
 153 probability of state-action pairs under policy  $\pi$ , i.e.,  $(\mathbf{e}_{s'} \otimes \mathbf{e}_{\mathbf{a}'})^\top (\mathbf{P} \mathbf{\Pi}^{\mathbf{Q}}) (\mathbf{e}_s \otimes \mathbf{e}_{\mathbf{a}}) =$   
 154  $\mathbb{P}[(s_{k+1}, \mathbf{a}_{k+1}) = (s', \mathbf{a}') \mid (s_k, \mathbf{a}_k) = (s, \mathbf{a}), \pi_{\mathbf{Q}}]$  for  $s, s' \in \mathcal{S}$  and  $\mathbf{a}, \mathbf{a}' \in \mathcal{A}$ . Now, we can  
 155 rewrite the Bellman equation in (1) using the matrix notations as follows:  $\mathbf{R}^{\text{avg}} + \gamma \mathbf{P} \mathbf{\Pi}^{\mathbf{Q}^*} \mathbf{Q}^* =$   
 156  $\mathbf{Q}^*$ , where  $\mathbf{R}^{\text{avg}} = \frac{1}{N} \sum_{i=1}^N \mathbf{R}^i \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$  and  $\mathbf{Q}^* \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$  represents optimal Q-function,  
 157  $\mathbf{Q}^*$ , i.e.,  $(\mathbf{e}_s \otimes \mathbf{e}_{\mathbf{a}})^\top \mathbf{Q}^* = Q^*(s, \mathbf{a})$  for  $s, \mathbf{a} \in \mathcal{S} \times \mathcal{A}$ .

## 159 2.2 DISTRIBUTED Q-LEARNING

160 In this section, we discuss a distributed Q-learning algorithm motivated from Nedic and  
 161 Ozdaglar (2009). The non-asymptotic behavior of the algorithm was first investigated

in Heredia et al. (2020); Zeng et al. (2022b) under linear function approximation scheme. Instead, we consider the tabular setup with mild assumptions, and detailed comparisons are given in Section 5. Each agent  $i \in \mathcal{V}$  at time  $k \in \mathbb{N}$  updates its estimate  $\mathbf{Q}_k^i \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$  upon observing  $s_k, \mathbf{a}_k, s'_k \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$  as follows:

$$\begin{aligned} \mathbf{Q}_{k+1}^i(s_k, \mathbf{a}_k) &= \sum_{j \in \mathcal{N}_i} [\mathbf{W}]_{ij} \mathbf{Q}_k^j(s_k, \mathbf{a}_k) + \alpha \left( r_{k+1}^i + \gamma \max_{\mathbf{a} \in \mathcal{A}} \mathbf{Q}_k^i(s'_k, \mathbf{a}) - \mathbf{Q}_k^i(s_k, \mathbf{a}_k) \right) \\ \mathbf{Q}_{k+1}^i(s, \mathbf{a}) &= \sum_{j \in \mathcal{N}_i} [\mathbf{W}]_{ij} \mathbf{Q}_k^j(s, \mathbf{a}), \quad s, \mathbf{a} \in \mathcal{S} \times \mathcal{A} \setminus \{(s_k, \mathbf{a}_k)\}, \end{aligned} \quad (2)$$

where  $\mathbf{Q}_k^i(s, \mathbf{a}) := (\mathbf{e}_s \otimes \mathbf{e}_\mathbf{a})^\top \mathbf{Q}_k^i$  for  $s, \mathbf{a} \in \mathcal{S} \times \mathcal{A}$ ,  $\alpha \in (0, 1)$  is the steps-size, and  $\mathbf{W} \in \mathbb{R}^{N \times N}$  is a non-negative matrix such that agent  $i$  assigns a weight  $[\mathbf{W}]_{ij}$  to its neighbour  $j \in \mathcal{N}_i$ . The agent  $i \in \mathcal{V}$  sends its estimate  $\mathbf{Q}_k^i$  to its neighbour  $j \in \mathcal{N}_i$ , and receives  $\mathbf{Q}_k^j$ , which is weighted by  $[\mathbf{W}]_{ij}$ . The update is different from that of distributed optimization over an objective function in sense that (2) does not use any gradient of a function. Furthermore, note that the memory space of each agent can be expensive due to exponential scaling in the action space, but one can choose linear or neural network approximation (Zhang et al., 2018b; Sunehag et al., 2017) to overcome such issue.

To ensure the consensus among the agents, i.e.,  $\mathbf{Q}_k^i \rightarrow \mathbf{Q}^*$  for all  $i \in [N]$ , where  $[N] := \{1, 2, \dots, N\}$ , a commonly adopted condition on  $\mathbf{W}$  is the so-called doubly stochastic matrix:

**Assumption 2.1.** For all  $i \in [N]$ ,  $[\mathbf{W}]_{ii} > 0$  and  $[\mathbf{W}]_{ij} > 0$  if  $(i, j) \in \mathcal{E}$ , otherwise  $[\mathbf{W}]_{ij} = 0$ . Furthermore,  $\sum_{j=1}^N [\mathbf{W}]_{ij} = \sum_{i=1}^N [\mathbf{W}]_{ji} = 1$ , and  $\mathbf{W}^\top = \mathbf{W}$ .

The assumption is widely adopted in the literature of distributed learning scheme (Heredia et al., 2020; Zeng et al., 2022b). In Appendix B, we provided a simple strategy to construct the doubly stochastic matrix by communicating only with its neighbour.

### 2.3 SWITCHED SYSTEM

In this paper, we consider a system, called the *switched affine system* (Liberzon, 2005),

$$\mathbf{x}_{k+1} = \mathbf{A}_{\sigma_k} \mathbf{x}_k + \mathbf{b}_{\sigma_k}, \quad \mathbf{x}_0 \in \mathbb{R}^n, \quad k \in \mathbb{N}, \quad (3)$$

where  $\mathbf{x}_k \in \mathbb{R}^n$  is the state,  $\mathcal{M} := \{1, 2, \dots, M\}$  is called the set of modes,  $\sigma_k \in \mathcal{M}$  is called the switching signal,  $\{\mathbf{A}_\sigma \in \mathbb{R}^{n \times n} \mid \sigma \in \mathcal{M}\}$  and  $\{\mathbf{b}_\sigma \in \mathbb{R}^n \mid \sigma \in \mathcal{M}\}$  are called the subsystem matrices, and the set of affine terms, respectively. The switching signal can be either arbitrary or controlled by the user under a certain switching policy. If the system in (3) evolves without the affine term, i.e.,  $\mathbf{b}_{\sigma_k} = \mathbf{0}$  for  $k \in \mathbb{N}$ , then it is called the switched linear system. The distributed Q-learning algorithm in (2) will be modeled as a switched affine system motivated from the recent connection of switched system and Q-learning (Lee and He, 2020), which will become clearer in Section 3.4

## 3 ERROR ANALYSIS : I.I.D. OBSERVATION MODEL

In this section, we first consider i.i.d. observation model, which provides simple and clear intuitive results. In the subsequent section, we will extend the result to the Markovian observation model. By an i.i.d. observation model, we refer to a sequence of trajectory  $\{(s_k, \mathbf{a}_k, s'_k)\}_{k \geq 0}$  where each  $(s_k, \mathbf{a}_k, s'_k)$  are an i.i.d. random variables. Suppose that each state-action pair is sampled from a distribution  $d \in \Delta^{|\mathcal{S} \times \mathcal{A}|}$ , i.e.,  $\mathbb{P}[(s_k, \mathbf{a}_k) = (s, \mathbf{a})] = d(s, \mathbf{a})$  and  $s'_k \sim \mathcal{P}(s_k, \mathbf{a}_k, \cdot)$ . The pseudo-code of the algorithm is given in Algorithm 1 in the Appendix J. We will adopt the following standard assumption in the literature:

**Assumption 3.1.** For all  $s, \mathbf{a} \in \mathcal{S} \times \mathcal{A}$ , we have  $d(s, \mathbf{a}) > 0$ .

### 3.1 MATRIX NOTATIONS

Let us introduce the following vector and matrix notations used throughout the paper to re-write (2) in matrix notations:  $\mathbf{D}_s := \text{diag}(d(s, 1), \dots, d(s, |\mathcal{A}|)) \in \mathbb{R}^{|\mathcal{A}| \times |\mathcal{A}|}$ ,  $\mathbf{D} =$

216  $\text{diag}(\mathbf{D}_1, \mathbf{D}_2, \dots, \mathbf{D}_{|\mathcal{S}|}) \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}$ , where  $\text{diag}(\cdot)$  is a diagonal matrix whose diag-  
 217 onal elements correspond to the input vector or matrix, and we will denote  $d_{\max} =$   
 218  $\max_{s, \mathbf{a} \in \mathcal{S} \times \mathcal{A}} d(s, \mathbf{a})$  and  $d_{\min} := \min_{s, \mathbf{a} \in \mathcal{S} \times \mathcal{A}} d(s, \mathbf{a})$ . Furthermore, for  $i \in [N]$ ,  $o =$   
 219  $(s, \mathbf{a}, s') \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$  and  $\mathbf{Q} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ , we define

$$\begin{aligned} 220 \delta^i(o, \mathbf{Q}) &:= (\mathbf{e}_s \otimes \mathbf{e}_a)(r^i(s, \mathbf{a}, s') + \mathbf{e}_{s'}^\top \gamma \mathbf{\Pi}^{\mathbf{Q}} \mathbf{Q} - (\mathbf{e}_s \otimes \mathbf{e}_a)^\top \mathbf{Q}), \\ 221 \Delta^i(\mathbf{Q}) &:= \mathbf{D}(\mathbf{R}^i + \gamma \mathbf{P} \mathbf{\Pi}^{\mathbf{Q}} \mathbf{Q} - \mathbf{Q}), \end{aligned}$$

222 which denotes the TD-error and expected TD-error in vector representation. For simplicity  
 223 of the notation, we denote  $\delta_k^i := \delta^i(o_k, \mathbf{Q}_k^i)$ ,  $\Delta_k^i := \Delta^i(\mathbf{Q}_k^i)$ , and

$$\begin{aligned} 224 \bar{\mathbf{Q}}_k &:= \begin{bmatrix} \mathbf{Q}_k^1 \\ \mathbf{Q}_k^2 \\ \vdots \\ \mathbf{Q}_k^N \end{bmatrix}, \quad \bar{\mathbf{\Pi}}^{\bar{\mathbf{Q}}_k} := \begin{bmatrix} \mathbf{\Pi}^{\mathbf{Q}_k^1} & & & \\ & \ddots & & \\ & & \ddots & \\ & & & \mathbf{\Pi}^{\mathbf{Q}_k^N} \end{bmatrix}, \quad \bar{\boldsymbol{\epsilon}}_k(o_k, \bar{\mathbf{Q}}_k) := \begin{bmatrix} \delta^1(o_k, \mathbf{Q}_k^1) - \Delta^1(\mathbf{Q}_k^1) \\ \delta^2(o_k, \mathbf{Q}_k^2) - \Delta^2(\mathbf{Q}_k^2) \\ \vdots \\ \delta^N(o_k, \mathbf{Q}_k^N) - \Delta^N(\mathbf{Q}_k^N) \end{bmatrix}, \\ 225 \bar{\mathbf{P}} &:= \mathbf{I}_N \otimes \mathbf{P}, \quad \bar{\mathbf{D}} := \mathbf{I}_N \otimes \mathbf{D}, \quad \bar{\mathbf{W}} := \mathbf{W} \otimes \mathbf{I}_{|\mathcal{S}||\mathcal{A}|}, \quad \bar{\mathbf{R}} := [\mathbf{R}^1 \quad \mathbf{R}^2 \quad \dots \quad \mathbf{R}^N]^\top, \end{aligned} \quad (4)$$

226 where  $\mathbf{I}_N$  is a  $N \times N$  identity matrix,  $\mathbf{Q}_k^i$  is defined in (2). Moreover, we denote  $\bar{\boldsymbol{\epsilon}}_k :=$   
 227  $\bar{\boldsymbol{\epsilon}}_k(o_k, \bar{\mathbf{Q}}_k)$ . With the above set of notations, we can re-write the update in (2) as follows:

$$228 \bar{\mathbf{Q}}_{k+1} = \bar{\mathbf{W}} \bar{\mathbf{Q}}_k + \alpha \bar{\mathbf{D}} \left( \bar{\mathbf{R}} + \gamma \bar{\mathbf{P}} \bar{\mathbf{\Pi}}^{\bar{\mathbf{Q}}_k} \bar{\mathbf{Q}}_k - \bar{\mathbf{Q}}_k \right) + \alpha \bar{\boldsymbol{\epsilon}}_k. \quad (5)$$

### 229 3.2 DISTRIBUTED Q-LEARNING : ERROR ANALYSIS

230 In this section, we provide a sketch of the proof to bound the error of distributed Q-learning.  
 231 Let us first decompose the error  $\bar{\mathbf{Q}}_k - \mathbf{1}_N \otimes \mathbf{Q}^*$  into consensus error and optimality error:

$$\begin{aligned} 232 \bar{\mathbf{Q}}_k - \mathbf{1}_N \otimes \mathbf{Q}^* &= \underbrace{\bar{\mathbf{Q}}_k - \mathbf{1}_N \otimes \left( \frac{1}{N} \sum_{i=1}^N \mathbf{Q}_k^i \right)}_{\text{Consensus Error}} + \underbrace{\mathbf{1}_N \otimes \left( \frac{1}{N} \sum_{i=1}^N \mathbf{Q}_k^i - \mathbf{Q}^* \right)}_{\text{Optimality Error}}, \end{aligned} \quad (6)$$

233 where  $\mathbf{1}_N$  is a  $N$ -dimensional vector whose elements are all one. The consensus error measures  
 234 the difference of  $\mathbf{Q}_k^i$  and the overall average,  $\frac{1}{N} \sum_{i=1}^N \mathbf{Q}_k^i$ . As the consensus error vanishes,  
 235 we will have  $\mathbf{Q}_k^1 = \mathbf{Q}_k^2 = \dots = \mathbf{Q}_k^N$ . Meanwhile, the optimality error denotes the difference  
 236 between the true solution  $\mathbf{Q}^*$  and the average,  $\frac{1}{N} \sum_{k=1}^N \mathbf{Q}_k^i$ . Together with the consensus  
 237 error, as optimality error vanishes, we should have  $\mathbf{Q}_k^i - \mathbf{Q}^* \rightarrow 0$  for all  $i \in [N]$ .

### 238 3.3 ANALYSIS OF CONSENSUS ERROR

239 Now, we provide an error bound on the consensus error in (6). We will represent the consensus  
 240 error as  $\boldsymbol{\Theta} \bar{\mathbf{Q}}_k = \bar{\mathbf{Q}}_k - \mathbf{1}_N \otimes \mathbf{Q}_k^{\text{avg}}$  where  $\mathbf{Q}_k^{\text{avg}} := \frac{1}{N} \sum_{i=1}^N \mathbf{Q}_k^i$  and  $\boldsymbol{\Theta} := \mathbf{I}_{N|\mathcal{S}||\mathcal{A}|} - \frac{1}{N} (\mathbf{1}_N \mathbf{1}_N^\top) \otimes$   
 241  $\mathbf{I}_{|\mathcal{S}||\mathcal{A}|}$ . Let us first provide an important lemma that characterizes the convergence of the  
 242 consensus error:

243 **Lemma 3.2.** *For  $k \in \mathbb{N}$ , we have  $\|\bar{\mathbf{W}}^k \boldsymbol{\Theta}\|_2 \leq \sigma_2(\mathbf{W})^k$ , where  $\sigma_2(\mathbf{W})$  is the second largest  
 244 singular value of  $\mathbf{W}$ , and it holds that  $\sigma_2(\mathbf{W}) < 1$ .*

245 The proof is given in Appendix D.1. Moving on, we show that  $\bar{\mathbf{Q}}_k$  will be remain bounded,  
 246 which will be useful throughout the paper:

247 **Lemma 3.3.** *For  $k \in \mathbb{N}$ , and  $\alpha \leq \min_{i \in [N]} [\mathbf{W}]_{ii}$ , we have :  $\|\bar{\mathbf{Q}}_k\|_\infty \leq \frac{R_{\max}}{1-\gamma}$ .*

248 The proof is given in Appendix D.2. The step-size depends on  $\min_{i \in [N]} [\mathbf{W}]_{ii}$ , which can be  
 249 considered as a global information. However, considering the method in Example B.1 in  
 250 Appendix, which requires only local information to construct  $\mathbf{W}$ , we have  $\min_{i \in [N]} [\mathbf{W}]_{ii} \geq \frac{1}{2}$ .  
 251 Therefore, it should be enough to choose  $\alpha \leq \frac{1}{2}$ . Furthermore, the step-size in many

distributed RL algorithms (Zeng et al., 2022b; Wang et al., 2020; Doan et al., 2021; Sun et al., 2020) depend on  $\sigma_2(\mathbf{W})$ , which also can be viewed as a global information. Moreover, we can use an agent-specific step-size, i.e., each agent keeps its own step-size,  $\alpha_i$ . Then, we only require  $\alpha_i < [\mathbf{W}]_{ii}$ , which only uses local information.

Now, we are ready to analyze the behavior of  $\Theta \bar{\mathbf{Q}}_k$ . Multiplying  $\Theta$  to (5), we get

$$\Theta \bar{\mathbf{Q}}_{k+1} = \prod_{i=0}^k \bar{\mathbf{W}}^i \Theta \bar{\mathbf{Q}}_0 + \alpha \sum_{j=0}^k \bar{\mathbf{W}}^{k-j} \Theta \left( \bar{D} \left( \bar{\mathbf{R}} + \gamma \bar{P} \bar{\Pi} \bar{\mathbf{Q}}_j - \bar{\mathbf{Q}}_j \right) + \bar{\epsilon}_j \right). \quad (7)$$

The equality results from recursively expanding the terms. Now, we are ready to bound  $\Theta \bar{\mathbf{Q}}_{k+1}$  using the fact that  $\|\bar{\mathbf{W}}^i \Theta\|_2$  for  $i \in \mathbb{N}$  will decay at a rate of  $\sigma_2(\mathbf{W})$  from Lemma 3.2, and the boundedness of  $\bar{\mathbf{Q}}_k$  in Lemma 3.3.

**Theorem 3.4.** *For  $k \in \mathbb{N}$ , and  $\alpha \leq \min_{i \in [N]} [\mathbf{W}]_{ii}$ , we have the following:*

$$\|\Theta \bar{\mathbf{Q}}_{k+1}\|_\infty \leq \sigma_2(\mathbf{W})^{k+1} \|\Theta \bar{\mathbf{Q}}_0\|_2 + \alpha \frac{8R_{\max}}{1-\gamma} \frac{\sqrt{N|\mathcal{S}||\mathcal{A}|}}{1-\sigma_2(\mathbf{W})}.$$

The proof is given in Appendix D.3. As we can expect, the convergence rate of the consensus error depends on the  $\sigma_2(\mathbf{W})$  with a constant error bound proportional to  $\alpha$ . Furthermore, we note that the above result also holds for the Markovian observation model in Section 4.

### 3.4 ANALYSIS OF OPTIMALITY ERROR

Throughout this section, we analyze the error bound on the optimality error,  $\mathbf{Q}_k^{\text{avg}} - \mathbf{Q}^*$ . Multiplying  $\frac{1}{N}(\mathbf{1}_N \mathbf{1}_N^\top) \otimes \mathbf{I}_{|\mathcal{S}||\mathcal{A}|}$  on (5), we can see that  $\mathbf{Q}_k^{\text{avg}}$  evolves via the following update:

$$\mathbf{Q}_{k+1}^{\text{avg}} = \mathbf{Q}_k^{\text{avg}} + \alpha D \left( \mathbf{R}^{\text{avg}} + \frac{\gamma}{N} \sum_{i=1}^N P \Pi \mathbf{Q}_k^i \mathbf{Q}_k^i - \mathbf{Q}_k^{\text{avg}} \right) + \alpha \epsilon^{\text{avg}}(o_k, \bar{\mathbf{Q}}_k), \quad (8)$$

where  $\epsilon^{\text{avg}}(o, \bar{\mathbf{Q}}) := \frac{1}{N}(\mathbf{1}_N \mathbf{1}_N^\top) \otimes \mathbf{I}_{|\mathcal{S}||\mathcal{A}|} \bar{\epsilon}(o, \bar{\mathbf{Q}})$  for  $o \in \mathcal{S} \times \mathcal{A} \times \mathcal{S}$ ,  $\bar{\mathbf{Q}} \in \mathbb{R}^{N|\mathcal{S}||\mathcal{A}|}$ , and  $\bar{\epsilon}(\cdot)$  is defined in (4). We will denote  $\epsilon_k^{\text{avg}} := \epsilon^{\text{avg}}(o_k, \bar{\mathbf{Q}}_k)$ . The update of (8) resembles that of Q-learning update in the single agent case, i.e.,  $N = 1$ , whose Q-function is  $\mathbf{Q}_k^{\text{avg}}$ . However, the difference with the update of single-agent case lies in the fact that we take average of the maximum of Q-function of each agent, i.e., the term  $\frac{1}{N} \sum_{i=1}^N \Pi \mathbf{Q}_k^i \mathbf{Q}_k^i$  in (8), rather than the maximum of average of Q-function of each agents, i.e.,  $\Pi \mathbf{Q}_k^{\text{avg}} \mathbf{Q}_k^{\text{avg}}$ . This poses difficulty in the analysis since  $\frac{1}{N} \sum_{i=1}^N \Pi \mathbf{Q}_k^i \mathbf{Q}_k^i$  cannot be represented in terms of  $\mathbf{Q}_k^{\text{avg}}$ . Consequently, it makes difficult to interpret it as switched affine system whose state-variable is  $\mathbf{Q}_k^{\text{avg}}$ , which is introduced in Section 2.3. To handle this issue, motivated from the approach in Kar et al. (2013), we introduce an additional error term  $\frac{1}{N} \sum_{i=1}^N \Pi \mathbf{Q}_k^i \mathbf{Q}_k^i - \Pi \mathbf{Q}_k^{\text{avg}} \mathbf{Q}_k^{\text{avg}}$ , which can be bounded by the consensus error discussed in Section 3.3. Therefore, we re-write (8) as:

$$\begin{aligned} \mathbf{Q}_{k+1}^{\text{avg}} = & \mathbf{Q}_k^{\text{avg}} + \alpha D \left( \mathbf{R}^{\text{avg}} + \gamma P \Pi \mathbf{Q}_k^{\text{avg}} \mathbf{Q}_k^{\text{avg}} - \mathbf{Q}_k^{\text{avg}} \right) + \alpha \epsilon_k^{\text{avg}} \\ & + \alpha \underbrace{\left( \frac{\gamma}{N} \sum_{i=1}^N D \left( P \Pi \mathbf{Q}_k^i \mathbf{Q}_k^i - \gamma P \Pi \mathbf{Q}_k^{\text{avg}} \mathbf{Q}_k^{\text{avg}} \right) \right)}_{:= \mathbf{E}_k}. \end{aligned} \quad (9)$$

Now, we can see that  $\mathbf{Q}_k^{\text{avg}}$  evolves via a single-agent Q-learning update whose estimator is  $\mathbf{Q}_k^{\text{avg}}$ , including an additional stochastic noise term,  $\epsilon_k^{\text{avg}}$ , and an error term,  $\mathbf{E}_k$  that can be bounded by the consensus error. In the following lemma, we use the contraction property of the max-operator to bound  $\mathbf{E}_k$  by the consensus error:

**Lemma 3.5.** *For  $k \in \mathbb{N}$ , we have  $\|\mathbf{E}_k\|_\infty \leq \gamma d_{\max} \|\Theta \bar{\mathbf{Q}}_k\|_\infty$ .*

The proof is given in Appendix D.4. We note that similar argument in Lemma 3.5 has been also considered in Kar et al. (2013). However, Kar et al. (2013) considered a different

distributed algorithm using two-time scale approach and focused on asymptotic convergence whereas we consider a single step-size and finite-time bounds.

Now, we follow the switched system approach (Lee and He, 2020) to bound the optimality error. In contrast to Lee and He (2020), we have an additional error term caused by  $\mathbf{E}_k$ , which will be bounded using Theorem 3.4. Using a coordinate transformation,  $\tilde{\mathbf{Q}}_k^{\text{avg}} = \mathbf{Q}_k^{\text{avg}} - \mathbf{Q}^*$ , we can re-write (9) as

$$\tilde{\mathbf{Q}}_{k+1}^{\text{avg}} = \mathbf{A}_{\mathbf{Q}^{\text{avg}}} \tilde{\mathbf{Q}}_k^{\text{avg}} + \alpha \mathbf{b}_{\mathbf{Q}^{\text{avg}}} + \alpha \boldsymbol{\epsilon}_k^{\text{avg}} + \alpha \mathbf{E}_k,$$

where, for  $\mathbf{Q} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ , we let

$$\mathbf{A}_{\mathbf{Q}} := \mathbf{I} + \alpha \mathbf{D}(\gamma \mathbf{P}\mathbf{\Pi}^{\mathbf{Q}} - \mathbf{I}) \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}, \quad \mathbf{b}_{\mathbf{Q}} := \gamma \mathbf{D}\mathbf{P}(\mathbf{\Pi}^{\mathbf{Q}} - \mathbf{\Pi}^{\mathbf{Q}^*})\mathbf{Q}^*. \quad (10)$$

We can see that  $\boldsymbol{\epsilon}_k^{\text{avg}}$  is a stochastic term, and we will bound the error caused by this term using concentration inequalities. The consensus error,  $\mathbf{E}_k$ , can be bounded from Theorem 3.4. However, the affine term,  $\mathbf{b}_{\mathbf{Q}^{\text{avg}}}$ , does not admit simple bounds. The approach in Lee and He (2020) provides a method to construct a system without an affine term, making the analysis simpler. In details, we introduce a lower and upper comparison system, denoted as  $\mathbf{Q}_k^{\text{avg},l}$  and  $\mathbf{Q}_k^{\text{avg},u}$ , respectively such that the following element-wise inequality holds:

$$\mathbf{Q}_k^{\text{avg},l} \leq \mathbf{Q}_k^{\text{avg}} \leq \mathbf{Q}_k^{\text{avg},u}, \quad \forall k \in \mathbb{N}, \quad (11)$$

Letting  $\tilde{\mathbf{Q}}_k^{\text{avg},l} := \mathbf{Q}_k^{\text{avg},l} - \mathbf{Q}^*$  and  $\tilde{\mathbf{Q}}_k^{\text{avg},u} := \mathbf{Q}_k^{\text{avg},u} - \mathbf{Q}^*$ , a candidate of update that satisfies (11), which is without the affine term  $\mathbf{b}_{\mathbf{Q}^{\text{avg}}}$ , is:

$$\tilde{\mathbf{Q}}_{k+1}^{\text{avg},l} = \mathbf{A}_{\mathbf{Q}^*} \tilde{\mathbf{Q}}_k^{\text{avg},l} + \alpha \boldsymbol{\epsilon}_k^{\text{avg}} + \alpha \mathbf{E}_k, \quad \tilde{\mathbf{Q}}_{k+1}^{\text{avg},u} = \mathbf{A}_{\mathbf{Q}^{\text{avg},u}} \tilde{\mathbf{Q}}_k^{\text{avg},u} + \alpha \boldsymbol{\epsilon}_k^{\text{avg}} + \alpha \mathbf{E}_k, \quad (12)$$

where  $\mathbf{Q}_0^{\text{avg},l} \leq \mathbf{Q}_0^{\text{avg}} \leq \mathbf{Q}_0^{\text{avg},u}$ . The detailed construction of each systems are given in Appendix E. Note that the lower comparison system,  $\tilde{\mathbf{Q}}_k^{\text{avg},l}$  follows a linear system governed by the matrix  $\mathbf{A}_{\mathbf{Q}^*}$  where as the upper comparison system,  $\tilde{\mathbf{Q}}_k^{\text{avg},u}$ , can be viewed as a switched linear system without an affine term. To prove the finite-time bound of  $\tilde{\mathbf{Q}}_k^{\text{avg}}$ , we will instead derive the finite-time bound of  $\tilde{\mathbf{Q}}_k^{\text{avg},l}$  and  $\tilde{\mathbf{Q}}_k^{\text{avg},u}$ , and using the relation in (11), we can obtain the desired result. Nonetheless, still the switching in the upper comparison system imposes difficulty in the analysis. Therefore, we consider the difference of upper and lower comparison system  $\tilde{\mathbf{Q}}_k^{\text{avg},l} - \tilde{\mathbf{Q}}_k^{\text{avg},u}$ , which gives the following bound:  $\|\tilde{\mathbf{Q}}_k^{\text{avg}}\|_{\infty} \leq \|\tilde{\mathbf{Q}}_k^{\text{avg},l}\|_{\infty} + \|\mathbf{Q}_{k+1}^{\text{avg},u} - \mathbf{Q}_{k+1}^{\text{avg},l}\|_{\infty}$ . The sketch of the proof for deriving the finite-time bound of each systems are as follows:

1. Bounding  $\tilde{\mathbf{Q}}_k^{\text{avg},l}$  (Proposition F.1 in the Appendix): We recursively expand the equation in (12). We have  $\|\mathbf{A}_{\mathbf{Q}^*}\|_{\infty} \leq 1 - (1 - \gamma)\alpha d_{\min}$  for any  $\mathbf{Q} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ , which is in Lemma C.1 in the Appendix, and the error induced by  $\boldsymbol{\epsilon}_k^{\text{avg}}$  can be bounded using Azuma-Hoeffding inequality in Lemma C.4 in the Appendix. Meanwhile, the error term  $\mathbf{E}_k$  can be bounded by the consensus error from Lemma 3.5, which is again bounded by using Theorem 3.4.
2. Bounding  $\tilde{\mathbf{Q}}_k^{\text{avg},u} - \tilde{\mathbf{Q}}_k^{\text{avg},l}$  (Proposition F.3 in the Appendix): Thanks to the fact that both the upper and lower comparison systems share  $\boldsymbol{\epsilon}_k^{\text{avg}}$  and  $\mathbf{E}_k$ , if we subtract  $\tilde{\mathbf{Q}}_k^{\text{avg},l}$  from  $\tilde{\mathbf{Q}}_k^{\text{avg},u}$  in (12), both terms are eliminated. Therefore, the iterate can be bounded with an additional error by  $\tilde{\mathbf{Q}}_k^{\text{avg},l}$ .

Now, we are ready to present the optimality error bound,  $\|\mathbf{Q}_k^{\text{avg}} - \mathbf{Q}^*\|_{\infty}$ , as follows:

**Theorem 3.6.** For  $k \in \mathbb{N}$ , and  $\alpha \leq \min_{i \in [N]} [\mathbf{W}]_{ii}$ , we have the following result :

$$\begin{aligned} \mathbb{E} [\|\mathbf{Q}_k^{\text{avg}} - \mathbf{Q}^*\|_{\infty}] &= \tilde{\mathcal{O}} \left( (1 - \alpha(1 - \gamma)d_{\min})^{\frac{k}{2}} + \sigma_2(\mathbf{W})^{\frac{k}{4}} \right) \\ &\quad + \tilde{\mathcal{O}} \left( \alpha^{\frac{1}{2}} \frac{d_{\max} R_{\max}}{(1 - \gamma)^{\frac{5}{2}} d_{\min}^{\frac{3}{2}}} + \alpha \frac{d_{\max}^2 \sqrt{|\mathcal{S}||\mathcal{A}|} R_{\max}}{(1 - \gamma)^3 d_{\min}^2 (1 - \sigma_2(\mathbf{W}))} \right), \end{aligned}$$

where the notation  $\tilde{\mathcal{O}}(\cdot)$  is used to hide the logarithmic factors.

The proof is given in Appendix F.1. Note that even the logarithmic terms are hidden, due to exponential scaling of the action space,  $\ln(|\mathcal{S}||\mathcal{A}|)$  could contribute  $\mathcal{O}(N)$  factor to the error bound. However, noting that  $d_{\min} \leq \frac{1}{|\mathcal{S}||\mathcal{A}|}$ ,  $\mathcal{O}\left(\frac{1}{d_{\min}}\right)$  already dominates the  $\mathcal{O}(N)$  if  $|\mathcal{A}_i| \geq 2$  for all  $i \in [N]$ , hence we omit the logarithmic terms. Likewise  $\mathcal{O}(|\mathcal{A}|)$  dominates  $\mathcal{O}(N)$ , which is hidden when both terms are multiplied.

### 3.5 FINAL ERROR

In this section, we present the error bound of the total error term  $\bar{\mathbf{Q}}_k - \mathbf{1}_N \otimes \mathbf{Q}^*$ . From (6), the bound follows from the decomposition into the consensus error and optimality error. In particular, collecting the results in Theorem 3.4 and Theorem 3.6 yields the following:

**Theorem 3.7.** *For  $k \in \mathbb{N}$ , and  $\alpha \leq \min_{i \in [N]} [\mathbf{W}]_{ii}$ , we have*

$$\begin{aligned} \mathbb{E} [\|\bar{\mathbf{Q}}_k - \mathbf{1}_N \otimes \mathbf{Q}^*\|_{\infty}] &= \tilde{\mathcal{O}} \left( (1 - \alpha(1 - \gamma)d_{\min})^{\frac{k}{2}} + \sigma_2(\mathbf{W})^{\frac{k}{4}} \right) \\ &\quad + \tilde{\mathcal{O}} \left( \alpha^{\frac{1}{2}} d_{\max} \frac{R_{\max}}{(1 - \gamma)^{\frac{5}{2}} d_{\min}^{\frac{3}{2}}} + \alpha \frac{d_{\max}^2 \sqrt{|\mathcal{S}||\mathcal{A}|} R_{\max}}{(1 - \gamma)^3 d_{\min}^2 (1 - \sigma_2(\mathbf{W}))} \right). \end{aligned}$$

The proof is given in Appendix F.2. One can see that the convergence rate has exponentially decaying terms,  $(1 - (1 - \gamma)d_{\min}\alpha)^{\frac{k}{2}}$  and  $\sigma_2(\mathbf{W})^{\frac{k}{4}}$ , with a bias term caused by using a constant step-size. Furthermore, we note that the bias term depends on  $\frac{1}{1 - \sigma_2(\mathbf{W})}$ . If we construct  $\mathbf{W}$  as in Example B.1 in the Appendix, then it will contribute  $\mathcal{O}(N^2)$  factor in the error bound (Olshevsky, 2014).

**Corollary 3.8.** *Suppose  $\alpha = \tilde{\mathcal{O}} \left( \min \left\{ \frac{(1 - \gamma)^5 d_{\min}^3}{R_{\max}^2 d_{\max}^2} \epsilon^2, \frac{(1 - \gamma)^3 d_{\min}^2 (1 - \sigma_2(\mathbf{W}))}{R_{\max} d_{\max}^2 \sqrt{|\mathcal{S}||\mathcal{A}|}} \epsilon \right\} \right)$ . Then, the following number of samples are required for  $\mathbb{E} [\|\bar{\mathbf{Q}}_k - \mathbf{1}_N \otimes \mathbf{Q}^*\|_{\infty}] \leq \epsilon$ :*

$$\tilde{\mathcal{O}} \left( \max \left\{ \frac{1}{\epsilon^2} \frac{d_{\max}^2}{(1 - \gamma)^6 d_{\min}^4}, \frac{1}{\epsilon} \frac{d_{\max}^2 \sqrt{|\mathcal{S}||\mathcal{A}|}}{(1 - \gamma)^4 d_{\min}^3 (1 - \sigma_2(\mathbf{W}))} \right\} \right).$$

The proof is given in Appendix Section F.3. As the known sample complexity of (single-agent) Q-learning, our bound depends on the factors,  $d_{\min}$  and  $\frac{1}{1 - \gamma}$ . The result is improvable in sense that the known tight dependency for single-agent case is  $\frac{1}{(1 - \gamma)^4 d_{\min}}$  by Li et al. (2020). Furthermore, we note that the dependency on the spectral property of the graph,  $\frac{1}{\epsilon} \frac{1}{1 - \sigma_2(\mathbf{W})}$  is common in the literature of distributed learning as can be seen in Table 1.

## 4 ERROR ANALYSIS : MARKOVIAN OBSERVATION MODEL

Now, we consider a Markovian observation model instead of the i.i.d. model. Starting from an initial distribution  $\boldsymbol{\mu}_0 \in \Delta^{|\mathcal{S}||\mathcal{A}|}$ , the samples are observed from a behavior policy  $\beta : \mathcal{S} \rightarrow \Delta^{|\mathcal{A}|}$ , i.e., from  $(s_k, \mathbf{a}_k)$ , transition occurs to  $s_{k+1} \sim \mathcal{P}(s_k, \mathbf{a}_k, \cdot)$  and the action is selected by  $\mathbf{a}_{k+1} \sim \beta(\cdot | s_{k+1})$ . This setting is closer to practical scenarios, but poses significant challenges in the analysis due to the dependence between the past observations and current estimates. To overcome this difficulty, we consider the so-called uniformly ergodic Markov chain (Paulin, 2015), which ensures that the Markov chain converges to its unique stationary distribution,  $\boldsymbol{\mu}_{\infty} \in \Delta^{|\mathcal{S}||\mathcal{A}|}$ , exponentially fast in sense of total variation distance, which is defined as  $d_{\text{TV}}(\mathbf{p}, \mathbf{q}) := \frac{1}{2} \sum_{x \in \mathcal{S} \times \mathcal{A}} |[\mathbf{p}]_x - [\mathbf{q}]_x|$  where  $\mathbf{p}, \mathbf{q} \in \Delta^{|\mathcal{S}||\mathcal{A}|}$ . That is, there exist positive real numbers  $m \in \mathbb{R}$  and  $\rho \in (0, 1)$  such that we have  $\max_{s, \mathbf{a} \in \mathcal{S} \times \mathcal{A}} d_{\text{TV}}(\boldsymbol{\mu}_k^{s, \mathbf{a}}, \boldsymbol{\mu}_{\infty}) \leq m\rho^k$ , where  $\boldsymbol{\mu}_k^{s, \mathbf{a}} := ((\mathbf{e}_s \otimes \mathbf{e}_{\mathbf{a}})^{\top} \mathbf{P}_{\beta}^k)^{\top}$  is the probability distribution of state-action pair after  $k$  number of transition occurs starting from  $s, \mathbf{a} \in \mathcal{S} \times \mathcal{A}$ , and  $\mathbf{P}_{\beta} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}$  is the transition matrix induced by behavior policy  $\beta$ , i.e.,  $(\mathbf{e}_s \otimes \mathbf{e}_{\mathbf{a}})^{\top} \mathbf{P}_{\beta} (\mathbf{e}_{s'} \otimes \mathbf{e}_{\mathbf{a}'})^{\top} = (\mathbf{e}_s \otimes \mathbf{e}_{\mathbf{a}})^{\top} \mathbf{P} \mathbf{e}_{s'} \cdot \beta(\mathbf{a}' | s')$ . Moreover, we will denote

$$\tau^{\text{mix}}(\epsilon) := \min\{t \in \mathbb{N} : m\rho^t \leq \epsilon\}, \quad \tau := \tau^{\text{mix}}(\alpha), \quad t_{\text{mix}} := \tau^{\text{mix}}(1/4), \quad (13)$$

for  $\epsilon > 0$ , and  $\tau$  is the so-called mixing time. The concept of mixing time is widely used in the literature (Zeng et al., 2022b; Bhandari et al., 2018). Note that  $\tau$  is approximately proportional to  $\log(\frac{1}{\alpha})$ , which is provided in Lemma C.7 in the Appendix. This contributes only logarithmic factor to the final error bound. Furthermore, we will denote

$$\mathbf{D}_\infty = \text{diag}(\boldsymbol{\mu}_\infty), \quad \mathbf{D}_k^{s,\mathbf{a}} = \text{diag}(\boldsymbol{\mu}_k^{s,\mathbf{a}}), \quad (14)$$

where  $\mathbf{D}_k^{s,\mathbf{a}}$  denotes the probability distribution of the state-action pair after  $k$  number of transitions from  $s, \mathbf{a} \in \mathcal{S} \times \mathcal{A}$ .  $\bar{\boldsymbol{\epsilon}}_k$  in (5) will be defined in terms of  $\mathbf{D}_\infty$  instead of  $\mathbf{D}$ , and the overall details are provided in Appendix G. To proceed, with slight abuse of notation, we will denote  $d_{\max} = \max_{s,\mathbf{a} \in \mathcal{S} \times \mathcal{A}} [\boldsymbol{\mu}_\infty]_{s,\mathbf{a}}$  and  $d_{\min} = \min_{s,\mathbf{a} \in \mathcal{S} \times \mathcal{A}} [\boldsymbol{\mu}_\infty]_{s,\mathbf{a}}$ .

Now, we provide the technical difference with the proof of i.i.d. case in Section 3. The challenge in the analysis lies in the fact that  $\mathbb{E}[\boldsymbol{\epsilon}_k^{\text{avg}} | \{(s_t, \mathbf{a}_t)\}_{t=0}^k, \mathbf{Q}_0] \neq \mathbf{0}$  due to Markovian observation scheme. Therefore, we cannot use Azuma-Hoeffding inequality as in the proof of i.i.d. case in the Appendix F.1. Instead, we consider the shifted sequence as in Qu and Wierman (2020). By shifted sequence, it means to consider the error by the stochastic observation at  $k$  with  $\bar{\mathbf{Q}}_{k-\tau}$  instead of  $\bar{\mathbf{Q}}_k$ , i.e.,  $\mathbf{w}_{k,1} := \delta^{\text{avg}}(o_k, \bar{\mathbf{Q}}_{k-\tau}) - \Delta_{k-\tau,k}^{\text{avg}}(\bar{\mathbf{Q}}_{k-\tau})$  where  $\Delta_{k-\tau,k}^{\text{avg}}(\bar{\mathbf{Q}}_{k-\tau}) := \mathbf{D}_\tau^{s_{k-\tau}, \mathbf{a}_{k-\tau}} \frac{1}{N} \sum_{i=1}^N (\mathbf{R}^i + \gamma \mathbf{P} \Pi \mathbf{Q}_k^i \mathbf{Q}_k^i - \mathbf{Q}_k^i)$ . Then, we have  $\mathbb{E}[\mathbf{w}_{k,1} | \{(s_t, \mathbf{a}_t)\}_{t=0}^{k-\tau}, \bar{\mathbf{Q}}_0] = \mathbf{0}$ . Now, we separately calculate the errors induced by  $\{\mathbf{w}_{\tau j+l,1}\}_{j \in \{t \in \mathbb{N} | \tau t + l \leq k\}}$  for each  $0 \leq l \leq \tau - 1$ , and invoke the Azuma-Hoeffding inequality. Overall details are given in Appendix G, and we have the following result:

**Theorem 4.1.** For  $k \geq \tau$ , and  $\alpha \leq \min\{\min_{i \in [N]} [\mathbf{W}]_{ii}, \frac{1}{2\tau}\}$ , we have

$$\begin{aligned} \mathbb{E}[\|\mathbf{Q}_{k+1} - \mathbf{Q}^*\|_\infty] &= \tilde{\mathcal{O}}\left((1 - \alpha(1 - \gamma)d_{\min})^{\frac{k-\tau}{2}} + \sigma_2(\mathbf{W})^{\frac{k-\tau}{4}}\right) \\ &\quad + \tilde{\mathcal{O}}\left(\alpha^{\frac{1}{2}} \frac{d_{\max} \sqrt{\tau} R_{\max}}{(1 - \gamma)^{\frac{5}{2}} d_{\min}^{\frac{3}{2}}} + \alpha \frac{R_{\max} d_{\max} \sqrt{|\mathcal{S}||\mathcal{A}|}}{(1 - \gamma)^3 d_{\min}^2 (1 - \sigma_2(\mathbf{W}))}\right). \end{aligned}$$

The proof is given in Appendix Section G.2.

**Corollary 4.2.** Suppose  $\alpha = \tilde{\mathcal{O}}\left(\frac{\epsilon^2}{\ln(\frac{1}{\epsilon^2})} \frac{(1-\gamma)^5 d_{\min}^3}{t_{\text{mix}} d_{\max}^3}\right)$ . Then, the following number of samples are required for  $\mathbb{E}[\|\bar{\mathbf{Q}}_k - \mathbf{1}_N \otimes \mathbf{Q}^*\|_\infty] \leq \epsilon$ :

$$\tilde{\mathcal{O}}\left(\max\left\{\frac{\ln^2(\frac{1}{\epsilon^2})}{\epsilon^2} \frac{t_{\text{mix}} d_{\max}^2}{(1-\gamma)^6 d_{\min}^4}, \frac{\ln(\frac{1}{\epsilon})}{\epsilon} \frac{d_{\max} \sqrt{|\mathcal{S}||\mathcal{A}|}}{(1-\gamma)^4 d_{\min}^3 (1-\sigma_2(\mathbf{W}))}\right\}\right).$$

The proof is given in Appendix Section G.3. As in the result of i.i.d. case in Corollary 3.8, we have the dependency on  $\frac{1}{1-\gamma}$ ,  $\frac{1}{d_{\min}}$ , and  $\frac{1}{1-\sigma_2(\mathbf{W})}$  with additional factor on mixing time. **The known tight sample complexity result in the single-agent case is  $\tilde{\mathcal{O}}\left(\frac{1}{(1-\gamma)^4 d_{\min} \epsilon^2} + \frac{t_{\text{mix}}}{(1-\gamma) d_{\min}}\right)$  by Li et al. (2024), and our result leaves room for improvement.** Assuming a uniform sampling scheme, i.e.,  $d_{\min} = d_{\max} = \frac{1}{|\mathcal{S}||\mathcal{A}|}$ , and  $|\mathcal{A}_i| = A$  for all  $i \in [N]$  and  $A \geq 2$ , the sample complexity becomes  $\tilde{\mathcal{O}}\left(\max\left\{\frac{t_{\text{mix}}}{\epsilon^2} \frac{|\mathcal{S}|^2 A^2 N}{(1-\gamma)^6}, \frac{1}{\epsilon} \frac{|\mathcal{S}|^{\frac{5}{2}} A^{\frac{5N}{2}}}{(1-\gamma)^4 (1-\sigma_2(\mathbf{W}))}\right\}\right)$ . We note that the exponential scaling in the action space is inevitable in the tabular setting unless we consider a near-optimal solution (Qu et al., 2022). Lastly, to verify the convergence of our algorithm, experiments are provided in Appendix Section I.

## 5 DISCUSSION

	Q-function	Assumption	Sample complexity	Bound type	Remarks
481	Ours	$\boldsymbol{\chi}$	$\max\left\{\frac{t_{\text{mix}}}{\epsilon^2} \frac{1}{(1-\gamma)^6 d_{\min}^3}, \frac{1}{\epsilon} \frac{\sqrt{ \mathcal{S}  \mathcal{A} }}{(1-\sigma_2(\mathbf{W})) (1-\gamma)^4 d_{\min}^2}\right\}$	Expectation	-
482	Wang et al. (2022)	$\boldsymbol{\chi}$	$\frac{1}{(1-\gamma)^2 d_{\min} \epsilon^2} + \frac{t_{\text{mix}}}{1-\gamma}$	High probability	$\epsilon \in [0, \frac{1}{1-\gamma})$
483	Heredia et al. (2020)	(15)	$\frac{R^2}{(d_{\min} - \gamma^2 d_{\max})^2 (1-\sigma_2(\mathbf{W}))}$	Expectation	Continuous state space
484	Zeng et al. (2022b)	(16)	$\frac{1}{\kappa^2 (1-\gamma)^2 (1-\sigma_2(\mathbf{W}))}$	Averaged squared error	$R$ is projection radius
				Expectation	-

Table 1: **LFA stands for linear function approximation.**

In this section, we provide comparison with recent works analyzing non-asymptotic behavior of distributed Q-learning algorithm. Our analysis relies on the minimal assumption in sense that we do not require any assumption further than standard assumptions in the literature, e.g., the state-action distribution induced by the behavior policy, is positive for all state-action pairs in Assumption 3.1.

Heredia et al. (2020) considered linear function approximation scheme to represent the Q-function with continuous state-space and finite-action space scenario. However, to prove the convergence, it requires the following condition:

$$d_{\min} > \gamma^2 d_{\max}^* := \max_s d(s, \pi^*(s)), \quad (15)$$

which is difficult to be met even in the tabular case, and an example is given in Appendix H.

Furthermore, Zeng et al. (2022b) considered a Q-learning model under linear function approximation with continuous-state space and finite action space. The work also covered the case when the features for linear function approximation is differently selected for each agents. However, it requires the following condition to hold for all  $\mathbf{Q} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ :

$$(\gamma DP(\Pi^{\mathbf{Q}} \mathbf{Q} - \Pi^{\mathbf{Q}^*} \mathbf{Q}^*) - D(\mathbf{Q} - \mathbf{Q}^*))^\top (\mathbf{Q} - \mathbf{Q}^*) \leq -\kappa \|\mathbf{Q} - \mathbf{Q}^*\|_2^2, \quad (16)$$

for some  $\kappa > 0$ . We have provided examples where the above conditions in (15) and (16) are not met even in the tabular case in Appendix Section H.

Overall, the assumptions used in Heredia et al. (2020); Zeng et al. (2022b) allows the analysis to follow similar lines to that of convex optimization literature. To the best of our knowledge, there is no existing literature that demonstrates how to extend convex optimization analysis, or an analogous approach, to the analysis of Q-learning under the tabular setup. This gap in the literature makes the analysis challenging and is the primary reason we rely on switched system analysis. Due to different settings, their sample complexity is not directly comparable with ours.

Wang et al. (2022) proposed a distributed Q-learning algorithm in the tabular setting, which is motivated from the adapt-then-combine algorithm, whereas our algorithm considers combine-and-adapt scheme (Chen and Sayed, 2012) in the distributed optimization literature. The work presents a sharper bound on the sample complexity  $\frac{1}{(1-\gamma)^5 d_{\min} \epsilon^2}$  compared to ours  $\frac{1}{(1-\gamma)^6 d_{\min}^4 \epsilon^2}$  but it only holds for restricted range of  $\epsilon$ , i.e.,  $\epsilon \in \left[0, \frac{1}{1-\gamma}\right)$  while our results do not have such restriction. More importantly, the algorithm proposed by Wang et al. (2022) requires two steps for a single update, whereas in our paper, we focus on a one-step algorithm that is algorithmically simpler and more efficient. Specifically, we analyze the traditional and widely adopted QD-learning algorithm proposed in Kar et al. (2013), for which a finite-time error analysis for the original form has been lacking in the literature. Additionally, we enhance the efficiency of QD-learning by employing a constant step-size, as opposed to the two-time-scale decaying step-size used in traditional QD-learning. This modification can significantly improve the convergence speed empirically.

## 6 CONCLUSION

In this paper, we have studied distributed version of Q-learning algorithm. We provided a sample complexity result of  $\tilde{\mathcal{O}} \left( \max \left\{ \frac{1}{\epsilon^2} \frac{1}{(1-\gamma)^6 d_{\min}^4}, \frac{1}{\epsilon} \frac{\sqrt{|\mathcal{S}||\mathcal{A}|}}{(1-\sigma_2(\mathbf{W}))(1-\gamma)^4 d_{\min}^3} \right\} \right)$ , which appears to be the first non-asymptotic result for tabular Q-learning. Future work would include improving the dependency on  $\frac{1}{1-\gamma}$  and  $d_{\min}$  to match the known tightest sample complexity bound of single-agent Q-learning (Li et al., 2020). Furthermore, to resolve the scalability issue, two promising approaches would be adopting a mean-field approach or exploring convergence to sub-optimal point.

## REFERENCES

R. Bellman. Dynamic programming. *science*, 153(3731):34–37, 1966.

- 540 J. Bhandari, D. Russo, and R. Singal. A finite time analysis of temporal difference learning  
541 with linear function approximation. In *Conference on learning theory*, pages 1691–1692.  
542 PMLR, 2018.
- 543 J. Blumenkamp, S. Morad, J. Gielis, Q. Li, and A. Prorok. A framework for real-world multi-  
544 robot systems running decentralized gnn-based policies. In *2022 International Conference*  
545 *on Robotics and Automation (ICRA)*, pages 8772–8778. IEEE, 2022.
- 547 J. Chen and A. H. Sayed. Diffusion adaptation strategies for distributed optimization and  
548 learning over networks. *IEEE Transactions on Signal Processing*, 60(8):4289–4305, 2012.
- 549 Z. Chen, S. T. Maguluri, S. Shakkottai, and K. Shanmugam. A lyapunov theory for finite-  
550 sample guarantees of asynchronous q-learning and td-learning variants. *arXiv preprint*  
551 *arXiv:2102.01567*, 2021.
- 553 F. R. Chung and L. Lu. *Complex graphs and networks*. American Mathematical Soc., 2006.
- 554 C. Claus and C. Boutilier. The dynamics of reinforcement learning in cooperative multiagent  
555 systems. *AAAI/IAAI*, 1998(746-752):2, 1998.
- 557 C. S. de Witt, B. Peng, P.-A. Kamienny, P. Torr, W. Böhmer, and S. Whiteson. Deep  
558 multi-agent reinforcement learning for decentralized continuous cooperative control. *arXiv*  
559 *preprint arXiv:2003.06709*, 19, 2020.
- 560 T. Doan, S. Maguluri, and J. Romberg. Finite-time analysis of distributed td (0) with linear  
561 function approximation on multi-agent reinforcement learning. In *International Conference*  
562 *on Machine Learning*, pages 1626–1635. PMLR, 2019.
- 564 T. T. Doan, S. T. Maguluri, and J. Romberg. Finite-time performance of distributed temporal-  
565 difference learning with linear function approximation. *SIAM Journal on Mathematics of*  
566 *Data Science*, 3(1):298–320, 2021.
- 567 Z. Dou, J. G. Kuba, and Y. Yang. Understanding value decomposition algorithms in deep  
568 cooperative multi-agent reinforcement learning. *arXiv preprint arXiv:2202.04868*, 2022.
- 569 R. Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.
- 570 E. Even-Dar, Y. Mansour, and P. Bartlett. Learning rates for q-learning. *Journal of machine*  
571 *learning Research*, 5(1), 2003.
- 572 H. Gu, X. Guo, X. Wei, and R. Xu. Mean-field multiagent reinforcement learning: A  
573 decentralized network approach. *Mathematics of Operations Research*, 2024.
- 574 P. Heredia, H. Ghadialy, and S. Mou. Finite-sample analysis of distributed q-learning for  
575 multi-agent networks. In *2020 American Control Conference (ACC)*, pages 3511–3516.  
576 IEEE, 2020.
- 577 S. Kar, J. M. Moura, and H. V. Poor.  $QD$ -learning: A collaborative distributed strategy for  
578 multi-agent reinforcement learning through Consensus + Innovations. *IEEE Transactions*  
579 *on Signal Processing*, 61(7):1848–1862, 2013.
- 580 S. Khodadadian, P. Sharma, G. Joshi, and S. T. Maguluri. Federated reinforcement learning:  
581 Linear speedup under markovian sampling. In *International Conference on Machine*  
582 *Learning*, pages 10997–11057. PMLR, 2022.
- 583 P. A. Knight. The sinkhorn–knopp algorithm: convergence and applications. *SIAM Journal*  
584 *on Matrix Analysis and Applications*, 30(1):261–275, 2008.
- 585 V. Konda and J. Tsitsiklis. Actor-critic algorithms. *Advances in neural information processing*  
586 *systems*, 12, 1999.
- 587 D. Lee and N. He. A unified switching system perspective and convergence analysis of q-  
588 learning algorithms. *Advances in Neural Information Processing Systems*, 33:15556–15567,  
589 2020.

- 594 D. Lee, H. Yoon, and N. Hovakimyan. Primal-dual algorithm for distributed reinforcement  
595 learning: Distributed gtd. In *2018 IEEE Conference on Decision and Control (CDC)*,  
596 pages 1967–1972. IEEE, 2018.
- 597 D. Lee, J. Hu, and N. He. A discrete-time switching system analysis of q-learning. *SIAM*  
598 *Journal on Control and Optimization*, 61(3):1861–1880, 2023.
- 600 G. Li, Y. Wei, Y. Chi, Y. Gu, and Y. Chen. Sample complexity of asynchronous q-learning:  
601 Sharper analysis and variance reduction. *Advances in neural information processing*  
602 *systems*, 33:7031–7043, 2020.
- 603 G. Li, C. Cai, Y. Chen, Y. Wei, and Y. Chi. Is q-learning minimax optimal? a tight sample  
604 complexity analysis. *Operations Research*, 72(1):222–236, 2024.
- 605 D. Liberzon. Switched systems. In *Handbook of networked and embedded control systems*,  
606 pages 559–574. Springer, 2005.
- 608 H.-D. Lim and D. Lee. A primal-dual perspective for distributed td-learning. *arXiv preprint*  
609 *arXiv:2310.00638*, 2023.
- 611 Y. Lin, G. Qu, L. Huang, and A. Wierman. Multi-agent reinforcement learning in stochastic  
612 networked systems. *Advances in neural information processing systems*, 34:7825–7837,  
613 2021.
- 614 M. L. Littman. Value-function reinforcement learning in markov games. *Cognitive systems*  
615 *research*, 2(1):55–66, 2001.
- 616 S. V. Macua, J. Chen, S. Zazo, and A. H. Sayed. Distributed policy evaluation under multiple  
617 behavior strategies. *IEEE Transactions on Automatic Control*, 60(5):1260–1274, 2014.
- 618 A. Nedic and A. Ozdaglar. Distributed subgradient methods for multi-agent optimization.  
619 *IEEE Transactions on Automatic Control*, 54(1):48–61, 2009.
- 620 A. Olshevsky. Linear time average consensus on fixed graphs and implications for decentralized  
621 optimization and multi-agent control. *arXiv preprint arXiv:1411.4186*, 2014.
- 622 D. Paulin. Concentration inequalities for markov chains by marton couplings and spectral  
623 methods. 2015.
- 624 S. U. Pillai, T. Suel, and S. Cha. The perron-frobenius theorem: some of its applications.  
625 *IEEE Signal Processing Magazine*, 22(2):62–75, 2005.
- 626 K. Prabuchandran, H. K. AN, and S. Bhatnagar. Multi-agent reinforcement learning for  
627 traffic signal control. In *17th International IEEE Conference on Intelligent Transportation*  
628 *Systems (ITSC)*, pages 2529–2534. IEEE, 2014.
- 629 S. Pu and A. Nedić. Distributed stochastic gradient tracking methods. *Mathematical*  
630 *Programming*, 187(1):409–457, 2021.
- 631 G. Qu and A. Wierman. Finite-time analysis of asynchronous stochastic approximation and  
632 q-learning. In *Conference on Learning Theory*, pages 3185–3205. PMLR, 2020.
- 633 G. Qu, A. Wierman, and N. Li. Scalable reinforcement learning for multiagent networked  
634 systems. *Operations Research*, 70(6):3601–3628, 2022.
- 640 T. Rashid, M. Samvelyan, C. S. De Witt, G. Farquhar, J. Foerster, and S. Whiteson.  
641 Monotonic value function factorisation for deep multi-agent reinforcement learning. *Journal*  
642 *of Machine Learning Research*, 21(178):1–51, 2020.
- 643 S. Shalev-Shwartz, S. Shammah, and A. Shashua. Safe, multi-agent, reinforcement learning  
644 for autonomous driving. *arXiv preprint arXiv:1610.03295*, 2016.
- 645 J. Sun, G. Wang, G. B. Giannakis, Q. Yang, and Z. Yang. Finite-time analysis of decen-  
646 tralized temporal-difference learning with linear function approximation. In *International*  
647 *Conference on Artificial Intelligence and Statistics*, pages 4485–4495. PMLR, 2020.

- 648 P. Sunehag, G. Lever, A. Gruslys, W. M. Czarnecki, V. Zambaldi, M. Jaderberg, M. Lanctot,  
649 N. Sonnerat, J. Z. Leibo, K. Tuyls, et al. Value-decomposition networks for cooperative  
650 multi-agent learning. *arXiv preprint arXiv:1706.05296*, 2017.
- 651 M. Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In  
652 *Proceedings of the tenth international conference on machine learning*, pages 330–337,  
653 1993.
- 654 G. Wang, S. Lu, G. Giannakis, G. Tesauro, and J. Sun. Decentralized td tracking with  
655 linear function approximation and its finite-time analysis. *Advances in neural information*  
656 *processing systems*, 33:13762–13772, 2020.
- 657 J. Wang and N. Elia. Control approach to distributed optimization. In *2010 48th Annual*  
658 *Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 557–561.  
659 IEEE, 2010.
- 660 Y. Wang, Y. Wang, Y. Zhou, A. Velasquez, and S. Zou. Data-driven robust multi-agent  
661 reinforcement learning. In *2022 IEEE 32nd International Workshop on Machine Learning*  
662 *for Signal Processing (MLSP)*, pages 1–6. IEEE, 2022.
- 663 C. J. Watkins and P. Dayan. Q-learning. *Machine learning*, 8:279–292, 1992.
- 664 J. Woo, G. Joshi, and Y. Chi. The blessing of heterogeneity in federated q-learning: Linear  
665 speedup and beyond. In *International Conference on Machine Learning*, pages 37157–37216.  
666 PMLR, 2023.
- 667 L. Xiao and S. Boyd. Fast linear iterations for distributed averaging. *Systems & Control*  
668 *Letters*, 53(1):65–78, 2004.
- 669 T. Yang, S. Cen, Y. Wei, Y. Chen, and Y. Chi. Federated natural policy gradient methods  
670 for multi-task reinforcement learning. 2023.
- 671 S. Zeng, T. Chen, A. Garcia, and M. Hong. Learning to coordinate in multi-agent systems:  
672 A coordinated actor-critic algorithm and finite-time guarantees. In *Learning for Dynamics*  
673 *and Control Conference*, pages 278–290. PMLR, 2022a.
- 674 S. Zeng, T. T. Doan, and J. Romberg. Finite-time convergence rates of decentralized  
675 stochastic approximation with applications in multi-agent and multi-task learning. *IEEE*  
676 *Transactions on Automatic Control*, 2022b.
- 677 K. Zhang, Z. Yang, and T. Basar. Networked multi-agent reinforcement learning in continuous  
678 spaces. In *2018 IEEE conference on decision and control (CDC)*, pages 2771–2776. IEEE,  
679 2018a.
- 680 K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Basar. Fully decentralized multi-agent  
681 reinforcement learning with networked agents. In *International Conference on Machine*  
682 *Learning*, pages 5872–5881. PMLR, 2018b.
- 683 K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Başar. Finite-sample analysis for decentralized  
684 batch multiagent reinforcement learning with networked agents. *IEEE Transactions on*  
685 *Automatic Control*, 66(12):5925–5940, 2021.
- 686 Y. Zhang and M. M. Zavlanos. Distributed off-policy actor-critic reinforcement learning  
687 with policy consensus. In *2019 IEEE 58th Conference on decision and control (CDC)*,  
688 pages 4674–4679. IEEE, 2019.
- 689 Y. Zhang, G. Qu, P. Xu, Y. Lin, Z. Chen, and A. Wierman. Global convergence of localized  
690 policy iteration in networked multi-agent reinforcement learning. *Proceedings of the ACM*  
691 *on Measurement and Analysis of Computing Systems*, 7(1):1–51, 2023.
- 692 C. Zhao, J. Liu, M. Sheng, W. Teng, Y. Zheng, and J. Li. Multi-uav trajectory planning for  
693 energy-efficient content coverage: A decentralized learning-based approach. *IEEE Journal*  
694 *on Selected Areas in Communications*, 39(10):3193–3207, 2021.
- 695 Z. Zheng, F. Gao, L. Xue, and J. Yang. Federated q-learning: Linear regret speedup with  
696 low communication cost. *arXiv preprint arXiv:2312.15023*, 2023.
- 697  
698  
699  
700  
701

## A APPENDIX : NOTATIONS

$\mathbb{R}^n$ : set of real-valued  $n$ -dimensional vectors;  $\mathbb{R}^{n \times m}$ : set of real-valued  $n \times m$ -dimensional matrices;  $\Delta^n$  for  $n \in \mathbb{N}$ : a probability simplex in  $\mathbb{R}^n$ ;  $[n]$  for  $n \in \mathbb{N}$ :  $\{1, 2, \dots, n\}$ ;  $\mathbf{1}_n$ :  $n$ -dimensional vector whose elements are all one;  $\mathbf{0}$ : a vector whose elements are all zero with appropriate dimension;  $[\mathbf{A}]_{ij}$ :  $i$ -th row and  $j$ -th column for any matrix  $\mathbf{A}$ ;  $\mathbf{e}_j$ : basis vector (with appropriate dimension) whose  $j$ -th element is one and others are all zero;  $|\mathcal{S}|$ : cardinality of any finite set  $\mathcal{S}$ ;  $\otimes$ : Kronecker product between two matrices;  $\mathbf{a} \geq \mathbf{b}$  for  $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ :  $[\mathbf{a}]_i \geq [\mathbf{b}]_i$  for all  $i \in [n]$ .

## B APPENDIX : CONSTRUCTING DOUBLY STOCHASTIC MATRIX

**Example B.1** (Lazy Metropolis matrix in Olshevsky (2014)). *To construct the doubly stochastic matrix  $\mathbf{W}$  with only local information, we can set  $[\mathbf{W}]_{ij} = \frac{1}{2 \max\{|\mathcal{N}_i|, |\mathcal{N}_j|\}}$  for  $i \neq j$  and  $i, j \in [N]$ , letting  $[\mathbf{W}]_{ii} = 1 - \sum_{j \in \mathcal{N}_i} [\mathbf{W}]_{ij}$ . This uses only local information, and does not require any global information sharing.*

One can formulate a semi-definite program to construct a doubly stochastic matrix (Xiao and Boyd, 2004). It finds the doubly stochastic matrix with minimum possible  $\sigma_2(W)$  but it requires a centralized controller to solve such system, and distributed the computed the result of each agents. Another choice is to use Sinkhorn-Knopp algorithm (Knight, 2008). However, it also requires a centralized computation scheme. Moreover, to our best knowledge, we are not aware of bound on the  $\sigma_2(W)$  of the output of Sinkhorn-Knopp algorithm.

## C APPENDIX : TECHNICAL DETAILS

**Lemma C.1.** *We have for  $\mathbf{Q} \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$ ,*

$$\|\mathbf{A}_Q\|_\infty \leq 1 - (1 - \gamma)d_{\min}\alpha.$$

*Proof.* For  $i \in [|\mathcal{S}| \times |\mathcal{A}|]$ , we have

$$\begin{aligned} \sum_{j=1}^{|\mathcal{S}| \times |\mathcal{A}|} \|[\mathbf{A}_Q]_{ij}\| &\leq 1 - [\mathbf{D}]_{ii}\alpha + \alpha[\mathbf{D}]_{ii}\gamma \sum_{j=1}^{|\mathcal{S}| \times |\mathcal{A}|} [\mathbf{P}\mathbf{\Pi}^Q]_{ij} \\ &= 1 - [\mathbf{D}]_{ii}(1 - \gamma)\alpha. \end{aligned}$$

The last equality follows from the fact that  $\mathbf{P}\mathbf{\Pi}^Q$  is a stochastic matrix, i.e., the row sum equals to one, and represents a probability distribution. Taking maximum over  $i \in [|\mathcal{S}| \times |\mathcal{A}|]$ , we complete the proof.  $\square$

**Lemma C.2.** *For  $k \in \mathbb{N}$ , we have*

$$\|\boldsymbol{\epsilon}_k^{\text{avg}}\|_\infty \leq \frac{4R_{\max}}{1 - \gamma}.$$

*Proof.* From the definition of  $\boldsymbol{\epsilon}_k^{\text{avg}} = \frac{1}{N} \sum_{i=1}^N \boldsymbol{\delta}_k^i - \boldsymbol{\Delta}_k^i$  in (4), we have

$$\begin{aligned} \|\boldsymbol{\epsilon}_k^{\text{avg}}\|_\infty &\leq 2 \left( R_{\max} + \gamma \frac{R_{\max}}{1 - \gamma} + \frac{R_{\max}}{1 - \gamma} \right) \\ &= \frac{4R_{\max}}{1 - \gamma}, \end{aligned}$$

where the first inequality comes from the boundedness of  $\bar{\mathbf{Q}}_k$  in Lemma 3.3. This completes the proof.  $\square$

756 **Lemma C.3.** For  $a, b \in (0, 1)$ , and for  $k \in \mathbb{N}$ , we have

$$757 \sum_{i=0}^k a^{k-i} b^i \leq a^{\frac{k}{2}} \frac{1}{1-b} + b^{\frac{k}{2}} \frac{1}{1-a}.$$

761 Furthermore, we have

$$762 \sum_{i=\tau}^k a^{k-i} b^{i-\tau} \leq a^{\frac{k-\tau}{2}} \frac{1}{1-b} + b^{\frac{k-\tau}{2}} \frac{1}{1-a}.$$

766 *Proof.* We have

$$767 \sum_{i=0}^k a^{k-i} b^i \leq \sum_{i=0}^{\lceil \frac{k}{2} \rceil} a^{k-i} b^i + \sum_{i=\lfloor \frac{k}{2} \rfloor}^k a^{k-i} b^i$$

$$770 \leq a^{\frac{k}{2}} \frac{1}{1-b} + b^{\frac{k}{2}} \frac{1}{1-a}.$$

774 The last inequality follows from the summation of geometric series. As for the second item, we have

$$775 \sum_{i=\tau}^k a^{k-i} b^{i-\tau} \leq \sum_{i=\tau}^{\lceil \frac{k+\tau}{2} \rceil} a^{k-i} b^{i-\tau} + \sum_{i=\lfloor \frac{k+\tau}{2} \rfloor}^k a^{k-i} b^{i-\tau}$$

$$778 \leq a^{\frac{k-\tau}{2}} \frac{1}{1-b} + b^{\frac{k-\tau}{2}} \frac{1}{1-a}.$$

782 This completes the proof.  $\square$

784 **Lemma C.4** (Azuma-Hoeffding Inequality, Theorem 2.19 in Chung and Lu (2006)). Let  $\{S_n\}_{n \in \mathbb{N}}$  be a Martingale sequence with  $S_0 = 0$ . Suppose  $|S_k - S_{k-1}| \leq c_k$  for  $k \in \mathbb{N}$ . Then, for  $\epsilon \geq 0$ , we have

$$785 \mathbb{P}[|S_k| \geq \epsilon] \leq 2 \exp\left(-\frac{\epsilon^2}{2 \sum_{j=1}^k c_j^2}\right).$$

791 **Lemma C.5.** Suppose  $X \geq 0$ ,  $\mathbb{P}[X \geq \epsilon] \leq \min\{a \exp(-b\epsilon^2), 1\}$ , and  $a \geq 2$ . Then, we have

$$792 \mathbb{E}[X] \leq 2\sqrt{\frac{\ln a}{b}}.$$

796 *Proof.* We have

$$797 \mathbb{E}[X] = \int_0^\infty \mathbb{P}[X \geq s] ds$$

$$798 \leq \int_0^\infty \min\{a \exp(-bs^2), 1\} ds$$

$$799 \leq \int_0^{\sqrt{\frac{\ln a}{b}}} 1 ds + \int_{\sqrt{\frac{\ln a}{b}}}^\infty a \exp(-bs^2) ds$$

$$800 \leq \sqrt{\frac{\ln a}{b}} + \frac{1}{2\sqrt{b \ln a}}$$

$$801 \leq 2\sqrt{\frac{\ln a}{b}}.$$

The last inequality follows from the fact that  $4 \ln a > 1/\ln a$ . The third inequality follows from the following relation:

$$\begin{aligned} \int_{\sqrt{\frac{\ln a}{b}}}^{\infty} a \exp(-bs^2) ds &= a \int_{\frac{\ln a}{b}}^{\infty} \frac{1}{2\sqrt{u}} \exp(-bu) du \\ &\leq \frac{a}{2} \sqrt{\frac{b}{\ln a}} \int_{\frac{\ln a}{b}}^{\infty} \exp(-bu) du \\ &= \frac{a}{2} \sqrt{\frac{b}{\ln a}} \frac{1}{b} [-\exp(-bu)]_{\frac{\ln a}{b}}^{\infty} \\ &= \frac{1}{2\sqrt{b \ln a}}. \end{aligned}$$

where we used the change of variables  $s^2 = u$  in the first equality.  $\square$

**Definition C.6** (Martingale sequence, Section 4.2 in Durrett (2019)). *Consider a sequence of random variables  $\{X_n\}_{n \in \mathbb{N}}$  and an increasing  $\sigma$ -field,  $\mathcal{F}_n$ , such that*

- 1)  $\mathbb{E}[|X_n|] < \infty$ ;
- 2)  $X_n$  is  $\mathcal{F}_n$ -measurable;
- 3)  $\mathbb{E}[X_{n+1} | \mathcal{F}_n] = X_n, \quad \forall n \in \mathbb{N}$ .

Then,  $X_n$  is said to be a Martingale sequence.

**Lemma C.7** (Proposition 3.4 in Paulin (2015)). *For uniformly ergodic Markov chain in Section 4, we have, for  $\epsilon > 0$ ,*

$$\tau(\epsilon) \leq t_{\text{mix}} \left( 1 + 2 \log \left( \frac{1}{\epsilon} \right) + \log \left( \frac{1}{d_{\text{min}}} \right) \right),$$

where  $\tau$  and  $t_{\text{mix}}$  are defined in (13).

## D APPENDIX : OMITTED PROOFS

### D.1 PROOF OF LEMMA 3.2

*Proof.* From the definition of  $\bar{\mathbf{W}}$  in (4), we have

$$\begin{aligned} (\bar{\mathbf{W}}^k \boldsymbol{\Theta})^\top \bar{\mathbf{W}}^k \boldsymbol{\Theta} &= \bar{\mathbf{W}}^{2k} - 2\bar{\mathbf{W}}^k \frac{1}{N} ((\mathbf{1}_N \mathbf{1}_N^\top) \otimes \mathbf{I}_{|S||\mathcal{A}|}) + \frac{1}{N} (\mathbf{1}_N \mathbf{1}_N^\top) \otimes \mathbf{I}_{|S||\mathcal{A}|} \\ &= \left( \mathbf{W}^{2k} - \frac{1}{N} \mathbf{1}_N \mathbf{1}_N^\top \right) \otimes \mathbf{I}_{|S||\mathcal{A}|}, \end{aligned}$$

where the second equality follows from the fact that  $\bar{\mathbf{W}} (\mathbf{1}_N \mathbf{1}_N^\top)^\top \otimes \mathbf{I}_{|S||\mathcal{A}|} = (\mathbf{1}_N \mathbf{1}_N^\top)^\top \otimes \mathbf{I}_{|S||\mathcal{A}|}$ . From the result, we can derive

$$\|\bar{\mathbf{W}}^k \boldsymbol{\Theta}\|_2 = \sqrt{\lambda_{\max}((\bar{\mathbf{W}}^k \boldsymbol{\Theta})^\top \bar{\mathbf{W}}^k \boldsymbol{\Theta})} = \sqrt{\lambda_{\max} \left( \mathbf{W}^{2k} - \frac{1}{N} \mathbf{1}_N \mathbf{1}_N^\top \right)} = \sigma_2(\mathbf{W})^k < 1. \quad (17)$$

To prove the inequality in (17), we first prove that 1 is the unique largest eigenvalue of  $\mathbf{W}$ . Noting that  $\mathbf{1}_N$  is an eigenvector of  $\mathbf{W}$  with eigenvalue of 1, and  $\rho(\mathbf{W}) \leq \|\mathbf{W}\|_\infty = 1$  where  $\rho(\cdot)$  is the spectral radius of a matrix, the largest eigenvalue of  $\mathbf{W}$  should be one. This implies that  $\sigma_2(\mathbf{W}) < 1$ . The multiplicity of the eigenvalue 1 is one, which follows from the fact that  $\mathbf{W}^k$  is a non-negative and irreducible matrix and that the largest eigenvalue of a non-negative and irreducible matrix is unique Pillai et al. (2005) from Perron-Frobenius theorem. Note that  $\mathbf{W}^k$  is a non-negative and irreducible matrix due to the fact that the graph  $\mathcal{G}$  is connected.

Next, we use the eigenvalue decomposition of a symmetric matrix to investigate the spectrum of  $\mathbf{W}^{2k} - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top$ . By eigendecomposition of a symmetric matrix, we have

$$\mathbf{W} = \lambda_1 \mathbf{v}_1 \mathbf{v}_1^\top + \sum_{j=2}^N \lambda_j \mathbf{v}_j \mathbf{v}_j^\top = \mathbf{T} \mathbf{\Lambda} \mathbf{T}^{-1},$$

where  $\mathbf{v}_j$  and  $\lambda_j$  are  $j$ -th eigenvector and eigenvalue of  $\mathbf{W}$ ,  $\lambda_1 = 1$ ,  $\mathbf{v}_1 = \frac{1}{\sqrt{N}}\mathbf{1}_N$ ,  $\mathbf{\Lambda}$  is a diagonal matrix whose diagonal elements are the eigenvalues of  $\mathbf{W}$ , and  $\mathbf{T}$  and  $\mathbf{T}^{-1}$  are formed from the eigenvectors of  $\mathbf{W}$ . From the uniqueness of the maximum eigenvalue of  $\mathbf{W}$ , we have  $\lambda_1 = 1 > \lambda_j, j \in \{2, 3, \dots, N\}$ . Therefore, we have

$$\mathbf{W}^{2k} = \mathbf{T} \mathbf{\Lambda}^{2k} \mathbf{T}^{-1} = \left( \frac{1}{\sqrt{N}} \mathbf{1}_N \right) \left( \frac{1}{\sqrt{N}} \mathbf{1}_N^\top \right) + \sum_{j=2}^N \lambda_j^{2k} \mathbf{v}_j \mathbf{v}_j^\top.$$

Therefore, we have  $\lambda_{\max}(\mathbf{W}^{2k} - \frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top) = \sigma_2(\mathbf{W}^{2k})$ . This completes the proof.  $\square$

## D.2 PROOF OF LEMMA 3.3

*Proof.* Let us first assume that for some  $k \in \mathbb{N}$ ,  $\|\mathbf{Q}_k^i\|_\infty \leq \frac{R_{\max}}{1-\gamma}$  for all  $i \in [N]$ . Then, considering (2), for all  $i \in [N]$ , we have

$$\begin{aligned} |\mathbf{Q}_{k+1}^i(s_k, \mathbf{a}_k)| &\leq ([\mathbf{W}]_{ii} - \alpha) \|\mathbf{Q}_k^i\|_\infty + \sum_{j \in [N] \setminus \{i\}} [\mathbf{W}]_{ij} \|\mathbf{Q}_k^j\|_\infty + \alpha (R_{\max} + \gamma \|\mathbf{Q}_k^i\|_\infty) \\ &\leq (1 - \alpha) \frac{R_{\max}}{1 - \gamma} + \alpha \frac{R_{\max}}{1 - \gamma} \\ &= \frac{R_{\max}}{1 - \gamma}. \end{aligned}$$

The first inequality follows from the fact that  $\alpha \leq \min_{i \in [N]} [\mathbf{W}]_{ii}$ . The second inequality follows from the induction hypothesis. For,  $s, \mathbf{a} \in \mathcal{S} \times \mathcal{A} \setminus \{s_k, \mathbf{a}_k\}$ , we have

$$|\mathbf{Q}_{k+1}^i(s, \mathbf{a})| \leq \sum_{j \in \mathcal{N}_i} [\mathbf{W}]_{ij} |\mathbf{Q}_k^j(s, \mathbf{a})| \leq \frac{R_{\max}}{1 - \gamma}.$$

The last line follows from the fact that  $\mathbf{W}$  is a doubly stochastic matrix, and the induction hypothesis. The proof is completed by applying the induction argument.  $\square$

## D.3 PROOF OF THEOREM 3.4

*Proof.* Taking infinity norm on (7), we get

$$\begin{aligned} \|\Theta \bar{\mathbf{Q}}_{k+1}\|_\infty &\leq \|\bar{\mathbf{W}}^{k+1} \Theta \bar{\mathbf{Q}}_0\|_2 + \alpha \sqrt{N|\mathcal{S}||\mathcal{A}|} \sum_{j=0}^k \|\bar{\mathbf{W}}^{k-j} \Theta\|_2 \left\| \left( \bar{\mathbf{D}} \left( \bar{\mathbf{R}} + \gamma \bar{\mathbf{P}} \bar{\mathbf{\Pi}}^{\bar{\mathbf{Q}}_j} \bar{\mathbf{Q}}_j - \bar{\mathbf{Q}}_j \right) + \bar{\epsilon}_j \right) \right\|_\infty \\ &\leq \|\bar{\mathbf{W}}^{k+1} \Theta \bar{\mathbf{Q}}_0\|_2 + \alpha \sqrt{N|\mathcal{S}||\mathcal{A}|} \sum_{j=0}^k \|\bar{\mathbf{W}}^{k-j} \Theta\|_2 \frac{8R_{\max}}{1 - \gamma} \\ &\leq \sigma_2(\mathbf{W})^{k+1} \|\Theta \bar{\mathbf{Q}}_0\|_2 + \alpha \sqrt{N|\mathcal{S}||\mathcal{A}|} \sum_{j=0}^k \sigma_2(\mathbf{W})^{k-j} \frac{8R_{\max}}{1 - \gamma} \\ &\leq \sigma_2(\mathbf{W})^{k+1} \|\Theta \bar{\mathbf{Q}}_0\|_2 + \alpha \frac{8R_{\max}}{1 - \gamma} \frac{\sqrt{N|\mathcal{S}||\mathcal{A}|}}{1 - \sigma_2(\mathbf{W})}. \end{aligned}$$

The first inequality follows from the inequality  $\|\mathbf{A}\|_\infty \leq \sqrt{N|\mathcal{S}||\mathcal{A}|} \|\mathbf{A}\|_2$  for  $\mathbf{A} \in \mathbb{R}^{N|\mathcal{S}||\mathcal{A}| \times N|\mathcal{S}||\mathcal{A}|}$ . The second inequality follows from the bound on  $\bar{\mathbf{Q}}_k$  in Lemma 3.3. The third inequality follows from Lemma 3.2. The last inequality follows from summation of geometric series. This completes the proof.  $\square$

## D.4 PROOF OF LEMMA 3.5

*Proof.* From the definition of  $\mathbf{E}_k$  in (9), we get

$$\begin{aligned}
\|\mathbf{E}_k\|_\infty &\leq \frac{\gamma}{N} \sum_{i=1}^N \left\| \mathbf{DP}(\Pi^{Q_k^i} \mathbf{Q}_k^i - \Pi^{Q_k^{\text{avg}}} \mathbf{Q}_k^{\text{avg}}) \right\|_\infty \\
&\leq \frac{\gamma d_{\max}}{N} \sum_{i=1}^N \left\| \begin{bmatrix} \max_{\mathbf{a} \in \mathcal{A}} \mathbf{Q}_k^i(1, \mathbf{a}) - \max_{\mathbf{a} \in \mathcal{A}} \mathbf{Q}_k^{\text{avg}}(1, \mathbf{a}) \\ \max_{\mathbf{a} \in \mathcal{A}} \mathbf{Q}_k^i(2, \mathbf{a}) - \max_{\mathbf{a} \in \mathcal{A}} \mathbf{Q}_k^{\text{avg}}(2, \mathbf{a}) \\ \vdots \\ \max_{\mathbf{a} \in \mathcal{A}} \mathbf{Q}_k^i(|\mathcal{S}|, \mathbf{a}) - \max_{\mathbf{a} \in \mathcal{A}} \mathbf{Q}_k^{\text{avg}}(|\mathcal{S}|, \mathbf{a}) \end{bmatrix} \right\|_\infty \\
&\leq \frac{\gamma d_{\max}}{N} \sum_{i=1}^N \|\mathbf{Q}_k^i - \mathbf{Q}_k^{\text{avg}}\|_\infty \\
&\leq \gamma d_{\max} \|\Theta \bar{\mathbf{Q}}_k\|_\infty.
\end{aligned}$$

The third inequality follows from the fact that  $|\max_{i \in [n]} [\mathbf{x}]_i - \max_{i \in [n]} [\mathbf{y}]_i| \leq \max_{i \in [n]} |\mathbf{x}_i - \mathbf{y}_i|$  for  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$  and  $n \in \mathbb{N}$ . The last inequality follows from the fact that

$$\|\mathbf{Q}_k^i - \mathbf{Q}_k^{\text{avg}}\|_\infty \leq \|\Theta \bar{\mathbf{Q}}_k\|_\infty, \quad \forall i \in [N].$$

This completes the proof.  $\square$

## E APPENDIX : CONSTRUCTION OF UPPER AND LOWER COMPARISON SYSTEM

## E.1 CONSTRUCTION OF LOWER COMPARISON SYSTEM

**Lemma E.1.** For  $k \in \mathbb{N}$ , if  $\mathbf{Q}_0^{\text{avg},l} \leq \mathbf{Q}_0^{\text{avg}}$ , we have

$$\mathbf{Q}_k^{\text{avg},l} \leq \mathbf{Q}_k^{\text{avg}}.$$

*Proof.* The proof follows from the induction argument. Suppose the statement holds for some  $k \in \mathbb{N}$ . Then, we have

$$\begin{aligned}
\mathbf{Q}_{k+1}^{\text{avg},l} &= \mathbf{Q}_k^{\text{avg},l} + \alpha D \left( \mathbf{R}^{\text{avg}} + \gamma \mathbf{P} \Pi^{Q^*} \mathbf{Q}_k^{\text{avg},l} - \mathbf{Q}_k^{\text{avg},l} \right) + \alpha \epsilon_k^{\text{avg}} + \alpha \mathbf{E}_k \\
&\leq \mathbf{Q}_k^{\text{avg}} + \alpha D \left( \mathbf{R}^{\text{avg}} + \gamma \mathbf{P} \Pi^{Q_k^{\text{avg}}} \mathbf{Q}_k^{\text{avg}} - \mathbf{Q}_k^{\text{avg}} \right) + \alpha \epsilon_k^{\text{avg}} + \alpha \mathbf{E}_k \\
&= \mathbf{Q}_{k+1}^{\text{avg}}.
\end{aligned}$$

The first inequality follows from the fact that  $\mathbf{Q}_k^{\text{avg},l} \leq \mathbf{Q}_k^{\text{avg}}$  and  $\Pi^{Q^*} \mathbf{Q}_k^{\text{avg},l} \leq \Pi^{Q^*} \mathbf{Q}_k^{\text{avg}} \leq \Pi^{Q_k^{\text{avg}}} \mathbf{Q}_k^{\text{avg}}$ . The proof is completed by the induction argument.  $\square$

## E.2 CONSTRUCTION OF UPPER COMPARISON SYSTEM

**Lemma E.2.** For  $k \in \mathbb{N}$ , if  $\tilde{\mathbf{Q}}_0^{\text{avg},u} \geq \tilde{\mathbf{Q}}_0^{\text{avg}}$ , we have

$$\tilde{\mathbf{Q}}_k^{\text{avg},u} \geq \tilde{\mathbf{Q}}_k^{\text{avg}}.$$

*Proof.* As in the construction of the lower comparison system in Lemma E.1 in Appendix, the proof follows from an induction argument. Suppose that the statement holds for some  $k \in \mathbb{N}$ . Then, we have

$$\begin{aligned}
\tilde{\mathbf{Q}}_{k+1}^{\text{avg}} &= \tilde{\mathbf{Q}}_k^{\text{avg}} + \alpha D \left( \gamma \mathbf{P} \Pi^{Q_k^{\text{avg}}} \tilde{\mathbf{Q}}_k^{\text{avg}} - \tilde{\mathbf{Q}}_k^{\text{avg}} \right) + \alpha \gamma \mathbf{DP}(\Pi^{Q_k^{\text{avg}}} \mathbf{Q}^* - \Pi^{Q^*} \mathbf{Q}^*) \\
&\quad + \alpha \epsilon_k^{\text{avg}} + \alpha D \mathbf{E}_k \\
&\leq (\mathbf{I} + \alpha D (\gamma \mathbf{P} \Pi^{Q_k^{\text{avg}}} - \mathbf{I})) \tilde{\mathbf{Q}}_k^{\text{avg},u} + \alpha \epsilon_k^{\text{avg}} + \alpha D \mathbf{E}_k \\
&= \tilde{\mathbf{Q}}_{k+1}^{\text{avg},u}.
\end{aligned}$$

The inequality follows from the fact that the elements of  $\mathbf{I} + \alpha \mathbf{D}(\gamma \mathbf{P} \mathbf{\Pi} \mathbf{Q}_k^{\text{avg}} - \mathbf{I})$  are all non-negative, and  $\mathbf{\Pi} \mathbf{Q}_k^{\text{avg}} \mathbf{Q}^* \leq \mathbf{\Pi} \mathbf{Q}^* \mathbf{Q}^*$ . The proof is completed by the induction argument.  $\square$

## F APPENDIX : I.I.D. OBSERVATION MODEL

**Proposition F.1.** *Assume i.i.d. observation model, and  $\alpha \leq \min_{i \in [N]} [\mathbf{W}]_{ii}$ . Then, we have, for  $k \in \mathbb{N}$ ,*

$$\begin{aligned} \mathbb{E} \left[ \left\| \tilde{\mathbf{Q}}_{k+1}^{\text{avg}, l} \right\|_{\infty} \right] &= \tilde{\mathcal{O}} \left( (1 - (1 - \gamma) d_{\min} \alpha)^{\frac{k}{2}} + \sigma_2(\mathbf{W})^{\frac{k}{2}} \right) \\ &\quad + \tilde{\mathcal{O}} \left( \alpha^{\frac{1}{2}} \frac{R_{\max}}{(1 - \gamma)^{\frac{3}{2}} d_{\min}^{\frac{1}{2}}} + \alpha d_{\max} \frac{R_{\max} \sqrt{N} |\mathcal{S}| |\mathcal{A}|}{(1 - \gamma)^2 d_{\min} (1 - \sigma_2(\mathbf{W}))} \right). \end{aligned}$$

Let us first introduce a key lemma to prove Proposition F.1:

**Lemma F.2.** *For  $k \in \mathbb{N}$ , we have*

$$\mathbb{E} \left[ \left\| \sum_{i=0}^k \mathbf{A}_{\mathbf{Q}^*}^{k-i} \boldsymbol{\epsilon}_i^{\text{avg}} \right\|_{\infty} \right] \leq \frac{8\sqrt{2} R_{\max}}{(1 - \gamma)^{\frac{3}{2}} d_{\min}^{\frac{1}{2}} \alpha^{\frac{1}{2}}} \sqrt{\ln(2|\mathcal{S}||\mathcal{A}|)}.$$

*Proof.* For the proof, we will apply Azuma-Hoeffding inequality in Lemma C.4. For simplicity, let  $\mathbf{S}_t = \sum_{i=0}^t \mathbf{A}_{\mathbf{Q}^*}^{k-i} \boldsymbol{\epsilon}_i^{\text{avg}}$ , for  $0 \leq t \leq k$ . Let  $\mathcal{F}_t := \sigma(\{(s_i, \mathbf{a}_i, s'_i)\}_{i=0}^t \cup \{\tilde{\mathbf{Q}}_0\})$ , which is the  $\sigma$ -algebra generated by  $\{(s_i, \mathbf{a}_i, s'_i)\}_{i=0}^t$  and  $\tilde{\mathbf{Q}}_0$ . Letting  $[\mathbf{S}_t]_{s, \mathbf{a}} = (\mathbf{e}_s \otimes \mathbf{e}_{\mathbf{a}})^{\top} \mathbf{S}_t$ , for  $s, \mathbf{a} \in \mathcal{S} \times \mathcal{A}$ , let us check that  $\{[\mathbf{S}_t]_{s, \mathbf{a}}\}_{t=0}^k$  is a Martingale sequence defined in Definition C.6. We can see that

$$\begin{aligned} \mathbb{E} [\mathbf{S}_t | \mathcal{F}_{t-1}] &= \mathbb{E} \left[ \mathbf{A}_{\mathbf{Q}^*}^{k-t} \boldsymbol{\epsilon}_t^{\text{avg}} + \mathbf{S}_{t-1} \middle| \mathcal{F}_{t-1} \right] \\ &= \mathbf{A}_{\mathbf{Q}^*}^{k-t} \mathbb{E} [\boldsymbol{\epsilon}_t^{\text{avg}} | \mathcal{F}_{t-1}] + \mathbf{S}_{t-1} \\ &= \mathbf{S}_{t-1}, \end{aligned}$$

where the second line is due to the fact that  $\mathbf{S}_{t-1}$  is  $\mathcal{F}_{t-1}$ -measurable, and the last line follows from  $\mathbb{E} [\boldsymbol{\epsilon}_t^{\text{avg}} | \mathcal{F}_{t-1}] = \mathbf{0}$  thanks to the i.i.d. observation model. Therefore, we have  $\mathbb{E} [[\mathbf{S}_t]_{s, \mathbf{a}} | \mathcal{F}_{t-1}] = [\mathbf{S}_{t-1}]_{s, \mathbf{a}}$ .

Moreover, we have

$$\begin{aligned} \mathbb{E} [\mathbf{S}_0] &= \mathbb{E} \left[ \frac{1}{N} \sum_{i=1}^N (\mathbf{e}_{s_0} \otimes \mathbf{e}_{\mathbf{a}_0}) (r_1^i + \mathbf{e}_{s'_0}^{\top} \gamma \mathbf{\Pi} \mathbf{Q}_0^i - (\mathbf{e}_{s_0} \otimes \mathbf{e}_{\mathbf{a}_0})^{\top} \mathbf{Q}_0^i) \right] \\ &\quad - \mathbb{E} \left[ \frac{1}{N} \sum_{i=1}^N \mathbf{D}(\mathbf{R}^i + \gamma \mathbf{P} \mathbf{\Pi} \mathbf{Q}_0^i - \mathbf{Q}_0^i) \right] \\ &= \mathbf{0}. \end{aligned}$$

The last line follows from that  $\mathbb{E} [\mathbf{e}_{s_0} \otimes \mathbf{e}_{\mathbf{a}_0}] = \mathbf{D}$  and  $\mathbb{E} [(\mathbf{e}_{s_0} \otimes \mathbf{e}_{\mathbf{a}_0}) \mathbf{e}_{s'_0}^{\top}] = \mathbf{D} \mathbf{P}$ .

Therefore,  $\{[\mathbf{S}_t]_{s, \mathbf{a}}\}_{t=0}^k$  is a Martingale sequence for any  $s, \mathbf{a} \in \mathcal{S} \times \mathcal{A}$ . Furthermore, we have

$$|[\mathbf{S}_t]_{s, \mathbf{a}} - [\mathbf{S}_{t-1}]_{s, \mathbf{a}}| \leq \|\mathbf{S}_t - \mathbf{S}_{t-1}\|_{\infty} = \left\| \mathbf{A}_{\mathbf{Q}^*}^{k-t} \boldsymbol{\epsilon}_t^{\text{avg}} \right\|_{\infty} \leq (1 - (1 - \gamma) d_{\min} \alpha)^{k-t} \frac{4R_{\max}}{1 - \gamma},$$

where the last inequality comes from Lemma C.1 and Lemma C.2. Furthermore, note that we have

$$\begin{aligned} \sum_{t=1}^k |[\mathbf{S}_t]_{s, \mathbf{a}} - [\mathbf{S}_{t-1}]_{s, \mathbf{a}}|^2 &\leq \sum_{t=0}^k (1 - (1 - \gamma) d_{\min} \alpha)^{2k-2t} \frac{16R_{\max}^2}{(1 - \gamma)^2} \\ &\leq \frac{16R_{\max}^2}{(1 - \gamma)^3 d_{\min} \alpha}. \end{aligned}$$

Therefore, applying the Azuma-Hoeffding inequality in Lemma C.4 in the Appendix, we have

$$\mathbb{P} [\|\mathbf{S}_k\|_{s,a} \geq \epsilon] \leq 2 \exp \left( -\frac{\epsilon^2(1-\gamma)^3 d_{\min} \alpha}{32R_{\max}^2} \right).$$

Noting that  $\{\|\mathbf{S}_k\|_{\infty} \geq \epsilon\} \subseteq \cup_{s,a \in \mathcal{S} \times \mathcal{A}} \{\|\mathbf{S}_k\|_{s,a} \geq \epsilon\}$ , using the union bound of the events, we get:

$$\mathbb{P} [\|\mathbf{S}_k\|_{\infty} \geq \epsilon] \leq \sum_{s,a \in \mathcal{S} \times \mathcal{A}} \mathbb{P} [\|\mathbf{S}_k\|_{s,a} \geq \epsilon] \leq 2|\mathcal{S}||\mathcal{A}| \exp \left( -\frac{\epsilon^2(1-\gamma)^3 d_{\min} \alpha}{32R_{\max}^2} \right).$$

Moreover, since a probability of an event is always smaller than one, we have

$$\mathbb{P} [\|\mathbf{S}_k\|_{\infty} \geq \epsilon] \leq \min \left\{ 2|\mathcal{S}||\mathcal{A}| \exp \left( -\frac{\epsilon^2(1-\gamma)^3 d_{\min} \alpha}{32R_{\max}^2} \right), 1 \right\}.$$

Now, we are ready to bound  $\mathbf{S}_k$  from Lemma C.5 in the Appendix:

$$\mathbb{E} [\|\mathbf{S}_k\|_{\infty}] = \int_0^{\infty} \mathbb{P} [\|\mathbf{S}_k\|_{\infty} \geq x] dx \leq \frac{8\sqrt{2}R_{\max}}{(1-\gamma)^{\frac{3}{2}} d_{\min}^{\frac{1}{2}} \alpha^{\frac{1}{2}}} \sqrt{\ln(2|\mathcal{S}||\mathcal{A}|)}.$$

This completes the proof.  $\square$

Now, we are ready prove Proposition F.1:

*Proof of Proposition F.1.* Recursively expanding the equation in (12), we get

$$\begin{aligned} \tilde{Q}_{k+1}^{\text{avg},l} &= \mathbf{A}_{Q^*} \tilde{Q}_k^{\text{avg},l} + \alpha \epsilon_k^{\text{avg}} + \alpha \mathbf{E}_k \\ &= \mathbf{A}_{Q^*}^2 \tilde{Q}_{k-1}^{\text{avg},l} + \alpha \mathbf{A}_{Q^*} \epsilon_{k-1}^{\text{avg}} + \alpha \mathbf{A}_{Q^*} \mathbf{E}_{k-1} + \alpha \epsilon_k^{\text{avg}} + \alpha \mathbf{E}_k \\ &= \mathbf{A}_{Q^*}^{k+1} \tilde{Q}_0^{\text{avg},l} + \alpha \sum_{i=0}^k \mathbf{A}_{Q^*}^{k-i} \epsilon_i^{\text{avg}} + \alpha \sum_{i=0}^k \mathbf{A}_{Q^*}^{k-i} \mathbf{E}_i. \end{aligned}$$

Taking infinity norm and expectation on both sides of the above equation, we get

$$\begin{aligned} &\mathbb{E} \left[ \left\| \tilde{Q}_{k+1}^{\text{avg},l} \right\|_{\infty} \right] \\ &\leq \mathbb{E} \left[ \left\| \mathbf{A}_{Q^*}^{k+1} \right\|_{\infty} \left\| \tilde{Q}_0^{\text{avg},l} \right\|_{\infty} + \alpha \left\| \sum_{i=0}^k \mathbf{A}_{Q^*}^{k-i} \epsilon_i^{\text{avg}} \right\|_{\infty} + \alpha \sum_{i=0}^k \left\| \mathbf{A}_{Q^*}^{k-i} \right\|_{\infty} \left\| \mathbf{E}_i \right\|_{\infty} \right] \\ &\leq (1 - (1-\gamma)d_{\min} \alpha)^{k+1} \left\| \tilde{Q}_0^{\text{avg},l} \right\|_{\infty} + \alpha \mathbb{E} \left[ \left\| \sum_{i=0}^k \mathbf{A}_{Q^*}^{k-i} \epsilon_i^{\text{avg}} \right\|_{\infty} \right] \\ &\quad + \alpha \mathbb{E} \left[ \sum_{i=0}^k \left\| \mathbf{A}_{Q^*}^{k-i} \right\|_{\infty} \left\| \mathbf{E}_i \right\|_{\infty} \right] \\ &\leq (1 - (1-\gamma)d_{\min} \alpha)^{k+1} \left\| \tilde{Q}_0^{\text{avg},l} \right\|_{\infty} + \alpha^{\frac{1}{2}} \frac{8\sqrt{2}R_{\max}}{(1-\gamma)^{\frac{3}{2}} d_{\min}^{\frac{1}{2}}} \sqrt{\ln(2|\mathcal{S}||\mathcal{A}|)} \\ &\quad + \alpha \mathbb{E} \left[ \sum_{i=0}^k \left\| \mathbf{A}_{Q^*}^{k-i} \right\|_{\infty} \left\| \mathbf{E}_i \right\|_{\infty} \right] \\ &\leq (1 - (1-\gamma)d_{\min} \alpha)^{k+1} \left\| \tilde{Q}_0^{\text{avg},l} \right\|_{\infty} + \alpha^{\frac{1}{2}} \frac{8\sqrt{2}R_{\max}}{(1-\gamma)^{\frac{3}{2}} d_{\min}^{\frac{1}{2}}} \sqrt{\ln(2|\mathcal{S}||\mathcal{A}|)} \\ &\quad + \gamma d_{\max} \left\| \Theta \tilde{Q}_0 \right\|_2 \left( (1 - (1-\gamma)d_{\min} \alpha)^{\frac{k}{2}} \frac{\alpha}{1 - \sigma_2(\mathbf{W})} + \sigma_2(\mathbf{W})^{\frac{k}{2}} \frac{1}{(1-\gamma)d_{\min}} \right) \\ &\quad + \alpha \gamma d_{\max} \frac{8R_{\max} \sqrt{N|\mathcal{S}||\mathcal{A}|}}{(1-\gamma)^2 d_{\min} (1 - \sigma_2(\mathbf{W}))}. \end{aligned}$$

The second inequality follows from Lemma C.1. The third inequality follows from Lemma F.2. The last line follows from bounding  $\sum_{i=0}^k \left\| \mathbf{A}_{\mathbf{Q}^*}^{k-i} \right\|_{\infty} \|\mathbf{E}_i\|_{\infty}$  as follows:

$$\begin{aligned}
& s \sum_{i=0}^k \left\| \mathbf{A}_{\mathbf{Q}^*}^{k-i} \right\|_{\infty} \|\mathbf{E}_i\|_{\infty} \\
& \leq \gamma d_{\max} \sum_{i=0}^k (1 - (1 - \gamma)d_{\min}\alpha)^{k-i} \left( \sigma_2(\mathbf{W})^i \|\Theta \bar{\mathbf{Q}}_0\|_2 + \alpha \frac{8R_{\max}}{1 - \gamma} \frac{\sqrt{N|\mathcal{S}||\mathcal{A}|}}{1 - \sigma_2(\mathbf{W})} \right) \\
& \leq \gamma d_{\max} \|\Theta \bar{\mathbf{Q}}_0\|_2 \left( (1 - (1 - \gamma)d_{\min}\alpha)^{\frac{k}{2}} \frac{1}{1 - \sigma_2(\mathbf{W})} + \sigma_2(\mathbf{W})^{\frac{k}{2}} \frac{1}{(1 - \gamma)d_{\min}\alpha} \right) \\
& \quad + \gamma d_{\max} \frac{8R_{\max}\sqrt{N|\mathcal{S}||\mathcal{A}|}}{(1 - \gamma)^2 d_{\min}(1 - \sigma_2(\mathbf{W}))}.
\end{aligned}$$

The first inequality follows from Lemma 3.5 and Theorem 3.4. The second inequality follows from Lemma C.3 in the Appendix. This completes the proof.  $\square$

Now, we bound  $\tilde{\mathbf{Q}}_k^{\text{avg},u}$  in (12). It is difficult to directly prove the convergence of upper comparison system. Therefore, we bound the difference of upper and lower comparison system,  $\mathbf{Q}_k^{\text{avg},u} - \mathbf{Q}_k^{\text{avg},l}$ . The good news is that since  $\mathbf{Q}_k^{\text{avg},u}$  and  $\mathbf{Q}_k^{\text{avg},l}$  shares the same error term  $\epsilon_k^{\text{avg}}$  and  $\mathbf{E}_k$ , such terms will be removed if we subtract each others.

**Proposition F.3.** For  $k \in \mathbb{N}$ , and  $\alpha \leq \min_{i \in [N]} [\mathbf{W}]_{ii}$ , we have

$$\begin{aligned}
\mathbb{E} \left[ \left\| \mathbf{Q}_{k+1}^{\text{avg},u} - \mathbf{Q}_{k+1}^{\text{avg},l} \right\|_{\infty} \right] &= \tilde{\mathcal{O}} \left( (1 - \alpha(1 - \gamma)d_{\min})^{\frac{k}{2}} + \sigma_2(\mathbf{W})^{\frac{k}{4}} \right) \\
& \quad + \tilde{\mathcal{O}} \left( \alpha^{\frac{1}{2}} \frac{d_{\max} R_{\max}}{(1 - \gamma)^{\frac{5}{2}} d_{\min}^{\frac{3}{2}}} + \alpha \frac{d_{\max}^2 \sqrt{N|\mathcal{S}||\mathcal{A}|} R_{\max}}{(1 - \gamma)^3 d_{\min}^2 (1 - \sigma_2(\mathbf{W}))} \right).
\end{aligned}$$

*Proof.* Subtracting  $\mathbf{Q}_{k+1}^{\text{avg},l}$  from  $\mathbf{Q}_{k+1}^{\text{avg},u}$  in (12), we have

$$\begin{aligned}
\mathbf{Q}_{k+1}^{\text{avg},u} - \mathbf{Q}_{k+1}^{\text{avg},l} &= \mathbf{A}_{\mathbf{Q}_k^{\text{avg}}} \tilde{\mathbf{Q}}_k^{\text{avg},u} - \mathbf{A}_{\mathbf{Q}^*} \tilde{\mathbf{Q}}_k^{\text{avg},l} \\
&= \mathbf{A}_{\mathbf{Q}_k^{\text{avg}}} (\mathbf{Q}_k^{\text{avg},u} - \mathbf{Q}_k^{\text{avg},l}) + (\mathbf{A}_{\mathbf{Q}_k^{\text{avg}}} - \mathbf{A}_{\mathbf{Q}^*}) \tilde{\mathbf{Q}}_k^{\text{avg},l} \\
&= \mathbf{A}_{\mathbf{Q}_k^{\text{avg}}} (\mathbf{Q}_k^{\text{avg},u} - \mathbf{Q}_k^{\text{avg},l}) + \alpha \gamma \mathbf{DP}(\Pi^{\mathbf{Q}_k^{\text{avg}}} - \Pi^{\mathbf{Q}^*}) \tilde{\mathbf{Q}}_k^{\text{avg},l}. \tag{18}
\end{aligned}$$

The last equality follows from the definition of  $\mathbf{A}_{\mathbf{Q}_k^{\text{avg}}}$  and  $\mathbf{A}_{\mathbf{Q}^*}$  in (10).

Recursively expanding the terms, we get

$$\begin{aligned}
\mathbf{Q}_{k+1}^{\text{avg},u} - \mathbf{Q}_{k+1}^{\text{avg},l} &= \prod_{i=0}^k \mathbf{A}_{\mathbf{Q}_i^{\text{avg}}} (\mathbf{Q}_0^{\text{avg},u} - \mathbf{Q}_0^{\text{avg},l}) \\
& \quad + \alpha \gamma \sum_{i=0}^{k-1} \prod_{j=i}^{k-1} \mathbf{A}_{\mathbf{Q}_{j+1}^{\text{avg}}} \mathbf{DP}(\Pi^{\mathbf{Q}_i^{\text{avg}}} - \Pi^{\mathbf{Q}^*}) \tilde{\mathbf{Q}}_i^{\text{avg},l} + \alpha \gamma \mathbf{DP}(\Pi^{\mathbf{Q}_k^{\text{avg}}} - \Pi^{\mathbf{Q}^*}) \tilde{\mathbf{Q}}_k^{\text{avg},l}.
\end{aligned}$$

Taking infinity norm on both sides of the above equation, and using triangle inequality yields

$$\begin{aligned}
\mathbb{E} \left[ \left\| \mathbf{Q}_{k+1}^{\text{avg},u} - \mathbf{Q}_{k+1}^{\text{avg},l} \right\|_{\infty} \right] &\leq (1 - \alpha(1 - \gamma)d_{\min})^{k+1} \left\| \mathbf{Q}_0^{\text{avg},u} - \mathbf{Q}_0^{\text{avg},l} \right\|_{\infty} \\
& \quad + \underbrace{2\alpha\gamma d_{\max} \sum_{i=0}^k (1 - \alpha(1 - \gamma)d_{\min})^{k-i} \mathbb{E} \left[ \left\| \tilde{\mathbf{Q}}_i^{\text{avg},l} \right\|_{\infty} \right]}_{(*)}. \tag{19}
\end{aligned}$$

The first inequality follows from Lemma C.1.

Now, we will use Proposition F.1 to bound  $(\star)$  in the above inequality. We have

$$\begin{aligned} \sum_{i=0}^k (1 - \alpha(1 - \gamma)d_{\min})^{k-i} \mathbb{E} \left[ \left\| \tilde{\mathbf{Q}}_i^{\text{avg},l} \right\|_{\infty} \right] &= \tilde{\mathcal{O}} \left( \sum_{j=0}^k (1 - \alpha(1 - \gamma)d_{\min})^{k-\frac{j}{2}} + (1 - \alpha(1 - \gamma)d_{\min})^{k-i} \sigma_2(\mathbf{W})^{\frac{i}{2}} \right) \\ &\quad + \tilde{\mathcal{O}} \left( \frac{R_{\max}}{\alpha^{\frac{1}{2}}(1 - \gamma)^{\frac{5}{2}}d_{\min}^{\frac{3}{2}}} + \frac{d_{\max}\sqrt{N|\mathcal{S}||\mathcal{A}|}2R_{\max}}{(1 - \gamma)^3d_{\min}^2(1 - \sigma_2(\mathbf{W}))} \right) \\ &= \tilde{\mathcal{O}} \left( (1 - \alpha(1 - \gamma)d_{\min})^{\frac{k}{2}} + \sigma_2(\mathbf{W})^{\frac{k}{4}} \right) \\ &\quad + \tilde{\mathcal{O}} \left( \frac{R_{\max}}{\alpha^{\frac{1}{2}}(1 - \gamma)^{\frac{5}{2}}d_{\min}^{\frac{3}{2}}} + \frac{d_{\max}\sqrt{N|\mathcal{S}||\mathcal{A}|}R_{\max}}{(1 - \gamma)^3d_{\min}^2(1 - \sigma_2(\mathbf{W}))} \right). \end{aligned}$$

The last inequality follows from Lemma C.3. Applying this result to (19), we get

$$\begin{aligned} \mathbb{E} \left[ \left\| \mathbf{Q}_{k+1}^{\text{avg},u} - \mathbf{Q}_{k+1}^{\text{avg},l} \right\|_{\infty} \right] &= \tilde{\mathcal{O}} \left( (1 - \alpha(1 - \gamma)d_{\min})^{\frac{k}{2}} + \sigma_2(\mathbf{W})^{\frac{k}{4}} \right) \\ &\quad + \tilde{\mathcal{O}} \left( \alpha^{\frac{1}{2}}d_{\max} \frac{R_{\max}}{(1 - \gamma)^{\frac{5}{2}}d_{\min}^{\frac{3}{2}}} + \alpha \frac{d_{\max}^2\sqrt{N|\mathcal{S}||\mathcal{A}|}R_{\max}}{(1 - \gamma)^3d_{\min}^2(1 - \sigma_2(\mathbf{W}))} \right). \end{aligned}$$

This completes the proof.  $\square$

## F.1 PROOF OF THEOREM 3.6

*Proof.*  $\left\| \tilde{\mathbf{Q}}_k^{\text{avg}} \right\|_{\infty}$  can be bounded using the fact that  $\tilde{\mathbf{Q}}_k^{\text{avg},l} \leq \tilde{\mathbf{Q}}_k^{\text{avg}} \leq \tilde{\mathbf{Q}}_k^{\text{avg},u}$ :

$$\begin{aligned} \left\| \tilde{\mathbf{Q}}_k^{\text{avg}} \right\|_{\infty} &\leq \max \left\{ \left\| \tilde{\mathbf{Q}}_k^{\text{avg},l} \right\|_{\infty}, \left\| \tilde{\mathbf{Q}}_k^{\text{avg},u} \right\|_{\infty} \right\} \\ &\leq \max \left\{ \left\| \tilde{\mathbf{Q}}_k^{\text{avg},l} \right\|_{\infty}, \left\| \tilde{\mathbf{Q}}_k^{\text{avg},l} \right\|_{\infty} + \left\| \tilde{\mathbf{Q}}_k^{\text{avg},u} - \tilde{\mathbf{Q}}_k^{\text{avg},l} \right\|_{\infty} \right\} \\ &\leq \left\| \tilde{\mathbf{Q}}_k^{\text{avg},l} \right\|_{\infty} + \left\| \tilde{\mathbf{Q}}_k^{\text{avg},u} - \tilde{\mathbf{Q}}_k^{\text{avg},l} \right\|_{\infty} \\ &= \left\| \tilde{\mathbf{Q}}_k^{\text{avg},l} \right\|_{\infty} + \left\| \mathbf{Q}_k^{\text{avg},u} - \mathbf{Q}_k^{\text{avg},l} \right\|_{\infty}. \end{aligned}$$

The second inequality follows from triangle inequality. Taking expectation, from Proposition F.1 and Proposition F.3, we have the desired result.  $\square$

## F.2 PROOF OF THEOREM 3.7

*Proof.* Using triangle inequality, we have

$$\begin{aligned} \mathbb{E} \left[ \left\| \tilde{\mathbf{Q}}_k - \mathbf{1}_N \otimes \mathbf{Q}^* \right\|_{\infty} \right] &\leq \mathbb{E} \left[ \left\| \tilde{\mathbf{Q}}_k - \mathbf{1}_N \otimes \mathbf{Q}_k^{\text{avg}} \right\|_{\infty} \right] + \mathbb{E} \left[ \left\| \mathbf{1}_N \otimes \mathbf{Q}_k^{\text{avg}} - \mathbf{1}_N \otimes \mathbf{Q}^* \right\|_{\infty} \right] \\ &= \mathbb{E} \left[ \left\| \tilde{\mathbf{Q}}_k - \mathbf{1}_N \otimes \mathbf{Q}_k^{\text{avg}} \right\|_{\infty} \right] + \mathbb{E} \left[ \left\| \mathbf{Q}_k^{\text{avg}} - \mathbf{Q}^* \right\|_{\infty} \right] \\ &= \tilde{\mathcal{O}} \left( \sigma_2(\mathbf{W})^k + \alpha \frac{\sqrt{N|\mathcal{S}||\mathcal{A}|}R_{\max}}{(1 - \gamma)(1 - \sigma_2(\mathbf{W}))} \right) \\ &\quad + \tilde{\mathcal{O}} \left( (1 - \alpha(1 - \gamma)d_{\min})^{\frac{k}{2}} + \sigma_2(\mathbf{W})^{\frac{k}{4}} \right) \\ &\quad + \tilde{\mathcal{O}} \left( \alpha^{\frac{1}{2}} \frac{d_{\max}R_{\max}}{(1 - \gamma)^{\frac{5}{2}}d_{\min}^{\frac{3}{2}}} + \alpha \frac{d_{\max}^2\sqrt{N|\mathcal{S}||\mathcal{A}|}R_{\max}}{(1 - \gamma)^3d_{\min}^2(1 - \sigma_2(\mathbf{W}))} \right) \\ &= \tilde{\mathcal{O}} \left( (1 - \alpha(1 - \gamma)d_{\min})^{\frac{k}{2}} + \sigma_2(\mathbf{W})^{\frac{k}{4}} \right) \\ &\quad + \tilde{\mathcal{O}} \left( \alpha^{\frac{1}{2}}d_{\max} \frac{R_{\max}}{(1 - \gamma)^{\frac{5}{2}}d_{\min}^{\frac{3}{2}}} + \alpha \frac{d_{\max}^2\sqrt{N|\mathcal{S}||\mathcal{A}|}R_{\max}}{(1 - \gamma)^3d_{\min}^2(1 - \sigma_2(\mathbf{W}))} \right). \end{aligned}$$

The first inequality comes from (6). The second inequality comes from Theorem 3.4 and 3.6. This completes the proof.  $\square$

## F.3 PROOF OF COROLLARY 3.8

*Proof.* Let us first bound the terms  $\alpha^{\frac{1}{2}} d_{\max} \frac{R_{\max}}{(1-\gamma)^{\frac{5}{2}} d_{\min}^{\frac{3}{2}}} + \alpha \frac{d_{\max}^2 \sqrt{|\mathcal{S}||\mathcal{A}|} R_{\max}}{(1-\gamma)^3 d_{\min}^2 (1-\sigma_2(\mathbf{W}))}$  in Theorem 3.7 with  $\epsilon$ . We require

$$\alpha = \tilde{\mathcal{O}} \left( \min \left\{ \frac{(1-\gamma)^5 d_{\min}^3 \epsilon^2}{R_{\max}^2 d_{\max}^2}, \frac{(1-\gamma)^3 d_{\min}^2 (1-\sigma_2(\mathbf{W}))}{R_{\max} d_{\max}^2 \sqrt{|\mathcal{S}||\mathcal{A}|}} \epsilon \right\} \right).$$

Next, we bound the terms  $(1 - \alpha(1 - \gamma)d_{\min})^{\frac{k}{2}} + \sigma_2(\mathbf{W})^{\frac{k}{4}}$ . Noting that

$$(1 - \alpha(1 - \gamma)d_{\min})^{\frac{k}{2}} \leq \exp \left( -\alpha(1 - \gamma)d_{\min} \frac{k}{2} \right),$$

we require

$$\begin{aligned} k &= \tilde{\mathcal{O}} \left( \frac{1}{(1-\gamma)d_{\min}\alpha} \ln \left( \frac{1}{\epsilon} \right) + \ln \left( \frac{1}{\epsilon} \right) / \ln \left( \frac{1}{\sigma_2(\mathbf{W})} \right) \right) \\ &= \tilde{\mathcal{O}} \left( \ln \left( \frac{1}{\epsilon} \right) \max \left\{ \frac{R_{\max}^2 d_{\max}^2}{\epsilon^2 (1-\gamma)^6 d_{\min}^4}, \frac{R_{\max} d_{\max}^2 \sqrt{|\mathcal{S}||\mathcal{A}|}}{\epsilon (1-\gamma)^4 d_{\min}^3 (1-\sigma_2(\mathbf{W}))} \right\} \right). \end{aligned}$$

This completes the proof.  $\square$

## G APPENDIX : MARKOVIAN OBSERVATION MODEL

In this section, we provide the analysis tools for the Markovian observation model in Section 4.

Considering a sequence of state-action trajectory  $\{(s_k, \mathbf{a}_k)\}_{k \in \mathbb{N}}$  induced by the behavior policy  $\beta$ , the update of Q-function at time  $k$  becomes

$$\begin{aligned} \mathbf{Q}_{k+1}^i(s_k, \mathbf{a}_k) &= \sum_{j \in \mathcal{N}_i} [\mathbf{W}]_{ij} \mathbf{Q}_k^j(s_k, \mathbf{a}_k) + \alpha \left( r_{k+1}^i + \gamma \max_{\mathbf{a} \in \mathcal{A}} \mathbf{Q}_k^i(s_{k+1}, \mathbf{a}) - \mathbf{Q}_k^i(s_k, \mathbf{a}_k) \right) \\ \mathbf{Q}_{k+1}^i(s, \mathbf{a}) &= \sum_{j \in \mathcal{N}_i} [\mathbf{W}]_{ij} \mathbf{Q}_k^j(s, \mathbf{a}), \quad s, \mathbf{a} \in \mathcal{S} \times \mathcal{A} \setminus \{(s_k, \mathbf{a}_k)\}, \end{aligned} \quad (20)$$

where we have replaced  $s'_k$  in (2) with  $s_{k+1}$ . The overall algorithm is given in Algorithm 2 in the Appendix Section J.

We follow the same definitions in Section 3 by letting  $\mathbf{D}$  to be  $\mathbf{D}_{\infty}$ . That is, we have

$$\mathbf{A}_Q = \mathbf{I} + \alpha \mathbf{D}_{\infty} (\gamma \mathbf{P} \mathbf{\Pi}^Q - \mathbf{I}), \quad \mathbf{b}_Q = \gamma \mathbf{D}_{\infty} \mathbf{P} (\mathbf{\Pi}^Q - \mathbf{\Pi}^{Q^*}) \mathbf{Q}^*,$$

which are defined in (10).

Furthermore, let us define for  $\mathbf{Q} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ ,  $\bar{\mathbf{Q}} \in \mathbb{R}^{N|\mathcal{S}||\mathcal{A}|}$ , and  $\bar{\mathbf{Q}}^i \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$  such that  $[\mathbf{Q}^i]_j = [\bar{\mathbf{Q}}]_{|\mathcal{S}||\mathcal{A}|(i-1)+j}$  for  $j \in [|\mathcal{S}||\mathcal{A}|]$ :

$$\begin{aligned} \Delta^{\text{avg}}(\bar{\mathbf{Q}}) &= \mathbf{D}_{\infty} \frac{1}{N} \sum_{i=1}^N \left( \mathbf{R}^i + \gamma \mathbf{P} \mathbf{\Pi}^{\mathbf{Q}^i} \mathbf{Q}^i - \mathbf{Q}^i \right), \\ \Delta_{k-\tau, \tau}^{\text{avg}}(\bar{\mathbf{Q}}) &:= \mathbf{D}_{\tau}^{s_{k-\tau}, \mathbf{a}_{k-\tau}} \frac{1}{N} \sum_{i=1}^N \left( \mathbf{R}^i + \gamma \mathbf{P} \mathbf{\Pi}^{\mathbf{Q}^i} \mathbf{Q}^i - \mathbf{Q}^i \right), \end{aligned}$$

where  $\mathbf{D}_{\tau}^{s_{k-\tau}, \mathbf{a}_{k-\tau}}$  is defined in (14).

Note that we did not use any property of the i.i.d. distribution in proving the consensus error. Therefore, we can directly use the result in Theorem 3.4 for the consensus error for Markovian observation model. Hence, in this section, we focus on bounding the optimality error,  $\mathbf{Q}_k^{\text{avg}} - \mathbf{Q}^*$ . As in the case of i.i.d. observation model in Section 3, we will analyze the error bound of lower and upper comparison system in the subsequent sections.

1242 G.1 ANALYSIS OF OPTIMALITY ERROR UNDER MARKOVIAN OBSERVATION MODEL  
1243

1244 As in Section 3.3, we will analyze the error bound for  $\tilde{Q}_k^{\text{avg},u}$  and  $\tilde{Q}_k^{\text{avg},l}$  to bound the  
1245 optimality error,  $\tilde{Q}_k^{\text{avg}}$ . We will present an error bound on the lower comparison system,  $\tilde{Q}_k^{\text{avg},l}$ ,  
1246 in Proposition G.5, and the error bound on  $\tilde{Q}_k^{\text{avg},u} - \tilde{Q}_k^{\text{avg},l}$  in Proposition G.6. Collecting  
1247 the results, the result on the optimality error,  $\tilde{Q}_k^{\text{avg}}$ , will be presented in Theorem G.7.  
1248

1249 Let us first investigate the lower comparison system.  $\tilde{Q}_k^{\text{avg},l}$  evolves via (12) where we replace  
1250  $\epsilon_k^{\text{avg}}$  with  $\epsilon_k^{\text{avg}}(o_k, \bar{Q}_k)$  where  $o_k = (s_k, \mathbf{a}_k, s_{k+1})$ . To analyze the error under Markovian  
1251 observation model, we decompose the terms, for  $k \geq \tau$  as follows:  
1252

$$\begin{aligned}
1253 \tilde{Q}_{k+1}^{\text{avg},l} &= \mathbf{A}_{Q^*} \tilde{Q}_k^{\text{avg},l} + \alpha \epsilon_k^{\text{avg}}(o_k, \bar{Q}_k) + \alpha \mathbf{E}_k \\
1254 &= \mathbf{A}_{Q^*} \tilde{Q}_k^{\text{avg},l} + \alpha \epsilon_k^{\text{avg}}(o_k, \bar{Q}_{k-\tau}) + \alpha (\epsilon_k^{\text{avg}}(o_k, \bar{Q}_k) - \epsilon_k^{\text{avg}}(o_k, \bar{Q}_{k-\tau})) + \alpha \mathbf{E}_k \\
1255 &= \mathbf{A}_{Q^*} \tilde{Q}_k^{\text{avg},l} + \underbrace{\alpha (\delta_k^{\text{avg}}(o_k, \bar{Q}_{k-\tau}) - \Delta_{k-\tau,\tau}^{\text{avg}}(\bar{Q}_{k-\tau}))}_{:=\mathbf{w}_{k,1}} + \underbrace{\alpha (\Delta_{k-\tau,\tau}^{\text{avg}}(\bar{Q}_{k-\tau}) - \Delta^{\text{avg}}(\bar{Q}_{k-\tau}))}_{:=\mathbf{w}_{k,2}} \\
1256 &\quad + \underbrace{\alpha (\epsilon_k^{\text{avg}}(o_k, \bar{Q}_k) - \epsilon_k^{\text{avg}}(o_k, \bar{Q}_{k-\tau}))}_{:=\mathbf{w}_{k,3}} + \alpha \mathbf{E}_k. \\
1257 & \\
1258 & \\
1259 & \\
1260 & \\
1261 & \tag{21}
\end{aligned}$$

1262 The decomposition is motivated to invoke Azuma-Hoeffding inequality as explained in  
1263 Section 4. Recursively expanding the terms in (21), we get  
1264

$$\begin{aligned}
1265 \tilde{Q}_{k+1}^{\text{avg},l} &= \mathbf{A}_{Q^*}^{k-\tau+1} \tilde{Q}_\tau^{\text{avg},l} + \alpha \sum_{j=\tau}^k \mathbf{A}_{Q^*}^{k-j} \mathbf{w}_{j,1} + \alpha \sum_{j=\tau}^k \mathbf{A}_{Q^*}^{k-j} \mathbf{w}_{j,2} + \alpha \sum_{j=\tau}^k \mathbf{A}_{Q^*}^{k-j} \mathbf{w}_{j,3} + \alpha \sum_{j=\tau}^k \mathbf{A}_{Q^*}^{k-j} \mathbf{E}_j. \\
1266 & \\
1267 & \\
1268 & \tag{22}
\end{aligned}$$

1269 Now, let us provide an analysis on the lower comparison system.  
1270

1271 We will provide the bounds of  $\sum_{j=\tau}^k \mathbf{A}_{Q^*}^{k-j} \mathbf{w}_{j,1}$ ,  $\sum_{j=\tau}^k \mathbf{A}_{Q^*}^{k-j} \mathbf{w}_{j,2}$ , and  $\sum_{j=\tau}^k \mathbf{A}_{Q^*}^{k-j} \mathbf{w}_{j,3}$  in  
1272 Lemma G.2, Lemma G.3, and Lemma G.4, respectively. We first provide an important  
1273 property to bound  $\sum_{j=\tau}^k \mathbf{A}_{Q^*}^{k-j} \mathbf{w}_{j,1}$ .  
1274

1275 **Lemma G.1.** For  $t \geq \tau$ , let  $\mathcal{F}_t := \sigma(\{\bar{Q}_0, s_0, \mathbf{a}_0, s_1, \mathbf{a}_1, \dots, s_t, \mathbf{a}_t\})$ . Then,  
1276

$$1277 \mathbb{E}[\mathbf{w}_{t,1} | \mathcal{F}_{t-\tau}] = \mathbf{0}.$$

1278 *Proof.* We have  
1279

$$\begin{aligned}
1280 \mathbb{E}[\mathbf{w}_{t,1} | \mathcal{F}_{t-\tau}] &= \mathbb{E}[\delta_k^{\text{avg}}(o_k, \bar{Q}_{k-\tau}) - \Delta_{k-\tau,\tau}^{\text{avg}}(\bar{Q}_{k-\tau}) | \mathcal{F}_{t-\tau}] \\
1281 &= \frac{1}{N} \sum_{i=1}^N \mathbb{E}[(\mathbf{e}_{s_t} \otimes \mathbf{e}_{\mathbf{a}_t})(r_{t+1} + \mathbf{e}_{s_{t+1}}^\top \gamma \mathbf{\Pi} \mathbf{Q}_{t-\tau}^i - (\mathbf{e}_{s_t} \otimes \mathbf{e}_{\mathbf{a}_t})^\top \mathbf{Q}_{t-\tau}^i) | \mathcal{F}_{t-\tau}] \\
1282 &\quad - \frac{1}{N} \mathbf{D}_\tau^{s_{t-\tau}, \mathbf{a}_{t-\tau}} \sum_{i=1}^N (\mathbf{R}^i + \gamma \mathbf{P} \mathbf{\Pi} \mathbf{Q}_{t-\tau}^i - \mathbf{Q}_{t-\tau}^i) \\
1283 & \\
1284 & \\
1285 & \\
1286 & \\
1287 & \\
1288 & = \mathbf{0}.
\end{aligned}$$

1289 The second equality follows from the fact that  $\mathbf{Q}_{t-\tau}^i$  is  $\mathcal{F}_{t-\tau}$ -measurable. This completes the  
1290 proof.  $\square$   
1291

1292 **Lemma G.2.** For  $k \in \mathbb{N}$ , and  $\alpha \leq \min\{\min_{i \in [N]} [\mathbf{W}]_{ii}, \frac{1}{2\tau}\}$ , we have  
1293

$$1294 \mathbb{E} \left[ \left\| \sum_{j=\tau}^k \mathbf{A}_{Q^*}^{k-j} \mathbf{w}_{j,1} \right\|_\infty \right] \leq 2\sqrt{\ln(2\tau|\mathcal{S}||\mathcal{A}|)} \frac{15\sqrt{\tau} R_{\max}}{(1-\gamma)^{\frac{3}{2}} d_{\min}^{\frac{1}{2}} \alpha^{\frac{1}{2}}}.$$

1295

1296 *Proof.* For  $0 \leq q \leq \tau - 1$ , let for  $t \in \mathbb{N}$  such that  $q \leq \tau t + q \leq k$ :

$$1297 \mathcal{F}_{k,t}^q := \mathcal{F}_{\tau t + q}.$$

1298 Then, let us consider the sequence  $\{\mathbf{S}_{k,t}^q\}_{t \in \{t \in \mathbb{N}: q \leq \tau t + q \leq k\}}$  as follows:

$$1300 \mathbf{S}_{k,t}^q := \sum_{j=1}^t \mathbf{A}_{\mathcal{Q}^*}^{k-\tau j - q} \mathbf{w}_{\tau j + q, 1}.$$

1301 Next, we will apply Azuma-Hoeffding inequality in Lemma C.4. Let us first check that  
1302  $\{\mathbf{S}_{k,t}^q\}_{t \in \{t \in \mathbb{N}: \tau t + q \leq k\}}$  is a Martingale sequence. We can see that

$$1303 \mathbb{E} \left[ \mathbf{S}_{k,t}^q \middle| \mathcal{F}_{k,t-1}^q \right] = \mathbb{E} \left[ \mathbf{A}_{\mathcal{Q}^*}^{k-\tau t - q} \mathbf{w}_{\tau t + q, 1} \middle| \mathcal{F}_{k,t-1}^q \right] + \mathbb{E} \left[ \sum_{j=1}^{t-1} \mathbf{A}_{\mathcal{Q}^*}^{k-\tau j - q} \mathbf{w}_{\tau j + q, 1} \middle| \mathcal{F}_{k,t-1}^q \right]$$

$$1304 = \mathbf{S}_{k,t-1}^q.$$

1305 The second equality follows from Lemma G.1, and the fact that  $\mathbf{S}_{k,t-1}^q$  is  $\mathcal{F}_{k,t-1}^q$ -measurable.

1306 Moreover, we have  $\mathbb{E} \left[ \mathbf{S}_{k,1}^q \middle| \mathcal{F}_q \right] = \mathbf{0}$ , and

$$1307 \left\| \mathbf{S}_{k,t}^q - \mathbf{S}_{k,t-1}^q \right\|_{\infty} = \left\| \mathbf{A}_{\mathcal{Q}^*}^{k-\tau t - q} \mathbf{w}_{\tau t + q, 1} \right\|_{\infty} \leq (1 - (1 - \gamma) d_{\min} \alpha)^{k - \tau t - q} \frac{4R_{\max}}{1 - \gamma},$$

1308 where the last inequality follows from Lemma C.1. Now, we have, for  $s, \mathbf{a} \in \mathcal{S} \times \mathcal{A}$ ,

$$1309 \sum_{j \in \{t \in \mathbb{N}: q < \tau t + q \leq k\}} \left| [\mathbf{S}_{k,j}^q]_{s, \mathbf{a}} - [\mathbf{S}_{k,j-1}^q]_{s, \mathbf{a}} \right| \leq \sum_{j \in \{t \in \mathbb{N}: \tau t + q \leq k\}} (1 - (1 - \gamma) d_{\min} \alpha)^{2k - 2\tau j - 2q} \frac{16R_{\max}^2}{(1 - \gamma)^2}$$

$$1310 \leq \frac{1}{(1 - (1 - (1 - \gamma) d_{\min} \alpha)^{2\tau})} \frac{16R_{\max}^2}{(1 - \gamma)^2}.$$

1311 Therefore, we can now apply Azuma-Hoeffding inequality in Lemma C.4, which yields

$$1312 \mathbb{P} \left[ \left\| \mathbf{S}_{k,t^*(q)}^q \right\|_{\infty} \geq \epsilon \right] \leq 2|\mathcal{S}||\mathcal{A}| \exp \left( - \frac{\epsilon^2 (1 - (1 - (1 - \gamma) d_{\min} \alpha)^{2\tau}) (1 - \gamma)^2}{2 \cdot 16R_{\max}^2} \right),$$

1313 where  $t^*(q) = \max\{t \in \mathbb{N} : \tau t + q \leq k\}$ . Considering that

$$1314 \bigcap_{q=0}^{\tau-1} \left\{ \left\| \mathbf{S}_{k,t^*(q)}^q \right\|_{\infty} < \epsilon / \tau \right\} \subset \{ \|\mathbf{S}_k\|_{\infty} < \epsilon \},$$

1315 taking the union bound of the events,

$$1316 \mathbb{P} \left[ \|\mathbf{S}_k\|_{\infty} \geq \epsilon \right] \leq \min \left\{ \sum_{0 \leq q \leq \tau - 1} \mathbb{P} \left[ \left\| \mathbf{S}_{k,t^*(q)}^q \right\|_{\infty} \geq \epsilon / \tau \right], 1 \right\}$$

$$1317 \leq \min \left\{ 2\tau |\mathcal{S}||\mathcal{A}| \exp \left( - \frac{\epsilon^2 (1 - (1 - (1 - \gamma) d_{\min} \alpha)^{2\tau}) (1 - \gamma)^2}{2\tau^2 \cdot 16R_{\max}^2} \right), 1 \right\}.$$

1318 Therefore, from Lemma C.5, we have

$$1319 \mathbb{E} \left[ \|\mathbf{S}_k\|_{\infty} \right] \leq 2\sqrt{\ln(2\tau|\mathcal{S}||\mathcal{A}|)} \frac{6\tau R_{\max}}{(1 - \gamma) \sqrt{(1 - (1 - (1 - \gamma) d_{\min} \alpha)^{2\tau})}}$$

$$1320 \leq 2\sqrt{\ln(2\tau|\mathcal{S}||\mathcal{A}|)} \frac{6\tau R_{\max}}{(1 - \gamma)^{\frac{3}{2}} d_{\min}^{\frac{1}{2}} \alpha^{\frac{1}{2}} \sqrt{(\sum_{j=0}^{2\tau-1} (1 - (1 - \gamma) d_{\min} \alpha)^j)}}$$

$$1321 \leq 2\sqrt{\ln(2\tau|\mathcal{S}||\mathcal{A}|)} \frac{6\tau R_{\max}}{(1 - \gamma)^{\frac{3}{2}} d_{\min}^{\frac{1}{2}} \alpha^{\frac{1}{2}} \sqrt{2\tau (1 - (1 - \gamma) d_{\min} \alpha)^{2\tau - 1}}}$$

$$1322 \leq 2\sqrt{\ln(2\tau|\mathcal{S}||\mathcal{A}|)} \frac{5\sqrt{\tau} R_{\max}}{(1 - \gamma)^{\frac{3}{2}} d_{\min}^{\frac{1}{2}} \alpha^{\frac{1}{2}}} \exp((1 - \gamma) d_{\min} \alpha (2\tau - 1))$$

$$1323 \leq 2\sqrt{\ln(2\tau|\mathcal{S}||\mathcal{A}|)} \frac{15\sqrt{\tau} R_{\max}}{(1 - \gamma)^{\frac{3}{2}} d_{\min}^{\frac{1}{2}} \alpha^{\frac{1}{2}}}.$$

The second inequality follows from  $1 - x^{2\tau} = (1 - x)(1 + x + x^2 + \dots + x^{2\tau-1})$  for  $x \in \mathbb{R}$ .  
 The third inequality follows from the fact that  $\sum_{j=0}^{2\tau-1} (1 - (1 - \gamma)d_{\min}\alpha)^j \geq \sum_{j=0}^{2\tau-1} (1 - (1 - \gamma)d_{\min}\alpha)^{2\tau-1}$ .

The second last inequality follows from the relation such that  $\exp(-2x) \leq 1 - x$  for  $x \in [0, 0.75]$ .  
 The condition  $\alpha \leq \frac{1}{2\tau}$  leads to  $\exp((1 - \gamma)d_{\min}\alpha(2\tau - 1)) \leq 3$ , yielding the last line. This completes the proof.  $\square$

Now, we bound  $\left\| \sum_{j=\tau}^k \mathbf{A}_{\mathbf{Q}^*}^{k-j} \mathbf{w}_{j,2} \right\|_{\infty}$ .

**Lemma G.3.** For  $k \geq \tau$ , we have

$$\mathbb{E} \left[ \left\| \sum_{j=\tau}^k \mathbf{A}_{\mathbf{Q}^*}^{k-j} \mathbf{w}_{j,2} \right\|_{\infty} \right] \leq \frac{8R_{\max}}{(1 - \gamma)^2 d_{\min}}.$$

*Proof.* Recalling the definition of  $\mathbf{D}_{\infty}$  and  $\mathbf{D}_{\tau}^{s_{j-\tau}, \mathbf{a}_{j-\tau}}$  in (14), we have

$$\begin{aligned} \|\mathbf{D}_{\infty} - \mathbf{D}_{\tau}^{s_{j-\tau}, \mathbf{a}_{j-\tau}}\|_{\infty} &= \max_{s, \mathbf{a} \in \mathcal{S} \times \mathcal{A}} |[(\mathbf{e}_{s_{j-\tau}} \otimes \mathbf{e}_{\mathbf{a}_{j-\tau}})^{\top} \mathbf{P}^{\tau}]^{\top}]_{s, \mathbf{a}} - [\boldsymbol{\mu}_{\infty}]_{s, \mathbf{a}}| \\ &\leq 2d_{\text{TV}}((\mathbf{e}_{s_{j-\tau}} \otimes \mathbf{e}_{\mathbf{a}_{j-\tau}})^{\top} \mathbf{P}^{\tau})^{\top}, \boldsymbol{\mu}_{\infty}) \\ &\leq 2m\rho^{\tau} \\ &\leq 2\alpha. \end{aligned}$$

The first inequality follows from the definition of the total variation distance, and the second and third inequalities follow from the definition of the mixing time in (13).

Now, we can see that

$$\begin{aligned} \|\mathbf{w}_{j,2}\|_{\infty} &= \left\| (\mathbf{D} - \mathbf{D}_{\tau}^{s_{j-\tau}, \mathbf{a}_{j-\tau}}) \frac{1}{N} \sum_{i=1}^N (\mathbf{R}^i + \gamma \mathbf{P} \Pi^{\mathbf{Q}_j^i} \mathbf{Q}_j^i - \mathbf{Q}_j^i) \right\|_{\infty} \\ &\leq \frac{1}{N} \|\mathbf{D} - \mathbf{D}_{\tau}^{s_{j-\tau}, \mathbf{a}_{j-\tau}}\|_{\infty} \left\| \sum_{i=1}^N \mathbf{R}^i + \gamma \mathbf{P} \Pi^{\mathbf{Q}_j^i} \mathbf{Q}_j^i - \mathbf{Q}_j^i \right\|_{\infty} \\ &\leq \alpha \frac{8R_{\max}}{1 - \gamma}, \end{aligned}$$

where the last inequality follows from Lemma 3.3.

Therefore, we have

$$\left\| \sum_{j=\tau}^k \mathbf{A}_{\mathbf{Q}^*}^{k-j} \mathbf{w}_{j,2} \right\|_{\infty} \leq \alpha \frac{8R_{\max}}{1 - \gamma} \sum_{j=\tau}^k (1 - \alpha(1 - \gamma)d_{\min})^{k-j} \leq \frac{8R_{\max}}{(1 - \gamma)^2 d_{\min}},$$

where the first inequality follows from Lemma C.1. This completes the proof.  $\square$

**Lemma G.4.** For  $k \geq \tau$ , we have

$$\begin{aligned} \left\| \sum_{j=\tau}^k \mathbf{A}_{\mathbf{Q}^*}^{k-j} \mathbf{w}_{j,3} \right\|_{\infty} &\leq 8 \|\bar{\mathbf{Q}}_0\|_2 \left( \sigma_2(\mathbf{W})^{\frac{k-\tau}{2}} \frac{1}{(1 - \gamma)d_{\min}\alpha} + (1 - (1 - \gamma)d_{\min}\alpha)^{\frac{k-\tau}{2}} \frac{1}{1 - \sigma_2(\mathbf{W})} \right) \\ &\quad + \frac{64R_{\max} \sqrt{N|\mathcal{S}||\mathcal{A}|}}{(1 - \gamma)^2 d_{\min} (1 - \sigma_2(\mathbf{W}))} + 4\tau \frac{2R_{\max}}{(1 - \gamma)^2 d_{\min}}. \end{aligned}$$

1404 *Proof.* Recalling the definition of  $\mathbf{w}_{j,3}$  in (21), we get

$$\begin{aligned}
1405 \mathbf{w}_{j,3} &= \delta^{\text{avg}}(o_j, \bar{\mathbf{Q}}_j) - \delta^{\text{avg}}(o_j, \bar{\mathbf{Q}}_{j-\tau}) - \Delta^{\text{avg}}(\bar{\mathbf{Q}}_j) + \Delta^{\text{avg}}(\bar{\mathbf{Q}}_{j-\tau}) \\
1406 &= \frac{1}{N} \sum_{i=1}^N \left( (\mathbf{e}_{s_j} \otimes \mathbf{e}_{\mathbf{a}_j}) \mathbf{e}_{s_{j+1}}^\top \gamma \left( \Pi^{\mathbf{Q}_j^i} \mathbf{Q}_j^i - \Pi^{\mathbf{Q}_{j-\tau}^i} \mathbf{Q}_{j-\tau}^i \right) - (\mathbf{e}_{s_j} \otimes \mathbf{e}_{\mathbf{a}_j}) (\mathbf{e}_{s_j} \otimes \mathbf{e}_{\mathbf{a}_j})^\top (\mathbf{Q}_j^i - \mathbf{Q}_{j-\tau}^i) \right) \\
1407 & \\
1408 & \\
1409 & \\
1410 & + D_\infty \frac{1}{N} \sum_{i=1}^N \left( \gamma \mathbf{P} \Pi^{\mathbf{Q}_j^i} \mathbf{Q}_j^i - \gamma \mathbf{P} \Pi^{\mathbf{Q}_{j-\tau}^i} \mathbf{Q}_{j-\tau}^i + \mathbf{Q}_j^i - \mathbf{Q}_{j-\tau}^i \right). \\
1411 & \\
1412 & \\
1413 &
\end{aligned}$$

1414 Taking infinity norm, we get

$$\begin{aligned}
1415 \|\mathbf{w}_{j,3}\|_\infty &\leq \frac{1}{N} \sum_{i=1}^N 2 \|\mathbf{Q}_j^i - \mathbf{Q}_{j-\tau}^i\|_\infty + \frac{d_{\max}}{N} \sum_{i=1}^N 2 \|\mathbf{Q}_j^i - \mathbf{Q}_{j-\tau}^i\|_\infty \\
1416 & \\
1417 & \\
1418 & \leq \frac{4}{N} \sum_{i=1}^N \left( \|\mathbf{Q}_j^i - \mathbf{Q}_j^{\text{avg}}\|_\infty + \|\mathbf{Q}_j^{\text{avg}} - \mathbf{Q}_{j-\tau}^{\text{avg}}\|_\infty + \|\mathbf{Q}_{j-\tau}^{\text{avg}} - \mathbf{Q}_{j-\tau}^i\|_\infty \right) \\
1419 & \\
1420 & \\
1421 & \leq 4 \|\Theta \bar{\mathbf{Q}}_j\|_\infty + 4 \|\Theta \bar{\mathbf{Q}}_{j-\tau}\|_\infty + 4 \|\mathbf{Q}_j^{\text{avg}} - \mathbf{Q}_{j-\tau}^{\text{avg}}\|_\infty. \quad (23) \\
1422 &
\end{aligned}$$

1423 The first inequality follows from the non-expansive property of max-operator. The second  
1424 inequality follows from the triangle inequality. The term  $\|\mathbf{Q}_j^{\text{avg}} - \mathbf{Q}_{j-\tau}^{\text{avg}}\|_\infty$  can be bounded  
1425 as follows:

$$\begin{aligned}
1426 \|\mathbf{Q}_j^{\text{avg}} - \mathbf{Q}_{j-\tau}^{\text{avg}}\|_\infty &\leq \sum_{t=j-\tau}^{j-1} \|\mathbf{Q}_{t+1}^{\text{avg}} - \mathbf{Q}_t^{\text{avg}}\|_\infty \\
1427 & \\
1428 & \leq \alpha \sum_{t=j-\tau}^{j-1} \frac{1}{N} \sum_{i=1}^N \left\| \mathbf{e}_{s_t, \mathbf{a}_t} \left( r_t^i + \gamma \max_{\mathbf{a} \in \mathcal{A}} \mathbf{Q}_t^i(s_{t+1}, \mathbf{a}) - \mathbf{Q}_t^i(s_t, \mathbf{a}_t) \right) \right\|_\infty \\
1429 & \\
1430 & \\
1431 & \leq \alpha \tau \frac{2R_{\max}}{1-\gamma}. \quad (24) \\
1432 & \\
1433 & \\
1434 &
\end{aligned}$$

1435 The second inequality follows from (2). The last inequality follows from Lemma 3.3.

1436 Applying the result in Theorem 3.4 together with (24) to (23), we get

$$\|\mathbf{w}_{j,3}\|_\infty \leq 8\sigma_2(\mathbf{W})^{j-\tau} \|\bar{\mathbf{Q}}_0\|_2 + 8\alpha \frac{8R_{\max}}{1-\gamma} \frac{\sqrt{N|\mathcal{S}||\mathcal{A}|}}{1-\sigma_2(\mathbf{W})} + 4\alpha\tau \frac{2R_{\max}}{1-\gamma}. \quad (25)$$

1441 Now, we are ready to derive our desired statement:

$$\begin{aligned}
1442 & \\
1443 & \left\| \sum_{j=\tau}^k \mathbf{A}_{\mathbf{Q}^*}^{k-j} \mathbf{w}_{j,3} \right\|_\infty \\
1444 & \\
1445 & \\
1446 & \leq \sum_{j=\tau}^k (1 - (1-\gamma)d_{\min}\alpha)^{k-j} \left( 8\sigma_2(\mathbf{W})^{j-\tau} \|\bar{\mathbf{Q}}_0\|_2 + 8\alpha \frac{8R_{\max}}{1-\gamma} \frac{\sqrt{N|\mathcal{S}||\mathcal{A}|}}{1-\sigma_2(\mathbf{W})} + 4\alpha\tau \frac{2R_{\max}}{1-\gamma} \right) \\
1447 & \\
1448 & \leq 8 \|\bar{\mathbf{Q}}_0\|_2 \left( \sigma_2(\mathbf{W})^{\frac{k-\tau}{2}} \frac{1}{(1-\gamma)d_{\min}\alpha} + (1 - (1-\gamma)d_{\min}\alpha)^{\frac{k-\tau}{2}} \frac{1}{1-\sigma_2(\mathbf{W})} \right) \\
1449 & \\
1450 & + \frac{64R_{\max}\sqrt{N|\mathcal{S}||\mathcal{A}|}}{(1-\gamma)^2 d_{\min}(1-\sigma_2(\mathbf{W}))} + 4\tau \frac{2R_{\max}}{(1-\gamma)^2 d_{\min}}. \\
1451 & \\
1452 & \\
1453 & \\
1454 &
\end{aligned}$$

1455 The first inequality follows from Lemma C.1 and (25). The last inequality follows from  
1456 Lemma C.3. This completes the proof.  $\square$

1457 Now, collecting the results we have the following bound for the lower comparison system:

**Proposition G.5.** For  $k \in \mathbb{N}$ , and  $\alpha \leq \min \{ \min_{i \in [N]} [\mathbf{W}]_{ii}, \frac{1}{2\tau} \}$ , we have

$$\begin{aligned} \mathbb{E} \left[ \left\| \tilde{\mathbf{Q}}_{k+1}^{\text{avg},l} \right\|_{\infty} \right] &= \tilde{\mathcal{O}} \left( (1 - (1 - \gamma)d_{\min}\alpha)^{\frac{k-\tau}{2}} + \sigma_2(\mathbf{W})^{\frac{k-\tau}{2}} \right) \\ &+ \tilde{\mathcal{O}} \left( \alpha^{\frac{1}{2}} \frac{\sqrt{\tau}R_{\max}}{(1-\gamma)^{\frac{3}{2}}d_{\min}^{\frac{1}{2}}} + \alpha \frac{R_{\max}\sqrt{N|\mathcal{S}||\mathcal{A}|}}{(1-\gamma)^2d_{\min}(1-\sigma_2(\mathbf{W}))} \right). \end{aligned}$$

*Proof.* Collecting the results in Lemma G.2, Lemma G.3, Lemma G.4, and Lemma 3.5, we can bound (22) as follows:

$$\begin{aligned} &\mathbb{E} \left[ \left\| \tilde{\mathbf{Q}}_{k+1}^{\text{avg},l} \right\|_{\infty} \right] \\ &\leq (1 - (1 - \gamma)d_{\min}\alpha)^{k-\tau+1} \mathbb{E} \left[ \left\| \tilde{\mathbf{Q}}_{\tau}^{\text{avg},l} \right\|_{\infty} \right] \\ &+ 2\alpha^{\frac{1}{2}} \sqrt{\ln(2\tau|\mathcal{S}||\mathcal{A}|)} \frac{15\sqrt{\tau}R_{\max}}{(1-\gamma)^{\frac{3}{2}}d_{\min}^{\frac{1}{2}}} \\ &+ \alpha \frac{8R_{\max}}{(1-\gamma)^2d_{\min}} \\ &+ 8 \left\| \tilde{\mathbf{Q}}_0 \right\|_2 \left( \sigma_2(\mathbf{W})^{\frac{k-\tau}{2}} \frac{1}{(1-\gamma)d_{\min}} + (1 - (1 - \gamma)d_{\min}\alpha)^{\frac{k-\tau}{2}} \frac{\alpha}{1 - \sigma_2(\mathbf{W})} \right) \\ &+ \alpha \frac{64R_{\max}\sqrt{N|\mathcal{S}||\mathcal{A}|}}{(1-\gamma)^2d_{\min}(1-\sigma_2(\mathbf{W}))} + 4\alpha\tau \frac{2R_{\max}}{(1-\gamma)^2d_{\min}} \\ &+ \gamma d_{\max} \left\| \Theta \tilde{\mathbf{Q}}_0 \right\|_2 \left( (1 - (1 - \gamma)d_{\min}\alpha)^{\frac{k-\tau}{2}} \frac{\alpha}{1 - \sigma_2(\mathbf{W})} + \sigma_2(\mathbf{W})^{\frac{k-\tau}{2}} \frac{1}{(1-\gamma)d_{\min}} \right) \\ &+ \alpha\gamma d_{\max} \frac{8R_{\max}\sqrt{N|\mathcal{S}||\mathcal{A}|}}{(1-\gamma)^2d_{\min}(1-\sigma_2(\mathbf{W}))}. \end{aligned}$$

That is,

$$\begin{aligned} \mathbb{E} \left[ \left\| \tilde{\mathbf{Q}}_{k+1}^{\text{avg},l} \right\|_{\infty} \right] &= \tilde{\mathcal{O}} \left( (1 - (1 - \gamma)d_{\min}\alpha)^{\frac{k-\tau}{2}} + \sigma_2(\mathbf{W})^{\frac{k-\tau}{2}} \right) \\ &+ \tilde{\mathcal{O}} \left( \alpha^{\frac{1}{2}} \frac{\sqrt{\tau}R_{\max}}{(1-\gamma)^{\frac{3}{2}}d_{\min}^{\frac{1}{2}}} + \alpha \frac{R_{\max}\sqrt{N|\mathcal{S}||\mathcal{A}|}}{(1-\gamma)^2d_{\min}(1-\sigma_2(\mathbf{W}))} \right). \end{aligned}$$

This completes the proof.  $\square$

The rest of the proof follows the same logic in Section 3. We consider the upper comparison system, and derive the convergence rate of  $\mathbf{Q}_k^{\text{avg},u} - \mathbf{Q}_k^{\text{avg},l}$ . As can be seen in (18), if we subtract  $\mathbf{Q}_{k+1}^{\text{avg},l}$  from  $\mathbf{Q}_{k+1}^{\text{avg},u}$ ,  $\epsilon_k^{\text{avg}}$  and  $\mathbf{E}_k$  are eliminated. Therefore, we can follow the same lines of the proof in Proposition F.3:

**Proposition G.6.** For  $k \in \mathbb{N}$ , and  $\alpha \leq \min \{ \min_{i \in [N]} [\mathbf{W}]_{ii}, \frac{1}{2\tau} \}$ , we have

$$\begin{aligned} \mathbb{E} \left[ \left\| \mathbf{Q}_{k+1}^{\text{avg},u} - \mathbf{Q}_{k+1}^{\text{avg},l} \right\|_{\infty} \right] &= \tilde{\mathcal{O}} \left( (1 - \alpha(1 - \gamma)d_{\min})^{\frac{k-\tau}{2}} + \sigma_2(\mathbf{W})^{\frac{k-\tau}{4}} \right) \\ &+ \tilde{\mathcal{O}} \left( \alpha^{\frac{1}{2}} d_{\max} \frac{\sqrt{\tau}R_{\max}}{(1-\gamma)^{\frac{5}{2}}d_{\min}^{\frac{3}{2}}} + \alpha \frac{d_{\max}R_{\max}\sqrt{N|\mathcal{S}||\mathcal{A}|}}{(1-\gamma)^3d_{\min}^2(1-\sigma_2(\mathbf{W}))} \right). \end{aligned}$$

*Proof.* As from the proof of Proposition F.3, we have

$$\begin{aligned} \mathbb{E} \left[ \left\| \mathbf{Q}_{k+1}^{\text{avg},u} - \mathbf{Q}_{k+1}^{\text{avg},l} \right\|_{\infty} \right] &\leq (1 - \alpha(1 - \gamma)d_{\min})^{k-\tau+1} \mathbb{E} \left[ \left\| \mathbf{Q}_{\tau}^{\text{avg},u} - \mathbf{Q}_{\tau}^{\text{avg},l} \right\|_{\infty} \right] \\ &+ 2\alpha\gamma d_{\max} \underbrace{\sum_{i=\tau}^k (1 - \alpha(1 - \gamma)d_{\min})^{k-i} \mathbb{E} \left[ \left\| \tilde{\mathbf{Q}}_i^{\text{avg},l} \right\|_{\infty} \right]}_{(*)}. \quad (26) \end{aligned}$$

We will use Proposition G.5 to bound  $(\star)$  in the above inequality. We have

$$\begin{aligned}
& \sum_{i=\tau}^k (1 - \alpha(1 - \gamma)d_{\min})^{k-i} \mathbb{E} \left[ \left\| \tilde{\mathbf{Q}}_i^{\text{avg},l} \right\|_{\infty} \right] \\
&= \tilde{\mathcal{O}} \left( \sum_{i=\tau}^k (1 - \alpha(1 - \gamma)d_{\min})^{k-i} + (1 - \alpha(1 - \gamma)d_{\min})^{k-i} \sigma_2(\mathbf{W})^{\frac{k-\tau}{2}} \right) \\
&+ \tilde{\mathcal{O}} \left( \alpha^{-\frac{1}{2}} \frac{\sqrt{\tau} R_{\max}}{(1 - \gamma)^{\frac{5}{2}} d_{\min}^{\frac{3}{2}}} + \frac{R_{\max} \sqrt{N|\mathcal{S}||\mathcal{A}|}}{(1 - \gamma)^3 d_{\min}^2 (1 - \sigma_2(\mathbf{W}))} \right) \\
&= \tilde{\mathcal{O}} \left( (1 - \alpha(1 - \gamma)d_{\min})^{\frac{k-\tau}{2}} + \sigma_2(\mathbf{W})^{\frac{k-\tau}{4}} \right) \\
&+ \tilde{\mathcal{O}} \left( \alpha^{-\frac{1}{2}} \frac{\sqrt{\tau} R_{\max}}{(1 - \gamma)^{\frac{5}{2}} d_{\min}^{\frac{3}{2}}} + \frac{R_{\max} \sqrt{N|\mathcal{S}||\mathcal{A}|}}{(1 - \gamma)^3 d_{\min}^2 (1 - \sigma_2(\mathbf{W}))} \right).
\end{aligned}$$

The last inequality follows from Lemma C.3. Applying this result to (26), we get

$$\begin{aligned}
\mathbb{E} \left[ \left\| \mathbf{Q}_{k+1}^{\text{avg},u} - \mathbf{Q}_{k+1}^{\text{avg},l} \right\|_{\infty} \right] &= \tilde{\mathcal{O}} \left( (1 - \alpha(1 - \gamma)d_{\min})^{\frac{k-\tau}{2}} + \sigma_2(\mathbf{W})^{\frac{k-\tau}{4}} \right) \\
&+ \tilde{\mathcal{O}} \left( \alpha^{\frac{1}{2}} d_{\max} \frac{\sqrt{\tau} R_{\max}}{(1 - \gamma)^{\frac{5}{2}} d_{\min}^{\frac{3}{2}}} + \alpha \frac{d_{\max} R_{\max} \sqrt{N|\mathcal{S}||\mathcal{A}|}}{(1 - \gamma)^3 d_{\min}^2 (1 - \sigma_2(\mathbf{W}))} \right).
\end{aligned}$$

This completes the proof.  $\square$

Now, we are ready to provide the optimality error under Markovian observation model:

**Theorem G.7.** For  $k \geq \tau$ , and  $\alpha \leq \min \{ \min_{i \in [N]} [\mathbf{W}]_{ii}, \frac{1}{2\tau} \}$ , we have

$$\begin{aligned}
\mathbb{E} \left[ \left\| \mathbf{Q}_k^{\text{avg}} - \mathbf{Q}^* \right\|_{\infty} \right] &= \tilde{\mathcal{O}} \left( (1 - \alpha(1 - \gamma)d_{\min})^{\frac{k-\tau}{2}} + \sigma_2(\mathbf{W})^{\frac{k-\tau}{4}} \right) \\
&+ \tilde{\mathcal{O}} \left( \alpha^{\frac{1}{2}} d_{\max} \frac{\sqrt{\tau} R_{\max}}{(1 - \gamma)^{\frac{5}{2}} d_{\min}^{\frac{3}{2}}} + \alpha \frac{R_{\max} d_{\max} \sqrt{N|\mathcal{S}||\mathcal{A}|}}{(1 - \gamma)^3 d_{\min}^2 (1 - \sigma_2(\mathbf{W}))} \right).
\end{aligned}$$

*Proof.* The proof follows the same logic as in Theorem 3.6 using the fact that  $\tilde{\mathbf{Q}}_k^{\text{avg},l} \leq \tilde{\mathbf{Q}}_k^{\text{avg}} \leq \tilde{\mathbf{Q}}_k^{\text{avg},u}$ . Therefore, we omit the proof.  $\square$

## G.2 PROOF OF THEOREM 4.1

*Proof.* The proof follows the same line as in Theorem 3.7. From Theorem 3.4 and Theorem G.7, we get

$$\begin{aligned}
\mathbb{E} \left[ \left\| \bar{\mathbf{Q}}_k - \mathbf{1}_N \otimes \mathbf{Q}^* \right\|_{\infty} \right] &\leq \mathbb{E} \left[ \left\| \bar{\mathbf{Q}}_k - \mathbf{1}_N \otimes \mathbf{Q}_k^{\text{avg}} \right\|_{\infty} \right] + \mathbb{E} \left[ \left\| \mathbf{Q}_k^{\text{avg}} - \mathbf{Q}^* \right\|_{\infty} \right] \\
&= \tilde{\mathcal{O}} \left( \sigma_2(\mathbf{W})^k + \alpha \frac{R_{\max} \sqrt{N|\mathcal{S}||\mathcal{A}|}}{(1 - \gamma)(1 - \sigma_2(\mathbf{W}))} \right) \\
&+ \tilde{\mathcal{O}} \left( (1 - \alpha(1 - \gamma)d_{\min})^{\frac{k-\tau}{2}} + \sigma_2(\mathbf{W})^{\frac{k-\tau}{4}} \right) \\
&+ \tilde{\mathcal{O}} \left( \alpha^{\frac{1}{2}} d_{\max} \frac{\sqrt{\tau} R_{\max}}{(1 - \gamma)^{\frac{5}{2}} d_{\min}^{\frac{3}{2}}} + \alpha \frac{R_{\max} d_{\max} \sqrt{N|\mathcal{S}||\mathcal{A}|}}{(1 - \gamma)^3 d_{\min}^2 (1 - \sigma_2(\mathbf{W}))} \right) \\
&= \tilde{\mathcal{O}} \left( (1 - \alpha(1 - \gamma)d_{\min})^{\frac{k-\tau}{2}} + \sigma_2(\mathbf{W})^{\frac{k-\tau}{4}} \right) \\
&+ \tilde{\mathcal{O}} \left( \alpha^{\frac{1}{2}} \frac{d_{\max} \sqrt{\tau} R_{\max}}{(1 - \gamma)^{\frac{5}{2}} d_{\min}^{\frac{3}{2}}} + \alpha \frac{R_{\max} d_{\max} \sqrt{|\mathcal{S}||\mathcal{A}|}}{(1 - \gamma)^3 d_{\min}^2 (1 - \sigma_2(\mathbf{W}))} \right).
\end{aligned}$$

This completes the proof.  $\square$

### 1566 G.3 PROOF OF COROLLARY 4.2

1567 *Proof.* For  $\mathbb{E} [\|\bar{\mathbf{Q}}_k - \mathbf{1}_N \otimes \mathbf{Q}^*\|_\infty] \leq \epsilon$ , we bound the each terms in Theorem 4.1 with  $\frac{\epsilon}{4}$ .  
1568 We require

$$1570 \alpha^{\frac{1}{2}} d_{\max} \frac{\sqrt{\tau} R_{\max}}{(1-\gamma)^{\frac{5}{2}} d_{\min}^{\frac{3}{2}}} \leq \epsilon/4,$$

1571 which is satisfied if

$$1572 \alpha = \tilde{\mathcal{O}} \left( \frac{\epsilon^2}{\ln(\frac{1}{\epsilon^2})} \frac{(1-\gamma)^5 d_{\min}^3}{t_{\text{mix}} d_{\max}^2} \right),$$

1573 where  $\tau$  is bounded by  $t_{\text{mix}}$  by Lemma C.7 in the Appendix. Likewise, bounding

1574  $\alpha \frac{R_{\max} d_{\max} \sqrt{|\mathcal{S}||\mathcal{A}|}}{(1-\gamma)^3 d_{\min}^2 (1-\sigma_2(\mathbf{W}))} \leq \epsilon/4$ , together with the above condition, we require

$$1575 \alpha = \tilde{\mathcal{O}} \left( \min \left\{ \frac{\epsilon^2}{\ln(\frac{1}{\epsilon^2})} \frac{(1-\gamma)^5 d_{\min}^3}{d_{\max}^2 t_{\text{mix}}}, \frac{\epsilon(1-\gamma)^3 d_{\min}^2 (1-\sigma_2(\mathbf{W}))}{d_{\max} \sqrt{|\mathcal{S}||\mathcal{A}|}} \right\} \right).$$

1576 Furthermore bounding the terms  $(1-\alpha(1-\gamma)d_{\min})^{\frac{k-\tau}{2}} + \sigma_2(\mathbf{W})^{\frac{k-\tau}{4}}$  in Theorem 4.1 with  
1577  $\frac{\epsilon}{4}$ , respectively, we require,

$$1578 k \geq \tilde{\mathcal{O}} \left( \min \left\{ \frac{\ln^2(\frac{1}{\epsilon^2})}{\epsilon^2} \frac{t_{\text{mix}} d_{\max}^2}{(1-\gamma)^6 d_{\min}^4}, \frac{\ln(\frac{1}{\epsilon})}{\epsilon} \frac{d_{\max} \sqrt{|\mathcal{S}||\mathcal{A}|}}{(1-\gamma)^4 d_{\min}^2 (1-\sigma_2(\mathbf{W}))} \right\} + \ln \left( \frac{1}{\epsilon} \right) / \ln \left( \frac{1}{\sigma_2(\mathbf{W})} \right) \right).$$

1579 This completes the proof.  $\square$

## 1591 H APPENDIX : EXAMPLES MENTIONED IN SECTION 5

1592 Let us provide an example where the condition (15) used in Heredia et al. (2020) is not met  
1593 in tabular MDP. Since the condition only depends on the state-action distribution, consider  
1594 an MDP that consists of two states and single action, where  $\mathcal{S} := \{1, 2\}$  and  $\mathcal{A} := \{1\}$   
1595 with  $d(1,1) = 0.1$ ,  $d(2,1) = 0.9$ , and  $\gamma = 0.5$ . Then,  $d_{\min} = 0.1$  and  $d_{\max} = 0.9$ , then  
1596  $d_{\min} < \gamma^2 d_{\max}$  which contradicts the condition in (15).

1597 Next, we provide an MDP where the condition (16) required in Zeng et al. (2022b) is not  
1598 met:

$$1600 \mathbf{P} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad \mathbf{R} = \begin{bmatrix} 0 \\ 0.1 \\ 0 \\ 0.1 \end{bmatrix}, \quad [\mathbf{D}]_{s,a} = \frac{1}{4}, \quad \forall s, a \in \mathcal{S} \times \mathcal{A}.$$

1601 Letting  $\gamma = 0.99$ , we can check that  $\mathbf{Q}^* = \begin{bmatrix} 9.9 \\ 10 \\ 9.9 \\ 10 \end{bmatrix}$  and  $\mathbf{\Pi}^{\mathbf{Q}^*} = \begin{bmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$ . Consider

1602  $\mathbf{Q} = \begin{bmatrix} 12 \\ 10 \\ 11 \\ 10 \end{bmatrix}$ . Then, we have

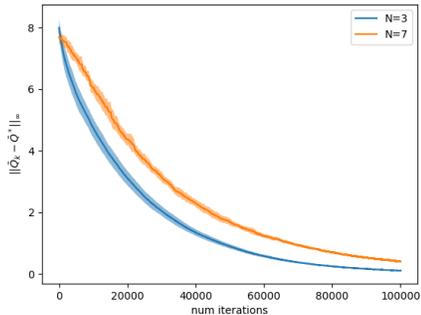
$$1603 (\gamma \mathbf{D} \mathbf{P} (\mathbf{\Pi}^{\mathbf{Q}} \mathbf{Q} - \mathbf{\Pi}^{\mathbf{Q}^*} \mathbf{Q}^*) - \mathbf{D} (\mathbf{Q} - \mathbf{Q}^*))^\top (\mathbf{Q} - \mathbf{Q}^*) = 0.179,$$

1604 which is contradiction to the condition in (16).

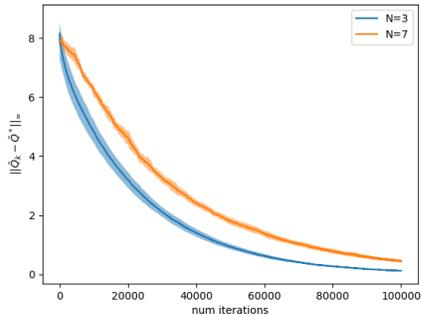
## 1616 I EXPERIMENTS

1617 The experiment used the MDP where and  $|\mathcal{A}_i| = 2$  for each agent  $i \in [N]$  where  $N$  denotes  
1618 the number of agents. For each run, we have randomly generated the transition and reward  
1619

1620  
1621  
1622  
1623  
1624  
1625  
1626  
1627  
1628  
1629  
1630  
1631



(a) Experiment under ring graph



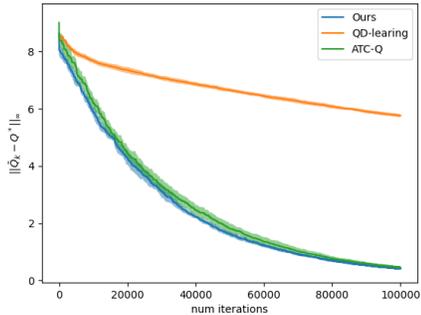
(b) Experiment under star graph

Figure 1:  $\alpha = 0.1$ . The result was averaged over five runs.

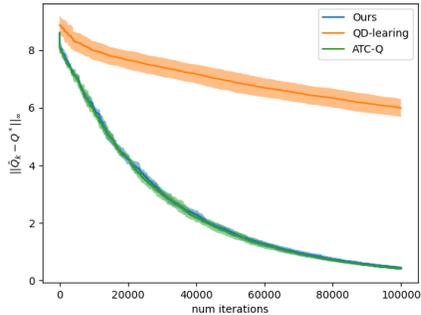
1632  
1633  
1634  
1635  
1636  
1637

matrix. Each elements were chosen uniformly random between zero and one, and for the transition matrix, each row is normalized to be a probability distribution. We can see that the distributed Q-learning algorithm converges to close to  $Q^*$ , where the constant bias is induced by using the constant step-size. As number of agents increase, the convergence rate becomes slower.

1642  
1643  
1644  
1645  
1646  
1647  
1648  
1649  
1650  
1651  
1652  
1653

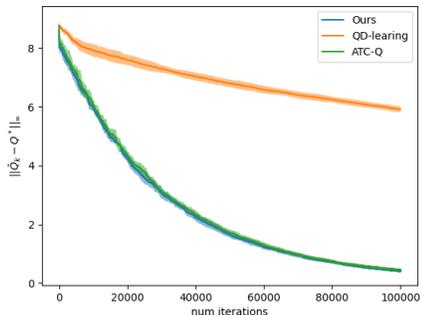


(a) Ring graph,  $|\mathcal{S}| = 2$

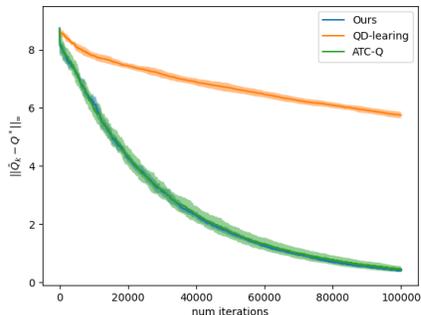


(b) Ring graph,  $|\mathcal{S}| = 5$

1654  
1655  
1656  
1657  
1658  
1659  
1660  
1661  
1662  
1663  
1664  
1665  
1666



(c) Star graph,  $|\mathcal{S}| = 2$



(d) Star graph,  $|\mathcal{S}| = 5$

Figure 2:  $\alpha = 0.1$ . The result was averaged over five runs and  $N = 7$

1667  
1668  
1669  
1670  
1671  
1672  
1673

The Figure 2 shows comparison with QD-learning developed in Kar et al. (2013). QD-learning uses a two-time scale approach, and therefore we have set the two-step-sizes as 0.1 and 0.01, where the faster time-scale matches the single-step-size of distributed Q-learning. As in

the figure, distributed Q-learning shows faster convergence rate compared to QD-learning. ATC-Q refers to the adapt-then-combine scheme in Wang et al. (2022).

## J APPENDIX : PSEUDO CODE

---

**Algorithm 1** Distributed Q-learning : i.i.d. observation model

---

**Require:** Initialize  $\mathbf{Q}_0^i \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$  such that  $\|\mathbf{Q}_0^i\| \leq \frac{R_{\max}}{1-\gamma}$  for all  $i \in [N]$ , and  $0 \leq \alpha \leq$

$\min_{i \in [N]} [\mathbf{W}]_{ii}$ .

**for**  $k = 0, 1, \dots$ , **do**

Observe  $s_k, \mathbf{a}_k \sim d(\cdot, \cdot), s'_k \sim \mathcal{P}(s_k, \mathbf{a}_k, \cdot)$ .

**for**  $i = 1, 2, \dots, N$  **do**

Update as follows:

$$\mathbf{Q}_{k+1}^i(s_k, \mathbf{a}_k) = \sum_{j \in \mathcal{N}_i} [\mathbf{W}]_{ij} \mathbf{Q}_k^j(s_k, \mathbf{a}_k) + \alpha \left( r_{k+1}^i + \gamma \max_{\mathbf{a} \in \mathcal{A}} \mathbf{Q}_k^i(s'_k, \mathbf{a}) - \mathbf{Q}_k^i(s_k, \mathbf{a}_k) \right).$$

**end for**

**end for**

---

**Algorithm 2** Distributed Q-learning : Markovian observation model

---

**Require:** Initialize  $\mathbf{Q}_0^i \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$  such that  $\|\mathbf{Q}_0^i\| \leq \frac{R_{\max}}{1-\gamma}$  for all  $i \in [N]$ , and  $0 \leq \alpha \leq$

$\min \left\{ \min_{i \in [N]} [\mathbf{W}]_{ii}, \frac{1}{2\tau} \right\}$ .

Observe  $s_0, \mathbf{a}_0 \sim \boldsymbol{\mu}_0$ .

**for**  $k = 0, 1, \dots$ , **do**

Observe  $s_{k+1} \sim \mathcal{P}(s_k, \mathbf{a}_k, \cdot)$  and  $\mathbf{a}_{k+1} \sim \beta(\cdot | s_k)$ .

**for**  $i = 1, 2, \dots, N$  **do**

Update as follows:

$$\mathbf{Q}_{k+1}^i(s_k, \mathbf{a}_k) = \sum_{j \in \mathcal{N}_i} [\mathbf{W}]_{ij} \mathbf{Q}_k^j(s_k, \mathbf{a}_k) + \alpha \left( r_{k+1}^i + \gamma \max_{\mathbf{a} \in \mathcal{A}} \mathbf{Q}_k^i(s_{k+1}, \mathbf{a}) - \mathbf{Q}_k^i(s_k, \mathbf{a}_k) \right).$$

**end for**

**end for**

---