

Methods

Scalable Reinforcement Learning for Multiagent Networked Systems

Guannan Qu,^{a,*} Adam Wierman,^b Na Li^c

^aDepartment of Electrical and Computer Engineering, Carnegie Mellon University, Pittsburgh, Pennsylvania 15213; ^bDepartment of Computing and Mathematical Sciences, California Institute of Technology, Pasadena, California 91125; ^cSchool of Engineering and Applied Sciences, Harvard University, Cambridge, Massachusetts 02138

*Corresponding author

Contact: gqu@andrew.cmu.edu,  <https://orcid.org/0000-0002-5466-3550> (GQ); adamw@caltech.edu,

 <https://orcid.org/0000-0002-5923-0199> (AW); nali@seas.harvard.edu,  <https://orcid.org/0000-0001-9545-3050> (NL)

Received: September 2, 2020

Revised: June 18, 2021

Accepted: October 13, 2021

Published Online in *Articles in Advance*:
February 23, 2022

Area of Review: Stochastic Models

<https://doi.org/10.1287/opre.2021.2226>

Copyright: © 2022 INFORMS

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 $O(\rho^{k+1})$ -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.

Funding: This work was supported by the Caltech Center for Autonomous Systems and Technologies; Office of Naval Research [Grant YIP N00014-19-1-2217]; Air Force Office of Scientific Research [Grant YIP FA9550-18-1-0150]; National Science Foundation [Grants AitF-1637598, CAREER 1553407, and CNS-1518941]; PIMCO [PIMCO Fellowship]; Amazon Web Services [Amazon AI4Science Fellowship]; and Resnick Sustainability Institute for Science, Energy and Sustainability, California Institute of Technology [Postdoctoral Fellowship].

Keywords: stochastic systems • networked systems • reinforcement learning

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 (Mnih et al. 2015, Silver et al. 2016), 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* multiagent 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, that

is, the distribution of $s_i(t+1)$ only depends on the current states of the local neighborhood of i as well as the local $a_i(t)$. Furthermore, 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 (Llas et al. 2003, Chakrabarti et al. 2008), wireless communication networks (Vogels et al. 2003, Zocca 2019), 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 $(s_1, \dots, s_n)$ and the action profile $(a_1, \dots, a_n)$ 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, for example, in multiagent 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, for example, the idea of “independent learners” in Tan (1993) and 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 (Blondel and Tsitsiklis 2000).

In addition to the scalability issue, another challenge is that, even if an optimal policy that maps a global state $(s_1, \dots, s_n)$ profile to a global action $(a_1, \dots, a_n)$ 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 (Rotkowitz and Lall 2005).

These challenges 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 as follows: *Can the network structure be utilized to develop scalable RL algorithms that provably find a (near-)optimal localized policy?*

1.1. 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 1) 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 multiagent 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 because of the large state-action space size, we introduce a *truncated Q -function* that only depends on a small spatial horizon (Lemma 3) that can be computed efficiently and can be used in an actor critic framework that 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.

1.2. 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 (Whittle 1988, Papadimitriou and Tsitsiklis 1999, Blondel and Tsitsiklis 2000). Despite this, there is a large amount of literature on RL/MDPs in multiagent settings, which we discuss here.

Multiagent RL dates back to the early work of Littman (1994), Claus and Boutilier (1998), Littman (2001), and Hu and Wellman (2003) (see Bu et al. 2008 for a review) and has been actively studied in Zhang et al. (2018), Kar et al. (2013), Macua et al. (2015), Mathkar and Borkar (2017), and Wai et al. (2018) (see a more recent review in Zhang et al. (2021)). Multiagent RL encompasses a broad range of settings including competitive agents and Markov games. The case most relevant to ours is the cooperative multiagent 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 on its own state. Despite the existence of a global state, multiagent RL still faces scalability issues because 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 uses a single-agent RL method. Although 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, for example, 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 have scalability issues. Similar approaches as in the case of multiagent RL are used. For example, 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.

1.2.1. 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 (Mezard and Montanari 2009, Lokhov et al. 2015), although 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) (Bertsekas 2005, Nair et al. 2005, Oliehoek and Amato 2016), although 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 and M and variable κ . The total variation distance for two distributions π, π' over a finite set S is defined as $\text{TV}(\pi, \pi') = \sup_{E \subset S} |\pi(E) - \pi'(E)|$.

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 $\mathcal{N} = \{1, \dots, n\}$ 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 $s = (s_1, \dots, s_n) \in \mathcal{S} := \mathcal{S}_1 \times \dots \times \mathcal{S}_n$ and similarly the global action $a = (a_1, \dots, a_n) \in \mathcal{A} := \mathcal{A}_1 \times \dots \times \mathcal{A}_n$. 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:

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

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 ; that is, 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 $\theta = (\theta_1, \dots, \theta_n)$ 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.

Furthermore, 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 :

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

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, that is, $r_i(s_{N_i}, a_{N_i})$, and each agent's localized policy can also be generalized to depend on its neighbors' states, that is, $\zeta_i^{\theta_i}(a_i | s_{N_i})$.

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), that is, i itself and the set of nodes that have a directed link pointing toward i . In addition, N_i^κ needs to be redefined as the κ -hop in-neighborhood of i , that is, 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.

2.2.1. Wireless Communication. We consider a wireless network with multiple access points (Zocca 2019), where there is a set of users $\mathcal{N} = \{1, 2, \dots, n\}$, and a set of network access points $Y = \{y_1, y_2, \dots, y_m\}$. 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; that is, 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 one. After this, the system moves to the next time step, with all deadlines of the remaining packets decreasing by one and packets with deadline zero 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 $s_i = (e_1, e_2, \dots, e_{d_i}) \in \mathcal{S}_i = \{0, 1\}^{d_i}$, where for each $\ell \in \{1, \dots, d_i\}$, $e_\ell \in \{0, 1\}$ 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. This model serves as a basis for more complex multiaccess wireless communication models studied in the literature, including those with multiple channels (Block and Van Houdt 2016) and (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), whereas our RL-based approach does not require such knowledge and learns to improve performance in a model-free manner.

2.2.2. Epidemic Network. We consider an SIS (susceptible-infected-susceptible) epidemic network model (Ahn 2014, Azizan Ruhi et al. 2016a, Mei et al. 2017), 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

Table 1. State Transition for the Wireless Communication Example

$s_i(t)$	$a_i(t)$	$s_{N_i \setminus \{i\}}(t), a_{N_i \setminus \{i\}}(t)$	$s_i(t+1)$
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	$y_k \in Y_i$	$\exists j \in N_i \setminus \{i\}$ 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 one in $s_i(t)$ to zero w.p. q_k , then left shift $s_i(t)$ and append Bernoulli(p_i)

Notes. Left shift means, for a binary tuple, discarding the left-most bit. Bernoulli(p_i) means a random variable sampled identical and independently distributed (i.i.d.) from the Bernoulli distribution that has probability p_i to be one and probability $1 - p_i$ to be zero.

well 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 and its control action in place (Preciado et al. 2013, Azizan Ruhi et al. 2016b); 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

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

where $\delta_i \in (0,1)$ is a given recovering rate parameter, and $\beta_i : \mathcal{A}_i \rightarrow (0,1)$ 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

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

which consists of two parts: a positive reward of one 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 previously 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 previous SIS model is a basis for more complex models, for example, those with “exposed” and “recovered” states (Kuznetsov and Piccardi 1994, Britton 2010) or more complex control interventions (Morris et al. 2021). These more complex models can also be captured using the previous framework. We reiterate that the approach in our paper does not require knowledge of model specifications and learns in a model-free manner.

2.2.3. Traffic Network. Last, 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 $s_i = (x_{i,j})_{j \in N_i/\{i\}}$ 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 $[S] = \{0, 1, \dots, S\}$. Correspondingly, the local state space is $\mathcal{S}_i = [S]^{N_i/\{i\}}$. Similarly, the local action $a_i = (y_{i,j})_{j \in N_i/\{i\}}$ is the binary traffic signal tuple with $y_{i,j}$ controlling the on-off of turn movement ($i \rightarrow j$), 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 is then assigned to each of the queues in $(x_{i,j})_{j \in N_i/\{i\}}$. Mathematically,

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

where $[x]_0^S$ means $\max(\min(x, S), 0)$, $C_{i,j}(t)$ (and similarly $C_{k,i}(t)$) is an identical and independently distributed (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

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

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 $C_{i,j}(t)$ and $R_{i,j}(t)$, whereas 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 $\theta = (\theta_1, \dots, \theta_n)$, 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) :

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

In the last step, we 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*

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

3. Algorithm Design and Results

In this paper, we propose an algorithm, SAC, which provably finds an $O(\rho^{\kappa+1})$ -stationary point of the objective $J(\theta)$ (i.e., a θ such that (s.t.) $\|\nabla J(\theta)\|^2 \leq \varepsilon$) for some $\rho \leq \gamma$, 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 1 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 because of 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 $N_{-i}^\kappa = \mathcal{N} / N_i^\kappa$, that is, the set of agents that are outside of i 'th κ -hop neighborhood. We write state s as $(s_{N_i^\kappa}, s_{N_{-i}^\kappa})$, that is, 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

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

It may not be immediately clear when the exponential decay property holds. Lemma 2(a) 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. Furthermore, 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 (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 $\left(\frac{\bar{r}}{1-\gamma}, \gamma\right)$ -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, $\text{TV}(\pi_{t,i}, \pi_{\infty,i}) \leq c' \mu^t$, $\forall t$, 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 $\text{TV}(\cdot, \cdot)$ is the total variation distance. Then, the $\left(\frac{2c'\bar{r}}{1-\gamma\mu}, \gamma\mu\right)$ -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 (Bamieh et al. 2002, Gamarnik 2013, Gamarnik et al. 2014), although 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 with these works where the optimization problem is static, we focus on an MDP setting that has

states that evolve on a time axis. Furthermore, 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) and 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:

$$\hat{Q}_i^\theta(s_{N_i^k}, a_{N_i^k}) = \sum_{s_{N_i^k} \in \mathcal{S}_{N_i^k}, a_{N_i^k} \in \mathcal{A}_{N_i^k}} w_i(s_{N_i^k}, a_{N_i^k}; s_{N_i^k}, a_{N_i^k}) Q_i^\theta(s_{N_i^k}, s_{N_i^k}, a_{N_i^k}, a_{N_i^k}), \quad (9)$$

where $w_i(s_{N_i^k}, a_{N_i^k}; s_{N_i^k}, a_{N_i^k})$ are any nonnegative weights satisfying

$$\sum_{s_{N_i^k} \in \mathcal{S}_{N_i^k}, a_{N_i^k} \in \mathcal{A}_{N_i^k}} w_i(s_{N_i^k}, a_{N_i^k}; s_{N_i^k}, a_{N_i^k}) = 1, \quad \forall (s_{N_i^k}, a_{N_i^k}) \in \mathcal{S}_{N_i^k} \times \mathcal{A}_{N_i^k}. \quad (10)$$

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

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

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

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

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

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: 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 section, we propose one particular approach using the actor critic framework. Our approach, 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.

3.2.1. 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.

3.2.1.1. The Critic: TD-Inner Loop. Line 4 through line 9 is the policy evaluation inner loop that obtains the truncated Q function, whereas line 7 and 8 are the TD update. Steps 7 and 8 use the same update equation as TD learning, except that it “pretends” $(s_{N_i^k}, a_{N_i^k})$ is the true state-action pair, whereas 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.

3.2.1.2. 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.

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  $m = 0, 1, 2, \dots$  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  $\hat{Q}_i^0 \in \mathbb{R}^{\mathcal{S}_{N_i^k} \times \mathcal{A}_{N_i^k}}$  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

```

3.2.2. 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 multiagent RL literature (Lowe et al. 2017), where in the training phase, global communication is used. 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, for example, potentially using consensus schemes like in Zhang et al. (2018) and techniques that only communicate quantized bits as opposed to real numbers (Magnússon et al. 2020).

3.2.3. 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, that is, initialize $\hat{Q}_i$ as the final estimate $\hat{Q}_i$ 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. Furthermore, 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). 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 $\delta_i, \beta_i(\cdot)$ 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) and 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 $|S_i| \leq S$, $|A_i| \leq A$, $\forall i$.*

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

Under Assumption 1, Assumption 2 automatically holds with $\rho = \gamma$ (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 (Lemma 2(b)).

Our third assumption can be interpreted as an ergodicity condition that 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 $(s, a) \in S \times \mathcal{A}$, $\forall i \in \mathcal{N}$,*

$\forall (s'_{N_i^\kappa}, a'_{N_i^\kappa}) \in \mathcal{S}_{N_i^\kappa} \times \mathcal{A}_{N_i^\kappa}$, we have $P((s_{N_i^\kappa}(\tau), a_{N_i^\kappa}(\tau)) = (s'_{N_i^\kappa}, a'_{N_i^\kappa}) \mid (s(1), a(1)) = (s, a)) \geq \sigma$.

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) and 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 that 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 multiagent 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 $\|\nabla_{\theta_i} \log \zeta_i^{\theta_i}(a_i \mid s_i)\| \leq L_i$. As a result, $\|\nabla_{\theta} \log \zeta^{\theta}(a \mid s)\| \leq L = \sqrt{\sum_{i=1}^n L_i^2}$. Furthermore, assume $\nabla J(\theta)$ is L' -Lipschitz continuous in θ .

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

Theorem 1. Under Assumptions 1–4, for any $\delta \in (0, 1)$, $M \geq 3$, suppose the critic step size $\alpha_t = \frac{h}{t+t_0}$ satisfies $h \geq \frac{1}{\sigma} \max\left(2, \frac{1}{1-\gamma}\right)$, $t_0 \geq \max(2h, 4\sigma h, \tau)$; and the actor step size satisfies $\eta_m = \frac{\eta}{\sqrt{m+1}}$ with $\eta \leq \frac{1}{4L}$. Furthermore, if the inner loop length T is large enough s.t. $T+1 \geq \log_{\gamma} \frac{c(1-\gamma)}{\bar{\tau}} + (\kappa+1)\log_{\gamma} \rho$ and

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

where $C_a(\delta, T) = \frac{6\bar{e}}{1-\sqrt{\gamma}} \sqrt{\frac{th}{\sigma} \left[\log\left(\frac{2\tau T^2}{\delta}\right) + f(\kappa) \log SA \right]}$ and $C'_a = \frac{2}{1-\sqrt{\gamma}} \max\left(\frac{16\bar{e}h\tau}{\sigma}, \frac{2\bar{\tau}}{1-\gamma}(\tau+t_0)\right)$, with $\bar{e} = 4\frac{\bar{\tau}}{1-\gamma} + 2\bar{\tau}$ and we recall that $f(\kappa) = \max_i |N_i^\kappa|$ is the size of the largest

κ -neighborhood. Then, with probability at least $1 - \delta$,

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

The proof of Theorem 1 is deferred to Section 4. To interpret the result, the first term in (13) converges to zero in the order of $\tilde{O}\left(\frac{1}{\sqrt{M}}\right)$ 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 $M \geq \tilde{\Omega}\left(\frac{1}{\varepsilon_\kappa} \text{poly}\left(\bar{\tau}, L, L', \frac{1}{(1-\gamma)}\right)\right)$, which scales polynomially with the parameters of the problem. We emphasize that it does not scale exponentially with n . Furthermore, because the left-hand side of (12) decays to zero as T increases in the order of $\tilde{O}\left(\frac{1}{\sqrt{T}}\right)$ and the right-hand side of (12) is in the same order as $O(\varepsilon_\kappa)$, the inner-loop length required is $T \geq \tilde{\Omega}\left(\frac{1}{\varepsilon_\kappa^2} \text{poly}\left(\tau, \frac{1}{\sigma}, \frac{1}{1-\gamma}, \bar{\tau}, f(\kappa)\right)\right)$. 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 1 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 (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 multiagent 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 multiagent RL, our contribution can be interpreted as follows. In multiagent 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 1 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 1.

4.1. 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

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

$$\begin{aligned} \hat{Q}_i^t(s_{N_i^\kappa}, a_{N_i^\kappa}) &= \hat{Q}_i^{t-1}(s_{N_i^\kappa}, a_{N_i^\kappa}) \text{ for } (s_{N_i^\kappa}, a_{N_i^\kappa}) \\ &\neq (s_{N_i^\kappa}(t-1), a_{N_i^\kappa}(t-1)), \end{aligned} \quad (14b)$$

where $\hat{Q}_i^0 \in \mathbb{R}^{S_{N_i^\kappa} \times A_{N_i^\kappa}}$ is initialized to be all zero, and $\alpha_t = \frac{h}{t+t_0}$ is the step size. 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 $Q_i^{\theta(m)} \in \mathbb{R}^{S \times A}$ be the true Q -function for reward r_i under this fixed policy $\theta(m)$ as defined in (7). Given the previous 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 2. Assume Assumptions 1–3 are true and suppose t_0, h satisfies, $h \geq \frac{1}{\sigma} \max\left(2, \frac{1}{1-\sqrt{\gamma}}\right)$ 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:

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

where

$$\begin{aligned} C_a &= \frac{6\bar{e}}{1-\sqrt{\gamma}} \sqrt{\frac{\tau h}{\sigma} \left[\log\left(\frac{2\tau T^2}{\delta}\right) + f(\kappa) \log SA \right]}, \\ C'_a &= \frac{2}{1-\sqrt{\gamma}} \max\left(\frac{16\bar{e}h\tau}{\sigma}, \frac{2\bar{r}}{1-\gamma}(\tau+t_0)\right), \end{aligned}$$

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

The proof of Theorem 2 is postponed to Section 4.1. The result in Theorem 2 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 $\tilde{O}\left(\frac{1}{\sqrt{T}} + \frac{1}{T}\right)$ and a steady-state error that is exponentially small in κ . In proving Theorem 2, we use two key techniques. The first is using the exponential decay property (cf. Definition 1 and 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 because of our results in Lemma 3, which show 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 with 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.

We view the actor step as a biased stochastic gradient step, with the bias characterized by our result in the critic (Theorem 2). Under this viewpoint, we finish the analysis of the actor and the proof of the main result in Section 4.2.

4.2. Analysis of the Critic: Proof of Theorem 2

Because Theorem 2 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 2 essentially says that the critic iterate $\hat{Q}_i^t$ in (14) will become a good estimate of Q_i^* as t increases. The full Q -function Q_i^* must satisfy the Bellman equation (Bertsekas and Tsitsiklis 1996):

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

where $\text{TD} : \mathbb{R}^{\mathcal{Z}} \rightarrow \mathbb{R}^{\mathcal{Z}}$ 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 $\theta(m)$. 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 three steps. In Step 1, we rewrite (14a) and (14b) in a linear update form (cf. (17)), study its behavior (Lemma 4), and decompose the error into a recursive form (Lemma 5). In Step 2, we bound the noise sequences in the error decomposition (Lemmas 8 and 9). Finally, in Step 3, we use the recursive error decomposition and the bound on the noise sequences to prove Theorem 2.

Step 1. Error Decomposition. Define $\mathbf{e}_{z_{N_i^\kappa}}$ to be the indicator vector in $\mathbb{R}^{\mathcal{Z}_{N_i^\kappa}}$; that is, the $z_{N_i^\kappa}$ th entry of $\mathbf{e}_{z_{N_i^\kappa}}$ is one and other entries are zero. Then, the critic update Equation (14) can be written as

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

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 $A_{z,z'}, b_z$):

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

and rewrite (16) in a linear form:

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

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 $z(0), \dots, z(t)$. 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 $\mathcal{F}_{t-\tau}$. Furthermore, 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

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

where in the last step, we have defined sequences $\epsilon_{t-1}, \phi_{t-1} \in \mathbb{R}^{Z_{N_i^k}}$, 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 4. The proof of Lemma 4 is similar to the analysis of fixed points for TD learning with linear function approximation (Van Roy and Tsitsiklis 1995). We postpone the detailed proof of Lemma 4 to Appendix C.1.

Lemma 4. For all t , there exists diagonal matrix $D_{t-1} \in \mathbb{R}^{Z_{N_i^k} \times Z_{N_i^k}}$ and operator $g_{t-1} : \mathbb{R}^{Z_{N_i^k}} \rightarrow \mathbb{R}^{Z_{N_i^k}}$ s.t.

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

where D_{t-1} satisfies $D_{t-1} \succeq \sigma I$ (recall $\sigma > 0$ is from Assumption 3) with its $Z_{N_i^k}$ th diagonal entry being $\bar{d}_{t-1}(Z_{N_i^k}) = \sum_{Z_{N_i^k} \in Z_{N_i^k}} d_{t-1}(Z_{N_i^k}, Z_{N_i^k})$, which means the marginalized distribution of $Z_{N_i^k}$ under distribution d_{t-1} . Furthermore, g_{t-1} is a γ -contraction in the infinity norm, with unique fixed point $\hat{Q}_i^{*,t-1}$ satisfying $\sup_{z \in Z} |\hat{Q}_i^{*,t-1}(Z_{N_i^k}) - Q_i^*(z)| \leq \frac{c\rho^{k+1}}{1-\gamma}$.

From Lemma 4, 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} . Although 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 4. As such, moving toward $\hat{Q}_i^{*,t-1}$ at each time step should eventually produce a good estimate of the full Q -function Q_i^* . In the following, we turn the previous intuition into an error decomposition. In details, we

unroll (18) and get

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

where we have used the following shorthand 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 $\beta_{k,t} = \alpha_k \prod_{\ell=k+1}^{t-1}(1 - \alpha_\ell \sigma)$, $\tilde{\beta}_{k,t} = \prod_{\ell=k+1}^{t-1}(1 - \alpha_\ell \sigma)$. Because every diagonal entry of D_ℓ is lower bounded by σ almost surely (Lemma 4), 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 shorthand notations, we state the error decomposition, which is a consequence of (20) and the property of operator $g_k(\cdot)$ and its fixed point $\hat{Q}_i^{*,k}$ in Lemma 4. The proof of Lemma 5 is postponed to Appendix C.2.

Lemma 5. Let $\xi_t = \sup_{z \in Z} |\hat{Q}_i^t(Z_{N_i^k}) - Q_i^*(z)|$. The following recursion holds almost surely:

$$\begin{aligned}\xi_t &\leq \tilde{\beta}_{\tau-1,t} \xi_\tau + \gamma \sup_{Z_{N_i^k} \in Z_{N_i^k}} \sum_{k=\tau}^{t-1} b_{k,t}(Z_{N_i^k}) \xi_k + \frac{2c\rho^{k+1}}{1-\gamma} \\ &\quad + \left\| \sum_{k=\tau}^{t-1} \alpha_k \tilde{B}_{k,t} \epsilon_k \right\|_\infty + \left\| \sum_{k=\tau}^{t-1} \alpha_k \tilde{B}_{k,t} \phi_k \right\|_\infty,\end{aligned}$$

where $b_{k,t}(Z_{N_i^k}) = \alpha_k \bar{d}_k(Z_{N_i^k}) \prod_{\ell=k+1}^{t-1}(1 - \alpha_\ell \bar{d}_\ell(Z_{N_i^k}))$ is the $Z_{N_i^k}$ th diagonal entry of $B_{k,t}$.

From Lemma 5, 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 $\hat{Q}_i^t$, ϵ_t and ϕ_t in Lemma 6. Because 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 7. The proofs of the two lemmas are postponed to Appendix C.3.

Lemma 6. We have almost surely, (a) $\|\hat{Q}_i^t\|_\infty \leq \frac{\bar{r}}{1-\gamma}$; (b) $\|\epsilon_t\|_\infty \leq \bar{\epsilon} = 4\frac{\bar{r}}{1-\gamma} + 2\bar{r}$; and (c) $\|\phi_t\|_\infty \leq 2\bar{\epsilon} \sum_{k=t-\tau+1}^{t-1} \alpha_k$.

Lemma 7. If $\alpha_t = \frac{h}{t+t_0}$, where $t_0 \geq h > \frac{2}{\sigma}$ and $t_0 \geq 4\sigma h$, and $t_0 \geq \tau$, then $\beta_{k,t}, \tilde{\beta}_{k,t}$ satisfies (a) $\beta_{k,t} \leq \frac{h}{k+t_0} \left(\frac{k+1+t_0}{t+t_0} \right)^{\sigma h}$,

$$\tilde{\beta}_{k,t} \leq \left(\frac{k+1+t_0}{t+t_0}\right)^{\sigma h}; \text{ (b) } \sum_{k=1}^{t-1} \beta_{k,t}^2 \leq \frac{2h}{\sigma} \frac{1}{(t+t_0)^{\gamma}} \text{ and (c) } \sum_{k=\tau}^{t-1} \beta_{k,t} \sum_{\ell=k-\tau+1}^{k-1} \alpha_{\ell} \leq \frac{8h\tau}{\sigma} \frac{1}{t+t_0}.$$

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

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

where the last equality is because of 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 8. 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 8 requires more than the direct application of the Azuma-Hoeffding bound. For more details, see the full proof of Lemma 8 in Appendix C.4.

Lemma 8. *We have with probability $1 - \delta$:*

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

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 because of the step size selection. The proof of Lemma 9 can also be found in Appendix C.4.

Lemma 9. *We have almost surely, $\|\sum_{k=\tau}^{t-1} \alpha_k \tilde{B}_{k,t} \phi_k\|_{\infty} \leq \frac{16\bar{\epsilon}h\tau}{\sigma} \frac{1}{t+t_0} := C_{\phi} \frac{1}{t+t_0}$.*

Step 3. Bound the Critic Error. We are now ready to use the error decomposition in Lemma 5 and the bounds on the ϵ_k, ϕ_k -sequences in Lemmas 8 and 9 to bound the error of the critic. Recall that Theorem 2 states with probability $1 - \delta$,

$$\xi_{\tau} \leq \frac{C_a}{\sqrt{T+t_0}} + \frac{C'_a}{T+t_0} + \frac{C_0}{1-\gamma}, \quad (22)$$

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

$$C_a = \frac{6\bar{\epsilon}}{1-\sqrt{\gamma}} \sqrt{\frac{\tau h}{\sigma} \left[\log\left(\frac{2\tau T^2}{\delta}\right) + f(\kappa) \log SA \right]},$$

$$C'_a = \frac{2}{1-\sqrt{\gamma}} \max\left(\frac{16\bar{\epsilon}h\tau}{\sigma}, \frac{2\bar{r}}{1-\gamma}(\tau+t_0)\right).$$

To prove (22), we start by applying Lemma 8 to $t \leq T$ with δ replaced by δ/T . Then, using a union bound, we get with probability $1 - \delta$, for any $t \leq T$, $\|\sum_{k=\tau}^{t-1} \alpha_k \tilde{B}_{k,t} \epsilon_k\|_{\infty} \leq C_{\epsilon} \frac{1}{\sqrt{t+t_0}}$ with $C_{\epsilon} = 6\bar{\epsilon} \sqrt{\frac{\tau h}{\sigma} \left[\log\left(\frac{2\tau T^2}{\delta}\right) + f(\kappa) \log SA \right]}$.

Combining this with Lemma 5 and using Lemma 9, we get with probability $1 - \delta$, for all $\tau \leq t \leq T$:

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

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

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

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

Lemma 10. *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 10, and using the bound on $\tilde{\beta}_{\tau-1,t}$ in Lemma 7(a), we have

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

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

$$F_t \frac{\sqrt{t+t_0}}{C_a} = \sqrt{\gamma} + \frac{C_e}{C_a},$$

$$F'_t \frac{t+t_0}{C'_a} = \sqrt{\gamma} + \frac{C_\phi}{C'_a} + \frac{\xi_\tau(\tau+t_0)(\tau+t_0)^{\sigma h-1}}{C'_a(t+t_0)^{\sigma h-1}}.$$

Therefore, we can require C_a, C'_a to be large enough such that $\frac{C_e}{C_a} \leq 1 - \sqrt{\gamma}$, $\frac{C_\phi}{C'_a} \leq \frac{1-\sqrt{\gamma}}{2}$, 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 previous three inequalities, and therefore the induction is finished and the proof of Theorem 2 is concluded. $\square$

4.3. Analysis of the Actor and Proof of the Main Result (Theorem 1)

With the error of the critic bounded in Theorem 2, the second part of the proof focuses on the actor, that is, 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

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

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

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

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:

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

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

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 Assumptions 1 and 4, whose proof is postponed to Appendix D.1.

Lemma 11. *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:

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

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 1).

4.2.1. Bounds on $e^1(m)$. 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 2, 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 12. The proof of Lemma 12 is postponed to Appendix D.2.

Lemma 12. *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*

$$C_a(\delta, T) = \frac{6\bar{e}}{1-\sqrt{\gamma}} \sqrt{\frac{\tau h}{\sigma} \left[\log \left(\frac{2\tau T^2}{\delta} \right) + f(\kappa) \log SA \right]},$$

$$C'_a = \frac{2}{1-\sqrt{\gamma}} \max \left(\frac{16\bar{e}h\tau}{\sigma}, \frac{2\bar{r}}{1-\gamma} (\tau + t_0) \right),$$

with $\bar{e} = 4\frac{\bar{r}}{1-\gamma} + 2\bar{r}$, then we have with probability at least

$$1 - \frac{\delta}{2}, \sup_{0 \leq m \leq M-1} \|e^1(m)\| \leq \frac{4cL\rho^{\kappa+1}}{(1-\gamma)^3}.$$

4.2.2. 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 $h_i(m)$. Furthermore, 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 with respect to (w.r.t.) $\mathcal{G}_m$, and we have the following bound in Lemma 13 that is a direct consequence of the Azuma-Hoeffding bound. The proof of Lemma 13 is postponed to Appendix D.3.

Lemma 13. *With probability at least $1 - \delta/2$, we have*

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

4.2.3. 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 14 to bound $\|e^3(m)\|$. Because of 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 cannot be directly used in the proof of Lemma 14. Nevertheless, the proof of Lemma 14 follows essentially the same arguments as in Lemma 3 and is postponed to Appendix D.4.

Lemma 14. *When $T + 1 \geq \frac{\log(\frac{d(1-\gamma)}{\epsilon}) + (\kappa+1)\log \rho}{\log \gamma}$, we have almost surely, $\|e^3(m)\| \leq 2 \frac{Lc}{(1-\gamma)} \rho^{\kappa+1}$.*

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

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

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

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

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

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

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

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

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

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

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

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

We now apply our bounds on $e^1(m)$, $e^2(m)$, $e^3(m)$. By Lemma 13, we have with probability $1 - \frac{\delta}{2}$:

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

By Lemmas 12 and 14, we have with probability $1 - \frac{\delta}{2}$:

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

By Lemma 11, we have almost surely, $\max(\|e^1(m)\|, \|e^2(m)\|, \|e^3(m)\|) \leq 2 \frac{\bar{r}L}{(1-\gamma)^2}$, and hence,

$$\begin{aligned} \sup_{m \leq M-1} \varepsilon_{m,2} &= 2L' (\|e^1(m)\|^2 + \|e^2(m)\|^2 + \|e^3(m)\|^2) \\ &\leq \frac{24\bar{r}^2 L'^2}{(1-\gamma)^4}. \end{aligned} \quad (35)$$

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

combined with (32) implies

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

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

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

This concludes the proof of Theorem 1. $\square$

5. Numerical Studies

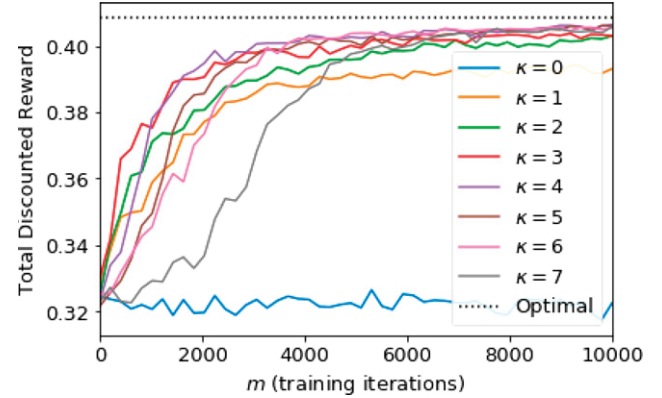
In this section, we first conduct numerical studies in Section 5.1 using a synthetic example where the optimal solution is known in closed form to verify our theoretic results. Then, in Section 5.2, we demonstrate our approach using the wireless communication example introduced in Section 2.2.

5.1. Synthetic Experiments

We first study a synthetic example where the interaction graph is a line of n nodes $\mathcal{N} = \{1, 2, \dots, n\}$ with the left most node labeled as one 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 of one whenever $s_1 = 1$, and all other reward values of node 1 and the rewards of all other nodes are zero. Furthermore, $s_1(t+1) = 1$ with probability of one when the second node has state $s_2(t) = 1$, and in all other cases, $s_1(t+1) = 0$ with probability of one. For the i th node with $2 \leq i \leq n-1$:

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

Figure 1. (Color online) Global Discounted Reward During the Training Process in the Synthetic Experiments



For the last node $i = n$, $s_n(t+1) = 1$ with probability (w.p.) one when $a_n(t) = 1$, and $s_n(t+1) = 0$ w.p. one 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 nonzero 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. 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 $\frac{1}{n(1-\gamma)}$.

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 and Barto 1998). We run the SAC algorithm with $\kappa = 0$ to $\kappa = 7$. We plot the global discounted reward throughout the training process in Figure 1, and we also plot the

Figure 2. (Color online) Optimality Gap as Function of κ in the Synthetic Experiments

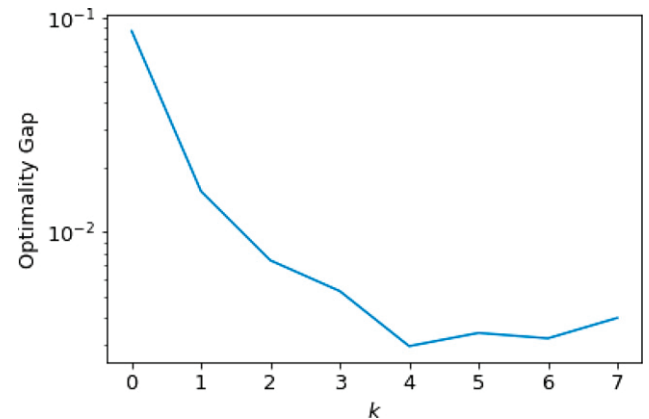
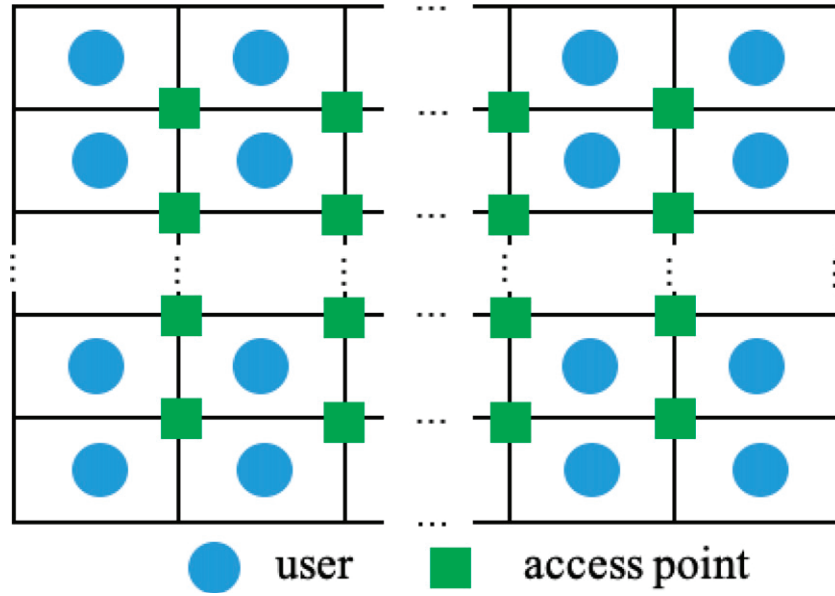


Figure 3. (Color online) Grid of Users and Access Points



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, although 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 1. The exponential decay appears to stop in Figure 2 when $\kappa > 4$, and this is because of the increasing complexity in training for large κ as discussed in Section 3.3 and may also be because any further potential improvement may be too small (in the order of 1×10^{-3}) to be noticed because of the noise.

5.2. Multiaccess Wireless Communication

We next study the multiaccess 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 and 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 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.

Figure 4. (Color online) Global Discounted Reward During Training

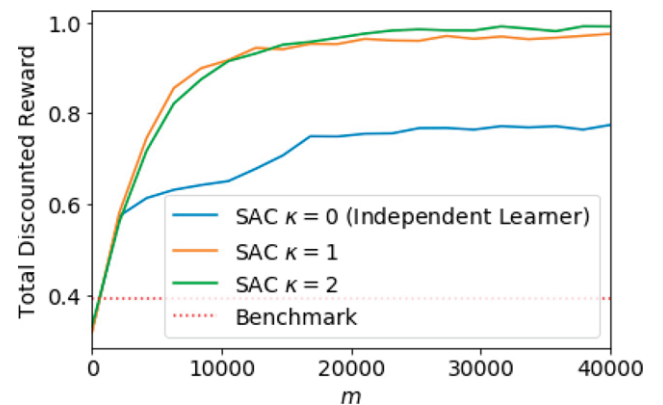
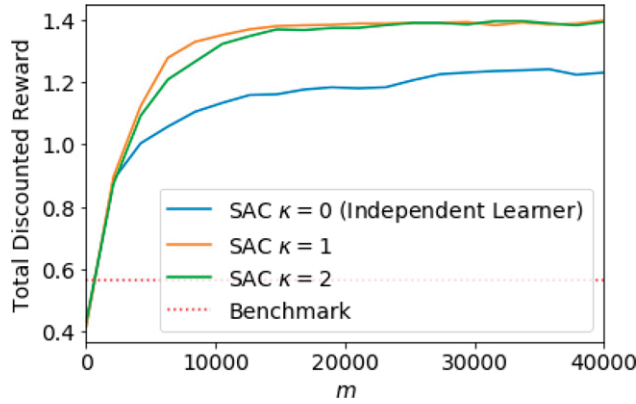


Figure 5. (Color online) Global Discounted Reward During Training When User Locations Are Assigned Randomly

We plot the results in Figure 5. Similar phenomena can be observed in Figure 5 as in Figure 4, where 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 multiagent networked systems with such a provable guarantee. There are many possible extensions and open questions, which we discuss here.

6.1. Average Reward

One extension is to consider average reward instead of discounted reward, that is, to consider

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

Under appropriate ergodicity assumptions, the average reward here is equivalent to the reward under the stationary distribution. Average reward is common in applications where the performance is measured in stationarity, for example, 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 (Tsitsiklis and Van Roy 1999, 2002). For example, the Q -function needs to be defined in a different way:

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

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 multiagent RL setting with the same local interaction structure (1) as in this paper. Although 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.

6.2. Time-Varying Dependence Structure

In this paper, the interaction graph that defines the dependence structure (1) is a fixed graph, whereas 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

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

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 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.

6.3. Other Future Directions

Although 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. Furthermore, our proof technique in showing the finite time convergence of TD learning can be of independent interest. In fact, we 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.

Acknowledgments

The authors thank Yiheng Lin of Caltech and Prof. Longbo Huang of Tsinghua University for suggesting the wireless communication example.

Appendix A. 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 $s_{N_i^k}, s_{N_{-i}^k}$ and $s'_{N_i^k}, s'_{N_{-i}^k}$, $a_{N_i^k}, a_{N_{-i}^k}$ and $a'_{N_i^k}, a'_{N_{-i}^k}$,

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

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 $s = (s_{N_i^k}, s_{N_{-i}^k})$, $a = (a_{N_i^k}, a_{N_{-i}^k})$; $s' = (s'_{N_i^k}, s'_{N_{-i}^k})$; and $a' = (a'_{N_i^k}, a'_{N_{-i}^k})$. 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 $t \leq \kappa$. The reason is that, because of 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 $(s'_{N_i^t}, a'_{N_i^t})$ when $t \leq \kappa$ 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):

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

where $\text{TV}(\pi_{t,i}, \pi'_{t,i})$ is the total variation distance between $\pi_{t,i}$ and $\pi'_{t,i}$ which is upper bounded by one. The previous inequality shows that the $\left(\frac{\bar{r}}{1-\gamma}, \gamma\right)$ -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 (A.1), we use $\text{TV}(\pi_{t,i}, \pi'_{t,i}) \leq 2c'\mu^t$. $\square$

Appendix 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):

$$\begin{aligned} & |\hat{Q}_i^\theta(s_{N_i^k}, a_{N_i^k}) - Q_i^\theta(s, a)| \\ & = \left| \sum_{s'_{N_{-i}^k}, a'_{N_{-i}^k}} w_i(s'_{N_{-i}^k}, a'_{N_{-i}^k}; s_{N_i^k}, a_{N_i^k}) Q_i^\theta(s_{N_i^k}, s'_{N_{-i}^k}, a_{N_i^k}, a'_{N_{-i}^k}) \right. \\ & \quad \left. - Q_i^\theta(s_{N_i^k}, s_{N_{-i}^k}, a_{N_i^k}, a_{N_{-i}^k}) \right| \\ & \leq \sum_{s'_{N_{-i}^k}, a'_{N_{-i}^k}} w_i(s'_{N_{-i}^k}, a'_{N_{-i}^k}; s_{N_i^k}, a_{N_i^k}) |Q_i^\theta(s_{N_i^k}, s'_{N_{-i}^k}, a_{N_i^k}, a'_{N_{-i}^k}) \\ & \quad - Q_i^\theta(s_{N_i^k}, s_{N_{-i}^k}, a_{N_i^k}, a_{N_{-i}^k})| \\ & \leq c\rho^{\kappa+1}, \end{aligned} \quad (\text{B.1})$$

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):

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

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 previous equation, we can compute $\hat{h}_i(\theta) - \nabla_{\theta_i} J(\theta)$:

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

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

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

where in the last equality, $\hat{Q}_j^\theta(s_{N_j^k}, a_{N_j^k})$ does not depend on a_i as $i \notin N_j^k$; 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:

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

where in the last step, we used (B.1) 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. $\square$

Appendix C. Proof of Auxiliary Results for the Analysis of the Critic

C.1. Proof of Lemma 4

The goal of Lemma 4 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 .

First, $\bar{A}_z$ is given by

$$\bar{A}_z = \mathbb{E}_{z' \sim P(\cdot|z)} A_{z, z'} = \mathbf{e}_{z_{N_i^k}} \left[\gamma P(\cdot | z) \Phi - \mathbf{e}_{z_{N_i^k}}^T \right], \quad (\text{C.1})$$

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

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

$$\begin{aligned} \bar{A}^d &= \mathbb{E}_{z \sim d} \bar{A}_z = \sum_{z \in \mathcal{Z}} d(z) \mathbf{e}_{z_{N_i^k}} \left[\gamma P(\cdot | z) \Phi - \mathbf{e}_{z_{N_i^k}}^T \right] \\ &= \Phi^T \text{diag}(d) [\gamma P \Phi - \Phi], \end{aligned} \quad (\text{C.2})$$

$$\bar{b}^d = \mathbb{E}_{z \sim d} b_z = \Phi^T \text{diag}(d) r_i, \quad (\text{C.3})$$

where $\text{diag}(d) \in \mathbb{R}^{Z \times Z}$ 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 $\mathcal{Z}$, although it only depends on z_i .

With these characterizations, we show the following property of $\bar{A}^d$ and $\bar{b}^d$ in Lemma C.1. Lemma C.1 (with d set as d_{t-1}) will directly lead to the results in Lemma 4, with D_{t-1} being the D in Lemma C.1, and it satisfies $D_{t-1} \succeq 0I$ because of Assumption 3; g_{t-1} and $\hat{Q}_i^{*,t-1}$ are $g^{d_{t-1}}$ and $\hat{Q}_i^{d_{t-1}}$ in Lemma C.1.

Lemma C.1. Given distribution d on state-action pair z whose marginalization onto $z_{N_i^k}$ is nonzero for every $z_{N_i^k}$, we have $\bar{A}^d \hat{Q}_i + \bar{b}^d$ can be written as

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

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

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

$$\sup_{z \in \mathcal{Z}} \left| \hat{Q}_i^d(z_{N_i^k}) - Q_i^*(z) \right| = \|\Phi \hat{Q}_i^d - Q_i^*\|_\infty \leq \frac{c \rho^{\kappa+1}}{1-\gamma}. \quad (\text{C.4})$$

Proof of Lemma C.1. It is easy to check that $D = \Phi^T \text{diag}(d) \Phi \in \mathbb{R}^{Z_{N_i^k} \times Z_{N_i^k}}$ is a diagonal matrix, and the $z_{N_i^k}$ th diagonal entry is the marginal probability of $z_{N_i^k}$ under d , which is nonzero by the assumption of the lemma. Therefore, $\Phi^T \text{diag}(d) \Phi \in \mathbb{R}^{Z_{N_i^k} \times Z_{N_i^k}}$ is invertible, and matrix $\Pi^d = (\Phi^T \text{diag}(d) \Phi)^{-1} \Phi^T \text{diag}(d)$ is well defined. Furthermore, the $z_{N_i^k}$ th row of Π^d is in fact the conditional distribution of the full state z given $z_{N_i^k}$. Therefore, Π^d must be a stochastic matrix and is nonexpansive in infinity norm.

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

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

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

Notice that Φ is nonexpansive in $\|\cdot\|_\infty$ norm because each row of Φ has precisely one entry being one and all

others are zero. Also, because Π^d is nonexpansive in $\|\cdot\|_\infty$ norm and TD is a γ -contraction in $\|\cdot\|_\infty$ norm, we have $g^d = \Pi^d \text{TD} \Phi$ 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 (C.4), 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

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

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

$$\|\Phi \hat{Q}_i^d - Q_i^*\|_\infty \leq \frac{1}{1-\gamma} \|\Phi \Pi^d Q_i^* - Q_i^*\|_\infty. \quad (\text{C.5})$$

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

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

Therefore,

$$(\Phi \Pi^d Q_i^*)(z_{N_i^k}, z_{N_i^k}) = \sum_{z_{N_i^k}} d(z_{N_i^k} | z_{N_i^k}) Q_i^*(z_{N_i^k}, z_{N_i^k}).$$

Furthermore, we have

$$\begin{aligned} &|(\Phi \Pi^d Q_i^*)(z_{N_i^k}, z_{N_i^k}) - Q_i^*(z_{N_i^k}, z_{N_i^k})| \\ &= \left| \sum_{z_{N_i^k}} d(z_{N_i^k} | z_{N_i^k}) Q_i^*(z_{N_i^k}, z_{N_i^k}) - \sum_{z_{N_i^k}} d(z_{N_i^k} | z_{N_i^k}) Q_i^*(z_{N_i^k}, z_{N_i^k}) \right| \\ &\leq \sum_{z_{N_i^k}} d(z_{N_i^k} | z_{N_i^k}) |Q_i^*(z_{N_i^k}, z_{N_i^k}) - Q_i^*(z_{N_i^k}, z_{N_i^k})| \\ &\leq c \rho^{\kappa+1}, \end{aligned}$$

where the last inequality is because of 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 this with (C.5), we get the desired result:

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

C.2. Proof of Lemma 5

Recall that the $z_{N_i^k}$ th diagonal entry of $B_{k,t}$ is $b_{k,t}(z_{N_i^k})$, and we define similarly the $z_{N_i^k}$ th diagonal entry of $\tilde{B}_{k,t}$ to

be $\tilde{b}_{k,t}(z_{N_i^k})$. Using these notations, Equation (20) can be written as

$$\begin{aligned} \hat{Q}_i^t(z_{N_i^k}) &= \tilde{b}_{\tau-1,t}(z_{N_i^k}) \hat{Q}_i^\tau(z_{N_i^k}) + \sum_{k=\tau}^{t-1} b_{k,t}(z_{N_i^k}) [g_k(\hat{Q}_i^k)](z_{N_i^k}) \\ &\quad + \sum_{k=\tau}^{t-1} \alpha_k \tilde{b}_{k,t}(z_{N_i^k}) (\epsilon_k(z_{N_i^k}) + \phi_k(z_{N_i^k})). \end{aligned} \quad (\text{C.6})$$

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

$$\begin{aligned} &|G(z_{N_i^k}) - Q_i^*(z)| \\ &\leq \tilde{b}_{\tau-1,t}(z_{N_i^k}) |\hat{Q}_i^\tau(z_{N_i^k}) - Q_i^*(z)| \\ &\quad + \sum_{k=\tau}^{t-1} b_{k,t}(z_{N_i^k}) |[g_k(\hat{Q}_i^k)](z_{N_i^k}) - Q_i^*(z)| \\ &\leq \tilde{b}_{\tau-1,t}(z_{N_i^k}) |\hat{Q}_i^\tau(z_{N_i^k}) - Q_i^*(z)| \\ &\quad + \sum_{k=\tau}^{t-1} b_{k,t}(z_{N_i^k}) |[g_k(\hat{Q}_i^k)](z_{N_i^k}) - \hat{Q}_i^{*,k}(z_{N_i^k})| \\ &\quad + \sum_{k=\tau}^{t-1} b_{k,t}(z_{N_i^k}) |Q_i^*(z) - \hat{Q}_i^{*,k}(z_{N_i^k})| \\ &\leq \tilde{b}_{\tau-1,t}(z_{N_i^k}) |\hat{Q}_i^\tau(z_{N_i^k}) - Q_i^*(z)| + \gamma \sum_{k=\tau}^{t-1} b_{k,t}(z_{N_i^k}) \|\hat{Q}_i^k - \hat{Q}_i^{*,k}\|_\infty \\ &\quad + \sum_{k=\tau}^{t-1} b_{k,t}(z_{N_i^k}) \|Q_i^* - \Phi \hat{Q}_i^{*,k}\|_\infty \\ &\leq \tilde{b}_{\tau-1,t}(z_{N_i^k}) |\hat{Q}_i^\tau(z_{N_i^k}) - Q_i^*(z)| + \gamma \sum_{k=\tau}^{t-1} b_{k,t}(z_{N_i^k}) \|\Phi \hat{Q}_i^k - Q_i^*\|_\infty \\ &\quad + 2 \sum_{k=\tau}^{t-1} b_{k,t}(z_{N_i^k}) \|Q_i^* - \Phi \hat{Q}_i^{*,k}\|_\infty \\ &\leq \tilde{\beta}_{\tau-1,t} \xi_\tau + \gamma \sum_{k=\tau}^{t-1} b_{k,t}(z_{N_i^k}) \xi_k + \frac{2c \rho^{\kappa+1}}{1-\gamma}, \end{aligned} \quad (\text{C.7})$$

where in the third inequality, we used that g_k is γ -contraction in infinity norm with fixed point $\hat{Q}_i^{*,k}$, and in the last inequality, we used the property of $\hat{Q}_i^{*,k}$ in Lemma 4. Combining this with (C.6), we have

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

C.3. Proof of Lemmas 6 and 7

In this section, we provide proofs of the two auxiliary lemmas: Lemmas 6 and 7. We start with the proof of Lemma 6.

Proof of Lemma 6. First, notice that $A_{z,z'} = \mathbf{e}_{z_{N_i^k}} [\gamma \mathbf{e}_{z_{N_i^k}}^T - \mathbf{e}_{z_{N_i^k}}^T]$ and $b_z = \mathbf{e}_{z_{N_i^k}} r_i(z_i)$. As such, $\|A_{z,z'}\|_\infty \leq 1 + \gamma < 2$, $\|b_z\|_\infty \leq \tilde{r}$.

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

$$\begin{aligned}\hat{Q}_i^t &= \hat{Q}_i^{t-1} + \alpha_{t-1}[r_i(z_i(t-1)) + \gamma\hat{Q}_i^{t-1}(z_{N_i^c}(t)) \\ &\quad - \hat{Q}_i^{t-1}(z_{N_i^c}(t-1))]\mathbf{e}_{z_{N_i^c}(t-1)},\end{aligned}$$

or in other words,

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

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

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

Notice that

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

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), $\|\epsilon_t\|_\infty \leq 4\frac{\bar{r}}{1-\gamma} + 2\bar{r} = \bar{\epsilon}$.

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

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

Therefore, by triangle inequality:

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

As a consequence,

$$\begin{aligned}\|\phi_t\|_\infty &\leq \|A_t - \bar{A}_t\|_\infty \|\hat{Q}_i^t - \hat{Q}_i^{t-\tau+1}\|_\infty \leq \left[8\frac{\bar{r}}{1-\gamma} + 4\bar{r}\right] \sum_{k=t-\tau+1}^{t-1} \alpha_k \\ &= 2\bar{\epsilon} \sum_{k=t-\tau+1}^{t-1} \alpha_k. \quad \square\end{aligned}$$

Proof of Lemma 7. Notice that $\log(1-x) \leq -x$ for all $x < 1$. Then,

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

Therefore,

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

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

For part (b):

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

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

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

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

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

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

where we used $(k+1+t_0)^{\sigma h} \leq 2(k+t_0)^{\sigma h}$ and $\sigma h - 1 > \frac{1}{2}\sigma h$. $\square$

C.4. Proof of Lemmas 8 and 9

Because $\|\sum_{k=\tau}^{t-1} \alpha_k \tilde{B}_{k,t} \phi_k\|_\infty \leq \sum_{k=\tau}^{t-1} \beta_{k,t} \|\phi_k\|_\infty$, Lemmas 6(c) and 7(c) directly lead to the bound in Lemma 9. Therefore, in this section, we focus on the proof of Lemma 8. We start by stating a variant of the Azuma-Hoeffding bound that handles our “shifted” Martingale difference sequence.

Lemma C.2. Let X_t be a $\mathcal{F}_t$ -adapted stochastic process, satisfying $\mathbb{E}X_t | \mathcal{F}_{t-\tau} = 0$. Furthermore, $|X_t| \leq \bar{X}_t$ almost surely. Then with probability $1 - \delta$, we have

$$\left| \sum_{k=0}^t X_k \right| \leq \sqrt{2\tau \sum_{k=0}^t \bar{X}_k^2 \log\left(\frac{2\tau}{\delta}\right)}.$$

Proof of Lemma C.2. Let ℓ be an integer between zero and $\tau - 1$. For each ℓ , define process $Y_k^\ell = X_{k\tau+\ell}$, scalar $\tilde{Y}_k^\ell = \tilde{X}_{k\tau+\ell}$, and define filtration $\tilde{\mathcal{F}}_k^\ell = \mathcal{F}_{k\tau+\ell}$. 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 the Azuma-Hoeffding bound on Y_k^ℓ , we have

$$P\left(\left|\sum_{k:k\tau+\ell \leq t} Y_k^\ell\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\sum_{k:k\tau+\ell \leq t} (\tilde{Y}_k^\ell)^2}\right),$$

that is, with probability at least $1 - \frac{\delta}{\tau}$,

$$\left|\sum_{k:k\tau+\ell \leq t} X_{k\tau+\ell}\right| = \left|\sum_{k:k\tau+\ell \leq t} Y_k^\ell\right| \leq \sqrt{2\sum_{k:k\tau+\ell \leq t} \tilde{X}_{k\tau+\ell}^2 \log\left(\frac{2\tau}{\delta}\right)}.$$

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

$$\begin{aligned} \left|\sum_{k=0}^t X_k\right| &\leq \sum_{\ell=0}^{\tau-1} \left|\sum_{k:k\tau+\ell \leq t} X_{k\tau+\ell}\right| \leq \sum_{\ell=0}^{\tau-1} \sqrt{2\sum_{k:k\tau+\ell \leq t} \tilde{X}_{k\tau+\ell}^2 \log\left(\frac{2\tau}{\delta}\right)} \\ &\leq \sqrt{2\tau \sum_{k=0}^t \tilde{X}_k^2 \log\left(\frac{2\tau}{\delta}\right)}, \end{aligned}$$

where the last inequality is because of Cauchy-Schwarz. $\square$

Recall that Lemma 8 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 $\mathbb{R}^{\mathcal{Z}_{N_i^c}}$, with its $\mathcal{Z}_{N_i^c}$ th entry being

$$\sum_{k=\tau}^{t-1} \alpha_k \epsilon_k(z_{N_i^c}) \prod_{\ell=k+1}^{t-1} (1 - \alpha_\ell \bar{d}_\ell(z_{N_i^c})), \quad (\text{C.8})$$

with $\bar{d}_\ell(z_{N_i^c}) \geq \sigma$ almost surely (cf. Lemma 4). Fixing $z_{N_i^c}$, as shown in (21), $\epsilon_k(z_{N_i^c})$ is a $\mathcal{F}_{k+1}$ adapted stochastic process satisfying $\mathbb{E}\epsilon_k(z_{N_i^c}) | \mathcal{F}_{k+1-\tau} = 0$. However, $\prod_{\ell=k+1}^{t-1} (1 - \alpha_\ell \bar{d}_\ell(z_{N_i^c}))$ is not $\mathcal{F}_{k+1-\tau}$ -measurable, and as such, we cannot directly apply the Azuma-Hoeffding bound in Lemma C.2 to quantity (C.8). In what follows, we first show in Lemma C.3 that, almost surely, the absolute value of quantity (C.8) can be upper bounded by the supremum of another quantity, to which we can directly apply Lemma C.2. With the help of Lemma C.3, we can use the Azuma-Hoeffding bound to control (C.8) and prove Lemma 8.

Lemma C.3. For each $z_{N_i^c}$, we have almost surely,

$$\begin{aligned} &\left|\sum_{k=\tau}^{t-1} \alpha_k \epsilon_k(z_{N_i^c}) \prod_{\ell=k+1}^{t-1} (1 - \alpha_\ell \bar{d}_\ell(z_{N_i^c}))\right| \\ &\leq \sup_{\tau \leq k_0 \leq t-1} \left(\left|\sum_{k=k_0+1}^{t-1} \epsilon_k(z_{N_i^c}) \beta_{k,t}\right| + 2\bar{\epsilon} \beta_{k_0,t}\right). \end{aligned}$$

Proof of Lemma C.3. Let p_k be a scalar sequence defined as follows. Set $p_\tau = 0$, and

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

Then $p_t = \sum_{k=\tau}^{t-1} \alpha_k \epsilon_k(z_{N_i^c}) \prod_{\ell=k+1}^{t-1} (1 - \alpha_\ell \bar{d}_\ell(z_{N_i^c}))$, and to prove Lemma C.3, we need to bound $|p_t|$. Let

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

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

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

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 $k - 1$. Without loss of generality, suppose both p_{k-1} and $\tilde{p}_{k-1}$ are non-negative. Because $k - 1 > k_0$ and by the definition of k_0 , we must have

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

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

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

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

Notice,

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

By the definition of k_0 , we have

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

where in the last step, we used the upper bound on $\|\epsilon_{k_0}\|_\infty$ in Lemma 6(b). As a result,

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

With these preparations, we are now ready to prove Lemma 8.

Proof of Lemma 8. Fix $z_{N_i^c}$ and $\tau \leq k_0 \leq t-1$. As shown in (21), $\epsilon_k(z_{N_i^c}) \beta_{k,t}$ is a $\mathcal{F}_{k+1}$ adapted stochastic process satisfying $\mathbb{E}\epsilon_k(z_{N_i^c}) \beta_{k,t} | \mathcal{F}_{k+1-\tau} = 0$. Also by Lemma 6(b), $|\epsilon_k(z_{N_i^c}) \beta_{k,t}| \leq \bar{\epsilon} \beta_{k,t}$ almost surely. As a result, we can use the Azuma-Hoeffding bound in Lemma C.2 to get, with probability $1 - \delta$:

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

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

$$\begin{aligned} \sup_{\tau \leq k_0 \leq t-1} \left| \sum_{k=\tau+1}^{t-1} \epsilon_k(Z_{N_i^k}) \beta_{k,t} \right| &\leq \sup_{\tau \leq k_0 \leq t-1} \bar{\epsilon} \sqrt{2\tau \sum_{k=k_0+1}^{t-1} \beta_{k,t}^2 \log\left(\frac{2\tau t}{\delta}\right)} \\ &\leq \bar{\epsilon} \sqrt{2\tau \sum_{k=\tau+1}^{t-1} \beta_{k,t}^2 \log\left(\frac{2\tau t}{\delta}\right)}. \end{aligned}$$

Then, by Lemma C.3, we have, with probability $1 - \delta$:

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

where in the third inequality, we used the bounds on $\beta_{k,t}$ in Lemma 7. Finally, apply the union bound over $Z_{N_i^k} \in \mathcal{Z}_{N_i^k}$, and noticing that $|N_i^k| \leq f(\kappa)$ and $|Z_{N_i^k}| \leq (SA)^{f(\kappa)}$ by Assumption 1, we have, with probability $1 - \delta$:

$$\begin{aligned} \left\| \sum_{k=\tau}^{t-1} \alpha_k \tilde{B}_{k,t} \epsilon_k \right\|_\infty &\leq 6\bar{\epsilon} \sqrt{\frac{\tau h}{\sigma(t+t_0)} \log\left(\frac{2\tau t(SA)^{f(\kappa)}}{\delta}\right)} \\ &= 6\bar{\epsilon} \sqrt{\frac{\tau h}{\sigma(t+t_0)} \left[\log\left(\frac{2\tau t}{\delta}\right) + f(\kappa) \log SA \right]}. \quad \square \end{aligned}$$

C.5. Proof of Lemma 10

Throughout the proof, we fix $Z_{N_i^k}$ and prove the desired upper bound. For notational simplicity, we drop the dependence on $Z_{N_i^k}$ and write $b_{k,t}$ and $\bar{d}_k$ instead, and we will use the property $\bar{d}_k \geq \sigma$. 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 $e_t \leq \frac{1}{\sqrt{\gamma}(t+t_0)^\omega}$. The statement is clearly true for $t = \tau + 1$, as $e_{\tau+1} = b_{\tau,\tau+1} \frac{1}{(\tau+t_0)^\omega} = \alpha_\tau \bar{d}_\tau \frac{1}{(\tau+t_0)^\omega} \leq \frac{1}{\sqrt{\gamma}(\tau+1+t_0)^\omega}$ (last step needs $\alpha_\tau \leq \frac{1}{2}, \left(1 + \frac{1}{t_0}\right)^\omega \leq \frac{2}{\sqrt{\gamma}}$, implied by

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

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

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

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

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

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

where in the last inequality, we 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.

Appendix D. Proof of Auxiliary Results for Analysis of Actor

D.1. Proof of Lemma 11

Recall that

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

Therefore,

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

where we have used that $\|\hat{Q}_j^{m,T}\|_\infty \leq \frac{\bar{r}}{1-\gamma}$ almost surely (cf. Lemma 6(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 12

Let $\mathcal{G}_m$ be the σ -algebra generated by the trajectories in the first m outer-loop iterations. Then, Theorem 2 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 $1 - \delta$:

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

where

$$C_a(\delta, T) = \frac{6\bar{\epsilon}}{1-\sqrt{\gamma}} \sqrt{\frac{\tau h}{\sigma} \left[\log\left(\frac{2\tau T^2}{\delta}\right) + f(\kappa) \log SA \right]},$$

$$C'_a = \frac{2}{1-\sqrt{\gamma}} \max\left(\frac{16\bar{\epsilon}h\tau}{\sigma}, \frac{2\bar{r}}{1-\gamma}(\tau+t_0)\right),$$

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 $1 - \frac{\delta}{2}$,

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

where in the last step, we 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 (D.1) being true, we have for any $m \leq M-1$ and any $i \in \mathcal{N}$:

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

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 (D.1) is true that happens with probability at least $1 - \frac{\delta}{2}$.

D.3. Proof of Lemma 13

By Lemma 11, we have almost surely,

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

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 the Azuma-Hoeffding bound, with probability of at least $1 - \frac{1}{2}\delta$,

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

D.4. Proof of Lemma 14

By (25), we have

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

where we 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):

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

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

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

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

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

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 $j \in N_{-i}^K$ and any t :

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

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

For E_3 , by the exponential decay property, the truncated Q -function has a small error (cf. (B.1)):

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

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 used

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

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

References

- Agarwal A, Kakade SM, Lee JD, Mahajan G (2021) On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *J. Machine Learn. Res.* 22(98):1–76.
- Ahn HJ (2014) Random propagation in complex systems: Nonlinear matrix recursions and epidemic spread. PhD thesis, California Institute of Technology, Pasadena.
- Azizan Ruhi N, Ahn HJ, Hassibi B (2016a) Analysis of exact and approximated epidemic models over complex networks. Preprint, submitted September 30, <https://arxiv.org/abs/1609.09565>.
- Azizan Ruhi N, Thrampoulidis C, Hassibi B (2016b) Improved bounds on the epidemic threshold of exact SIS models on complex networks. *Proc. 55th IEEE Conf. on Decision and Control* (IEEE, New York), 3560–3565.
- Bamieh B, Paganini F, Dahleh MA (2002) Distributed control of spatially invariant systems. *IEEE Trans. Automated Control* 47(7):1091–1107.
- Bertsekas DP (2005) *Dynamic Programming and Optimal Control*. Athena Scientific Optimization and Computation Series, 3rd ed. (Athena Scientific, Belmont, MA).
- Bertsekas DP, Tsitsiklis JN (1996) *Neuro-Dynamic Programming*, vol. 5 (Athena Scientific, Belmont, MA).
- Bhandari J, Russo D (2019) Global optimality guarantees for policy gradient methods. Preprint, submitted June 5, 2019, updated October 29, 2020, <https://arxiv.org/abs/1906.01786>.
- Block R, Van Houdt B (2016) Spatial fairness in multi-channel CSMA line networks. *Performance Evaluation* 103:69–85.
- Blondel VD, Tsitsiklis JN (2000) A survey of computational complexity results in systems and control. *Automatica J. IFAC* 36(9):1249–1274.
- Britton T (2010) Stochastic epidemic models: A survey. *Math. Biosci.* 225(1):24–35.
- Busoniu L, Babuska R, De Schutter B (2008) A comprehensive survey of multiagent reinforcement learning. *IEEE Trans. Systems Man Cybernetics C* 38(2):156–172.
- Chakrabarti D, Wang Y, Wang C, Leskovec J, Faloutsos C (2008) Epidemic thresholds in real networks. *ACM Trans. Inform. Systems Security*. 10(4):1.
- Claus C, Boutillier C (1998) The dynamics of reinforcement learning in cooperative multiagent systems. *AAAI Conf. Artificial Intelligence, Madison, WI (AAAI, Palo Alto, CA)*, 746–752. <https://www.aaai.org/Library/AAAI/aaai98contents.php>.
- Duan Y, Chen X, Houthooft R, Schulman J, Abbeel P (2016) Benchmarking deep reinforcement learning for continuous control. Balcan MF, Weinberger KQ, eds. *Proc. Internat. Conf. on Machine Learn.* (PMLR, New York), 1329–1338. <http://proceedings.mlr.press/v48/>.
- Gamarnik D (2013) Correlation decay method for decision, optimization, and inference in large-scale networks. *Theory Driven by Influential Applications* (INFORMS), 108–121. <https://pubsonline.informs.org/doi/abs/10.1287/educ.2013.0119>.
- Gamarnik D, Goldberg DA, Weber T (2014) Correlation decay in random decision networks. *Math. Oper. Res.* 39(2):229–261.
- Guestrin C, Koller D, Parr R, Venkataraman S (2003) Efficient solution algorithms for factored MDPs. *J. Artificial Intelligence Res.* 19:399–468.
- Hu J, Wellman MP (2003) Nash Q-learning for general-sum stochastic games. *J. Machine Learn. Res.* 4(11):1039–1069.

- Jin C, Allen-Zhu Z, Bubeck S, Jordan MI (2018) Is Q-learning provably efficient? *Proc. 32nd Internat. Conf. on Neural Information Processing Systems*, 4868–4878. <https://dl.acm.org/doi/proceedings/10.5555/3326943>.
- Kar S, Moura JM, Poor HV (2013) QD-learning: A collaborative distributed strategy for multi-agent reinforcement learning through consensus + innovations. *IEEE Trans. Signal Processing* 61(7):1848–1862.
- Kearns M, Koller D (1999) Efficient reinforcement learning in factored MDPs. *Proc. 16th Internat. Joint Conf. Artificial Intelligence (IJCAI'99)*, vol. 2 (Morgan Kaufmann Publishers Inc., San Francisco), 740–747. <https://dl.acm.org/doi/abs/10.5555/1624312.1624325>.
- Kim TH, Ni J, Srikant R, Vaidya NH (2011) On the achievable throughput of CSMA under imperfect carrier sensing. 2011. *Proc. IEEE INFOCOM (IEEE, New York)*, 1674–1682.
- Konda VR, Tsitsiklis JN (2000) Actor-critic algorithms. *Proc. 12th Internat. Conf. Neural Inform. Processing Systems (NIPS'99)* (MIT Press, Cambridge, MA), 1008–1014.
- Kuznetsov YA, Piccardi C (1994) Bifurcation analysis of periodic SEIR and SIR epidemic models. *J. Math. Biology* 32(2):109–121.
- Li D, Zhao D, Zhang Q, Chen Y (2019) Reinforcement learning and deep learning based lateral control for autonomous driving. *IEEE Comput. Intelligence Magazine* 14(2):83–98.
- Li G, Wei Y, Chi Y, Gu Y, Chen Y (2020) Sample complexity of asynchronous q-learning: Sharper analysis and variance reduction. Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, eds. *Advances in Neural Information Processing Systems*, vol. 33 (Curran Associates, Inc.), 7031–7043. <https://proceedings.neurips.cc/paper/2020/file/4eab60e55fe4c7dd567a0be28016bff3-Paper.pdf>.
- Lin Y, Qu G, Huang L, Wierman A (2021) Multi-agent reinforcement learning in stochastic networked systems. *Adv. Neural Inform. Processing Systems*. <https://proceedings.neurips.cc/paper/2021/hash/412604be30f701b1b1e3124c252065e6-Abstract.html>.
- Littman ML (1994) Markov games as a framework for multi-agent reinforcement learning. *Proc. Machine Learn.* (Elsevier, New York), 157–163.
- Littman ML (2001) Value-function reinforcement learning in Markov games. *Cognitive Systems Res.* 2(1):55–66.
- Llas M, Gleiser PM, López JM, Díaz-Guilera A (2003) Nonequilibrium phase transition in a model for the propagation of innovations among economic agents. *Phys. Rev. E* 68(6):066101.
- Lokhov AY, Mézard M, Zdeborová L (2015) Dynamic message-passing equations for models with unidirectional dynamics. *Phys. Rev. E* 91:012811.
- Lowe R, Wu Y, Tamar A, Harb J, Abbeel OP, Mordatch I (2017) Multi-agent actor-critic for mixed cooperative-competitive environments. Guyon I, Luxburg UV, Bengio S, Wallach H, Fergus R, Vishwanathan S, Garnett R, eds. *Adv. Neural Inform. Processing Systems*, vol. 30 (Curran Associates, Inc.), 6379–6390. <https://proceedings.neurips.cc/paper/2017/file/68a9750337a418a86fe06c1991a1d64c-Paper.pdf>.
- Macua SV, Chen J, Zazo S, Sayed AH (2015) Distributed policy evaluation under multiple behavior strategies. *IEEE Trans. Automated Control* 60(5):1260–1274.
- Magnússon S, Shokri-Ghadikolaei H, Li N (2020) On maintaining linear convergence of distributed learning and optimization under limited communication. *IEEE Trans. Signal Process ing* 68: 6101–6116.
- Mathkar A, Borkar VS (2017) Distributed reinforcement learning via gossip. *IEEE Trans. Automated Control* 62(3):1465–1470.
- Matignon L, Laurent GJ, Le Fort-Piat N (2012) Independent reinforcement learners in cooperative Markov games: A survey regarding coordination problems. *Knowledge Engng. Rev.* 27(1):1–31.
- Mei W, Mohagheghi S, Zampieri S, Bullo F (2017) On the dynamics of deterministic epidemic propagation over networks. *Annu. Rev. Control* 44:116–128.
- Meuleau N, Hauskrecht M, Kim KE, Peshkin L, Kaelbling LP, Dean TL, Boutilier C (1998) Solving very large weakly coupled Markov decision processes. (AAAI, Palo Alto, CA), 165–172.
- Mezard M, Montanari A (2009) *Information, Physics, and Computation* (Oxford University Press, Oxford, UK).
- Mnih V, Kavukcuoglu K, Silver D, Rusu AA, Veness J, Bellemare MG, Graves A, et al. (2015) Human-level control through deep reinforcement learning. *Nature*. 518(7540):529.
- Morris DH, Rossine FW, Plotkin JB, Levin SA (2021) Optimal, near-optimal, and robust epidemic control. *Commun. Phys.* 4:78.
- Nair R, Varakantham P, Tambe M, Yokoo M (2005) Networked distributed POMDPs: A synthesis of distributed constraint optimization and POMDPs. *Proc. 20th Natl. Conf. Artificial Intelligence, AAAI'05, Pittsburgh*, vol. 1 (AAAI Press, Palo Alto, CA), 133–139.
- Oliehoek FA, Amato C (2016) *A Concise Introduction to Decentralized POMDPs* (Springer, Berlin).
- Osband I, Van Roy B (2014) Near-optimal reinforcement learning in factored MDPs. Ghahramani Z, Welling M, Cortes C, Lawrence N, Weinberger KQ, eds. *Advances in Neural Information Processing Systems*, vol. 24 (Curran Associates, Inc.), 604–612. <https://proceedings.neurips.cc/paper/2014/file/0deb1c54814305ca9ad266f53bc82511-Paper.pdf>.
- Papadimitriou CH, Tsitsiklis JN (1999) The complexity of optimal queueing network control. *Math. Oper. Res.* 24(2):293–305.
- Preciado VM, Zargham M, Enyioha C, Jadbabaie A, Pappas G (2013) Optimal vaccine allocation to control epidemic outbreaks in arbitrary networks. *Proc. 52nd IEEE Conf. on Decision and Control (IEEE, New York)*, 7486–7491.
- Qu G, Li N (2019) Exploiting fast decaying and locality in multi-agent MDP with tree dependence structure. *Proc. IEEE 58th Conf. on Decision and Control (IEEE, New York)*, 6479–6486.
- Qu G, Lin Y, Wierman A, Li N (2020a) Scalable multi-agent reinforcement learning for networked systems with average reward. Larochelle H, Ranzato M, Hadsell R, Balcan MF, Lin H, eds. *Advances in Neural Information Processing Systems*, vol. 33 (Curran Associates, Inc.), 2074–2086. <https://proceedings.neurips.cc/paper/2020/file/168efc366c449fab9c2843e9b54e2a18-Paper.pdf>.
- Qu G, Yu C, Low S, Wierman A (2020b) Combining model-based and model-free methods for nonlinear control: A provably convergent policy gradient approach. Preprint, submitted June 12, 2020, <https://arxiv.org/abs/2006.07476>.
- Roberts L (1975) ALOHA packet system with and without slots and capture. *ACM SIGCOMM Comput. Comm. Rev.* 5:28–42.
- Rotkowitz M, Lall S (2005) A characterization of convex problems in decentralized control. *IEEE Trans. Automated Control* 50(12):1984–1996.
- Silver D, Huang A, Maddison CJ, Guez A, Sifre L, Van Den Driessche G, Schrittwieser J, et al. (2016) Mastering the game of go with deep neural networks and tree search. *Nature* 529(7587):484.
- Srikant R, Ying L (2019) Finite-time error bounds for linear stochastic approximation and TD learning. Beygelzimer A, Hsu D, eds. *Proc. 32nd Conf. on Learn. Theory*, Proceedings of Machine Learning Research, vol. 99 (PMLR), 2803–2830. <http://proceedings.mlr.press/v99/srikant19a/srikant19a.pdf>.
- Sutton RS, Barto AG, et al. (1998) *Introduction to Reinforcement Learning*, vol. 135 (MIT Press, Cambridge, MA). <https://mitpress.mit.edu/books/reinforcement-learning-second-edition>.
- Sutton RS, McAllester DA, Singh SP, Mansour Y (2000) Policy gradient methods for reinforcement learning with function approximation. Solla S, Leen T, Müller K, eds. *Advances in Neural Information Processing Systems*, vol. 12 (MIT Press, Cambridge, MA), 1057–1063. <https://proceedings.neurips.cc/paper/1999/file/464d828b85b0bed98e80ade0a5c43b0f-Paper.pdf>.
- Tan M (1993) Multi-agent reinforcement learning: Independent vs. cooperative agents. *Proc. of the 10th Internat. Conf. on Machine Learn.*

- (Morgan Kaufmann, San Francisco), 330–337. <https://www.sciencedirect.com/science/article/pii/B9781558603073500496>.
- Tassioulas L, Ephremides A (1990) Stability properties of constrained queueing systems and scheduling policies for maximum throughput in multihop radio networks. *Proc. 29th IEEE Conf. on Decision and Control* (IEEE, New York), 2130–2132.
- Tsitsiklis JN, Van Roy B (1997) Analysis of temporal-difference learning with function approximation. Mozer MC, Jordan M, Petsche T, eds. *Adv. Neural Inform. Processing Systems*, vol. 9 (MIT Press, Cambridge, MA), 1075–1081. <https://proceedings.neurips.cc/paper/1996/file/e00406144c1e7e35240afed70f34166a-Paper.pdf>.
- Tsitsiklis JN, Van Roy B (1999) Average cost temporal-difference learning. *Automatica J. IFAC* 35(11):1799–1808.
- Tsitsiklis JN, Van Roy B (2002) On average vs. discounted reward temporal-difference learning. *Machine Learn.* 49(2):179–191.
- Tu S, Recht B (2019) The gap between model-based and model-free methods on the linear quadratic regulator: An asymptotic viewpoint. *Proc. Conf. on Learn. Theory* (PMLR), 3036–3083.
- Van Roy B, Tsitsiklis JN (1995) Stable linear approximations to dynamic programming for stochastic control problems with local transitions. *Proc. 8th Internat. Conf. on Neural Inform. Processing Systems, NIPS'95, Denver* (MIT Press, Cambridge, MA), 1045–1051.
- Varaiya P (2013) Max pressure control of a network of signalized intersections. *Transportation Res., Part C Emerging Tech.* 36:177–195.
- Vogels W, van Renesse R, Birman K (2003) The power of epidemics: Robust communication for large-scale distributed systems. *SIGCOMM Comput. Comm. Rev.* 33(1):131–135.
- Wai HT, Yang Z, Wang Z, Hong M (2018) Multi-agent reinforcement learning via double averaging primal-dual optimization. Bengio S, Wallach H, Larochelle H, Grauman K, Cesa-Bianchi N, Garnett R, eds. *Proc. 32nd Internat. Conf. on Neural Inform. Processing Systems. Advances in Neural Information Processing Systems*, vol. 31 (Curran Associates, Inc.), 9672–9683. <https://proceedings.neurips.cc/paper/2018/file/5a378f8490c8d6af8647a753812f6e31-Paper.pdf>.
- Whittle P (1988) Restless bandits: Activity allocation in a changing world. *J. Appl. Probability* 25(A):287–298.
- Wu Z, Jia QS, Guan X (2016) Optimal control of multiroom HVAC system: An event-based approach. *IEEE Trans. Control Systems Tech.* 24(2):662–669.
- Yun SY, Yi Y, Shin J, Eun DY (2012) Optimal CSMA: A survey. *2012 IEEE Internat. Conf. Comm. Systems (ICCS)* (IEEE, New York), 199–204.
- Zhang R, Pavone M (2016) Control of robotic mobility-on-demand systems: A queueing-theoretical perspective. *Internat. J. Robotics Res.* 35(1-3):186–203.
- Zhang K, Yang Z, Başar T (2021) Multi-agent reinforcement learning: A selective overview of theories and algorithms. Vamvoudakis KG, Wan Y, Lewis FL, Cansever D, eds. *Handbook of Reinforcement Learning and Control*, vol. 325 (Springer, Cham, Switzerland), 321–384. https://doi.org/10.1007/978-3-030-60990-0_12.
- Zhang X, Shi W, Yan B, Malkawi A, Li N (2017) Decentralized and distributed temperature control via HVAC systems in energy efficient buildings. Preprint, submitted February 10, 2017, <https://arxiv.org/abs/1702.03308>.
- Zhang K, Yang Z, Liu H, Zhang T, Basar T (2018) Fully decentralized multi-agent reinforcement learning with networked agents. *Proc. Internat. Conf. on Machine Learn.* (PMLR), 5872–5881.
- Zocca A (2019) Temporal starvation in multi-channel CSMA networks: An analytical framework. *Queueing Systems* 91(3-4):241–263.
- Zou S, Xu T, Liang Y (2019) Finite-sample analysis for SARSA with linear function approximation. Wallach H, Larochelle H, Beygelzimer A, d'Alch'e-Buc F, Fox E, Garnett R, eds. *Adv. Neural Inform. Processing Systems*, vol. 32 (Curran Associates, Inc.), 8665–8675. <https://proceedings.neurips.cc/paper/2019/file/9f9e8cba3700df6a947a8cf91035ab84-Paper.pdf>.

Guannan Qu is an assistant professor in the Electrical and Computer Engineering Department of Carnegie Mellon University. He is the recipient of Caltech Simoudis Discovery Award, PIMCO Fellowship, Amazon AI4Science Fellowship, and Institute of Electrical and Electronics Engineers Smart-GridComm Best Student Paper Reward. His research interest lies in control, optimization, and machine/reinforcement learning with applications to power systems, multiagent systems, Internet of things, smart city, and so on.

Adam Wierman is a professor in the Department of Computing and Mathematical Sciences at the California Institute of Technology. His research strives to make the networked systems that govern our world sustainable and resilient. He is a recipient of multiple awards, including the ACM Sigmetrics Rising Star award, ACM Sigmetrics Test of Time award, and IEEE Communications Society William R. Bennett Prize.

Na Li is a Gordon McKay professor in electrical engineering and applied mathematics at Harvard University. Her research lies in control, learning, and optimization of networked systems. She received an National Science Foundation Career Award (2016), Air Force Office of Scientific Research Young Investigator Award (2017), Office of Naval Research Young Investigator Award (2019), Donald P. Eckman Award (2019), and McDonald Mentoring Award (2020).