arXiv:1912.02906v3 [math.OC] 1 Nov 2021

Scalable Reinforcement Learning for Multi-Agent Networked Systems

Guannan Qu* Adam Wierman† Na Li‡

Abstract

We study reinforcement learning (RL) in a setting with a network of agents whose states and actions interact in a local manner where the objective is to find localized policies such that the (discounted) global reward is maximized. A fundamental challenge in this setting is that the state-action space size scales exponentially in the number of agents, rendering the problem intractable for large networks. In this paper, we propose a Scalable Actor Critic (SAC) framework that exploits the network structure and finds a localized policy that is an $O(\rho^{\kappa+1})$-approximation of a stationary point of the objective for some $\rho \in (0,1)$, with complexity that scales with the local state-action space size of the largest κ-hop neighborhood of the network. We illustrate our model and approach using examples from wireless communication, epidemics and traffic.

1 Introduction

The modeling and optimization of networked systems such as wireless communication networks and traffic networks is a long-standing challenge. Typically analytic models must make numerous assumptions to obtain tractable models as a result of the complexity of the systems, which include many unknown, or unmodeled dynamics. Given the success of Reinforcement Learning (RL) in a wide array of domains such as game play [Silver et al., 2016; Mnih et al., 2015], robotics [Duan et al., 2016], and autonomous driving [Li et al., 2019], it has emerged as a promising tool for tackling the complexity of networked systems. However, when seeking to use RL in the context of the control and optimization of large-scale networked systems, scalability quickly becomes an issue. The goal of this paper is to develop *scalable* multi-agent RL for networked systems.

Motivated by real-world networked systems like wireless communication, epidemics, and traffic, we consider an RL model of n agents with *local interaction structure*. Specifically, each agent i has local state s_i, local action a_i and the agents are associated with an underlying dependence graph $\mathcal{G}$ and interact locally, i.e, the distribution of $s_i(t+1)$ only depends on the current states of the local neighborhood of i as well as the local $a_i(t)$. Further, each agent is associated with stage reward r_i that is a function of s_i, a_i, and the global stage reward is the average of r_i. In this setting, the design goal is to find a decision policy that maximizes the (discounted) global reward. This setting captures a wide range of applications, e.g. epidemics [Mei et al., 2017], social networks [Chakrabarti

*Department of Electrical and Computer Engineering, Carnegie Mellon University. Email: gqu@andrew.cmu.edu

†Department of Computing and Mathematical Sciences, California Institute of Technology. Email: adamw@caltech.edu

‡School of Engineering and Applied Sciences, Harvard University. Email: nali@seas.harvard.edu

et al., 2008; Llas et al., 2003], wireless communication networks [Zocca, 2019; Vogels et al., 2003], queueing networks [Papadimitriou and Tsitsiklis, 1999], smart transportation [Zhang and Pavone, 2016], and smart building systems [Wu et al., 2016; Zhang et al., 2017].

A fundamental difficulty when applying RL to such networked systems is that, even if individual state and action spaces are small, the entire state profile $(s_1, \ldots, s_n)$ and the action profile $(a_1, \ldots, a_n)$ can take values from a set of size exponentially large in n. This "curse of dimensionality" renders the problem unscalable. For example, most RL algorithms such as temporal difference (TD) learning or Q-learning require storage of a Q-function [Bertsekas and Tsitsiklis, 1996] whose size is the same as the state-action space, which is exponentially large in n. Such scalability issues have indeed been observed in previous research on variants of the problem we study, e.g. in multi-agent RL [Littman, 1994; Bu et al., 2008] and factored Markov Decision Process (MDP) [Kearns and Koller, 1999; Guestrin et al., 2003]. A variety of approaches have been proposed to manage this issue, e.g. the idea of "independent learners" in Tan [1993]; Claus and Boutilier [1998]; or function approximation schemes [Tsitsiklis and Van Roy, 1997]. However, such approaches lack rigorous optimality guarantees. In fact, it has been suggested that such MDPs with exponentially large state spaces may be fundamentally intractable, e.g., see Blondel and Tsitsiklis [2000].

In addition to the scalability issue, another challenge is that, even if an optimal policy that maps a global state $(s_1, \ldots, s_n)$ profile to a global action $(a_1, \ldots, a_n)$ can be found, it is usually impractical to implement such a policy for real-world networked systems because of the limited information and communication among agents. For example, in large scale networks, each agent i may only be able to implement *localized policies*, where its action a_i only depends on its own state s_i. Designing such localized policies with global network performance guarantees can also be challenging, e.g., see Rotkowitz and Lall [2005].

The challenges described above highlight the difficulty of applying RL to control large scale networked systems; however, the network itself provides some structure, particularly the local interaction structure, that can potentially be exploited. The question that motivates this paper is: *Can the network structure be utilized to develop scalable RL algorithms that provably find a (near-)optimal localized policy?*

Contributions. In this work we propose a framework that exploits properties of the network structure to develop RL to learn *localized* policies for large-scale networked systems in a *scalable* manner. Specifically, our main result (Theorem 4) shows that our algorithm, Scalable Actor Critic (SAC), finds a localized policy that is a $O(\rho^{\kappa+1})$-approximation of a stationary point of the objective function, with complexity that scales with the local state-action space size of the largest κ-hop neighborhood. To the best of our knowledge, our results are the first to provide such provable guarantees for scalable RL of localized policies in multi-agent networked settings.

The key technique underlying our results is we prove that, under the local interaction structure, the Q-function satisfies an *exponential decay property* (Definition 1), where the Q-function's dependence on far away nodes shrink exponentially in their graph distance with rate $\rho \leq \gamma$, where γ is the discounting factor. This leads to a tractable approximation of the Q-function. In particular, despite the Q-function itself being intractable to compute due to the large state-action space size, we introduce a *truncated Q-function* which only depends on a small spatial horizon, (see Lemma 3) that can be computed efficiently and can be used in an actor critic framework which yields an $O(\rho^{\kappa})$-approximation. This technique is novel and is a contribution in its own right. It can be used broadly to develop RL for networked settings beyond the specific actor critic algorithm we propose in this paper.

To illustrate our model and our results, we provide stylized examples of applications in three areas: multi-access wireless communication, epidemics, and traffic signal control in Section 2.2. We conduct numerical experiments to demonstrate the performance of the approach using both synthetic examples and an application to wireless communication in Section 5.

Related Literature. Our problem falls into the category of the "succinctly described" MDPs in Blondel and Tsitsiklis [2000, Section 5.2], where the state/action space is a product space formed by the individual state/action space of multiple agents. As the state/action space is exponentially large, such problems are not scalable in general, even when the problem has structure [Blondel and Tsitsiklis, 2000; Whittle, 1988; Papadimitriou and Tsitsiklis, 1999]. Despite this, there is a large literature on RL/MDPs in multi-agent settings, which we discuss below.

Multi-agent RL dates back to the early work of Littman [1994]; Claus and Boutilier [1998]; Littman [2001]; Hu and Wellman [2003] (see Bu et al. [2008] for a review) and has been actively studied, e.g. Zhang et al. [2018]; Kar et al. [2013]; Macua et al. [2015]; Mathkar and Borkar [2017]; Wai et al. [2018], see a more recent review in Zhang et al. [2021]. Multi-agent RL encompasses a broad range of settings including competitive agents and Markov games. The case most relevant to ours is the cooperative multi-agent RL where typically, the agents can take their own actions but they share a common global state and maximize a global reward [Bu et al., 2008]. This is in contrast to the model we study, in which each agent has its own state and acts upon its own state. Despite the existence of a global state, multi-agent RL still faces scalability issues since the joint-action space is exponentially large. Methods have been proposed to deal with this, including independent learners [Tan, 1993; Claus and Boutilier, 1998; Matignon et al., 2012], where each agent employs a single-agent RL method. While successful in some cases, the independent learner approach can suffer from instability [Matignon et al., 2012]. Alternatively, one can use function approximation schemes to approximate the large Q-table, e.g. linear function approximation [Zhang et al., 2018] or neural networks [Lowe et al., 2017]. Such methods can reduce computation complexity significantly, but it is unclear whether the performance loss caused by the function approximation is small. In contrast, our technique not only reduces computation but also guarantees small performance loss.

Factored MDPs are problems where every agent has its own state and the state transition factorizes in a way similar to our model [Kearns and Koller, 1999; Guestrin et al., 2003; Osband and Van Roy, 2014]. However, they differ from the model we consider in that each agent does not have its own action. Instead, there is a global action affecting every agent. Despite the difference, Factored MDPs still suffer from scalability issues. Similar approaches as in the case of multi-agent RL are used, e.g., Guestrin et al. [2003] proposes a class of "factored" linear function approximators; however, it is unclear whether the loss caused by the approximation is small.

Other Related Work. Our work is also related to weakly coupled MDPs, where every agent has its own state and action but their transition is decoupled [Meuleau et al., 1998]. Additionally, our model shares some similarity with Glauber dynamics in physics [Lokhov et al., 2015; Mezard and Montanari, 2009], though our focus is very different from these works. As we consider the class of localized policies, another related line of work is Partially Observable MDP (POMDP) [Nair et al., 2005; Oliehoek and Amato, 2016; Bertsekas, 2005], though the formulations and results we have are very different from those works.

Finally, this work is related to our earlier work Qu and Li [2019], which assumes the full knowledge of the MDP model (not RL) and imposes strong assumptions on the graph. In contrast, our work here does not need knowledge of the MDP and significantly relaxes the assumptions.

2 Preliminaries

In this section, we introduce our model, provide a few illustrative examples, and provide important background in RL that underlies our analysis. Throughout this paper, $\|\cdot\|$ denotes Euclidean norm and $\|\cdot\|_\infty$ denotes infinity norm. Notation t and T are reserved as iteration counters for the inner loop of the algorithm to be introduced later, m and M for the outer loop, and κ is used for counting the hops of neighbors. Notation $O(\cdot)$ hides constants and $\tilde{O}(\cdot)$ hides log factors with respect to iteration variables T, M and variable κ. The total variation distance for two distributions π, π' over a finite set $\mathcal{S}$ is defined as $\mathrm{TV}(\pi, \pi') = \sup_{E \subset \mathcal{S}} |\pi(E) - \pi'(E)|$.

2.1 Model

We consider a network of n agents that are associated with an underlying undirected graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N} = \{1, \ldots, n\}$ is the set of agents and $\mathcal{E} \subset \mathcal{N} \times \mathcal{N}$ is the set of edges. Each agent i is associated with state $s_i \in \mathcal{S}_i$, $a_i \in \mathcal{A}_i$ where $\mathcal{S}_i$ and $\mathcal{A}_i$ are finite sets. The global state is denoted as $s = (s_1, \ldots, s_n) \in \mathcal{S} := \mathcal{S}_1 \times \cdots \times \mathcal{S}_n$ and similarly the global action $a = (a_1, \ldots, a_n) \in \mathcal{A} := \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$. At time t, given current state $s(t)$ and action $a(t)$, the next individual state $s_i(t+1)$ is independently generated and is only dependent on neighbors:

$$P(s(t+1)|s(t), a(t)) = \prod_{i=1}^{n} P(s_i(t+1)|s_{N_i}(t), a_i(t)), \tag{1}$$

where notation N_i means the neighborhood of i (including i itself) and notation s_{N_i} means the states of the agents in N_i. In addition, for integer $\kappa \geq 0$, we use N_i^κ to denote the κ-hop neighborhood of i, i.e. the nodes whose graph distance to i has length less than or equal to κ. We also let $f(\kappa) = \sup_i |N_i^\kappa|$.

Each agent is associated with a class of localized policies $\zeta_i^{\theta_i}$ parameterized by θ_i. The localized policy $\zeta_i^{\theta_i}(a_i|s_i)$ is a distribution on the local action a_i conditioned on the local state s_i, and each agent, conditioned on observing $s_i(t)$, takes an action $a_i(t)$ independently drawn from $\zeta_i^{\theta_i}(\cdot|s_i(t))$. We use $\theta = (\theta_1, \ldots, \theta_n)$ to denote the tuple of the localized policies $\zeta_i^{\theta_i}$, and also use $\zeta^\theta(a|s) = \prod_{i=1}^n \zeta_i^{\theta_i}(a_i|s_i)$ to denote the joint policy, which is a product distribution of the localized policies as each agent acts independently.

Further, each agent is associated with a stage reward function $r_i(s_i, a_i)$ that depends on the local state and action, and the global stage reward is $r(s, a) = \frac{1}{n}\sum_{i=1}^n r_i(s_i, a_i)$. The objective is to find localized policy tuple θ such that the discounted global stage reward is maximized, starting from some initial state distribution π_0,

$$\max_\theta J(\theta) := \mathbb{E}_{s\sim\pi_0}\mathbb{E}_{a(t)\sim\zeta^\theta(\cdot|s(t))}\left[\sum_{t=0}^{\infty}\gamma^t r(s(t), a(t))\middle| s(0) = s\right]. \tag{2}$$

Remark 1. *In the state transition (1) of our model, the distribution of each node's next state is allowed to depend on its neighbors' states $s_{N_i}(t)$, but only on its own action $a_i(t)$ as opposed to its neighbors' actions $a_{N_i}(t)$. This restriction is imposed only for simplicity of exposition. With a simple change of notation, our model, algorithm and analysis can be extended to the more general dependence on $a_{N_i}(t)$. Similarly, each agent's reward function $r_i(s_i, a_i)$ can be generalized to depend on its neighbors' state-action pairs, i.e. $r_i(s_{N_i}, a_{N_i})$, and each agent's localized policy can also be generalized to depend on its neighbors' states, i.e. $\zeta_i^{\theta_i}(a_i|s_{N_i})$.*

Remark 2. *In the paper, the interaction graph $\mathcal{G}$ is undirected, but our model and results can be easily generalized to the directed graph setting without essential changes in the algorithm and the analysis. In detail, to generalize to the directed graph case, the only change needed is to redefine the "neighbors". Specifically, N_i needs to be redefined as the "in-neighborhood" of i (including i itself), i.e. i itself and the set of nodes that have a directed link pointing towards i. In addition, N_i^κ needs to be redefined as the κ-hop "in-neighborhood" of i, i.e. the nodes whose shortest directed link to i has a length less than or equal to κ (including i itself).*

2.2 Examples

In this section, we provide three networked system examples in wireless communication, epidemics, and traffic that feature the local dependence structure we study in this paper. For ease of exposition, we present simple versions of these examples, keeping the essence of the model and highlighting the dependence structure while ignoring some application-specific details.

Wireless Communication. We consider a wireless network with multiple access points [Zocca, 2019], where there is a set of users $\mathcal{N} = \{1, 2, \cdots, n\}$, and a set of network access points $Y = \{y_1, y_2, \cdots, y_m\}$. Each user i only has access to a subset $Y_i \subseteq Y$ of the access points. We define the interaction graph as the conflict graph, in which two users i and j are neighbors if and only if they share an access point, i.e. the neighbors of user i is $N_i = \{j \in \mathcal{N} : Y_i \cap Y_j \neq \emptyset\}$. Each user i maintains a queue of packets defined as follows. At time step t, with probability p_i, user i receives a new packet with an initial deadline d_i. Then, user i can choose to send the earliest packet in its queue to one access point in its available set Y_i, or not send anything at all. If an action of sending to $y_k \in Y_i$ is taken, and if no other users send to the same access point at this time, then the earliest packet in user i's queue is transmitted with success probability q_k which depends on the access point y_k; however, if another user also chooses to send to y_k, then there is a conflict and no transmission occurs. If the packet is successfully transmitted, it will be removed from user i's queue and user i will get a reward of 1. After this, the system moves to the next time step, with all deadlines of the remaining packets decreasing by 1 and packets with deadline 0 being discarded. In this example, the local state s_i of user i is a characterization of its queue of packets, and is represented by a d_i binary tuple $s_i = (e_1, e_2, \cdots, e_{d_i}) \in \mathcal{S}_i = \{0, 1\}^{d_i}$, where for each $\ell \in \{1, \ldots, d_i\}$, $e_\ell \in \{0, 1\}$ indicates whether user i has a packet with remaining deadline ℓ. The action space is $\mathcal{A}_i = \{\text{null}\} \cup Y_i$, where null represents the action of not sending. The detailed transition is provided in Table 1, where $s_i(t+1)$ only depends on $s_{N_i}(t), a_{N_i}(t)$, which fits into the local interaction structure we consider. The local reward is given by $r_i(s_{N_i}(t), a_{N_i}(t)) = 1$ in the case of the last row of Table 1, and $r_i(s_{N_i}(t), a_{N_i}(t)) = 0$ in all other cases. The local state space $\mathcal{S}_i$, the local action space $\mathcal{A}_i$, the local transition probabilities in Table 1 and the local reward function $r_i(\cdot)$ form a networked MDP model described in Section 2.1. The above model serves as a basis for more complex multi-access wireless communication models studied in the literature, including those with multiple channels [Block and Van Houdt, 2016], (imperfect) carrier sensing [Kim et al., 2011]. Existing analytical approaches typically require knowledge of modeling details and parameters such as the packet arrival rate [Tassiulas and Ephremides, 1990; Yun et al., 2012], while our RL-based approach does not require such knowledge and learns to improve performance in a model-free manner.

Epidemic Network. We consider an SIS (Susceptible-Infected-Susceptible) epidemic network model [Mei et al., 2017; Ahn, 2014; Azizan Ruhi et al., 2016a], where there is a undirected graph of nodes $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, and each node has a binary state space $\mathcal{S}_i = \{\text{susceptible}, \text{infected}\}$, as well

$s_i(t)$	$a_i(t)$	$s_{N_i/\{i\}}(t), a_{N_i/\{i\}}(t)$	$s_i(t+1)$
any	null	any	Left shift $s_i(t)$ and append Bernoulli(p_i).
all zero	not null	any	Left shift $s_i(t)$ and append Bernoulli(p_i).
not all zero	$y_k \in Y_i$	$\exists j \in N_i/\{i\}$ s.t. $s_j(t)$ not all zero, $a_j(t) = y_k$	Left shift $s_i(t)$ and append Bernoulli(p_i).
all other cases (denote $a_i(t) = y_k$)			Flip left most "1" in $s_i(t)$ to "0" w.p. q_k, then left shift $s_i(t)$ and append Bernoulli(p_i).

Table 1: State transition for the wireless communication example. "Left shift" means for a binary tuple, discarding the left most bit. Bernoulli(p_i) means a random variable sampled i.i.d. from the Bernoulli distribution that has probability p_i to be 1, and probability $1 - p_i$ to be 0.

as a finite action space $\mathcal{A}_i$ with action $a_i \in \mathcal{A}_i$ representing epidemic control measures like different levels of vaccination [Preciado et al., 2013]. The evolution of the states follows a local interaction structure: the probability of a node turning from susceptible to infected depends on the whether its neighboring nodes are infected or not as well as its control action in place [Azizan Ruhi et al., 2016b; Preciado et al., 2013]; the probability of a node turning from infected to susceptible depends on the recovering rate. More precisely, $s_i(t+1)$ only depends on $s_{N_i}(t)$ and $a_i(t)$, and the state transition is provided by,

$$P(s_i(t+1) = \text{susceptible} | s_{N_i}(t), a_i(t)) = \begin{cases} [1 - \beta_i(a_i(t))]^{|\{j \in N_i/\{i\} : s_j(t)=1\}|}, & \text{if } s_i(t) = \text{susceptible}, \\ \delta_i, & \text{if } s_i(t) = \text{infected}, \end{cases} \tag{3}$$

where $\delta_i \in (0,1)$ is a given recovering rate parameter, and $\beta_i : \mathcal{A}_i \to (0,1)$ is a given transmission rate function and it depends on the control action $a_i(t)$, and $|\{j \in N_i/\{i\} : s_j(t) = 1\}|$ is the number of neighboring nodes excluding i itself that are infected. The local reward of each node is given by,

$$r_i(s_i, a_i) = \mathbf{1}(s_i = \text{susceptible}) - c_i(a_i), \tag{4}$$

which consists of two parts: a positive reward of 1 if the node is free from infection, subtracting a given cost function $c_i(a_i)$ on the epidemic control measure a_i. In this setup, the expected global reward is a weighted balance between the overall infection level and the epidemic control cost. The above defined local state space $\mathcal{S}_i$, local action space $\mathcal{A}_i$, local transition probabilities in (3) and local rewards in (4) form a networked MDP model in Section 2.1. We comment that the above SIS model is a basis for more complex models, e.g. those with "exposed" and "recovered" states [Kuznetsov and Piccardi, 1994; Britton, 2010] or more complex control interventions [Morris et al., 2020]. These more complex models can also be captured using the framework above. We reiterate that the approach in our paper does not require knowledge of model specifications and learns in a model-free manner.

Traffic Network. Lastly, we consider a traffic signal control problem adapted from Varaiya [2013]. In this setting, each node i represents a road link and the interaction graph represents the physical connection of the road links. Given road link i, the local state $s_i = (x_{i,j})_{i \to j}$ is a tuple of variables with j ranging from neighboring links i can turn to, and $x_{i,j}$ is the number of vehicles on link i that intend to turn to link j, and can only take values in $[S] = \{0, 1, \ldots, S\}$. Correspondingly,

the local state space is $\mathcal{S}_i = [S]^{N_i/\{i\}}$. Similarly, the local action $a_i = (y_{i,j})_{i\to j}$ is the binary traffic signal tuple with $y_{i,j}$ controlling the on-off of turn movement $(i \to j)$, and the local action space is $\mathcal{A}_i = \{0,1\}^{\{N_i\}/\{i\}}$. At each time, a random amount of vehicles on the queue $x_{i,j}$ will flow into link j when the traffic signal $y_{i,j}$ is on. Meanwhile, link i will receive vehicles from other incoming links, a random fraction of which are then assigned to each of the queues in $(x_{i,j})_{i\to j}$. Mathematically,

$$x_{i,j}(t+1) = \Big[x_{i,j}(t) - \min(C_{i,j}(t)y_{i,j}(t), x_{i,j}(t)) + \sum_{k\to i} \min(C_{k,i}(t)y_{k,i}(t), x_{k,i}(t))R_{i,j}(t)\Big]_0^S, \tag{5}$$

where $[x]_0^S$ means $\max(\min(x,S),0)$, $C_{i,j}(t)$ (and similarly $C_{k,i}(t)$) is an i.i.d. random variable indicating the random amount of vehicles leaving i for j, and $R_{i,j}(t)$ is an i.i.d. random variable that controls the split of the inflow to link i to the queue $x_{i,j}(t)$. See Varaiya [2013] for the complete details. Given a fixed distribution on the random variables $C_{i,j}(t), R_{i,j}(t)$, (5) provides a complete characterization of the distribution of $s_i(t+1)$ conditioned on $s_{N_i}(t)$ and $a_{N_i}(t)$, in which the local state at each link $s_i(t+1)$ only depends on its neighbors' current states and current actions, which fits into our local interaction structure (equation (1) and Remark 1). The local reward is a characterization of the congestion level at link i, and one version of the reward is the negative queue length

$$r_i(s_i, a_i) = -\sum_{j\in N_i/\{i\}} x_{i,j}. \tag{6}$$

The local state space $\mathcal{S}_i$, the local action space $\mathcal{A}_i$, the local transition probabilities (5) and the local rewards (6) form a networked MDP discussed in Section 2.1. We comment that policies for traffic signal control, like the max pressure policy in Varaiya [2013], typically require knowledge of the statistics of the random variables $C_{ij}(t)$ and $R_{ij}(t)$, while our approach learns from data in a model-free manner.

2.3 Background in RL

To provide background for the analysis in this paper, we review a few key concepts in RL. First, fixing a localized policy tuple $\theta = (\theta_1, \ldots, \theta_n)$, an important notion is the Q-function, which is defined for policy θ as a "table" of values for each state-action pair $(s,a) \in \mathcal{S}\times\mathcal{A}$ and it is the expected infinite horizon discounted reward under policy θ conditioned on the initial state and action being (s,a):

$$\begin{aligned} Q^\theta(s,a) &= \mathbb{E}_{a(t)\sim\zeta^\theta(\cdot|s(t))}\left[\sum_{t=0}^{\infty}\gamma^t r(s(t),a(t))\Big| s(0)=s, a(0)=a\right] \\ &= \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_{a(t)\sim\zeta^\theta(\cdot|s(t))}\left[\sum_{t=0}^{\infty}\gamma^t r_i(s_i(t),a_i(t))\Big| s(0)=s, a(0)=a\right] := \frac{1}{n}\sum_{i=1}^{n} Q_i^\theta(s,a). \end{aligned} \tag{7}$$

In the last step, we have defined $Q_i^\theta(s,a)$ which is the Q function for the individual reward r_i. Both Q^θ and Q_i^θ are exponentially large tables and, therefore, are intractable to compute and store.

Additionally, another important concept we use is the policy gradient theorem, which provides a characterization of the gradient of the objective $J(\theta)$ and is the basis of many algorithmic results in RL. The policy gradient theorem shows that the gradient of $J(\theta)$ depends on Q^θ and, therefore, is intractable to compute using the form in Lemma 1.

Lemma 1 (Sutton et al. [2000]). *Let π^θ be a distribution on the state space given by $\pi^\theta(s) = (1-\gamma)\sum_{t=0}^{\infty}\gamma^t\pi_t^\theta(s)$, where π_t^θ is the distribution of $s(t)$ under a fixed policy θ when $s(0)$ is drawn from π_0. Then, we have,*

$$\nabla J(\theta) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim\pi^\theta, a\sim\zeta^\theta(\cdot|s)}\Big[Q^\theta(s,a)\nabla\log\zeta^\theta(a|s)\Big]. \tag{8}$$

3 Algorithm Design and Results

In this paper we propose an algorithm, Scalable Actor Critic (SAC), which provably finds an $O(\rho^{\kappa+1})$-stationary point of the objective $J(\theta)$ (i.e. a θ s.t. $\|\nabla J(\theta)\|^2 \le \varepsilon$) for some $\rho \le \gamma$, with complexity scaling in the size of the local state-action space of the largest κ-hop neighborhood. We state our main result formally in Theorem 4 after introducing the details of SAC and the key idea underlying its design.

3.1 Key Idea: Exponential Decay of Q-function Leads to Efficient Approximation

Recall that the policy gradient in Lemma 1 is intractable to compute due to the dimension of the Q-function. Our key idea is that exponential decay of the Q function allows efficient approximation of the Q-function via truncation. To illustrate this, we start with the definition of the exponential decay property. Recall that N_i^κ is the set of κ-hop neighborhood of node i and define $N_{-i}^\kappa = \mathcal{N}/N_i^\kappa$, i.e. the set of agents that are outside of i'th κ-hop neighborhood. We write state s as $(s_{N_i^\kappa}, s_{N_{-i}^\kappa})$, i.e. the states of agents that are in the κ-hop neighborhood of i and outside of the κ-hop neighborhood respectively. Similarly, we write a as $(a_{N_i^\kappa}, a_{N_{-i}^\kappa})$. The exponential decay property is then defined as follows.

Definition 1. *The (c,ρ)-exponential decay property holds if, for any localized policy θ, for any $i\in\mathcal{N}$, $s_{N_i^\kappa}\in\mathcal{S}_{N_i^\kappa}$, $s_{N_{-i}^\kappa}, s'_{N_{-i}^\kappa}\in\mathcal{S}_{N_{-i}^\kappa}$, $a_{N_i^\kappa}\in\mathcal{A}_{N_i^\kappa}$, $a_{N_{-i}^\kappa}, a'_{N_{-i}^\kappa}\in\mathcal{A}_{N_{-i}^\kappa}$, Q_i^θ satisfies,*

$$|Q_i^\theta(s_{N_i^\kappa}, s_{N_{-i}^\kappa}, a_{N_i^\kappa}, a_{N_{-i}^\kappa}) - Q_i^\theta(s_{N_i^\kappa}, s'_{N_{-i}^\kappa}, a_{N_i^\kappa}, a'_{N_{-i}^\kappa})| \le c\rho^{\kappa+1}.$$

It may not be immediately clear when the exponential decay property holds. Lemma 2 (a) below highlights that the exponential decay property holds generally with $\rho=\gamma$, without any assumption on the transition probabilities except for the factorization structure (1) and the localized policy structure. Further, many MDPs in practice have ergodicity and fast mixing properties, and Lemma 2 (b) shows that when such fast mixing property holds, the (c,ρ)-exponential decay property holds for some $\rho<\gamma$ depending on the mixing rate. The proof of Lemma 2 is postponed to Appendix A. The condition on mixing rate in Lemma 2 (b) is similar to those used in the literature on the finite time analysis of RL methods, e.g. Zou et al. [2019]. In fact, our condition is weaker than the common mixing rate condition in that we only require the distribution of the local state-action pair $(s_i(t), a_i(t))$ to mix, instead of the full state-action pair $(s(t), a(t))$. We leave it as future work to study such "local" mixing behavior and its relation to the local transition probabilities (1).

Lemma 2. *Assume $\forall i$, r_i is upper bounded by $\bar{r}$. Then the following holds.*

(a) The $(\frac{\bar{r}}{1-\gamma}, \gamma)$-exponential decay property holds.

(b) If there exists $c' > 0$ and $\mu \in (0,1)$ s.t. under any policy θ, the Markov chain is ergodic and starting from any initial state, $\mathrm{TV}(\pi_{t,i}, \pi_{\infty,i}) \le c'\mu^t, \forall t$, where $\pi_{t,i}$ is the distribution of $(s_i(t), a_i(t))$ and $\pi_{\infty,i}$ is the distribution for (s_i, a_i) in stationarity, and recall $\mathrm{TV}(\cdot,\cdot)$ is the total variation distance. Then, the $(\frac{2c'\bar{r}}{1-\gamma\mu}, \gamma\mu)$-exponential decay property holds.

Our definition of exponential decay is similar in spirit to the "correlation decay", or "spatial decay" that has been studied in the literature [Gamarnik, 2013; Gamarnik et al., 2014; Bamieh et al., 2002], though these works consider very different settings. For example, Gamarnik [2013] and Gamarnik et al. [2014] study optimization in a graphical model setting (no concept of state and/or time), and show that the effect of cost functions far away on the optimal solution at a particular node shrinks exponentially in their graph distance, under certain weak interaction assumptions. Compared to these works where the optimization problem is static, we focus on an MDP setting which has states that evolve on a time axis. Further, our exponential decay is in terms of the Q functions, as opposed to the optimal solution. That being said, we believe there are deep connections between our results and that in Gamarnik [2013]; Gamarnik et al. [2014], and we leave the investigation of it as future work.

The power of the exponential decay property is that such properties usually lead to scalable and distributed algorithm design, as in Gamarnik [2013]. In our context, the exponential decay property guarantees that the dependence of Q_i^θ on other agents shrinks quickly as the distance between them grows. This motivates us to consider the following class of truncated Q-functions,

$$\hat{Q}_i^\theta(s_{N_i^\kappa}, a_{N_i^\kappa}) = \sum_{s_{N_{-i}^\kappa} \in \mathcal{S}_{N_{-i}^\kappa}, a_{N_{-i}^\kappa} \in \mathcal{A}_{N_{-i}^\kappa}} w_i(s_{N_{-i}^\kappa}, a_{N_{-i}^\kappa}; s_{N_i^\kappa}, a_{N_i^\kappa}) Q_i^\theta(s_{N_i^\kappa}, s_{N_{-i}^\kappa}, a_{N_i^\kappa}, a_{N_{-i}^\kappa}), \tag{9}$$

where $w_i(s_{N_{-i}^\kappa}, a_{N_{-i}^\kappa}; s_{N_i^\kappa}, a_{N_i^\kappa})$ are *any* non-negative weights satisfying

$$\sum_{s_{N_{-i}^\kappa} \in \mathcal{S}_{N_{-i}^\kappa}, a_{N_{-i}^\kappa} \in \mathcal{A}_{N_{-i}^\kappa}} w_i(s_{N_{-i}^\kappa}, a_{N_{-i}^\kappa}; s_{N_i^\kappa}, a_{N_i^\kappa}) = 1, \qquad \forall (s_{N_i^\kappa}, a_{N_i^\kappa}) \in \mathcal{S}_{N_i^k} \times \mathcal{A}_{N_i^k}. \tag{10}$$

With the definition of the truncated Q-function, our key insight is the following Lemma 3, which says when the exponential decay property holds, the truncated Q-function (9) approximates the full Q-function with high accuracy and can be used to approximate the policy gradient. The proof of Lemma 3 is postponed to Appendix B.

Lemma 3. *Under the (c,ρ)-exponential decay property, the following holds:*

(a) Any truncated Q-function in the form of (9) satisfies,

$$\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} |\hat{Q}_i^\theta(s_{N_i^\kappa}, a_{N_i^\kappa}) - Q_i^\theta(s,a)| \le c\rho^{\kappa+1}.$$

(b) Given i, define the following truncated policy gradient,

$$\hat{h}_i(\theta) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim\pi^\theta, a\sim\zeta^\theta(\cdot|s)}\Big[\frac{1}{n}\nabla_{\theta_i}\log\zeta_i^{\theta_i}(a_i|s_i)\sum_{j\in N_i^\kappa}\hat{Q}_j^\theta(s_{N_j^\kappa}, a_{N_j^\kappa})\Big], \tag{11}$$

where $\hat{Q}_j^\theta$ can be any truncated Q-function in the form of (9). Then, if $\|\nabla_{\theta_i}\log\zeta_i^{\theta_i}(a_i|s_i)\| \le L_i, \forall a_i, s_i$, we have $\|\hat{h}_i(\theta) - \nabla_{\theta_i}J(\theta)\| \le \frac{cL_i}{1-\gamma}\rho^{\kappa+1}$.

Algorithm 1: SAC: Scalable Actor Critic

Input: $\theta_i(0)$; parameter κ; T, length of each episode; step size parameters h, t_0, η.

1 **for** $m = 0, 1, 2, \ldots$ **do**

2 Sample initial state $s(0) \sim \pi_0$, each agent i takes action $a_i(0) \sim \zeta_i^{\theta_i(m)}(\cdot|s_i(0))$, receives reward $r_i(0) = r_i(s_i(0), a_i(0))$.

3 Initialize $\hat{Q}_i^0 \in \mathbb{R}^{\mathcal{S}_{N_i^\kappa} \times \mathcal{A}_{N_i^\kappa}}$ to be the all zero vector.

4 **for** $t = 1$ **to** T **do**

5 Get state $s_i(t)$, take action $a_i(t) \sim \zeta_i^{\theta_i(m)}(\cdot|s_i(t))$, get reward $r_i(t) = r_i(s_i(t), a_i(t))$.

6 Update the truncated Q function with step size $\alpha_{t-1} = \frac{h}{t-1+t_0}$,

7 $\hat{Q}_i^t(s_{N_i^\kappa}(t-1), a_{N_i^\kappa}(t-1)) =$
$(1-\alpha_{t-1})\hat{Q}_i^{t-1}(s_{N_i^\kappa}(t-1), a_{N_i^\kappa}(t-1)) + \alpha_{t-1}(r_i(t-1) + \gamma \hat{Q}_i^{t-1}(s_{N_i^\kappa}(t), a_{N_i^\kappa}(t)))$,

8 $\hat{Q}_i^t(s_{N_i^\kappa}, a_{N_i^\kappa}) = \hat{Q}_i^{t-1}(s_{N_i^\kappa}, a_{N_i^\kappa})$ for $(s_{N_i^\kappa}, a_{N_i^\kappa}) \neq (s_{N_i^\kappa}(t-1), a_{N_i^\kappa}(t-1))$.

9 **end**

10 Each agent i calculates approximated gradient,

11 $\hat{g}_i(m) = \sum_{t=0}^{T} \gamma^t \frac{1}{n} \sum_{j \in N_i^\kappa} \hat{Q}_j^T(s_{N_j^\kappa}(t), a_{N_j^\kappa}(t)) \nabla_{\theta_i} \log \zeta_i^{\theta_i(m)}(a_i(t)|s_i(t))$.

12 Each agent i conducts gradient step $\theta_i(m+1) = \theta_i(m) + \eta_m \hat{g}_i(m)$ with $\eta_m = \frac{\eta}{\sqrt{m+1}}$.

13 **end**

The power of this lemma is that the truncated Q function has a much smaller dimension than the true Q function, and is thus scalable to compute and store. However, despite the reduction in dimension, the error resulting from the approximation is small. In the next section, we use this idea to design a scalable algorithm.

3.2 Algorithm Design: Scalable Actor Critic (SAC)

The good properties of the truncated Q-function open many possibilities for algorithm design. For instance, one can first obtain the truncated Q-function in some way (which could be much easier than directly computing the full Q-function) and then do a policy gradient step using the Lemma 3. In this subsection, we propose one particular approach using the actor critic framework. Our approach, Scalable Actor Critic (SAC), uses temporal difference (TD) learning to obtain the truncated Q-function and then uses policy gradient for policy improvement. The pseudocode of the proposed algorithm is given in Algorithm 1.

Overall structure. The overall structure of SAC is a for-loop from line 1 to line 13. Inside the outer loop, there is an inner loop (line 4 through line 9) that uses temporal difference learning to get the truncated Q-function, which is followed by a policy gradient step that does policy improvement.

The Critic: TD-inner loop. Line 4 through line 9 is the policy evaluation inner loop that obtains the truncated Q function, where line 7 and 8 are the temporal difference update. We note that steps 7 and 8 use the same update equation as TD learning, except that it "pretends" $(s_{N_i^\kappa}, a_{N_i^\kappa})$ is the true state-action pair while the true state-action pair should be (s, a). As will be shown in the theoretic analysis, such a TD update implicitly gives an estimate of a truncated Q function.

The Actor: policy gradient. Line 10 through line 12 define the actor actions. Here, each agent calculates an estimate of the truncated gradient based on (11), and then conducts a gradient step.

Communication. To implement our training algorithm, each agent needs to communicate with other agents in its κ-hop neighborhood in line 7 and line 11; after training is done, each agent implements its localized policy that does not need communication. This communication requirement is weaker than the "centralized training with decentralized execution" paradigm in the multi-agent RL literature [Lowe et al., 2017], where in the training phase, global communication is used. We also comment that when $\kappa = 0$, our algorithm does not need communication and is effectively the same as the independent learner approach in the literature [Tan, 1993; Lowe et al., 2017], as each agent simply runs a single-agent actor critic method based on its local state and local action. When $\kappa > 1$, our algorithm requires communication with agents beyond the direct 1-hop neighbors, which may be unrealistic for some applications. An interesting future direction is to reduce the communication requirements, e.g. potentially using consensus schemes like in Zhang et al. [2018], and also techniques that only communicate quantized bits as opposed to real numbers [Magnússon et al., 2020].

Discussion. Our algorithm serves as an initial concrete demonstration of how to make use of the truncated Q-functions to develop a scalable RL method for networked systems. There are many extensions and other approaches that could be pursued, either within the actor critic framework or beyond. One immediate extension is to do a warm start, i.e., initialize $\hat{Q}_i^0$ as the final estimate $\hat{Q}_i^T$ in the previous outer-loop. Additionally, one can use the TD-λ variant of TD learning, incorporate variance reduction schemes like the advantage function (Advantage Actor Critic), or incorporate function approximation. Further, beyond the actor critic framework, another direction is to develop Q-learning/SARSA type algorithms based on the truncated Q-functions. An appealing aspect of Q-learning/SARSA algorithms is that they may exhibit better convergence properties, but the challenge is that, unlike the actor critic framework, it is not straightforward to enforce the policy to be local in Q-learning/SARSA algorithms. These are interesting topics for future work.

Remark 3 (Model-based vs model-free)**.** *We note that by using an actor critic framework, the proposed approach is model-free, meaning it does not explicitly estimate the transition probabilities and the reward function. This is in contrast to model-based RL, which explicitly estimates the transition probabilities (or parameters that determine the transition probabilities, like the δ_i, $\beta_i(\cdot)$ parameter in the epidemic example in Section 2.2). On one hand, it is known that model-based RL can be more sample efficient than model-free RL in certain circumstances [Tu and Recht, 2019]. Additionally, for specific applications, model-based control design may come with properties like robustness [Varaiya, 2013]. On the other hand, model-free RL offers more flexibility since it does not impose assumptions on the model class. The comparison and tradeoff between model-based and model-free approaches is an open research question and is beyond the scope of this paper. We refer the reader to Tu and Recht [2019]; Qu et al. [2020b] for more details.*

3.3 Approximation Bound

In this section, we state and discuss the formal approximation guarantee for SAC. Before stating the theorem, we first state the assumptions we use. The first assumption is standard in the RL literature and bounds the reward and state/action space size.

Assumption 1 (Bounded reward and state/action space size)**.** *The reward is upper bounded as $0 \leq r_i(s_i, a_i) \leq \bar{r}, \forall i, s_i, a_i$. The individual state and action space size are upper bounded as $|\mathcal{S}_i| \leq S, |\mathcal{A}_i| \leq A, \forall i$.*

Assumption 2 (Exponential decay). *The (c, ρ)-exponential decay property holds for some $\rho \leq \gamma$.*

Note that under Assumption 1, Assumption 2 automatically holds with $\rho = \gamma$, cf. Lemma 2 (a). However, we state the exponential decay property as an assumption to account for the more general case that ρ could be strictly less than γ, cf. Lemma 2 (b).

Our third assumption can be interpreted as an ergodicity condition which ensures that the state-action pairs are sufficiently visited.

Assumption 3 (Sufficient local exploration). *There exists positive integer τ and $\sigma \in (0, 1)$ s.t. under any fixed policy θ and any initial state-action $(s, a) \in \mathcal{S} \times \mathcal{A}$, $\forall i \in \mathcal{N}, \forall (s'_{N_i^\kappa}, a'_{N_i^\kappa}) \in \mathcal{S}_{N_i^\kappa} \times \mathcal{A}_{N_i^\kappa}$, we have $P((s_{N_i^\kappa}(\tau), a_{N_i^\kappa}(\tau)) = (s'_{N_i^\kappa}, a'_{N_i^\kappa}) | (s(1), a(1)) = (s, a)) \geq \sigma$.*

Assumption 3 requires that every state action pair in the κ-hop neighborhood must be visited with some positive probability after some time. This type of assumption is common for finite time convergence results in RL. For example, in Srikant and Ying [2019]; Li et al. [2020], it is assumed that every state-action pair is visited with positive probability in the stationary distribution and the state-action distribution converges to the stationary distribution with some rate. This implies our assumption which is weaker in the sense that we only require local state-action pair $(s_{N_i^\kappa}, a_{N_i^\kappa})$ to be visited as opposed to the full state-action pair (s, a). Having said that, we note that by making Assumption 3, we do not consider the exploration-exploitation tradeoff, which is a challenging issue even in single-agent RL. One potential way to relax Assumption 3 is to use Upper Confidence Bound (UCB) bonuses to encourage exploration, which has been proposed in single-agent RL [Jin et al., 2018]. We leave the study of the exploration-exploitation tradeoff in the multi-agent networked setting as future work.

Finally, we assume boundedness and Lipschitz continuity of the gradients, which is standard in the RL literature.

Assumption 4 (Bounded and Lipschitz continuous gradient). *For any i, a_i, s_i and θ_i, we assume $\|\nabla_{\theta_i} \log \zeta_i^{\theta_i}(a_i|s_i)\| \leq L_i$. As a result, $\|\nabla_\theta \log \zeta^\theta(a|s)\| \leq L = \sqrt{\sum_{i=1}^n L_i^2}$. Further, assume $\nabla J(\theta)$ is L'-Lipschitz continuous in θ.*

With these assumptions in hand, we are ready to state our convergence result.

Theorem 4. *Under Assumption 1, 2, 3 and 4, for any $\delta \in (0, 1)$, $M \geq 3$, suppose the critic step size $\alpha_t = \frac{h}{t+t_0}$ satisfies $h \geq \frac{1}{\sigma} \max(2, \frac{1}{1-\sqrt{\gamma}})$, $t_0 \geq \max(2h, 4\sigma h, \tau)$; and the actor step size satisfies $\eta_m = \frac{\eta}{\sqrt{m+1}}$ with $\eta \leq \frac{1}{4L'}$. Further, if the inner loop length T is large enough s.t. $T + 1 \geq \log_\gamma \frac{c(1-\gamma)}{\bar{r}} + (\kappa + 1) \log_\gamma \rho$ and*

$$\frac{C_a(\frac{\delta}{2nM}, T)}{\sqrt{T + t_0}} + \frac{C'_a}{T + t_0} \leq \frac{2c\rho^{\kappa+1}}{(1-\gamma)^2}, \tag{12}$$

where $C_a(\delta, T) = \frac{6\bar{\epsilon}}{1-\sqrt{\gamma}} \sqrt{\frac{\tau h}{\sigma}[\log(\frac{2\tau T^2}{\delta}) + f(\kappa) \log SA]}$ and $C'_a = \frac{2}{1-\sqrt{\gamma}} \max(\frac{16\bar{\epsilon}h\tau}{\sigma}, \frac{2\bar{r}}{1-\gamma}(\tau + t_0))$, with $\bar{\epsilon} = 4\frac{\bar{r}}{1-\gamma} + 2\bar{r}$ and we recall that $f(\kappa) = \max_i |N_i^\kappa|$ is the size of the largest κ-neighborhood. Then, with probability at least $1 - \delta$,

$$\frac{\sum_{m=0}^{M-1} \eta_m \|\nabla J(\theta(m))\|^2}{\sum_{m=0}^{M-1} \eta_m} \leq \frac{\frac{2\bar{r}}{\eta(1-\gamma)} + \frac{8\bar{r}^2 L^2}{(1-\gamma)^4} \sqrt{\log M \log \frac{4}{\delta}} + \frac{96\bar{r}^2 L' L^2}{(1-\gamma)^4} \eta \log M}{\sqrt{M+1}} + \frac{12L^2 c\bar{r}}{(1-\gamma)^5} \rho^{\kappa+1}. \tag{13}$$

The proof of Theorem 4 is deferred to Section 4. To interpret the result, note that the first term in (13) converges to 0 in the order of $\tilde{O}(\frac{1}{\sqrt{M}})$ and the second term, which we denote as ε_κ, is the bias caused by the truncation of the Q-function and it scales in the order of $O(\rho^{\kappa+1})$. As such, our method SAC will eventually find an $O(\rho^{\kappa+1})$-approximation of a stationary point of the objective function $J(\theta)$, which could be very close to a true stationary point even for small κ as ε_κ decays exponentially in κ.

In terms of complexity, (13) gives that, to reach a $O(\varepsilon_\kappa)$-approximate stationary point, the number of outer-loop iterations required is $M \geq \tilde{\Omega}(\frac{1}{\varepsilon_\kappa^2} poly(\bar{r}, L, L', \frac{1}{(1-\gamma)}))$, which scales polynomially with the parameters of the problem. We emphasize that it does not scale exponentially with n. Further, since the left hand side of (12) decays to 0 as T increases in the order of $\tilde{O}(\frac{1}{\sqrt{T}})$ and the right hand side of (12) is in the same order as $O(\varepsilon_\kappa)$, the inner-loop length required is $T \geq \tilde{\Omega}(\frac{1}{\varepsilon_k^2} poly(\tau, \frac{1}{\sigma}, \frac{1}{1-\gamma}, \bar{r}, f(\kappa)))$. This iteration complexity for the inner loop can potentially be further reduced if we do a warm start for the inner-loop, as the Q-estimate from the previous outer-loop should be already a good estimate for the current outer-loop. We leave the finite time analysis of the warm start variant as future work.

In the complexity bound, a key parameter is σ, which we recall is defined in Assumption 3 and it roughly means the probability that a state-action pair in a κ-hop neighborhood is visited. Suppose we interpret σ to scale with $\sigma \sim \frac{1}{(|S||A|)^{f(\kappa)}}$, where we recall $f(\kappa)$ is the size of the largest κ-hop neighborhood around any node, and $(|S||A|)^{f(\kappa)}$ is the largest state-action space size of κ-hop neighborhoods of any node. Then, the iteration complexity scales with $\frac{1}{\sigma} \sim (|S||A|)^{f(\kappa)}$, whereas the steady state error depends on $\rho^{\kappa+1}$. Therefore, κ is a parameter that balances between complexity and performance – the larger κ is, the higher the complexity but the smaller the steady state error. Exactly how the complexity grows depends on $f(\kappa)$, the size of κ-hop neighborhoods, which in turn depends on the topology of the interaction graph. On one hand, for a sparse graph where $f(\kappa)$ is a constant much smaller than the number of nodes n, the state-action space size of κ-hop neighborhoods is much smaller than the global state-action space size, in which case our algorithm can avoid the exponential scaling in n and is scalable to implement. In the case where the graph is very dense or even complete, we have $f(\kappa) = \Omega(n)$ for any $\kappa > 0$ and our algorithm still suffers from the curse of dimensionality as the complexity scales with $(|S||A|)^{\Omega(n)}$. However, in the case of dense or complete graphs, the local interaction structure becomes degenerate as it takes exponentially many parameters to even specify the local transition probabilities in (1). How to handle the dense or complete graph case remains an interesting future direction, and we believe more structural assumptions are needed to break the curse of dimensionality in that case.

Another thing to note in Theorem 4 is that our guarantee is an upper bound on the running average of the squared norm of the gradient and is essentially a local convergence guarantee. This kind of local convergence is typical for actor critic methods, see e.g. Konda and Tsitsiklis [2000]; Zhang et al. [2018]. Recently, there have been works studying the optimization landscape for policy optimization, showing that in single-agent settings and for certain parameterizations of the policy, the objective of the policy optimization problem $J(\theta)$ may satisfy the gradient dominance property, indicating any stationary point will be a global optimum [Bhandari and Russo, 2019; Agarwal et al., 2021]. One interesting future direction is to show whether similar properties hold in our multi-agent networked setting with the local interaction structure, and whether the local convergence guarantee can imply a global optimality guarantee.

When put in the context of the broader literature in multi-agent RL, our contribution can be

interpreted as follows. In multi-agent RL, the typical way to handle the curse of dimensionality is to use function approximation of the Q function. For example, Zhang et al. [2018] uses linear function to approximate the Q-function. Alternatively, it has also been popular to use neural networks to do the approximation [Lowe et al., 2017]. However, in these approximation methods, the resulting steady state error in the actor critic framework depends on the approximation error, and it is generally unclear how to choose the function approximator that is both computationally tractable and also accurate in representing the true Q function. In the context of these works, our truncated Q-functions can be viewed as a specific way of function approximation that exploits the local interaction structure in our problem, and our method is not only computationally tractable but also has a small approximation error.

4 Proof of Main Result

In this section, we provide the proof of the main result Theorem 4 with some auxiliary derivations postponed to the appendix. As our algorithm is an actor critic algorithm, the proof is divided into two parts: firstly, we provide an analysis of the critic, i.e. TD learning that estimates the truncated Q-function; secondly, we analyze the actor and finish the proof of Theorem 4.

Analysis of the critic. The first part of the analysis concerns the critic, and we show that the critic inner loop converges to an estimate with steady-state error exponentially small in κ. Specifically, recall that within iteration m the inner loop update is

$$\begin{aligned}\hat{Q}_i^t(s_{N_i^\kappa}(t-1), a_{N_i^\kappa}(t-1)) &= (1-\alpha_{t-1})\hat{Q}_i^{t-1}(s_{N_i^\kappa}(t-1), a_{N_i^\kappa}(t-1)) \\ &\quad + \alpha_{t-1}(r_i(s_i(t-1), a_i(t-1)) + \gamma \hat{Q}_i^{t-1}(s_{N_i^\kappa}(t), a_{N_i^\kappa}(t))), \qquad (14a)\\ \hat{Q}_i^t(s_{N_i^\kappa}, a_{N_i^\kappa}) &= \hat{Q}_i^{t-1}(s_{N_i^\kappa}, a_{N_i^\kappa}) \text{ for } (s_{N_i^\kappa}, a_{N_i^\kappa}) \neq (s_{N_i^\kappa}(t-1), a_{N_i^\kappa}(t-1)), \qquad (14b)\end{aligned}$$

where $\hat{Q}_i^0 \in \mathbb{R}^{\mathcal{S}_{N_i^\kappa} \times \mathcal{A}_{N_i^\kappa}}$ is initialized to be all zero, and $\alpha_t = \frac{h}{t+t_0}$ is the step size. We note that when implementing (14) within outer loop iteration m, trajectory $(s(t), a(t))$ is generated by the agents taking a fixed policy $\theta(m)$. Let $Q_i^{\theta(m)} \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ be the true Q-function for reward r_i under this fixed policy $\theta(m)$ as defined in (7). Given the above notation, we prove the following theorem on the critic, which bounds the error between the approximation $\hat{Q}_i^T$ generated by (14) and the true $Q_i^{\theta(m)}$.

Theorem 5. *Assume Assumption 1, 2, 3 are true and suppose t_0, h satisfies, $h \geq \frac{1}{\sigma}\max(2, \frac{1}{1-\sqrt{\gamma}})$ and $t_0 \geq \max(2h, 4\sigma h, \tau)$. Then, inside outer loop iteration m, for each $i \in \mathcal{N}$, with probability at least $1-\delta$, we have the following error bound,*

$$\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left|Q_i^{\theta(m)}(s,a) - \hat{Q}_i^T(s_{N_i^\kappa}, a_{N_i^\kappa})\right| \leq \frac{C_a}{\sqrt{T+t_0}} + \frac{C_a'}{T+t_0} + \frac{2c\rho^{\kappa+1}}{(1-\gamma)^2},$$

where

$$C_a = \frac{6\bar{\epsilon}}{1-\sqrt{\gamma}}\sqrt{\frac{\tau h}{\sigma}[\log(\frac{2\tau T^2}{\delta}) + f(\kappa)\log SA]}, \quad C_a' = \frac{2}{1-\sqrt{\gamma}}\max(\frac{16\bar{\epsilon}h\tau}{\sigma}, \frac{2\bar{r}}{1-\gamma}(\tau+t_0)),$$

with $\bar{\epsilon} = 4\frac{\bar{r}}{1-\gamma} + 2\bar{r}$.

The proof of Theorem 5 is postponed to Section 4.1. The result in Theorem 5 is an upper bound of the infinity-norm error between the truncated Q-function $\hat{Q}_i^T$ obtained by TD learning and the true Q-function. This error bound can be further decomposed into two parts - a transient part that converges to zero in the order of $\tilde{O}(\frac{1}{\sqrt{T}} + \frac{1}{T})$, and a steady state error that is exponentially small in κ. In proving Theorem 5, we use two key techniques. The first is using the exponential decay property (cf. Definition 1, Lemma 2) to show that in the "steady state", the error of the truncated Q-function is bounded by $O(\rho^{\kappa+1})$. This is possible due to our results in Lemma 3 which shows the class of truncated Q-functions are good approximations of the full Q-function. Our second proof technique is that we develop novel finite time analysis tools for TD learning to obtain the finite time error bound. Our proof uses a novel recursive decomposition of the error. Compared to existing work on finite time analysis on TD learning in Srikant and Ying [2019], our result does not need the ergodicity assumption (the weaker Assumption 3 instead), and obtains error bounds in the infinity norm as opposed to the (weighted) Euclidean norm in Srikant and Ying [2019]. This finite time proof technique could be of independent interest.

Analysis of the actor. We view the actor step as a biased stochastic gradient step, with the bias characterized by our result in the critic (Theorem 5). Under this viewpoint, we finish the analysis of the actor and the proof of the main result in Section 4.2.

4.1 Analysis of the Critic: Proof of Theorem 5

Since Theorem 5 is entirely about a particular outer-loop iteration m, inside which the policy is fixed to be $\theta(m)$, to simplify notation we drop the dependence on m and $\theta(m)$ throughout this section. Particularly, we refer to $Q_i^{\theta(m)}$ as Q_i^*, which is the true Q-function for reward r_i under policy $\theta(m)$ (cf. (7)). We also introduce short-hand notation $z = (s, a) \in \mathcal{Z} = \mathcal{S} \times \mathcal{A}$ to represent a particular state action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$. Similarly, we define $z_i = (s_i, a_i) \in \mathcal{Z}_i = \mathcal{S}_i \times \mathcal{A}_i$, and $z_{N_i^\kappa} = (s_{N_i^\kappa}, a_{N_i^\kappa}) \in \mathcal{Z}_{N_i^\kappa} = \mathcal{S}_{N_i^\kappa} \times \mathcal{A}_{N_i^\kappa}$. A vector $v \in \mathbb{R}^{\mathcal{Z}}$ means a vector of dimension $|\mathcal{Z}|$ that is indexed by $z \in \mathcal{Z}$, with its z'th entry denoted by $v(z)$. For example, the full Q-function Q_i^* will be treated as a vector in $\mathbb{R}^{\mathcal{Z}}$ with its z'th entry denoted by $Q_i^*(z)$. Similarly, a vector $v \in \mathbb{R}^{\mathcal{Z}_{N_i^\kappa}}$ means a vector of dimension $|\mathcal{Z}_{N_i^\kappa}|$ indexed by $z_{N_i^\kappa} \in \mathcal{Z}_{N_i^\kappa}$, and its $z_{N_i^\kappa}$'th entry will be denoted by $v(z_{N_i^\kappa})$. For example, the truncated Q-functions $\hat{Q}_i^t$ will be treated as vectors in $\mathbb{R}^{\mathcal{Z}_{N_i^\kappa}}$. Following a similar convention, a matrix $A \in \mathbb{R}^{\mathcal{Z} \times \mathcal{Z}}$ will be a $|\mathcal{Z}|$-by-$|\mathcal{Z}|$ matrix indexed by $(z, z') \in \mathcal{Z} \times \mathcal{Z}$, with its (z, z')'th entry denoted by $A(z, z')$.

Theorem 5 essentially says that the critic iterate $\hat{Q}_i^t$ in (14) will become a good estimate of Q_i^* as t increases. We note that the full Q-function Q_i^* must satisfy the Bellman equation [Bertsekas and Tsitsiklis, 1996],

$$Q_i^* = \mathrm{TD}(Q_i^*) := r_i + \gamma P Q_i^*, \tag{15}$$

where $\mathrm{TD} : \mathbb{R}^{\mathcal{Z}} \to \mathbb{R}^{\mathcal{Z}}$ is the standard Bellman operator for reward r_i and $P \in \mathbb{R}^{\mathcal{Z} \times \mathcal{Z}}$ is the transition probability matrix from $z(t)$ to $z(t+1)$ under policy $\theta(m)$. Note in (15), without causing any confusion, r_i is interpreted as a vector in $\mathbb{R}^{\mathcal{Z}}$ although r_i only depends on z_i.

Our proof is divided into 3 steps. In Step 1, we rewrite (14a) and (14b) in a linear update form (cf. (17)), study its behavior (Lemma 6), and decompose the error into a recursive form (Lemma 7). In Step 2, we bound the noise sequences in the error decomposition (Lemma 10 and Lemma 11). Finally, in Step 3, we use the recursive error decomposition and the bound on the noise sequences to prove Theorem 5.

Step 1: error decomposition. Define $\mathbf{e}_{z_{N_i^\kappa}}$ to be the indicator vector in $\mathbb{R}^{\mathcal{Z}_{N_i^\kappa}}$, i.e. the $z_{N_i^\kappa}$'th entry of $\mathbf{e}_{z_{N_i^\kappa}}$ is 1 and other entries are zero. Then, the critic update equation (14) can be written as,

$$\hat{Q}_i^t = \hat{Q}_i^{t-1} + \alpha_{t-1} \cdot \mathbf{e}_{z_{N_i^\kappa}(t-1)} \cdot [r_i(z_i(t-1)) + \gamma \hat{Q}_i^{t-1}(z_{N_i^\kappa}(t)) - \hat{Q}_i^{t-1}(z_{N_i^\kappa}(t-1))], \tag{16}$$

with $\hat{Q}_i^0$ being the all zero vector in $\mathbb{R}^{\mathcal{Z}_{N_i^\kappa}}$. Notice that $\hat{Q}_i^{t-1}(z_{N_i^\kappa}) = \mathbf{e}_{z_{N_i^\kappa}}^\top \hat{Q}_i^{t-1}$, we can make the following definition (where we omit the dependence on i in notation $A_{z,z'}$, b_z),

$$A_{z,z'} = \mathbf{e}_{z_{N_i^\kappa}} [\gamma \mathbf{e}_{z'_{N_i^\kappa}}^\top - \mathbf{e}_{z_{N_i^\kappa}}^\top] \in \mathbb{R}^{\mathcal{Z}_{N_i^\kappa} \times \mathcal{Z}_{N_i^\kappa}}, \quad b_z = \mathbf{e}_{z_{N_i^\kappa}} r_i(z_i) \in \mathbb{R}^{\mathcal{Z}_{N_i^\kappa}},$$

and rewrite (16) in a linear form

$$\hat{Q}_i^t = \hat{Q}_i^{t-1} + \alpha_{t-1} \big[A_{z(t-1),z(t)} \hat{Q}_i^{t-1} + b_{z(t-1)} \big]. \tag{17}$$

We define the following simplifying notations, $A_{t-1} = A_{z(t-1),z(t)}$, $b_{t-1} = b_{z(t-1)}$. Let $\mathcal{F}_t$ be the σ-algebra generated by $z(0), \ldots, z(t)$. Then, clearly A_{t-1} is $\mathcal{F}_t$-measurable and b_{t-1} is $\mathcal{F}_{t-1}$ measurable. As a result, $\hat{Q}_i^t$ is $\mathcal{F}_t$-measurable. Let $\tau > 0$ to be the integer in Assumption 3. Let $d_{t-1} \in \mathbb{R}^{\mathcal{Z}}$ be the distribution of $z(t-1)$ conditioned on $\mathcal{F}_{t-\tau}$. Further define, $\bar{A}_{t-1} = \mathbb{E} A_{t-1} | \mathcal{F}_{t-\tau}$, $\bar{b}_{t-1} = \mathbb{E} b_{t-1} | \mathcal{F}_{t-\tau}$. It is clear that d_{t-1}, $\bar{A}_{t-1}$, $\bar{b}_{t-1}$ are all $\mathcal{F}_{t-\tau}$ measurable random vectors (matrices). With these notations, (17) can be rewritten as,

$$\begin{aligned}
\hat{Q}_i^t &= \hat{Q}_i^{t-1} + \alpha_{t-1} \big[A_{t-1} \hat{Q}_i^{t-1} + b_{t-1} \big] \\
&= \hat{Q}_i^{t-1} + \alpha_{t-1} \big[\bar{A}_{t-1} \hat{Q}_i^{t-1} + \bar{b}_{t-1} \big] + \alpha_{t-1} [(A_{t-1} - \bar{A}_{t-1}) \hat{Q}_i^{t-1} + b_{t-1} - \bar{b}_{t-1}] \\
&= \hat{Q}_i^{t-1} + \alpha_{t-1} \big[\bar{A}_{t-1} \hat{Q}_i^{t-1} + \bar{b}_{t-1} \big] \\
&\quad + \alpha_{t-1} \underbrace{[(A_{t-1} - \bar{A}_{t-1}) \hat{Q}_i^{t-\tau} + b_{t-1} - \bar{b}_{t-1}]}_{:=\epsilon_{t-1}} + \alpha_{t-1} \underbrace{(A_{t-1} - \bar{A}_{t-1})(\hat{Q}_i^{t-1} - \hat{Q}_i^{t-\tau})}_{:=\phi_{t-1}},
\end{aligned} \tag{18}$$

where in the last step, we have defined sequences $\epsilon_{t-1}, \phi_{t-1} \in \mathbb{R}^{\mathcal{Z}_{N_i^\kappa}}$, which are noise sequences that we will later control in Step 2. For now, we focus on the term $\bar{A}_{t-1} \hat{Q}_i^{t-1} + \bar{b}_{t-1}$, and show the following Lemma 6. The proof of Lemma 6 is similar to the analysis of fixed points for TD learning with linear function approximation, see e.g. Van Roy and Tsitsiklis [1995]. We postpone the detailed proof of Lemma 6 to Appendix C.1.

Lemma 6. *$\forall t$, there exists diagonal matrix $D_{t-1} \in \mathbb{R}^{\mathcal{Z}_{N_i^\kappa} \times \mathcal{Z}_{N_i^\kappa}}$ and operator $g_{t-1} : \mathbb{R}^{\mathcal{Z}_{N_i^\kappa}} \to \mathbb{R}^{\mathcal{Z}_{N_i^\kappa}}$ s.t.*

$$\bar{A}_{t-1} \hat{Q}_i^{t-1} + \bar{b}_{t-1} = -D_{t-1} \hat{Q}_i^{t-1} + D_{t-1} g_{t-1}(\hat{Q}_i^{t-1}), \tag{19}$$

where D_{t-1} satisfies $D_{t-1} \succeq \sigma I$ (recall $\sigma > 0$ is from Assumption 3) with its $z_{N_i^\kappa}$'th diagonal entry being $\bar{d}_{t-1}(z_{N_i^\kappa}) = \sum_{z_{N_{-i}^\kappa} \in \mathcal{Z}_{N_{-i}^\kappa}} d_{t-1}(z_{N_i^\kappa}, z_{N_{-i}^\kappa})$, which means the marginalized distribution of $z_{N_i^\kappa}$ under distribution d_{t-1}. Further, g_{t-1} is a γ-contraction in the infinity norm, with unique fixed point $\hat{Q}_i^{*,t-1}$ satisfying $\sup_{z \in \mathcal{Z}} |\hat{Q}_i^{*,t-1}(z_{N_i^\kappa}) - Q_i^*(z)| \le \frac{c\rho^{\kappa+1}}{1-\gamma}$.

From Lemma 6, we can see the first two terms in (18) can be written as $(I-\alpha_{t-1}D_{t-1})\hat{Q}_i^{t-1}+\alpha_{t-1}D_{t-1}g_{t-1}(\hat{Q}_i^{t-1})$, which roughly speaking drives the iterate to $\hat{Q}_i^{*,t-1}$, the fixed point of operator g_{t-1}. Though depending on t, $\hat{Q}_i^{*,t-1}$ is a good approximation of the true Q-function Q_i^* regardless of t, as shown in Lemma 6. As such, moving towards $\hat{Q}_i^{*,t-1}$ at each time step should eventually produce a good estimate of the full Q-function Q_i^*. In the following, we turn the above intuition into an error decomposition. In details, we unroll (18) and get,

$$\begin{aligned}\hat{Q}_i^t &= (I-\alpha_{t-1}D_{t-1})\hat{Q}_i^{t-1}+\alpha_{t-1}D_{t-1}g_{t-1}(\hat{Q}_i^{t-1})+\alpha_{t-1}\epsilon_{t-1}+\alpha_{t-1}\phi_{t-1}\\ &= \underbrace{\prod_{k=\tau}^{t-1}(I-\alpha_k D_k)}_{\tilde{B}_{\tau-1,t}}\hat{Q}_i^{\tau}+\sum_{k=\tau}^{t-1}\underbrace{\alpha_k D_k\prod_{\ell=k+1}^{t-1}(I-\alpha_\ell D_\ell)}_{B_{k,t}}g_k(\hat{Q}_i^k)+\sum_{k=\tau}^{t-1}\alpha_k\underbrace{\prod_{\ell=k+1}^{t-1}(I-\alpha_\ell D_\ell)}_{\tilde{B}_{k,t}}(\epsilon_k+\phi_k),\end{aligned} \tag{20}$$

where we have used the following short-hand notations $B_{k,t}=\alpha_k D_k\prod_{\ell=k+1}^{t-1}(I-\alpha_\ell D_\ell)$, $\tilde{B}_{k,t}=\prod_{\ell=k+1}^{t-1}(I-\alpha_\ell D_\ell)$. We also define $\beta_{k,t}=\alpha_k\prod_{\ell=k+1}^{t-1}(1-\alpha_\ell\sigma)$, $\tilde{\beta}_{k,t}=\prod_{\ell=k+1}^{t-1}(1-\alpha_\ell\sigma)$. Since every diagonal entry of D_ℓ is lower bounded by σ almost surely (Lemma 6), we have every entry of $B_{k,t}$ is upper bounded by $\beta_{k,t}$ and every entry of $\tilde{B}_{k,t}$ is upper bounded by $\tilde{\beta}_{k,t}$ almost surely. With these short-hand notations, we state the error decomposition below, which is a consequence of (20) and the property of operator $g_k(\cdot)$ and its fixed point $\hat{Q}_i^{*,k}$ in Lemma 6. The proof of Lemma 7 is postponed to Appendix C.2.

Lemma 7. *Let $\xi_t=\sup_{z\in\mathcal{Z}}|\hat{Q}_i^t(z_{N_i^\kappa})-Q_i^*(z)|$. The following recursion holds almost surely,*

$$\xi_t\le\tilde{\beta}_{\tau-1,t}\xi_\tau+\gamma\sup_{z_{N_i^\kappa}\in\mathcal{Z}_{N_i^\kappa}}\sum_{k=\tau}^{t-1}b_{k,t}(z_{N_i^\kappa})\xi_k+\frac{2c\rho^{\kappa+1}}{1-\gamma}+\|\sum_{k=\tau}^{t-1}\alpha_k\tilde{B}_{k,t}\epsilon_k\|_\infty+\|\sum_{k=\tau}^{t-1}\alpha_k\tilde{B}_{k,t}\phi_k\|_\infty,$$

where $b_{k,t}(z_{N_i^\kappa})=\alpha_k\bar{d}_k(z_{N_i^\kappa})\prod_{\ell=k+1}^{t-1}(1-\alpha_\ell\bar{d}_\ell(z_{N_i^\kappa}))$ is the $z_{N_i^\kappa}$'th diagonal entry of $B_{k,t}$

From Lemma 7, it is clear that to bound the error ξ_t, we need to bound $\|\sum_{k=\tau}^{t-1}\alpha_k\tilde{B}_{k,t}\epsilon_k\|_\infty$ and $\|\sum_{k=\tau}^{t-1}\alpha_k\tilde{B}_{k,t}\phi_k\|_\infty$, which is the focus of the next step.

Step 2: bound the ϵ_k and the ϕ_k-sequence. The goal of this step is to bound $\|\sum_{k=\tau}^{t-1}\alpha_k\tilde{B}_{k,t}\epsilon_k\|_\infty$ and $\|\sum_{k=\tau}^{t-1}\alpha_k\tilde{B}_{k,t}\phi_k\|_\infty$. We first show some simple properties of $\hat{Q}_i^t,\epsilon_t$ and ϕ_t in Lemma 8. Since every entry of $\alpha_k\tilde{B}_{k,t}$ is upper bounded by $\beta_{k,t}$, we also show some properties of $\beta_{k,t}$, $\tilde{\beta}_{k,t}$ in Lemma 9. The proofs of the two lemmas are postponed to Appendix C.3.

Lemma 8. *We have almost surely, (a) $\|\hat{Q}_i^t\|_\infty\le\frac{\bar{r}}{1-\gamma}$; (b) $\|\epsilon_t\|_\infty\le\bar{\epsilon}=4\frac{\bar{r}}{1-\gamma}+2\bar{r}$; (c) $\|\phi_t\|_\infty\le 2\bar{\epsilon}\sum_{k=t-\tau+1}^{t-1}\alpha_k$.*

Lemma 9. *If $\alpha_t=\frac{h}{t+t_0}$, where $t_0\ge h>\frac{2}{\sigma}$ and $t_0\ge 4\sigma h$, and $t_0\ge\tau$, then $\beta_{k,t},\tilde{\beta}_{k,t}$ satisfies (a) $\beta_{k,t}\le\frac{h}{k+t_0}\left(\frac{k+1+t_0}{t+t_0}\right)^{\sigma h}$, $\tilde{\beta}_{k,t}\le\left(\frac{k+1+t_0}{t+t_0}\right)^{\sigma h}$; (b) $\sum_{k=1}^{t-1}\beta_{k,t}^2\le\frac{2h}{\sigma}\frac{1}{(t+t_0)}$; (c) $\sum_{k=\tau}^{t-1}\beta_{k,t}\sum_{\ell=k-\tau+1}^{k-1}\alpha_\ell\le\frac{8h\tau}{\sigma}\frac{1}{t+t_0}$.*

We now bound $\|\sum_{k=\tau}^{t-1}\alpha_k\tilde{B}_{k,t}\epsilon_k\|_\infty$. Clearly, ϵ_{t-1} is $\mathcal{F}_t$-measurable, and satisfies

$$\mathbb{E}[\epsilon_{t-1}|\mathcal{F}_{t-\tau}]=\mathbb{E}[(A_{t-1}-\bar{A}_{t-1})\hat{Q}_i^{t-\tau}+b_{t-1}-\bar{b}_{t-1}|\mathcal{F}_{t-\tau}]=0, \tag{21}$$

where the last equality is due to the definition of $\bar{A}_{t-1}$ and $\bar{b}_{t-1}$ and the fact $\hat{Q}_i^{t-\tau}$ is $\mathcal{F}_{t-\tau}$-measurable. Equation (21) shows that ϵ_{t-1} is a "shifted" martingale difference sequence (it is not a standard martingale difference sequence which would require $\mathbb{E}\epsilon_{t-1}|\mathcal{F}_{t-1}=0$). Therefore, $\|\sum_{k=\tau}^{t-1}\alpha_k\tilde{B}_{k,t}\epsilon_k\|_\infty$ can be controlled by Azuma-Hoeffding type inequalities, as shown by Lemma 10. We comment that $\tilde{B}_{k,t}$ is also random and $\tilde{B}_{k,t}\epsilon_k$ is no longer a martingale difference sequence. As a result, to prove Lemma 10 requires more than the direct application of the Azuma-Hoeffding bound. For more details, see the full proof of Lemma 10 in Appendix C.4.

Lemma 10. *We have with probability* $1-\delta$,

$$\Big\|\sum_{k=\tau}^{t-1}\alpha_k\tilde{B}_{k,t}\epsilon_k\Big\|_\infty \le 6\bar{\epsilon}\sqrt{\frac{\tau h}{\sigma(t+t_0)}[\log(\frac{2\tau t}{\delta})+f(\kappa)\log SA]}.$$

Finally we bound sequence $\|\sum_{k=\tau}^{t-1}\alpha_k\tilde{B}_{k,t}\phi_k\|_\infty$ using the fact $\phi_{t-1}=(A_{t-1}-\bar{A}_{t-1})(\hat{Q}_i^{t-1}-\hat{Q}_i^{t-\tau})$ can be bounded by the movement of $\hat{Q}_i^t$ after τ steps (i.e. $\|\hat{Q}_i^{t-1}-\hat{Q}_i^{t-\tau}\|_\infty$), which is small due to the step size selection. The proof of Lemma 11 can also be found in Appendix C.4.

Lemma 11. *We have almost surely,* $\|\sum_{k=\tau}^{t-1}\alpha_k\tilde{B}_{k,t}\phi_k\|_\infty \le \frac{16\bar{\epsilon}h\tau}{\sigma}\frac{1}{t+t_0} := C_\phi\frac{1}{t+t_0}$.

Step 3: bound the critic error. We are now ready to use the error decomposition in Lemma 7 as well as the bounds on the ϵ_k, ϕ_k-sequences in Lemma 10 and Lemma 11 to bound the error of the critic. Recall that Theorem 5 states with probability $1-\delta$,

$$\xi_T \le \frac{C_a}{\sqrt{T+t_0}}+\frac{C_a'}{T+t_0}+\frac{C_0}{1-\gamma}, \tag{22}$$

where $C_0=\frac{2c\rho^{\kappa+1}}{1-\gamma}$, and

$$C_a=\frac{6\bar{\epsilon}}{1-\sqrt{\gamma}}\sqrt{\frac{\tau h}{\sigma}[\log(\frac{2\tau T^2}{\delta})+f(\kappa)\log SA]},\quad C_a'=\frac{2}{1-\sqrt{\gamma}}\max(\frac{16\bar{\epsilon}h\tau}{\sigma},\frac{2\bar{r}}{1-\gamma}(\tau+t_0)).$$

To prove (22), we start by applying Lemma 10 to $t\le T$ with δ replaced by δ/T. Then, using a union bound, we get with probability $1-\delta$, for any $t\le T$, $\Big\|\sum_{k=\tau}^{t-1}\alpha_k\tilde{B}_{k,t}\epsilon_k\Big\|_\infty \le C_\epsilon\frac{1}{\sqrt{t+t_0}}$ with $C_\epsilon=6\bar{\epsilon}\sqrt{\frac{\tau h}{\sigma}[\log(\frac{2\tau T^2}{\delta})+f(\kappa)\log SA]}$. Combining this with Lemma 7 and using Lemma 11, we get with probability $1-\delta$, for all $\tau\le t\le T$,

$$\xi_t \le \tilde{\beta}_{\tau-1,t}\xi_\tau+\gamma\sup_{z_{N_i^\kappa}}\sum_{k=\tau}^{t-1}b_{k,t}(z_{N_i^\kappa})\xi_k+C_\epsilon\frac{1}{\sqrt{t+t_0}}+C_\phi\frac{1}{t+t_0}+C_0. \tag{23}$$

We now condition on (23) is true and use induction to show (22). Eq. (22) is true for $t=\tau$, as $\frac{C_a'}{\tau+t_0}\ge\frac{2}{1-\sqrt{\gamma}}\frac{2\bar{r}}{1-\gamma}>\xi_\tau$, where we have used $|\xi_\tau|\le\|Q_i^*\|_\infty+\|\hat{Q}_i^\tau\|_\infty\le\frac{2\bar{r}}{1-\gamma}$. Then, assume (22) is true for up to $k\le t-1$, we have by (23),

$$\begin{aligned}
\xi_t &\le \tilde{\beta}_{\tau-1,t}\xi_\tau+\gamma\sup_{z_{N_i^\kappa}}\sum_{k=\tau}^{t-1}b_{k,t}(z_{N_i^\kappa})[\frac{C_a}{\sqrt{k+t_0}}+\frac{C_a'}{k+t_0}+\frac{C_0}{1-\gamma}]+C_\epsilon\frac{1}{\sqrt{t+t_0}}+C_\phi\frac{1}{t+t_0}+C_0\\
&\le \tilde{\beta}_{\tau-1,t}\xi_\tau+\gamma C_a\sup_{z_{N_i^\kappa}}\sum_{k=\tau}^{t-1}b_{k,t}(z_{N_i^\kappa})\frac{1}{\sqrt{k+t_0}}+\gamma C_a'\sup_{z_{N_i^\kappa}}\sum_{k=\tau}^{t-1}b_{k,t}(z_{N_i^\kappa})\frac{1}{k+t_0}\\
&\quad+C_\epsilon\frac{1}{\sqrt{t+t_0}}+C_\phi\frac{1}{t+t_0}+\frac{C_0}{1-\gamma}.
\end{aligned}$$

We use the following auxiliary Lemma, whose proof is provided in Section C.5.

Lemma 12. *Recall* $\alpha_k = \frac{h}{k+t_0}$, *and* $b_{k,t}(z_{N_i^\kappa}) = \alpha_k \bar{d}_k(z_{N_i^\kappa}) \prod_{\ell=k+1}^{t-1}(1-\alpha_\ell \bar{d}_\ell(z_{N_i^\kappa}))$, *here* $\bar{d}_k(z_{N_i^\kappa}) \geq \sigma$. *If* $\sigma h(1-\sqrt{\gamma}) \geq 1$, $t_0 \geq 1$, *and* $\alpha_0 \leq \frac{1}{2}$, *then, for any* $z_{N_i^\kappa}$, *and any* $0 < \omega \leq 1$,

$$\sum_{k=\tau}^{t-1} b_{k,t}(z_{N_i^\kappa}) \frac{1}{(k+t_0)^\omega} \leq \frac{1}{\sqrt{\gamma}(t+t_0)^\omega}.$$

With Lemma 12, and using the bound on $\tilde{\beta}_{\tau-1,t}$ in Lemma 9 (a), we have

$$\begin{aligned}
\xi_t &\leq \tilde{\beta}_{\tau-1,t}\xi_\tau + \sqrt{\gamma}C_a \frac{1}{\sqrt{t+t_0}} + \sqrt{\gamma}C_a' \frac{1}{t+t_0} + C_\epsilon \frac{1}{\sqrt{t+t_0}} + C_\phi \frac{1}{t+t_0} + \frac{C_0}{1-\gamma} \\
&\leq \underbrace{\sqrt{\gamma}C_a \frac{1}{\sqrt{t+t_0}} + C_\epsilon \frac{1}{\sqrt{t+t_0}}}_{:=F_t} + \underbrace{\sqrt{\gamma}C_a' \frac{1}{t+t_0} + C_\phi \frac{1}{t+t_0} + \left(\frac{\tau+t_0}{t+t_0}\right)^{\sigma h} \xi_\tau}_{:=F_t'} + \frac{C_0}{1-\gamma}.
\end{aligned}$$

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

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

So, we can require C_a, C_a' to be large enough such that $\frac{C_\epsilon}{C_a} \leq 1-\sqrt{\gamma}$, $\frac{C_\phi}{C_a'} \leq \frac{1-\sqrt{\gamma}}{2}$, and $\frac{\xi_\tau(\tau+t_0)}{C_a'} \leq \frac{1-\sqrt{\gamma}}{2}$. Using $\xi_\tau \leq \frac{2\bar{r}}{1-\gamma}$, one can check our selection of C_a and C_a' satisfies the above three inequalities, and so the induction is finished and the proof of Theorem 5 is concluded. □

4.2 Analysis of the Actor and Proof of Main Result (Theorem 4)

With the error of the critic bounded in Theorem 5, the second part of the proof focuses on the actor, i.e. the policy gradient step. Recall that at iteration m, the policy gradient step is given by $\theta_i(m+1) = \theta_i(m) + \eta_m \hat{g}_i(m)$ with $\eta_m = \frac{\eta}{\sqrt{m+1}}$ and $\hat{g}_i(m)$ is given by

$$\hat{g}_i(m) = \sum_{t=0}^{T} \gamma^t \frac{1}{n} \sum_{j \in N_i^\kappa} \hat{Q}_j^{m,T}(s_{N_j^\kappa}(t), a_{N_j^\kappa}(t)) \nabla_{\theta_i} \log \zeta_i^{\theta_i(m)}(a_i(t)|s_i(t)), \tag{24}$$

where $\hat{Q}_i^{m,T}$ is the final estimate of the Q-function for r_i at the end of the critic loop in iteration m, where we have added an additional superscript m to $\hat{Q}_i^{m,T}$ to indicate its dependence on m; $\{s(t), a(t)\}_{t=0}^T$ is the state-action trajectory with $s(0)$ drawn from π_0 (the initial state distribution defined in the objective function $J(\theta)$, cf. (2)) and the agents taking policy $\theta(m)$. Our goal is to show that $\hat{g}_i(m)$ is approximately the right gradient direction, $\nabla_{\theta_i} J(\theta(m))$, which by Lemma 1 can be written as,

$$\nabla_{\theta_i} J(\theta(m)) = \sum_{t=0}^{\infty} \mathbb{E}_{s \sim \pi_t^{\theta(m)}, a \sim \zeta^{\theta(m)}(\cdot|s)} \left[\gamma^t Q^{\theta(m)}(s,a) \nabla_{\theta_i} \log \zeta^{\theta(m)}(a|s) \right], \tag{25}$$

where $\pi_t^{\theta(m)}$ is the distribution of $s(t)$ under fixed policy $\theta(m)$ when the initial state is drawn from π_0; $Q^{\theta(m)}$ is the true Q function for the global reward r under policy $\theta(m)$, cf. (7).

To bound the difference between $\hat{g}_i(m)$ and the true gradient $\nabla_{\theta_i} J(\theta(m))$, we define the following additional sequences,

$$g_i(m) = \sum_{t=0}^{T} \gamma^t \frac{1}{n} \sum_{j \in N_i^\kappa} Q_j^{\theta(m)}(s(t), a(t)) \nabla_{\theta_i} \log \zeta_i^{\theta_i(m)}(a_i(t)|s_i(t)), \tag{26}$$

$$h_i(m) = \sum_{t=0}^{T} \mathbb{E}_{s \sim \pi_t^{\theta(m)}, a \sim \zeta^{\theta(m)}(\cdot|s)} \left[\gamma^t \frac{1}{n} \sum_{j \in N_i^\kappa} Q_j^{\theta(m)}(s, a) \nabla_{\theta_i} \log \zeta_i^{\theta_i(m)}(a_i|s_i) \right], \tag{27}$$

where $Q_i^{\theta(m)}$ is the true Q function for r_i under policy $\theta(m)$. We also use notation $h(m)$, $g(m)$, $\hat{g}(m)$ to denote the respective $h_i(m), g_i(m), \hat{g}_i(m)$ stacked into a larger vector. The following result is an immediate consequence of Assumption 1 and Assumption 4, whose proof is postponed to Appendix D.1.

Lemma 13. *We have almost surely,* $\forall m \leq M$, $\max(\|\hat{g}(m)\|, \|g(m)\|, \|h(m)\|, \|\nabla J(\theta(m))\|) \leq \frac{\bar{r}L}{(1-\gamma)^2}$.

With these definitions, we decompose the error between the gradient estimator $\hat{g}(m)$ and the true gradient $\nabla J(\theta(m))$ into the following three terms,

$$\hat{g}(m) = \underbrace{\hat{g}(m) - g(m)}_{e^1(m)} + \underbrace{g(m) - h(m)}_{e^2(m)} + \underbrace{h(m) - \nabla J(\theta(m))}_{e^3(m)} + \nabla J(\theta(m)). \tag{28}$$

In the following, we will provide bounds on $e^1(m), e^2(m), e^3(m)$, and then combine these bounds to prove our main result Theorem 4.

Bounds on $e^1(m)$. Notice that the difference between $\hat{g}_i(m)$ and $g_i(m)$ is that the critic estimate $\hat{Q}_j^{m,T}$ is replaced with the true Q-function $Q_j^{\theta(m)}$. By Theorem 5, we have $\hat{Q}_j^{m,T}$ will be very close to $Q_j^{\theta(m)}$ with high probability when T is large enough, based on which we can bound $\|e^1(m)\|$, which is formally provided in Lemma 14. The proof of Lemma 14 is postponed to Appendix D.2.

Lemma 14. *When T is large enough s.t.* $\frac{C_a(\frac{\delta}{2nM}, T)}{\sqrt{T+t_0}} + \frac{C_a'}{T+t_0} \leq \frac{2c\rho^{\kappa+1}}{(1-\gamma)^2}$, *where*

$$C_a(\delta, T) = \frac{6\bar{\epsilon}}{1-\sqrt{\gamma}} \sqrt{\frac{\tau h}{\sigma}[\log(\frac{2\tau T^2}{\delta}) + f(\kappa) \log SA]}, \quad C_a' = \frac{2}{1-\sqrt{\gamma}} \max(\frac{16\bar{\epsilon} h \tau}{\sigma}, \frac{2\bar{r}}{1-\gamma}(\tau + t_0)),$$

with $\bar{\epsilon} = 4\frac{\bar{r}}{1-\gamma} + 2\bar{r}$, *then we have with probability at least* $1 - \frac{\delta}{2}$, $\sup_{0 \leq m \leq M-1} \|e^1(m)\| \leq \frac{4cL\rho^{\kappa+1}}{(1-\gamma)^3}$.

Bounds on $e^2(m)$. Let $\mathcal{G}_m$ be the σ-algebra generated by the trajectories in the first m outer-loop iterations. Then, $\theta(m)$ is $\mathcal{G}_{m-1}$ measurable, and so is $h_i(m)$. Further, by the way that the trajectory $\{(s(t), a(t))\}_{t=0}^T$ is generated, we have $\mathbb{E}[g(m)|\mathcal{G}_{m-1}] = h(m)$. As such, $\eta_m \langle \nabla J(\theta(m)), e^2(m) \rangle$ is a martingale difference sequence w.r.t. $\mathcal{G}_m$, and we have the following bound in Lemma 15 which is a direct consequence of Azuma-Hoeffding bound. The proof of Lemma 15 is postponed to Section D.3.

Lemma 15. *With probability at least* $1 - \delta/2$, *we have*

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

Bounds on $e^3(m)$. The term $e^3(m)$ is the error between $h(m)$ and the true gradient $\nabla J(\theta(m))$, where $h(m)$ is similar to the truncated policy gradient in Lemma 3 and therefore should be close to the true gradient by Lemma 3. We provide Lemma 16 below to bound $\|e^3(m)\|$. Due to technical issues, $h(m)$ is not exactly the same as the truncated policy gradient defined in Lemma 3, but a variant of it instead. Therefore the conclusion of Lemma 3 can not be directly used in the proof of Lemma 16. Nevertheless, the proof of Lemma 16 follows essentially the same arguments as in Lemma 3 and is postponed to Appendix D.4.

Lemma 16. *When* $T+1 \geq \frac{\log\left(\frac{c(1-\gamma)}{\bar{r}}\right)+(\kappa+1)\log\rho}{\log\gamma}$, *we have almost surely,* $\|e^3(m)\| \leq 2\frac{Lc}{(1-\gamma)}\rho^{\kappa+1}$.

Combining the bounds and proof of Theorem 4. With the above bounds on $e^1(m)$, $e^2(m)$ and $e^3(m)$, we are now ready to prove the main result Theorem 4. Since $\nabla J(\theta)$ is L' Lipschitz continuous, we have

$$\begin{aligned} J(\theta(m+1)) &\geq J(\theta(m)) + \langle \nabla J(\theta(m)), \theta(m+1)-\theta(m)\rangle - \frac{L'}{2}\|\theta(m+1)-\theta(m)\|^2 \\ &= J(\theta(m)) + \eta_m \langle \nabla J(\theta(m)), \hat{g}(m)\rangle - \frac{L'\eta_m^2}{2}\|\hat{g}(m)\|^2. \end{aligned} \tag{29}$$

Using the decomposition of $\hat{g}(m)$ in (28), we get,

$$\|\hat{g}(m)\|^2 \leq 4\|e^1(m)\|^2 + 4\|e^2(m)\|^2 + 4\|e^3(m)\|^2 + 4\|\nabla J(\theta(m))\|^2.$$

Further, we can bound $\langle \nabla J(\theta(m)), \hat{g}(m)\rangle$,

$$\begin{aligned} \langle \nabla J(\theta(m)), \hat{g}(m)\rangle &= \|\nabla J(\theta(m))\|^2 + \langle \nabla J(\theta(m)), e^1(m)+e^2(m)+e^3(m)\rangle \\ &\geq \|\nabla J(\theta(m))\|^2 + \langle \nabla J(\theta(m)), e^2(m)\rangle - \|\nabla J(\theta(m))\|(\|e^1(m)\|+\|e^3(m)\|). \end{aligned}$$

Plugging the above bounds on $\|\hat{g}(m)\|^2$ and $\langle \nabla J(\theta(m)), \hat{g}(m)\rangle$ into (29), we have,

$$J(\theta(m+1)) \geq J(\theta(m)) + (\eta_m - 2L'\eta_m^2)\|\nabla J(\theta(m))\|^2 + \eta_m\varepsilon_{m,0} - \eta_m\varepsilon_{m,1} - \eta_m^2\varepsilon_{m,2}, \tag{30}$$

where $\varepsilon_{m,0} = \langle \nabla J(\theta(m)), e^2(m)\rangle$, $\varepsilon_{m,1} = \|\nabla J(\theta(m))\|(\|e^1(m)\|+\|e^3(m)\|)$, $\varepsilon_{m,2} = 2L'(\|e^1(m)\|^2 + \|e^2(m)\|^2 + \|e^3(m)\|^2)$. Doing a telescope sum for (30), we get

$$\begin{aligned} J(\theta(M)) &\geq J(\theta(0)) + \sum_{m=0}^{M-1}(\eta_m - 2L'\eta_m^2)\|\nabla J(\theta(m))\|^2 + \sum_{m=0}^{M-1}\eta_m\varepsilon_{m,0} - \sum_{m=0}^{M-1}\eta_m\varepsilon_{m,1} - \sum_{m=0}^{M-1}\eta_m^2\varepsilon_{m,2} \\ &\geq J(\theta(0)) + \sum_{m=0}^{M-1}\frac{1}{2}\eta_m\|\nabla J(\theta(m))\|^2 + \sum_{m=0}^{M-1}\eta_m\varepsilon_{m,0} - \sum_{m=0}^{M-1}\eta_m\varepsilon_{m,1} - \sum_{m=0}^{M-1}\eta_m^2\varepsilon_{m,2}, \end{aligned} \tag{31}$$

where we have used $\eta_m - 2L'\eta_m^2 = \eta_m(1-2L'\eta_m) \geq \frac{1}{2}\eta_m$, which is true because $\eta_m \leq \eta \leq \frac{1}{4L'}$. After rearranging, we get

$$\sum_{m=0}^{M-1}\frac{1}{2}\eta_m\|\nabla J(\theta(m))\|^2 \leq J(\theta(M)) - J(\theta(0)) - \sum_{m=0}^{M-1}\eta_m\varepsilon_{m,0} + \sum_{m=0}^{M-1}\eta_m\varepsilon_{m,1} + \sum_{m=0}^{M-1}\eta_m^2\varepsilon_{m,2}. \tag{32}$$

We now apply our bounds on $e^1(m), e^2(m), e^3(m)$. By Lemma 15, we have with probability $1-\frac{\delta}{2}$,

$$\Big|\sum_{m=0}^{M-1}\eta_m\varepsilon_{m,0}\Big| \le \frac{2\bar{r}^2L^2}{(1-\gamma)^4}\sqrt{2\sum_{m=0}^{M-1}\eta_m^2\log\frac{4}{\delta}}. \tag{33}$$

By Lemma 14 and Lemma 16, we have with probability $1-\frac{\delta}{2}$,

$$\begin{aligned}\sup_{m\le M-1}\varepsilon_{m,1} &\le \frac{\bar{r}L}{(1-\gamma)^2}\Big(\sup_{m\le M-1}\|e^1(m)\| + \sup_{m\le M-1}\|e^3(m)\|\Big) \le \frac{\bar{r}L}{(1-\gamma)^2}\Big(\frac{4cL\rho^{\kappa+1}}{(1-\gamma)^3} + 2\frac{Lc}{(1-\gamma)}\rho^{\kappa+1}\Big)\\ &\le \frac{6L^2c\bar{r}}{(1-\gamma)^5}\rho^{\kappa+1}.\end{aligned} \tag{34}$$

By Lemma 13, we have almost surely, $\max(\|e^1(m)\|, \|e^2(m)\|, \|e^3(m)\|) \le 2\frac{\bar{r}L}{(1-\gamma)^2}$, and hence,

$$\sup_{m\le M-1}\varepsilon_{m,2} = 2L'(\|e^1(m)\|^2 + \|e^2(m)\|^2 + \|e^3(m)\|^2) \le \frac{24\bar{r}^2L'L^2}{(1-\gamma)^4}. \tag{35}$$

Using a union bound, we have with probability $1-\delta$, all three events (33), (34) and (35) hold, which when combined with (32) implies

$$\begin{aligned}&\frac{\sum_{m=0}^{M-1}\eta_m\|\nabla J(\theta(m))\|^2}{\sum_{m=0}^{M-1}\eta_m}\\ &\le \frac{2(J(\theta(M)) - J(\theta(0))) + 2\Big|\sum_{m=0}^{M-1}\eta_m\varepsilon_{m,0}\Big| + 2\sup_{m\le M-1}\varepsilon_{m,2}\sum_{m=0}^{M-1}\eta_m^2}{\sum_{m=0}^{M-1}\eta_m} + 2\sup_{m\le M-1}\varepsilon_{m,1}\\ &\le \frac{2(J(\theta(M)) - J(\theta(0))) + \frac{4\bar{r}^2L^2}{(1-\gamma)^4}\sqrt{2\sum_{m=0}^{M-1}\eta_m^2\log\frac{4}{\delta}} + \frac{48\bar{r}^2L'L^2}{(1-\gamma)^4}\sum_{m=0}^{M-1}\eta_m^2}{\sum_{m=0}^{M-1}\eta_m} + \frac{12L^2c\bar{r}}{(1-\gamma)^5}\rho^{\kappa+1}.\end{aligned} \tag{36}$$

Since $\eta_m = \frac{\eta}{\sqrt{m+1}}$, we have, $\sum_{m=0}^{M-1}\eta_m > 2\eta(\sqrt{M+1}-1) \ge \eta\sqrt{M+1}$ and $\sum_{m=0}^{M-1}\eta_m^2 < \eta^2(1+\log(M)) < 2\eta^2\log(M)$ (using $M \ge 3$). Further we use the bound $J(\theta(M)) \le \frac{\bar{r}}{1-\gamma}$ and $J(\theta(0)) \ge 0$ almost surely. Combining these results, we get with probability $1-\delta$,

$$\frac{\sum_{m=0}^{M-1}\eta_m\|\nabla J(\theta(m))\|^2}{\sum_{m=0}^{M-1}\eta_m} \le \frac{\frac{2\bar{r}}{\eta(1-\gamma)} + \frac{8\bar{r}^2L^2}{(1-\gamma)^4}\sqrt{\log M\log\frac{4}{\delta}} + \frac{96\bar{r}^2L'L^2}{(1-\gamma)^4}\eta\log M}{\sqrt{M+1}} + \frac{12L^2c\bar{r}}{(1-\gamma)^5}\rho^{\kappa+1}.$$

This concludes the proof of Theorem 4. □

5 Numerical Studies

In this section, we first conduct numerical studies in Section 5.1 using a synthetic example where the optimal solution is known in closed form to verify our theoretic results. Then, in Section 5.2, we demonstrate our approach using the wireless communication example introduced in Section 2.2.

5.1 Synthetic Experiments

We first study a synthetic example where the interaction graph is a line of n nodes $\mathcal{N} = \{1, 2, \ldots, n\}$ with the left most node labeled as 1 and the right most as n. Each node has a binary local state space and local action space $\mathcal{S}_i = \mathcal{A}_i = \{0, 1\}$. The left most node (node 1) has a reward 1 whenever $s_1 = 1$, and all other reward values of node 1 and the rewards of all other nodes are 0. Further, $s_1(t+1) = 1$ with probability 1 when the second node has state $s_2(t) = 1$, and in all other cases $s_1(t+1) = 0$ with probability 1. For the i'th node with $2 \leq i \leq n-1$,

$$P(s_i(t+1) = 1 | s_{i-1}(t), s_i(t), s_{i+1}(t), a_i(t)) = \begin{cases} 1, & \text{if } s_{i+1}(t) = 1, a_i(t) = 1, \\ 0.8, & \text{if } s_{i+1}(t) = 0, a_i(t) = 1, \\ 0, & \text{all other cases.} \end{cases}$$

For the last node $i = n$, $s_n(t+1) = 1$ w.p. 1 when $a_n(t) = 1$, and $s_n(t+1) = 0$ w.p. 1 when $a_n(t) = 0$. The initial states of all nodes are $s_i(0) = 1$.

In this example, the rewards and transitions are designed in a way so that the optimal policy can be determined explicitly. Only agent 1 has a non-zero reward when its state is $s_1 = 1$, and it will stay in state $s_1 = 1$ only when $s_2 = 1$, which happens with high probability when $a_2 = 1$ and $s_3 = 1$, and so on so forth. Therefore, it is clear that the optimal policy is for all nodes to always take action $a_i(t) = 1$ regardless of the local state, in which case the states of all agents will stay as $s_i(t) = 1$, and the resulting optimal global discounted reward is $\frac{1}{n}\frac{1}{1-\gamma}$.

In our experiments, we set $n = 8$, $\gamma = 0.7$. We use the softmax policy for the localized policies, which is a standard policy parameterization that encompasses all stochastic policies from $\mathcal{S}_i$ to $\mathcal{A}_i$ [Sutton et al., 1998]. We run the SAC algorithm with $\kappa = 0$ up to $\kappa = 7$. We plot the global discounted reward throughout the training process in Figure 1, and we also plot the optimality gap in Figure 2 computed as the difference of the optimal discounted reward ($\frac{1}{n}\frac{1}{1-\gamma}$) and the discounted reward achieved by the algorithm with the respective κ values. Figure 1 shows that when κ increases, the performance of the algorithm also increases, though the improvement becomes small when $\kappa > 1$. This is also confirmed by Figure 2, which shows the optimality gap is decaying exponentially when κ increases. This is consistent with our theoretic result in Theorem 4. We note that the exponential decay appears to stop in Figure 2 when $\kappa > 4$, and this is due to the increasing complexity in training for large κ as discussed in Section 3.3, and may also be due to the fact that any further potential improvement may be too small (in the order of 1×10^{-3}) to be noticed because of the noise.

5.2 Multi-Access Wireless Communication

We next study the multi-access wireless communication example discussed in Section 2.2. We consider a grid of users in Figure 3, where each user has access points on the corners of the block it is in. In the experiments, we set the grid size as 6×6, deadline as $d_i = 2$, and all parameters p_i (packet arrival probability for user i) and q_k (success transmission probability for access point y_k) are generated uniformly random from $[0, 1]$. We set $\gamma = 0.7$ and the initial state to be uniformly random, and run the SAC algorithm with $\kappa = 0, 1, 2$ to learn a localized softmax policy, starting from an initial policy where the action is chosen uniformly random. We compare the proposed method with a benchmark based on the localized ALOHA protocol [Roberts, 1975], where each user has a certain probability of sending the earliest packet and otherwise not sending at all. When it sends, it sends the packet to a random access point in its available set, with probability proportional

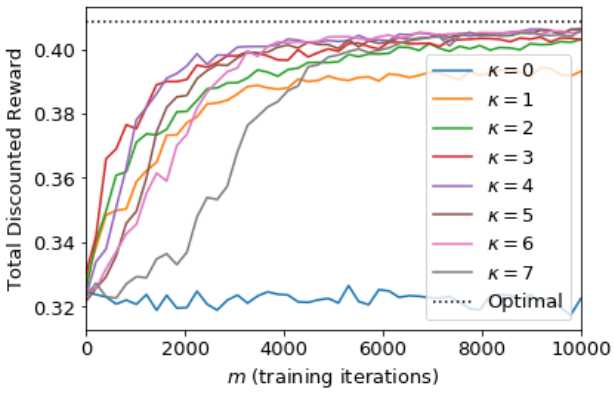

Figure 1: Global discounted reward during the training process in the synthetic experiments.

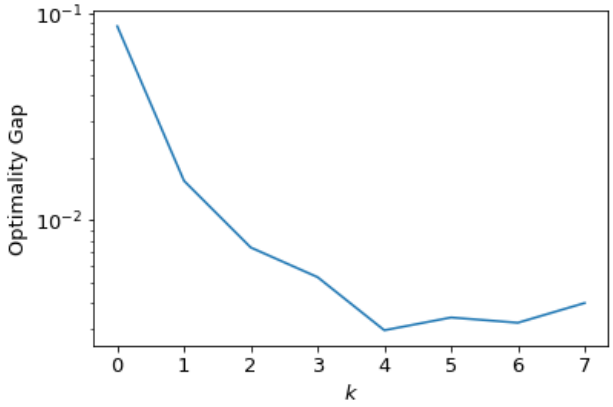

Figure 2: Optimality gap as function of κ in the synthetic experiments.

to the success transmission probability of this access point and inverse proportional to the number of users that share this access point. The results are shown in Figure 4. It shows that the proposed algorithm can outperform the ALOHA based benchmark, despite the proposed algorithm does not have access to the transmission probability q_k which the benchmark has access to. It also shows that the SAC with $\kappa = 1, 2$ outperforms $\kappa = 0$, which as we mentioned in Section 3.2 corresponds to the independent learner approach in the literature [Tan, 1993; Lowe et al., 2017]. SAC with $\kappa = 2$ outperforms $\kappa = 1$, but the improvement is small, which is consistent with the results in the synthetic experiments in Section 5.1. We also study a case with the same 6×6 grid of access points, but the 36 users are assigned randomly to the square blocks in the grid. We plot the results in Figure 5. Similar phenomenons can be observed in Figure 5 as in Figure 4, with SAC outperforms the benchmark and SAC with $\kappa = 1, 2$ outperforms $\kappa = 0$, the independent learner approach.

6 Concluding Remarks and Extensions

This paper proposes a SAC algorithm that provably finds a close-to-stationary point of $J(\theta)$ in time that scales with the local state-action space size of the largest κ-hop neighborhood, which can be much smaller than the full state-action space size when the graph is sparse. This represents the first scalable RL method for localized control of multi-agent networked systems with such a provable guarantee. There are many possible extensions and open questions, which we discuss below.

Average Reward. One extension is to consider average reward instead of discounted reward, i.e., to consider

$$J_{\text{ave}}(\theta) = \lim_{T\to\infty} \frac{1}{T} \mathbb{E}_{s\sim\pi_0} \mathbb{E}_{a(t)\sim\zeta^\theta(\cdot|s(t))} \Big[\sum_{t=0}^{T-1} r(s(t), a(t)) | s(0) = s \Big].$$

Under appropriate ergodicity assumptions, the average reward above is equivalent to the reward under the stationary distribution. Average reward is common in applications where the performance is measured in stationarity, e.g. throughput or waiting time in communication and queueing networks. Despite the importance in applications, average reward RL is known to be more challenging even in single-agent settings, see e.g. Tsitsiklis and Van Roy [1999, 2002]. For example,

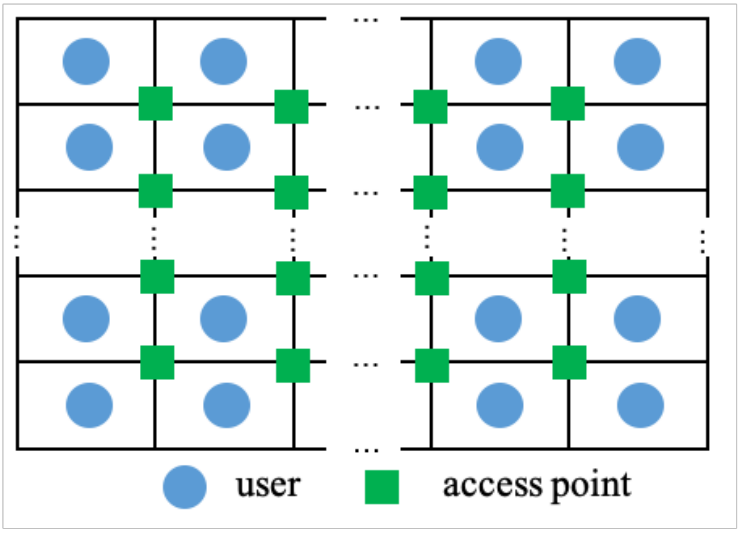

Figure 3: Grid of users and access points.

the Q function needs to be defined in a different way,

$$Q^{\theta}(s,a) = \mathbb{E}_{a(t)\sim\zeta^{\theta}(\cdot|s(t))}\Big[\sum_{t=0}^{\infty}(r(s(t),a(t)) - J_{\text{ave}}(\theta))|s(0)=s, a(0)=a\Big],$$

and because of the lack of a discounting factor γ, the associated Bellman operator no longer has γ as the natural contraction factor. In ongoing work summarized in Qu et al. [2020a], we have begun to study an average reward multi-agent RL setting with the same local interaction structure (1) as in this paper. While similar exponential decay properties on the Q-functions can be defined, in the average reward setting, Qu et al. [2020a] shows that the exponential decay only holds under a form of bounded interaction strength assumption. Under this assumption, Qu et al. [2020a] proposes a variant of the SAC algorithm that achieves similar guarantees as the algorithm in this paper. However, the bounded interaction strength assumption in Qu et al. [2020a] may be restrictive, and searching for weaker assumptions that guarantee the exponential decay (or weaker forms of decay) is an interesting open future direction.

Time-varying dependence structure. In this paper, the interaction graph that defines the dependence structure (1) is a fixed graph, while in many real world applications the graph is time-varying. In other words, if the interaction graph at time t is $\mathcal{G}_t$, then (1) becomes,

$$P(s(t+1)|s(t),a(t)) = \prod_{i=1}^{n} P_i(s_i(t+1)|s_{N_i(\mathcal{G}_t)}(t), a_i(t)),$$

where $N_i(\mathcal{G}_t)$ means the set of neighbors of i in graph $\mathcal{G}_t$. In another of our extensions of this work [Lin et al., 2021], we study a stochastic graph setting where, at each time step, the graph $\mathcal{G}_t$ is sampled from a fixed graph distribution in which each link is present with probability exponentially decreasing in a predefined distance measure between the two nodes. Lin et al. [2021] shows that a weaker form of exponential decay holds in this setting, termed μ-decay, and a similar SAC algorithm can be used. Beyond Lin et al. [2021], an open question is how to handle the case where $\mathcal{G}_t$ can be

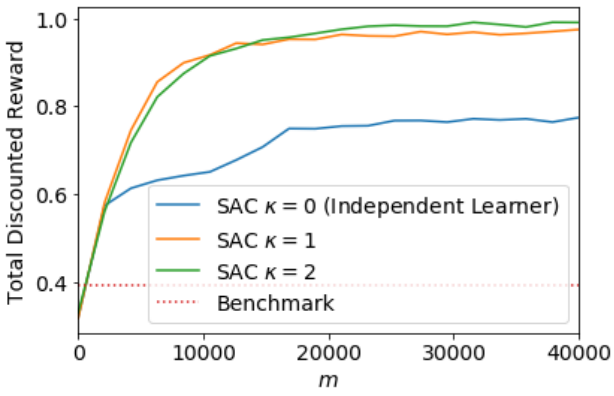

Figure 4: Global discounted reward during training.

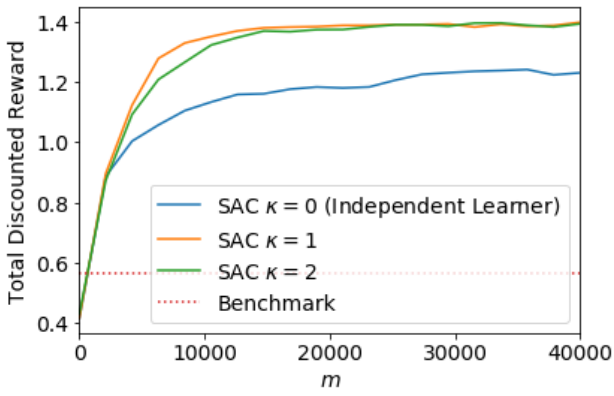

Figure 5: Global discounted reward during training when user locations are assigned randomly.

arbitrarily chosen, rather than stochastically sampled. An issue in this setting is that the transition kernel may not be time-homogeneous anymore, which means that is a challenging open problem.

Other future directions. While SAC is an actor critic based algorithm, the idea underlying SAC, including the exponential decay property and the truncated Q-function (9), is a contribution in its own right, and can potentially lead to other scalable RL methods for networked systems. For example, within the actor critic framework, one idea is to change the critic to variants like TD-λ, change the actor to include the advantage function, or use the simultaneous update version of the actor critic algorithm (as opposed to the inner-loop structure used in this paper). Beyond the actor critic framework, one can develop policy-iteration type algorithms (e.g. Q-learning/SARSA type methods) on our truncated Q-functions. Further, our proof technique in showing the finite time convergence of TD learning can be of independent interest. In fact, we have already considered preliminary applications of our technique in the context of Q-learning and TD learning with state aggregation [Lin et al., 2021], and we expect other applications to emerge in the coming years. Additionally, the setting we consider here does not provide an investigation of the tradeoff between exploration and exploitation. Adding consideration of this tradeoff is an interesting direction for future work. Finally, other future directions include studying the landscape of our policy optimization problem, and studying the robustness of the trained policies.

A The Exponential Decay Property

Our main results depend on the (c, ρ)-exponential decay of the Q-function (cf. Definition 1), which means that for any i, any $s_{N_i^\kappa}$, $s_{N_{-i}^\kappa}$ and $s'_{N_{-i}^\kappa}$, $a_{N_i^\kappa}$, $a_{N_{-i}^\kappa}$ and $a'_{N_{-i}^\kappa}$,

$$|Q_i^\theta(s_{N_i^\kappa}, s_{N_{-i}^\kappa}, a_{N_i^\kappa}, a_{N_{-i}^\kappa}) - Q_i^\theta(s_{N_i^\kappa}, s'_{N_{-i}^\kappa}, a_{N_i^\kappa}, a'_{N_{-i}^\kappa})| \leq c\rho^{\kappa+1}.$$

In Section 3.1, we have pointed out in Lemma 2 that the (c, ρ)-exponential decay property always holds with $\rho \leq \gamma$, assuming the rewards r_i are upper bounded. We now provide the proof of Lemma 2.

Proof of Lemma 2. We first prove part (a). For notational simplicity, denote $s = (s_{N_i^\kappa}, s_{N_{-i}^\kappa})$,

$a = (a_{N_i^\kappa}, a_{N_{-i}^\kappa})$; $s' = (s_{N_i^\kappa}, s'_{N_{-i}^\kappa})$ and $a' = (a_{N_i^\kappa}, a'_{N_{-i}^\kappa})$. Let $\pi_{t,i}$ be the distribution of $(s_i(t), a_i(t))$ conditioned on $(s(0), a(0)) = (s, a)$ under policy θ, and let $\pi'_{t,i}$ be the distribution of $(s_i(t), a_i(t))$ conditioned on $(s(0), a(0)) = (s', a')$ under policy θ. Then, we must have $\pi_{t,i} = \pi'_{t,i}$ for all $t \le \kappa$. The reason is that, due to the local dependence structure (1) and the localized policy structure, $\pi_{t,i}$ only depends on $(s_{N_i^t}, a_{N_i^t})$ (the initial state-action of agent i'th t-hop neighborhood) which is the same as $(s'_{N_i^t}, s'_{N_i^t})$ when $t \le \kappa$ per the way the initial state (s, a), (s', a') are chosen. With these definitions, we expand the definition of Q_i^θ in (7),

$$
\begin{aligned}
&|Q_i^\theta(s,a) - Q_i^\theta(s',a')| \\
&\le \sum_{t=0}^{\infty} \left| \mathbb{E}\big[\gamma^t r_i(s_i(t), a_i(t)) \big| (s(0), a(0)) = (s,a)\big] - \mathbb{E}\big[\gamma^t r_i(s_i(t), a_i(t)) \big| (s(0), a(0)) = (s',a')\big] \right| \\
&= \sum_{t=0}^{\infty} \left| \gamma^t \mathbb{E}_{(s_i,a_i)\sim\pi_{t,i}} r_i(s_i,a_i) - \gamma^t \mathbb{E}_{(s_i,a_i)\sim\pi'_{t,i}} r_i(s_i,a_i) \right| \\
&= \sum_{t=\kappa+1}^{\infty} \left| \gamma^t \mathbb{E}_{(s_i,a_i)\sim\pi_{t,i}} r_i(s_i,a_i) - \gamma^t \mathbb{E}_{(s_i,a_i)\sim\pi'_{t,i}} r_i(s_i,a_i) \right| \\
&\le \sum_{t=\kappa+1}^{\infty} \gamma^t \bar{r} \mathrm{TV}(\pi_{t,i}, \pi'_{t,i}) \le \frac{\bar{r}}{1-\gamma} \gamma^{\kappa+1},
\end{aligned} \tag{37}
$$

where $\mathrm{TV}(\pi_{t,i}, \pi'_{t,i})$ is the total variation distance between $\pi_{t,i}$ and $\pi'_{t,i}$ which is upper bounded by 1. The above inequality shows that the $(\frac{\bar{r}}{1-\gamma}, \gamma)$-exponential decay property holds and concludes the proof of Lemma 2 (a).

The proof of part (b) is almost identical to that of part(a). The only change is that in step (37), we use $\mathrm{TV}(\pi_{t,i}, \pi'_{t,i}) \le 2c'\mu^t$. □

B Proof of Lemma 3

We first show part (a), the truncated Q function is a good approximation of the true Q function. To see that, we have for any $(s, a) \in \mathcal{S} \times \mathcal{A}$, by (9) and (10),

$$
\begin{aligned}
&|\hat{Q}_i^\theta(s_{N_i^\kappa}, a_{N_i^\kappa}) - Q_i^\theta(s,a)| \\
&= \Big| \sum_{s'_{N_{-i}^\kappa}, a'_{N_{-i}^\kappa}} w_i(s'_{N_{-i}^\kappa}, a'_{N_{-i}^\kappa}; s_{N_i^\kappa}, a_{N_i^\kappa}) Q_i^\theta(s_{N_i^\kappa}, s'_{N_{-i}^\kappa}, a_{N_i^\kappa}, a'_{N_{-i}^\kappa}) - Q_i^\theta(s_{N_i^\kappa}, s_{N_{-i}^\kappa}, a_{N_i^\kappa}, a_{N_{-i}^\kappa}) \Big| \\
&\le \sum_{s'_{N_{-i}^\kappa}, a'_{N_{-i}^\kappa}} w_i(s'_{N_{-i}^\kappa}, a'_{N_{-i}^\kappa}; s_{N_i^\kappa}, a_{N_i^\kappa}) \Big| Q_i^\theta(s_{N_i^\kappa}, s'_{N_{-i}^\kappa}, a_{N_i^\kappa}, a'_{N_{-i}^\kappa}) - Q_i^\theta(s_{N_i^\kappa}, s_{N_{-i}^\kappa}, a_{N_i^\kappa}, a_{N_{-i}^\kappa}) \Big| \\
&\le c\rho^{\kappa+1},
\end{aligned} \tag{38}
$$

where in the last step, we have used the (c, ρ)-exponential decay property, cf. Definition 1.

Next, we show part (b). Recall by the policy gradient theorem (Lemma 1),

$$
\begin{aligned}
\nabla_{\theta_i} J(\theta) &= \frac{1}{1-\gamma} \mathbb{E}_{s\sim\pi^\theta, a\sim\zeta^\theta(\cdot|s)} \left[Q^\theta(s,a) \nabla_{\theta_i} \log \zeta^\theta(a|s) \right] \\
&= \frac{1}{1-\gamma} \mathbb{E}_{s\sim\pi^\theta, a\sim\zeta^\theta(\cdot|s)} \left[Q^\theta(s,a) \nabla_{\theta_i} \log \zeta_i^{\theta_i}(a_i|s_i) \right],
\end{aligned}
$$

where we have used $\nabla_{\theta_i} \log \zeta^\theta(a|s) = \nabla_{\theta_i} \sum_{j\in\mathcal{N}} \log \zeta_j^{\theta_j}(a_j|s_j) = \nabla_{\theta_i} \log \zeta_i^{\theta_i}(a_i|s_i)$ by the localized policy structure. With the above equation, we can compute $\hat{h}_i(\theta) - \nabla_{\theta_i} J(\theta)$,

$$
\begin{aligned}
&\hat{h}_i(\theta) - \nabla_{\theta_i} J(\theta) \\
&= \frac{1}{1-\gamma} \mathbb{E}_{s\sim\pi^\theta, a\sim\zeta^\theta(\cdot|s)} \left[\left(\frac{1}{n} \sum_{j\in N_i^\kappa} \hat{Q}_j^\theta(s_{N_j^\kappa}, a_{N_j^\kappa}) - Q^\theta(s,a) \right) \nabla_{\theta_i} \log \zeta_i^{\theta_i}(a_i|s_i) \right] \\
&= \frac{1}{1-\gamma} \mathbb{E}_{s\sim\pi^\theta, a\sim\zeta^\theta(\cdot|s)} \left[\left(\frac{1}{n} \sum_{j\in\mathcal{N}} \hat{Q}_j^\theta(s_{N_j^\kappa}, a_{N_j^\kappa}) - \frac{1}{n} \sum_{j\in\mathcal{N}} Q_j^\theta(s,a) \right) \nabla_{\theta_i} \log \zeta_i^{\theta_i}(a_i|s_i) \right] \\
&\quad - \frac{1}{1-\gamma} \mathbb{E}_{s\sim\pi^\theta, a\sim\zeta^\theta(\cdot|s)} \left[\frac{1}{n} \sum_{j\in N_{-i}^\kappa} \hat{Q}_j^\theta(s_{N_j^\kappa}, a_{N_j^\kappa}) \nabla_{\theta_i} \log \zeta_i^{\theta_i}(a_i|s_i) \right] \\
&:= E_1 - E_2.
\end{aligned}
$$

We claim that $E_2 = 0$. To see this, consider for any $j \in N_{-i}^\kappa$,

$$
\begin{aligned}
&\mathbb{E}_{s\sim\pi^\theta, a\sim\zeta^\theta(\cdot|s)} \left[\nabla_{\theta_i} \log \zeta_i^{\theta_i}(a_i|s_i) \hat{Q}_j^\theta(s_{N_j^\kappa}, a_{N_j^\kappa}) \right] \\
&= \sum_{s,a} \pi^\theta(s) \prod_{\ell=1}^n \zeta_\ell^{\theta_\ell}(a_\ell|s_\ell) \frac{\nabla_{\theta_i} \zeta_i^{\theta_i}(a_i|s_i)}{\zeta_i^{\theta_i}(a_i|s_i)} \hat{Q}_j^\theta(s_{N_j^\kappa}, a_{N_j^\kappa}) \\
&= \sum_{s,a} \pi^\theta(s) \prod_{\ell\neq i} \zeta_\ell^{\theta_\ell}(a_\ell|s_\ell) \nabla_{\theta_i} \zeta_i^{\theta_i}(a_i|s_i) \hat{Q}_j^\theta(s_{N_j^\kappa}, a_{N_j^\kappa}) \\
&= \sum_{s,a_1,\ldots,a_{i-1},a_{i+1},\ldots,a_n} \pi^\theta(s) \prod_{\ell\neq i} \zeta_\ell^{\theta_\ell}(a_\ell|s_\ell) \hat{Q}_j^\theta(s_{N_j^\kappa}, a_{N_j^\kappa}) \sum_{a_i} \nabla_{\theta_i} \zeta_i^{\theta_i}(a_i|s_i) \\
&= 0,
\end{aligned}
$$

where in the last equality, we have used $\hat{Q}_j^\theta(s_{N_j^\kappa}, a_{N_j^\kappa})$ does not depend on a_i as $i \notin N_j^\kappa$; and $\sum_{a_i} \nabla_{\theta_i} \zeta_i^{\theta_i}(a_i|s_i) = \nabla_{\theta_i} \sum_{a_i} \zeta_i^{\theta_i}(a_i|s_i) = \nabla_{\theta_i} 1 = 0$. Now that we have shown $E_2 = 0$, we can bound E_1 as follows

$$
\begin{aligned}
&\|\hat{h}_i(\theta) - \nabla_{\theta_i} J(\theta)\| = \|E_1\| \\
&\leq \frac{1}{1-\gamma} \mathbb{E}_{s\sim\pi^\theta, a\sim\zeta^\theta(\cdot|s)} \left[\frac{1}{n} \sum_{j\in\mathcal{N}} \left| \hat{Q}_j^\theta(s_{N_j^\kappa}, a_{N_j^\kappa}) - Q_j^\theta(s,a) \right| \|\nabla_{\theta_i} \log \zeta_i^{\theta_i}(a_i|s_i)\| \right] \\
&\leq \frac{1}{1-\gamma} c\rho^{\kappa+1} L_i,
\end{aligned}
$$

where in the last step, we have used (38) and the upper bound $\|\nabla_{\theta_i} \log \zeta_i^{\theta_i}(a_i|s_i)\| \leq L_i$. This concludes the proof of Lemma 3. □

C Proof of Auxiliary Results for the Analysis of the Critic

C.1 Proof of Lemma 6

The goal of Lemma 6 is to understand $\bar{A}_{t-1}\hat{Q}_i^{t-1} + \bar{b}_{t-1}$. Recall that d_{t-1} is the distribution of $z(t-1)$ conditioned on $\mathcal{F}_{t-\tau}$ and $\bar{A}_{t-1} = \mathbb{E}[A_{t-1}|\mathcal{F}_{t-\tau}] = \mathbb{E}[A_{z(t-1),z(t)}|\mathcal{F}_{t-\tau}]$, $\bar{b}_{t-1} = \mathbb{E}[b_{t-1}|\mathcal{F}_{t-\tau}] = \mathbb{E}[b_{z(t-1)}|\mathcal{F}_{t-\tau}]$.

Recall that P is the transition matrix from $z(t-1)$ to $z(t)$. Given any distribution d on the state-action space $\mathcal{Z}$, we define $\tilde{A}_z = \mathbb{E}_{z'\sim P(\cdot|z)}A_{z,z'}$ and $\bar{A}^d = \mathbb{E}_{z\sim d}\tilde{A}_z$, $\bar{b}^d = \mathbb{E}_{z\sim d}b_z$. Then, $\bar{A}_{t-1}$ and $\bar{b}_{t-1}$ can be rewritten as $\bar{A}_{t-1} = \bar{A}^{d_{t-1}}$, $\bar{b}_{t-1} = \bar{b}^{d_{t-1}}$. In what follows, we provide characterizations of $\bar{A}^d$, $\bar{b}^d$ for general distributions d.

Firstly, $\tilde{A}_z$ is given by,

$$\tilde{A}_z = \mathbb{E}_{z'\sim P(\cdot|z)}A_{z,z'} = \mathbf{e}_{z_{N_i^\kappa}}[\gamma P(\cdot|z)\Phi - \mathbf{e}_{z_{N_i^\kappa}}^T], \tag{39}$$

where $P(\cdot|z)$ is understood as the z'th row of P and is treated as a row vector. Also, we have defined $\Phi \in \mathbb{R}^{\mathcal{Z}\times\mathcal{Z}_{N_i^\kappa}}$ to be a matrix with each row indexed by $z \in \mathcal{Z}$ and each column indexed by $z'_{N_i^\kappa} \in \mathcal{Z}_{N_i^\kappa}$. Further, the z'th row of Φ is the indicator vector $\mathbf{e}_{z_{N_i^\kappa}}^\top$, in other words $\Phi(z, z'_{N_i^\kappa}) = 1$ if $z'_{N_i^\kappa} = z_{N_i^\kappa}$ and $\Phi(z, z'_{N_i^\kappa}) = 0$ elsewhere.

Then, $\bar{A}^d$, $\bar{b}^d$ are given by,

$$\bar{A}^d = \mathbb{E}_{z\sim d}\tilde{A}_z = \sum_{z\in\mathcal{Z}} d(z)\mathbf{e}_{z_{N_i^\kappa}}[\gamma P(\cdot|z)\Phi - \mathbf{e}_{z_{N_i^\kappa}}^\top] = \Phi^\top \mathrm{diag}(d)\big[\gamma P\Phi - \Phi\big], \tag{40}$$

$$\bar{b}^d = \mathbb{E}_{z\sim d}b_z = \Phi^\top \mathrm{diag}(d) r_i, \tag{41}$$

where $\mathrm{diag}(d) \in \mathbb{R}^{\mathcal{Z}\times\mathcal{Z}}$ is a diagonal matrix with the z'th diagonal entry being $d(z)$; in the last equation, r_i is understood as a vector over the entire state-action space $\mathcal{Z}$, though it only depends on z_i.

With the above characterizations, we show the following property of $\bar{A}^d$ and $\bar{b}^d$ in Lemma 17. Lemma 17 (with d set as d_{t-1}) will directly lead to the results in Lemma 6, with D_{t-1} being the D in Lemma 17 and it satisfies $D_{t-1} \succeq \sigma I$ due to Assumption 3; g_{t-1} and $\hat{Q}_i^{*,t-1}$ being $g^{d_{t-1}}$ and $\hat{Q}_i^{d_{t-1}}$ in Lemma 17.

Lemma 17. *Given distribution d on state-action pair z whose marginalization onto $z_{N_i^\kappa}$ is non-zero for every $z_{N_i^\kappa}$, we have $\bar{A}^d\hat{Q}_i + \bar{b}^d$ can be written as*

$$\bar{A}^d\hat{Q}_i + \bar{b}^d = -D\hat{Q}_i + Dg^d(\hat{Q}_i),$$

where $D = \Phi^\top\mathrm{diag}(d)\Phi \in \mathbb{R}^{\mathcal{Z}_{N_i^\kappa}\times\mathcal{Z}_{N_i^\kappa}}$ is a diagonal matrix, with the $z_{N_i^\kappa}$'th entry being the marginalized distribution of $z_{N_i^\kappa}$ under distribution d; $g^d(\cdot)$ is given by $g^d(\hat{Q}_i) = \Pi^d\mathrm{TD}\Phi\hat{Q}_i$, where $\Pi^d = (\Phi^\top\mathrm{diag}(d)\Phi)^{-1}\Phi^\top\mathrm{diag}(d)$ and $\mathrm{TD}(Q_i) = r_i + \gamma PQ_i$ is the Bellman operator in (15).

Further, $g^d(\cdot)$ is γ contractive in infinity norm, and has a unique fixed point $\hat{Q}_i^d \in \mathbb{R}^{\mathcal{Z}_{N_i^\kappa}}$ depending on d, and the fixed point satisfies

$$\sup_{z\in\mathcal{Z}}|\hat{Q}_i^d(z_{N_i^\kappa}) - Q_i^*(z)| = \|\Phi\hat{Q}_i^d - Q_i^*\|_\infty \le \frac{c\rho^{\kappa+1}}{1-\gamma}. \tag{42}$$

Proof of Lemma 17. It is easy to check that $D = \Phi^\top \mathrm{diag}(d)\Phi \in \mathbb{R}^{\mathcal{Z}_{N_i^\kappa} \times \mathcal{Z}_{N_i^\kappa}}$ is a diagonal matrix, and the $z_{N_i^\kappa}$'th diagonal entry is the marginal probability of $z_{N_i^\kappa}$ under d, which is non-zero by the assumption of the lemma. Therefore, $\Phi^\top \mathrm{diag}(d)\Phi \in \mathbb{R}^{\mathcal{Z}_{N_i^\kappa} \times \mathcal{Z}_{N_i^\kappa}}$ is invertable and matrix $\Pi^d = (\Phi^\top \mathrm{diag}(d)\Phi)^{-1}\Phi^\top \mathrm{diag}(d)$ is well defined. Further, the $z_{N_i^\kappa}$'th row of Π^d is in fact the conditional distribution of the full state z given $z_{N_i^\kappa}$. So, Π^d must be a stochastic matrix and is non-expansive in infinity norm.

By the definition of $\bar{A}^d$ and $\bar{b}^d$, we have,

$$\begin{aligned}
\bar{A}^d\hat{Q}_i + \bar{b}^d &= \Phi^\top \mathrm{diag}(d)\big[\gamma P\Phi - \Phi\big]\hat{Q}_i + \Phi^\top \mathrm{diag}(d) r_i \\
&= \Phi^\top \mathrm{diag}(d)[r_i + \gamma P\Phi\hat{Q}_i] - \Phi^\top \mathrm{diag}(d)\Phi\hat{Q}_i \\
&= \Phi^\top \mathrm{diag}(d)\mathrm{TD}(\Phi\hat{Q}_i) - \Phi^\top \mathrm{diag}(d)\Phi\hat{Q}_i \\
&= -D\hat{Q}_i + D\Pi^d \mathrm{TD}(\Phi\hat{Q}_i) \\
&= -D\hat{Q}_i + Dg^d(\hat{Q}_i),
\end{aligned}$$

where TD is the Bellman operator for reward r_i defined in (15), and operator g^d is given by $g^d(\hat{Q}_i) = \Pi^d \mathrm{TD}\Phi\hat{Q}_i$.

Notice that Φ is non-expansive in $\|\cdot\|_\infty$ norm since each row of Φ has precisely one entry being 1 and all others are zero. Also since Π^d is non-expansive in $\|\cdot\|_\infty$ norm and TD is a γ-contraction in $\|\cdot\|_\infty$ norm, we have $g^d = \Pi^d \mathrm{TD}\Phi$ is a γ contraction in $\|\cdot\|_\infty$ norm. As a result, g^d has a unique fixed point $\hat{Q}_i^d$.

Finally, we show (42), which bounds the distance between $\Phi\hat{Q}_i^d$ and Q_i^*, where Q_i^* is the true Q-function for reward r_i and it is the unique fixed point of TD operator (15). We have,

$$\begin{aligned}
\|\Phi\hat{Q}_i^d - Q_i^*\|_\infty &\le \|\Phi\hat{Q}_i^d - \Phi\Pi^d Q_i^*\|_\infty + \|\Phi\Pi^d Q_i^* - Q_i^*\|_\infty \\
&= \|\Phi\Pi^d \mathrm{TD}(\Phi\hat{Q}_i^d) - \Phi\Pi^d \mathrm{TD}(Q_i^*)\|_\infty + \|\Phi\Pi^d Q_i^* - Q_i^*\|_\infty \\
&\le \gamma\|\Phi\hat{Q}_i^d - Q_i^*\|_\infty + \|\Phi\Pi^d Q_i^* - Q_i^*\|_\infty,
\end{aligned}$$

where the equality follows from the fact that $\hat{Q}_i^d$ is the fixed point of $\Pi^d\mathrm{TD}\Phi$, Q_i^* is the fixed point of TD; the last inequality is due to $\Phi\Pi^d\mathrm{TD}$ is a γ contration in infinity norm. Therefore,

$$\|\Phi\hat{Q}_i^d - Q_i^*\|_\infty \le \frac{1}{1-\gamma}\|\Phi\Pi^d Q_i^* - Q_i^*\|_\infty. \tag{43}$$

Next, recall that the $z_{N_i^\kappa}$'s row of Π^d is the distribution of the state-action pair z conditioned on its N_i^κ coordinates being fixed to be $z_{N_i^\kappa}$. We denote this conditional distribution of the states outside of N_i^κ, $z_{N_{-i}^\kappa}$, given $z_{N_i^\kappa}$, as $d(z_{N_{-i}^\kappa}|z_{N_i^\kappa})$. With this notation,

$$(\Pi^d Q_i^*)(z_{N_i^\kappa}) = \sum_{z_{N_{-i}^\kappa}} d(z_{N_{-i}^\kappa}|z_{N_i^k}) Q_i^*(z_{N_i^\kappa}, z_{N_{-i}^\kappa}).$$

And therefore,

$$(\Phi\Pi^d Q_i^*)(z_{N_i^\kappa}, z_{N_{-i}^\kappa}) = \sum_{z'_{N_{-i}^\kappa}} d(z'_{N_{-i}^\kappa}|z_{N_i^\kappa}) Q_i^*(z_{N_i^\kappa}, z'_{N_{-i}^\kappa}).$$

Further, we have

$$
\begin{aligned}
&|(\Phi\Pi^d Q_i^*)(z_{N_i^\kappa}, z_{N_{-i}^\kappa}) - Q_i^*(z_{N_i^\kappa}, z_{N_{-i}^\kappa})| \\
&= \Big| \sum_{z'_{N_{-i}^\kappa}} d(z'_{N_{-i}^\kappa}|z_{N_i^\kappa}) Q_i^*(z_{N_i^\kappa}, z'_{N_{-i}^\kappa}) - \sum_{z'_{N_{-i}^\kappa}} d(z'_{N_{-i}^\kappa}|z_{N_i^\kappa}) Q_i^*(z_{N_i^\kappa}, z_{N_{-i}^\kappa}) \Big| \\
&\leq \sum_{z'_{N_{-i}^\kappa}} d(z'_{N_{-i}^\kappa}|z_{N_i^\kappa}) \big| Q_i^*(z_{N_i^\kappa}, z'_{N_{-i}^\kappa}) - Q_i^*(z_{N_i^\kappa}, z_{N_{-i}^\kappa}) \big| \\
&\leq c\rho^{\kappa+1},
\end{aligned}
$$

where the last inequality is due to the exponential decay property (cf. Definition 1 and Assumption 2). Therefore,

$$\|\Phi\Pi^d Q_i^* - Q_i^*\|_\infty \leq c\rho^{\kappa+1}.$$

Combining the above with (43), we get the desired result

$$\|\Phi\hat{Q}_i^d - Q_i^*\|_\infty \leq \frac{c\rho^{\kappa+1}}{1-\gamma}.$$

□

C.2 Proof of Lemma 7

Recall that the $z_{N_i^\kappa}$'th diagonal entry of $B_{k,t}$ is $b_{k,t}(z_{N_i^\kappa})$, and we define similarly the $z_{N_i^\kappa}$'th diagonal entry of $\tilde{B}_{k,t}$ to be $\tilde{b}_{k,t}(z_{N_i^\kappa})$. Using these notations, equation (20) can be written as,

$$
\begin{aligned}
\hat{Q}_i^t(z_{N_i^\kappa}) = \overbrace{\tilde{b}_{\tau-1,t}(z_{N_i^\kappa})\hat{Q}_i^\tau(z_{N_i^\kappa}) + \sum_{k=\tau}^{t-1} b_{k,t}(z_{N_i^\kappa})[g_k(\hat{Q}_i^k)](z_{N_i^\kappa})}^{:=G(z_{N_i^\kappa})} \\
+ \sum_{k=\tau}^{t-1} \alpha_k \tilde{b}_{k,t}(z_{N_i^\kappa})(\epsilon_k(z_{N_i^\kappa}) + \phi_k(z_{N_i^\kappa})).
\end{aligned} \tag{44}
$$

Notice that by definition, $\tilde{b}_{\tau-1,t}(z_{N_i^\kappa}) + \sum_{k=\tau}^{t-1} b_{k,t}(z_{N_i^\kappa}) = 1$. Then,

$$
\begin{aligned}
|G(z_{N_i^\kappa}) - Q_i^*(z)| &\le \tilde{b}_{\tau-1,t}(z_{N_i^\kappa})|\hat{Q}_i^\tau(z_{N_i^\kappa}) - Q_i^*(z)| + \sum_{k=\tau}^{t-1} b_{k,t}(z_{N_i^\kappa})|[g_k(\hat{Q}_i^k)](z_{N_i^\kappa}) - Q_i^*(z)| \\
&\le \tilde{b}_{\tau-1,t}(z_{N_i^\kappa})|\hat{Q}_i^\tau(z_{N_i^\kappa}) - Q_i^*(z)| + \sum_{k=\tau}^{t-1} b_{k,t}(z_{N_i^\kappa})|[g_k(\hat{Q}_i^k)](z_{N_i^\kappa}) - \hat{Q}_i^{*,k}(z_{N_i^\kappa})| \\
&\quad + \sum_{k=\tau}^{t-1} b_{k,t}(z_{N_i^\kappa})|Q_i^*(z) - \hat{Q}_i^{*,k}(z_{N_i^\kappa})| \\
&\le \tilde{b}_{\tau-1,t}(z_{N_i^\kappa})|\hat{Q}_i^\tau(z_{N_i^\kappa}) - Q_i^*(z)| + \gamma \sum_{k=\tau}^{t-1} b_{k,t}(z_{N_i^\kappa})\|\hat{Q}_i^k - \hat{Q}_i^{*,k}\|_\infty \\
&\quad + \sum_{k=\tau}^{t-1} b_{k,t}(z_{N_i^\kappa})\|Q_i^* - \Phi\hat{Q}_i^{*,k}\|_\infty \\
&\le \tilde{b}_{\tau-1,t}(z_{N_i^\kappa})|\hat{Q}_i^\tau(z_{N_i^\kappa}) - Q_i^*(z)| + \gamma \sum_{k=\tau}^{t-1} b_{k,t}(z_{N_i^\kappa})\|\Phi\hat{Q}_i^k - Q_i^*\|_\infty \\
&\quad + 2\sum_{k=\tau}^{t-1} b_{k,t}(z_{N_i^\kappa})\|Q_i^* - \Phi\hat{Q}_i^{*,k}\|_\infty \\
&\le \tilde{\beta}_{\tau-1,t}\xi_\tau + \gamma \sum_{k=\tau}^{t-1} b_{k,t}(z_{N_i^\kappa})\xi_k + \frac{2c\rho^{\kappa+1}}{1-\gamma}, \qquad (45)
\end{aligned}
$$

where in the thrid inequality, we have used that g_k is γ-contraction in infinity norm with fixed point $\hat{Q}_i^{*,k}$, and in the last inequality, we have used the property of $\hat{Q}_i^{*,k}$ in Lemma 6. Combining the above with (44), we have

$$
\begin{aligned}
\xi_t &= \|\Phi\hat{Q}_i^t - Q_i^*\|_\infty \\
&\le \tilde{\beta}_{\tau-1,t}\xi_\tau + \gamma \sup_{z_{N_i^\kappa}} \sum_{k=\tau}^{t-1} b_{k,t}(z_{N_i^\kappa})\xi_k + \frac{2c\rho^{\kappa+1}}{1-\gamma} + \|\sum_{k=\tau}^{t-1} \alpha_k \tilde{B}_{k,t}\epsilon_k\|_\infty + \|\sum_{k=\tau}^{t-1} \alpha_k \tilde{B}_{k,t}\phi_k\|_\infty.
\end{aligned}
$$

□

C.3 Proof of Lemma 8 and Lemma 9

In this section, we provide proofs of the two auxiliary lemmas, Lemma 8 and Lemma 9. We start with the proof of Lemma 8.

Proof of Lemma 8. First, notice that $A_{z,z'} = \mathbf{e}_{z_{N_i^\kappa}}[\gamma \mathbf{e}_{z'_{N_i^\kappa}}^T - \mathbf{e}_{z_{N_i^\kappa}}^T]$ and $b_z = \mathbf{e}_{z_{N_i^\kappa}} r_i(z_i)$. As such, $\|A_{z,z'}\|_\infty \le 1 + \gamma < 2$, $\|b_z\|_\infty \le \bar{r}$.

Part (a) can be proved by induction. Part (a) is true for $t = 0$ as $\hat{Q}_i^0 = 0$. Assume $\|\hat{Q}_i^{t-1}\|_\infty \le \frac{\bar{r}}{1-\gamma}$. Recall the update equation (16),

$$
\hat{Q}_i^t = \hat{Q}_i^{t-1} + \alpha_{t-1}[r_i(z_i(t-1)) + \gamma \hat{Q}_i^{t-1}(z_{N_i^\kappa}(t)) - \hat{Q}_i^{t-1}(z_{N_i^\kappa}(t-1))]\mathbf{e}_{z_{N_i^\kappa}(t-1)},
$$

or in other words,

$$\begin{aligned}\hat{Q}_i^t(z_{N_i^\kappa}(t-1)) &= \hat{Q}_i^{t-1}(z_{N_i^\kappa}(t-1)) + \alpha_{t-1}[r_i(z_i(t-1)) + \gamma \hat{Q}_i^{t-1}(z_{N_i^\kappa}(t)) - \hat{Q}_i^{t-1}(z_{N_i^\kappa}(t-1))] \\ &= (1-\alpha_{t-1})\hat{Q}_i^{t-1}(z_{N_i^\kappa}(t-1)) + \alpha_{t-1}[r_i(z_i(t-1)) + \gamma \hat{Q}_i^{t-1}(z_{N_i^\kappa}(t))].\end{aligned}$$

And for other entries of $\hat{Q}_i^t$, it stays the same as $\hat{Q}_i^{t-1}$. For this reason,

$$\|\hat{Q}_i^t\|_\infty \le \max(\|\hat{Q}_i^{t-1}\|_\infty, |\hat{Q}_i^t(z_{N_i^\kappa}(t-1))|).$$

Notice that

$$|\hat{Q}_i^t(z_{N_i^\kappa}(t-1))| \le (1-\alpha_{t-1})\frac{\bar{r}}{1-\gamma} + \alpha_{t-1}(\bar{r} + \gamma\frac{\bar{r}}{1-\gamma}) = \frac{\bar{r}}{1-\gamma},$$

which finishes the induction and the proof of part (a).

For part (b), notice that $\epsilon_t = (A_t - \bar{A}_t)\hat{Q}_i^{t+1-\tau} + b_t - \bar{b}_t$. Therefore, it is easy to check that by part (a), $\|\epsilon_t\|_\infty \le 4\frac{\bar{r}}{1-\gamma} + 2\bar{r} = \bar{\epsilon}$.

For part (c), notice that, for any k

$$\|\hat{Q}_i^k - \hat{Q}_i^{k-1}\|_\infty = \alpha_{k-1}\|A_{k-1}\hat{Q}_i^{k-1} + b_{k-1}\|_\infty \le \alpha_{k-1}[2\frac{\bar{r}}{1-\gamma} + \bar{r}].$$

Therefore, by triangle inequality,

$$\|\hat{Q}_i^{t-1} - \hat{Q}_i^{t-\tau}\|_\infty \le [2\frac{\bar{r}}{1-\gamma} + \bar{r}] \sum_{k=t-\tau}^{t-2} \alpha_k.$$

As a consequence,

$$\|\phi_t\|_\infty \le \|A_t - \bar{A}_t\|_\infty \|\hat{Q}_i^t - \hat{Q}_i^{t-\tau+1}\|_\infty \le [8\frac{\bar{r}}{1-\gamma} + 4\bar{r}] \sum_{k=t-\tau+1}^{t-1} \alpha_k = 2\bar{\epsilon} \sum_{k=t-\tau+1}^{t-1} \alpha_k.$$

□

Proof of Lemma 9. Notice that $\log(1-x) \le -x$ for all $x < 1$. Then,

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

Therefore,

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

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

For part (b),

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

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

$$\begin{aligned}\sum_{k=1}^{t-1} \beta_{k,t}^2 &\le \frac{2h^2}{(t+t_0)^{2\sigma h}} \sum_{k=1}^{t-1} (k+t_0)^{2\sigma h-2} \le \frac{2h^2}{(t+t_0)^{2\sigma h}} \int_1^t (y+t_0)^{2\sigma h-2} dy \\ &< \frac{2h^2}{(t+t_0)^{2\sigma h}} \frac{1}{2\sigma h-1} (t+t_0)^{2\sigma h-1} < \frac{2h}{\sigma} \frac{1}{(t+t_0)},\end{aligned}$$

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

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

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

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

C.4 Proof of Lemma 10 and Lemma 11

Since $\|\sum_{k=\tau}^{t-1} \alpha_k \tilde{B}_{k,t} \phi_k\|_\infty \le \sum_{k=\tau}^{t-1} \beta_{k,t} \|\phi_k\|_\infty$, Lemma 8 (c) and Lemma 9 (c) directly leads to the bound in Lemma 11. So, in this section, we focus on the proof of Lemma 10. We start by stating a variant of the Azuma-Hoeffding bound that handles our "shifted" Martingale difference sequence.

Lemma 18. *Let X_t be a $\mathcal{F}_t$-adapted stochastic process, satisfying $\mathbb{E}X_t|\mathcal{F}_{t-\tau} = 0$. Further, $|X_t| \le \bar{X}_t$ almost surely. Then with probability $1-\delta$, we have,*

$$|\sum_{k=0}^{t} X_t| \le \sqrt{2\tau \sum_{k=0}^{t} \bar{X}_k^2 \log(\frac{2\tau}{\delta})}.$$

Proof of Lemma 18. Let ℓ be an integer between 0 and $\tau-1$. For each ℓ, define process $Y_k^\ell = X_{\tau k+\ell}$, scalar $\bar{Y}_k^\ell = \bar{X}_{k\tau+\ell}$, and define Filtration $\tilde{\mathcal{F}}_k^\ell = \mathcal{F}_{\tau k+\ell}$. Then, Y_k^ℓ is $\tilde{\mathcal{F}}_k^\ell$-adapted, and satisfies

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

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

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

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

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

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

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

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

Recall that Lemma 10 is an upper bound on $\|\sum_{k=\tau}^{t-1}\alpha_k\tilde{B}_{k,t}\epsilon_k\|$, where $\sum_{k=\tau}^{t-1}\alpha_k\tilde{B}_{k,t}\epsilon_k$ is a random vector in $\mathbb{R}^{\mathcal{Z}_{N_i^\kappa}}$, with its $z_{N_i^\kappa}$'th entry being

$$\sum_{k=\tau}^{t-1}\alpha_k\epsilon_k(z_{N_i^\kappa})\prod_{\ell=k+1}^{t-1}(1-\alpha_\ell\bar{d}_\ell(z_{N_i^\kappa})), \tag{46}$$

with $\bar{d}_\ell(z_{N_i^\kappa}) \ge \sigma$ almost surely, cf. Lemma 6. Fixing $z_{N_i^\kappa}$, as have been shown in (21), $\epsilon_k(z_{N_i^\kappa})$ is a $\mathcal{F}_{k+1}$ adapted stochastic process satisfying $\mathbb{E}\epsilon_k(z_{N_i^\kappa})|\mathcal{F}_{k+1-\tau} = 0$. However, $\prod_{\ell=k+1}^{t-1}(1-\alpha_\ell\bar{d}_\ell(z_{N_i^\kappa}))$ is not $\mathcal{F}_{k+1-\tau}$-measurable, and as such we cannot directly apply the Azuma-Hoeffding bound in Lemma 18 to quantity (46). In what follows, we first show in Lemma 19 that almost surely, the absolute value of quantity (46) can be upper bounded by the sup of another quantity, to which we can directly apply Lemma 18. With the help of Lemma 19, we can use the Azuma-Hoeffding bound to control (46) and prove Lemma 10.

Lemma 19. *For each $z_{N_i^\kappa}$, we have almost surely,*

$$\Big|\sum_{k=\tau}^{t-1}\alpha_k\epsilon_k(z_{N_i^\kappa})\prod_{\ell=k+1}^{t-1}(1-\alpha_\ell\bar{d}_\ell(z_{N_i^\kappa}))\Big| \le \sup_{\tau\le k_0\le t-1}\Big(|\sum_{k=k_0+1}^{t-1}\epsilon_k(z_{N_i^\kappa})\beta_{k,t}| + 2\bar{\epsilon}\beta_{k_0,t}\Big).$$

Proof of Lemma 19. Let p_k be a scalar sequence defined as follows. Set $p_\tau = 0$, and

$$p_k = (1-\alpha_{k-1}\bar{d}_{k-1}(z_{N_i^\kappa}))p_{k-1} + \alpha_{k-1}\epsilon_{k-1}(z_{N_i^\kappa}).$$

Then $p_t = \sum_{k=\tau}^{t-1}\alpha_k\epsilon_k(z_{N_i^\kappa})\prod_{\ell=k+1}^{t-1}(1-\alpha_\ell\bar{d}_\ell(z_{N_i^\kappa}))$, and to prove Lemma 19 we need to bound $|p_t|$. Let

$$k_0 = \sup\{k \le t-1 : (1-\alpha_k\bar{d}_k(z_{N_i^\kappa}))|p_k| \le \alpha_k|\epsilon_k(z_{N_i^\kappa})|\}.$$

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

$$\tilde{p}_k = (1-\alpha_{k-1}\sigma)\tilde{p}_{k-1} + \alpha_{k-1}\epsilon_{k-1}(z_{N_i^\kappa}).$$

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

$$(1 - \alpha_{k-1}\bar{d}_{k-1}(z_{N_i^\kappa}))p_{k-1} > |\alpha_{k-1}\epsilon_{k-1}(z_{N_i^\kappa})|.$$

Therefore, $p_k > 0$. Further, since $\bar{d}_{k-1}(z_{N_i^\kappa}) \geq \sigma$, we also have

$$(1 - \alpha_{k-1}\sigma)\tilde{p}_{k-1} \geq (1 - \alpha_{k-1}\bar{d}_{k-1}(z_{N_i^\kappa}))p_{k-1} > |\alpha_{k-1}\epsilon_{k-1}(z_{N_i^\kappa})|.$$

These imply $\tilde{p}_k \geq p_k > 0$. The case where both p_{k-1} and $\tilde{p}_{k-1}$ are negative are similar. This finishes the induction, and as a result, $|p_t| \leq |\tilde{p}_t|$.

Notice,

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

By the definition of k_0, we have

$$|p_{k_0+1}| \leq (1 - \alpha_{k_0}\bar{d}_{k_0}(z_{N_i^\kappa}))|p_{k_0}| + \alpha_{k_0}|\epsilon_{k_0}(z_{N_i^\kappa})| \leq 2\alpha_{k_0}|\epsilon_{k_0}(z_{N_i^\kappa})| \leq 2\alpha_{k_0}\bar{\epsilon},$$

where in the last step, we have used the upper bound on $\|\epsilon_{k_0}\|_\infty$ in Lemma 8 (b). As a result,

$$\begin{aligned}
|p_t| \leq |\tilde{p}_t| &\leq \Big| \sum_{k=k_0+1}^{t-1} \epsilon_k(z_{N_i^\kappa})\beta_{k,t} \Big| + \big|\tilde{p}_{k_0+1}\tilde{\beta}_{k_0,t}\big| \\
&\leq \Big| \sum_{k=k_0+1}^{t-1} \epsilon_k(z_{N_i^\kappa})\beta_{k,t} \Big| + \big|2\alpha_{k_0}\bar{\epsilon}\tilde{\beta}_{k_0,t}\big| \\
&= \Big| \sum_{k=k_0+1}^{t-1} \epsilon_k(z_{N_i^\kappa})\beta_{k,t} \Big| + 2\bar{\epsilon}\beta_{k_0,t}.
\end{aligned}$$

□

With the above preparations, we are now ready to prove Lemma 10.

Proof of Lemma 10. Fix $z_{N_i^\kappa}$ and $\tau \leq k_0 \leq t - 1$. As have been shown in (21), $\epsilon_k(z_{N_i^\kappa})\beta_{k,t}$ is a $\mathcal{F}_{k+1}$ adapted stochastic process satisfying $\mathbb{E}\epsilon_k(z_{N_i^\kappa})\beta_{k,t}|\mathcal{F}_{k+1-\tau} = 0$. Also by Lemma 8(b), $|\epsilon_k(z_{N_i^\kappa})\beta_{k,t}| \leq \bar{\epsilon}\beta_{k,t}$ almost surely. As a result, we can use the Azuma-Hoeffding bound in Lemma 18 to get with probability $1 - \delta$,

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

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

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

Then, by Lemma 19, we have with probability $1-\delta$,

$$\begin{aligned}
\Big|\sum_{k=\tau}^{t-1}\alpha_k\epsilon_k(z_{N_i^\kappa})\prod_{\ell=k+1}^{t-1}(1-\alpha_\ell\bar{d}_\ell(z_{N_i^\kappa}))\Big| &\le \sup_{\tau\le k_0\le t-1}\Big(\Big|\sum_{k=k_0+1}^{t-1}\epsilon_k(z_{N_i^\kappa})\beta_{k,t}\Big| + 2\bar{\epsilon}\beta_{k_0,t}\Big) \\
&\le \bar{\epsilon}\sqrt{2\tau\sum_{k=\tau+1}^{t-1}\beta_{k,t}^2\log(\frac{2\tau t}{\delta})} + \sup_{\tau\le k_0\le t-1}2\bar{\epsilon}\beta_{k_0,t} \\
&\le 2\bar{\epsilon}\sqrt{\frac{\tau h}{\sigma(t+t_0)}\log(\frac{2\tau t}{\delta})} + \sup_{\tau\le k_0\le t-1}2\bar{\epsilon}\frac{h}{k_0+t_0}\Big(\frac{k_0+1+t_0}{t+t_0}\Big)^{\sigma h} \\
&\le 2\bar{\epsilon}\sqrt{\frac{\tau h}{\sigma(t+t_0)}\log(\frac{2\tau t}{\delta})} + 2\bar{\epsilon}\frac{h}{t-1+t_0} \\
&\le 6\bar{\epsilon}\sqrt{\frac{\tau h}{\sigma(t+t_0)}\log(\frac{2\tau t}{\delta})},
\end{aligned}$$

where in the third inequality, we have used the bounds on $\beta_{k,t}$ in Lemma 9. Finally, apply the union bound over $z_{N_i^\kappa}\in\mathcal{Z}_{N_i^\kappa}$, and noticing that $|N_i^\kappa|\le f(\kappa)$ and $|\mathcal{Z}_{N_i^\kappa}|\le(SA)^{f(\kappa)}$ by Assumption 1, we have with probability $1-\delta$,

$$\Big\|\sum_{k=\tau}^{t-1}\alpha_k\tilde{B}_{k,t}\epsilon_k\Big\|_\infty \le 6\bar{\epsilon}\sqrt{\frac{\tau h}{\sigma(t+t_0)}\log(\frac{2\tau t(SA)^{f(\kappa)}}{\delta})} = 6\bar{\epsilon}\sqrt{\frac{\tau h}{\sigma(t+t_0)}[\log(\frac{2\tau t}{\delta})+f(\kappa)\log SA]}.$$

□

C.5 Proof of Lemma 12

Throughout the proof, we fix $z_{N_i^\kappa}$ and prove the desired upper bound. For notational simplicity, we drop the dependence on $z_{N_i^\kappa}$ and write $b_{k,t}$ and $\bar{d}_k$ instead, and we will use the property $\bar{d}_k\ge\sigma$. Define the sequence

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

We use induction to show that $e_t\le\frac{1}{\sqrt{\gamma}(t+t_0)^\omega}$. The statement is clearly true for $t=\tau+1$, as $e_{\tau+1}=b_{\tau,\tau+1}\frac{1}{(\tau+t_0)^\omega}=\alpha_\tau\bar{d}_\tau\frac{1}{(\tau+t_0)^\omega}\le\frac{1}{\sqrt{\gamma}(\tau+1+t_0)^\omega}$ (last step needs $\alpha_\tau\le\frac{1}{2},(1+\frac{1}{t_0})^\omega\le\frac{2}{\sqrt{\gamma}}$, implied

by $t_0 \geq 1$, $\omega \leq 1$). Let the statement be true for $t-1$. Then, notice that,

$$\begin{aligned}
e_t &= \sum_{k=\tau}^{t-2} b_{k,t} \frac{1}{(k+t_0)^\omega} + b_{t-1,t} \frac{1}{(t-1+t_0)^\omega} \\
&= (1-\alpha_{t-1}\bar{d}_{t-1}) \sum_{k=\tau}^{t-2} b_{k,t-1} \frac{1}{(k+t_0)^\omega} + \alpha_{t-1}\bar{d}_{t-1} \frac{1}{(t-1+t_0)^\omega} \\
&= (1-\alpha_{t-1}\bar{d}_{t-1}) e_{t-1} + \alpha_{t-1}\bar{d}_{t-1} \frac{1}{(t-1+t_0)^\omega} \\
&\leq (1-\alpha_{t-1}\bar{d}_{t-1}) \frac{1}{\sqrt{\gamma}(t-1+t_0)^\omega} + \alpha_{t-1}\bar{d}_{t-1} \frac{1}{(t-1+t_0)^\omega} \\
&= \Big[1 - \alpha_{t-1}\bar{d}_{t-1}(1-\sqrt{\gamma})\Big] \frac{1}{\sqrt{\gamma}(t-1+t_0)^\omega},
\end{aligned}$$

where the inequality is based on induction assumption. Then, plug in $\alpha_{t-1} = \frac{h}{t-1+t_0}$ and use $\bar{d}_{t-1} \geq \sigma$, we have,

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

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

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

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

D Proof of Auxiliary Results for Analysis of Actor

D.1 Proof of Lemma 13

Recall that

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

Therefore,

$$\begin{aligned}
\|\hat{g}_i(m)\| &\leq \sum_{t=0}^{T} \gamma^t \frac{1}{n} \sum_{j \in N_i^\kappa} |\hat{Q}_j^{m,T}(s_{N_j^\kappa}(t), a_{N_j^\kappa}(t))| \|\nabla_{\theta_i} \log \zeta_i^{\theta_i(m)}(a_i(t)|s_i(t))\| \\
&\leq \sum_{t=0}^{T} \gamma^t \frac{\bar{r}}{1-\gamma} L_i < \frac{\bar{r}}{(1-\gamma)^2} L_i,
\end{aligned}$$

where we have used that $\|\hat{Q}_j^{m,T}\|_\infty \leq \frac{\bar{r}}{1-\gamma}$ almost surely (cf. Lemma 8 (a)). As a result,

$$\|\hat{g}(m)\| = \sqrt{\sum_{i=1}^{n} \|\hat{g}_i(m)\|^2} < \frac{\bar{r}}{(1-\gamma)^2} L.$$

The upper bounds for $\|g(m)\|$, $\|h(m)\|$ and $\|\nabla J(\theta(m))\|$ can be obtained in an almost identical way and their proof is therefore omitted. □

D.2 Proof of Lemma 14

Let $\mathcal{G}_m$ be the σ-algebra generated by the trajectories in the first m outer-loop iterations. Then, Theorem 5 implies that, fixing each $m \leq M$ and $i \in \mathcal{N}$, conditioned on $\mathcal{G}_{m-1}$, the following event happens with probability at least $1-\delta$:

$$\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left|Q_i^{\theta(m)}(s,a) - \hat{Q}_i^{m,T}(s_{N_i^\kappa}, a_{N_i^\kappa})\right| \leq \frac{C_a(\delta,T)}{\sqrt{T+t_0}} + \frac{C_a'}{T+t_0} + \frac{2c\rho^{\kappa+1}}{(1-\gamma)^2},$$

where

$$C_a(\delta,T) = \frac{6\bar{\epsilon}}{1-\sqrt{\gamma}}\sqrt{\frac{\tau h}{\sigma}[\log(\frac{2\tau T^2}{\delta}) + f(\kappa)\log SA]}, C_a' = \frac{2}{1-\sqrt{\gamma}}\max(\frac{16\bar{\epsilon}h\tau}{\sigma}, \frac{2\bar{r}}{1-\gamma}(\tau+t_0)),$$

with $\bar{\epsilon} = 4\frac{\bar{r}}{1-\gamma} + 2\bar{r}$.

We can take expectation and average out $\mathcal{G}_{m-1}$, and apply union bound over $0 \leq m \leq M-1$ and $i \in \mathcal{N}$, getting with probability at least $1-\frac{\delta}{2}$,

$$\begin{aligned}\sup_{m\leq M-1}\sup_{i\in\mathcal{N}}\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left|Q_i^{\theta(m)}(s,a) - \hat{Q}_i^{m,T}(s_{N_i^\kappa}, a_{N_i^\kappa})\right| &\leq \frac{C_a(\frac{\delta}{2nM},T)}{\sqrt{T+t_0}} + \frac{C_a'}{T+t_0} + \frac{2c\rho^{\kappa+1}}{(1-\gamma)^2} \\ &\leq \frac{4c\rho^{\kappa+1}}{(1-\gamma)^2}, \end{aligned} \tag{47}$$

where in the last step, we have used that our lower bound on T implies $\frac{C_a(\frac{\delta}{2nM},T)}{\sqrt{T+t_0}} + \frac{C_a'}{T+t_0} \leq \frac{2c\rho^{\kappa+1}}{(1-\gamma)^2}$. Therefore, conditioned on (47) being true, we have for any $m \leq M-1$ and any $i \in \mathcal{N}$,

$$\begin{aligned}&\|\hat{g}_i(m) - g_i(m)\| \\ &\leq \Big\|\sum_{t=0}^{T}\gamma^t \frac{1}{n}\sum_{j\in N_i^\kappa}\left[Q_j^{\theta(m)}(s(t),a(t)) - \hat{Q}_j^{m,T}(s_{N_j^\kappa}(t), a_{N_j^\kappa}(t))\right]\nabla_{\theta_i}\log\zeta_i^{\theta_i(m)}(a_i(t)|s_i(t))\Big\| \\ &\leq \sum_{t=0}^{T}\gamma^t \frac{1}{n}\sum_{j\in N_i^\kappa}\left|Q_j^{\theta(m)}(s(t),a(t)) - \hat{Q}_j^{m,T}(s_{N_j^\kappa}(t), a_{N_j^\kappa}(t))\right|\left\|\nabla_{\theta_i}\log\zeta_i^{\theta_i(m)}(a_i(t)|s_i(t))\right\| \\ &\leq \sum_{t=0}^{T}\gamma^t \frac{4c\rho^{\kappa+1}}{(1-\gamma)^2}L_i < \frac{4cL_i\rho^{\kappa+1}}{(1-\gamma)^3}.\end{aligned}$$

As a result,

$$\sup_{0\leq m\leq M-1}\|\hat{g}(m) - g(m)\| \leq \frac{4cL\rho^{\kappa+1}}{(1-\gamma)^3},$$

which is true conditioned on event (47) is true that happens with probability at least $1-\frac{\delta}{2}$. □

D.3 Proof of Lemma 15

By Lemma 13, we have almost surely,

$$|\eta_m\langle\nabla J(\theta(m)), e^2(m)\rangle| \le \eta_m\|\nabla J(\theta(m))\|\|h(m)-g(m)\| \le \eta_m\frac{2\bar{r}^2L^2}{(1-\gamma)^4}.$$

As $\eta_m\langle\nabla J(\theta(m)), e^2(m)\rangle$ is a martingale difference sequence w.r.t. $\mathcal{G}_m$, we have by Azuma Hoeffding bound, with probability at least $1-\frac{1}{2}\delta$,

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

□

D.4 Proof of Lemma 16

By (25), we have

$$\begin{aligned}\nabla_{\theta_i}J(\theta(m)) &= \sum_{t=0}^{\infty}\mathbb{E}_{s\sim\pi_t^{\theta(m)},a\sim\zeta^{\theta(m)}(\cdot|s)}\left[\gamma^tQ^{\theta(m)}(s,a)\nabla_{\theta_i}\log\zeta^{\theta(m)}(a|s)\right]\\ &= \sum_{t=0}^{\infty}\mathbb{E}_{s\sim\pi_t^{\theta(m)},a\sim\zeta^{\theta(m)}(\cdot|s)}\left[\gamma^tQ^{\theta(m)}(s,a)\nabla_{\theta_i}\log\zeta_i^{\theta_i(m)}(a_i|s_i)\right]\end{aligned}$$

where we have used $\nabla_{\theta_i}\log\zeta^{\theta(m)}(a|s) = \nabla_{\theta_i}\sum_{j\in\mathcal{N}}\log\zeta_j^{\theta_j(m)}(a_j|s_j) = \nabla_{\theta_i}\log\zeta_i^{\theta_i(m)}(a_i|s_i)$. Also recall the definition of $h_i(\theta)$ in (27),

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

The rest of the proof is essentially the same as Lemma 3. For completeness we provide a proof below. Combining the above two equations, we have,

$$\begin{aligned}&\nabla_{\theta_i}J(\theta(m)) - h_i(m)\\ &= \sum_{t=0}^{T}\mathbb{E}_{s\sim\pi_t^{\theta(m)},a\sim\zeta^{\theta(m)}(\cdot|s)}\left[\gamma^t\nabla_{\theta_i}\log\zeta_i^{\theta_i(m)}(a_i|s_i)\left(Q^{\theta(m)}(s,a)-\frac{1}{n}\sum_{j\in N_i^\kappa}Q_j^{\theta(m)}(s,a)\right)\right]\\ &\quad+\sum_{t=T+1}^{\infty}\mathbb{E}_{s\sim\pi_t^{\theta(m)},a\sim\zeta^{\theta(m)}(\cdot|s)}\left[\gamma^t\nabla_{\theta_i}\log\zeta_i^{\theta_i(m)}(a_i|s_i)Q^{\theta(m)}(s,a)\right]\\ &:= E_1+E_2.\end{aligned}$$

Clearly, the second term satisfies $\|E_2\| \leq \frac{L_i \bar{r}}{(1-\gamma)^2}\gamma^{T+1}$. For E_1, we have

$$\begin{aligned}
E_1 &= \sum_{t=0}^{T} \mathbb{E}_{s\sim\pi_t^{\theta(m)}, a\sim\zeta^{\theta(m)}(\cdot|s)} \left[\gamma^t \nabla_{\theta_i} \log \zeta_i^{\theta_i(m)}(a_i|s_i) \left(\frac{1}{n} \sum_{j\in N_{-i}^{\kappa}} Q_j^{\theta(m)}(s,a) \right) \right] \\
&= \sum_{t=0}^{T} \mathbb{E}_{s\sim\pi_t^{\theta(m)}, a\sim\zeta^{\theta(m)}(\cdot|s)} \left[\gamma^t \nabla_{\theta_i} \log \zeta_i^{\theta_i(m)}(a_i|s_i) \frac{1}{n} \sum_{j\in N_{-i}^{\kappa}} \left(Q_j^{\theta(m)}(s,a) - \hat{Q}_j^{\theta(m)}(s_{N_j^{\kappa}}, a_{N_j^{\kappa}}) \right) \right] \\
&\quad + \sum_{t=0}^{T} \mathbb{E}_{s\sim\pi_t^{\theta(m)}, a\sim\zeta^{\theta(m)}(\cdot|s)} \left[\gamma^t \nabla_{\theta_i} \log \zeta_i^{\theta_i(m)}(a_i|s_i) \frac{1}{n} \sum_{j\in N_{-i}^{\kappa}} \hat{Q}_j^{\theta(m)}(s_{N_j^{\kappa}}, a_{N_j^{\kappa}}) \right] \\
&:= E_3 + E_4,
\end{aligned}$$

where $\hat{Q}_j^{\theta(m)}$ is any truncated Q function for $Q_j^{\theta(m)}$ as defined in (9). We claim E_4 is zero. To see this, consider for any $j \in N_{-i}^{\kappa}$ and any t,

$$\begin{aligned}
&\mathbb{E}_{s\sim\pi_t^{\theta(m)}, a\sim\zeta^{\theta(m)}(\cdot|s)} \left[\nabla_{\theta_i} \log \zeta_i^{\theta_i(m)}(a_i|s_i) \hat{Q}_j^{\theta(m)}(s_{N_j^{\kappa}}, a_{N_j^{\kappa}}) \right] \\
&= \sum_{s,a} \pi_t^{\theta(m)}(s) \prod_{\ell=1}^{n} \zeta_\ell^{\theta_\ell(m)}(a_\ell|s_\ell) \frac{\nabla_{\theta_i} \zeta_i^{\theta_i(m)}(a_i|s_i)}{\zeta_i^{\theta_i(m)}(a_i|s_i)} \hat{Q}_j^{\theta(m)}(s_{N_j^{\kappa}}, a_{N_j^{\kappa}}) \\
&= \sum_{s,a} \pi_t^{\theta(m)}(s) \prod_{\ell\neq i} \zeta_\ell^{\theta_\ell(m)}(a_\ell|s_\ell) \nabla_{\theta_i} \zeta_i^{\theta_i(m)}(a_i|s_i) \hat{Q}_j^{\theta(m)}(s_{N_j^{\kappa}}, a_{N_j^{\kappa}}) \\
&= \sum_{s,a_{1:i-1},a_{i+1:n}} \pi_t^{\theta(m)}(s) \prod_{\ell\neq i} \zeta_\ell^{\theta_\ell(m)}(a_\ell|s_\ell) \hat{Q}_j^{\theta(m)}(s_{N_j^{\kappa}}, a_{N_j^{\kappa}}) \sum_{a_i} \nabla_{\theta_i} \zeta_i^{\theta_i(m)}(a_i|s_i) \\
&= 0,
\end{aligned}$$

where in the last equality, we have used $\hat{Q}_j^{\theta(m)}(s_{N_j^{\kappa}}, a_{N_j^{\kappa}})$ does not depend on a_i as $i \notin N_j^{\kappa}$; and $\sum_{a_i} \nabla_{\theta_i} \zeta_i^{\theta_i(m)}(a_i|s_i) = \nabla_{\theta_i} \sum_{a_i} \zeta_i^{\theta_i(m)}(a_i|s_i) = \nabla_{\theta_i} 1 = 0$.

For E_3, by the exponential decay property, the truncated Q function has a small error, cf. (38),

$$\sup_{s,a} |Q_j^{\theta(m)}(s,a) - \hat{Q}_j^{\theta(m)}(s_{N_j^{\kappa}}, a_{N_j^{\kappa}})| \leq c\rho^{\kappa+1},$$

and as a result,

$$\|E_3\| \leq \frac{1-\gamma^{T+1}}{1-\gamma} L_i c \rho^{\kappa+1} < \frac{L_i c}{(1-\gamma)} \rho^{\kappa+1}.$$

Therefore,

$$\begin{aligned}
\|\nabla_{\theta_i} J(\theta(m)) - h_i(m)\| = \|E_2 + E_3\| &\leq \frac{L_i \bar{r}}{(1-\gamma)^2} \gamma^{T+1} + \frac{L_i c}{(1-\gamma)} \rho^{\kappa+1}, \\
&\leq 2 \frac{L_i c}{(1-\gamma)} \rho^{\kappa+1},
\end{aligned}$$

where in the last step, we have used

$$T+1 \geq \frac{\log \frac{c(1-\gamma)}{\bar{r}} + (\kappa+1)\log\rho}{\log\gamma},$$

and as a result, $\|\nabla J(\theta(m)) - h(m)\| \leq 2\frac{Lc}{(1-\gamma)}\rho^{\kappa+1}$. □

Acknowledgement

We would like to thank Yiheng Lin of Caltech and Prof. Longbo Huang of Tsinghua University for suggesting the wireless communication example. This work was supported by Resnick Sustainability Institute Fellowship, NSF CAREER 1553407, ONR YIP N00014-19-1-2217, AFOSR YIP FA9550-18-1-0150, the PIMCO Fellowship, NSF AitF-1637598, NSF CNS-1518941, Amazon AI4Science Fellowship, and Caltech Center for Autonomous Systems and Technologies (CAST).

References

A. Agarwal, S. M. Kakade, J. D. Lee, and G. Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research*, 22 (98):1–76, 2021.

H. J. Ahn. *Random propagation in complex systems: nonlinear matrix recursions and epidemic spread*. PhD thesis, California Institute of Technology, 2014.

N. Azizan Ruhi, H. J. Ahn, and B. Hassibi. Analysis of exact and approximated epidemic models over complex networks. *arXiv preprint arXiv:1609.09565*, 2016a.

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

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

D. P. Bertsekas. *Dynamic programming and optimal control*. Athena Scientific optimization and computation series. Athena Scientific, Belmont, Mass., 3rd ed. edition, 2005. ISBN 1886529086.

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

J. Bhandari and D. Russo. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786*, 2019.

R. Block and B. Van Houdt. Spatial fairness in multi-channel CSMA line networks. *Performance Evaluation*, 103:69–85, 2016.

V. D. Blondel and J. N. Tsitsiklis. A survey of computational complexity results in systems and control. *Automatica*, 36(9):1249–1274, 2000.

T. Britton. Stochastic epidemic models: a survey. *Mathematical biosciences*, 225(1):24–35, 2010.

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.

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.

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

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.

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.

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

C. Guestrin, D. Koller, R. Parr, and S. Venkataraman. Efficient solution algorithms for factored MDPs. *Journal of Artificial Intelligence Research*, 19:399–468, 2003.

J. Hu and M. P. Wellman. Nash Q-learning for general-sum stochastic games. *Journal of machine learning research*, 4(Nov):1039–1069, 2003.

C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan. Is Q-learning provably efficient? In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 4868–4878, 2018.

S. Kar, J. M. Moura, and H. V. Poor. QD-learning: A collaborative distributed strategy for multi-agent reinforcement learning through consensus + innovations. *IEEE Transactions on Signal Processing*, 61(7):1848–1862, 2013.

M. Kearns and D. Koller. Efficient reinforcement learning in factored MDPs. In *IJCAI*, volume 16, pages 740–747, 1999.

T. H. Kim, J. Ni, R. Srikant, and N. H. Vaidya. On the achievable throughput of CSMA under imperfect carrier sensing. In *2011 Proceedings IEEE Infocom*, pages 1674–1682. IEEE, 2011.

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

Y. A. Kuznetsov and C. Piccardi. Bifurcation analysis of periodic SEIR and SIR epidemic models. *Journal of mathematical biology*, 32(2):109–121, 1994.

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.

G. Li, Y. Wei, Y. Chi, Y. Gu, and Y. Chen. Sample complexity of asynchronous q-learning: Sharper analysis and variance reduction. *Advances in neural information processing systems*, 2020.

Y. Lin, G. Qu, L. Huang, and A. Wierman. Multi-agent reinforcement learning in stochastic networked systems. *Advances in neural information processing systems*, 2021.

M. L. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Machine learning proceedings 1994*, pages 157–163. Elsevier, 1994.

M. L. Littman. Value-function reinforcement learning in Markov games. *Cognitive Systems Research*, 2(1):55–66, 2001.

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.

A. Y. Lokhov, M. Mézard, and L. Zdeborová. Dynamic message-passing equations for models with unidirectional dynamics. *Phys. Rev. E*, 91:012811, Jan 2015. doi: 10.1103/PhysRevE.91.012811. URL https://link.aps.org/doi/10.1103/PhysRevE.91.012811.

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.

S. V. Macua, J. Chen, S. Zazo, and A. H. Sayed. Distributed policy evaluation under multiple behavior strategies. *IEEE Transactions on Automatic Control*, 60(5):1260–1274, 2015.

S. Magnússon, H. Shokri-Ghadikolaei, and N. Li. On maintaining linear convergence of distributed learning and optimization under limited communication. *IEEE Transactions on Signal Processing*, 68:6101–6116, 2020.

A. Mathkar and V. S. Borkar. Distributed reinforcement learning via gossip. *IEEE Transactions on Automatic Control*, 62(3):1465–1470, 2017.

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.

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.

N. Meuleau, M. Hauskrecht, K.-E. Kim, L. Peshkin, L. P. Kaelbling, T. L. Dean, and C. Boutilier. Solving very large weakly coupled Markov decision processes. In *AAAI/IAAI*, pages 165–172, 1998.

M. Mezard and A. Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

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.

D. H. Morris, F. W. Rossine, J. B. Plotkin, and S. A. Levin. Optimal, near-optimal, and robust epidemic control. *arXiv preprint arXiv:2004.02209*, 2020.

R. Nair, P. Varakantham, M. Tambe, and M. Yokoo. Networked distributed POMDPs: A synthesis of distributed constraint optimization and POMDPs. In *AAAI*, volume 5, pages 133–139, 2005.

F. A. Oliehoek and C. Amato. *A concise introduction to decentralized POMDPs*. Springer, 2016.

I. Osband and B. Van Roy. Near-optimal reinforcement learning in factored MDPs. In *Advances in Neural Information Processing Systems*, pages 604–612, 2014.

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

V. M. Preciado, M. Zargham, C. Enyioha, A. Jadbabaie, and G. Pappas. Optimal vaccine allocation to control epidemic outbreaks in arbitrary networks. In *52nd IEEE conference on decision and control*, pages 7486–7491. IEEE, 2013.

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.

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, 2020a.

G. Qu, C. Yu, S. Low, and A. Wierman. Combining model-based and model-free methods for nonlinear control: A provably convergent policy gradient approach. *arXiv preprint arXiv:2006.07476*, 2020b.

L. Roberts. ALOHA packet system with and without slots and capture. *ACM SIGCOMM Computer Communication Review*, 5:28–42, 04 1975. doi: 10.1145/1024916.1024920.

M. Rotkowitz and S. Lall. A characterization of convex problems in decentralized control. *IEEE transactions on Automatic Control*, 50(12):1984–1996, 2005.

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.

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

R. S. Sutton, A. G. Barto, et al. *Introduction to reinforcement learning*, volume 135. MIT press Cambridge, 1998.

R. S. Sutton, D. A. McAllester, S. P. Singh, and Y. Mansour. Policy gradient methods for reinforcement learning with function approximation. In *Advances in neural information processing systems*, pages 1057–1063, 2000.

M. Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*, pages 330–337, 1993.

L. Tassiulas and A. Ephremides. Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. In *29th IEEE Conference on Decision and Control*, pages 2130–2132. IEEE, 1990.

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.

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

J. N. Tsitsiklis and B. Van Roy. On average versus discounted reward temporal-difference learning. *Machine Learning*, 49(2):179–191, 2002.

S. Tu and B. Recht. The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint. In *Conference on Learning Theory*, pages 3036–3083. PMLR, 2019.

B. Van Roy and J. N. Tsitsiklis. Stable linear approximations to dynamic programming for stochastic control problems with local transitions. In *Proceedings of the 8th International Conference on Neural Information Processing Systems*, pages 1045–1051, 1995.

P. Varaiya. Max pressure control of a network of signalized intersections. *Transportation Research Part C: Emerging Technologies*, 36:177–195, 2013.

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. ISSN 0146-4833. doi: 10.1145/774763.774784. URL http://doi.acm.org/10.1145/774763.774784.

H.-T. Wai, Z. Yang, Z. Wang, and M. Hong. Multi-agent reinforcement learning via double averaging primal-dual optimization. In *Proceedings of the 32nd International Conference on Neural Information Processing Systems*, pages 9672–9683, 2018.

P. Whittle. Restless bandits: Activity allocation in a changing world. *Journal of applied probability*, 25(A):287–298, 1988.

Z. Wu, Q.-S. Jia, and X. Guan. Optimal control of multiroom HVAC system: An event-based approach. *IEEE Transactions on Control Systems Technology*, 24(2):662–669, 2016.

S.-Y. Yun, Y. Yi, J. Shin, et al. Optimal CSMA: a survey. In *2012 IEEE international conference on communication systems (ICCS)*, pages 199–204. IEEE, 2012.

K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Basar. Fully decentralized multi-agent reinforcement learning with networked agents. In *International Conference on Machine Learning*, pages 5872–5881. PMLR, 2018.

K. Zhang, Z. Yang, and T. Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control*, pages 321–384, 2021.

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.

X. Zhang, W. Shi, B. Yan, A. Malkawi, and N. Li. Decentralized and distributed temperature control via HVAC systems in energy efficient buildings. *arXiv preprint arXiv:1702.03308*, 2017.

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

S. Zou, T. Xu, and Y. Liang. Finite-sample analysis for SARSA with linear function approximation. In *Advances in Neural Information Processing Systems*, pages 8665–8675, 2019.