Multi-Agent Reinforcement Learning in Stochastic Networked Systems

Yiheng Lin
CMS, Caltech
yihengl@caltech.edu

Guannan Qu
CMS, Caltech
gqu@caltech.edu

Longbo Huang
IIIS, Tsinghua University
longbohuang@tsinghua.edu.cn

Adam Wierman
CMS, Caltech
adamw@caltech.edu

Abstract

We study multi-agent reinforcement learning (MARL) in a stochastic network of agents. The objective is to find localized policies that maximize the (discounted) global reward. In general, scalability is a challenge in this setting because the size of the global state/action space can be exponential in the number of agents. Scalable algorithms are only known in cases where dependencies are static, fixed and local, e.g., between neighbors in a fixed, time-invariant underlying graph. In this work, we propose a Scalable Actor Critic framework that applies in settings where the dependencies can be non-local and stochastic, and provide a finite-time error bound that shows how the convergence rate depends on the speed of information spread in the network. Additionally, as a byproduct of our analysis, we obtain novel finite-time convergence results for a general stochastic approximation scheme and for temporal difference learning with state aggregation, which apply beyond the setting of MARL in networked systems.

1 Introduction

Multi-Agent Reinforcement Learning (MARL) has achieved impressive performance in a wide array of applications including multi-player game play [42, 31], multi-robot systems [13], and autonomous driving [25]. In comparison to single-agent reinforcement learning (RL), MARL poses many challenges, chief of which is scalability [57]. Even if each agent's local state/action spaces are small, the size of the global state/action space can be large, potentially exponentially large in the number of agents, which renders many RL algorithms such as Q-learning not applicable.

A promising approach for addressing the scalability challenge that has received attention in recent years is to exploit application-specific structures, e.g., [18, 35, 38]. A particularly important example of such a structure is a networked structure, e.g., applications in multi-agent networked systems such as social networks [7, 27], communication networks [60, 51], queueing networks [34], and smart transportation networks [59]. In these networked systems, it is often possible to exploit *static, local* dependency structures [16, 17, 1, 32], e.g., the fact that agents only interact with a fixed set of neighboring agents throughout the game. This sort of dependency structure often leads to scalable, distributed algorithms for optimization and control [16, 1, 32], and has proven effective for designing scalable and distributed MARL algorithms, e.g. [35, 38].

This work was supported by NSF grants CNS-2106403 and NGSDI-2105648, with additional support from Amazon AWS, PIMCO, and the Resnick Sustainability Insitute. Yiheng Lin was supported by Kortschak Scholars program. The work of Longbo Huang was supported by the Technology and Innovation Major Project of the Ministry of Science and Technology of China under Grants 2020AAA0108400 and 2020AAA0108403.

35th Conference on Neural Information Processing Systems (NeurIPS 2021), Sydney, Australia.

However, many real-world networked systems have inherently *time-varying, non-local* dependencies. For example, in the context of wireless networks, each node can send packets to other nodes within a fixed transmission range. However, the interference range, in which other nodes can interfere the transmission, can be larger than the transmission range [53]. As a result, due to potential collisions, the local reward of each node not only depends on its own local state/action, but also depends on the actions of other nodes within the interference range, which may be more than one-hop away. In addition, a node may be able to observe other nodes' local states before picking its local action [33]. Things become even more complex when mobility and stochastic network conditions are considered. These lead to dependencies that are both stochastic and non-local. Although one can always fix and localize the dependence model, this leads to considerably reduced performance. Beyond wireless networks, similar stochastic and non-local dependencies exists in epidemics [30], social networks [7, 27], and smart transportation networks [59].

A challenging open question in MARL is to understand how to obtain algorithms that are scalable in settings where the dependencies are stochastic and non-local. Prior work considers exclusively static and local dependencies, e.g., [35, 38]. It is clear that hardness results apply when the dependencies are too general [24]. Further, results in the static, local setting to this point rely on the concept of exponential decay [35, 16], meaning the agents' impact on each other decays exponentially in their graph distance. This property relies on the fact that the dependencies are purely local and static, and it is not clear whether it can still be exploited when the interactions are more general. This motivates an important open question: *Is it possible to design scalable algorithms for stochastic, non-local networked MARL?*

Contributions. In this paper, we introduce a class of stochastic, non-local dependency structures where every agent is allowed to depend on a random subset of agents. In this context, we propose and analyze a Scalable Actor Critic (SAC) algorithm that provably learns a near-optimal local policy in a scalable manner (Theorem 2.5). This result represents the *first* provably scalable method for stochastic networked MARL. Key to our approach is that the class of dependencies we consider leads to a μ-decay property (Definition 2.1). This property generalizes the exponential decay property underlying recent results such as [35, 16], which does not apply to stochastic non-local dependencies, and enables the design of an efficient and scalable algorithm for settings with stochastic, non-local dependencies. Our analysis of the algorithm reveals an important trade-off: as deeper interactions appear more frequently, the "information" can spread more quickly from one part of the network to another, which leads to the efficiency of the proposed method to degrade. This is to be expected, as when the agents are allowed to interact globally, the problem becomes a single-agent tabular Q-learning problem with an exponentially large state space, which is known to be intractable since the sample complexity is polynomial in the size of the state/action space [12, 24].

The key technical result underlying our analysis of the Scalable Actor Critic algorithm is a finite-time analysis of a general stochastic approximation scheme featuring infinite-norm contraction and state aggregation (Theorem 3.1). We apply this result to networked MARL using the local neighborhood of each agent to provide state aggregation (SA). This result also applies beyond MARL. Specifically, we show that it yields finite-time bounds on Temporal Difference (TD)/Q learning with state aggregation (Theorem 3.2). To the best of our knowledge the resulting bound is the first finite-time bound on asynchronous Q-learning with state aggregation. Additionally, it yields a novel analysis for TD-learning with state aggregation (the first error bound in the infinity norm) that sheds new insight into how the error depends on the quality of state abstraction. These two results are important contributions in their own right. Due to space constraints, we discuss asynchronous Q-learning with state aggregation in Appendix D.4.

Related literature. The prior work that is most related to our paper is [38], which also studies MARL in a networked setting. The key difference is that we allow the dependency structure among agents to be non-local and stochastic, while [38] requires the dependency structure to be local and static. The generality of setting means techniques from [38] do not apply and adds considerable complexity to the proof in two aspects. First, instead of analyzing the algorithm directly like [38], we derive a finite-time error bound for TD learning with state aggregation (Section 3.1 and 3.2), and then establish its connection with the algorithm (Section 2.3). Second, we need a more general decay property (Definition 2.1) than the exponential one used in [38]. Defining and establishing this general decay property for the non-local and stochastic setting is highly non-trivial (Section 2.1).

More broadly, MARL has received considerable attention in recent years, see [57] for a survey. The line of work most relevant to the current paper focuses on cooperative MARL. In the cooperative setting, each agent can decide its local actions but share a common global state with other agents. The objective is to maximize a global reward by working cooperatively. Notable examples of this approach include [6, 10] and the references therein. In contrast, we study a situation where each agent has its own state that it acts upon. Despite the differences, like our situation, cooperative MARL problems still face scalability issues since the joint-action space is exponentially large. A variety of methods have been proposed to deal with this, including independent learners [8, 29], where each agent employs a single-agent RL policy. Function approximation is another approach that can significantly reduce the space/computational complexity. One can use linear functions [58] or neural networks [28] in the approximation. A limitation of these approaches is the lack of theoretical guarantees on the approximation error. In contrast, our technique not only reduces the space/computational complexity significantly, but also has theoretical guarantees on the performance loss in settings with stochastic and non-local dependencies.

The mean-field approach [45, 56, 19] provides another way to address the scalability issue, but under very different settings compared to ours. Specifically, the mean-field approach typically assumes homogeneous agents with identical local state/action space and policies, and each agent depends on other agents through their population or "mean" behavior. In contrast, our approach considers a local-interaction model, where there is an underlying graph and each agent depends on neighboring agents in the graph. Further, our approach allows heterogeneous agents, which means that the local state/action spaces and policies can differ among the agents.

Another related line of work uses centralized training with decentralized execution, e.g., [28, 15], where there is a centralized coordinator that can communicate with all the agents and keep track of their experiences and policies. In contrast, our work only requires distributed training, where we constrain the scale of communication in training within the κ-hop neighborhood of each agent.

More broadly, this paper contributes to a growing literature that uses exponential decay to derive scalable algorithms for learning in networked systems. The specific form of exponential decay that we generalize is related to the idea of "correlation decay" studied in [16, 17], though their focus is on solving static combinatorial optimization problems whereas ours is on learning policies in dynamic environments. Most related to the current paper is [38], which shows an exponential decay property in a restricted networked MARL model with purely local dependencies. In contrast, we show a more general μ-decay property holds for a general form of stochastic, non-local dependencies.

The technical work in this paper contributes to the analysis of stochastic approximation (SA), which has received considerable attention over the past decade [54, 44, 11, 55]. Our work is most related to [37], which uses an asynchronous nonlinear SA to study the finite-time convergence rate for asynchronous Q-learning on a single trajectory. Beyond [37], there are many other works that use SA schemes to study TD learning and Q-learning, e.g. [44, 52, 20]. The finite-time error bound for TD learning with state aggregation in our work is most related to the asymptotic convergence limit given in [49] and the application of SA scheme to asynchronous Q-learning in [37]. Beyond these papers, other related work in the broader area of RL with state aggregation includes [26, 23, 22, 9, 43]. We add to this literature with a novel finite-time convergence bound for a general SA with state aggregation. This result, in turn, yields the first finite-time error bound in the infinity norm for both TD learning with state aggregation and Q-learning with state aggregation.

2 Networked MARL

We consider a network of agents that are associated with an underlying undirected graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N} = \{1, 2, \cdots, n\}$ denotes the set of agents and $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$ denotes the set of edges. The distance $d_{\mathcal{G}}(i, j)$ between two agents i and j is defined as the number of edges on the shortest path that connects them on graph $\mathcal{G}$. Each agent is associated with its local state $s_i \in \mathcal{S}_i$ and local action $a_i \in \mathcal{A}_i$ where $\mathcal{S}_i$ and $\mathcal{A}_i$ are finite sets. The global state/action is defined as the combination of all local states/actions, i.e., $s = (s_1, \cdots, s_n) \in \mathcal{S} := \mathcal{S}_1 \times \cdots \times \mathcal{S}_n$, and $a = (a_1, \cdots, a_n) \in \mathcal{A} := \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$. We use N_i^κ to denote the κ-hop neighborhood of agent i on $\mathcal{G}$, i.e., $N_i^\kappa := \{j \in \mathcal{N} \mid d_{\mathcal{G}}(i, j) \leq \kappa\}$. Let $f(\kappa) := \sup_i |N_i^\kappa|$. For a subset $M \subseteq \mathcal{N}$, we use s_M / a_M to denote the tuple formed by the states/actions of agents in M.

Before we define the transitions and rewards, we first define the notion of active link sets, which are directed graphs on the agents $\mathcal{N}$ and they characterize the interaction structure among the agents. More specifically, an active link set is a set of directed edges that contains all self-loops, i.e., a subset of $\mathcal{N} \times \mathcal{N}$ and a super set of $\{(i,i) \mid i \in \mathcal{N}\}$. Generally speaking, $(j,i) \in L$ means agent j can affect agent i in the active link set L. Given an active link set L, we also use $N_i(L) := \{j \in \mathcal{N} \mid (j,i) \in L\}$ to denote the set of all agents (include itself) who can affect agent i in the active link set L. In this paper, we consider a pair of active link sets (L_t^s, L_t^r) that is independently drawn from some joint distribution $\mathcal{D}$ at each time step t,[1] where the distribution $\mathcal{D}$ will be defined using the underlying graph $\mathcal{G}$ later in Section 2.1. The role of L_t^s/L_t^r is that they define the dependence structure of state transition/reward at time t, which we detail below.

Transitions. At time t, given the current state, action $s(t), a(t)$ and the active link set L_t^s, the next individual state $s_i(t+1)$ is independently generated and only depends on the state/action of the agents in $N_i(L_t^s)$. In other words, we have,

$$P(s(t+1)|s(t),a(t),L_t^s) = \prod_{i\in\mathcal{N}} P_i(s_i(t+1)|s_{N_i(L_t^s)}(t), a_{N_i(L_t^s)}(t), L_t^s). \tag{1}$$

Rewards. Each agent is associated with a local reward function r_i. At time t, it is a function of L_t^r and the state/action of agents in $N_i(L_t^r)$: $r_i(L_t^r, s_{N_i(L_t^r)}(t), a_{N_i(L_t^r)}(t))$. The global reward $r(t)$ is defined to be the summation of the local rewards $r_i(t)$.

Policy. Each agent follows a localized policy that depends on its β-hop neighborhood, where $\beta \geq 0$ is a fixed integer. Specifically, at time step t, given the global state $s(t)$, agent i adopts a local policy ζ_i parameterized by θ_i to decide the distribution of $a_i(t)$ based on the the states of agents in N_i^β.

Our objective is for all the agents to *cooperatively* maximize the discounted global reward, i.e., $J(\theta) = \mathbb{E}_{s\sim\pi_0}\left[\sum_{t=0}^\infty \gamma^t r(s(t),a(t)) \mid s(0)=s\right]$, where π_0 is a given distribution on the initial global state, and we recall $r(s(t),a(t))$ is the global stage reward defined as the sum of all local rewards at time t.

Examples. To highlight the applicability of the general model, we include two examples of networked systems that feature the dependence structure captured by our model in Appendix A: a wireless communication example and an example of controlling a process that spreads over a network.

Note that a limitation of our setting is that the dependence structure we consider is stationary, in the sense that dependencies are sampled i.i.d. from the distribution $\mathcal{D}$. It is important to consider more general time-varying forms (e.g. Markovian) in future research.

Background. Before moving on, we review a few key concepts in RL which will be useful in the rest of the section. We use π_t^θ to denote the distribution of $s(t)$ under policy θ given that $s(0) \sim \pi_0$. A well-known result [47] is that the gradient of the objective $\nabla J(\theta)$ can be computed by $\frac{1}{1-\gamma}\mathbb{E}_{s\sim\pi^\theta, a\sim\zeta^\theta(\cdot|s)} Q^\theta(s,a)\nabla \log \zeta^\theta(a \mid s)$, where distribution $\pi^\theta(s) = (1-\gamma)\sum_{t=0}^\infty \gamma^t \pi_t^\theta(s)$ is the *discounted state visitation distribution*. Evaluating the Q-function $Q^\theta(s,a)$ plays a key role in approximating $\nabla J(\theta)$. The local Q-function for agent i is the discounted local reward, i.e. $Q_i^\theta(s,a) = \mathbb{E}_{\zeta^\theta}\left[\sum_{t=0}^\infty \gamma^t r_i(t) \mid s(0)=s, a(0)=a\right]$, where we use $r_i(t)$ to denote the local reward of agent i at time step t. Using local Q-functions, we can decompose the global Q-function as $Q^\theta(s,a) = \frac{1}{n}\sum_{i=1}^n Q_i^\theta(s,a)$, which allows each node to evaluate its local Q-function separately.

A key challenge in our MARL setting is that directly estimating the Q-functions is not scalable since the size of the Q-functions is exponentially large in the number of agents. Therefore, in Section 2.1, we study structural properties of the Q-functions resulting from the dependence structure in the transition (1), which enables us to design a scalable RL algorithm in Section 2.2.

2.1 μ-decay Property

One of the core challenges for MARL is that the size of the Q function is exponentially large in the number of agents. The key to our algorithm and its analysis is the identification of a novel structural

[1] Here, correlations between L_t^s and L_t^r are possible

decay property for the Q-function, which says that the local Q-function of each agent i is mainly decided by the states of the agents who are near i. This property is critical for the design of scalable algorithms because it enables the agents to reduce the dimension of the Q-function by truncating its dependence of the states and actions of far away agents. Recently, exponential decay has been shown to hold in networked MARL when the network is static [38, 36], which is exploited to design a scalable RL algorithm. However, in stochastic network settings it is too much to hope for exponential decay in general [14], and so we introduce the more general notion of μ-decay here, where μ is a function that converges to 0 as κ tends to infinity. The case of exponential decay that has been studied previously corresponds to $\mu(\kappa) = \gamma^\kappa/(1-\gamma)$. The formal definition of μ-decay is given below, where for simplicity, we use $i \xrightarrow{L} j$ to denote $(i,j) \in L$ and denote $N^\kappa_{-i} := \mathcal{N} \setminus N^\kappa_i$.

Definition 2.1. *For a function $\mu : \mathbb{N} \to \mathbb{R}^+$ that satisfies $\lim_{\kappa\to+\infty} \mu(\kappa) = 0$, the μ-decay property holds if for any policy θ and any $i \in \mathcal{N}$, the local Q function Q^θ_i satisfies $\left|Q^\theta_i(s,a) - Q^\theta_i(s',a')\right| \leq \mu(\kappa)$ for any $(s,a),(s',a')$ that are identical within N^κ_i, i.e. $s_{N^\kappa_i} = s'_{N^\kappa_i}, a_{N^\kappa_i} = a'_{N^\kappa_i}$.*

Intuitively, if the μ-decay property holds and $\mu(\kappa)$ decays quickly as κ increases, we can approximately decompose the global Q function as $Q^\theta(s,a) = \frac{1}{n}\sum_{i=1}^n Q^\theta_i(s,a) \approx \frac{1}{n}\sum_{i=1}^n \hat{Q}^\theta_i(s_{N^\kappa_i}, a_{N^\kappa_i})$, where $\hat{Q}_i$ only depends on the states and actions within the κ-hop neighborhood of agent i. Before our work, [46] empirically showed that such a value decomposition allows efficient training of MARL. Under the assumption that such decomposition exists, [46] propose an approach to learn this decomposition. In contrast, as we prove in this section, the μ decay property holds provably and therefore, the global Q function can be directly decomposed in the networked MARL model and that the error of such decomposition is provably small.

Our first result is Theorem 2.1 which shows the relationship between the random active link sets and the μ-decay property. The proof of Theorem 2.1 is deferred to Appendix B.1.

Theorem 2.1. *Define L^a as the static active link set that contains all pairs (i,j) whose graph distance on $\mathcal{G}$ is less than or equal to β, which is the dependency of local policy. Let random variable $X_i(\kappa)$ denote the smallest $t \in \mathbb{N}$ such that there exists a chain of agents*

$$j^a_0 \xrightarrow{L^s_0} j^s_1 \xrightarrow{L^a} j^a_1 \xrightarrow{L^s_1} \cdots \xrightarrow{L^s_{t-1}} j^s_t \xrightarrow{L^a} j^a_t,$$

that satisfies $j^a_0 \in N^\kappa_{-i}$ and $j^a_t \xrightarrow{L^r_t} i$. The μ-decay property holds for $\mu(\kappa) = \frac{1}{1-\gamma}\mathbb{E}\left[\gamma^{X_i(\kappa)}\right]$.

To make the μ-decay result more concrete, we provide several scenarios that yield different upper bounds on the term $\mathbb{E}\left[\gamma^{X_i(\kappa)}\right]$. In the first scenario, we study the case where long range links do not exist in Corollary 2.2. In this case, we obtain an exponential decay property that generalizes the result in [38]. A proof is in Appendix B.2.

Corollary 2.2 (Exponential Decay)**.** *Consider a distribution $\mathcal{D}$ of active link sets that satisfies*

$$P_{(L^s,L^r)\sim\mathcal{D}}\{(i,j) \in L^s\} = 0, \text{ for all } i,j \in \mathcal{N} \text{ s.t. } d_\mathcal{G}(i,j) \geq \alpha_1,$$
$$P_{(L^s,L^r)\sim\mathcal{D}}\{(i,j) \in L^r\} = 0, \text{ for all } i,j \in \mathcal{N} \text{ s.t. } d_\mathcal{G}(i,j) \geq \alpha_2.$$

Then, $\mathbb{E}\left[\gamma^{X_i(\kappa)}\right] \leq C\rho^\kappa$, where $\rho = \gamma^{1/(\alpha_1+\beta)}, C = \gamma^{-\alpha_2/(\alpha_1+\beta)}$.

In the second scenario, long range active links can occur, but with exponentially small probability with respect to their distance. In this case, we can obtain a near-exponential decay property where $\mu(\kappa) = O(\rho^{\kappa/\log\kappa}))$ for some $\rho \in (0,1)$. A proof can be found in Appendix B.3.

Theorem 2.3 (Near-Exponential Decay)**.** *Suppose the distribution $\mathcal{D}$ of active link sets satisfies*

$$P_{(L^s,L^r)\sim\mathcal{D}}\{(i,j) \in L^s \cup L^r\} \leq c\lambda^{d_\mathcal{G}(i,j)}, \text{ for all } i,j \in \mathcal{N},$$

where $c \geq 1, 1 > \lambda > 0$ are constants. If the largest size of the κ neighborhood in the underlying graph $\mathcal{G}$ can be bounded by a polynomial of κ, i.e., there exists some constants $c_0 \geq 1, n_0 \in \mathbb{N}$ such that $|\{j \in \mathcal{N} \mid d_\mathcal{G}(i,j) = \kappa\}| \leq c_0(\kappa+1)^{n_0}$ holds for all i, then $\mathbb{E}\left[\gamma^{X_i(\kappa-1)}\right] \leq C\rho^{\kappa/(1+\ln(\kappa+1))}$ for some positive constant C and decay rate $\rho < 1$. [2]

It is interesting to compare the result above with models of the so-called "small world phenomena" in social networks, e.g., [14]. In these models, a link (i,j) occurs with probability $1/poly(d_\mathcal{G}(i,j))$,

[2] The explicit expression of C and ρ can be found in Appendix B.3.

Algorithm 1 Scalable Actor Critic

1: **for** $m = 0, 1, 2, \cdots$ **do**
2: Sample initial global state $s(0) \sim \pi_0$.
3: Each node i takes action $a_i(0) \sim \zeta_i^{\theta_i(m)}(\cdot \mid s_{N_i^\beta}(0))$ to obtain the global state $s(1)$.
4: Each node i records $s_{N_i^\kappa}(0), a_{N_i^\kappa}(0), r_i(0)$ and initialize $\hat{Q}_i^0$ to be all zero vector.
5: **for** $t = 1, \cdots, T$ **do**
6: Each node i takes action $a_i(t) \sim \zeta_i^{\theta_i(m)}(\cdot \mid s_{N_i^\beta}(t))$ to obtain the global state $s(t+1)$.
7: Each node i update the local estimation $\hat{Q}_i$ with step size $\alpha_{t-1} = \frac{H}{t-1+t_0}$,

$$\hat{Q}_i^t\big(s_{N_i^\kappa}(t-1), a_{N_i^\kappa}(t-1)\big) =$$
$$(1-\alpha_{t-1})\hat{Q}_i^{t-1}\big(s_{N_i^\kappa}(t-1), a_{N_i^\kappa}(t-1)\big) + \alpha_{t-1}\Big(r_i(t) + \gamma\hat{Q}_i^{t-1}\big(s_{N_i^\kappa}(t), a_{N_i^\kappa}(t)\big)\Big),$$
$$\hat{Q}_i^t\big(s_{N_i^\kappa}, a_{N_i^\kappa}\big) = \hat{Q}_i^{t-1}\big(s_{N_i^\kappa}, a_{N_i^\kappa}\big) \text{ for } \big(s_{N_i^\kappa}, a_{N_i^\kappa}\big) \neq \big(s_{N_i^\kappa}(t-1), a_{N_i^\kappa}(t-1)\big).$$

8: Each node i approximate $\nabla_{\theta_i} J(\theta)$ by
$\hat{g}_i(m) = \sum_{t=0}^T \gamma^t \frac{1}{n} \sum_{j \in N_i^\kappa} \hat{Q}_j^T\big(s_{N_j^\kappa}(t), a_{N_j^\kappa}(t)\big) \nabla_{\theta_i} \log \zeta_i^{\theta_i(m)}\big(a_i(t) \mid s_{N_i^\beta}(t)\big)$.
9: Each node i conducts gradient ascent by $\theta_i(m+1) = \theta_i(m) + \eta_m \hat{g}_i(m)$.

as opposed to the exponential dependence in Lemma 2.3. In this case, one can see function $\mu(\kappa)$ is lower bounded by $1/poly(\kappa)$, which leads us to conjecture that $\mu(\kappa)$ is also upper bounded by $O(1/poly(\kappa))$. Thus, when information spreads "slowly" it helps a localized algorithm to learn efficiently.

2.2 A Scalable Actor Critic Algorithm

Motivated by the μ-decay property of the Q-functions, we design a novel Scalable Actor Critic algorithm (Algorithm 1) for networked MARL problem, which exploits the μ-decay result in the previous section. The Critic part (from line 2 to line 7) uses the local trajectory $\{(s_{N_i^\kappa}, a_{N_i^\kappa}, r_i)\}$ to evaluate the local Q-functions under parameter $\theta(m)$. Intuitively, the μ-decay property guarantees that we can achieve good approximation error even when κ is not large. The Actor part (from line 8 to line 9) computes the estimated partial derivative using the estimated local Q-functions, and uses the partial derivative to update local parameter θ_i. The step size sequence $\{\eta_m\}$ will be defined in Theorem B.2. Compared with the Scalable Actor Critic algorithm proposed in [38], Algorithm 1 extends the policy dependency structure considered. No longer is the dependency completely local; it now extends to all agents within the β-hop neighborhood. Interestingly, the time-varying dependencies do not add complexity into the algorithm (though the analysis is more complex).

Algorithm 1 is highly scalable. Each agent i needs only to query and store the information within its κ-hop neighborhood during the learning process. The parameter κ can be set to balance accuracy and complexity. Specifically, as κ increases, the error bound becomes tighter at the expense of increasing computation, communication, and space complexity.

2.3 Convergence

We now present our main result, a finite-time error bound for the Scalable Actor Critic algorithm (Algorithm 1) that holds under general (non-local) dependencies. To that end, we first describe the assumption needed in our result. It focuses on the Markov chain formed by the global state-action pair (s, a) under a fixed policy parameter θ and is standard for finite-time convergence results in RL, e.g., [44, 5, 37].

Assumption 2.1. *Under any fixed policy θ, $\{z(t) := (s(t), a(t))\}$ is an aperiodic and irreducible Markov chain on state space $\mathcal{Z} := \mathcal{S} \times \mathcal{A}$ with a unique stationary distribution $d^\theta = (d_z^\theta, z \in \mathcal{Z})$, which satisfies $d_z^\theta > 0, \forall z \in \mathcal{Z}$. Define $d^\theta(z') = \sum_{z \in \mathcal{Z}: z_{N_i^\kappa} = z'} d^\theta(z)$ and $\sigma'(\kappa) := \inf_{z' \in \mathcal{Z}_{N_i^\kappa}} d^\theta(z')$. There exists positive constants K_1, K_2 such that $K_2 \geq 1$ and $\forall z' \in \mathcal{Z}, \forall t \geq 0, \sup_{\mathcal{K} \subseteq \mathcal{Z}} \big|\sum_{z \in \mathcal{K}} d_z^\theta - \sum_{z \in \mathcal{K}} \mathbb{P}(z(t) = z \mid z(0) = z')\big| \leq K_1 e^{-t/K_2}$.*

We next analyze the Critic part of Algorithm 1 within a given outer loop iteration m. Since the policy is fixed in the inner loop, the global state/action pair (s, a) in the original MDP can be viewed as the state of a Markov chain. We observe that each local estimate $\hat{Q}_i^t\big(s_{N_i^\kappa}, a_{N_i^\kappa}\big)$ can be viewed as a form of state aggregation, where the global state (s, a) is "compressed" to $h(s, a) := (s_{N_i^\kappa}, a_{N_i^\kappa})$. Broadly speaking, the technique of state aggregation is one of the easiest-to-deploy schemes for state space compression [21, 43], while its final performance relies heavily on whether the state aggregation map h only aggregates "similar" states. To have a good approximate equivalence, we need to find a good h, i.e., if two states are mapped to the same abstract state, their value functions are required to be close (to be discussed in Theorem 3.2). In the context of networked MARL, the μ decay property (Definition 2.1) provides a natural mapping for state aggregation $h(s, a) := (s_{N_i^\kappa}, a_{N_i^\kappa})$ which we defined earlier. This mapping h maps the global state/action to the local states/actions in agent i's κ-hop neighborhood and the μ-decay property guarantees that if $h(s, a) = h(s', a')$, the difference in their Q-functions is upper bounded by $\mu(\kappa)$, which is vanishing as κ increases. This shows that the mapping h we used is "good" in the sense it aggregates very similar global state-action pairs. This idea leads to the following theorem about the Critic part of Scalable Actor Critic (Algorithm 1).

Theorem 2.4. *Suppose Assumption 2.1 and μ-decay property (Definition 2.1) hold. Let the step size be $\alpha_t = \frac{H}{t+t_0}$ with $t_0 = \max(4H, 2K_2 \log T)$, and $H \geq \frac{2}{(1-\gamma)\sigma'(\kappa)}$. Define constant $C_b := 4K_1(1 + 2K_2 + 4H)$. Then, inside outer loop iteration m, for each $i \in \mathcal{N}$, with probability at least $1 - \delta$, we have $\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left|Q_i^{\theta(m)}(s, a) - \hat{Q}_i^T(s_{N_i^\kappa}, a_{N_i^\kappa})\right| \leq \frac{C_a}{\sqrt{T+t_0}} + \frac{C'_a}{T+t_0} + \frac{\mu(\kappa)}{1-\gamma}$, where the constants are given by $C_a = \frac{40H}{(1-\gamma)^2}\sqrt{K_2 \log T\Big(\log\Big(\frac{4f(\kappa)K_2T}{\delta}\Big) + \log\log T\Big)}$ and $C'_a = \frac{8}{(1-\gamma)^2}\max\{\frac{144K_2H\log T}{\sigma'(\kappa)} + C_b, 2K_2 \log T + t_0\}$.*

The proof of Theorem 2.4 can be found in Appendix B.4. The most related result in the literature to Theorem 2.4 is Theorem 7 in [38]. In comparison, Theorem 2.4 applies for more general, potentially non-local, dependencies and, also, improves the constant term by a factor of $1/(1 - \gamma)$.

To analyze the Actor part of Algorithm 1, we make the following additional boundedness and Lipschitz continuity assumptions on the gradients. These are standard assumptions in the literature.

Assumption 2.2. *For any $i, a_i, s_{N_i^\beta}$ and θ_i, we assume $\Big\|\nabla_{\theta_i} \log \zeta_i^{\theta_i}(a_i \mid s_{N_i^\beta})\Big\| \leq W_i$. Then, for any L_t^a, $\big\|\nabla_\theta \log \zeta^\theta(a \mid s)\big\| \leq W := \sqrt{\sum_{i=1}^n W_i^2}$. We further assume $\nabla J(\theta)$ is W'-Lipschitz in θ.*

Intuitively, since the quality of the estimated policy gradient depends on the quality of the estimation of Q-functions, if every agent i has learned a good approximation of its local Q-function in the Critic part of Algorithm 1, the policy gradient can be approximated well. Therefore, the Actor part can obtain a good approximation of a stationary point of the objective function. We state the sample complexity result in Theorem 2.5 and defer the detailed bounds and a proof to Appendix B.5.

Theorem 2.5. *Under Assumption 2.2, to reach an $O(\epsilon)$-approximate stationary point with probability at least $1 - \delta$, we need to choose κ such that $\mu(\kappa) = O\big(W^{-2}(1 - \gamma)^4\epsilon\big)$. The number of required iterations of the outer loop should satisfy $M = \tilde{\Omega}\big(\epsilon^{-2} poly(W, W', \frac{1}{1-\gamma})\big)$ and the number of required iterations of the inner loop is $T = \tilde{\Omega}\big(\epsilon^{-2} poly(W, \frac{1}{\sigma'(\kappa)}, K_2, \frac{1}{1-\gamma}, \log f(\kappa), \log(1/\delta))\big)$.*

Note that W scales with the number of agents n. Thus, Theorem 2.5 shows that the complexity of our algorithm scales with the largest state-action space size of any κ-hop neighborhood and the number of agents n, which avoids the exponential blowup in n when the graph is sparse and achieves scalable RL for networked agents even under stochastic, non-local settings.

3 Proof Idea: Stochastic Approximation and State Aggregation

In this section, we present the key technical innovation underlying our results on MARL in Theorem 2.4: a new finite-time analysis of a general asynchronous stochastic approximation (SA) scheme. As we mention in Section 2, the truncation enabled by μ-decay provides a form of state aggregation, which we analyze via a general SA scheme in Section 3.1. Further, this SA scheme is of interest more broadly, e.g., to the settings of TD learning with state aggregation (Section 3.2) and asynchronous Q-learning with state aggregation (Appendix D.4).

3.1 Stochastic Approximation

Consider a finite-state Markov chain whose state space is given by $\mathcal{N} = \{1, 2, \cdots, n\}$. Let $\{i_t\}_{t=0}^{\infty}$ be the sequence of states visited by this Markov chain. Our focus is generalizing the following asynchronous stochastic approximation (SA) scheme, which is studied in [48, 41, 52]: Let parameter $x \in \mathbb{R}^{\mathcal{N}}$, and $F : \mathbb{R}^{\mathcal{N}} \to \mathbb{R}^{\mathcal{N}}$ be a γ-contraction in the infinity norm. The update rule of the SA scheme is given by

$$\begin{aligned} x_{i_t}(t+1) &= x_{i_t}(t) + \alpha_t(F_{i_t}(x(t)) - x_{i_t}(t) + w(t)), \\ x_j(t+1) &= x_j(t) \text{ for } j \neq i_t, j \in \mathcal{N}, \end{aligned} \tag{2}$$

where $w(t)$ is a noise sequence. It is shown in [37] that parameter $x(t)$ converges to the unique fixed point of F at the rate of $O\big(1/\sqrt{t}\big)$.

While general, in many cases, including networked MARL, we do not wish to calculate an entry for every state in $\mathcal{N}$ in parameter x, but instead, wish to calculate "aggregated entries." Specifically, at each time step, after i_t is generated, we use a surjection h to decide which dimension of parameter x should be updated. This technique, referred to as state aggregation, is one of the easiest-to-deploy schemes for state space compression in the RL literature [21, 43]. In the generalized SA scheme, our objective is to specify the convergence point as well as obtain a finite-time error bound.

Formally, to define the generalization of (2), let $\mathcal{N} = \{1, \cdots, n\}$ be the state space of $\{i_t\}$ and $\mathcal{M} = \{1, \cdots, m\}, (m \leq n)$ be the *abstract* state space. The surjection $h : \mathcal{N} \to \mathcal{M}$ is used to convert every state in $\mathcal{N}$ to its abstraction in $\mathcal{M}$. Given parameter $x \in \mathbb{R}^{\mathcal{M}}$ and function $F : \mathbb{R}^{\mathcal{N}} \to \mathbb{R}^{\mathcal{N}}$, we consider the generalized SA scheme that updates $x(t) \in \mathbb{R}^{\mathcal{M}}$ starting from $x(0) = \mathbf{0}$,

$$\begin{aligned} x_{h(i_t)}(t+1) &= x_{h(i_t)}(t) + \alpha_t\big(F_{i_t}(\Phi x(t)) - x_{h(i_t)}(t) + w(t)\big), \\ x_j(t+1) &= x_j(t) \text{ for } j \neq h(i_t), j \in \mathcal{M}, \end{aligned} \tag{3}$$

where the feature matrix $\Phi \in \mathbb{R}^{\mathcal{N} \times \mathcal{M}}$ is defined as

$$\Phi_{ij} = \begin{cases} 1 & \text{if } h(i) = j \\ 0 & \text{otherwise} \end{cases}, \forall i \in \mathcal{N}, j \in \mathcal{M}. \tag{4}$$

In order to state our main result characterizing the convergence of (3), we must first state a few definitions and assumptions. To begin, we define the weighted infinity norm as in [37], except that we extend its definition so as to define the contraction of function F. The reason we use the weighted infinity norm as opposed to the standard infinity norm is that its generality can be used in certain settings for undiscounted RL, as shown in [48, 2].

Definition 3.1 (Weighted Infinity Norm)**.** *Fix a positive vector $v \in \mathbb{R}^{\mathcal{M}}$. For $x \in \mathbb{R}^{\mathcal{M}}$, we define $\|x\|_v := \sup_{i \in \mathcal{M}} \frac{|x_i|}{v_i}$. For $x \in \mathbb{R}^{\mathcal{N}}$, we define $\|x\|_v := \sup_{i \in \mathcal{N}} \frac{|x_i|}{v_{h(i)}}$.*

Next, we state our assumption on the mixing rate of the Markov chain $\{i_t\}$, which is common in the literature [50, 44]. It holds for any finite-state Markov chain which is aperiodic and irreducible [5].

Assumption 3.1 (Stationary Distribution and Geometric Mixing Rate)**.** *$\{i_t\}$ is an aperiodic and irreducible Markov chain on state space $\mathcal{N}$ with stationary distribution $d = (d_1, d_2, \cdots, d_n)$. Let $d'_j = \sum_{i \in h^{-1}(j)} d_i$ and $\sigma' = \inf_{j \in \mathcal{M}} d'_j$. There exists positive constants K_1, K_2 which satisfy that $\sup_{\mathcal{S} \subseteq \mathcal{N}} \left| \sum_{i \in \mathcal{S}} d_i - \sum_{i \in \mathcal{S}} \mathbb{P}(i_t = i \mid i_0 = j) \right| \leq K_1 \exp(-t/K_2), \forall j \in \mathcal{N}, \forall t \geq 0$ and $K_2 \geq 1$.*

Our next assumption ensures contraction of F. It is also standard, e.g., [48, 52, 37], and ensures that F has a unique fixed point y^*.

Assumption 3.2 (Contraction)**.** *Operator F is a γ contraction in $\|\cdot\|_v$, i.e., for any $x, y \in \mathbb{R}^{\mathcal{N}}$, we have $\|F(x) - F(y)\|_v \leq \gamma \|x - y\|_v$. Further, there exists some constant $C > 0$ such that for any $x \in \mathbb{R}^{\mathcal{N}}$, we have $\|F(x)\|_v \leq \gamma \|x\|_v + C$.*

In Assumption 3.2, notice that the first sentence directly implies the second with $C = (1 + \gamma)\|y^*\|_v$, where $y^* \in \mathbb{R}^{\mathcal{N}}$ is the unique fixed point of F. Further, while Assumption 3.2 implies that F has a unique fixed point y^*, we do not expect our stochastic approximation scheme to converge to it. Instead, we show that the convergence is to the unique x^* that solves

$$\Pi F(\Phi x^*) = x^*, \text{ where } \Pi := \left(\Phi^\top D \Phi\right)^{-1} \Phi^\top D. \tag{5}$$

Here $D = diag(d_1, d_2, \cdots, d_n)$ denotes the steady-state probabilities for the process $\{i_t\}$. Note that x^* is well-defined because the operator $\Pi F(\Phi \cdot)$, which defines a mapping from $\mathbb{R}^{\mathcal{M}}$ to $\mathbb{R}^{\mathcal{M}}$, is also a contraction in $\|\cdot\|_v$. We state and prove this as Proposition C.1 in Appendix C.1.

Our last assumption is on the noise sequence $w(t)$. It is also standard, e.g., [41, 37].

Assumption 3.3 (Martingale Difference Sequence)**.** *w_t is $\mathcal{F}_{t+1}$ measurable and satisfies $\mathbb{E}w(t) \mid \mathcal{F}_t = 0$. Further, $|w(t)| \leq \bar{w}$ almost surely for constant $\bar{w}$.*

We are now ready to state our finite-time convergence result for stochastic approximation.

Theorem 3.1. *Suppose Assumptions 3.1, 3.2, 3.3 hold. Further, assume there exists constant $\bar{x} \geq \|x^*\|_v$ such that $\forall t, \|x(t)\|_v \leq \bar{x}$ almost surely.*[3] *Let the step size be $\alpha_t = \frac{H}{t+t_0}$ with $t_0 = \max(4H, 2K_2 \log T)$, and $H \geq \frac{2}{\sigma'(1-\gamma)}$. Let x^* be the unique solution of equation $\Pi F(\Phi x^*) = x^*$, and define constants $C_1 := 2\bar{x} + C + \frac{\bar{w}}{\underline{v}}, C_2 := 4\bar{x} + 2C + \frac{\bar{w}}{\underline{v}}, C_3 := 2K_1(2\bar{x} + C)(1 + 2K_2 + 4H)$. Then, with probability at least $1 - \delta$,*

$$\|x(T) - x^*\|_v \leq \frac{C_a}{\sqrt{T + t_0}} + \frac{C_a'}{T + t_0} = \tilde{O}\left(\frac{1}{\sqrt{T}}\right),$$

where the constants are given by $C_a = \frac{4HC_2}{1-\gamma}\sqrt{K_2 \log T\big(\log\big(\frac{4mK_2T}{\delta}\big) + \log\log T\big)}$ and $C_a' = 4\max\{\frac{48K_2C_1H\log T + \sigma' C_3}{(1-\gamma)\sigma'}, \frac{2\bar{x}(2K_2 \log T + t_0)}{1-\gamma}\}$.

A proof of Theorem 3.1 can be found in Appendix C.2. Compared with Theorem 4 in [37], Theorem 3.1 holds for a more general SA scheme where state aggregation is used to reduce the dimension of the parameter x. The proof technique used in [37] does not apply to our setting because our stationary point x^* has a more complex form (4). To do the generalization, we need to use a different error decomposition method compared to [37] that leverages the stationary distribution D rather than the distribution of i_t condition on $i_{t-\tau}$ (see Appendix C.2 for details). Because of this generality, Theorem 3.1 requires a stronger but standard assumption on the mixing rate of the Markov chain $\{i_t\}$.

3.2 State Aggregation

To illustrate the impact of our analysis of SA (Theorem 3.1) beyond the network setting, we study a simpler application to the cases of TD-learning and Q-learning with state aggregation in this section. Understanding state aggregation methods is a foundational goal of analysis in the RL literature and it has been studied in many previous works, e.g., [26, 23, 22, 9, 43]. Further, the result is extremely useful in the analysis in networked MARL that follows since the μ-decay property we introduce (Definition 2.1) provides a natural state aggregation in the network setting (see Corollary 2.4). Due to space constraints, in this section we only introduce the results on TD-learning; the results on Q-learning are given in Appendix D.4.

In TD learning with state aggregation [43, 49], given the sequence of states visited by the Markov chain is $\{i_t\}$, the update rule of TD(0) is given by

$$\begin{aligned} \theta_{h(i_t)}(t+1) &= \theta_{h(i_t)}(t) + \alpha_t\big(r_t + \gamma\theta_{h(i_{t+1})}(t) - \theta_{h(i_t)}(t)\big), \\ \theta_j(t+1) &= \theta_j(t) \text{ for } j \neq h(i_t), j \in \mathcal{M}, \end{aligned} \tag{6}$$

where $h : \mathcal{N} \to \mathcal{M}$ is a surjection that maps each state in $\mathcal{N}$ to an abstract state in $\mathcal{M}$ and r_t is the reward at time step t such that $\mathbb{E}[r_t] = r(i_t, i_{t+1})$.

Taking F as the Bellman Policy Operator, i.e., the i'th dimension of function F is given by

$$F_i(V) = \mathbb{E}_{i' \sim \mathbb{P}(\cdot|i)}[r(i, i') + \gamma V_{i'}], \forall i \in \mathcal{N}, V \in \mathbb{R}^{\mathcal{N}}.$$

The value function (vector) V^* is defined as $V_i^* = \mathbb{E}[\sum_{t=0}^{\infty} \gamma^t r(i_t, i_{t+1}) \mid i_0 = i], i \in \mathcal{N}$ [49]. By defining the feature matrix Φ as (4) and the noise sequence as

$$w(t) = r_t + \gamma\theta_{h(i_{t+1})}(t) - \mathbb{E}_{i' \sim \mathbb{P}(\cdot|i_t)}[r(i_t, i') + \gamma\theta_{h(i')}(t)],$$

we can rewrite the update rule of TD(0) in (6) in the form of an SA scheme (3). Therefore, we can apply Theorem 3.1 to obtain a finite-time error bound for TD learning with state aggregation. A proof of Theorem 3.2 can be found in Appendix D.2.

[3] The assumption on $\bar{x}$ follows from Assumptions 3.2 and 3.3. See Proposition C.2 in Appendix C.3.

Theorem 3.2. *Let Assumption 3.1 hold for the Markov chain $\{i_t\}$ and let the stage reward r_t be upper bounded by $\bar{r}$ almost surely. Assume that if $h(i) = h(i')$ for $i, i' \in \mathcal{N}$, we have $|V_i^* - V_{i'}^*| \leq \zeta$ for a constant ζ. Consider TD(0) with the step size $\alpha_t = \frac{H}{t+t_0}$, where $t_0 = \max(4H, 2K_2 \log T)$ and $H \geq \frac{2}{\sigma'(1-\gamma)}$. Define constant $C_4 := 4K_1(1 + 2K_2 + 4H)$. Then, with probability at least $1 - \delta$,*

$$\|\Phi \cdot \theta(T) - V^*\|_\infty \leq \frac{C_a}{\sqrt{T+t_0}} + \frac{C_a'}{T+t_0} + \frac{\zeta}{1-\gamma},$$

where the constants are given by $C_a = \frac{40H\bar{r}}{(1-\gamma)^2}\sqrt{K_2 \log T\big(\log\big(\frac{4mK_2T}{\delta}\big) + \log\log T\big)}$ and $C_a' = \frac{8\bar{r}}{(1-\gamma)^2}\max\{\frac{144K_2H\log T}{\sigma'} + C_4, 2K_2 \log T + t_0\}$.

The most related prior results to Theorem 3.2 are [44, 4]. In contrast to these, Theorem 3.2 considers the infinity norm, which is more natural for measuring error when using state aggregation. Further, our analysis is different and extends to the case of Q-learning with state aggregation (see Appendix D.4), where we obtain the first finite-time error bound. Moreover, unlike [4], our TD-learning algorithm does not require a projection step.

4 Concluding Remarks

In this paper, we propose and analyze the Scalable Actor Critic Algorithm that provably learns a near-optimal local policy in a setting where every agent is allowed to interact with a random subset of agents. The μ-decay property, which enables the decentralized approximation of local Q functions, is the key to our approach.

There are a number of future directions motivated by the results in this paper. For example, we allow the interaction structure among the agents to change in a stochastic way in this work. It is interesting to see if such structure can be time-varying in more general ways (e.g., Markovian or adversarial). Besides, although our Scalable Actor Critic algorithm consumes much less memory than a centralized tabular approach, the memory space required by each agent i to store $\hat{Q}_i$ grows exponentially with respect to $f(\kappa)$, which denotes the size of the largest κ-hop neighborhood. Thus, memory problems may still arise if f grows quickly as κ increases. Therefore, an interesting open problem is whether we can apply additional function approximations on truncated state/action pair $(s_{N_i^\kappa}, a_{N_i^\kappa})$, and obtain similar finite-time convergence guarantees as Scalable Actor Critic.

References

[1] B. Bamieh, F. Paganini, and M. A. Dahleh. Distributed control of spatially invariant systems. *IEEE Transactions on automatic control*, 47(7):1091–1107, 2002.

[2] D. P. Bertsekas. *Dynamic Programming and Optimal Control, Vol. II*. Athena Scientific, 3rd edition, 2007.

[3] D. P. Bertsekas and J. N. Tsitsiklis. *Neuro-dynamic programming*, volume 5. Athena Scientific Belmont, MA, 1996.

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

[5] P. Bremaud. *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. Texts in Applied Mathematics. Springer New York, 2013.

[6] L. Bu, R. Babu, B. De Schutter, et al. A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172, 2008.

[7] D. Chakrabarti, Y. Wang, C. Wang, J. Leskovec, and C. Faloutsos. Epidemic thresholds in real networks. *ACM Transactions on Information and System Security (TISSEC)*, 10(4):1, 2008.

[8] C. Claus and C. Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. *AAAI/IAAI*, 1998:746–752, 1998.

[9] C. Dann, N. Jiang, A. Krishnamurthy, A. Agarwal, J. Langford, and R. E. Schapire. On oracle-efficient pac rl with rich observations. In *Advances in Neural Information Processing Systems*, pages 1422–1432, 2018.

[10] T. Doan, S. Maguluri, and J. Romberg. Finite-time analysis of distributed TD(0) with linear function approximation on multi-agent reinforcement learning. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 1626–1635, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[11] T. T. Doan. Finite-time analysis and restarting scheme for linear two-time-scale stochastic approximation, 2019.

[12] K. Dong, Y. Wang, X. Chen, and L. Wang. Q-learning with ucb exploration is sample efficient for infinite-horizon mdp. *arXiv preprint arXiv:1901.09311*, 2019.

[13] Y. Duan, X. Chen, R. Houthooft, J. Schulman, and P. Abbeel. Benchmarking deep reinforcement learning for continuous control. In *International Conference on Machine Learning*, pages 1329–1338, 2016.

[14] D. Easley, J. Kleinberg, et al. Networks, crowds, and markets: Reasoning about a highly connected world. *Significance*, 9:43–44, 2012.

[15] J. Foerster, G. Farquhar, T. Afouras, N. Nardelli, and S. Whiteson. Counterfactual multi-agent policy gradients. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

[16] D. Gamarnik. Correlation decay method for decision, optimization, and inference in large-scale networks. In *Theory Driven by Influential Applications*, pages 108–121. INFORMS, 2013.

[17] D. Gamarnik, D. A. Goldberg, and T. Weber. Correlation decay in random decision networks. *Mathematics of Operations Research*, 39(2):229–261, 2014.

[18] H. Gu, X. Guo, X. Wei, and R. Xu. Q-learning for mean-field controls, 2020.

[19] H. Gu, X. Guo, X. Wei, and R. Xu. Q-learning for mean-field controls. *arXiv preprint arXiv:2002.04131*, 2020.

[20] D. hwan Lee and N. He. A unified switching system perspective and o.d.e. analysis of q-learning algorithms. *ArXiv*, abs/1912.02270, 2019.

[21] N. Jiang. Notes on state abstractions. `http://nanjiang.web.engr.illinois.edu/files/cs598/note4.pdf`, 2018.

[22] N. Jiang, A. Kulesza, and S. Singh. Abstraction selection in model-based reinforcement learning. In *International Conference on Machine Learning*, pages 179–188, 2015.

[23] N. K. Jong and P. Stone. State abstraction discovery from irrelevant state variables. In *IJCAI*, volume 8, pages 752–757, 2005.

[24] T. Lattimore and M. Hutter. Pac bounds for discounted mdps. In N. H. Bshouty, G. Stoltz, N. Vayatis, and T. Zeugmann, editors, *Algorithmic Learning Theory*, pages 320–334, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[25] D. Li, D. Zhao, Q. Zhang, and Y. Chen. Reinforcement learning and deep learning based lateral control for autonomous driving [application notes]. *IEEE Computational Intelligence Magazine*, 14(2):83–98, 2019.

[26] L. Li, T. J. Walsh, and M. L. Littman. Towards a unified theory of state abstraction for MDPs. In *ISAIM*, 2006.

[27] M. Llas, P. M. Gleiser, J. M. López, and A. Díaz-Guilera. Nonequilibrium phase transition in a model for the propagation of innovations among economic agents. *Physical Review E*, 68(6):066101, 2003.

[28] R. Lowe, Y. Wu, A. Tamar, J. Harb, O. P. Abbeel, and I. Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, pages 6379–6390, 2017.

[29] L. Matignon, G. J. Laurent, and N. Le Fort-Piat. Independent reinforcement learners in cooperative markov games: a survey regarding coordination problems. *The Knowledge Engineering Review*, 27(1):1–31, 2012.

[30] W. Mei, S. Mohagheghi, S. Zampieri, and F. Bullo. On the dynamics of deterministic epidemic propagation over networks. *Annual Reviews in Control*, 44:116–128, 2017.

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

[32] N. Motee and A. Jadbabaie. Optimal control of spatially distributed systems. *IEEE Transactions on Automatic Control*, 53(7):1616–1629, 2008.

[33] M. J. Neely. Optimal backpressure routing for wireless networks with multi-receiver diversity. In *2006 40th Annual Conference on Information Sciences and Systems*, pages 18–25, 2006.

[34] C. H. Papadimitriou and J. N. Tsitsiklis. The complexity of optimal queuing network control. *Mathematics of Operations Research*, 24(2):293–305, 1999.

[35] G. Qu and N. Li. Exploiting fast decaying and locality in multi-agent mdp with tree dependence structure. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 6479–6486. IEEE, 2019.

[36] G. Qu, Y. Lin, A. Wierman, and N. Li. Scalable multi-agent reinforcement learning for networked systems with average reward. *Advances in Neural Information Processing Systems*, 33, 2020.

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

[38] G. Qu, A. Wierman, and N. Li. Scalable reinforcement learning of localized policies for multi-agent networked systems. In *Learning for Dynamics and Control*, pages 256–266. PMLR, 2020.

[39] L. G. Roberts. Aloha packet system with and without slots and capture. *ACM SIGCOMM Computer Communication Review*, 5(2):28–42, 1975.

[40] N. A. Ruhi, C. Thrampoulidis, and B. Hassibi. Improved bounds on the epidemic threshold of exact sis models on complex networks. In *2016 IEEE 55th Conference on Decision and Control (CDC)*, pages 3560–3565. IEEE, 2016.

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

[42] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

[43] S. P. Singh, T. Jaakkola, and M. I. Jordan. Reinforcement learning with soft state aggregation. In *Advances in neural information processing systems*, pages 361–368, 1995.

[44] R. Srikant and L. Ying. Finite-time error bounds for linear stochastic approximation andtd learning. pages 2803–2830, 2019.

[45] J. Subramanian and A. Mahajan. Reinforcement learning in stationary mean-field games. In *Proceedings of the 18th International Conference on Autonomous Agents and MultiAgent Systems*, pages 251–259, 2019.

[46] P. Sunehag, G. Lever, A. Gruslys, W. M. Czarnecki, V. F. Zambaldi, M. Jaderberg, M. Lanctot, N. Sonnerat, J. Z. Leibo, K. Tuyls, et al. Value-decomposition networks for cooperative multi-agent learning based on team reward. In *AAMAS*, pages 2085–2087, 2018.

[47] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Proceedings of the 12th International Conference on Neural Information Processing Systems*, NIPS'99, page 1057–1063, Cambridge, MA, USA, 1999. MIT Press.

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

[49] J. N. Tsitsiklis and B. Van Roy. An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690, 1997.

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

[51] W. Vogels, R. van Renesse, and K. Birman. The power of epidemics: Robust communication for large-scale distributed systems. *SIGCOMM Comput. Commun. Rev.*, 33(1):131–135, Jan. 2003.

[52] M. J. Wainwright. Stochastic approximation with cone-contractive operators: Sharp $\ell_i nfty$-bounds for Q-learning. *arXiv preprint arXiv:1905.06265*, 2019.

[53] S. Wang, V. Venkateswaran, and X. Zhang. Fundamental analysis of full-duplex gains in wireless networks. *IEEE/ACM Transactions on Networking*, 25(3):1401–1416, 2017.

[54] Y. Wu, W. Zhang, P. Xu, and Q. Gu. A finite time analysis of two time-scale actor critic methods, 2020.

[55] T. Xu, S. Zou, and Y. Liang. Two time-scale off-policy td learning: Non-asymptotic analysis over markovian samples. In *Advances in Neural Information Processing Systems 32*, pages 10634–10644. Curran Associates, Inc., 2019.

[56] Y. Yang, R. Luo, M. Li, M. Zhou, W. Zhang, and J. Wang. Mean field multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 5571–5580. PMLR, 2018.

[57] K. Zhang, Z. Yang, and T. Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *arXiv preprint arXiv:1911.10635*, 2019.

[58] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Başar. Fully decentralized multi-agent reinforcement learning with networked agents. *arXiv preprint arXiv:1802.08757*, 2018.

[59] R. Zhang and M. Pavone. Control of robotic mobility-on-demand systems: a queueing-theoretical perspective. *The International Journal of Robotics Research*, 35(1-3):186–203, 2016.

[60] A. Zocca. Temporal starvation in multi-channel csma networks: an analytical framework. *Queueing Systems*, 91(3-4):241–263, 2019.

A Examples

A.1 Wireless Networks

We consider a wireless network with multiple access points setting shown in Fig. 1, where a set of user nodes in a wireless network, denoted by $U = \{u_1, u_2, \cdots, u_n\}$, share a set of access points $Y = \{y_1, y_2, \cdots, y_m\}$ [60]. Each access point y_i is associated with a probability p_i of successful transmission. Each user node u_i only has access to a subset $Y_i \subseteq Y$ of the access points. Typically, this available set is determined by each user node's physical connections to the access points. To apply the networked MARL model, we identify the set of user nodes U as the set of agents $\mathcal{N}$ in Section 2. The underlying graph $G = (\mathcal{N}, \mathcal{E})$ is defined as the conflict graph, i.e., edge $(u_i, u_j) \in \mathcal{E}$ if and only if $Y_i \cap Y_j \neq \emptyset$.

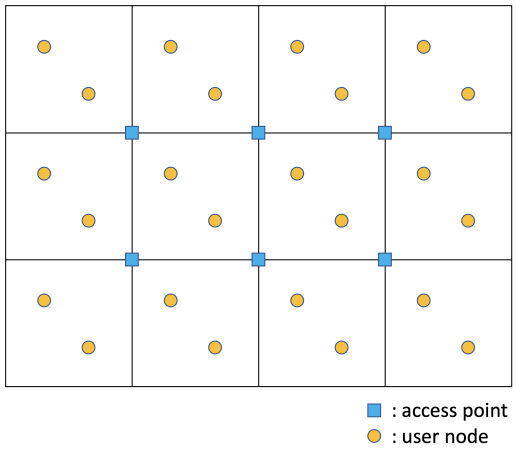

Figure 1: An example setup of wireless networks. Each user node can send packets to the access points at the corners of its grid.

At each time step t, each user u_i receives a packet with initial life span d with probability q. Each user maintains a queue to cache the packets it receives. At each time step, if the packet is successfully sent to an access point, it will be removed from the queue. Otherwise, its life span will decrease by 1. A packet is discarded from the queue immediately if its remaining life span is 0. At each time step t, a user node u_i can choose to send one of the packets in its queue to one of the access point $y_{i,t} \in Y_i$. If no other user node sends packets to access point $y_{i,t}$ at time step t, the packet from user i can be delivered successfully with probability p_i. Otherwise, the sending action will fail. A user u_i receives a local reward of $r_{i,t} = 1$ immediately after successfully sending a packet at time step t, and receives $r_{i,t} = 0$ otherwise. Our objective is to find a policy that maximizes the global discounted reward under a discounted factor $0 \leq \gamma < 1$:

$$\mathbb{E}\left[\sum_{i=1}^{n}\sum_{t=0}^{\infty}\gamma^t r_{i,t}\right].$$

To see how this setting fits into our model, we first define the local state/action and specify the parameters. Since each packet has a life span of d, and each user node receives at most one packet at a time step, we use a d-tuple $s_i = (e_1, e_2, \cdots, e_d) \in \mathcal{S}_i := \{0, 1\}^d$ to denote the local state of user node i. Specifically, e_j indicates whether user node u_i has a packet with remaining life span j in its queue. A local action of user node u_i is 2-tuple (l, y), which means sending the packet with remaining life span $l \in \{1, 2, \cdots, d\}$ to an access point $y \in Y_i$. Note that we define an empty action that does nothing at all. If a user node performs an action (l, y) when there is no packet with life span l in its queue, we view this as an empty action. This setting falls into the category we studied in Corollary 2.2, where long range links do not exist. Specifically, in this setting, the next local state of user node u_i depends on the current local states/actions in its 1-hop neighborhood ($\alpha_1 = 1$ in Corollary 2.2). We assume each user node can choose its action only based on its current local state ($\beta = 0$). Due to potential collisions, the local reward of user u_i also depends on the states/actions in

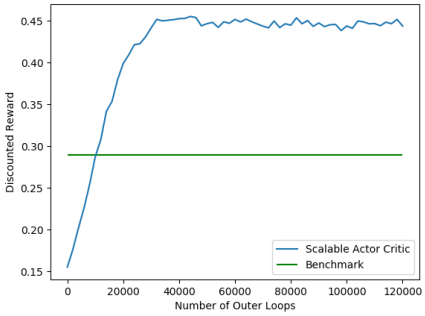

Figure 2: Discounted reward in the training process. 5×5 grid, 1 user per grid.

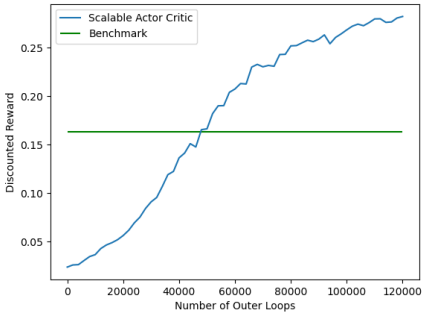

Figure 3: Discounted reward in the training process. 3×4 grid, 2 users per grid.

its 1-hop neighborhood ($\alpha_2 = 1$ in Corollary 2.2). Though this is a static setting, note that the results of [38] do not apply.

The detailed setting we use is as follows. We consider the setting where the user nodes are located in $h \times w$ grids (see Fig. 1). There are c user nodes in each grid, and each user can send packets to an access point on the corner of its grid. We set the initial life span $d = 2$, the arrival probability $q = 0.5$, and the discounted factor $\gamma = 0.7$. The successful transmission probability p_i for each access point y_i is sampled uniformly randomly from $[0, 1]$. We run the Scalable Actor Critic algorithm with parameter $\kappa = 1$ to learn a localized stochastic policy in two cases $(h, w, c) = (5, 5, 1)$ (see Fig. 2) and $(h, w, c) = (3, 4, 2)$ (see Fig. 3). For comparison, we use a benchmark based on the localized ALOHA protocol [39]. Specifically, the benchmark policy works as following: At time step t, each user node u_i takes the empty action with a certain probability p'; otherwise, it sends the packet with the minimum remaining life span to a random access point in Y_i, with the probability proportional to the successful transmission probability of this access point and inverse proportional to the number of users sharing this access point. In Fig. 2 and Fig. 3, we have tuned the parameter p' to find the one with the highest discounted reward.

As shown in Fig. 2 and Fig. 3, starting from the initial policy that chooses an local action uniformly at random, the Scalable Actor Critic algorithm with parameter $\kappa = 1$ can learn a policy that performs better than the benchmark. As a remark, the benchmark policy requires the set $\{p_i\}_{1 \le i \le m}$, the probability of successful transmission, as input. Moreover, in the benchmark policy, the probability of performing an empty action also needs to be tuned manually. In contrast, the Scalable Actor Critic algorithm can learn a better policy without these specific inputs by interacting with the system.

A.2 Spreading Networks

We consider a spreading network with n agents and an underlying graph $\mathcal{G}$. See Fig. 4 for an illustration of $n = wh$ agents on a $w \times h$ grid network. For each agent i, the local state/action space is given by $\mathcal{S}_i = \{0, 1\}$ and $\mathcal{A}_i = \{0, 1\}$. To make the discussion more concrete, in the following we present the spreading network model in the context of SIS epidemic network. This version of the SIS model has been studied in, for example, [40]. Our setting is more general and can be generalized to other types of spreading networks like opinion networks, social networks, etc. At time step t, the local state $s_i(t) = 0$ means agent i is "susceptible", while the local state $s_i(t) = 1$ means the agent i is "infected". By taking action $a_i(t) = 1$, agent i can suppress its infection probability at the expense of incurring an action cost. In the meantime, agent i will incur an infection cost if $s_i(t) = 1$. The interaction among agents is modeled by a set of undirected links, where two agents can affect each other if they are connected by a link. To model the influence of physical distance on the pattern of social contact, we assume the short range links occur more frequently than long range links. An illustration of the spreading network is shown in Fig. 4 (a), where the black nodes denote the agents with state 1; the white nodes denote the agents with state 0; the blue edges denote the set of active links at some time step.

Mathematically, the model can be described as follows. At each time step t, each agent i can decide her/his local action $a_i(t)$ based on the information of local states in the 1-hop neighborhood N_i^1, i.e.,

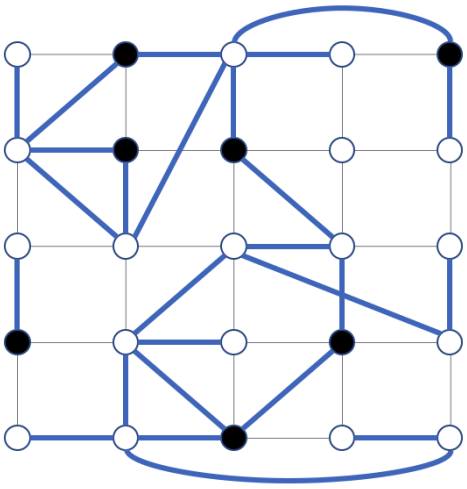

Figure 4: An illustration of the spreading network with 25 agents on a 5×5 grid network. The black nodes denote "infected" agents; The white nodes denote "susceptible" agents; The blue edges denote the active links at some time step.

$\beta = 1$. The local reward $r_i(t)$ is a function of the local state $s_i(t)$ and the local action $a_i(t)$, i.e., L_t^r is static and only contains self loops. Specifically, we define

$$r_i(t) = -c_i^{(a)} \mathbf{1}(a_i(t) = 1) - c_i^{(s)} \mathbf{1}(s_i(t) = 1),$$

where $\left(c_i^{(s)}, c_i^{(a)}\right)$ are parameters associated with agent i and can be different among agents. As mentioned earlier, $c_i^{(s)}$ penalizes the agent for being "infected", while $c_i^{(a)}$ is the cost of taking epidemic control measure. The stage reward is the sum of these two costs.

To describe the state transition rule, we first define the way the active link set L_t^s is generated: independently for each pair of agents $(i, j) \in \mathcal{N} \times \mathcal{N}$ with $i \neq j$, with probability $2^{-d_{\mathcal{G}}(i,j)}$, we include edges (i, j) and (j, i) in the set L_t^s; otherwise, neither edge is included in the set, i.e. $(i, j), (j, i) \notin L_t^s$. Given L_t^s, the next local state $s_i(t + 1)$ is sampled from a distribution that depends on the local states in $N_i(L_t^s)$. Specifically, define the quantities

$$n_i(t) = |\{j \mid j \in N_i(L_t) \setminus \{i\}, s_j(t) = 1, a_j(t) = 0\}|,$$
$$m_i(t) = |\{j \mid j \in N_i(L_t) \setminus \{i\}, s_j(t) = 1, a_j(t) = 1\}|.$$

Then, the probability that $s_i(t + 1) = 0$ is given by

$$P(s_i(t+1) = 0 \mid s_{N_i(L_t)}, a_{N_i(L_t)}) = \begin{cases} p_i^{(r)} & \text{if } s_i(t) = 1; \\ \left(1 - p_i^{(h)}\right)^{n_i(t)} \left(1 - p_i^{(m)}\right)^{m_i(t)} & \text{if } s_i(t) = 0, a_i(t) = 1; \\ \left(1 - p_i^{(m)}\right)^{n_i(t)} \left(1 - p_i^{(l)}\right)^{m_i(t)} & \text{if } s_i(t) = 0, a_i(t) = 0, \end{cases}$$

where $\left(p_i^{(r)}, p_i^{(h)}, p_i^{(m)}, p_i^{(l)}\right)$ are parameters associated with agent i and can be different among agents. Due to control actions, we assume $p_i^{(h)} > p_i^{(m)} > p_i^{(l)}$. This provides the transition rule, and the underlying intuition is that the local state of agent i turns from "infected" ($s_i(t) = 1$) to "susceptible" ($s_i(t+1) = 0$) with a fixed recovering probability $p_i^{(r)}$; the probability that agent i turns from "susceptible" ($s_i(t) = 0$) to "infected" ($s_i(t + 1) = 1$) depends on the number of neighboring agents in the active link set that are already infected, and further, whether agent i or the nearby agents j take epidemic control measures ($a_i(t) = 1, a_j(t) = 1$) or not. Roughly speaking, the more nearby infected agents, the more likely agent i will become infected; however, if epidemic control measures are taken by agent i and nearby agents in $N_i(L_t^s)$, the probability of agent i getting infected will be smaller.

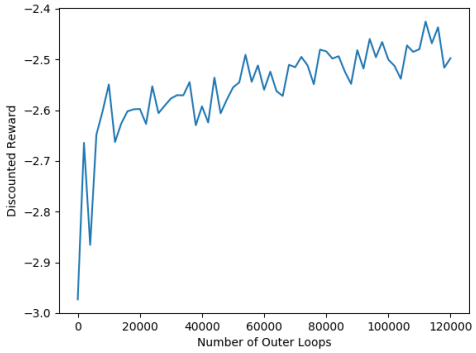

Figure 5: Discounted reward in the training process. 5×5 grid.

We run the Scalable Actor Critic algorithm with parameter $\kappa = 1$ to learn a localized stochastic policy in the case $(h, w) = (5, 5)$ (Fig. 5). For each agent i, parameters $\left(c_i^{(s)}, c_i^{(a)}, p_i^{(r)}, p_i^{(h)}\right)$ are sampled independently from the distribution

$$c_i^{(s)} \sim U[1.0, 3.0], c_i^{(a)} \sim U[0.01, 0.20], p_i^{(r)} \sim U[0.1, 0.5], p_i^{(h)} \sim U[0.5, 0.9],$$

and we set $p_i^{(m)} = p_i^{(h)}/4, p_i^{(l)} = p_i^{(m)}/4$. At time step 0, for each $i \in \mathcal{N}$, we initialize local state $s_i(0)$ to be 1 with probability 0.3.

B Stochastic Networked MARL

B.1 Proof of Theorem 2.1

For ease of exposition, let A, B be two subsets of the agent set $\mathcal{N}$ and we use $A \xrightarrow{\tau} B$ to denote the event that there exists a chain

$$j_0^a \xrightarrow{L_0^s} j_1^s \xrightarrow{L^a} j_1^a \xrightarrow{L_1^s} \cdots \xrightarrow{L_{\tau-1}^s} j_\tau^s \xrightarrow{L^a} j_\tau^a,$$

whose head and tail satisfies $j_0^a \in A$ and $j_\tau^a \in B$.

Given a sequence of active link sets $\{L_t^s\}_{t=0}^\infty$ and under fixed global policy θ, we say the information at set $A \subseteq \mathcal{N}$ spread to another set $B \subseteq \mathcal{N}$ in τ time steps (denoted by $I(A) \xrightarrow{\tau} I(B)$) if there exists (s, a) and (s', a') such that $(s_{\mathcal{N}\setminus A}, a_{\mathcal{N}\setminus A}) = (s'_{\mathcal{N}\setminus A}, a'_{\mathcal{N}\setminus A})$ and the distribution of $(s_B(\tau), a_B(\tau))$ given $(s(0), a(0)) = (s, a)$ is different with that given $(s(0), a(0)) = (s', a')$.

We show by induction that $I(A) \xrightarrow{\tau} I(B)$ happens only if $A \xrightarrow{\tau} B$ happens.

If $\tau = 0$, since $I(A) \xrightarrow{0} I(B)$, we see that $A \cap B \neq \emptyset$. Therefore, we can let j_0^a be any agent in $A \cap B$. Hence we also have $A \xrightarrow{0} B$.

Suppose the statement holds for $\tau = t$. When $\tau = t + 1$, suppose that $I(A) \xrightarrow{t+1} I(B)$. Define sets

$$B' := \{j \in \mathcal{N} \mid \exists k \in B, s.t. j \xrightarrow{L^a} k\}, B'' := \{j \in \mathcal{N} \mid \exists k \in B', s.t. j \xrightarrow{L_t^s} k\}.$$

Notice that $B \subseteq B' \subseteq B''$. By the definition of transition probability and policy dependence, we know that the distribution of $a_B(t+1)$ is decided by $s_{B'}(t+1)$, and the distribution of $s_{B'}(t+1)$ is decided by $(s_{B''}(t), a_{B''}(t))$. Therefore, we must have $I(A) \xrightarrow{t} I(B'')$. By the induction hypothesis, we have $A \xrightarrow{t} B''$, which further implies $A \xrightarrow{t+1} B$. This finishes the induction.

Given a sequence of active link sets $\{(L_t^s, L_t^r)\}$, we use $\pi_{t,i}$ to denote the distribution of $\left(s_{N_i(L_t^r)}(t), a_{N_i(L_t^r)}(t)\right)$ given that $(s(0), a(0)) = (s, a)$; we use $\pi'_{t,i}$ to denote the distribution of $\left(s_{N_i(L_t^r)}(t), a_{N_i(L_t^r)}(t)\right)$ given that $(s(0), a(0)) = (s', a')$. We notice that $\pi_{t,i} \neq \pi'_{t,i}$ happens only

if $I(N^{\kappa}_{-i}) \xrightarrow{t} I(N_i(L^r_t))$, which is true only if $N^{\kappa}_{-i} \xrightarrow{t} N_i(L^r_t)$. Recall that $X_i(\kappa)$ is defined as the smallest t such that $N^{\kappa}_{-i} \xrightarrow{t} N_i(L^r_t)$ holds. Hence, we obtain that

$$
\begin{aligned}
&\left|Q^{\theta}_i(s,a) - Q^{\theta}_i(s',a')\right| \\
&\leq \mathbb{E}_{\{(L^s_t, L^r_t)\}} \sum_{t=0}^{\infty} \left|\gamma^t \mathbb{E}_{\pi_{t,i}} r_i(s_{N_i(L^r_t)}, a_{N_i(L^r_t)}) - \gamma^t \mathbb{E}_{\pi'_{t,i}} r_i(s_{N_i(L^r_t)}, a_{N_i(L^r_t)})\right| \\
&\leq \mathbb{E}_{\{(L^s_t, L^r_t)\}} \sum_{t=X_i(\kappa)}^{\infty} \left|\gamma^t \mathbb{E}_{\pi_{t,i}} r_i(s_{N_i(L^r_t)}, a_{N_i(L^r_t)}) - \gamma^t \mathbb{E}_{\pi'_{t,i}} r_i(s_{N_i(L^r_t)}, a_{N_i(L^r_t)})\right| \\
&\leq \frac{1}{1-\gamma} \mathbb{E}\left[\gamma^{X_i(\kappa)}\right],
\end{aligned}
$$

where we use the definition of $X_i(\kappa)$ in the second step.

B.2 Proof of Corollary 2.2

Given a sequence of active link sets $\{(L^s_t, L^r_t)\}$, let $t = X_i(\kappa)$. By the definition of $X_i(\kappa)$, we assume that a chain of agents

$$
j^a_0 \xrightarrow{L^s_0} j^s_1 \xrightarrow{L^a} j^a_1 \xrightarrow{L^s_1} \cdots \xrightarrow{L^s_{t-1}} j^s_t \xrightarrow{L^a} j^a_t
$$

satisfies $j^a_0 \in N^{\kappa}_{-i}$ and $j^a_t \xrightarrow{L^r_t} i$.

By the triangle inequality and the assumptions of Lemma 2.2, we obtain that

$$
\begin{aligned}
d_{\mathcal{G}}(j^a_0, i) &\leq \sum_{\tau=0}^{t-1} \big(d_{\mathcal{G}}(j^a_\tau, j^s_{\tau+1}) + d_{\mathcal{G}}(j^s_{\tau+1}, j^a_{\tau+1})\big) + d_{\mathcal{G}}(j^a_t, i) \\
&\leq t(\beta + \alpha_1) + \alpha_2.
\end{aligned}
$$

Therefore, we see that t is lower bounded by $\frac{\kappa - \alpha_2}{\beta + \alpha_1}$, which also gives a lower bound of $X_i(\kappa)$.

B.3 Proof of Theorem 2.3

To simplify notation, we adopt the same notations as in the proof of Theorem 2.1 (Appendix B.1). Specifically, recall that we use $A \xrightarrow{\tau} B$ to denote the event that there exists a chain

$$
j^a_0 \xrightarrow{L^s_0} j^s_1 \xrightarrow{L^a} j^a_1 \xrightarrow{L^s_1} \cdots \xrightarrow{L^s_{\tau-1}} j^s_\tau \xrightarrow{L^a} j^a_\tau,
$$

whose head and tail satisfies $j^a_0 \in A$ and $j^a_\tau \in B$. We will use ∂N^{κ}_i to denote the set of neighbors whose distance to i is κ, i.e., $\partial N^{\kappa}_i := \{j \in \mathcal{N} \mid d_{\mathcal{G}}(i,j) = \kappa\} = N^{\kappa}_i \setminus N^{\kappa-1}_i$. Define $a_\kappa := \mathbb{E}\left[\gamma^{X_i(\kappa-1)}\right]$. Define function cat (concatenation) such that for a pair of active link sets (L^s, L^a), $(x,y) \in cat(L^s, L^a)$ if and only if $\exists z \in \mathcal{N}$ such that $x \xrightarrow{L^s} z \xrightarrow{L^a} y$.

Before proving Theorem 2.3, we first give an upper bound for the sum of an infinite sequence $\{poly(k+i) \cdot \nu^i\}_{i \in \mathbb{N}}$, where $\nu < 1$ is a positive constant. This result is helpful for showing an upper bound of $P(N^{\kappa}_{-i} \to N^j_i)$.

Lemma B.1. *If $m \in \mathbb{N}^*$ and $0 < \nu < 1$ are constants, for all $k \geq \frac{2m}{\ln(1/\nu)}$, we have*

$$
\sum_{i=0}^{\infty} (k+i)^m \nu^i \leq \frac{1}{1-\sqrt{\nu}} \cdot k^m.
$$

Proof of Lemma B.1. Define function $f : \mathbb{R}^+ \cup \{0\} \to \mathbb{R}^+$ as

$$
f(t) = (k+t)^m \cdot \nu^{t/2}.
$$

The derivative of function f is given by

$$f'(t) = (k+t)^{m-1} \cdot \nu^{t/2} \bigg(m + \frac{1}{2} \ln \nu \cdot (k+t) \bigg).$$

Since $k \geq \frac{2m}{\ln(1/\nu)}$, $f'(t) \leq 0$ holds for all $t \geq 0$, hence we have $f(t) \leq f(0) = k^m$.

Therefore, we obtain that

$$\begin{aligned}
\sum_{i=0}^{\infty} (k+i)^m \nu^i &\leq \sum_{i=0}^{\infty} f(i) \cdot \nu^{i/2} \\
&\leq k^m \sum_{i=0}^{\infty} \nu^{i/2} \\
&\leq \frac{1}{1-\sqrt{\nu}} \cdot k^m.
\end{aligned}$$

□

Now we come back to the proof of Theorem 2.3.

By union bound, we derive an upper bound of the probability that a link (x, y) is in $cat(L^s, L^a)$. Suppose $d \in \mathbb{N}$ is constant that satisfies $d_{\mathcal{G}}(x, y) \geq d$, and the probability P is taken over $(L^s, L^r) \sim \mathcal{D}$:

$$\begin{aligned}
P((x,y) \in cat(L^s, L^a)) &= P(\exists z \in \mathcal{N}, (x,z) \in L^s \wedge (z,y) \in L^a) \\
&\leq \sum_{z: d_{\mathcal{G}}(z,y) \leq \beta} P((x,z) \in L^s) \\
&\leq c_0 (\beta+1)^{n_0+1} \cdot c \lambda^{d-\beta} \\
&= c_g \lambda^d, \qquad (7)
\end{aligned}$$

where constant c_g is defined as $c_0 c (\beta+1)^{n_0+1} \lambda^{-\beta}$.

By the assumption on the size of κ-hop neighborhood, we know that for some constant c_0 and $n_0 \in \mathbb{N}^*$, $|\partial N_i^\kappa| \leq c_0(\kappa+1)^{n_0}$ holds for all $\kappa \geq 1$. Let $n_1 := 2n_0$. With the help of Lemma B.1, we show that for some constant $c_2 > 0$, $P\Big(N_{-i}^{\kappa-1} \xrightarrow{1} \partial N_i^j\Big)$ is upper bounded by $c_2(\kappa+1)^{n_1} \lambda^{\kappa-j}$ for all $j \leq \kappa - 1$ when $\kappa \geq \frac{2n_0}{\ln(1/\lambda)}$:

$$\begin{aligned}
P\Big(N_{-i}^{\kappa-1} \xrightarrow{1} \partial N_i^j\Big) &\leq P\Big(\exists x \in N_{-i}^{\kappa-1}, y \in \partial N_i^j \text{ s.t. } (x,y) \in cat(L^s, L^a)\Big) && \text{(8a)} \\
&\leq \sum_{q=0}^{\infty} P\Big(\exists x \in \partial N_i^{\kappa+q}, y \in \partial N_i^j \text{ s.t. } (x,y) \in cat(L^s, L^a)\Big) && \text{(8b)} \\
&\leq \sum_{q=0}^{\infty} \sum_{x \in \partial N_i^{\kappa+q}, y \in \partial N_i^j} P((x,y) \in cat(L^s, L^a)) && \text{(8c)} \\
&\leq \sum_{q=0}^{\infty} \sum_{x \in \partial N_i^{\kappa+q}, y \in \partial N_i^j} c_g \lambda^{(\kappa+q-j)} && \text{(8d)} \\
&\leq c_g \lambda^{\kappa-j} \sum_{q=0}^{\infty} \Big|\partial N_i^{\kappa+q}\Big| \cdot \Big|\partial N_i^j\Big| \cdot \lambda^q \\
&\leq c_g c_0^2 (\kappa+1)^{n_0} \lambda^{\kappa-j} \sum_{q=0}^{\infty} (\kappa+q+1)^{n_0} \lambda^q && \text{(8e)} \\
&\leq c_2 (\kappa+1)^{n_1} \lambda^{\kappa-j}, && \text{(8f)}
\end{aligned}$$

where we use the definition of $N_{-i}^{\kappa-1} \xrightarrow{1} \partial N_i^j$ in (8a); we use union bound in (8b) and (8c); we use the fact that $d_{\mathcal{G}}(x,y) \geq \kappa + q - j, \forall x \in \partial N_i^{\kappa+q}, y \in \partial N_i^j$ and (7) in (8d); we use the bounds

$\left|\partial N_i^j\right| \le c_0 j^{n_0} \le c_0 \kappa^{n_0}$ and $\left|\partial N_i^{\kappa+q}\right| \le c_0(\kappa+q)^{n_0}$ in (8e); we define $c_2 := \frac{c_g c_0^2}{1-\sqrt{\lambda}}$ and use Lemma B.1 in (8f).

Let constants c_3 and q be defined as

$$c_3 := \frac{1}{2}\sqrt[4]{\lambda}(1-\sqrt{\lambda})\left(\frac{1}{\sqrt{\gamma}}-1\right),$$
$$q := \frac{1}{\ln(1/\lambda)}\max\{(\ln c_2 - \ln c_3 - 2\ln(1-\sqrt{\gamma})), (2n_1+4)\},$$

and define function $p(\kappa) := [q(1+\ln(\kappa+1))]+1$. We can find $\kappa_0 \in \mathbb{Z}^+$ such that $p(\kappa) \ge \kappa$ for all $\kappa \le \kappa_0$, and $p(\kappa) > \kappa$ for all $\kappa > \kappa_0$.

Let ρ be a constant such that $1 > \rho > \max\{\gamma^{1/(2q)}, \sqrt[4]{\lambda}\}$. Let $C := \rho^{-\max\{q+1, \frac{2n_0}{\ln(1/\lambda)}\}}$. Recall that we define $a_\kappa := \mathbb{E}\left[\gamma^{X_i(\kappa-1)}\right]$, where $X_i(\kappa-1)$ denotes the smallest t such that $N_{-i}^{\kappa-1} \xrightarrow{t} N_i(L_t^r)$ holds. Now we show by induction that

$$a_\kappa \le C\rho^{\kappa/(1+\ln(\kappa+1))}, \forall \kappa \ge 1. \tag{9}$$

Since $a_\kappa \le 1$, (9) clearly holds when $\kappa \le \kappa_0$. To see this, recall that we have $\kappa \le p(\kappa)$ and $C \ge \rho^{-(q+1)}$ by definition, thus the right hand side of (9) can be lower bounded by

$$C\rho^{\kappa/(1+\ln(\kappa+1))} \ge \rho^{-(q+1)} \cdot \rho^{p(\kappa)/(1+\ln(\kappa+1))} \ge \rho^{-(q+1)} \cdot \rho^{q+1} = 1.$$

When $\kappa > \kappa_0$, we have $\kappa > p(\kappa)$. Recall that $a_\kappa := \mathbb{E}\left[\gamma^{X_i(\kappa-1)}\right]$. Notice that $X_i(\kappa-1) = 0$ if and only if $N_{-i}^{\kappa-1} \cap N_i(L_0^r) \ne \emptyset$. To simplify the notation, we denote the event $N_{-i}^{\kappa-1} \cap N_i(L_0^r) \ne \emptyset$ by E_0. Using this and the idea of dynamic programming, we see that

$$\begin{aligned} a_\kappa &\le \gamma\left(P\{\left(\neg N_{-i}^{\kappa-1} \xrightarrow{1} N_i^{\kappa-1}\right) \wedge \neg E_0\}a_\kappa + \sum_{j=0}^{\kappa-1} P\{\left(N_{-i}^{\kappa} \xrightarrow{1} \partial N_i^j\right) \wedge \left(\neg N_{-i}^{\kappa} \xrightarrow{1} N_i^{j-1}\right) \wedge \neg E_0\}a_j\right) \\ &\quad + P(E_0) \\ &\le \gamma\left(P\{\neg N_{-i}^{\kappa-1} \xrightarrow{1} N_i^{\kappa-1}\}a_\kappa + \sum_{j=0}^{\kappa-1} P\{\left(N_{-i}^{\kappa} \xrightarrow{1} \partial N_i^j\right) \wedge \left(\neg N_{-i}^{\kappa} \xrightarrow{1} N_i^{j-1}\right)\}a_j\right) + P(E_0), \end{aligned} \tag{10}$$

where the probability P are taken over $(L_0^s, L_0^r) \sim D$.

Since $\kappa \ge p(\kappa) \ge q \ge \frac{2n_1}{\ln(1/\lambda)} \ge \frac{2n_0}{\ln(1/\lambda)}$, by Lemma B.1, we see that

$$P(E_0) = P\{\exists j \in N_{-i}^{\kappa-1} \text{ s.t. } (j,i) \in L^r\} \le \sum_{q=0}^{\infty} cc_0(\kappa+q+1)^{n_0}\lambda^{\kappa+q} \le \frac{cc_0}{1-\sqrt{\lambda}}(\kappa+1)^{n_0+1}\lambda^\kappa.$$

Substituting this into (10) and rearranging the terms gives

$$\begin{aligned} \left(1-\gamma P\{\neg N_{-i}^{\kappa-1} \xrightarrow{1} N_i^{\kappa-1}\}\right)a_\kappa \le \gamma &\sum_{j=\kappa-p(\kappa)+1}^{\kappa-1} P\{\left(N_{-i}^{\kappa-1} \xrightarrow{1} \partial N_i^j\right) \wedge \left(\neg N_{-i}^{\kappa-1} \xrightarrow{1} N_i^{j-1}\right)\}a_j \\ &+ \gamma \sum_{j=0}^{\kappa-p(\kappa)} P\{\left(N_{-i}^{\kappa-1} \xrightarrow{1} \partial N_i^j\right) \wedge \left(\neg N_{-i}^{\kappa-1} \xrightarrow{1} N_i^{j-1}\right)\}a_j \\ &+ \frac{cc_0}{1-\sqrt{\lambda}}(\kappa+1)^{n_0+1}\lambda^\kappa. \end{aligned} \tag{11}$$

For simplicity, we define $\rho_\kappa := \rho^{1/(1+\ln(\kappa+1))}$. By the induction assumption, we have that

$$a_j \le C\rho^{j/(\ln(j+1)+1)} \le C\rho^{j/(\ln(\kappa+1)+1)} = C\rho_\kappa^j.$$

Substituting this into (11) gives that

$$\Big(1-\gamma P\{\neg N_{-i}^{\kappa-1} \xrightarrow{1} N_i^{\kappa-1}\}\Big)a_\kappa \le C\gamma \sum_{j=\kappa-p(\kappa)+1}^{\kappa-1} P\{\Big(N_{-i}^{\kappa-1} \xrightarrow{1} \partial N_i^j\Big) \wedge \Big(\neg N_{-i}^{\kappa-1} \xrightarrow{1} N_i^{j-1}\Big)\}\rho_\kappa^j + C\gamma \sum_{j=0}^{\kappa-p(\kappa)} P\{\Big(N_{-i}^{\kappa-1} \xrightarrow{1} \partial N_i^j\Big) \wedge \Big(\neg N_{-i}^{\kappa-1} \xrightarrow{1} N_i^{j-1}\Big)\}\rho_\kappa^j + \frac{c_0}{1-\sqrt{\lambda}}(\kappa+1)^{n_0+1}\lambda^\kappa. \tag{12}$$

By the definition of $p(\kappa)$ and q, we see that

$$\lambda^{-p(\kappa)} \ge \lambda^{-q(1+\ln(\kappa+1))} = \lambda^{-q}\cdot(\kappa+1)^{q\ln(1/\lambda)} \ge \frac{c_2}{c_3(1-\sqrt{\gamma})^2}\cdot(\kappa+1)^{n_1} \ge \frac{c_2}{c_3(1-\gamma)}\cdot(\kappa+1)^{n_1}.$$

Therefore, we obtain the upper bound

$$\begin{aligned} P\{\Big(N_{-i}^{\kappa-1} \xrightarrow{1} \partial N_i^j\Big) \wedge \Big(\neg N_{-i}^{\kappa} \xrightarrow{1} N_i^{j-1}\Big)\} &\le P\{N_{-i}^{\kappa-1} \xrightarrow{1} \partial N_i^j\} \\ &\le c_2(\kappa+1)^{n_1}\lambda^{(\kappa-j)} \\ &\le (1-\gamma)c_3\lambda^{(\kappa-p(\kappa)-j)}. \end{aligned}$$

Using this and divide both sides of (12) by $\Big(1-\gamma P\{\neg N_{-i}^{\kappa} \xrightarrow{1} N_i^{\kappa-1}\}\Big)$, we see that

$$\begin{aligned} a_\kappa \le{}& \gamma\Big(C\rho_\kappa^{\kappa-p(\kappa)+1} + Cc_3(\rho_\kappa^{\kappa-p(\kappa)} + \lambda^1\cdot\rho_\kappa^{\kappa-p(\kappa)-1} + \lambda^2\cdot\rho_\kappa^{\kappa-p(\kappa)-2} + \cdots)\Big) \\ &+ \frac{c_0}{(1-\gamma)(1-\sqrt{\lambda})}(\kappa+1)^{n_0+1}\lambda^\kappa, \end{aligned} \tag{13}$$

where we also use the fact that

$$\sum_{j=\kappa-p+1}^{\kappa-1} P\{\Big(N_{-i}^{\kappa-1} \xrightarrow{1} \partial N_i^j\Big) \wedge \Big(\neg N_{-i}^{\kappa-1} \xrightarrow{1} N_i^{j-1}\Big)\} \le 1-\gamma P\{\neg N_{-i}^{\kappa-1} \xrightarrow{1} N_i^{\kappa-1}\}.$$

By the definition of $p(\kappa), q$ and c_2, we have that

$$\lambda^{\frac{\kappa}{4}} \le \lambda^{\frac{p(\kappa)}{4}} \le (\kappa+1)^{-\frac{q\ln(1/\lambda)}{4}} \le (\kappa+1)^{-n_0-1}$$

and

$$\lambda^{\frac{\kappa}{2}} \le \lambda^{\frac{p(\kappa)}{2}} \le \lambda^{\frac{q}{2}} \le \frac{(1-\sqrt{\gamma})(1-\gamma)(1-\sqrt{\lambda})}{2c_0},$$

which implies

$$\lambda^{\frac{3\kappa}{4}} \le \frac{(1-\sqrt{\gamma})(1-\gamma)(1-\sqrt{\lambda})}{2c_0(\kappa+1)^{n_0+1}}. \tag{14}$$

Dividing both sides of (13) by $C\rho_\kappa^\kappa$ gives that

$$\begin{aligned} \frac{a_\kappa}{C\rho_\kappa^\kappa} &\le \gamma\left(\frac{1}{\rho_\kappa^{p(\kappa)-1}} + \frac{c_3}{\rho_\kappa^{p(\kappa)}}\cdot\frac{1}{1-(\lambda/\rho_\kappa)}\right) + \frac{c_0}{(1-\gamma)(1-\sqrt{\lambda})}(\kappa+1)^{n_0+1}\lambda^{\frac{3\kappa}{4}} && \text{(15a)} \\ &\le \gamma\left(\frac{1}{\rho^q} + \frac{1}{\rho^{q+1}}\cdot\frac{c_3}{1-\sqrt{\lambda}}\right) + \frac{1}{2}(1-\sqrt{\gamma}) && \text{(15b)} \\ &= \frac{\gamma}{\rho^q}\left(1+\frac{c_3}{\rho(1-\sqrt{\lambda})}\right) + \frac{1}{2}(1-\sqrt{\gamma}) \\ &\le \sqrt{\gamma}\cdot\frac{1}{2}\left(1+\frac{1}{\sqrt{\gamma}}\right) + \frac{1}{2}(1-\sqrt{\gamma}) && \text{(15c)} \\ &= 1, \end{aligned}$$

where we use $\rho_\kappa = \rho^{1/(1+\ln\kappa)} \ge \rho \ge \sqrt[4]{\lambda}$ in (15a); we use $\rho_\kappa \ge \sqrt[4]{\lambda}$, $p = [q(1+\ln\kappa)]+1$, and (14) in (15b); we use $c_3 = \sqrt{\lambda}(1-\sqrt{\lambda})(\sqrt{\gamma}-1) \le \rho(1-\sqrt{\lambda})(\sqrt{\gamma}-1)$ and $\rho \ge \gamma^{1/(2q)}$ in (15c).

B.4 Proof of Theorem 2.4

In the Critic part of Algorithm 1, since the policy is fixed to be $\theta(m)$, the pair (s,a) can be viewed as the state of a Markov chain $\mathcal{C}$, and $Q^{\theta(m)}(s,a)$ in the original MDP corresponds to the value function $V^*((s,a))$ on $\mathcal{C}$. Define the state aggregation map h such that $h((s,a)) = (s_{N_i^\kappa}, a_{N_i^\kappa})$. By the μ-decay property, we see that if $h((s,a)) = h((s',a'))$, then

$$|V^*((s,a)) - V^*((s',a'))| = \left|Q^{\theta(m)}(s,a) - Q^{\theta(m)}(s',a')\right| \leq \mu(\kappa).$$

Note that Assumption 2.1 implies that Assumption 3.1 holds for $\mathcal{C}$. Thus, we can apply Theorem 3.2 to finish the proof of Theorem 2.4.

B.5 Proof of Theorem 2.5

Before showing Theorem 2.5, we first state a theorem concerning the actor part of Algorithm 1. The proof is deferred to Appendix B.6.

Theorem B.2. *Under the same assumption as Theorem 2.5, suppose inner loop length T is sufficiently large such that $T+1 \geq \log_\gamma((1-\gamma)\mu(\kappa))$ and with probability at least $1-\frac{\delta}{2}$, the following inequality holds for all agents $i \in \mathcal{N}$:*

$$\sup_{m\leq M-1}\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left|Q_i^{\theta(m)}(s,a) - \hat{Q}^T(s_{N_i^\kappa}, a_{N_i^\kappa})\right| \leq \frac{\iota\mu(\kappa)}{1-\gamma},$$

where ι is a positive constant. Suppose the actor step size satisfies $\eta_m = \frac{\eta}{\sqrt{m+1}}$ with $\eta \leq \frac{1}{4W'}$. Define $C_M := \frac{2}{\eta(1-\gamma)} + \frac{8W^2\sqrt{\log M \log\frac{4}{\delta}} + 96W'W^2\eta\log M}{(1-\gamma)^4}$. Then, with probability at least $1-\delta$,

$$\frac{\sum_{m=0}^{M-1}\eta_m\|\nabla J(\theta(m))\|^2}{\sum_{m=0}^{M-1}\eta_m} \leq \frac{C_M}{\sqrt{M+1}} + \frac{2(2+\iota)W^2\mu(\kappa)}{(1-\gamma)^4}. \tag{16}$$

As a remark, note that the left hand side of (16) is a weighted average of the squared norm of the gradients $\nabla J(\theta(m))$. We say the algorithm has reached an $O(\epsilon)$-approximate stationary point if the left hand side of (16) is in the order of $O(\epsilon)$.

Now we come back to the proof of Theorem 2.5. Let constant $\iota = 2$ in Theorem B.2. By Theorem B.2, to satisfy

$$\frac{\sum_{m=0}^{M-1}\eta_m\|\nabla J(\theta(m))\|^2}{\sum_{m=0}^{M-1}\eta_m} \leq O(\epsilon),$$

it suffices to guarantee that

$$\frac{C_M}{\sqrt{M+1}} = O(\epsilon), \text{ and } \frac{2(2+\iota)W^2\mu(\kappa)}{(1-\gamma)^4} = O(\epsilon).$$

These can be satisfied by letting

$$M = \tilde{\Omega}\left(\epsilon^{-2}\left(\frac{(W')^2}{(1-\gamma)^2} + \frac{W^4(1+\log(1/\delta))}{(1-\gamma)^8}\right)\right), \mu(\kappa) = O\big(W^{-2}(1-\gamma)^4\epsilon\big).$$

To satisfy

$$\sup_{m\leq M-1}\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left|Q_i^{\theta(m)}(s,a) - \hat{Q}^T(s_{N_i^\kappa}, a_{N_i^\kappa})\right| \leq \frac{\iota\mu(\kappa)}{1-\gamma},$$

with probability at least $1-\frac{\delta}{2}$, by Corollary 2.4, it suffices to select T such that

$$\begin{aligned}
&\frac{1}{\sqrt{T+t_0}}\cdot\frac{40H}{(1-\gamma)^2}\sqrt{K_2\log T\left(\log\left(\frac{4f(\kappa)K_2T}{\delta}\right) + \log\log T\right)} \\
&\quad + \frac{1}{T+t_0}\cdot\frac{8}{(1-\gamma)^2}\max\{\frac{144K_2H\log T}{\sigma'(\kappa)} + C_3, 2K_2\log T + t_0\} \\
&\leq \frac{\mu(\kappa)}{1-\gamma}.
\end{aligned}$$

Recall that

$$\mu(\kappa) = O\big(W^{-2}(1-\gamma)^4\epsilon\big), H \geq \frac{2}{(1-\gamma)\sigma'(\kappa)}, t_0 = \max(4H, 2K_2 \log T).$$

Hence the required number of inner loop is

$$T = \tilde{\Omega}\bigg(\frac{W^4(K_2(\log f(\kappa) + \log(1/\delta) + 1) + K_1)}{\epsilon^2(1-\gamma)^{12}\sigma'(\kappa)^2}\bigg).$$

B.6 Proof of Theorem B.2

While Theorem 5 in [38] studies the error bound of Scalable Actor Critic as a whole, we want to decouple the effect of the inner loop and the outer loop in Theorem B.2. Our proof of Theorem B.2 uses similar techniques with the proof in [38], but we extend the analysis to a more general dependence model.

According to Algorithm 1, at iteration m, agent i performs gradient ascent by

$$\theta_i(m+1) = \theta_i(m) + \eta_m \hat{g}_i(m),$$

with step size $\eta_m = \frac{\eta}{\sqrt{m+1}}$. The approximate local gradient $\hat{g}_i(m)$ is given by

$$\hat{g}_i(m) = \sum_{t=0}^{T} \gamma^t \frac{1}{n} \sum_{j \in N_i^\kappa} \hat{Q}_j^{m,T}\Big(s_{N_j^\kappa}(t), a_{N_j^\kappa}(t)\Big) \nabla_{\theta_i} \log \zeta_i^{\theta_i(m)}\Big(a_i(t) \mid s_{N_i^\beta}(t)\Big).$$

Recall that the true local gradient is given by

$$\nabla_{\theta_i} J(\theta(m)) = \sum_{t=0}^{\infty} \mathbb{E}_{s\sim\pi_t^{\theta(m)}, a\sim\zeta_i^{\theta(m)}(\cdot|s)} \gamma^t Q^{\theta(m)}(s,a) \nabla_{\theta_i} \log \zeta^{\theta_i(m)}\Big(a_i(t) \mid s_{N_i^\beta}(t)\Big),$$

where we use π_t^θ to denote the distribution of global state $s(t)$ under fixed policy θ.

To bound $\|\hat{g}(m) - \nabla_\theta J(\theta(m))\|$, we define intermediate quantities $g(m)$ and $h(m)$ whose i'th component is given by

$$g_i(m) = \sum_{t=0}^{T} \gamma^t \frac{1}{n} \sum_{j \in N_i^\kappa} Q_j^{\theta(m)}(s(t), a(t)) \nabla_{\theta_i} \log \zeta_i^{\theta_i(m)}\Big(a_i(t) \mid s_{N_i^\beta}(t)\Big),$$

$$h_i(m) = \sum_{t=0}^{T} \mathbb{E}_{s\sim\pi_t^{\theta(m)}, a\sim\zeta^{\theta(m)}(\cdot|s)} \gamma^t \frac{1}{n} \sum_{j \in N_i^\kappa} Q_j^{\theta(m)}(s,a) \nabla_{\theta_i} \log \zeta_i^{\theta_i(m)}\Big(a_i(t) \mid s_{N_i^\beta}(t)\Big).$$

Lemma B.3. *We have almost surely, $\forall m \leq M$,*

$$\max(\|\hat{g}(m)\|, \|g(m)\|, \|h(m)\|, \|\nabla J(\theta(m))\|) \leq \frac{W}{(1-\gamma)^2}.$$

To show Lemma B.3, we only need to replace $\zeta_i^{\theta_i(m)}(a_i(t) \mid s_i(t))$ by $\zeta_i^{\theta_i(m)}\Big(a_i(t) \mid s_{N_i^\beta}(t)\Big)$ in the proof of Lemma 17 in [38].

Notice that

$$\hat{g}(m) - \nabla J(\theta(m)) = e^1(m) + e^2(m) + e^3(m),$$

where

$$e^1(m) := \hat{g}(m) - g(m), e^2(m) := g(m) - h(m), e^3(m) := h(m) - \nabla J(\theta(m)).$$

To bound $\|\hat{g}(m) - \nabla J(\theta(m))\|$, we only need to bound $e_1(m), e_2(m), e_3(m)$ separately.

Lemma B.4. *With probability at least $1 - \frac{\delta}{2}$, we have*

$$\sup_{0\leq m\leq M-1} \big\|e^1(m)\big\| \leq \frac{\iota W \mu(\kappa)}{(1-\gamma)^2}.$$

Proof of Lemma B.4. By the assumption that

$$\sup_{m\le M-1}\sup_{i\in\mathcal{N}}\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left|Q_i^{\theta(m)}(s,a)-\hat{Q}^T(s_{N_i^\kappa},a_{N_i^\kappa})\right|\le\frac{\iota\cdot\mu(\kappa)}{1-\gamma},$$

we have for all $m\le M-1$ and $i\in\mathcal{N}$,

$$\begin{aligned}
&\|\hat{g}_i(m)-g_i(m)\|\\
\le&\left\|\sum_{t=0}^{T}\gamma^t\frac{1}{n}\sum_{j\in N_i^\kappa}\left[\hat{Q}_j^{m,T}\left(s_{N_j^\kappa}(t),a_{N_j^\kappa}(t)\right)-Q_j^{\theta(m)}(s(t),a(t))\right]\nabla_{\theta_i}\log\zeta_i^{\theta_i(m)}\left(a_i(t)\mid s_{N_i^\beta}(t)\right)\right\|\\
\le&\sum_{t=0}^{T}\gamma^t\frac{1}{n}\sum_{j\in N_i^\kappa}\left|\hat{Q}_j^{m,T}\left(s_{N_j^\kappa}(t),a_{N_j^\kappa}(t)\right)-Q_j^{\theta(m)}(s(t),a(t))\right|\left\|\nabla_{\theta_i}\log\zeta_i^{\theta_i(m)}\left(a_i(t)\mid s_{N_i^\beta}(t)\right)\right\|\\
\le&\sum_{t=0}^{T}\gamma^t\frac{\iota\cdot\mu(\kappa)}{1-\gamma}W_i\\
<&\frac{2\iota W_i\cdot\mu(\kappa)}{(1-\gamma)^2}.
\end{aligned}$$

Combining all n dimensions finishes the proof. □

Lemma B.5. *With probability at least $1-\frac{\delta}{2}$, we have*

$$\left|\sum_{m=0}^{M-1}\eta_m\langle\nabla J(\theta(m)),e^2(m)\rangle\right|\le\frac{2W^2}{(1-\gamma)^4}\sqrt{2\sum_{m=0}^{M-1}\eta_m^2\log\frac{4}{\delta}}.$$

To show Lemma B.5, we only need to replace $\zeta_i^{\theta_i(m)}(a_i(t)\mid s_i(t))$ by $\zeta_i^{\theta_i(m)}\left(a_i(t)\mid s_{N_i^\beta}(t)\right)$ in the proof of Lemma 19 in [38].

Lemma B.6. *When $T+1\ge\log_\gamma((1-\gamma)\mu(\kappa))$, we have almost surely*

$$\left\|e^3(m)\right\|\le\frac{2W\mu(\kappa)}{1-\gamma}.$$

To show Lemma B.6, we only need to replace $\zeta_i^{\theta_i(m)}(a_i(t)\mid s_i(t))$ with $\zeta_i^{\theta_i(m)}\left(a_i(t)\mid s_{N_i^\beta}(t)\right)$ and replace $c\rho^{\kappa+1}$ with $\mu(\kappa)$ in the proof of Lemma 20 in [38].

Now we come back to the proof of Theorem B.2. Using the identical steps with the proof of Theorem 5 in [38], we can obtain that (equation (44) in [38])

$$\sum_{m=0}^{M-1}\frac{1}{2}\eta_m\|\nabla J(\theta(m))\|^2\le J(\theta(m))-J(\theta(0))-\sum_{m=0}^{M-1}\eta_m\epsilon_{m,0}+\sum_{m=0}^{M-1}\eta_m\epsilon_{m,1}+\sum_{m=0}^{M-1}\eta_m^2\epsilon_{m,2},\quad(17)$$

where

$$\begin{aligned}
\epsilon_{m,0}&=\langle\nabla J(\theta(m)),e^2(m)\rangle,\\
\epsilon_{m,1}&=\|\nabla J(\theta(m))\|(\left\|e^1(m)\right\|+\left\|e^3(m)\right\|),\\
\epsilon_{m,2}&=2W'(\left\|e^1(m)\right\|^2+\left\|e^2(m)\right\|^2+\left\|e^3(m)\right\|^2).
\end{aligned}$$

By Lemma B.5, we have with probability at least $1-\frac{\delta}{2}$,

$$\left|\sum_{m=0}^{M-1}\eta_m\epsilon_{m,0}\right|\le\frac{2W^2}{(1-\gamma)^4}\sqrt{2\sum_{m=0}^{M-1}\eta_m^2\log\frac{4}{\delta}}.\quad(18)$$

By Lemma B.4 and Lemma B.6, we have with probability at least $1 - \frac{\delta}{2}$,

$$\sup_{m \le M-1} \epsilon_{m,1} \le \frac{W}{(1-\gamma)^2} \left(\sup_{m \le M-1} \left\| e^1(m) \right\| + \sup_{m \le M-1} \left\| e^3(m) \right\| \right) \le \frac{(2+\iota) W^2 \mu(\kappa)}{(1-\gamma)^4}. \tag{19}$$

By Lemma B.3, we have almost surely $\max(\left\| e^1(m) \right\|, \left\| e^2(m) \right\|, \left\| e^3(m) \right\|) \le 2\frac{W}{(1-\gamma)^2}$, and hence almost surely

$$\sup_{m \le M-1} \epsilon_{m,2} = 2W' \Big(\left\| e^1(m) \right\|^2 + \left\| e^2(m) \right\|^2 + \left\| e^3(m) \right\|^2 \Big) \le \frac{24 W' W^2}{(1-\gamma)^4}. \tag{20}$$

By union bound, (18), (19), and (20) hold simultaneously with probability $1 - \delta$. Combining them with (17) gives

$$\begin{aligned} &\frac{\sum_{m=0}^{M-1} \eta_m \| \nabla J(\theta(m)) \|^2}{2 \sum_{m=0}^{M-1} \eta_m} \\ &\le \frac{(J(\theta(M)) - J(\theta(0))) + \left| \sum_{m=0}^{M-1} \eta_m \epsilon_{m,0} \right| + \sup_{m \le M-1} \epsilon_{m,2} \sum_{m=0}^{M-1} \eta_m^2}{\sum_{m=0}^{M-1} \eta_m} + 2 \sup_{m \le M-1} \epsilon_{m,1}. \end{aligned} \tag{21}$$

We can use identical steps with the proof of Theorem 5 in [38] to bound the first term in (21), and use (19) to bound the second term in (21). This completes the proof.

C Stochastic Approximation Scheme

C.1 Contraction of the Update Operator

To show that the equation $\Pi F(\Phi x) = x$ has a unique solution x^*, by the Banach–Caccioppoli fixed-point theorem, it suffices to show that operator $\Pi F(\Phi \cdot)$ is a γ-contraction in $\|\cdot\|_v$.

Proposition C.1. *If Assumption 3.2 holds, operator $\Pi F(\Phi \cdot)$ is a contraction in $\|\cdot\|_v$, i.e., for any $x, y \in \mathbb{R}^{\mathcal{M}}$, $\|\Pi F(\Phi x) - \Pi F(\Phi y)\|_v \le \gamma \|x - y\|_v$.*

To prove this proposition, we first show both operator Π and operator Φ are non-expansive in $\|\cdot\|_v$ before combining them with F.

Proof of Proposition C.1. We first show that operator Π is non-expansive in $\|\cdot\|_v$, i.e. for any $x, y \in \mathbb{R}^{\mathcal{N}}$, we have

$$\|\Pi x - \Pi y\|_v \le \|x - y\|_v. \tag{22}$$

Since Π is a linear operator, it suffices to show that for any $x \in \mathbb{R}^{\mathcal{N}}$, $\|\Pi x\|_v \le \|x\|_v$.

Recall that $\forall j \in \mathcal{M}, h^{-1}(j) := \{i \in \mathcal{N} \mid h(i) = j\}$. Using this notation, the j th element of vector Πx is given by

$$(\Pi x)_j = \frac{1}{\sum_{i \in h^{-1}(j)} d_i} \big(\Phi^\top D x \big)_j = \frac{1}{\sum_{i \in h^{-1}(j)} d_i} \cdot \sum_{i \in h^{-1}(j)} d_i x_i.$$

Hence we see that

$$\frac{\left| (\Pi x)_j \right|}{v_j} \le \frac{1}{\sum_{i \in h^{-1}(j)} d_i} \cdot \sum_{i \in h^{-1}(j)} d_i \frac{|x_i|}{v_j} \le \sup_{i \in h^{-1}(j)} \frac{|x_i|}{v_j}. \tag{23}$$

By taking $\sup_j$ on both sides of (23), we see that

$$\|\Pi x\|_v = \sup_{j\in\mathcal{M}} \frac{\left|(\Pi x)_j\right|}{v_j} \leq \sup_{j\in\mathcal{M}} \sup_{i\in h^{-1}(j)} \frac{|x_i|}{v_j} = \sup_{i\in\mathcal{N}} \frac{|x_i|}{v_{h(i)}} = \|x\|_v, \tag{24}$$

where we use the definition of $\|\cdot\|_v$ on $\mathbb{R}^{\mathcal{N}}$ in the last equation. Hence we have shown that Π is non-expansive in $\|\cdot\|_v$ (inequality (22)).

We can also show that for any $x, y \in \mathbb{R}^{\mathcal{M}}$, we have

$$\|\Phi x - \Phi y\|_v = \|x - y\|_v. \tag{25}$$

Since Φ is a linear operator, we only need to show that for any $x \in \mathbb{R}^{\mathcal{M}}$, $\|\Phi x\|_v = \|x\|_v$.

Since $(\Phi x)_i = x_{h(i)}, \forall i \in \mathcal{N}$, by the definition of $\|\cdot\|_v$ on $\mathbb{R}^{\mathcal{N}}$, we see that

$$\|\Phi x\|_v = \sup_{i\in\mathcal{N}} \frac{|(\Phi x)_i|}{v_{h(i)}} = \sup_{i\in\mathcal{N}} \frac{\left|x_{h(i)}\right|}{v_{h(i)}} = \sup_{j\in\mathcal{M}} \frac{|x_j|}{v_j} = \|x\|_v.$$

Hence we have shown that Φ is non-expansive in $\|\cdot\|_v$ (equation (25)).

Therefore, for any $x, y \in \mathbb{R}^{\mathcal{M}}$, we have

$$\begin{aligned} \|\Pi F(\Phi x) - \Pi F(\Phi y)\|_v &\leq \|F(\Phi x) - F(\Phi y)\|_v && \text{(26a)} \\ &\leq \gamma \|\Phi x - \Phi y\|_v && \text{(26b)} \\ &= \gamma \|x - y\|_v, && \text{(26c)} \end{aligned}$$

where we use (22) in (26a); Assumption 3.2 in (26b); (25) in (26c). □

C.2 Proof of Theorem 3.1

The proof approach of Theorem 3.1 is similar to the proof of Theorem 4 in [37]. Specifically, we show an upper bound for $\|x(t) - x^*\|_v$ by induction on time step t. To do so, we divide the whole proof into three steps: In Step 1, we manipulate the update rule (3) so that it can be written in a recursive form of sequence $\|x(t) - x^*\|_v$ (see Lemma C.1); In Step 2, we bound the effect of noise terms in the recursive form we obtained in Step 1; In Step 3, we combine the first two steps to finish the induction.

For simplicity of notation, we use e_i to denote the indicator vector in $\mathbb{R}^n$, i.e. the i th entry is 1 and all other entries are 0. We also use ξ_i to denote the indicator vector in $\mathbb{R}^m$.

One of the main proof techniques used in [37] is to consider $D_t = \mathbb{E} e_{i_t} e_{i_t}^\top \mid \mathcal{F}_{t-\tau}$, which is the distribution of i_t condition on $\mathcal{F}_{t-\tau}$, in the coefficients of the recursive relationship of sequence $\|x(t) - x^*\|_v$. However, this approach does not work in the more general setting we consider because x^* may not be the stationary point of operator $(\Phi^\top D_t \Phi)^{-1} \phi^\top D_t F(\Phi \cdot)$. As a result, we cannot decompose $\|x(t) - x^*\|_v$ recursively if we use D_t in the coefficients. To overcome this difficulty, we use $D = diag(d_1, \cdots, d_n)$, which is the stationary distribution of i_t, in the coefficients of the recursive relationship (Lemma C.1).

Now we begin the technical part of our proof.

Step 1: Decomposition of Error. Let $D_t = \mathbb{E} e_{i_t} e_{i_t}^\top \mid \mathcal{F}_{t-\tau}$, where τ is a parameter that we will tune later. Then D_t is a $\mathcal{F}_{t-\tau}$-measurable n-by-n diagonal random matrix, with its i'th entry being $d_{t,i} = \mathbb{P}(i_t = i \mid \mathcal{F}_{t-\tau})$. Recall that $D = diag(d_1, \cdots, d_n)$, where d is the stationary distribution of the Markov Chain $\{i_t\}$.

Notice that for all $i \in \mathcal{N}$, we have $\xi_{h(i)} = \Phi^\top e_i$. We can rewrite the update rule as

$$\begin{aligned}
x(t+1) &= x(t) + \alpha_t[e_{i_t}^\top F(\Phi x(t)) - \xi_{h(i_t)}^\top x(t) + w(t)]\xi_{h(i_t)} \\
&= x(t) + \alpha_t[\xi_{h(i_t)} e_{i_t}^\top F(\Phi x(t)) - \xi_{h(i_t)}\xi_{h(i_t)}^\top x(t) + w(t)\xi_{h(i_t)}] \\
&= x(t) + \alpha_t \Phi^\top \big[e_{i_t} e_{i_t}^\top (F(\Phi x(t)) - \Phi x(t)) + w(t) e_{i_t}\big] \qquad (27a) \\
&= x(t) + \alpha_t \big[\Phi^\top D F(\Phi x(t)) - \Phi^\top D \Phi x(t)\big] \\
&\quad + \alpha_t \Phi^\top \big[(e_{i_t} e_{i_t}^\top - D)(F(\Phi x(t)) - \Phi x(t)) + w(t) e_{i_t}\big] \\
&= x(t) + \alpha_t \big[\Phi^\top D F(\Phi x(t)) - \Phi^\top D \Phi x(t)\big] \\
&\quad + \alpha_t \Phi^\top \big[(e_{i_t} e_{i_t}^\top - D)(F(\Phi x(t-\tau)) - \Phi x(t-\tau)) + w(t) e_{i_t}\big] \\
&\quad + \alpha_t \Phi^\top (e_{i_t} e_{i_t}^\top - D)[F(\Phi x(t)) - F(\Phi x(t-\tau)) - \Phi(x(t) - x(t-\tau))] \\
&= (I - \alpha_t \Phi^\top D \Phi) x(t) + \alpha_t \Phi^\top D F(\Phi x(t)) + \alpha_t(\epsilon(t) + \psi(t)), \qquad (27b)
\end{aligned}$$

where in (27a), we use $\xi_{h(i_t)} = \Phi^\top e_{i_t}$. Additionally, in (27b), we define

$$\epsilon(t) = \Phi^\top \big[(e_{i_t} e_{i_t}^\top - D)(F(\Phi x(t-\tau)) - \Phi x(t-\tau)) + w(t) e_{i_t}\big]$$

and

$$\psi(t) = \Phi^\top (e_{i_t} e_{i_t}^\top - D)[F(\Phi x(t)) - F(\Phi x(t-\tau)) - \Phi(x(t) - x(t-\tau))].$$

We further decompose $\epsilon(t)$ as $\epsilon(t) = \epsilon_1(t) + \epsilon_2(t)$, where $\epsilon_1(t)$ and $\epsilon_2(t)$ are defined as

$$\epsilon_1(t) = \Phi^\top \big[(e_{i_t} e_{i_t}^\top - D_t)(F(\Phi x(t-\tau)) - \Phi x(t-\tau)) + w(t) e_{i_t}\big]$$

and

$$\epsilon_2(t) = \Phi^\top (D_t - D)(F(\Phi x(t-\tau)) - \Phi x(t-\tau)).$$

We see that condition on $\mathcal{F}_{t-\tau}$, the expected value of $\epsilon_1(t)$ is zero, i.e.

$$\begin{aligned}
&\mathbb{E}\epsilon_1(t) \mid \mathcal{F}_{t-\tau} \\
&= \Phi^\top \mathbb{E}\big[(e_{i_t} e_{i_t}^\top - D_t) \mid \mathcal{F}_{t-\tau}\big][F(\Phi x(t-\tau)) - \Phi x(t-\tau)] + \Phi^\top \mathbb{E}[\mathbb{E}[w(t) \mid \mathcal{F}_t] e_{i_t} \mid \mathcal{F}_{t-\tau}] \\
&= 0.
\end{aligned}$$

Recall that matrix Π is defined as

$$\Pi = \big(\Phi^\top D \Phi\big)^{-1} \Phi^\top D.$$

By expanding (27) recursively, we obtain that

$$\begin{aligned}
x(t+1) = &\prod_{k=\tau}^{t} \big(I - \alpha_k \Phi^\top D \Phi\big) x(\tau) + \sum_{k=\tau}^{t} \alpha_k \left(\prod_{l=k+1}^{t} (I - \alpha_l \Phi^\top D \Phi) \right) \Phi^\top D F(\Phi x(k)) \\
&+ \sum_{k=\tau}^{t} \alpha_k \left(\prod_{l=k+1}^{t} (I - \alpha_l \Phi^\top D \Phi) \right) (\epsilon(k) + \psi(k)) \\
= &\tilde{B}_{\tau-1,t} x(\tau) + \sum_{k=\tau}^{t} B_{k,t} \Pi F(\Phi x(k)) + \sum_{k=\tau}^{t} \alpha_k \tilde{B}_{k,t} (\epsilon(k) + \psi(k)), \qquad (28)
\end{aligned}$$

where $B_{k,t} = \alpha_k \big(\Phi^\top D \Phi\big) \prod_{l=k+1}^{t} (I - \alpha_l \Phi^\top D \Phi)$ and $\tilde{B}_{k,t} = \prod_{l=k+1}^{t} \big(I - \alpha_l \Phi^\top D \Phi\big)$.

For simplicity of notation, we define $D' = \Phi^\top D \Phi \in \mathbb{R}^{\mathcal{M} \times \mathcal{M}}$. Notice that D' is a diagonal matrix in $\mathbb{R}^{\mathcal{M} \times \mathcal{M}}$ with the j'th entry $d'_j = \sum_{j \in h^{-1}(i)} d_i$. Clearly, $B_{k,t}$ and $\tilde{B}_{k,t}$ are m-by-m diagonal matrices, with the i'th diagonal entry given by $b_{k,t,i}$ and $\tilde{b}_{k,t,i}$, where $b_{k,t,i} = \alpha_k d'_i \prod_{l=k+1}^{t} (1 - \alpha_l d'_i)$ and $\tilde{b}_{k,t,i} = \prod_{l=k+1}^{t} (1 - \alpha_l d'_i)$. Therefore, for any $i \in \mathcal{M}$, we have

$$\tilde{b}_{\tau-1,t,i} + \sum_{k=\tau}^{t} b_{k,t,i} = 1. \qquad (29)$$

Also, by the definition of σ', we have that for any i, almost surely

$$b_{k,t,i} \le \beta_{k,t} := \alpha_k \prod_{l=k+1}^{t} (1-\alpha_l \sigma'), \tilde{b}_{k,t,i} \le \tilde{\beta}_{k,t} = \prod_{l=k+1}^{t} (1-\alpha_l \sigma'),$$

where $\sigma' = \min\{d'_1, \cdots, d'_m\}$.

Recall that x^* is the unique solution of the equation $\Pi F(\Phi x^*) = x^*$. Lemma C.1 shows that we can expand the error term $\|x(t) - x^*\|_v$ recursively.

Lemma C.1. *Let $\Upsilon_t = \|x(t) - x^*\|_v$, we have almost surely,*

$$\Upsilon_{t+1} \le \tilde{\beta}_{\tau-1,t} \Upsilon_\tau + \gamma \sup_{i \in \mathcal{M}} \sum_{k=\tau}^{t} b_{k,t,i} \Upsilon_k + \left\| \sum_{k=\tau}^{t} \alpha_k \tilde{B}_{k,t} \epsilon(k) \right\|_v + \left\| \sum_{k=\tau}^{t} \alpha_k \tilde{B}_{k,t} \psi(k) \right\|_v.$$

Proof of Lemma C.1. By (28) and the triangle inequality of $\|\cdot\|_v$, we have

$$\begin{aligned} &\|x(t+1) - x^*\|_v \\ \le& \sup_{i \in \mathcal{M}} \frac{1}{v_i} \left| \tilde{b}_{\tau-1,t,i} x_i(\tau) + \sum_{k=\tau}^{t} b_{k,t,i} (\Pi F(\Phi x(k)))_i - x_i^* \right| \\ &+ \left\| \sum_{k=\tau}^{t} \alpha_k \tilde{B}_{k,t} \epsilon(k) \right\|_v + \left\| \sum_{k=\tau}^{t} \alpha_k \tilde{B}_{k,t} \psi(k) \right\|_v. \end{aligned} \tag{30}$$

We also see that for each $i \in \mathcal{M}$,

$$\begin{aligned} &\frac{1}{v_i} \left| \tilde{b}_{\tau-1,t,i} x_i(\tau) + \sum_{k=\tau}^{t} b_{k,t,i} (\Pi F(\Phi x(k)))_i - x_i^* \right| \\ \le& \tilde{b}_{\tau-1,t,i} \frac{1}{v_i} |x_i(\tau) - x_i^*| + \sum_{k=\tau}^{t} b_{k,t,i} \frac{1}{v_i} |(\Pi F(\Phi x(k)))_i - x_i^*| \qquad (31a) \\ \le& \tilde{b}_{\tau-1,t,i} \|x(\tau) - x^*\|_v + \sum_{k=\tau}^{t} b_{k,t,i} \|(\Pi F(\Phi x(k))) - x^*\|_v \\ \le& \tilde{b}_{\tau-1,t,i} \|x(\tau) - x^*\|_v + \gamma \sum_{k=\tau}^{t} b_{k,t,i} \|x(k) - x^*\|_v, \qquad (31b) \end{aligned}$$

where in (31a), we use (29) which says $\tilde{b}_{\tau-1,t,i} + \sum_{k=\tau}^{t} b_{k,t,i} = 1$ holds for all $i \in \mathcal{M}$; in (31b), we use Proposition C.1, which says $\Pi F(\Phi \cdot)$ is γ-contraction in $\|\cdot\|_v$ with fixed point x^*.

Therefore, by substituting (31) into (30), we obtain that

$$\Upsilon_{t+1} \le \tilde{\beta}_{\tau-1,t} \Upsilon_\tau + \gamma \sup_{i \in \mathcal{M}} \sum_{k=\tau}^{t} b_{k,t,i} \Upsilon_k + \left\| \sum_{k=\tau}^{t} \alpha_k \tilde{B}_{k,t} \epsilon(k) \right\|_v + \left\| \sum_{k=\tau}^{t} \alpha_k \tilde{B}_{k,t} \psi(k) \right\|_v.$$

□

Step 2: Bounding $\left\| \sum_{k=\tau}^{t} \alpha_k \tilde{B}_{k,t} \epsilon(k) \right\|_v$ and $\left\| \sum_{k=\tau}^{t} \alpha_k \tilde{B}_{k,t} \psi(k) \right\|_v$.

We start with a bound on each individual $\epsilon_1(k), \epsilon_2(k)$, and $\psi(k)$ in Lemma C.2. For simplicity of notation, we define $\underline{v} := \inf_{j \in \mathcal{M}} v_j$.

Lemma C.2. *The following bounds hold almost surely.*

1. $\|\epsilon_1(t)\|_v \le 4\bar{x} + 2C + \frac{\bar{w}}{\underline{v}} := \bar{\epsilon}$.
2. $\|\epsilon_2(t)\|_v \le (2\bar{x} + C) \cdot 2K_1 \exp(-\tau/K_2)$.
3. $\|\psi(t)\|_v \le 3\left(2\bar{x} + C + \frac{\bar{w}}{\underline{v}}\right) \sum_{k=t-\tau+1}^{t} \alpha_{k-1}$.

Proof of Lemma C.2. By the definition of $\|\cdot\|_v$ in $\mathbb{R}^{\mathcal{M}}$ and its extension to $\mathbb{R}^{\mathcal{N}}$, the induced matrix norm of $\|\cdot\|$ for a matrix $A = [a_{ij}]_{i\in\mathcal{M},j\in\mathcal{N}}$ is given by $\|A\|_v = \sup_{i\in\mathcal{M}} \sum_{j\in\mathcal{N}} \frac{v_{h(j)}}{v_i}|a_{ij}|$. Recall that the i'th entry of the diagonal matrix D_t is given by $d_{t,i} = \mathbb{P}(i_t = i \mid \mathcal{F}_{t-\tau})$. Hence we have that

$$\left\|\Phi^\top(e_{i_t}e_{i_t}^\top - D_t)\right\|_v = \sup_{j\in\mathcal{M}} \sum_{i\in\mathcal{N}} 1(h(i)=j)\cdot|1(i=i_t) - d_{t,i}| \le 2. \tag{32}$$

Therefore, we can upper bound $\|\epsilon_1(t)\|_v$ by

$$\begin{aligned}
\|\epsilon_1(t)\|_v &= \left\|\Phi^\top\left[(e_{i_t}e_{i_t}^\top - D_t)(F(\Phi x(t-\tau)) - \Phi x(t-\tau)) + w(t)e_{i_t}\right]\right\|_v \\
&\le \left\|\Phi^\top(e_{i_t}e_{i_t}^\top - D_t)\right\|_v \|F(\Phi x(t-\tau)) - \Phi x(t-\tau)\|_v + |w(t)|\left\|\Phi^\top e_{i_t}\right\|_v \\
&\le 2\|F(\Phi x(t-\tau)) - \Phi x(t-\tau)\|_v + |w(t)|\left\|\Phi^\top e_{i_t}\right\|_v && \text{(33a)} \\
&\le 2\|F(\Phi x(t-\tau))\|_v + 2\|x(t-\tau)\|_v + \frac{\bar{w}}{\underline{v}} && \text{(33b)} \\
&\le 4\bar{x} + 2C + \frac{\bar{w}}{\underline{v}}, && \text{(33c)}
\end{aligned}$$

where we use (32) in (33a); the triangle inequality, the definition of $\bar{v}$, and Assumption 3.3 in (33b); Assumption 3.2 in (33c).

For $\|\epsilon_2(t)\|_v$, recall that

$$\begin{aligned}
\|\epsilon_2(t)\|_v &= \left\|\Phi^\top(D_t - D)(F(\Phi x(t-\tau)) - \Phi x(t-\tau))\right\|_v \\
&= \sup_{j\in\mathcal{M}} \frac{1}{v_j}\left|\sum_{i\in\mathcal{N}} 1(h(i)=j)(d_{t,i}-d_i)(F(\Phi x(t-\tau)) - \Phi x(t-\tau))_i\right| \\
&= \sup_{j\in\mathcal{M}} \frac{1}{v_j}\left|\sum_{i\in h^{-1}(j)} (d_{t,i}-d_i)(F(\Phi x(t-\tau)) - \Phi x(t-\tau))_i\right|. && \text{(34)}
\end{aligned}$$

By Assumption 3.1, we have that

$$\sup_{\mathcal{S}\subseteq\mathcal{N}}\left|\sum_{i\in\mathcal{S}} d_i - \sum_{i\in\mathcal{S}} d_{t,i}\right| \le K_1\exp(-\tau/K_2). \tag{35}$$

Our objective is to bound the following term in (34) for all $j\in\mathcal{M}$:

$$\left|\sum_{i\in h^{-1}(j)} (d_{t,i}-d_i)(F(\Phi x(t-\tau)) - \Phi x(t-\tau))_i\right|.$$

Let $M_j := \sup_{i\in h^{-1}(j)} |(F(\Phi x(t-\tau)) - \Phi x(t-\tau))_i|$. Define function $g : [-M_j, M_j]^{\mathcal{N}} \to \mathbb{R}$ as

$$g(y) = \left|\sum_{i\in h^{-1}(j)} (d_{t,i}-d_i)y_i\right|.$$

Suppose $y_{max} \in \arg\max_y g(y)$. We know that for $i\in h^{-1}(j)$, $(y_{max})_i$ is either M_j or $-M_j$ if $d_{t,i}-d_i \neq 0$. Let $S_j := \{i\in h^{-1}(j) \mid (y_{max})_i = M_j\}$ and $S_j' := \{i\in h^{-1}(j) \mid (y_{max})_i = -M_j\}$. Therefore, we see that

$$\begin{aligned}
&\left|\sum_{i\in h^{-1}(j)} (d_{t,i}-d_i)(F(\Phi x(t-\tau)) - \Phi x(t-\tau))_i\right| \\
&\le \max_{y\in[-M_j,M_j]^{\mathcal{N}}} g(y) && \text{(36a)} \\
&= \left|\sum_{i\in S_j}(d_{t,i}-d_i)\right|M_j + \left|\sum_{i\in S_j'}(d_{t,i}-d_i)\right|M_j \\
&\le 2K_1\exp(-\tau/K_2)M_j. && \text{(36b)}
\end{aligned}$$

where we use the definition of function g in (36a); we use (35) in (36b).

Substituting (36) into (34) gives that

$$\begin{aligned}\|\epsilon_2(t)\|_v &\le \|F(\Phi x(t-\tau)) - \Phi x(t-\tau)\|_v \cdot 2K_1 \exp(-\tau/K_2) \\ &\le (\|F(\Phi x(t-\tau))\|_v + \|\Phi x(t-\tau)\|_v) \cdot 2K_1 \exp(-\tau/K_2) \quad (37a) \\ &\le (2\bar{x} + C) \cdot 2K_1 \exp(-\tau/K_2), \quad (37b)\end{aligned}$$

where we use the triangle inequality in (37a); we use Assumption 3.2 in (37b).

As for $\|\psi(t)\|_v$, we have the following bound

$$\begin{aligned}&\|\psi(t)\|_v \\ &= \left\|\Phi^\top(e_{i_t}e_{i_t}^\top - D)(F(\Phi x(t)) - F(\Phi x(t-\tau))) - \Phi^\top(e_{i_t}e_{i_t}^\top - D)\Phi(x(t) - x(t-\tau))\right\|_v \\ &\le \left\|\Phi^\top(e_{i_t}e_{i_t}^\top - D)(F(\Phi x(t)) - F(\Phi x(t-\tau)))\right\|_v + \left\|\Phi^\top(e_{i_t}e_{i_t}^\top - D)\Phi(x(t) - x(t-\tau))\right\|_v \\ &\le \left\|\Phi^\top(e_{i_t}e_{i_t}^\top - D)\right\|_v \cdot \|(F(\Phi x(t)) - F(\Phi x(t-\tau)))\|_v \\ &\quad + \left\|\Phi^\top(e_{i_t}e_{i_t}^\top - D)\Phi\right\|_v \cdot \|(x(t) - x(t-\tau))\|_v. \quad (38)\end{aligned}$$

Notice that

$$\left\|\Phi^\top(e_{i_t}e_{i_t}^\top - D)\Phi\right\|_v = \left\|\xi_{h(i_t)}\xi_{h(i_t)}^\top - D'\right\|_v = \sup_{j\in\mathcal{M}} \left|1(h(i_t) = j) - d'_j\right| \le 1.$$

Substituting this into (38) and use (32), we obtain that

$$\begin{aligned}\|\psi(t)\|_v &\le 2\|F(\Phi x(t)) - F(\Phi x(t-\tau))\|_v + \|x(t) - x(t-\tau)\|_v \\ &\le 3\|x(t) - x(t-\tau)\|_v \\ &\le 3\sum_{k=t-\tau+1}^{t} \|x(k) - x(k-1)\|_v. \quad (39)\end{aligned}$$

By the update rule of x and Assumption 3.2, we have that

$$\begin{aligned}\|x(t) - x(t-1)\|_v &\le \alpha_{t-1}\left(\|F(\Phi x(t-1))\|_v + \|x(t-1)\|_v + \frac{\bar{w}}{\underline{v}}\right) \\ &\le \alpha_{t-1}\left(2\bar{x} + C + \frac{\bar{w}}{\underline{v}}\right). \quad (40)\end{aligned}$$

Substituting (40) into (39), we obtain that

$$\|\psi(t)\|_v \le 3\left(2\bar{x} + C + \frac{\bar{w}}{\underline{v}}\right)\sum_{k=t-\tau+1}^{t} \alpha_{k-1}.$$

□

Lemma C.3. *If $\alpha_t = \frac{H}{t+t_0}$, where $H > \frac{2}{\sigma'}$ and $t_0 \ge \max(4H, \tau)$, then $\beta_{k,t}, \tilde{\beta}_{k,t}$ satisfies the following*

1. $\beta_{k,t} \le \frac{H}{k+t_0}\left(\frac{k+1+t_0}{t+1+t_0}\right)^{\sigma' H}, \tilde{\beta}_{k,t} \le \left(\frac{k+1+t_0}{t+1+t_0}\right)^{\sigma' H}$.
2. $\sum_{k=1}^{t} \beta_{k,t}^2 \le \frac{2H}{\sigma'}\frac{1}{t+1+t_0}$.
3. $\sum_{k=\tau}^{t} \beta_{k,t}\sum_{l=k-\tau+1}^{k} \alpha_{l-1} \le \frac{8H\tau}{\sigma'}\frac{1}{t+1+t_0}$.

Proof of Lemma C.3. To show Lemma C.3, we only need to substitute σ' for σ in the proof of [37][Lemma 10]. □

Lemma C.4. *The following inequality holds almost surely*

$$\left\|\sum_{k=\tau}^{t} \alpha_k \tilde{B}_{k,t}\psi(k)\right\|_v \le \frac{24\left(2\bar{x} + C + \frac{\bar{w}}{\underline{v}}\right)H\tau}{\sigma'}\frac{1}{t+1+t_0} := C_\psi\frac{1}{t+1+t_0}.$$

Proof of Lemma C.4. We have that

$$\begin{aligned}\left\|\sum_{k=\tau}^{t}\alpha_k\tilde{B}_{k,t}\psi(k)\right\|_v &\le \sum_{k=\tau}^{t}\alpha_k\left\|\tilde{B}_{k,t}\right\|_v\|\psi(k)\|_v \\ &\le 3\left(2\bar{x}+C+\frac{\bar{w}}{\underline{v}}\right)\sum_{k=\tau}^{t}\beta_{k,t}\sum_{l=k-\tau+1}^{k}\alpha_{l-1} \qquad (41a)\\ &\le \frac{24\left(2\bar{x}+C+\frac{\bar{w}}{\underline{v}}\right)H\tau}{\sigma'}\frac{1}{t+1+t_0}, \qquad (41b)\end{aligned}$$

where we use Lemma C.2 in (41a); Lemma C.3 in (41b). □

Lemma C.5. *For each t, with probability at least $1-\delta$, we have*

$$\left\|\sum_{k=\tau}^{t}\alpha_k\tilde{B}_{k,t}\epsilon_1(k)\right\|_v \le \frac{H\bar{\epsilon}}{t+t_0}\sqrt{2\tau t\log\left(\frac{2\tau m}{\delta}\right)}.$$

To show Lemma C.5, we need to use Lemma C.6, which is Lemma 13 in [37].

Lemma C.6. *Let X_t be a $\mathcal{F}_t$-adapted stochastic process which satisfies $\mathbb{E}X_t \mid \mathcal{F}_{t-\tau} = 0$. Further, $|X_t| \le \bar{X}_t$ almost surely. Then with probability $1-\delta$, we have, $\left|\sum_{k=0}^{t}X_t\right| \le \sqrt{2\tau\sum_{k=0}^{t}\bar{X}_k^2\log\left(\frac{2\tau}{\delta}\right)}$.*

Proof of Lemma C.5. Recall that $\sum_{k=\tau}\alpha_k\tilde{B}_{k,t}\epsilon_1(k)$ is a random vector in $\mathbb{R}^{\mathcal{M}}$, with its i'th entry

$$\sum_{k=\tau}^{t}\alpha_k(\epsilon_1)_i(k)\prod_{l=k+1}^{t}(1-\alpha_l d_i').$$

Since step sizes $\{\alpha_l\}$ are deterministic, we see that

$$\mathbb{E}\left[\alpha_k(\epsilon_1)_i(k)\prod_{l=k+1}^{t}(1-\alpha_l d_i') \mid \mathcal{F}_{k-\tau}\right] = \alpha_k\prod_{l=k+1}^{t}(1-\alpha_l d_i')\mathbb{E}[(\epsilon_1)_i(k) \mid \mathcal{F}_{k-\tau}] = 0.$$

Notice that

$$\begin{aligned}\alpha_k\prod_{l=k+1}^{t}(1-\alpha_l d_i') &= \frac{H}{k+t_0}\prod_{l=k+1}^{t}\left(1-\frac{Hd_i'}{l+t_0}\right) \qquad (42a)\\ &\le \frac{H}{k+t_0}\prod_{l=k+1}^{t}\left(1-\frac{2}{l+t_0}\right) \qquad (42b)\\ &\le \frac{H}{k+t_0}\prod_{l=k+1}^{t}\left(1-\frac{1}{l+t_0}\right)\\ &\le \frac{H}{t+t_0},\end{aligned}$$

where we use $\alpha_l = \frac{H}{l+t_0}$ in (42a); we use $H > \frac{2}{\sigma'}$ in (42b).

By the definition of $\bar{\epsilon}$, we also see that $|(\epsilon_1)_i(k)| \le v_i\bar{\epsilon}$. Therefore, by Lemma C.6, we obtain that

$$\left|\sum_{k=\tau}^{t}\alpha_k(\epsilon_1)_i(k)\prod_{l=k+1}^{t}(1-\alpha_l d_i')\right| \le \frac{Hv_i\bar{\epsilon}}{t+t_0}\sqrt{2\tau t\log\left(\frac{2\tau}{\delta}\right)}$$

holds with probability at least $1-\delta$. By union bound, we see that with probability at least $1-\delta$,

$$\left\|\sum_{k=\tau}^{t}\alpha_k\tilde{B}_{k,t}\epsilon_1(k)\right\|_v \le \frac{H\bar{\epsilon}}{t+t_0}\sqrt{2\tau t\log\left(\frac{2\tau m}{\delta}\right)}.$$

□

Lemma C.7. *If we set τ to be an integer such that*

$$\tau \geq 2K_2 \max(\log t, 1),$$

we have that

$$\left\| \sum_{k=\tau}^{t} \alpha_k \tilde{B}_{k,t} \epsilon_2(k) \right\|_v \leq \frac{C_{\epsilon_2}}{t + t_0 + 1},$$

where $t_0 = \max(\tau, 4H)$ and $C_{\epsilon_2} = (2\bar{x} + C) \cdot 2K_1(1 + 2K_2 + 4H)$.

Proof of Lemma C.7. Since $K_2 \geq 1$, the bound is trivial when $t = 1$. We consider the case when $t \geq 2$ below.

Since $\alpha_k \tilde{B}_{k,t}$ is a diagonal matrix and its entries are positive and less than 1, we have that

$$\begin{aligned}
\left\| \sum_{k=\tau}^{t} \alpha_k \tilde{B}_{k,t} \epsilon_2(k) \right\|_v &\leq \sum_{k=\tau}^{t} \left\| \alpha_k \tilde{B}_{k,t} \right\|_v \cdot \|\epsilon_2(k)\|_v \\
&\leq t \|\epsilon_2(k)\|_v \qquad (43a) \\
&\leq t(2\bar{x} + C) \cdot 2K_1 \exp(-\tau/K_2). \qquad (43b)
\end{aligned}$$

where we use $\left\| \alpha_k \tilde{B}_{k,t} \right\|_v \leq 1$ in (43a); Lemma C.2 in (43b).

To show Lemma C.7, we only need to show

$$t(2\bar{x} + C) \cdot 2K_1(t + \tau + 4H) \exp(-\tau/K_2) \leq C_{\epsilon_2} \tag{44}$$

holds for all $\tau \geq 2K_2 \log t$ because $t + t_0 + 1 \leq t + \tau + 4H$.

To study how the left hand side of (44) changes with τ, we define function

$$g(\tau) = (\tau + t + 4H) \exp(-\tau/K_2).$$

Notice that we view τ as real number in function g, so we can get the derivative of g:

$$g'(\tau) = \frac{\exp(-\tau/K_2)}{K_2}(K_2 - t - 4H - \tau).$$

Therefore, when $\tau \geq 2K_2 \log t$, we always have $g'(\tau) < 0$. Hence we obtain that

$$g(\tau) \leq g(2K_2 \log t) = \frac{2K_2 \log t + t + 4H}{t^2} \leq \frac{1 + 2K_2 + 4H}{t} \tag{45}$$

holds for all $\tau \geq 2K_2 \log t$.

Substituting (45) into (44) finishes the proof. □

Step 3: Bounding the error sequence. Based on the recursive relationship we derived in Lemma C.1 and the bounds we obtained in Step 2, we want to show that, with probability $1 - \delta$,

$$\Upsilon_t \leq \frac{C_a}{\sqrt{t + t_0}} + \frac{C'_a}{t + t_0}, \tag{46}$$

holds for all $\tau \leq t \leq T$, where

$$C_a = \frac{2H\bar{\epsilon}}{1 - \gamma} \sqrt{2\tau \log\left(\frac{2\tau m T}{\delta}\right)}, C'_a = \frac{4}{1 - \gamma} \max(C_\psi + C_{\epsilon_2}, 2\bar{x}(\tau + t_0)).$$

Notice that C_a and C'_a are independent of t but may dependent on T. We set $\tau = 2K_2 \log T$.

By applying union bound to Lemma C.5, we see that with probability at least $1 - \delta$, for any $t \leq T$,

$$\left\| \sum_{k=\tau}^{t} \alpha_k \tilde{B}_{k,t} \epsilon_1(k) \right\|_v \leq \frac{C_{\epsilon_1}}{\sqrt{t + 1 + t_0}},$$

where $C_{\epsilon_1} = H\bar{\epsilon}\sqrt{2\tau \log\left(\frac{2\tau m T}{\delta}\right)}$.

Therefore, we get with probability $1-\delta$, (47) holds for all $\tau \le t \le T$:

$$\Upsilon_{t+1} \le \tilde{\beta}_{\tau-1,t}\Upsilon_\tau + \gamma \sup_{i\in\mathcal{M}} \sum_{k=\tau}^{t} b_{k,t,i}\Upsilon_k + \frac{C_{\epsilon_1}}{\sqrt{t+1+t_0}} + \frac{C_\psi + C_{\epsilon_2}}{t+1+t_0}. \tag{47}$$

We now condition on (47) to show (46) by induction. (46) is true for $t=\tau$, as $\frac{C'_a}{\tau+t_0} \ge \frac{8}{1-\gamma}\bar{x} \ge \Upsilon_\tau$, where we have used $\Upsilon_\tau = \|x(\tau) - x^*\|_v \le \|x(\tau)\|_v + \|x^*\|_v \le 2\bar{x}$. Then, assuming (46) is true for up to $k \le t$. By (47), we have that

$$\begin{aligned}
\Upsilon_{t+1} &\le \tilde{\beta}_{\tau-1,t}\Upsilon_\tau + \gamma \sup_{i\in\mathcal{M}} \sum_{k=\tau}^{t} b_{k,t,i}\left[\frac{C_a}{\sqrt{k+t_0}} + \frac{C'_a}{k+t_0}\right] + \frac{C_{\epsilon_1}}{\sqrt{t+1+t_0}} + \frac{C_\psi + C_{\epsilon_2}}{t+1+t_0} \\
&\le \tilde{\beta}_{\tau-1,t}\Upsilon_\tau + \gamma C_a \sup_{i\in\mathcal{M}} \sum_{k=\tau}^{t} b_{k,t,i}\frac{1}{\sqrt{k+t_0}} + \gamma C'_a \sup_{i\in\mathcal{M}} \sum_{k=\tau}^{t} \frac{1}{k+t_0} b_{k,t,i} \\
&\quad + \frac{C_{\epsilon_1}}{\sqrt{t+1+t_0}} + \frac{C_\psi + C_{\epsilon_2}}{t+1+t_0}.
\end{aligned} \tag{48}$$

We use the following auxiliary lemma to handle the second and the third term in (48).

Lemma C.8. *If $\sigma' H(1-\sqrt{\gamma}) \ge 1, t_0 \ge 1$, and $\alpha_0 \le \frac{1}{2}$, then, for any $i \in \mathcal{N}$, and any $0 < \omega \le 1$, we have*

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

Proof of Lemma C.8. Recall that $\alpha_k = \frac{H}{k+t_0}$, and $b_{k,t,i} = \alpha_k d'_i \prod_{l=k+1}^{t}(1-\alpha_l d'_i)$, where $d'_i \ge \sigma'$. Define $e_t = \sum_{k=\tau}^{t} b_{k,t,i}\frac{1}{(k+t_0)^\omega}$. We use induction on t to show that $e_t \le \frac{1}{\sqrt{\gamma}(t+1+t_0)^\omega}$.

The statement is clearly true for $t=\tau$. Assume it is true for $t-1$. Notice that

$$\begin{aligned}
e_t &= \sum_{k=\tau}^{t-1} b_{k,t,i}\frac{1}{(k+t_0)^\omega} + b_{t,t,i}\frac{1}{(t+t_0)^\omega} \\
&= (1-\alpha_t d'_i)\sum_{k=\tau}^{t-1} b_{k,t-1,i}\frac{1}{(k+t_0)^\omega} + \alpha_t d'_i \frac{1}{(t+t_0)^\omega} && \text{(49a)} \\
&= (1-\alpha_t d'_i)e_{t-1} + \alpha_t d'_i \frac{1}{(t+t_0)^\omega} \\
&\le (1-\alpha_t d'_i)\frac{1}{\sqrt{\gamma}(t+t_0)^\omega} + \alpha_t d'_i \frac{1}{(t+t_0)^\omega} && \text{(49b)} \\
&= [1-\alpha_t d'_i(1-\sqrt{\gamma})]\frac{1}{\sqrt{\gamma}(t+t_0)^\omega},
\end{aligned}$$

where we use $b_{t,t,i} = \alpha_t d'_i$ in (49a); we use the induction assumption in (49b).

Plugging in $\alpha_t = \frac{H}{t+t_0}$, we see that

$$
\begin{aligned}
e_t &\le \left[1 - \frac{\sigma' H}{t+t_0}(1-\sqrt{\gamma})\right] \frac{1}{\sqrt{\gamma}(t+t_0)^\omega} && \text{(50a)} \\
&= \left[1 - \frac{\sigma' H}{t+t_0}(1-\sqrt{\gamma})\right] \left(1 + \frac{1}{t+t_0}\right)^\omega \frac{1}{\sqrt{\gamma}(t+1+t_0)^\omega} \\
&\le \left(1 - \frac{1}{t+t_0}\right)\left(1 + \frac{1}{t+t_0}\right)^\omega \frac{1}{\sqrt{\gamma}(t+1+t_0)^\omega} && \text{(50b)} \\
&\le \left(1 - \frac{1}{t+t_0}\right)\left(1 + \frac{1}{t+t_0}\right) \frac{1}{\sqrt{\gamma}(t+1+t_0)^\omega} && \text{(50c)} \\
&\le \frac{1}{\sqrt{\gamma}(t+1+t_0)^\omega},
\end{aligned}
$$

where we use $d_i' \ge \sigma'$ in (50a); we use the assumption that $\sigma' H(1-\sqrt{\gamma}) \ge 1$ in (50b); we use $0 < \omega \le 1$ in (50c). □

Applying Lemma C.8 to (48), we see that

$$
\begin{aligned}
\Upsilon_{t+1} &\le \tilde{\beta}_{\tau-1,t}\Upsilon_\tau + \sqrt{\gamma}C_a \frac{1}{\sqrt{t+1+t_0}} + \sqrt{\gamma}C_a' \frac{1}{t+1+t_0} \\
&\quad + C_{\epsilon_1}\frac{1}{\sqrt{t+1+t_0}} + (C_\psi + C_{\epsilon_2})\frac{1}{t+1+t_0} && \text{(51a)} \\
&\le \left(\sqrt{\gamma}C_a \frac{1}{\sqrt{t+1+t_0}} + C_{\epsilon_1}\frac{1}{\sqrt{t+1+t_0}}\right) \\
&\quad + \left(\sqrt{\gamma}C_a' \frac{1}{t+1+t_0} + (C_\psi + C_{\epsilon_2})\frac{1}{t+1+t_0} + \left(\frac{\tau+t_0}{t+1+t_0}\right)^{\sigma' H}\Upsilon_\tau\right), && \text{(51b)}
\end{aligned}
$$

where we use Lemma C.8 in (51a); we use the bound on $\tilde{\beta}_{\tau-1,t}$ in Lemma C.3 in (51b).

To bound the two terms in (51b), we define

$$
\chi_t := \sqrt{\gamma}C_a \frac{1}{\sqrt{t+1+t_0}} + C_{\epsilon_1}\frac{1}{\sqrt{t+1+t_0}}
$$

and

$$
\chi_t' = \sqrt{\gamma}C_a' \frac{1}{t+1+t_0} + (C_\psi + C_{\epsilon_2})\frac{1}{t+1+t_0} + \left(\frac{\tau+t_0}{t+1+t_0}\right)^{\sigma' H} a_\tau.
$$

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

$$
\chi_t \frac{\sqrt{t+1+t_0}}{C_a} = \sqrt{\gamma} + \frac{C_{\epsilon_1}}{C_a}, \chi_t' \frac{t+1+t_0}{C_a'} = \sqrt{\gamma} + \frac{C_\psi + C_{\epsilon_2}}{C_a'} + \frac{\Upsilon_\tau(\tau+t_0)}{C_a'}\left(\frac{\tau+t_0}{t+1+t_0}\right)^{\sigma' H - 1}.
$$

It suffices to show that $\frac{C_{\epsilon_1}}{C_a} \le 1 - \sqrt{\gamma}$, $\frac{C_\psi + C_{\epsilon_2}}{C_a'} \le \frac{1-\sqrt{\gamma}}{2}$, and $\frac{\Upsilon_\tau(\tau+t_0)}{C_a'} \le \frac{1-\sqrt{\gamma}}{2}$. Recall that

$$
C_a = \frac{2H\bar{\epsilon}}{1-\gamma}\sqrt{2\tau \log\left(\frac{2\tau m T}{\delta}\right)}, C_a' = \frac{4}{1-\gamma}\max(C_\psi + C_{\epsilon_2}, 2\bar{x}(\tau+t_0)),
$$

and

$$
C_{\epsilon_1} = H\bar{\epsilon}\sqrt{2\tau \log\left(\frac{2\tau m T}{\delta}\right)}.
$$

Using that $\Upsilon_\tau \le 2\bar{x}$, one can check that C_a and C_a' satisfy the above three inequalities.

C.3 Parameter Upper Bound

Proposition C.2. *Suppose Assumptions 3.2 and 3.3 hold. Then for all t,*

$$\|x(t)\|_v \leq \frac{1}{1-\gamma}\left((1+\gamma)\|y^*\|_v + \frac{\bar{w}}{\underline{v}}\right)$$

holds almost surely, where $y^* \in \mathbb{R}^{\mathcal{N}}$ is the stationary point of F.

Proof of Proposition C.2. By Assumption 3.2, we have that for all $x \in \mathbb{R}^{\mathcal{M}}$,

$$\begin{aligned}
\|F(\Phi x)\|_v &\leq \|F(\Phi x) - F(y^*)\|_v + \|F(y^*)\|_v && \text{(52a)}\\
&\leq \gamma\|\Phi x - y^*\|_v + \|y^*\|_v && \text{(52b)}\\
&\leq \gamma\|x\|_v + (1+\gamma)\|y^*\|_v, && \text{(52c)}
\end{aligned}$$

where we use the triangle inequality in (52a) and (52c); we use Assumption 3.2 in (52b).

Let $\bar{x} = \frac{1}{1-\gamma}\left((1+\gamma)\|y^*\|_v + \frac{\bar{w}}{\underline{v}}\right)$. We prove $\|x(t)\|_v \leq \bar{x}$ by induction on t. Since we initialize $x(0)$ to be $\mathbf{0}$, the statement is true for $t = 0$.

Suppose the statement is true for t. By the update rule of x, we see that

$$\begin{aligned}
\frac{1}{v_{h(i_t)}}\left|x_{h(i_t)}(t+1)\right| &\leq (1-\alpha_t)\frac{1}{v_{h(i_t)}}\left|x_{h(i_t)}(t)\right| + \alpha_t\left(\frac{1}{v_{h(i_t)}}|F_{i_t}(\Phi x(t))| + \frac{1}{v_{h(i_t)}}|w(t)|\right)\\
&\leq (1-\alpha_t)\|x(t)\|_v + \alpha_t\left(\|F(\Phi x(t))\|_v + \frac{\bar{w}}{\underline{v}}\right) && \text{(53a)}\\
&\leq (1-\alpha_t)\|x(t)\|_v + \alpha_t\left(\gamma\|x(t)\|_v + (1+\gamma)\|y^*\|_v + \frac{\bar{w}}{\underline{v}}\right) && \text{(53b)}\\
&\leq (1-\alpha_t)\bar{x} + \alpha_t\left(\gamma\bar{x} + (1+\gamma)\|y^*\|_v + \frac{\bar{w}}{\underline{v}}\right) && \text{(53c)}\\
&= \bar{x},
\end{aligned}$$

where we use Assumption 3.3 in (53a); (52) in (53b); the induction assumption in (53c).

For $j \neq h(i_t), j \in \mathcal{M}$, we have that

$$\frac{1}{v_j}|x_j(t+1)| = \frac{1}{v_j}|x_j(t)| \leq \|x(t)\|_v \leq \bar{x}. \tag{54}$$

Combining (53) and (54), we see that the statement also holds for $t+1$. Hence we have showed $\|x(t)\|_v \leq \bar{x}$ by induction. □

D TD/Q-Learning with State Aggregation

D.1 Asymptotic Convergence of TD Learning with State Aggregation

Our asymptotic convergence result for TD learning with state aggregation builds upon the asymptotic convergence result for TD learning with linear function approximation shown in [49]. For completeness, we first present the main result of [49] in Theorem D.1. In order to do this, we must first state a few definitions and assumptions made in [49].

We use $\phi(i) \in \mathbb{R}^m$ to denote the feature vector associated with state $i \in \mathcal{N}$. Feature matrix Φ is a n-by-m matrix whose i'th row is $\phi(i)^\top$. Starting from $\theta(0) = \mathbf{0}$, the $TD(\lambda)$ algorithm keeps updating θ, ψ by the following update rule,

$$\begin{aligned}
\theta(t+1) &= \theta(t) + \alpha_t d_t \psi_t,\\
\psi_{t+1} &= \gamma\lambda\psi_t + \phi(i_{t+1}),
\end{aligned}$$

where ψ_t is named *eligible vector* in [49] and satisfies $\psi_0 = \phi(i_0)$.

Recall that $D = diag(d_1, d_2, \cdots, d_n)$ denotes the stationary distribution of Markov chain $\{i_t\}$. For vectors $x, y \in \mathbb{R}^n$, we define inner product $\langle x, y\rangle = x^\top Dy$. The induced norm of this inner product is $\|\cdot\|_D = \sqrt{\langle \cdot, \cdot\rangle_D}$. Let $L_2(\mathcal{N}, D)$ denote the set of vectors $V \in \mathbb{R}^n$ such that $\|V\|_D$ is finite.

Recall that we define $\Pi = (\Phi^\top D\Phi)^{-1}\Phi^\top D$. As shown in [49], the projection matrix that projects an arbitrary vector in $\mathbb{R}^n$ to the set $\{\Phi\theta \mid \theta \in \mathbb{R}^m\}$ is given by $\Phi\Pi$, i.e. for any $V \in L_2(\mathcal{N}, D)$, we have

$$\Phi\Pi V = \underset{\bar{V}\in\{\Phi\theta|\theta\in\mathbb{R}^m\}}{\arg\min} \left\|V - \bar{V}\right\|_D.$$

Notice that our definition of matrix Π is slightly different with [49] because we want to be consistent with Section 3.1.

To characterize the TD(λ) algorithm's dynamics, [49] defines $T^{(\lambda)} : L_2(\mathcal{N}, D) \to L_2(\mathcal{N}, D)$ operator as following: for all $V \in \mathbb{R}^n$, let the i'th dimension of $\left(T^{(\lambda)}V\right)$ be defined as

$$\left(T^{(\lambda)}V\right)_i = \begin{cases} (1-\lambda)\sum_{m=0}^{\infty}\lambda^m\mathbb{E}\left[\sum_{t=0}^{m}\gamma^t r(i_t, i_{t+1}) + \gamma^{m+1}V_{i_{m+1}} \mid i_0 = i\right] & \text{if } \lambda < 1 \\ \mathbb{E}[\sum_{t=0}^{\infty}\gamma^t r(i_t, i_{t+1}) \mid i_0 = i] & \text{if } \lambda = 1. \end{cases}$$

If V is an approximation of the value function V^*, $T^{(\lambda)}$ can be viewed as an improved approximation to V^*. Notice that when $\lambda = 0$, $T^{(\lambda)}$ is identical with the Bellman operator.

Formally, [49] made four necessary assumptions for their main result (Theorem D.1). We omit the third assumption ([49][Assumption 3]) in our summary because it must hold when the state space $\mathcal{N}$ is finite.

The first assumption ([49][Assumption 1]) concerns the stationary distribution and the reward function of the Markov chain $\{i_t\}$. It must hold when Assumption 3.1 holds and every stage reward r_t is upper bounded by $\bar{r}$, as assumed by Theorem 3.2.

Assumption D.1. *The transition probability and cost function satisfies the following two conditions:*

1. *The Markov chain $\{i_t\}$ is irreducible and aperiodic. Furthermore, there is a unique distribution d that satisfies $d^\top P = d^\top$ with $d_i > 0$ for all $i \in \mathcal{N}$. Let $\mathbb{E}_0$ stand for expectation with respect to this distribution.*
2. *The reward function $r(i_t, i_{t+1})$ satisfies $\mathbb{E}_0\left[r^2(i_t, i_{t+1})\right] < \infty$.*

The second assumption ([49][Assumption 2]) concerns the feature vectors and the feature matrix. It must hold when Φ is defined as (4).

Assumption D.2. *The following two conditions hold for Φ:*

1. *The matrix Φ has full column rank; that is, the m columns (named basis functions in [49]) $\{\phi_k \mid k = 1, \cdots, m\}$ are linearly independent.*
2. *For every k, the basis function ϕ_k satisfies $\mathbb{E}_0\left[\phi_k^2(i_t)\right] < \infty$.*

The third assumption ([49][Assumption 4]) concerns the learning step size. It must hold if the learning step sizes are as defined in Theorem 3.2.

Assumption D.3. *The step sizes α_t are positive, nonincreasing, and chosen prior to execution of the algorithm. Furthermore, they satisfy $\sum_{t=0}^{\infty}\alpha_t = \infty$ and $\sum_{t=0}^{\infty}\alpha_t^2 < \infty$.*

Now we are ready to present the main asymptotic convergence result given in [49].

Theorem D.1. *Under Assumptions D.1, D.2, D.3, the following hold.*

1. *The value function V is in $L_2(\mathcal{N}, D)$.*
2. *For any $\lambda \in [0, 1]$, the TD(λ) algorithm with linear function approximation converges with probability one.*
3. *The limit of convergence θ^* is the unique solution of the equation*

$$\Pi T^{(\lambda)}(\Phi\theta^*) = \theta^*.$$

4. Furthermore, θ^* satisfies

$$\|\Phi\theta^* - V^*\|_D \leq \frac{1-\lambda\gamma}{1-\gamma}\|\Phi\Pi V^* - V^*\|_D. \tag{55}$$

Notice that (55) is not exactly the result we want to obtain. Specifically, we want the both sides of (55) to be in $\|\cdot\|_\infty$ instead of $\|\cdot\|_D$. Although this kind of result is not obtainable for general TD learning with linear function approximation, we can leverage the special assumptions for state aggregation, which are summarized below:

Assumption D.4. *$h : \mathcal{N} \to \mathcal{M}$ is a surjective function from set $\mathcal{N}$ to $\mathcal{M}$. The feature matrix Φ is as defined in* (4)*, i.e. the feature vector associated with state $i \in \mathcal{N}$ is given by*

$$\phi_k(i) = \begin{cases} 1 & \text{if } k = h(i) \\ 0 & \text{otherwise} \end{cases}, \forall k \in \mathcal{M}.$$

Further, if $h(i) = h(i')$ for $i, i' \in \mathcal{N}$, we have $|V^*(i) - V^*(i')| \leq \zeta$ for a fixed positive constant ζ.

Under Assumption D.4, we can show the asymptotic error bound in the infinity norm as we desired:

Theorem D.2. *Under Assumptions D.1, D.2, D.3, if Assumption D.4 also holds, the limit of convergence θ^* of the $TD(\lambda)$ algorithm satisfies*

$$\|\Phi\theta^* - V^*\|_\infty \leq \frac{(1-\lambda\gamma)}{1-\gamma}\|\Phi\Pi V^* - V^*\|_\infty \leq \frac{(1-\lambda\gamma)}{1-\gamma}\zeta.$$

To show Theorem D.2, we need to prove several auxiliary lemmas first.

Lemma D.3. *Under Assumption D.1, for any $V \in L_2(\mathcal{N}, D)$, we have $\|PV\|_\infty \leq \|V\|_\infty$.*

Proof of Lemma D.3. This lemma holds because the transition matrix P is non-expansive in infinity norm. □

Lemma D.4. *Under Assumption D.1, for any $V, \bar{V} \in L_2(\mathcal{N}, D)$, we have*

$$\left\|T^{(\lambda)}V - T^{(\lambda)}\bar{V}\right\|_\infty \leq \frac{\gamma(1-\lambda)}{1-\gamma\lambda}\left\|V - \bar{V}\right\|_\infty.$$

Proof of Lemma D.4. By the definition of $T^{(\lambda)}$, we have that

$$\begin{aligned}
\left\|T^{(\lambda)}V - T^{(\lambda)}\bar{V}\right\|_\infty &= \left\|(1-\lambda)\sum_{m=0}^{\infty}\lambda^m(\gamma P)^{m+1}\big(V - \bar{V}\big)\right\|_\infty \\
&\leq (1-\lambda)\sum_{m=0}^{\infty}\lambda^m\gamma^{m+1}\left\|V - \bar{V}\right\|_\infty \qquad (56a) \\
&\frac{\gamma(1-\lambda)}{1-\gamma\lambda}\left\|V - \bar{V}\right\|_\infty,
\end{aligned}$$

where inequality (56a) holds because $\left\|V - \bar{V}\right\|_\infty < \infty$ so we use Lemma D.3. □

Lemma D.5. *Under Assumption D.1 and D.4, we have*

$$\|\Phi\Pi V^* - V^*\|_\infty \leq \zeta \tag{57}$$

and for any $V \in L_2(\mathcal{N}, D)$

$$\|\Phi\Pi V\|_\infty \leq \|V\|_\infty. \tag{58}$$

Proof of Lemma D.5. For $j \in \mathcal{M}$, we use $h^{-1}(j) \subseteq \mathcal{N}$ to denote all the elements in $\mathcal{N}$ whose feature is e_j, i.e. $h^{-1}(j) = \{i \mid i \in \mathcal{N}, h(i) = j\}$. Since h is surjection, $h^{-1}(j) \neq \emptyset, \forall j \in \mathcal{M}$. Since $\Phi\Pi$ is the projection matrix that projects a vector in $\mathbb{R}^n$ to the set $\{\Phi\theta \mid \theta \in \mathbb{R}^m\}$, we have

$$\Pi V = \arg\min_{\theta\in\mathbb{R}^m}\sum_{j\in\mathcal{M}}\sum_{i\in h^{-1}(j)} d_i(V_i - \theta_j).$$

Hence the optimal θ_j must be in the range $\left[\min_{i\in h^{-1}(j)} V_i, \max_{i\in h^{-1}(j)} V_i\right]$. Therefore, we see that

$$|(\Phi\Pi V)_i| = \left|(\Pi V)_{h(i)}\right| \leq \max_{i'\in h^{-1}(h(i))} |V_{i'}|,$$

which shows (58). Besides, we also have

$$|(\Phi\Pi V)_i - V_i| \leq \max\left(\left|\min_{i'\in h^{-1}(h(i))} V_{i'} - V_i\right|, \left|\max_{i'\in h^{-1}(h(i))} V_{i'} - V_i\right|\right). \tag{59}$$

holds for all $z \in \mathcal{Z}$. Let $V = V^*$ and use Assumption D.4 in (59) gives (57). □

Now we come back to the proof of Theorem D.2.

Notice that

$$\begin{aligned}
\|\Phi\theta^* - V^*\|_\infty &\leq \|\Phi\theta^* - \Phi\Pi V^*\|_\infty + \|\Phi\Pi V^* - V^*\|_\infty && \text{(60a)}\\
&= \left\|\Phi\Pi T^{(\lambda)}(\Phi\theta^*) - \Phi\Pi V^*\right\|_\infty + \|\Phi\Pi V^* - V^*\|_\infty && \text{(60b)}\\
&\leq \left\|T^{(\lambda)}(\Phi\theta^*) - V^*\right\|_\infty + \|\Phi\Pi V^* - V^*\|_\infty && \text{(60c)}\\
&\leq \frac{\gamma(1-\lambda)}{1-\gamma\lambda}\|\Phi\theta^* - V^*\|_\infty + \|\Phi\Pi V^* - V^*\|_\infty, && \text{(60d)}
\end{aligned}$$

where we use the triangle inequality in (60a); Theorem D.1 in (60b); Lemma D.5 in (60c); Lemma D.4 in (60d).

Therefore, we obtain that

$$\|\Phi\theta^* - V^*\|_\infty \leq \frac{(1-\lambda\gamma)}{1-\gamma}\|\Pi V^* - V^*\|_\infty \leq \frac{(1-\lambda\gamma)}{1-\gamma}\zeta,$$

where we use Lemma D.5 in the second inequality.

D.2 Proof of Theorem 3.2

Before presenting the proof of Theorem 3.2, we first show two upper bounds that are needed in the assumptions of Theorem 3.1. We defer the proof of this result to Appendix D.3.

Proposition D.1. *Under the same assumptions as Theorem 3.2, we have $\|\theta(t)\|_\infty \leq \bar{\theta} := \frac{\bar{r}}{1-\gamma}$ holds for all t almost surely and $\|\theta^*\|_\infty \leq \bar{\theta}$. $|w(t)| \leq \bar{w} := \frac{2\bar{r}}{1-\gamma}$ also holds for all t almost surely.*

Now we come back to the proof of Theorem 3.2. Recall that we define F as the Bellman Policy Operator and the noise sequence $w(t)$ as

$$w(t) = r_t + \gamma\theta_{h(i_{t+1})}(t) - \mathbb{E}_{i'\sim\mathbb{P}(\cdot|i_t)}\left[r(i_t, i') + \gamma\theta_{h(i')}(t)\right].$$

Let θ^* be the unique solution of the equation

$$\Pi F(\Phi\theta^*) = \theta^*.$$

By the triangle inequality, we have that

$$\begin{aligned}
\|\Phi\cdot\theta(T) - V^*\|_\infty &\leq \|\Phi\cdot\theta(T) - \Phi\cdot\theta^*\|_\infty + \|\Phi\cdot\theta^* - V^*\|_\infty\\
&\leq \|\theta(T) - \theta^*\|_\infty + \|\Phi\cdot\theta^* - V^*\|_\infty. && \text{(61)}
\end{aligned}$$

We first bound the first term of (61) by Theorem 3.1. To do this, we first rewrite the update rule of TD learning with state aggregation (6) in the form of the SA update rule (3):

$$\begin{aligned}
\theta_{h(i_t)}(t+1) &= \theta_{h(i_t)}(t) + \alpha_t\big(F_{i_t}(\Phi\theta(t)) - \theta_{h(i_t)}(t) + w(t)\big),\\
\theta_j(t+1) &= \theta_j(t) \text{ for } j \neq h(i_t), j \in \mathcal{M}.
\end{aligned}$$

Now we verify all the assumptions of Theorem 3.1. Assumption 3.1 is assumed to be satisfied in the body of Theorem 3.2. As for Assumption 3.2, F is γ-contraction in the infinity norm because it is the

Bellman operator, and we can set $C = \frac{2\bar{r}}{1-\gamma}$ so that $C \geq (1+\gamma)\|y^*\|_\infty$ (see the discussion below Assumption 3.2). As for Assumption 3.3, by the definition of noise sequence $w(t)$, we see that

$$\begin{aligned}\mathbb{E}[w(t) \mid \mathcal{F}_t] &= \mathbb{E}\big[r_t + \gamma\theta_{h(i_{t+1})}(t) - \mathbb{E}_{i'\sim\mathbb{P}(\cdot|i_t)}\big[r(i_t,i') + \gamma\theta_{h(i')}(t)\big] \mid \mathcal{F}_t\big] \\ &= \mathbb{E}\big[r_t + \gamma\theta_{h(i_{t+1})}(t) \mid \mathcal{F}_t\big] - \mathbb{E}_{i'\sim\mathbb{P}(\cdot|i_t)}\big[r(i_t,i') + \gamma\theta_{h(i')}(t)\big] \\ &= 0.\end{aligned}$$

In addition, we can set $\bar{w} = \frac{2\bar{r}}{1-\gamma}$ according to Proposition D.1. Finally, we can set $\bar{\theta} = \frac{\bar{r}}{1-\gamma}$ according to Proposition D.1.

Therefore, by Theorem 3.1, we see that

$$\|\theta(T) - \theta^*\|_\infty \leq \frac{C_a}{\sqrt{T+t_0}} + \frac{C_a'}{T+t_0}, \text{ where} \tag{62}$$

$$\begin{aligned}C_a &= \frac{40H\bar{r}}{(1-\gamma)^2}\sqrt{K_2\log T}\cdot\sqrt{\log T + \log\log T + \log\left(\frac{4mK_2}{\delta}\right)}, \\ C_a' &= \frac{8\bar{r}}{(1-\gamma)^2}\max\left(\frac{144K_2H\log T}{\sigma'} + 4K_1(1+2K_2+4H), 2K_2\log T + t_0\right).\end{aligned}$$

As for the second term of (61), by Theorem D.2, we have that

$$\|\Phi\cdot\theta^* - V^*\|_\infty \leq \frac{\zeta}{1-\gamma}. \tag{63}$$

Substituting (62) and (63) into (61) finishes the proof.

D.3 Proof of Proposition D.1

We show $\|\theta(t)\|_\infty \leq \frac{\bar{r}}{1-\gamma}$ by induction on t. The statement holds for $t=0$ because we initialize $\theta(0) = \mathbf{0}$. Suppose the statement holds for t. By the induction assumption, we see that

$$\begin{aligned}\theta_{h(i_t)}(t+1) &= (1-\alpha_t)\theta_{h(i_t)}(t) + \alpha_t\big[r_t + \gamma\theta_{h(i_{t+1})}(t)\big] \\ &\leq (1-\alpha_t)\|\theta(t)\|_\infty + \alpha_t[r_t + \gamma\|\theta(t)\|_\infty] \\ &\leq (1-\alpha_t)\frac{\bar{r}}{1-\gamma} + \alpha_t\left[r_t + \gamma\cdot\frac{\bar{r}}{1-\gamma}\right] \\ &\leq \frac{\bar{r}}{1-\gamma}.\end{aligned}$$

For $j \neq h(i_t), j \in \mathcal{M}$, we have that

$$\theta_j(t+1) = \theta_j(t) \leq \|\theta(t)\|_\infty \leq \frac{\bar{r}}{1-\gamma}.$$

Hence the statement also holds for $t+1$. Therefore, we have showed $\|\theta(t)\|_\infty \leq \frac{\bar{r}}{1-\gamma}$ by induction.

By Theorem D.1, we know $\theta^* = \lim_{t\to\infty}\theta(t)$. Since we have already shown that $\|\theta(t)\|_\infty \leq \frac{\bar{r}}{1-\gamma}$ holds for all t, we must have $\|\theta^*\|_\infty \leq \frac{\bar{r}}{1-\gamma}$.

Using $\|\theta(t)\|_\infty \leq \frac{\bar{r}}{1-\gamma}$, we see that

$$\begin{aligned}|w(t)| &\leq |r_t| + \gamma\big|\theta_{h(i_{t+1})}(t)\big| - \big|\mathbb{E}_{i'\sim\mathbb{P}(\cdot|i_t)}\big[r(i_t,i') + \gamma\theta_{h(i')}(t)\big]\big| \\ &\leq 2\bar{r} + 2\gamma\bar{\theta} \\ &= \frac{2\bar{r}}{1-\gamma}.\end{aligned}$$

D.4 Application of the SA Scheme to Q-learning with State and Action Aggregation

We study Q-learning with state and action aggregation in a setting that is a generalization of the tabular setting studied in [37]. Specifically, we consider an MDP M with a finite state space $\mathcal{S}$ and finite action space $\mathcal{A}$. Suppose the transition probability is given by $\mathbb{P}(s_{t+1} = s' \mid s_t = s, a_t = a) = \mathbb{P}(s' \mid s, a)$, and the stage reward at time step t is a random variable r_t with its expectation given by R_{s_t,a_t}. Under a stochastic policy π, the Q function (vector) $Q^\pi \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ is defined as

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

where $0 \leq \gamma < 1$ is the discounting factor. We use Q^* to denote the Q function corresponding to the optimal policy π^*.

Similar to [37], we assume the trajectory $\{(s_t, a_t, r_t)\}_{t=0}^{\infty}$ is sampled by implementing a fixed behavioral stochastic policy π. In Q-learning with state and action aggregation, the state abstraction ψ_1 operates on the state space $\mathcal{S}$ and the action abstraction ψ_2 operates on action space $\mathcal{A}$. For simplicity of notation, we define the abstraction space as $\mathcal{M} = \psi_1(\mathcal{S}) \times \psi_2(\mathcal{A})$ and the abstraction operator $h : \mathcal{S} \times \mathcal{A} \to \mathcal{M}$ as $h(s, a) = (\psi_1(s), \psi_2(a))$. The update rule for Q-learning with state and action aggregation is then given by

$$\begin{aligned} \theta_{h(s_t,a_t)}(t+1) &= (1-\alpha_t)\theta_{h(s_t,a_t)}(t) + \alpha_t\left[r_t + \gamma \max_{a\in\mathcal{A}} \theta_{h(s_{t+1},a)}(t)\right], \\ \theta_j(t+1) &= \theta_j(t) \text{ for } j \neq h(s_t, a_t). \end{aligned} \tag{64}$$

As a remark, some previous work considers abstraction on the state space $\mathcal{S}$ but does not compress the action space (see [21]). In contrast, our setting also compresses the action space, and when ψ_2 is the identity map, our setting reduces to the case with only state aggregation.

To apply the result in Section 3.1, we define function F as the *Bellman Optimality Operator*, i.e.

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

It is shown in [3] that Q^* is the unique fixed point of function F. By viewing $\mathcal{S} \times \mathcal{A}$ as $\mathcal{N}$, we can define matrix $\Phi \in \mathcal{N} \times \mathcal{M}$ as in (4). We can rewrite the update rule (64) as

$$\begin{aligned} \theta_{h(s_t,a_t)}(t+1) &= \theta_{h(s_t,a_t)}(t) + \alpha_t\big[F_{s_t,a_t}(\Phi\theta(t)) - \theta_{h(s_t,a_t)}(t) + w(t)\big], \\ \theta_j(t+1) &= \theta_j(t) \text{ for } j \neq h(s_t, a_t), \end{aligned}$$

where

$$\begin{aligned} w(t) &= r_t + \gamma \max_{a\in\mathcal{A}} \theta_{h(s_{t+1},a)}(t) - F_{s_t,a_t}(\Phi\theta(t)) \\ &= (r_t - R_{s_t,a_t}) + \gamma\left[\max_{a\in\mathcal{A}} \theta_{h(s_{t+1},a)}(t) - \mathbb{E}_{s'\sim\mathbb{P}(\cdot|s_t,a_t)} \max_{a'\in\mathcal{A}} \theta_{h(s',a')}(t)\right]. \end{aligned}$$

Hence we have $\mathbb{E}[w(t) \mid \mathcal{F}_t] = 0$. In order to apply Theorem 3.1, we need the following assumption on the induced Markov chain of stochastic policy π which is standard, cf. [37].

Assumption D.5. *The following conditions hold:*

1. *For each time step t, the stage reward r_t satisfies $|r_t| \leq \bar{r}$ almost surely.*
2. *Under the behavioral policy π, the induced Markov chain (s_t, a_t) with state space $\mathcal{S} \times \mathcal{A}$ satisfies Assumption 3.1 with stationary distribution d and parameters σ', K_1, K_2.*

The next assumption is approximate Q^*-irrelevant abstraction, which measures the quality of the abstraction map and is standard in the literature (see [21]).

Assumption D.6. *There exists an abstract Q function $q : \mathcal{M} \to \mathbb{R}$ such that $\|\Phi q - Q^*\|_\infty \leq \epsilon_{Q^*}$.*

We can now state our theorem for Q-learning with state aggregation.

Theorem D.6. *Under Assumption D.5 and D.6, suppose the step size of Q-learning with state aggregation is given by $\alpha_t = \frac{H}{t+t_0}$, where $t_0 = \max(4H, 2K_2 \log T)$ and $H \geq \frac{2}{\sigma'(1-\gamma)}$. Then, with probability at least $1-\delta$,*

$$\|\Phi \cdot \theta(T) - Q^*\|_\infty \leq \frac{C_a}{\sqrt{T+t_0}} + \frac{C_a'}{T+t_0} + \frac{2\epsilon_{Q^*}}{1-\gamma}, \text{ where}$$

$$C_a = \frac{40H\bar{r}}{(1-\gamma)^2}\sqrt{K_2 \log T} \cdot \sqrt{\log T + \log\log T + \log\left(\frac{4mK_2}{\delta}\right)},$$

$$C_a' = \frac{8\bar{r}}{(1-\gamma)^2}\max\left(\frac{144K_2 H \log T}{\sigma'} + 4K_1(1+2K_2+4H), 2K_2 \log T + t_0\right).$$

Proof of Theorem D.6. Define θ^* as the unique solution of equation $\theta = \Pi F(\Phi\theta)$, where the definition of Π is given in (5). Under Assumption D.5, we see that $\|\theta^*\|_\infty \leq \frac{\bar{r}}{1-\gamma}$: otherwise, by assuming that $|\theta_i^*| = \|\theta^*\|_\infty > \frac{\bar{r}}{1-\gamma}$, we can derive a contradiction that $\|\Pi F(\Phi\theta^*)\|_\infty < |\theta_i^*|$. To see this, recall that linear operators Π and Φ are non-expansions in the infinity norm (see Appendix C.1), and $\|F(v)\|_\infty < \|v\|_\infty$ for a vector $v \in \mathbb{R}^{\mathcal{N}}$ if $\|v\|_\infty > \frac{\bar{r}}{1-\gamma}$.

Further, using a similar approach with the proof of Proposition D.1, we also see that

$$\|\theta(t)\|_\infty \leq \bar{\theta} := \frac{\bar{r}}{1-\gamma}, |w(t)| \leq \bar{w} := \frac{2\bar{r}}{1-\gamma}$$

hold for all t almost surely.

Therefore, by Theorem 3.1, we obtain that

$$\|\theta(T) - \theta^*\|_\infty \leq \frac{C_a}{\sqrt{T+t_0}} + \frac{C_a'}{T+t_0}. \tag{65}$$

To finish the proof of Theorem D.6, we only need to show that

$$\|\Phi\theta^* - Q^*\| \leq \frac{2\epsilon_{Q^*}}{1-\gamma}. \tag{66}$$

Given the behavioral policy π, we use $\{d_{s,a} \mid (s,a) \in \mathcal{S} \times \mathcal{A}\}$ to denote the stationary distribution under policy π. Recall that we define $\mathcal{M} = \psi_1(\mathcal{S}) \times \psi_2(\mathcal{A})$. For each abstract state-action pair $(x,y) \in \mathcal{M}$, we define a distribution $p_{(x,y)}$ over $h^{-1}(x,y)$ such that

$$p_{(x,y)}(s,a) = \frac{d_{s,a}}{\sum_{(\tilde{s},\tilde{a}) \in h^{-1}(x,y)} d_{\tilde{s},\tilde{a}}}, \forall (s,a) \in h^{-1}(x,y).$$

Using the set of distributions $\{p_{(x,y)} \mid (x,y) \in \mathcal{M}\}$, we define two new MDPs:

$$M_\psi = (\psi_1(\mathcal{S}), \psi_2(\mathcal{A}), P_\psi, R_\psi, \gamma), \tag{67}$$

where $(R_\psi)_{x,y} = \mathbb{E}_{(s,a)\sim p_{(x,y)}}[R_{s,a}]$, and $P_\psi(x' \mid x,y) = \mathbb{E}_{(s,a)\sim p_{(x,y)}}[P(x' \mid s,a)]$; and

$$M_\psi' = (\mathcal{S}, \mathcal{A}, P_\psi', R_\psi', \gamma), \tag{68}$$

where $(R_\psi')_{s,a} = \mathbb{E}_{(\tilde{s},\tilde{a})\sim p_{h(s,a)}}[R_{\tilde{s},\tilde{a}}], P_\psi'(s' \mid s,a) = \mathbb{E}_{(\tilde{s},\tilde{a})\sim p_{h(s,a)}}[P(s' \mid \tilde{s},\tilde{a})]$.

We use Γ to denote the Bellman Optimality Operator. For simplicity, we use the subscript to distinguish the value functions (V^*), the state-action value functions (Q^*), and the Bellman Optimality Operators (Γ) of the three MDPs M, M_ψ and M_ψ'. Notice that Γ_M is identical with F.

We can show that θ^* is identical with the state-action value function of M_ψ, i.e.,

$$\theta^* = Q^*_{M_\psi}. \tag{69}$$

To see this, we notice that $(\Phi\theta^*)_{s,a} = \theta^*_{h(s,a)}$. Hence we get that

$$\begin{aligned} F(\Phi\theta^*)_{s,a} &= [\Gamma_M \Phi\theta^*]_{s,a} \\ &= R_{s,a} + \mathbb{E}_{s'\sim P(s,a)}\left[\max_a (\Phi\theta^*)_{s',a}\right] \\ &= R_{s,a} + \mathbb{E}_{s'\sim P(s,a)}\left[\max_a \theta^*_{h(s',a)}\right]. \end{aligned}$$

Using this, we further obtain that

$$\begin{aligned}
(\Pi F(\Phi\theta^*))_{x,y} &= \sum_{(s,a)\in h^{-1}(x,y)} \frac{d_{s,a}}{\sum_{(\tilde{s},\tilde{a})\in h^{-1}(x,y)} d_{\tilde{s},\tilde{a}}} \Big(R_{s,a} + \mathbb{E}_{s'\sim P(s,a)}\Big[\max_a \theta^*_{h(s',a)}\Big]\Big) \\
&= \sum_{(s,a)\in h^{-1}(x,y)} p_{(x,y)}(s,a)\Big(R_{s,a} + \mathbb{E}_{s'\sim P(s,a)}\Big[\max_a \theta^*_{h(s',a)}\Big]\Big) \\
&= (R_\psi)_{x,y} + \sum_{(s,a)\in h^{-1}(x,y)} p_{(x,y)}(s,a) \sum_{x'\in\psi_1(\mathcal{S})} P(x' \mid s,a) \max_a \theta^*_{x',\psi_2(a)} \\
&= (R_\psi)_{x,y} + \sum_{x'\in\psi_1(\mathcal{S})} P_\psi(x' \mid x,y) \max_{y'} \theta^*_{x',y'} \\
&= [\Gamma_{M_\psi}\theta^*]_{x,y}.
\end{aligned}$$

Since we have $\Pi F(\Phi\theta^*) = \theta^*$ by definition, we see that

$$[\Gamma_{M_\psi}\theta^*]_{x,y} = \theta^*_{x,y}, \forall (x,y) \in \mathcal{M}.$$

Thus we have shown that $\theta^* = Q^*_{M_\psi}$.

Next, we observe that the state-value function of MDP M'_ψ is given by

$$Q^*_{M'_\psi} = \Phi Q^*_{M_\psi}. \tag{70}$$

This is because

$$\begin{aligned}
\Big(\Gamma_{M'_\psi}(\Phi Q^*_{M_\psi})\Big)_{s,a} &= (R'_\psi)_{s,a} + \gamma \sum_{s'\in\mathcal{S}} P'_\psi(s' \mid s,a) \max_{a'} (\Phi Q^*_{M_\psi})_{s',a'} \\
&= (R'_\psi)_{s,a} + \gamma \langle P'_\psi(s,a), \Phi V^*_{M_\psi}\rangle \\
&= \sum_{(\tilde{s},\tilde{a})\in h^{-1}(h(s,a))} p_{h(s,a)}(\tilde{s},\tilde{a})\Big(R_{\tilde{s},\tilde{a}} + \gamma\langle P(\tilde{s},\tilde{a}), \Phi V^*_{M_\psi}\rangle\Big) \qquad \text{(71a)} \\
&= \sum_{(\tilde{s},\tilde{a})\in h^{-1}(h(s,a))} p_{h(s,a)}(\tilde{s},\tilde{a}) R_{\tilde{s},\tilde{a}} \\
&\quad + \sum_{(\tilde{s},\tilde{a})\in h^{-1}(h(s,a))} p_{h(s,a)}(\tilde{s},\tilde{a})\gamma\langle P(\tilde{s},\tilde{a}), \Phi V^*_{M_\psi}\rangle \\
&= (R_\psi)_{h(s,a)} + \gamma\langle P_\psi(h(s,a)), V^*_{M_\psi}\rangle \qquad \text{(71b)} \\
&= (Q^*_{M_\psi})_{h(s,a)} \\
&= (\Phi Q^*_{M_\psi})_{s,a},
\end{aligned}$$

where we use the definition of M'_ψ (see (68)) in (71a); we use the definition of M_ψ (see (67)) in (71b).

By (70), we see that

$$\Big\|\Phi Q^*_{M_\psi} - Q^*_M\Big\|_\infty = \Big\|Q^*_{M'_\psi} - Q^*_M\Big\|_\infty \le \frac{1}{1-\gamma}\Big\|\Gamma_{M'_\psi} Q^*_M - Q^*_M\Big\|_\infty. \tag{72}$$

We further notice that

$$
\begin{aligned}
&\left|(\Gamma^*_{M_\psi} Q^*_M)_{s,a} - (Q^*_M)_{s,a}\right| \\
&= \left|(R'_\psi)_{s,a} + \gamma\langle P_\psi(s,a), V^*_M\rangle - (Q^*_M)_{s,a}\right| \\
&= \left|\left(\sum_{(\tilde{s},\tilde{a})\in h^{-1}(h(s,a))} p_{h(s,a)}(\tilde{s},\tilde{a})(R_{\tilde{s},\tilde{a}} + \gamma\langle P(\tilde{s},\tilde{a}), V^*_M\rangle)\right) - (Q^*_M)_{s,a}\right| && \text{(73a)} \\
&= \left|\sum_{(\tilde{s},\tilde{a})\in h^{-1}(h(s,a))} p_{h(s,a)}(\tilde{s},\tilde{a})((Q^*_M)_{\tilde{s},\tilde{a}} - (Q^*_M)_{s,a})\right| \\
&\leq \sum_{(\tilde{s},\tilde{a})\in h^{-1}(h(s,a))} p_{h(s,a)}(\tilde{s},\tilde{a})|(Q^*_M)_{\tilde{s},\tilde{a}} - (Q^*_M)_{s,a}| \\
&\leq \sum_{(\tilde{s},\tilde{a})\in h^{-1}(h(s,a))} p_{h(s,a)}(\tilde{s},\tilde{a})(2\epsilon_{Q^*}) && \text{(73b)} \\
&= 2\epsilon_{Q^*},
\end{aligned}
$$

where we use the definition of M_ψ in (73a); we use Assumption D.6 in (73b).

Substituting (73) into (72) gives that

$$
\left\|\Phi Q^*_{M_\psi} - Q^*_M\right\|_\infty \leq \frac{2\epsilon_{Q^*}}{1-\gamma}. \tag{74}
$$

Combining (69) and (74) finishes the proof. □