Proceedings of Machine Learning Research vol xxx:1–38, 2020

Scalable Reinforcement Learning of Localized Policies for Multi-Agent Networked Systems

Guannan Qu GQU@CALTECH.EDU
Adam Wierman ADAMW@CALTECH.EDU
Department of Computing and Mathematical Sciences
California Institute of Technology
Pasadena, CA 91125, USA

Na Li NALI@SEAS.HARVARD.EDU
School of Engineering and Applied Sciences
Harvard University
Cambridge, MA 02138, USA

Abstract

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

1. Introduction

Having demonstrated impressive performance in a wide array of domains such as game play (Silver et al., 2016; Mnih et al., 2015), robotics (Duan et al., 2016; Levine et al., 2016), autonomous driving (Li et al., 2019), Reinforcement Learning (RL) has emerged as a promising tool for decision and control. However, in order to use RL in the context of control of large scale networked systems, such as those in cyber-physical systems, it is necessary to develop scalable RL algorithms for networked systems.

In this paper, we consider a RL problem for a network of n agents, each with state s_i and action a_i, both taking values from finite sets. The agents are associated with an underlying dependence graph $\mathcal{G}$ and interact locally, i.e, the distribution of $s_i(t+1)$ only depends on the current states of the local neighborhood of i as well as the local $a_i(t)$. Further, each agent is associated with stage reward r_i that is a function of s_i, a_i, and the global stage reward is the average of r_i. In this setting, the design goal is to find a decision policy that maximizes the (discounted) global reward. This setting captures a wide range of applications. For example, such models have been used in the literature on epidemics (Mei et al., 2017), social networks (Chakrabarti et al., 2008; Llas et al., 2003), communication networks (Zocca, 2019; Vogels et al., 2003), queueing networks (Papadimitriou and Tsitsiklis, 1999), smart transportation (Zhang and Pavone, 2016), smart building systems (Wu et al., 2016; Zhang et al., 2017), and multi-agent game play (Borovikov et al., 2019).

A fundamental difficulty when applying RL to such networked systems is that, even if individual state and action spaces are small, the entire state profile $(s_1, \ldots, s_n)$ and the action profile $(a_1, \ldots, a_n)$ can take values from a set of size exponentially large in n. This "curse of dimensionality" renders the problem unscalable. For example, most RL algorithms like temporal difference (TD) learning or Q-learning require storage of a value function or Q-function (Bertsekas and Tsitsiklis, 1996) whose size is the same as the state space (or state-action space), which in our problem is exponentially large in n. Such scalability issues have indeed been observed in previous research on variants of the problem we study, e.g. in multi-agent RL (Littman, 1994; Bu et al., 2008) and factored Markov Decision Proccess (MDP) (Kearns and Koller, 1999; Guestrin et al., 2003). A variety of approaches have been proposed to manage this issue, e.g. the idea of "independent learners" in 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 in general, e.g., see Papadimitriou and Tsitsiklis (1999); Blondel and Tsitsiklis (2000).

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

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

Contributions. In this work we propose a framework that exploits properties of the network structure to develop RL to learn localized policies for large-scale networked systems in a scalable manner. Specifically, our main result (Theorem 5) 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 perhaps the first to provide such provable guarantee for scalable RL of localized policies in multi-agent network settings.

The key technique underlying our results is the observation that, when the size of κ-hop neighborhood is bounded, the network structure implies that the Q-function satisfies an *exponential decay property* (Definition 2), which leads to a tractable approximation of the policy gradient. In particular, despite the policy gradient itself being intractable to compute due to the large state-action space size, we introduce a *truncated policy gradient* (see Lemma 4) that can be computed efficiently and can be used in an actor-critic framework which yields an $O(\rho^{\kappa+1})$-approximation. This technique is novel and is a contribution in its own right. It can be used broadly to develop RL in network settings beyond the specific actor-critic algorithm we propose in this paper.

Related Literature. Our problem falls under category of the "succinctly described" MDPs in Blondel and Tsitsiklis (2000, Section 5.2), where the state and/or action space is a product space formed by the individual state and/or action space of multiple agents. As the state/action space is exponentially large, such problems are unscalable in general, even when the problem has structure (Blondel and Tsitsiklis, 2000; Whittle, 1988; Papadimitriou and Tsitsiklis, 1999). Despite this, there is a large literature on RL and MDPs in multi-agent settings under various structural assumptions.

Multi-agent RL dates back to the early work of Littman (1994); Claus and Boutilier (1998); Littman (2001); Hu and Wellman (2003) (see Bu et al. (2008) for a review) and has been actively studied, e.g. Zhang et al. (2018); Kar et al. (2013); Macua et al. (2015); Mathkar and Borkar (2017); Wai et al. (2018), see a more recent review in Zhang et al. (2019). Multi-agent RL encompasses a broad range of settings including competitive agents and Markov games. The case most relevant to ours is the cooperative multi-agent RL where typically, the agents can take their own actions but they share a common global state and maximize a global reward (Bu et al., 2008). This is contrast to the model we study, in which each agent has its own state and acts upon its own state. Despite the existence of a global state, multi-agent RL still faces scalability issues since the joint-action space is exponentially large. A number of techniques have been proposed to deal with this, including independent learners (Claus and Boutilier, 1998; Matignon et al., 2012), where each agent employs a single-agent RL method. While successful in some cases, the independent learner approach can suffer from instability (Matignon et al., 2012). Alternatively, one can use function approximation schemes to approximate the large Q-table, e.g., linear function approximation (Zhang et al., 2018) or neuro networks (Lowe et al., 2017). Such methods can reduce computation complexity significantly, but it is unclear whether the performance loss caused by the function approximation is small. In contrast, our technique not only reduces computation but also guarantees small performance loss.

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

Other Related Work. Beyond the above, our work is also connected to a few other classes of problems. The first is the class of *weakly coupled MDPs*, where every agent has its own state and action but their transition is decoupled (Meuleau et al., 1998). While similar to our model, our model differs in that the transition probability is coupled among the agents. Additionally, our model shares some similarity with the work of control of dynamical systems over graphs, e.g., the epidemics (Cator and Van Mieghem, 2012; Sahneh et al., 2013; Mei et al., 2017) and Glauber dynamics in physics (Lokhov et al., 2015; Mezard and Montanari, 2009), though our focus is very different from these works. Finally, this work is related to Qu and Li (2019), which assumes the full knowledge of 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 network assumptions.

2. Preliminaries

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

$\mathcal{A}_1 \times \cdots \times \mathcal{A}_n$. At time t, given current state $s(t)$ and action $a(t)$, the next individual state $s_i(t+1)$ is independently generated and is only dependent on neighbors:

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

where notation N_i means the neighborhood of i (including i itself) and s_{N_i} is the states of i's neighbors. In addition, for integer $\kappa \geq 1$, we let N_i^κ denote the κ-hop neighborhood of i, i.e. the nodes whose graph distance to i is less than or equal to κ, including i itself. We also let $f(\kappa) = \sup_i |N_i^\kappa|$.

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

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

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

To provide context for what follows, we review a few key concepts in RL. First, fixing a localized policy tuple $\theta = (\theta_1, \ldots, \theta_n)$, the Q-function for this policy θ is:

$$\begin{aligned} Q^\theta(s,a) &= \mathbb{E}_{a(t)\sim\zeta^\theta(\cdot|s(t))}\left[\sum_{t=0}^{\infty}\gamma^t r(s(t), a(t))\middle| s(0) = s, a(0) = a\right] \\ &= \frac{1}{n}\sum_{i=1}^n \mathbb{E}_{a(t)\sim\zeta^\theta(\cdot|s(t))}\left[\sum_{t=0}^{\infty}\gamma^t r_i(s_i(t), a_i(t))\middle| s(0) = s, a(0) = a\right] := \frac{1}{n}\sum_{i=1}^n Q_i^\theta(s,a). \end{aligned} \tag{3}$$

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

Finally, we recall the policy gradient theorem, which is the basis of many algorithmic results in RL. We emphasize that the lemma 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 fixed policy θ when $s(0)$ is drawn from π_0. Then*

$$\nabla J(\theta) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim\pi^\theta, a\sim\zeta^\theta(\cdot|s)} Q^\theta(s,a)\nabla \log \zeta^\theta(a|s). \tag{4}$$

3. Algorithm Design and Results

In this paper we propose an algorithm, Scalable Actor Critic (SAC), which provably finds an $O(\rho^{\kappa+1})$-stationary point of the objective $J(\theta)$ for some $\rho \leq \gamma$,[1] 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 5 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 Gradient Approximation

Recall that the policy gradient in Lemma 1 is intractable to compute due to the dimension of the Q-function. Our key idea is that exponential decay of the Q function allows efficient approximation of the policy gradient via truncation. To illustrate this, we start with the definition of the exponential decay property. Recall that N_i^κ is the set of κ-hop neighborhood of node i and define $N_{-i}^\kappa = \mathcal{N}/N_i^\kappa$, i.e. the set of agents that are outside of i'th κ-hop neighborhood. We write state s as $(s_{N_i^\kappa}, s_{N_{-i}^\kappa})$, i.e. the states of agents that are in the κ-hop neighborhood of i and outside of 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 2 *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 3 highlights that the exponential decay property holds generally, with $\rho = \gamma$. Further, under some mixing time assumptions, the exponential decay property holds with $\rho < \gamma$. For more details on the generality of the exponential decay property, see Appendix A.

Lemma 3 *If $\forall i$, r_i is upper bounded by $\bar{r}$, then the $(\frac{\bar{r}}{1-\gamma}, \gamma)$-exponential decay property holds.*

The power of the exponential decay property is that it guarantees that the dependence of Q_i^θ on other agents shrinks quickly as the distance between them grows. This motivates us to consider the following class of truncated Q-functions,

$$\hat{Q}_i^\theta(s_{N_i^\kappa}, a_{N_i^\kappa}) = \sum_{s_{N_{-i}^\kappa}, a_{N_{-i}^\kappa}} w_i(s_{N_{-i}^\kappa}, a_{N_{-i}^\kappa}; s_{N_i^\kappa}, a_{N_i^\kappa}) Q_i^\theta(s_{N_i^\kappa}, s_{N_{-i}^\kappa}, a_{N_i^\kappa}, a_{N_{-i}^\kappa}), \tag{5}$$

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

$$\sum_{s_{N_{-i}^\kappa} \in \mathcal{S}_{N_{-i}^\kappa}, a_{N_{-i}^\kappa} \in \mathcal{A}_{N_{-i}^\kappa}} w_i(s_{N_{-i}^\kappa}, a_{N_{-i}^\kappa}; s_{N_i^\kappa}, a_{N_i^\kappa}) = 1, \forall (s_{N_i^\kappa}, a_{N_i^\kappa}) \in \mathcal{S}_{N_i^k} \times \mathcal{A}_{N_i^k}. \tag{6}$$

Finally, our key insight is the following Lemma 4, which says when the exponential decay property holds, the truncated Q-function (5) can be used to accurately approximate the policy gradient. The proof of Lemma 4 is postponed to Appendix B.

1. In this paper, a ε-stationary point of $J(\theta)$ refers to a θ s.t. $\|\nabla J(\theta)\|^2 \leq \varepsilon$.

Lemma 4 (Truncated Policy Gradient) *Given i, define the following truncated policy gradient*

$$\hat{h}_i(\theta) = \frac{1}{1-\gamma}\mathbb{E}_{s\sim\pi^\theta, a\sim\zeta^\theta(\cdot|s)}\Big[\frac{1}{n}\sum_{j\in N_i^\kappa}\hat{Q}_j^\theta(s_{N_j^\kappa}, a_{N_j^\kappa})\Big]\nabla_{\theta_i}\log\zeta_i^{\theta_i}(a_i|s_i), \tag{7}$$

where $\hat{Q}_j^\theta$ can be any truncated Q-function in the form of (5). Then, if (c,ρ)-exponential decay property holds and if $\|\nabla_{\theta_i}\log\zeta_i^{\theta_i}(a_i|s_i)\| \le L_i$ for any a_i, s_i, we have $\|\hat{h}_i(\theta) - \nabla_{\theta_i}J(\theta)\| \le \frac{cL_i}{1-\gamma}\rho^{\kappa+1}$.

The power of this lemma is that the truncated Q function has much smaller dimension than the true Q function, and is thus scalable. However, despite the reduction in dimension, the error of the approximated gradient (7) is small. In the next section, we use this idea to design a scalable algorithm.

3.2. Algorithm Design: Scalable Actor Critic (SAC)

The good properties of the truncated Q-function open many possibilities for algorithm design. For instance, one can first obtain the truncated Q-function in some way (which could be much easier than directly computing the full Q-function) and then do a policy gradient step using the Lemma 4. In this subsection, we propose one particular approach using the actor-critic framework. Our approach, Scalable Actor Critic (SAC), uses temporal difference (TD) learning to obtain the truncated Q-function and then uses policy gradient for policy improvement. Psuedocode of the proposed algorithm is given in Algorithm 1.

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

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

The Actor: Policy Gradient. Steps 10 through 12 define the the actor actions. Here, each agent calculates an estimate of the truncated gradient based on (7), and then conducts a gradient step.

Discussion. Our algorithm serves as an initial concrete demonstration of how to make use of the truncated policy gradient to develop a scalable RL method for networked systems. There are many extensions and other approaches that could be pursued, either within the actor-critic framework or beyond. One immediate extension is to do a warm start, i.e., initialize $\hat{Q}_i^0$ as the final estimate $\hat{Q}_i^T$ in the previous outer-loop. Additionally, one can use the TD-λ variant of TD learning with variance reduction schemes like the advantage function. Further, beyond the actor-critic framework, another direction is to develop Q-learning/SARSA type algorithms based on the truncated Q-functions. These are interesting topics for future work.

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.

Algorithm 1: SAC: Scalable Actor Critic

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

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

  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 $r_i(0) = r_i(s_i(0), a_i(0))$.

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

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

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

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

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

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

  end

  Each agent i calculates approximated gradient,

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

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

end

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

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

Note that under Assumption 1, Assumption 2 automatically holds with $\rho = \gamma$, cf. Lemma 3. 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, as detailed in Appendix A.

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

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

Assumption 3 requires that every state action pair in the κ-hop neighborhood must be visited with some positive probability after some time. This type of assumption is common for finite time convergence results in RL. For example, in Srikant and Ying (2019), it is assumed that every state-action pair is visited with positive probability in the stationary distribution and the state-action distribution converges to the stationary distribution with some rate. This implies our assumption which is weaker in the sense that we only require local state-action pair $(s_{N_i^\kappa}, a_{N_i^\kappa})$ to be visited as opposed to the full state-action pair (s, a).

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

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

Theorem 5 *Under Assumption 1, 2, 3 and 4, for any $\delta \in (0,1)$, $M \geq 3$, suppose the critic step size $\alpha_t = \frac{h}{t+t_0}$ satisfies $h \geq \frac{1}{\sigma}\max(2, \frac{1}{1-\sqrt{\gamma}})$, $t_0 \geq \max(2h, 4\sigma h, \tau)$; and the actor step size satisfies $\eta_m = \frac{\eta}{\sqrt{m+1}}$ with $\eta \leq \frac{1}{4L'}$. Further, if the inner loop length T is large enough s.t. $T+1 \geq \log_\gamma \frac{c(1-\gamma)}{\bar{r}} + (\kappa+1)\log_\gamma \rho$ and*

$$\frac{C_a(\frac{\delta}{2nM}, T)}{\sqrt{T+t_0}} + \frac{C_a'}{T+t_0} \leq \frac{2c\rho^{\kappa+1}}{(1-\gamma)^2}, \tag{8}$$

where

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

with $\bar{\epsilon} = 4\frac{\bar{r}}{1-\gamma} + 2\bar{r}$ and we recall that $f(\kappa) = \max_i |N_i^\kappa|$ is the size of the largest κ-neighborhood. Then, with probability at least $1-\delta$,

$$\frac{\sum_{m=0}^{M-1}\eta_m\|\nabla J(\theta(m))\|^2}{\sum_{m=0}^{M-1}\eta_m} \leq \frac{\frac{2\bar{r}}{\eta(1-\gamma)} + \frac{8\bar{r}^2L^2}{(1-\gamma)^4}\sqrt{\log M \log\frac{4}{\delta}} + \frac{96\bar{r}^2L'L^2}{(1-\gamma)^4}\eta\log M}{\sqrt{M+1}} + \frac{12L^2c\bar{r}}{(1-\gamma)^5}\rho^{\kappa+1}. \tag{9}$$

The proof of Theorem 5 can be found in Appendix-D. To interpret the result, note that the first term in (9) converges to 0 in the order of $\tilde{O}(\frac{1}{\sqrt{M}})$ and the second term, which we denote as ε_κ, is the bias caused by the truncation of the Q-function and it scales in the order of $O(\rho^{\kappa+1})$. As such, our method SAC will eventually find an $O(\rho^{\kappa+1})$-approximation of a stationary point of the objective function $J(\theta)$, which could be very close to a true stationary point even for small κ as ε_κ decays exponentially in κ.

In terms of complexity, (9) gives that, to reach a $O(\varepsilon_\kappa)$-approximate stationary point, the number of outer-loop iterations required is $M \geq \tilde{\Omega}(\frac{1}{\varepsilon_\kappa{}^2}poly(\bar{r}, L, L', \frac{1}{(1-\gamma)}))$, which scales polynomially with the parameters of the problem. We emphasize that it does not scale exponentially with n. Further, since the left hand side of (8) decays to 0 as T increases in the order of $\tilde{O}(\frac{1}{\sqrt{T}})$ and the right hand side of (8) is in the same order as $O(\varepsilon_\kappa)$, the inner-loop length required is $T \geq \tilde{\Omega}(\frac{1}{\varepsilon_k^2}poly(\tau, \frac{1}{\sigma}, \frac{1}{1-\gamma}, \bar{r}, f(\kappa)))$. Parameters τ and $\frac{1}{\sigma}$ are from Assumption 3 and they scale with the local state-action space size of the largest κ-hop neighborhood. Therefore, the inner-loop length required scale with the size of the local state-action space of the largest κ-neighborhood, which is much smaller than the full state-action space size when the graph is sparse.[2]

4. Conclusion and Discussion

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 neighbor, which can be much

2. This requirement on T could 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.

smaller than the full state-action space size when the graph is sparse. This perhaps represents the first scalable RL method for localized control of multi-agent networked systems with such provable guarantee. In addition, the framework underlying SAC, including the truncated Q-function (5) and truncated policy gradient (Lemma 7), is a contribution in its own right and could potentially lead to other scalable RL methods for networked systems, including the warm start, TD-λ variants and Q-learning/SARSA type methods. We leave these directions as future work.

References

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

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

Igor Borovikov, Yunqi Zhao, Ahmad Beirami, Jesse Harder, John Kolen, James Pestrak, Jervis Pinto, Reza Pourabolghasem, Harold Chaput, Mohsen Sardari, Long Lin, Navid Aghdaie, and Kazi Zaman. Winning isn't everything: Training agents to playtest modern games. In *AAAI 2019 Workshop on Reinforcement Learning in Games*, 01 2019.

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

Eric Cator and Piet Van Mieghem. Second-order mean-field susceptible-infected-susceptible epidemic threshold. *Physical review E*, 85(5):056111, 2012.

Deepayan Chakrabarti, Yang Wang, Chenxi Wang, Jurij Leskovec, and Christos Faloutsos. Epidemic thresholds in real networks. *ACM Transactions on Information and System Security (TISSEC)*, 10(4):1, 2008.

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

Yan Duan, Xi Chen, Rein Houthooft, John Schulman, and Pieter Abbeel. Benchmarking deep reinforcement learning for continuous control. In *International Conference on Machine Learning*, pages 1329–1338, 2016.

Carlos Guestrin, Daphne Koller, Ronald Parr, and Shobha Venkataraman. Efficient solution algorithms for factored mdps. *Journal of Artificial Intelligence Research*, 19:399–468, 2003.

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

Soummya Kar, José MF Moura, and H Vincent Poor. Qd-learning: A collaborative distributed strategy for multi-agent reinforcement learning through consensus + innovations. *IEEE Transactions on Signal Processing*, 61(7):1848–1862, 2013.

Michael Kearns and Daphne Koller. Efficient reinforcement learning in factored mdps. In *IJCAI*, volume 16, pages 740–747, 1999.

Sergey Levine, Chelsea Finn, Trevor Darrell, and Pieter Abbeel. End-to-end training of deep visuomotor policies. *The Journal of Machine Learning Research*, 17(1):1334–1373, 2016.

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

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

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

Mateu Llas, Pablo M Gleiser, Juan M López, and Albert Díaz-Guilera. Nonequilibrium phase transition in a model for the propagation of innovations among economic agents. *Physical Review E*, 68(6):066101, 2003.

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

Ryan Lowe, Yi Wu, Aviv Tamar, Jean Harb, OpenAI Pieter Abbeel, and Igor Mordatch. Multi-agent actor-critic for mixed cooperative-competitive environments. In *Advances in Neural Information Processing Systems*, pages 6379–6390, 2017.

Sergio Valcarcel Macua, Jianshu Chen, Santiago Zazo, and Ali H Sayed. Distributed policy evaluation under multiple behavior strategies. *IEEE Transactions on Automatic Control*, 60(5):1260–1274, 2015.

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

Laetitia Matignon, Guillaume J Laurent, and Nadine Le Fort-Piat. Independent reinforcement learners in cooperative markov games: a survey regarding coordination problems. *The Knowledge Engineering Review*, 27(1):1–31, 2012.

Wenjun Mei, Shadi Mohagheghi, Sandro Zampieri, and Francesco Bullo. On the dynamics of deterministic epidemic propagation over networks. *Annual Reviews in Control*, 44:116–128, 2017.

Nicolas Meuleau, Milos Hauskrecht, Kee-Eung Kim, Leonid Peshkin, Leslie Pack Kaelbling, Thomas L Dean, and Craig Boutilier. Solving very large weakly coupled markov decision processes. In *AAAI/IAAI*, pages 165–172, 1998.

Marc Mezard and Andrea Montanari. *Information, physics, and computation*. Oxford University Press, 2009.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A Rusu, Joel Veness, Marc G Bellemare, Alex Graves, Martin Riedmiller, Andreas K Fidjeland, Georg Ostrovski, et al. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529, 2015.

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

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

Guannan Qu and Na Li. Exploiting fast decaying and locality in multi-agent mdp with tree dependence structure. *arXiv preprint arXiv:1909.06900*, 2019.

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

Faryad Darabi Sahneh, Caterina Scoglio, and Piet Van Mieghem. Generalized epidemic mean-field model for spreading processes over multilayer complex networks. *IEEE/ACM Transactions on Networking (TON)*, 21(5):1609–1620, 2013.

David Silver, Aja Huang, Chris J Maddison, Arthur Guez, Laurent Sifre, George Van Den Driessche, Julian Schrittwieser, Ioannis Antonoglou, Veda Panneershelvam, Marc Lanctot, et al. Mastering the game of go with deep neural networks and tree search. *nature*, 529(7587):484, 2016.

R Srikant and Lei Ying. Finite-time error bounds for linear stochastic approximation and td learning. *arXiv preprint arXiv:1902.00923*, 2019.

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

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

Werner Vogels, Robbert van Renesse, and Ken Birman. The power of epidemics: Robust communication for large-scale distributed systems. *SIGCOMM Comput. Commun. Rev.*, 33(1):131–135, January 2003. ISSN 0146-4833. doi: 10.1145/774763.774784. URL http://doi.acm.org/10.1145/774763.774784.

Hoi-To Wai, Zhuoran Yang, Zhaoran Wang, and Mingyi Hong. Multi-agent reinforcement learning via double averaging primal-dual optimization. *arXiv preprint arXiv:1806.00877*, 2018.

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

Zijian Wu, Qing-Shan Jia, and Xiaohong Guan. Optimal control of multiroom hvac system: An event-based approach. *IEEE Transactions on Control Systems Technology*, 24(2):662–669, 2016.

Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Başar. Fully decentralized multi-agent reinforcement learning with networked agents. *arXiv preprint arXiv:1802.08757*, 2018.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *arXiv preprint arXiv:1911.10635*, 2019.

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

Xuan Zhang, Wenbo Shi, Bin Yan, Ali Malkawi, and Na Li. Decentralized and distributed temperature control via hvac systems in energy efficient buildings. *arXiv preprint arXiv:1702.03308*, 2017.

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

Shaofeng Zou, Tengyu Xu, and Yingbin Liang. Finite-sample analysis for sarsa with linear function approximation. In *Advances in Neural Information Processing Systems*, pages 8665–8675, 2019.

Appendix A. The Exponential Decay Property

Our main results depend on the (c,ρ)-exponential decay of the Q-function (cf. Definition 2), which means that for any i, any $s_{N_i^\kappa}$, $s_{N_{-i}^\kappa}$ and $s'_{N_{-i}^\kappa}$, $a_{N_i^\kappa}$, $a_{N_{-i}^\kappa}$ and $a'_{N_{-i}^\kappa}$,

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

In Section 3.1, we have pointed out in Lemma 3 that the (c,ρ)-exponential decay property always holds with ρ being set to the discounting factor γ, assuming the rewards r_i are upper bounded. We now provide the proof of Lemma 3.

Proof of Lemma 3. For notational simplicity, denote $s = (s_{N_i^\kappa}, s_{N_{-i}^\kappa})$, $a = (a_{N_i^\kappa}, a_{N_{-i}^\kappa})$; $s' = (s_{N_i^\kappa}, s'_{N_{-i}^\kappa})$ and $a' = (a_{N_i^\kappa}, a'_{N_{-i}^\kappa})$. Let $\pi_{t,i}$ be the distribution of $(s_i(t), a_i(t))$ conditioned on $(s(0), a(0)) = (s, a)$ under policy θ, and let $\pi'_{t,i}$ be the distribution of $(s_i(t), a_i(t))$ conditioned on $(s(0), a(0)) = (s', a')$ under policy θ. Then, we must have $\pi_{t,i} = \pi'_{t,i}$ for all $t \le \kappa$. The reason is that, due to the local dependence structure (1) and the localized policy structure, $\pi_{t,i}$ only depends on $(s_{N_i^t}, a_{N_i^t})$ (the initial state-action of agent i'th t-hop neighborhood) which is the same as $(s'_{N_i^t}, s'_{N_i^t})$ when $t \le \kappa$ per the way the initial state (s,a), (s',a') are chosen. With these definitions, we expand the definition of Q_i^θ in (3),

$$\begin{aligned}
&|Q_i^\theta(s,a) - Q_i^\theta(s',a')| \\
&\le \sum_{t=0}^\infty \Big| \mathbb{E}\big[\gamma^t r_i(s_i(t), a_i(t)) \big| (s(0), a(0)) = (s,a)\big] - \mathbb{E}\big[\gamma^t r_i(s_i(t), a_i(t)) \big| (s(0), a(0)) = (s',a')\big] \Big| \\
&= \sum_{t=0}^\infty \Big| \gamma^t \mathbb{E}_{(s_i,a_i)\sim\pi_{t,i}} r_i(s_i,a_i) - \gamma^t \mathbb{E}_{(s_i,a_i)\sim\pi'_{t,i}} r_i(s_i,a_i) \Big| \\
&= \sum_{t=\kappa+1}^\infty \Big| \gamma^t \mathbb{E}_{(s_i,a_i)\sim\pi_{t,i}} r_i(s_i,a_i) - \gamma^t \mathbb{E}_{(s_i,a_i)\sim\pi'_{t,i}} r_i(s_i,a_i) \Big| \\
&\le \sum_{t=\kappa+1}^\infty \gamma^t \bar{r} \mathrm{TV}(\pi_{t,i}, \pi'_{t,i}) \le \frac{\bar{r}}{1-\gamma}\gamma^{\kappa+1},
\end{aligned} \tag{10}$$

where $\mathrm{TV}(\pi_{t,i}, \pi'_{t,i})$ is the total variation distance between $\pi_{t,i}$ and $\pi'_{t,i}$ which is upper bounded by 1. The above inequality shows that the $(\frac{\bar{r}}{1-\gamma}, \gamma)$-exponential decay property holds and concludes the proof of Lemma 3. ■

Lemma 3 shows that the (c, ρ)-exponential decay property automatically holds with ρ being the discounting factor γ, without any assumption on the transition probabilities except for the factorization structure (1) and the localized policy structure. However, in practice, typically the Markov chain is ergodic and has fast mixing property. The following Lemma 6 shows that when some fast mixing holds, then the (c, ρ)-exponential decay property holds for some $\rho < \gamma$.

Lemma 6 *Suppose r_i is upper bounded by $\bar{r}$ for all i, and assume there exists $c' > 0$ and $\mu \in (0,1)$ s.t. under any policy θ, the Markov chain is ergodic and starting from any initial state, $\mathrm{TV}(\pi_{t,i}, \pi_{\infty,i}) \le c'\mu^t$ 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. Then, the $(\frac{2c'\bar{r}}{1-\gamma\mu}, \gamma\mu)$-exponential decay property holds.*

Proof The proof is almost identical to that of Lemma 3. The only change is that in step (10), we use $\mathrm{TV}(\pi_{t,i}, \pi'_{t,i}) \le 2c'\mu^t$. ■

The condition on mixing rate in Lemma 6 is similar to those used in the literature on finite-time analysis of RL methods, e.g. Zou et al. (2019). In fact, our condition is weaker than the common mixing rate condition in that we only require the distribution of the local state-action pair $(s_i(t), a_i(t))$ to mix, instead of the full state-action pair $(s(t), a(t))$. We leave it as future work to study such "local" mixing behavior and its relation to the local transition probabilities (1).

Appendix B. Proof of Lemma 4

We first show that 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 (5) and (6),

$$\begin{aligned}
&|\hat{Q}_i^\theta(s_{N_i^\kappa}, a_{N_i^\kappa}) - Q_i^\theta(s,a)| \\
&= \Big| \sum_{s'_{N_{-i}^\kappa}, a'_{N_{-i}^\kappa}} w_i(s'_{N_{-i}^\kappa}, a'_{N_{-i}^\kappa}; s_{N_i^\kappa}, a_{N_i^\kappa}) Q_i^\theta(s_{N_i^\kappa}, s'_{N_{-i}^\kappa}, a_{N_i^\kappa}, a'_{N_{-i}^\kappa}) - Q_i^\theta(s_{N_i^\kappa}, s_{N_{-i}^\kappa}, a_{N_i^\kappa}, a_{N_{-i}^\kappa}) \Big| \\
&\le \sum_{s'_{N_{-i}^\kappa}, a'_{N_{-i}^\kappa}} w_i(s'_{N_{-i}^\kappa}, a'_{N_{-i}^\kappa}; s_{N_i^\kappa}, a_{N_i^\kappa}) \Big| Q_i^\theta(s_{N_i^\kappa}, s'_{N_{-i}^\kappa}, a_{N_i^\kappa}, a'_{N_{-i}^\kappa}) - Q_i^\theta(s_{N_i^\kappa}, s_{N_{-i}^\kappa}, a_{N_i^\kappa}, a_{N_{-i}^\kappa}) \Big| \\
&\le c\rho^{\kappa+1},
\end{aligned} \tag{11}$$

where in the last step, we have used the (c, ρ) exponential decay property, cf. Definition 2.

Next, recall by the policy gradient theorem (Lemma 1),

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

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

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

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

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

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

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

where in the last step, we have used (11) 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 4. ■

Appendix C. Analysis of the Critic

In this section we provide an analysis of the error bound associated with the critic component of our framework. More specifically, recall that within iteration m the inner loop update is

$$\begin{aligned}\hat{Q}_i^t(s_{N_i^\kappa}(t-1), a_{N_i^\kappa}(t-1)) &= (1-\alpha_{t-1})\hat{Q}_i^{t-1}(s_{N_i^\kappa}(t-1), a_{N_i^\kappa}(t-1)) \\ &\quad + \alpha_{t-1}(r_i(s_i(t-1), a_i(t-1)) + \gamma \hat{Q}_i^{t-1}(s_{N_i^\kappa}(t), a_{N_i^\kappa}(t))), \quad (12)\\ \hat{Q}_i^t(s_{N_i^\kappa}, a_{N_i^\kappa}) &= \hat{Q}_i^{t-1}(s_{N_i^\kappa}, a_{N_i^\kappa}) \text{ for } (s_{N_i^\kappa}, a_{N_i^\kappa}) \neq (s_{N_i^\kappa}(t-1), a_{N_i^\kappa}(t-1)),\end{aligned}\tag{13}$$

where $\hat{Q}_i^0 \in \mathbb{R}^{\mathcal{S}_{N_i^\kappa}\times\mathcal{A}_{N_i^\kappa}}$ is initialized to be all zero vector, and $\alpha_t = \frac{h}{t+t_0}$ is the step size. We note that when implementing (12) and (13) within outer loop iteration m, $(s(t), a(t))$ is a random trajectory generated by the agents taking a fixed policy $\theta(m)$. Let $Q_i^{\theta(m)} \in \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ be the true Q-function for reward r_i under this fixed policy $\theta(m)$ as defined in (3).

Given the above notation, the specific goal of this section is to prove the following theorem, which bounds the error between the approximation $\hat{Q}_i^T$ generated by (12), (13) and the true $Q_i^{\theta(m)}$.

Theorem 7 *Assume Assumption 1, 2, 3 are true and suppose t_0, h satisfies, $h \geq \frac{1}{\sigma}\max(2, \frac{1}{1-\sqrt{\gamma}})$ and $t_0 \geq \max(2h, 4\sigma h, \tau)$. Then, inside outer loop iteration m, for each $i \in \mathcal{N}$, with probability at least $1-\delta$, we have the following error bound,*

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

where

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

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

C.1. Overview of the proof of Theorem 7

Since Theorem 7 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^*. Since Q_i^* is the true Q-function for reward r_i under policy $\theta(m)$, it must satisfy the Bellman equation (Bertsekas and Tsitsiklis, 1996),

$$Q_i^* = \text{TD}(Q_i^*) := r_i + \gamma P Q_i^*, \tag{14}$$

where $\text{TD} : \mathbb{R}^{\mathcal{S}\times\mathcal{A}} \to \mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ is the standard Bellman operator for reward r_i and P is the transition probability from $s(t), a(t)$ to $s(t+1), a(t+1)$ under policy $\theta(m)$. Note in (14), without causing any confusion, r_i is interpreted as a vector in $\mathbb{R}^{\mathcal{S}\times\mathcal{A}}$ although r_i only depends on (s_i, a_i).

Theorem 7 essentially says that the critic iterate $\hat{Q}_i^t$ in (12) (13) will become a good estimate of Q_i^* as t increases. Our proof is divided into 5 steps. In Step 1, we rewrite (12) and (13) in a linear update form (cf. (16)). Then, the averaged behavior of the linear update form will be studied in Step 2 (cf. Lemma 8). In Step 3, we decompose the error into a recursive form (cf.

Lemma 11), and in Step 4, we bound a certain martingale difference-like sequence (cf. Lemma 12 and Lemma 13). Finally, in Step 5, we use the recursive error decomposition and the bound on the martingale difference-like sequence to prove Theorem 7.

Step 1: Writing the critic update in linear form. To simplify notation, we use the following definitions. We use $z = (s, a) \in \mathcal{Z} = \mathcal{S} \times \mathcal{A}$ 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 $z_{N_i^\kappa} = (s_{N_i^\kappa}, a_{N_i^\kappa}) \in \mathcal{Z}_{N_i^\kappa} = \mathcal{S}_{N_i^\kappa} \times \mathcal{A}_{N_i^\kappa}$. Also, define $\mathbf{e}_{z_{N_i^\kappa}}$ to be the indicator vector in $\mathbb{R}^{\mathcal{Z}_{N_i^\kappa}}$, i.e. the $z_{N_i^\kappa}$'th entry of $\mathbf{e}_{z_{N_i^\kappa}}$ is 1 and other entries are zero. Then, the critic update equations (12) and (13) can be written as,

$$\hat{Q}_i^t = \hat{Q}_i^{t-1} + \alpha_{t-1}[r_i(z_i(t-1)) + \gamma \hat{Q}_i^{t-1}(z_{N_i^\kappa}(t)) - \hat{Q}_i^{t-1}(z_{N_i^\kappa}(t-1))]\mathbf{e}_{z_{N_i^\kappa}(t-1)}, \tag{15}$$

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

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

and rewrite (15) in a linear form

$$\hat{Q}_i^t = \hat{Q}_i^{t-1} + \alpha_{t-1}\big[A(z(t-1), z(t))\hat{Q}_i^{t-1} + b(z(t-1))\big]. \tag{16}$$

Step 2: Analyze the average behavior of A, b. Recall that P is transition matrix from $z(t-1)$ to $z(t)$. We define,

$$\tilde{A}(z) = \mathbb{E}_{z' \sim P(\cdot|z)} A(z, z') = \mathbf{e}_{z_{N_i^\kappa}}[\gamma P(\cdot|z)\Phi - \mathbf{e}_{z_{N_i^\kappa}}^T], \tag{17}$$

where $P(\cdot|z)$ is understood as the z'th row of P and is treated as a row vector. Also, we have defined $\Phi \in \mathbb{R}^{\mathcal{Z} \times \mathcal{Z}_{N_i^\kappa}}$ to be a matrix with each row indexed by $z \in \mathcal{Z}$ and each column indexed by $z'_{N_i^\kappa} \in \mathcal{Z}_{N_i^\kappa}$. Further, the z'th row of Φ is the indicator vector $\mathbf{e}_{z_{N_i^\kappa}}^\top$, in other words $\Phi(z, z'_{N_i^\kappa}) = 1$ if $z'_{N_i^\kappa} = z_{N_i^\kappa}$ and $\Phi(z, z'_{N_i^\kappa}) = 0$ elsewhere. We further define, given any distribution d on the state-action pair z, the "averaged" A and b,

$$\begin{aligned} \bar{A}^d &= \mathbb{E}_{z \sim d} \tilde{A}(z) \\ &= \sum_{z \in \mathcal{Z}} d(z) \mathbf{e}_{z_{N_i^\kappa}}[\gamma P(\cdot|z)\Phi - \mathbf{e}_{z_{N_i^\kappa}}^\top] \\ &= \Phi^\top \text{diag}(d)\big[\gamma P \Phi - \Phi\big], \end{aligned} \tag{18}$$

$$\bar{b}^d = \mathbb{E}_{z \sim d} b(z) = \Phi^\top \text{diag}(d) r_i, \tag{19}$$

where $\text{diag}(d) \in \mathbb{R}^{\mathcal{Z} \times \mathcal{Z}}$ is a diagonal matrix with the z'th diagonal entry being $d(z)$; in the last equation, r_i is understood as a vector over the entire state-action space $\mathcal{Z}$, though it only depends on z_i. The goal of this step is to show the following lemma, which shows a certain contraction property for the "averaged" A and b. The proof is postponed to Section C.2.

Lemma 8 *Given distribution d on state-action pair z whose marginalization onto $z_{N_i^\kappa}$ is non-zero for every $z_{N_i^\kappa}$, we have $\bar{A}^d\hat{Q}_i + \bar{b}^d$ can be written as*

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

where $D = \Phi^\top \mathrm{diag}(d)\Phi \in \mathbb{R}^{\mathcal{Z}_{N_i^\kappa}\times\mathcal{Z}_{N_i^\kappa}}$ is a diagonal matrix, with the $z_{N_i^\kappa}$'th entry being the marginalized distribution of $z_{N_i^\kappa}$ under distribution d; $g^d(\cdot)$ is given by $g^d(\hat{Q}_i) = \Pi^d \mathrm{TD}\Phi\hat{Q}_i$, where $\Pi^d = (\Phi^\top\mathrm{diag}(d)\Phi)^{-1}\Phi^\top\mathrm{diag}(d)$ and $\mathrm{TD}(Q_i) = r_i + \gamma P Q_i$ is the Bellman operator in (14).

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

$$\|\Phi\hat{Q}_i^d - Q_i^*\|_\infty \le \frac{c\rho^{\kappa+1}}{1-\gamma}. \tag{20}$$

Step 3: Decomposition of the error. Recall the update for $\hat{Q}_i^t$ is

$$\hat{Q}_i^t = \hat{Q}_i^{t-1} + \alpha_{t-1}\big[A(z(t-1), z(t))\hat{Q}_i^{t-1} + b(z(t-1))\big]. \tag{21}$$

We define the following simplifying notations,

$$A_{t-1} = A(z(t-1), z(t)),$$
$$b_{t-1} = b(z(t-1)).$$

Let $\mathcal{F}_t$ be the σ-algebra generated by $z(0),\ldots,z(t)$. Then, clearly A_{t-1} is $\mathcal{F}_t$-measurable and b_{t-1} is $\mathcal{F}_{t-1}$ measurable. As a result, $\hat{Q}_i^t$ is $\mathcal{F}_t$-measurable. Let $\tau > 0$ to be the integer in Assumption 3. Let d_{t-1} be the distribution of $z(t-1)$ conditioned on $\mathcal{F}_{t-\tau}$. Further define,

$$\bar{A}_{t-1} = \bar{A}^{d_{t-1}}, \quad \bar{b}_{t-1} = \bar{b}^{d_{t-1}},$$

i.e. the "averaged" A and b under distribution d_{t-1}. 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, (21) can be rewritten as,

$$\begin{aligned}
\hat{Q}_i^t &= \hat{Q}_i^{t-1} + \alpha_{t-1}\big[A_{t-1}\hat{Q}_i^{t-1} + b_{t-1}\big]\\
&= \hat{Q}_i^{t-1} + \alpha_{t-1}\big[\bar{A}_{t-1}\hat{Q}_i^{t-1} + \bar{b}_{t-1}\big] + \alpha_{t-1}[(A_{t-1}-\bar{A}_{t-1})\hat{Q}_i^{t-1} + b_{t-1} - \bar{b}_{t-1}]\\
&= \hat{Q}_i^{t-1} + \alpha_{t-1}\big[\bar{A}_{t-1}\hat{Q}_i^{t-1} + \bar{b}_{t-1}\big]\\
&\quad + \alpha_{t-1}\underbrace{\big[(A_{t-1}-\bar{A}_{t-1})\hat{Q}_i^{t-\tau} + b_{t-1} - \bar{b}_{t-1}\big]}_{:=\epsilon_{t-1}} + \alpha_{t-1}\underbrace{(A_{t-1}-\bar{A}_{t-1})(\hat{Q}_i^{t-1} - \hat{Q}_i^{t-\tau})}_{:=\phi_{t-1}},
\end{aligned} \tag{22}$$

where in the last step, we have defined sequence ϵ_{t-1} and ϕ_{t-1}. We have the following auxiliary lemma that provides upper bounds for $\hat{Q}_i^t$, ϵ_t and ϕ_t, which will be frequently used in the rest of the proof. The proof of Lemma 9 is postponed to Section C.3.

Lemma 9 *We have the following upper bounds.*

(a) $\|\hat{Q}_i^t\|_\infty \le \frac{\bar{r}}{1-\gamma}$ *almost surely.*

(b) $\|\epsilon_t\|_\infty \leq \bar{\epsilon} = 4\frac{\bar{r}}{1-\gamma} + 2\bar{r}$ *almost surely.*

(c) $\|\phi_t\|_\infty \leq 2\bar{\epsilon}\sum_{k=t-\tau+1}^{t-1}\alpha_k$ *almost surely.*

By Lemma 8, we have for each t, there exists diagonal matrix D_{t-1} and operator g_{t-1} s.t.

$$\bar{A}_{t-1}\hat{Q}_i^{t-1} + \bar{b}_{t-1} = -D_{t-1}\hat{Q}_i^{t-1} + D_{t-1}g_{t-1}(\hat{Q}_i^{t-1}), \tag{23}$$

where by Lemma 8, g_{t-1} is a γ-contraction in infinity norm, with unique fixed point $\hat{Q}_i^{d_{t-1}}$ satisfying

$$\|\Phi\hat{Q}_i^{d_{t-1}} - Q_i^*\|_\infty \leq \frac{c\rho^{\kappa+1}}{1-\gamma}. \tag{24}$$

Further, by Lemma 8 $D_{t-1} \in \mathbb{R}^{\mathcal{Z}_{N_i^\kappa}\times\mathcal{Z}_{N_i^\kappa}}$ is a diagonal matrix, with the $z_{N_i^\kappa}$'th entry being $d_{t-1}(z_{N_i^\kappa})$, the marginalized distribution of $z_{N_i^\kappa}$ under d_{t-1}. Since d_{t-1} is the distribution of $z(t-1)$ conditioned on $\mathcal{F}_{t-\tau}$, by Assumption 3, we have almost surely,

$$D_{t-1} \succeq \sigma I, \tag{25}$$

where $\sigma > 0$ is from Assumption 3.

With these preparations, we plug (23) into (22) and expand it recursively, getting,

$$\begin{aligned}\hat{Q}_i^t &= (I-\alpha_{t-1}D_{t-1})\hat{Q}_i^{t-1} + \alpha_{t-1}D_{t-1}g_{t-1}(\hat{Q}_i^{t-1}) + \alpha_{t-1}\epsilon_{t-1} + \alpha_{t-1}\phi_{t-1}\\ &= \prod_{k=\tau}^{t-1}(I-\alpha_k D_k)\hat{Q}_i^\tau + \sum_{k=\tau}^{t-1}\alpha_k D_k\prod_{\ell=k+1}^{t-1}(I-\alpha_\ell D_\ell)g_k(\hat{Q}_i^k) + \sum_{k=\tau}^{t-1}\alpha_k\prod_{\ell=k+1}^{t-1}(I-\alpha_\ell D_\ell)(\epsilon_k+\phi_k).\end{aligned} \tag{26}$$

We use the following notation:

$$B_{k,t} = \alpha_k D_k\prod_{\ell=k+1}^{t-1}(I-\alpha_\ell D_\ell), \tilde{B}_{k,t} = \prod_{\ell=k+1}^{t-1}(I-\alpha_\ell D_\ell).$$

It is then immediately clear that

$$\tilde{B}_{\tau-1,t} + \sum_{k=\tau}^{t-1}B_{k,t} = I. \tag{27}$$

We also define

$$\beta_{k,t} = \alpha_k\prod_{\ell=k+1}^{t-1}(1-\alpha_\ell\sigma), \quad \tilde{\beta}_{k,t} = \prod_{\ell=k+1}^{t-1}(1-\alpha_\ell\sigma).$$

Since every diagonal entry of D_ℓ is lower bounded by σ almost surely (cf. (25)), 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. We have the following lemma on the $\beta_{k,t}$, $\tilde{\beta}_{k,t}$ sequence which we will frequently use later. The proof of Lemma 10 is provided in Section C.3.

Lemma 10 *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 the following.*

(a) $\beta_{k,t} \le \frac{h}{k+t_0}\left(\frac{k+1+t_0}{t+t_0}\right)^{\sigma h}$, $\tilde{\beta}_{k,t} \le \left(\frac{k+1+t_0}{t+t_0}\right)^{\sigma h}$.

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

(c) $\sum_{k=\tau}^{t-1} \beta_{k,t} \sum_{\ell=k-\tau+1}^{k-1} \alpha_\ell \le \frac{8h\tau}{\sigma}\frac{1}{t+t_0}$.

Next, (26) can be rewritten as

$$\hat{Q}_i^t = \tilde{B}_{\tau-1,t}\hat{Q}_i^\tau + \sum_{k=\tau}^{t-1} B_{k,t} g_k(\hat{Q}_i^k) + \sum_{k=\tau}^{t-1} \alpha_k \tilde{B}_{k,t}\epsilon_k + \sum_{k=\tau}^{t-1} \alpha_k \tilde{B}_{k,t}\phi_k. \tag{28}$$

The goal of this step is to decompose the error. Let $a_t = \|\Phi\hat{Q}_i^t - Q_i^*\|_\infty = \sup_{z\in\mathcal{Z}} |\hat{Q}_i^t(z_{N_i^\kappa}) - Q_i^*(z)|$ be the error at time t. From (28), and also utilizing the γ-contraction of g_k as well as the property of the fixed point of g_k (24), we have the following Lemma, which decomposes the error in a resursive form. The proof of Lemma 11 is postponed to Section C.4.

Lemma 11 *Let $a_t = \|\Phi\hat{Q}_i^t - Q_i^*\|_\infty$. The following recursion holds almost surely,*

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

where $b_{k,t}(z_{N_i^\kappa})$ is the $z_{N_i^\kappa}$'th diagonal entry of $B_{k,t}$, and $b_{k,t}(z_{N_i^\kappa}) = \alpha_k d_k(z_{N_i^\kappa}) \prod_{\ell=k+1}^{t-1}(1 - \alpha_\ell d_\ell(z_{N_i^\kappa}))$, where $d_k(z_{N_i^\kappa})$ is the $z_{N_i^\kappa}$'th diagonal entry of D_k satisfying $d_k(z_{N_i^\kappa}) \ge \sigma$.

From Lemma 11, it is clear that to bound the error a_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 4: 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 $\|\sum_{k=\tau}^{t-1} \alpha_k \tilde{B}_{k,t}\phi_k\|_\infty$. Recall that,

$$\begin{aligned}
\epsilon_{t-1} &= (A_{t-1} - \bar{A}_{t-1})\hat{Q}_i^{t-\tau} + b_{t-1} - \bar{b}_{t-1}, \\
\phi_{t-1} &= (A_{t-1} - \bar{A}_{t-1})(\hat{Q}_i^{t-1} - \hat{Q}_i^{t-\tau}).
\end{aligned}$$

Clearly, ϵ_{t-1} is $\mathcal{F}_t$-measurable, and satisfies

$$\begin{aligned}
\mathbb{E}\epsilon_{t-1}|\mathcal{F}_{t-\tau} &= \mathbb{E}[(A_{t-1} - \bar{A}_{t-1})\hat{Q}_i^{t-\tau} + b_{t-1} - \bar{b}_{t-1}|\mathcal{F}_{t-\tau}] \\
&= \mathbb{E}[(A_{t-1} - \bar{A}_{t-1})|\mathcal{F}_{t-\tau}]\hat{Q}_i^{t-\tau} + \mathbb{E}[b_{t-1} - \bar{b}_{t-1}|\mathcal{F}_{t-\tau}] \\
&= 0,
\end{aligned} \tag{29}$$

where in the last equality we have used

$$\begin{aligned}
\mathbb{E}[A_{t-1}|\mathcal{F}_{t-\tau}] &= \mathbb{E}[A(z(t-1), z(t))|\mathcal{F}_{t-\tau}] = \mathbb{E}\tilde{A}(z(t-1))|\mathcal{F}_{t-\tau} = \bar{A}^{d_{t-1}} = \bar{A}_{t-1}, \\
\mathbb{E}[b_{t-1}|\mathcal{F}_{t-\tau}] &= \mathbb{E}b(z(t-1))|\mathcal{F}_{t-\tau} = \bar{b}^{d_{t-1}} = \bar{b}_{t-1},
\end{aligned}$$

per the definition of d_{t-1}.

Equation (29) shows that ϵ_{t-1} is a "shifted" martingale difference sequence.[3] 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 12. 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 12 requires more than direct application of the Azuma-Hoeffding bound. For more details, see the full proof of Lemma 12 in Appendix C.5.

Lemma 12 *We have with probability* $1-\delta$*,*

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

Finally we bound sequence $\| \sum_{k=\tau}^{t-1} \alpha_k \tilde{B}_{k,t}\phi_k \|_\infty$, primarily using the fact each $\phi_{t-1} = (A_{t-1} - \bar{A}_{t-1})(\hat{Q}_i^{t-1} - \hat{Q}_i^{t-\tau})$ can be bounded by the movement of the $\hat{Q}_i^t$ function after τ steps (i.e. $\|\hat{Q}_i^{t-1} - \hat{Q}_i^{t-\tau}\|_\infty$), which is quite small due to the step size selection. The proof of Lemma 13 can also be found in Section C.5.

Lemma 13 *The following inequality holds 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 5: bounding the critic error and proof of Theorem 7. We are now ready to use the error decomposition in Lemma 11 as well as the bound on ϵ_k, ϕ_k-sequences in Lemma 12 and Lemma 13 to bound the error of the critic. Recall that Theorem 7 states with probability $1-\delta$,

$$a_T \leq \frac{C_a}{\sqrt{T+t_0}} + \frac{C_a'}{T+t_0} + \frac{C_0}{1-\gamma}, \tag{30}$$

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

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

To prove (30), we start by applying Lemma 12 to $t \leq T$ with δ replaced by δ/T. Then, using a union bound, we get with probability $1-\delta$, for any $t \leq T$,

$$\Big\| \sum_{k=\tau}^{t-1} \alpha_k \tilde{B}_{k,t}\epsilon_k \Big\|_\infty \leq C_\epsilon \frac{1}{\sqrt{t+t_0}},$$

where $C_\epsilon = 6\bar{\epsilon}\sqrt{\frac{\tau h}{\sigma}[\log(\frac{2\tau T^2}{\delta}) + f(\kappa)\log SA]}$. Combine the above with Lemma 11 and use Lemma 13, we get with probability $1-\delta$, for all $\tau \leq t \leq T$,

$$a_t \leq \tilde{\beta}_{\tau-1,t}a_\tau + \gamma \sup_{z_{N_i^\kappa}} \sum_{k=\tau}^{t-1} b_{k,t}(z_{N_i^\kappa})a_k + C_\epsilon \frac{1}{\sqrt{t+t_0}} + C_\phi \frac{1}{t+t_0} + C_0. \tag{31}$$

3. It is not a standard martingale difference sequence, which would require $\mathbb{E}\epsilon_{t-1}|\mathcal{F}_{t-1} = 0$.

We now condition on (31) is true and use induction to show (30). Eq. (30) is true for $t = \tau$, as $\frac{C_a'}{\tau+t_0} \geq \frac{2}{1-\sqrt{\gamma}}\frac{2\bar{r}}{1-\gamma} > a_\tau$, where we have used $|a_\tau| \leq \|Q_i^*\|_\infty + \|\hat{Q}_i^\tau\|_\infty \leq \frac{2\bar{r}}{1-\gamma}$. Then, assume (30) is true for up to $k \leq t-1$, we have by (31),

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

We use the following auxiliary Lemma, whose proof is provided in Section C.6.

Lemma 14 *Recall $\alpha_k = \frac{h}{k+t_0}$, and $b_{k,t}(z_{N_i^\kappa}) = \alpha_k d_k(z_{N_i^\kappa}) \prod_{\ell=k+1}^{t-1}(1 - \alpha_\ell d_\ell(z_{N_i^\kappa}))$, here $d_k(z_{N_i^\kappa}) \geq \sigma$. 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 14, and using the bound on $\tilde{\beta}_{\tau-1,t}$ in Lemma 10 (a), we have

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

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

$$\begin{aligned} F_t \frac{\sqrt{t+t_0}}{C_a} &= \sqrt{\gamma} + \frac{C_\epsilon}{C_a}, \\ F_t' \frac{t+t_0}{C_a'} &= \sqrt{\gamma} + \frac{C_\phi}{C_a'} + \frac{a_\tau(\tau+t_0)}{C_a'} \frac{(\tau+t_0)^{\sigma h-1}}{(t+t_0)^{\sigma h-1}}. \end{aligned}$$

So, we can require C_a, C_a' to be large enough such that

$$\frac{C_\epsilon}{C_a} \leq 1 - \sqrt{\gamma}, \quad \frac{C_\phi}{C_a'} \leq \frac{1-\sqrt{\gamma}}{2}, \quad \frac{a_\tau(\tau+t_0)}{C_a'} \leq \frac{1-\sqrt{\gamma}}{2}.$$

Using $a_\tau \leq \frac{2\bar{r}}{1-\gamma}$, one can check our selection of C_a and C_a' satisfies the above three inequalities, and so the induction is finished and the proof of Theorem 7 is concluded.

C.2. Proof of Lemma 8

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

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

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

where TD is the Bellman operator for reward r_i defined in (14), and operator g^d is given by $g^d(\hat{Q}_i) = \Pi^d\text{TD}\Phi\hat{Q}_i$.

Notice that Φ is non-expansive in $\|\cdot\|_\infty$ norm since each row of Φ has precisely one entry being 1 and all others are zero. Also since Π^d is non-expansive in $\|\cdot\|_\infty$ norm and TD is a γ-contraction in $\|\cdot\|_\infty$ norm, we have $g^d = \Pi^d\text{TD}\Phi$ 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 (20), 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 (14). We have,

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

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

$$\|\Phi\hat{Q}_i^d - Q_i^*\|_\infty \le \frac{1}{1-\gamma}\|\Phi\Pi^d Q_i^* - Q_i^*\|_\infty. \tag{32}$$

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

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

And therefore,

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

Further, we have

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

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

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

Combining the above with (32), we get the desired result

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

■

C.3. Proof of Lemma 9 and Lemma 10

In this section, we provide proofs of the two auxiliary lemmas, Lemma 9 and Lemma 10. We start with the proof of Lemma 9.

Proof of Lemma 9. First, notice that $A(z, z') = \mathbf{e}_{z_{N_i^\kappa}}[\gamma \mathbf{e}^T_{z'_{N_i^\kappa}} - \mathbf{e}^T_{z_{N_i^\kappa}}]$ and $b(z) = \mathbf{e}_{z_{N_i^\kappa}} r_i(z_i)$. As such, $\|A(z, z')\|_\infty \le 1 + \gamma < 2$, $\|b(z)\|_\infty \le \bar{r}$.

Part (a) can be proved by induction. Part (a) is true for $t = 0$ as $\hat{Q}_i^0 = 0$. Assume $\|\hat{Q}_i^{t-1}\|_\infty \le \frac{\bar{r}}{1-\gamma}$. Recall the update equation (15),

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

or in other words,

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

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

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

Notice that

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

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

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

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

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

Therefore, by triangle inequality,

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

As a consequence,

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

■

Proof of Lemma 10. Notice that $\log(1-x) \leq -x$ for all $x < 1$. Then,

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

Therefore,

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

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

For part (b),

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

where we have 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 have used $2\sigma h - 1 > \sigma h$.

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

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

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

C.4. Proof of Lemma 11

Let the $z_{N_i^\kappa}$'th diagonal entry of $B_{k,t}$ be $b_{k,t}(z_{N_i^\kappa})$, and that of $\tilde{B}_{k,t}$ be $\tilde{b}_{k,t}(z_{N_i^\kappa})$. Using these notations, equation (28) can be written as,

$$\begin{aligned}
\hat{Q}_i^t(z_{N_i^\kappa}) = \overbrace{\tilde{b}_{\tau-1,t}(z_{N_i^\kappa}) \hat{Q}_i^\tau(z_{N_i^\kappa}) + \sum_{k=\tau}^{t-1} b_{k,t}(z_{N_i^\kappa}) [g_k(\hat{Q}_i^k)](z_{N_i^\kappa})}^{:= G(z_{N_i^\kappa})} \\
+ \sum_{k=\tau}^{t-1} \alpha_k \tilde{b}_{k,t}(z_{N_i^\kappa}) (\epsilon_k(z_{N_i^\kappa}) + \phi_k(z_{N_i^\kappa})). \qquad (33)
\end{aligned}$$

Notice that by (27), $\tilde{b}_{\tau-1,t}(z_{N_i^\kappa}) + \sum_{k=\tau}^{t-1} b_{k,t}(z_{N_i^\kappa}) = 1$. Then,

$$
\begin{aligned}
|G(z_{N_i^\kappa}) - Q_i^*(z)| &\leq \tilde{b}_{\tau-1,t}(z_{N_i^\kappa})|\hat{Q}_i^\tau(z_{N_i^\kappa}) - Q_i^*(z)| + \sum_{k=\tau}^{t-1} b_{k,t}(z_{N_i^\kappa})|[g_k(\hat{Q}_i^k)](z_{N_i^\kappa}) - Q_i^*(z)| \\
&\leq \tilde{b}_{\tau-1,t}(z_{N_i^\kappa})|\hat{Q}_i^\tau(z_{N_i^\kappa}) - Q_i^*(z)| + \sum_{k=\tau}^{t-1} b_{k,t}(z_{N_i^\kappa})|[g_k(\hat{Q}_i^k)](z_{N_i^\kappa}) - \hat{Q}_i^{d_k}(z_{N_i^\kappa})| \\
&\quad + \sum_{k=\tau}^{t-1} b_{k,t}(z_{N_i^\kappa})|Q_i^*(z) - \hat{Q}_i^{d_k}(z_{N_i^\kappa})| \\
&\leq \tilde{b}_{\tau-1,t}(z_{N_i^\kappa})|\hat{Q}_i^\tau(z_{N_i^\kappa}) - Q_i^*(z)| + \gamma \sum_{k=\tau}^{t-1} b_{k,t}(z_{N_i^\kappa})\|\hat{Q}_i^k - \hat{Q}_i^{d_k}\|_\infty \\
&\quad + \sum_{k=\tau}^{t-1} b_{k,t}(z_{N_i^\kappa})\|Q_i^* - \Phi\hat{Q}_i^{d_k}\|_\infty \\
&\leq \tilde{b}_{\tau-1,t}(z_{N_i^\kappa})|\hat{Q}_i^\tau(z_{N_i^\kappa}) - Q_i^*(z)| + \gamma \sum_{k=\tau}^{t-1} b_{k,t}(z_{N_i^\kappa})\|\Phi\hat{Q}_i^k - Q_i^*\|_\infty \\
&\quad + 2\sum_{k=\tau}^{t-1} b_{k,t}(z_{N_i^\kappa})\|Q_i^* - \Phi\hat{Q}_i^{d_k}\|_\infty \\
&\leq \tilde{\beta}_{\tau-1,t}a_\tau + \gamma \sum_{k=\tau}^{t-1} b_{k,t}(z_{N_i^\kappa})a_k + \frac{2c\rho^{\kappa+1}}{1-\gamma},
\end{aligned} \tag{34}
$$

where in the thrid inequality, we have used that g_k is γ-contraction in infinity norm with fixed point $\hat{Q}_i^{d_k}$, and in the last inequality, we have used (24). Combining the above with (33), we have

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

■

C.5. Proof of Lemma 12 and Lemma 13

Given the work done above, notice that Lemma 9 (c) and Lemma 10 (c) imply the bound on $\|\sum_{k=\tau}^{t-1}\alpha_k\tilde{B}_{k,t}\phi_k\|_\infty$ in Lemma 13, and so the lemma follows directly. So, in this section, we focus on the proof of Lemma 12. We start by stating a variant of the Azuma-Hoeffding bound that handles our "shifted" Martingale difference sequence.

Lemma 15 *Let X_t be a $\mathcal{F}_t$-adapted stochastic process, satisfying $\mathbb{E}X_t|\mathcal{F}_{t-\tau} = 0$. Further, $|X_t| \leq \bar{X}_t$ almost surely. Then with probability $1-\delta$, we have,*

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

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

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

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

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

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

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

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

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

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

Recall that Lemma 12 is an upper bound on $\|\sum_{k=\tau}^{t-1}\alpha_k\tilde{B}_{k,t}\epsilon_k\|$, where $\sum_{k=\tau}^{t-1}\alpha_k\tilde{B}_{k,t}\epsilon_k$ is a random vector in $\mathbb{R}^{\mathcal{Z}_{N_i^\kappa}}$, with its $z_{N_i^\kappa}$'th entry being

$$\sum_{k=\tau}^{t-1}\alpha_k\epsilon_k(z_{N_i^\kappa})\prod_{\ell=k+1}^{t-1}(1-\alpha_\ell d_\ell(z_{N_i^\kappa})), \tag{35}$$

with $d_\ell(z_{N_i^\kappa}) \geq \sigma$ almost surely, cf. (25). Fixing $z_{N_i^\kappa}$, as have been shown in (29), $\epsilon_k(z_{N_i^\kappa})$ is a $\mathcal{F}_{k+1}$ adapted stochastic process satisfying $\mathbb{E}\epsilon_k(z_{N_i^\kappa})|\mathcal{F}_{k+1-\tau} = 0$. However, $\prod_{\ell=k+1}^{t-1}(1-\alpha_\ell d_\ell(z_{N_i^\kappa}))$ is not $\mathcal{F}_{k+1-\tau}$-measurable, and as such we cannot directly apply the Azuma-Hoeffding bound in Lemma 15 to quantity (35). In what follows, we first show in Lemma 16 that almost surely, the absolute value of quantity (35) can be upper bounded by the sup of another quantity, to which we can directly apply Lemma 15. With the help of Lemma 16, we can use the Azuma-Hoeffding bound to control (35) and prove Lemma 12.

Lemma 16 *For each $z_{N_i^\kappa}$, we have almost surely,*

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

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

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

Then $p_t = \sum_{k=\tau}^{t-1} \alpha_k \epsilon_k(z_{N_i^\kappa}) \prod_{\ell=k+1}^{t-1}(1-\alpha_\ell d_\ell(z_{N_i^\kappa}))$, and to prove Lemma 16 we need to bound $|p_t|$. Let

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

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

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

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

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

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

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

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

Notice,

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

By the definition of k_0, we have

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

where in the last step, we have used the upper bound on $\|\epsilon_{k_0}\|_\infty$ in Lemma 9 (b). As a result,

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

■

With the above preparations, we are now ready to prove Lemma 12.

Proof of Lemma 12. Fix $z_{N_i^\kappa}$ and $\tau \le k_0 \le t-1$. As have been shown in (29), $\epsilon_k(z_{N_i^\kappa})\beta_{k,t}$ is a $\mathcal{F}_{k+1}$ adapted stochastic process satisfying $\mathbb{E}\epsilon_k(z_{N_i^\kappa})\beta_{k,t}|\mathcal{F}_{k+1-\tau} = 0$. Also by Lemma 9(b), $|\epsilon_k(z_{N_i^\kappa})\beta_{k,t}| \le \bar{\epsilon}\beta_{k,t}$ almost surely. As a result, we can use the Azuma-Hoeffding bound in Lemma 15 to get with probability $1-\delta$,

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

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

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

Then, by Lemma 16, we have with probability $1-\delta$,

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

where in the third inequality, we have used the bounds on $\beta_{k,t}$ in Lemma 10. Finally, apply the union bound over $z_{N_i^\kappa} \in \mathcal{Z}_{N_i^\kappa}$, and noticing that $|N_i^\kappa| \leq f(\kappa)$ and $|\mathcal{Z}_{N_i^\kappa}| \leq (SA)^{f(\kappa)}$ by Assumption 1, we have with probability $1-\delta$,

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

■

C.6. Proof of Lemma 14

Throughout the proof, we fix $z_{N_i^\kappa}$ and prove the desired upper bounded. For notational simplicity, we drop the dependence on $z_{N_i^\kappa}$ and write $b_{k,t}$ and d_k instead, and we will use the property $d_k \geq \sigma$. 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 $e_{\tau+1} = b_{\tau,\tau+1}\frac{1}{(\tau+t_0)^\omega} = \alpha_\tau d_\tau \frac{1}{(\tau+t_0)^\omega} \leq \frac{1}{\sqrt{\gamma}(\tau+1+t_0)^\omega}$ (last step needs $\alpha_\tau \leq \frac{1}{2}, (1+\frac{1}{t_0})^\omega \leq \frac{2}{\sqrt{\gamma}}$, implied by $t_0 \geq 1, \omega \leq 1$). Let the statement be true for $t-1$. Then, notice that,

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

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

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

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

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

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

Appendix D. Analysis of the Actor and Proof of Theorem 5

In this section, we analyze the actor step. Recall that at iteration m,

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

with $\eta_m = \frac{\eta}{\sqrt{m+1}}$ and $\hat{g}_i(m)$ is given by

$$\hat{g}_i(m) = \sum_{t=0}^{T} \gamma^t \frac{1}{n} \sum_{j \in N_i^\kappa} \hat{Q}_j^{m,T}(s_{N_j^\kappa}(t), a_{N_j^\kappa}(t)) \nabla_{\theta_i} \log \zeta_i^{\theta_i(m)}(a_i(t)|s_i(t)), \tag{36}$$

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,

$$\nabla_{\theta_i} J(\theta(m)) = \sum_{t=0}^{\infty} \mathbb{E}_{s\sim\pi_t^{\theta(m)}, a\sim\zeta^{\theta(m)}(\cdot|s)} \gamma^t Q^{\theta(m)}(s,a) \nabla_{\theta_i} \log \zeta^{\theta(m)}(a|s), \tag{37}$$

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. (3).

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

$$g_i(m) = \sum_{t=0}^{T} \gamma^t \frac{1}{n} \sum_{j\in N_i^\kappa} Q_j^{\theta(m)}(s(t), a(t)) \nabla_{\theta_i} \log \zeta_i^{\theta_i(m)}(a_i(t)|s_i(t)), \tag{38}$$

$$h_i(m) = \sum_{t=0}^{T} \mathbb{E}_{s\sim\pi_t^{\theta(m)}, a\sim\zeta^{\theta(m)}(\cdot|s)} \gamma^t \frac{1}{n} \sum_{j\in N_i^\kappa} Q_j^{\theta(m)}(s,a) \nabla_{\theta_i} \log \zeta_i^{\theta_i(m)}(a_i|s_i), \tag{39}$$

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

Lemma 17 *We have almost surely, $\forall m \le M$,*

$$\max(\|\hat{g}(m)\|, \|g(m)\|, \|h(m)\|, \|\nabla J(\theta(m))\|) \le \frac{\bar{r}L}{(1-\gamma)^2}.$$

Proof Overview. Our main proof idea is the following decomposition,

$$\hat{g}(m) = \underbrace{\hat{g}(m) - g(m)}_{e^1(m)} + \underbrace{g(m) - h(m)}_{e^2(m)} + \underbrace{h(m) - \nabla J(\theta(m))}_{e^3(m)} + \nabla J(\theta(m)), \tag{40}$$

where the error between the gradient estimator $\hat{g}(m)$ and the true gradient $\nabla J(\theta(m))$ is decomposed into the sum of three terms. In Step 1, we bound the first term $\|e^1(m)\|$ which is a direct consequence of our result in the analysis of the critic, cf. Theorem 7 in Appendix C. In Step 2, we study $e^2(m)$, which turns out to be a martingale difference sequence and can be controlled by Azuma-Hoeffding bound. In Step 3, we bound $e^3(m)$, and finally in Step 4, we combine the bounds on $e^1(m), e^2(m)$ and $e^3(m)$ to prove our main result Theorem 5.

Step 1: bounds on $e^1(m)$. Notice that the difference between $\hat{g}_i(m)$ and $g_i(m)$ is that the critic estimate $\hat{Q}_j^{m,T}$ is replaced with the true Q-function $Q_j^{\theta(m)}$. By Theorem 7, 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 18. The proof of Lemma 18 is postponed to Appendix D.2.

Lemma 18 *When T is large enough s.t. $\frac{C_a(\frac{\delta}{2nM},T)}{\sqrt{T+t_0}} + \frac{C_a'}{T+t_0} \leq \frac{2c\rho^{\kappa+1}}{(1-\gamma)^2}$, where*

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

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

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

Step 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 $h_i(m)$. Further, by the way that the trajectory $\{(s(t),a(t))\}_{t=0}^T$ is generated, we have $\mathbb{E}g(m)|\mathcal{G}_{m-1} = h(m)$. As such, $\eta_m\langle\nabla J(\theta(m)), e^2(m)\rangle$ is a martingale difference sequence w.r.t. $\mathcal{G}_m$, and we have the following bound in Lemma 19 which is a direct consequence of Azuma-Hoeffding bound. The proof of Lemma 19 is postponed to Section D.3.

Lemma 19 *With probability at least $1-\delta/2$,*

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

Step 3: bounds on $e^3(m)$. We have the following Lemma 20 that bounds $\|e^3(m)\|$. Its proof is quite similar to that of Lemma 4 and is postponed to Appendix D.4.

Lemma 20 *When $T+1 \geq \frac{\log\frac{c(1-\gamma)}{\bar{r}} + (\kappa+1)\log\rho}{\log\gamma}$, we have almost surely,*

$$\|e^3(m)\| \leq 2\frac{Lc}{(1-\gamma)}\rho^{\kappa+1}.$$

Step 4: Proof of Theorem 5. With the above bounds on $e^1(m)$, $e^2(m)$ and $e^3(m)$, we are now ready to prove the main result Theorem 5. Since $\nabla J(\theta)$ is L' Lipschitz continuous, we have

$$\begin{aligned} J(\theta(m+1)) &\geq J(\theta(m)) + \langle\nabla J(\theta(m)), \theta(m+1)-\theta(m)\rangle - \frac{L'}{2}\|\theta(m+1)-\theta(m)\|^2 \\ &= J(\theta(m)) + \eta_m\langle\nabla J(\theta(m)), \hat{g}(m)\rangle - \frac{L'\eta_m^2}{2}\|\hat{g}(m)\|^2. \end{aligned} \tag{41}$$

Recall the decomposition of $\hat{g}(m)$,

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

Then,

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

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

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

Plug the above bounds on $\|\hat{g}(m)\|^2$ and $\langle \nabla J(\theta(m)), \hat{g}(m) \rangle$ into (41), we have,

$$J(\theta(m+1)) \geq J(\theta(m)) + (\eta_m - 2L'\eta_m^2)\|\nabla J(\theta(m))\|^2 + \eta_m \varepsilon_{m,0} - \eta_m \varepsilon_{m,1} - \eta_m^2 \varepsilon_{m,2}, \quad (42)$$

where

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

Doing a telescope sum for (42), we get

$$\begin{aligned}J(\theta(M)) &\geq J(\theta(0)) + \sum_{m=0}^{M-1} (\eta_m - 2L'\eta_m^2)\|\nabla J(\theta(m))\|^2 + \sum_{m=0}^{M-1} \eta_m \varepsilon_{m,0} - \sum_{m=0}^{M-1} \eta_m \varepsilon_{m,1} - \sum_{m=0}^{M-1} \eta_m^2 \varepsilon_{m,2} \\ &\geq J(\theta(0)) + \sum_{m=0}^{M-1} \frac{1}{2}\eta_m \|\nabla J(\theta(m))\|^2 + \sum_{m=0}^{M-1} \eta_m \varepsilon_{m,0} - \sum_{m=0}^{M-1} \eta_m \varepsilon_{m,1} - \sum_{m=0}^{M-1} \eta_m^2 \varepsilon_{m,2},\end{aligned} \quad (43)$$

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

$$\sum_{m=0}^{M-1} \frac{1}{2}\eta_m \|\nabla J(\theta(m))\|^2 \leq J(\theta(M)) - J(\theta(0)) - \sum_{m=0}^{M-1} \eta_m \varepsilon_{m,0} + \sum_{m=0}^{M-1} \eta_m \varepsilon_{m,1} + \sum_{m=0}^{M-1} \eta_m^2 \varepsilon_{m,2}. \quad (44)$$

We now apply our results in the first three steps. By Lemma 19, we have with probability $1 - \frac{\delta}{2}$,

$$\Big| \sum_{m=0}^{M-1} \eta_m \varepsilon_{m,0} \Big| \leq \frac{2\bar{r}^2 L^2}{(1-\gamma)^4} \sqrt{2 \sum_{m=0}^{M-1} \eta_m^2 \log \frac{4}{\delta}}. \quad (45)$$

By Lemma 18 and Lemma 20, we have with probability $1 - \frac{\delta}{2}$,

$$\begin{aligned}\sup_{m \leq M-1} \varepsilon_{m,1} &\leq \frac{\bar{r}L}{(1-\gamma)^2} (\sup_{m \leq M-1} \|e^1(m)\| + \sup_{m \leq M-1} \|e^3(m)\|) \\ &\leq \frac{\bar{r}L}{(1-\gamma)^2} (\frac{4cL\rho^{\kappa+1}}{(1-\gamma)^3} + 2\frac{Lc}{(1-\gamma)}\rho^{\kappa+1}) \\ &\leq \frac{6L^2 c\bar{r}}{(1-\gamma)^5}\rho^{\kappa+1}.\end{aligned} \quad (46)$$

By Lemma 17, 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 almost surely,

$$\sup_{m\leq M-1} \varepsilon_{m,2} = 2L'(\|e^1(m)\|^2 + \|e^2(m)\|^2 + \|e^3(m)\|^2)$$
$$\leq \frac{24\bar{r}^2 L' L^2}{(1-\gamma)^4}. \quad (47)$$

Using a union bound, we have with probability $1-\delta$, all three events (45), (46) and (47) hold, which when combined with (44) implies

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

Since $\eta_m = \frac{\eta}{\sqrt{m+1}}$, 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 $M \geq 3$). Further 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 $1-\delta$,

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

This concludes the proof of the main Theorem 5. ■

D.1. Proof of Lemma 17

Recall that

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

Therefore,

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

where we have used that $\|\hat{Q}_j^{m,T}\|_\infty \le \frac{\bar{r}}{1-\gamma}$ almost surely (cf. Lemma 9 (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 18

In this section, we prove Lemma 18.

Proof of Lemma 18 Let $\mathcal{G}_m$ be the σ-algebra generated by the trajectories in the first m outer-loop iterations. Then, Theorem 7 implies that, fixing each $m \le M$ and $i \in \mathcal{N}$, conditioned on $\mathcal{G}_{m-1}$, the following event happens with probability at least $1-\delta$:

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

where

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

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

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

$$\begin{aligned}\sup_{m\le M-1}\sup_{i\in\mathcal{N}}\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left| Q_i^{\theta(m)}(s,a) - \hat{Q}_i^{m,T}(s_{N_i^\kappa}, a_{N_i^\kappa}) \right| &\le \frac{C_a(\frac{\delta}{2nM},T)}{\sqrt{T+t_0}} + \frac{C_a'}{T+t_0} + \frac{2c\rho^{\kappa+1}}{(1-\gamma)^2} \\ &\le \frac{4c\rho^{\kappa+1}}{(1-\gamma)^2},\end{aligned} \tag{49}$$

where in the last step, we have used that our lower bound on T implies $\frac{C_a(\frac{\delta}{2nM},T)}{\sqrt{T+t_0}} + \frac{C_a'}{T+t_0} \le \frac{2c\rho^{\kappa+1}}{(1-\gamma)^2}$. Therefore, conditioned on (49) being true, we have for any $m \le M-1$ and any $i \in \mathcal{N}$,

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

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

D.3. Proof of Lemma 19

By Lemma 17, we have almost surely,

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

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

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

■

D.4. Proof of Lemma 20

In this section, we provide the proof of Lemma 20.

Proof of Lemma 20 By (37), we have

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

where we have used $\nabla_{\theta_i}\log\zeta^{\theta(m)}(a|s)=\nabla_{\theta_i}\sum_{j\in\mathcal{N}}\log\zeta_j^{\theta_j(m)}(a_j|s_j)=\nabla_{\theta_i}\log\zeta_i^{\theta_i(m)}(a_i|s_i)$. Also recall the definition of $h_i(\theta)$ in (39),

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

Combining the above two equations, we have,

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

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

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

where $\hat{Q}_j^{\theta(m)}$ is any truncated Q function for $Q_j^{\theta(m)}$ as defined in (5). We claim E_4 is zero. To see this, consider for any $j \in N_{-i}^{\kappa}$ and any t,

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

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

For E_3, by the exponential decay property, the truncated Q function has a small error, cf. (11),

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

and as a result,

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

Therefore,

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

where in the last step, we have used

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

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