Global Convergence of Localized Policy Iteration in Networked Multi-Agent Reinforcement Learning

YIZHOU ZHANG*, Tsinghua University, China
GUANNAN QU*, Carnegie Mellon University, United States
PAN XU*, Duke University, United States
YIHENG LIN, California Institute of Technology, United States
ZAIWEI CHEN, California Institute of Technology, United States
ADAM WIERMAN, California Institute of Technology, United States

We study a multi-agent reinforcement learning (MARL) problem where the agents interact over a given network. The goal of the agents is to cooperatively maximize the average of their entropy-regularized long-term rewards. To overcome the curse of dimensionality and to reduce communication, we propose a Localized Policy Iteration (LPI) algorithm that provably learns a near-globally-optimal policy using only local information. In particular, we show that, despite restricting each agent's attention to only its κ-hop neighborhood, the agents are able to learn a policy with an optimality gap that decays polynomially in κ. In addition, we show the finite-sample convergence of LPI to the global optimal policy, which explicitly captures the trade-off between optimality and computational complexity in choosing κ. Numerical simulations demonstrate the effectiveness of LPI.

ACM Reference Format:
Yizhou Zhang, Guannan Qu, Pan Xu, Yiheng Lin, Zaiwei Chen, and Adam Wierman. 2023. Global Convergence of Localized Policy Iteration in Networked Multi-Agent Reinforcement Learning. *Proc. ACM Meas. Anal. Comput. Syst.* 7, 1, Article 13 (March 2023), 51 pages. https://doi.org/10.1145/3579443

1 INTRODUCTION

Reinforcement learning (RL) has seen remarkable successes in recent years, many of which fall into the multi-agent setting, such as playing multi-agent games [26, 31], smart grid [9], queueing networks [37], etc. In this work, we focus on a form of *networked* Multi-Agent RL (MARL) where the agents interact according to a given network graph.

Compared to single-agent RL, MARL faces many additional challenges. First of all, the curse of dimensionality (which is already a major challenge in single-agent RL) becomes a more severe issue in MARL because the complexity of the problem scales exponentially with the number of agents [6, 16, 18, 22]. This is because the size of the state and action space scales exponentially with the number of agents and, as a result, the dimension of the value/Q-functions and the policies all scale exponentially with the number of agents, which is hard to compute and store in moderately large networks [4]. Moreover, since the agents are coupled by the global state, the training process requires extensive communication among the agents in the entire network.

*Yizhou Zhang, Guannan Qu, Pan Xu contributed equally to this work.

Authors' addresses: Yizhou Zhang, Tsinghua University, China; Guannan Qu, Carnegie Mellon University, United States; Pan Xu, Duke University, United States; Yiheng Lin, California Institute of Technology, United States; Zaiwei Chen, California Institute of Technology, United States; Adam Wierman, California Institute of Technology, United States.

2476-1249/2023/3-ART13
https://doi.org/10.1145/3579443

To overcome the aforementioned challenges, the existing literature considers performing MARL with only local information. For example, it has been proposed that each agent's policy depends only on the states of itself and, potentially, its neighboring agents. An example of this type is the independent learners approach [10, 35], where each agent learns a policy for itself while treating the states and actions of other agents as part of the environment. Another example is the recent work of Lin et al. [21], Qu et al. [28, 29], which focuses on learning localized policies where each agent is allowed to choose its action based on its own and neighbors' states. Beyond the localization in policy, it has also been proposed to approximate each agent's value or Q functions in a way that they only depend on the local and nearby agents' states and actions [16, 21, 28, 29, 39], as opposed to the full state. These approaches greatly relieve the computation and communication burden both empirically [13, 32, 35] and theoretically [21, 28, 29].

Despite this progress, there are still major limitations. As discussed above, in order to speed up learning, many previous works, e.g., [10, 21, 28, 29, 35], restrict consideration to localized policies where each agent makes decisions only based on the local state of the agent and, potentially, neighbors, as opposed to the global state. This leads to a fundamental performance gap between the class of localized policies considered and the optimal centralized policy. Even when the best localized policy can be found, there is a performance degradation compared to the best centralized policy, as the centralized policy class is a strict superset of the localized policy class. An important open question that remains is the following:

How large is the gap between localized policies and the optimal centralized policy?
How much information must be available to each local agent
in order to achieve a near-optimal performance?

This question has received increasing attention in linear control settings [3, 27, 30, 42], but is still open in MARL settings. In this work, we provide the first bounds on the gap between localized and centralized policies under a MARL model in networked systems.

Another limitation of prior work studying MARL is that, while existing results have provided convergence bounds for local policies, the convergence of their methods is typically only to a suboptimal policy in the localized policy class. For example, in Lin et al. [21], Qu et al. [28], the policies are shown to converge to a stationary point of the objective function defined on localized policies, as opposed to the global optimum. This is unsatisfying as converging to a stationary point does not even guarantee converging to the best localized policy. Therefore, another important and open question that remains is the following:

Is it possible to design a MARL algorithm that provably
finds a near-globally-optimal policy using only local information?

This question has received increasing attention in single-agent settings, where people have studied the convergence to the global optimum for policy-gradient methods under various policy classes [2, 5]. However, it is open in the context of learning localized policies in MARL and in this work, we provide an affirmative answer to this question under an MARL model in networked systems.

1.1 Contributions

Motivated by the open questions above, our work proposes and analyzes a new class of localized policies for networked MARL and proposes a Localized Policy Iteration (LPI) algorithm that converges to a near-globally-optimal policy. Our main contributions are summarized in the following.

- **Near-Globally-Optimal κ-Hop Localized Policies.** We show that a class of κ-hop localized policies are nearly globally optimal, where a κ-hop localized policy means that each agent is allowed to choose its action based on the states of its κ-hop local neighborhood. More specifically, in Theorem 1, we show that there exists a κ-hop localized policy whose optimality gap is

polynomially small in κ, where the optimality gap is with respect to the best centralized policy that is allowed to depend on the global state. As a result, even with a small κ, the class of κ-hop localized policies is near optimal despite the fact that each agent only uses information from within a small κ-hop neighborhood to make its decision. This result justifies using localized policies in networked MARL.

- **Localized Policy Iteration.** Motivated by the result described above, we propose a localized MARL algorithm, a.k.a. LPI, given in Algorithm 1. At a high level, LPI iteratively performs policy improvement (and policy evaluation), but restricted to κ-hop policies. While standard policy improvement requires using the information of global states and actions in order to conduct the policy improvement step, we develop a *soft* policy improvement that approximately performs policy improvement to κ-hop localized policies using only local information. Therefore, our proposed algorithm can be implemented in a truly localized manner.
- **Finite-Sample Analysis of LPI.** We provide a global convergence guarantee for LPI in Theorem 2. The guarantee holds for any policy evaluation method used as a subroutine in the evaluation step of LPI (denoted as `PolicyEvaluation`) that satisfies a mild condition specified in Definition 7. Furthermore, in Corollary 3, we provide the finite-sample complexity (cf. (14)) for LPI when choosing a specific `PolicyEvaluation` method proposed by Lin et al. [21]. Specifically, we show that in order to achieve an ε optimality (compared to the best centralized policy), one needs to use a κ-hop localized policy with $\kappa = \Theta(poly(\frac{1}{\varepsilon}))$ in LPI, and the sample complexity in learning such a κ-hop localized policy with ε optimality gap scales polynomially with the largest state-action space size of local neighborhoods, as opposed to the global network.

The key technical novelty in this paper is a policy closure argument (Theorem 6), where we identify a class of policies that satisfy a form of spatial decaying properties and show that they are closed under the entropy regularized Bellman operator in MARL. This key observation implies that when starting from a policy in this class with spatial decaying properties, the policy improvement procedure will result in a new policy within this class, which further reveals that the optimal policy is also in the spatial decaying policy class. We then show that LPI is an inexact version of the above policy iteration given by the regularized Bellman operator, where the learned policies and Q-functions are truncated to only depend on a localized neighborhood of each agent, and the exact policy evaluation and improvement steps are approximated by finite-sample estimations. Therefore, combining the policy closure argument for the exact version of LPI and error bound analyses for the approximation made in LPI, we are able to show its convergence to a near global optimal policy, even though we only use localized information in LPI.

While the main contribution of this work is to theoretically show that LPI can converge to the global optimal policy using only localized information, we also verify its empirical performance on a simulated networked MARL problem. In particular, we design an example of a spreading process over a network where the optimal policy depends on the global states of each agent (not just the local information). We run LPI on this example and the results highlight that LPI can perform well even outside the region suggested by the theory.

1.2 Related Literature

Reinforcement Learning (RL) studies a dynamical environment where the agent decides its current action based on the current state and the current state/action affects the distribution of its next state. This environment is usually modeled as a *Markov Decision Process* (MDP) where the key assumption is the *Markov Property*, which means the current state and action are statistically sufficient for deciding the next state (see, e.g., Sutton and Barto [34]). Specifically, an MDP (with discounted cumulative reward) can be characterized as a tuple $\langle \mathcal{S}, \mathcal{A}, P, \gamma, r \rangle$, where $\mathcal{S}$ and $\mathcal{A}$ denote the

state/action space. The transition function $P : \mathcal{S} \times \mathcal{A} \to \Delta_{\mathcal{S}}$ characterizes the transition probability $\mathbf{P}(s' \mid s, a)$ from the current state/action pair to next state s'. $\gamma \in (0, 1)$ denotes the discount factor, and the reward function $r : \mathcal{S} \times \mathcal{A} \to [0, \bar{r}]$ is nonnegative for the agent with the state/action pair (s, a).

In single-agent RL, the goal of the agent is to learn a policy $\zeta : \mathcal{S} \to \Delta(\mathcal{A})$ to maximize the expected discounted accumulative reward $J(\zeta) := \mathbb{E}_{s(0)\sim\rho, a(t)\sim\zeta(\cdot|s(t))} \left[\sum_{t=0}^{T} \gamma^t r(s(t), a(t))\right]$. Two classic learning algorithms have been proposed to learn the optimal policy when the transition probabilities are known: value iteration and policy iteration [34]. *Value Iteration* (VI) iteratively updates the estimated optimal value function by the Bellman equation and derives the optimal policy from the learned optimal value function. In contrast, *Policy Iteration* (PI) works by evaluating the current policy and iteratively updating the policy using this policy's value function. Our algorithm, LPI, can be viewed as a generalization of PI to a networked setting where the transition probabilities are unknown. Lower bound results show tabular RL algorithms inevitably suffer from *the curse of dimensionality* when the transition probabilities are unknown (see, e.g., Jin et al. [17]), which means the sample complexity (i.e., the number of samples to find a near-optimal policy) grows with respect to the size of state space $\mathcal{S}$ and action space $\mathcal{A}$.

Multi-Agent Reinforcement Learning (MARL) generalizes the single-agent RL setting to a situation where the global MDP evolves based on the joint action/policy of n agents. The generalization to include more agents is necessary for solving practical problems that involve large-scale networks and/or games. Specifically, each agent i has its local action space $\mathcal{A}_i$ and its local reward function r_i, and the global action space $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$. At every time step t, agent i receives a partial observation $o_i(s(t))$ to decide its local action $a_i(t)$. More details about this setting can be found in Section 2. Depending on each agent's learning objective, MARL can be divided into two categories: cooperative or competitive MARL (see [40] for a survey). In *cooperative MARL*, all agents work together to optimize a shared global objective [21, 28, 29], which is also the setting studied in this work. In *competitive MARL*, each agent optimizes its local accumulative reward, and the goal is to find an approximate Nash Equilibrium [11, 19].

A major challenge for cooperative MARL is that the size of the global action space $\mathcal{A}$ can grow exponentially with respect to the number of agents n, which makes centralized training intractable. Fully distributed training is not practical either, because its reward and transition probability depend on other agents' policies. To address this challenge and derive finite-sample complexity bounds that are not exponential in n, a common assumption made by previous works [12, 33] is that the global state is observable to all agents (i.e., $o_i(s) = s$), which thus eliminates the need for communicating local states in the network. This works in special settings such as cooperative navigation and Predator-Prey where a joint state space is easy to observe or all agents share the same state space [23, 41]. Nevertheless, such an assumption is not practical in more general MARL settings like the one studied in our paper, where each agent has its own state space and the global state space $\mathcal{S} = \mathcal{S}_1 \times \cdots \times \mathcal{S}_n$ is exponentially large with respect to n. In this setting, huge communication costs are usually unavoidable and thus it is more challenging and less studied.

Our work is most related to a class of cooperative MARL problems where the agents are located in a network graph $\mathcal{G}$ [21, 28, 29], which makes similar structural assumptions on the MDP as this paper, i.e., each agent has its own local state/action space and the local transition probabilities depend only on the agent's neighbors. Each agent i observes the local states of the agents whose graph distance to i is less than or equal to κ to decide its local action (i.e., $o_i(s) = s_{\mathcal{N}_i^\kappa}$, where $\mathcal{N}_i^\kappa := \{j \mid d_{\mathcal{G}}(i, j) \le \kappa\}$. $d_{\mathcal{G}}(i, j)$ denotes the length of the shortest path between i and j on $\mathcal{G}$). As a remark, all local policies become centralized when the dependence parameter κ exceeds the diameter of graph $\mathcal{G}$. Previous works [21, 28, 29] propose decentralized policy iteration algorithms

to learn local policies for each agent[1] and prove finite-time convergence bounds. However, there is still an open gap between the localized policy and centralized policy where each agent makes decisions based on the global state. Specifically, before our work, there is no result on how large the gap between the optimal κ-hop localized policy and the optimal centralized policy is, nor how fast it decays with respect to κ. Besides, even if the objective is defined with respect to the κ-hop localized policy class, the algorithms proposed by Lin et al. [21], Qu et al. [28, 29] are only guaranteed to converge to stationary points of their local objective functions rather than global optima of the objective.

Another challenge widely encountered in RL is that the policy might rapidly become deterministic during the training process, which further leads to a slow convergence. To speed up the convergence and also encourage exploration for the training algorithm to escape suboptimal points, entropy regularization has often been added to the value function. This approach has yielded both good empirical performance [25, 36, 38] and strong theoretical guarantees [1, 8, 24]. In this paper, we focus on the entropy regularized problem in multi-agent reinforcement learning.

2 MODEL & PRELIMINARIES

In this section, we introduce the model we study and provide preliminaries that are used throughout our paper. In particular, we introduce networked multi-agent Markov decision processes in Section 2.1. Then, in Section 2.2, we define the (state) value function and action-state value function (i.e., Q function) in the multi-agent setting with an entropy regularization. In Section 2.3, we introduce a novel class of polices where the dependence of local policies on other agents' actions decay polynomially as the graph distance increases. This property is key to our analysis. Note that a summary table of notation is given in Table 1 in the appendix for the reader's reference.

2.1 Multi-Agent Markov Decision Process

We consider a network of n agents that are associated with an undirected graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N} = \{1, \dots, n\}$ is the set of nodes and $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$ is the set of edges. We denote the state-space and the action-space of agent i by $\mathcal{S}_i$ and $\mathcal{A}_i$, respectively, and they are both finite sets. We also denote $A_{\max} = \max_{i \in \mathcal{N}} |\mathcal{A}_i|$. The global state is denoted as $s = (s_1, \dots, s_n) \in \mathcal{S} := \mathcal{S}_1 \times \cdots \times \mathcal{S}_n$ and similarly the global action is denoted as $a = (a_1, \dots, a_n) \in \mathcal{A} := \mathcal{A}_1 \times \cdots \times \mathcal{A}_n$. At time $t \geq 0$, given current state $s(t)$ and action $a(t)$, for each $i \in \mathcal{N}$, the next individual state $s_i(t+1)$ is independently generated and is only dependent on its neighbors' states and its own action:

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

where $\mathcal{N}_i = \{i\} \cup \{j \in \mathcal{N} \mid (i,j) \in \mathcal{E}\}$ denotes the neighborhood of i (including i itself) and $s_{\mathcal{N}_i}$ represents the states of the agents in $\mathcal{N}_i$. In addition, for integer $\kappa \geq 0$, we use $\mathcal{N}_i^\kappa$ to denote the κ-hop neighborhood of i, i.e., the nodes of which the graph distance to i have length less than or equal to κ. We also use $\mathcal{N}_{-i}^\kappa = \mathcal{N} / \mathcal{N}_i^\kappa$ to denote the agents that are not in $\mathcal{N}_i^\kappa$. We use $s_{\mathcal{N}_i^\kappa}$ and $a_{\mathcal{N}_i^\kappa}$ to denote the states and actions of the agents in $\mathcal{N}_i^\kappa$ respectively, use s_{-i}, a_{-i} to denote the states and actions of all agents other than i, and we use $f(\kappa) = \sup_{i \in \mathcal{N}} |\mathcal{N}_i^\kappa|$ to denote the size of the largest κ-hop neighborhood.

Each agent is associated with a policy ζ_i, which maps each global state $s \in \mathcal{S}$ to a probability distribution supported on the set of local actions $\mathcal{A}_i$, and we use $\Delta_{\mathcal{A}_i|\mathcal{S}}$ to denote the space ζ_i lies in. Each agent, conditioned on observing $s(t)$, takes an action $a_i(t)$ according to $\zeta_i(\cdot|s(t))$. We use $\zeta(a|s) = \prod_{i=1}^n \zeta_i(a_i|s)$ to denote the joint policy, which is the product of all the individual policies.

[1]The local policy defined in these papers focuses on a local agent that chooses local actions based only on the agent's state.

Since ζ is uniquely determined by the tuple of the ζ_i's, we also slightly abuse the notation and denote $\zeta = (\zeta_1, \zeta_2, \cdots, \zeta_n) \in \Delta_{\mathcal{A}_1|\mathcal{S}} \times \Delta_{\mathcal{A}_2|\mathcal{S}} \times \cdots \times \Delta_{\mathcal{A}_n|\mathcal{S}} := \Delta_{\text{policy}}$.

2.2 Value Function and Q-Function

Each agent $i \in \mathcal{N}$ is associated with a stage reward function $r_i(s_i, a_i)$ that depends on its local state and action. The global stage reward is defined as $r(s, a) = \frac{1}{n}\sum_{i=1}^n r_i(s_i, a_i)$. We assume that all rewards r_i are bounded above by $\bar{r}$ throughout the paper (which can always be satisfied as the local state and action spaces are finite). Given any joint policy $\zeta \in \Delta_{\text{policy}}$, we define the entropy regularized value function at state s as

$$V^\zeta(s) = \mathbb{E}_{a(t)\sim\zeta(\cdot|s(t))}\left[\sum_{t=0}^\infty \gamma^t \left[r(s(t), a(t)) - \tau \log(\zeta(a(t)|s(t)))\right] \,\middle|\, s(0) = s\right], \tag{2}$$

where $\tau > 0$ is a tunable parameter. Based on $V^\zeta(\cdot)$, we further define the objective function as

$$J(\zeta) = \mathbb{E}_{s\sim\rho}[V^\zeta(s)],$$

where $\rho(\cdot)$ is a given initial state distribution. In this work, we aim to find a global optimal policy $\zeta^* \in \Delta_{\text{policy}}$ that maximizes the objective function $J(\zeta)$.

In both the definition of the value function and the objective function, there is a weighted entropy term $-\tau \log(\zeta(a(t)|s(t)))$ with positive weight τ. Entropy regularization has gained popularity in the RL literature for both practical and theoretical reasons. Practically, it is known that adding regularization encourages randomness and exploration of the policy [25, 36, 38], which is necessary in RL. To see this, observe that the optimal policy becomes the uniform policy as the parameter τ goes to infinity. Theoretically, due to the strong concavity of the negative entropy function, it has been shown that entropy regularization helps improving the convergence rate of several RL algorithms, e.g., natural policy gradient [7, 8]. Moreover, we show in Section 3 that the entropy regularization term plays an important role in our multi-agent RL setting as it will affect the suboptimality gap between localized policies and the best centralized policy.

Given the definition of the value function (2), the Q function of a joint policy ζ is defined as

$$Q^\zeta(s, a) = r(s, a) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)} V^\zeta(s'). \tag{3}$$

The value function and the Q-function satisfy the following relationship

$$\begin{aligned} V^\zeta(s) &= \mathbb{E}_{a\sim\zeta(\cdot|s)}\left[r(s, a) - \tau \log \zeta(a|s) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)} V^\zeta(s')\right] \\ &= \mathbb{E}_{a\sim\zeta(\cdot|s)}\left[Q^\zeta(s, a) - \tau \log \zeta(a|s)\right], \end{aligned} \tag{4}$$

where the first line follows from the Bellman equation for $V^\zeta(\cdot)$. In view of the additive structure of the stage rewards $r(s, a)$, we also define in the following the local value function and the local Q function for all $i \in \mathcal{N}$:

$$V_i^\zeta(s) = \mathbb{E}_{a(t)\sim\zeta(\cdot|s(t))}\left[\sum_{t=0}^\infty \gamma^t \left[r_i(s_i(t), a_i(t)) - n\tau \log(\zeta_i(a_i(t)|s(t)))\right] \middle| s(0) = s\right], \tag{5}$$

$$Q_i^\zeta(s, a) = r_i(s_i, a_i) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)} V_i^\zeta(s'). \tag{6}$$

We sometimes use the vector $\mathbf{Q} = (Q_1, \ldots, Q_n)$ to denote the complete profile of all the local Q functions. It is clear from the definition that

$$V^\zeta(s) = \frac{1}{n}\sum_{i=1}^n V_i^\zeta(s), \quad \text{and} \quad Q^\zeta(s, a) = \frac{1}{n}\sum_{i=1}^n Q_i^\zeta(s, a).$$

Similar to the standard Bellman optimal operator, we define the Bellman optimal operator with entropy regularization $\mathcal{T}: \mathbb{R}^{|\mathcal{S}|} \mapsto \mathbb{R}^{|\mathcal{S}|}$ as

$$[\mathcal{T}V](s) = \max_{\zeta \in \Delta_{\text{policy}}} \mathbb{E}_{a\sim\zeta(\cdot|s)} \left[r(s,a) - \tau \log \zeta(a|s) + \gamma \mathbb{E}_{s'\sim P(\cdot|s,a)} V(s') \right], \ \forall\, V \in \mathbb{R}^{|\mathcal{S}|},\ s \in \mathcal{S}. \quad (7)$$

The following result regarding the properties of $\mathcal{T}(\cdot)$ is an immediate extension to that of the standard Bellman optimal operator. For completeness, a proof is presented in Appendix D.1.

Proposition 1. *$\mathcal{T}(\cdot)$ is a contraction mapping with respect to the ℓ_∞-norm, with contraction factor γ. In addition, suppose a policy ζ^* satisfies $V^{\zeta^*} = \mathcal{T}(V^{\zeta^*})$, then ζ^* is an optimal policy.*

2.3 Spatial Decay Properties

Throughout this paper, we use spatial decay properties heavily, i.e. the fact that a certain quantity decays as the distance between two agents increases. Spatial decay properties have been investigated in the literature in various forms, including the exponential decay of Q-functions in [28, 29] and correlation decay in combinatorial optimization [3, 14, 15]. In this paper, we introduce the classes of (ν, μ)-decay polices and (ν, μ)-decay Q-functions. The decay properties of these quantities eventually enable us to control the optimality gap (compared to the best centralized policy) when restricting the agents to using only κ-hop localized policies (to be defined in Definition 5).

Compared to the exponential decay of Q-functions in the literature, our work considers a different type of decay whose rate is polynomial. Further, we extend the decay property from Q-functions to policies which, broadly speaking, is similar in concept to Shin et al. [30], Zhang et al. [42], which studies a similar decaying policy class in linear dynamical systems.

To start, we first introduce the (ν, μ)-decay matrices.

Definition 1 ((ν, μ)-decay matrix). *For $\nu, \mu > 0$, a matrix $A \in \mathbb{R}^{n\times n}$ is said to be (ν, μ)-decay with respect to graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$ with $\mathcal{V} = \{1, 2, \cdots, n\}$ if every entry of A is non-negative and*

$$\max\left\{ \sup_i \sum_{j=1}^n A_{ij}(\text{dist}(i,j)+1)^\mu, \sup_j \sum_{i=1}^n A_{ij}(\text{dist}(i,j)+1)^\mu \right\} \le \nu,$$

where $\text{dist}(i, j)$ is the graph distance between i and j in $\mathcal{G}$.

Intuitively speaking, for a (ν, μ)-decay matrix A, A_{ij} decays polynomially as the graph distance between i and j increases. To see this, observe that we have for all $i \in \mathcal{N}$ and $\kappa \in \mathbb{N}$:

$$\sum_{j:\text{dist}(i,j)>\kappa} A_{ij} \le \sum_{j:\text{dist}(i,j)>\kappa} A_{ij} \frac{(\text{dist}(i,j)+1)^\mu}{(\kappa+1)^\mu} \le \frac{\nu}{(\kappa+1)^\mu}. \quad (8)$$

Based on Definition 1, we next define the (ν, μ)-decay policy class as follows.

Definition 2 ((ν, μ)-decay policy). *Let $\zeta = (\zeta_1, \ldots, \zeta_n)$ be a joint policy, where ζ_i is the policy of agent $i \in \{1, 2, \cdots, n\}$. The interaction matrix of ζ, denoted by $Z^\zeta \in \mathbb{R}^{n\times n}$, is defined as*

$$Z^\zeta_{ij} = \max_{s_{-j}} \max_{s_j, s'_j} \text{TV}(\zeta_i(\cdot|s_j, s_{-j}), \zeta_i(\cdot|s'_j, s_{-j})), \quad \text{for all } i, j \in \{1, 2, \cdots, n\}.$$

A policy ζ is said to be (ν,μ)-decay if Z^ζ is a (ν, μ)-decay matrix.

In a (ν, μ)-decay policy, agent i's action a_i depends on the state of agent j in a manner that the dependence decays polynomially in the graph distance between i and j.

The next definition introduces σ-regular policies.

Definition 3. *A distribution $d = (d_1, \ldots, d_M)$ over a set $\{1, 2, \ldots, M\}$ is called σ-regular if $\max_{m,m' \in [M]} \log(d_m/d_{m'}) \leq \sigma$. A joint policy $\zeta = (\zeta_1, \ldots, \zeta_n)$ is σ-regular if $\forall i, s \in \mathcal{S}$, the distribution $\zeta_i(\cdot|s)$ is σ-regular.*

Throughout the paper, we use $\Delta_{\nu,\mu,\sigma}$ to denote the set of all joint policies that are (ν, μ)-decay and σ-regular. Similarly, we also define a class of decaying Q-functions as follows.

Definition 4 ((ν,μ)-decay Q-function class). *Let $Q_i \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, $i \in \mathcal{N}$ be the local Q-functions defined in (6), and let $\mathbf{Q} = \{Q_i\}_{i\in\mathcal{N}}$. The interaction matrix of $\mathbf{Q}$, denoted by $Z^{\mathbf{Q}} \in \mathbb{R}^{n\times n}$, is defined as*

$$Z^{\mathbf{Q}}_{ij} = \max_{s_{-j},a_{-j}} \max_{s_j,a_j,s'_j,a'_j} |Q_i(s_j, a_j, s_{-j}, a_{-j}) - Q_i(s'_j, a'_j, s_{-j}, a_{-j})|, \quad \forall\, i, j \in \{1, 2, \cdots, n\}.$$

For any given ν and μ, a local Q function tuple $\mathbf{Q} = \{Q_i\}_{i\in\mathcal{N}}$ is said to be (ν, μ)-decay if $Z^{\mathbf{Q}}$ is a (ν, μ)-decay matrix.

One benefit of the (ν, μ)-decay property of $\mathbf{Q} = \{Q_i\}_{i\in\mathcal{N}}$ is that it allows "truncation" of the local Q functions. Specifically, given a positive integer β, and a local Q-function tuple $\mathbf{Q} = \{Q_i\}_{i\in\mathcal{N}}$, we can define the following truncated local Q-functions:

$$\hat{Q}_i(s_{\mathcal{N}_i^\beta}, a_{\mathcal{N}_i^\beta}) = \sum_{s_{\mathcal{N}_{-i}^\beta}, a_{\mathcal{N}_{-i}^\beta}} w(s_{\mathcal{N}_{-i}^\beta}, a_{\mathcal{N}_{-i}^\beta}; s_{\mathcal{N}_i^\beta}, a_{\mathcal{N}_i^\beta}) Q_i(s_{\mathcal{N}_i^\beta}, s_{\mathcal{N}_{-i}^\beta}, a_{\mathcal{N}_i^\beta} a_{\mathcal{N}_{-i}^\beta}) \tag{9}$$

where weight parameters $w(s_{\mathcal{N}_{-i}^\beta}, a_{\mathcal{N}_{-i}^\beta}; s_{\mathcal{N}_i^\beta}, a_{\mathcal{N}_i^\beta}) \geq 0$ satisfy $\sum_{s_{\mathcal{N}_{-i}^\beta}, a_{\mathcal{N}_{-i}^\beta}} w(s_{\mathcal{N}_{-i}^\beta}, a_{\mathcal{N}_{-i}^\beta}; s_{\mathcal{N}_i^\beta}, a_{\mathcal{N}_i^\beta}) = 1$.

It can be shown that despite the reduction in dimension, the truncated local Q functions are a good approximation of the original local Q functions under the (ν, μ)-decay property.

Proposition 2 (Adapted from Lemma 4 in [29]). *Suppose the local Q functions $\{Q_i\}_{i\in\mathcal{N}}$ satisfy (ν, μ) decay property. Then, $\forall i$,*

$$\sup_{s,a} \left| Q_i(s, a) - \hat{Q}_i(s_{\mathcal{N}_i^\beta}, a_{\mathcal{N}_i^\beta}) \right| \leq \frac{\nu}{(\beta + 1)^\mu}.$$

This truncation of the local Q functions has been adopted in the literature [28, 29]. We also use this in our approach.

3 ALGORITHM DESIGN

We now present the design of Localized Policy Iteration, i.e., LPI. As discussed in the introduction, it is impractical to implement ζ_i because it needs access to the global state $s = (s_1, \ldots, s_n)$. To see this, note that it requires $\Omega(|\mathcal{S}|) = \Omega(\prod_{i=1}^n |\mathcal{S}_i|)$ parameters to even specify such a policy, which is impractically large and hard to store. Moreover, each agent would need access to the states of all the agents in order to implement such a policy, which is often hard to achieve in a large networked system due to communication challenges.

To address these issues, our focus is on localized policies, where agent i's action depends only on the states of the agents who are close to i in the network graph $\mathcal{G}$. Given this restriction on the policies, one may immediately ask what is lost when we restrict to localized policies compared with the centralized policies where the agents are allowed access to the global state. We show in Section 3.1 that, under a structural assumption, using κ-hop localized policies is almost as good as using centralized policies, where the parameter κ captures the level of localization. Then, in Section 3.2, we introduce our LPI framework that learns a near optimal κ-hop localized policy.

3.1 Performance Gap between Localized Policies and Centralized Policies

We begin by formally defining the class of κ-hop localized policies we consider. Recall that $\kappa > 0$ is a positive integer, and we use $s_{\mathcal{N}_i^\kappa}$ to denote the states of the agents in $\mathcal{N}_i^\kappa$.

Definition 5 (κ-hop policies). *A policy $\zeta = (\zeta_1, \ldots, \zeta_n) \in \Delta_{\text{policy}}$ is called a κ-hop (localized) policy if the following equation holds for all $i \in \mathcal{N}$, $s_{\mathcal{N}_i^\kappa} \in \mathcal{S}_{\mathcal{N}_i^\kappa}$, and $s_{\mathcal{N}_{-i}^\kappa}, s'_{\mathcal{N}_{-i}^\kappa} \in \mathcal{S}_{\mathcal{N}_{-i}^\kappa}$:*

$$\zeta_i(\cdot \mid s_{\mathcal{N}_i^\kappa}, s_{\mathcal{N}_{-i}^\kappa}) = \zeta_i(\cdot \mid s_{\mathcal{N}_i^\kappa}, s'_{\mathcal{N}_{-i}^\kappa}).$$

In other words, there exists $\hat{\zeta}_i : \mathcal{S}_{\mathcal{N}_i^\kappa} \mapsto \Delta_{\mathcal{A}_i}$ such that $\zeta_i(\cdot|s_{\mathcal{N}_i^\kappa}, s_{\mathcal{N}_{-i}^\kappa}) = \hat{\zeta}_i(\cdot|s_{\mathcal{N}_i^\kappa})$.

From Definition 5, we see that, when using a κ-hop localized policy, each agent only needs to know the states of the agents that are within its κ-hop neighborhood. Note that κ is a tunable parameter in practice, and captures the trade-off between communication, optimality, and computational complexity. Such a policy is more practical to implement than a centralized policy because of both the reduction in dimension and the reduction in communication.

To state our result on the performance gap, the following definition regarding the system transition matrix is needed. Following Qu et al. [28], we define a matrix $C \in \mathbb{R}^{n\times n}$ that characterizes the interaction strength between agents via the total variation of the transition probabilities.

$$C_{ij} = \begin{cases} 0, & \text{if } j \notin \mathcal{N}_i, \\ \sup_{s_{\mathcal{N}_i/j}, a_i} \sup_{s_j, s'_j} \text{TV}(P_i(\cdot|s_j, s_{\mathcal{N}_i/j}, a_i), P_i(\cdot|s'_j, s_{\mathcal{N}_i/j}, a_i)), & \text{if } j \in \mathcal{N}_i/i, \\ \sup_{s_{\mathcal{N}_i/i}} \sup_{s_i, s'_i, a_i, a'_i} \text{TV}(P_i(\cdot|s_i, s_{\mathcal{N}_i/i}, a_i), P_i(\cdot|s'_i, s_{\mathcal{N}_i/i}, a'_i)), & \text{if } j = i. \end{cases} \tag{10}$$

Our first main result states that κ-hop polices are nearly as good as centralized policies, with a gap that decays polynomially in κ.

Theorem 1. *Suppose that $\gamma < 0.8$, $\tau \geq 6\bar{r}\frac{4-3\gamma}{4-5\gamma}A_{\max}^2 e$, and $\sum_{j\in\mathcal{N}_i} C_{ij} < 1/2$ for all $i \in \{1, 2, \cdots, n\}$. Then there exists a κ-hop policy $\hat{\zeta}^{*,\kappa}$ that satisfies*

$$J(\zeta^*) - J(\hat{\zeta}^{*,\kappa}) \leq \frac{\bar{r}(4-3\gamma)}{(1-\gamma)(4-5\gamma)} \frac{1}{(\kappa+1)^\mu},$$

where

$$\mu = \min\left\{ \log_2 \frac{\tau(4-5\gamma)}{6\bar{r}A_{\max}^2 e(4-3\gamma)}, \log_2 \frac{1}{2\sup_{i\in\mathcal{N}} \sum_{j\in\mathcal{N}_i} C_{ij}} \right\}.$$

The above theorem states that when κ increases, the optimality gap of κ-hop localized policies decays polynomially in $1/\kappa$. This suggests that, even for κ-hop policies with a relatively small κ, one can achieve reasonably good performance and also avoid the curse of dimensionality and the communication issue in large networked systems. Theorem 1 motivates us to learn κ-hop localized policies, which is presented in the next section.

The proof of Theorem 1 is provided in Section 5. The key idea is to show that the optimal centralized policy turns out to be a (ν, μ)-decay policy defined in Definition 2. This means that in the optimal policy, each agent's action a_i's dependence on other states s_j shrinks polynomially in the distance between i and j. Therefore, by "truncating" the dependence on the agents that are more than κ-hop away, we will obtain a κ-hop policy that is $O(\frac{1}{(\kappa+1)^\mu})$ away from the optimal centralized policy, leading to a proof of Theorem 1.

We note that the polynomial decay rate μ in Theorem 1 depends on several factors. The factor $\sum_{j\in\mathcal{N}_i} C_{ij}$ can be viewed as the interaction strength between node i and its neighbors. The smaller the total interaction strength is for all agents, the larger the decay parameter μ. This is intuitive since weaker interaction means that κ-hop localized policies should perform better as there is less

coupling between the nodes. Another important factor is τ. The larger τ is, the larger the decay rate μ. Intuitively, a larger τ generally means entropy regularization will play a more important role, which effectively "dampens" the interaction among agents. To understand this, note that the policy that maximizes the entropy is the uniform policy, and therefore, the larger the entropy regularization is, the less incentive there is for an agent's policy to depend on the states of far away agents.

It is also interesting to compare our result to the results in the Linear Quadratic Control (LQC) setting [3, 27, 30, 42], where it has been shown that the performance gap between the optimal κ-hop policy and the optimal centralized policy decays (quasi-)exponentially in κ. This is a faster decay rate than the polynomial rate in our result. One reason for this difference is that the LQC setting is inherently a linear setting, whereas our MARL setting is inherently a combinatorial setting, which is typically more complicated. It remains an interesting question to understand the fundamental reason behind the discrepancy between the LQC setting and our setting.

Lastly, we note that, for Theorem 1 to hold, we need a lower bound on τ and an upper bound on $\sup_i \sum_{j\in\mathcal{N}_i} C_{ij}$. We believe these two bounds are hard to avoid. Bounds like these are common in the spatial decay literature in combinatorial settings. As an example, in [14, 15], the ratio I_2/I_1 is assumed to be small enough, where I_1 can be viewed as the bias of each agent towards a specific action, and I_2 can be viewed as the interaction strength between neighboring agents. Put in the context of our paper, I_1 corresponds to the τ parameter, as the entropy regularization can be viewed as a bias towards the uniform policy, and I_2 corresponds to the total interaction strength $\sup_i \sum_{j\in\mathcal{N}_i} C_{ij}$. As a result, our assumptions on τ and $\sup_i \sum_{j\in\mathcal{N}_i} C_{ij}$ are consistent with those in the literature. On a different note, we also have an upper bound on $\gamma < 0.8$, which we believe is an artifact of the proof. In Section 6, we show our approach also works when $\gamma > 0.8$.

3.2 Algorithm Design: Localized Policy Iteration

The details of LPI are presented in Algorithm 1. It is a policy iteration style algorithm where each agent updates a κ-hop policy $\hat{\zeta}_i^m$, with m being the iteration counter. Within each iteration, the algorithm is divided into two major steps: policy evaluation and policy improvement. Each is described in detail below, followed by a discussion of the communication and computation requirements of the algorithm.

Policy evaluation. Each agent implements the current policy (Line 4) to collect samples (Line 5). Then, in Line 7, each agent i estimates a *truncated* local Q function (defined in (9)). Such a truncated local Q-function is much smaller in dimension than the full local Q-function, and the error caused by the truncation is small (cf. Proposition 2). We note that Line 7 uses a subroutine `PolicyEvaluation`, which we leave unspecified for now because our algorithm can accommodate many popular policy evaluation schemes such as Temporal Difference (TD) learning. Further, our final convergence guarantee holds as long as the policy evaluation subroutine satisfies a exactness property (Definition 7) to be defined in Section 4. Moreover, in Section 4.2, we provide a version of TD learning as the `PolicyEvaluation` subroutine that can learn the truncated local Q functions and satisfy Definition 7.

Policy improvement. The policy improvement step runs from Line 8 to Line 10. It is an iterative procedure with $p_{\max}$ iterations, and each agent updates a κ-hop policy $\hat{\pi}_i^p$. The core step is for the agents to collectively implement (11) and (12), where each agent i sets $\hat{\pi}_i^{p+1}$ to be the softmax policy according to $\boldsymbol{Q}^i$. We explain this step in detail below.

Calculating $\boldsymbol{Q}^i$ in (11). $\boldsymbol{Q}^i$ is a *locally aggregated Q function*, where it averages over the truncated local Q functions of agent i's κ hop neighborhood: $\{\hat{Q}_j\}_{j\in\mathcal{N}_i^\kappa}$. We note that $\boldsymbol{Q}^i$ only depends on the states and actions of i's κ-hop neighborhood $s_{\mathcal{N}_i^\kappa}, a_{\mathcal{N}_i^\kappa}$, but in the meanwhile $\hat{Q}_j$ depends on

Algorithm 1 Localized Policy Iteration (LPI)

for $m = 0, 1, 2, \cdots$ **do**
  Sample initial global state $s(0) \sim \rho$.

  Policy Evaluation
  for $t = 0, 1, \cdots, T$ **do**
    Each agent i takes action $a_i(t) \sim \hat{\zeta}_i^m(\cdot \mid s_{\mathcal{N}_i^\kappa}(t))$ to obtain the next global state $s(t+1)$.
    Each agent i records $s_{\mathcal{N}_i^\beta}(t), a_{\mathcal{N}_i^\beta}(t), r_i(t) := r_i(s_i(t), a_i(t))$.
  end for
  Each agent i conducts policy evaluation subroutine to estimate its truncated local Q-function $\hat{Q}_i^m \leftarrow \texttt{PolicyEvaluation}(\beta, \{s_{\mathcal{N}_i^\beta}(t), a_{\mathcal{N}_i^\beta}(t), r_i(t)\}_{t=0}^T, \hat{\zeta}_i^m)$.

  Soft Policy Improvement
  for $p = 0, 1, \cdots, p_{\max}$ **do**
    Each agent i runs the following iterative procedure to calculate a policy, where $\hat{\pi}_i^0$ is initialized at a uniformly random policy.
    For all $(a_{\mathcal{N}_i^\kappa}, s_{\mathcal{N}_i^\kappa}) \in \mathcal{A}_{\mathcal{N}_i^\kappa} \times \mathcal{S}_{\mathcal{N}_i^\kappa}$, update

$$Q^i(a_i, a_{\mathcal{N}_i^\kappa/i}, s_{\mathcal{N}_i^\kappa}) = \frac{1}{n} \sum_{j \in \mathcal{N}_i^\kappa} \hat{Q}_j^m([\overline{s_{\mathcal{N}_i^\kappa}}]_{\mathcal{N}_j^\beta}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{\mathcal{N}_j^\beta \setminus i}, a_i), \tag{11}$$

$$\hat{\pi}_i^{p+1}(a_i \mid s_{\mathcal{N}_i^\kappa}) = \frac{\left(\hat{\pi}_i^p(a_i \mid s_{\mathcal{N}_i^\kappa})\right)^{1-\eta\tau}}{Z_i^p(s_{\mathcal{N}_i^\kappa})} \exp\left(\eta \mathbb{E}_{a_j \sim \hat{\pi}_j^p([\overline{s_{\mathcal{N}_i^\kappa}}]_{\mathcal{N}_j^\kappa}), j \in \mathcal{N}_i^\kappa/\{i\}}\left[Q^i(a_i, a_{\mathcal{N}_i^\kappa/i}, s_{\mathcal{N}_i^\kappa})\right]\right), \tag{12}$$

    where $Z_i^p(s_{\mathcal{N}_i^\kappa})$ is the normalization term.
  end for
  Each agent sets $\hat{\zeta}_i^{m+1} \leftarrow \hat{\pi}_i^{p_{\max}+1}$.
end for

the state and action of node j's β-hop neighborhood, which might not be in $\mathcal{N}_i^\kappa$. Therefore, when evaluating $\hat{Q}_j$, for the nodes $\ell \in \mathcal{N}_j^\beta / \mathcal{N}_i^\kappa$, we use a default value for their state and action denoted as $s_\ell^{\text{default}}, a_\ell^{\text{default}}$. More formally, we define the following extension operator:

Definition 6 (Extension Operator). *Define a state tuple $s_{\mathcal{J}} = (s_j)_{j \in \mathcal{J}} \in \mathcal{S}_{\mathcal{J}}$ for a subset of agents $\mathcal{J} \subseteq \mathcal{N}$. We define a corresponding global state $\overline{s_{\mathcal{J}}}$ based on $s_{\mathcal{J}}$:*

$$[\overline{s_{\mathcal{J}}}]_j = \begin{cases} s_j & \text{if } j \in \mathcal{J}, \\ s_j^{\text{default}} & \text{if} j \notin \mathcal{J}. \end{cases} \tag{13}$$

where s^{default} is a default global state. The same notation is used to extend local actions to global actions.

It is worth noting that throughout this paper, the default state s^{default} is fixed and can be any state tuple in $\mathcal{S}$. More importantly, the value of s^{default} does not affect our final convergence result. This is because when i and ℓ are far away, the choice of s_ℓ, a_ℓ will almost play no role in the policy improvement for agent i. More precisely, we show later in the proof (Lemma 22) that the difference caused by changing a_i, $Q(s, a_i, a_{-i}) - Q(s, a_i', a_{-i})$, is insensitive to s_ℓ, a_ℓ. This allows us to use an arbitrary default value $s_\ell^{\text{default}}, a_\ell^{\text{default}}$ without changing the result. With Definition 6, we can write the evaluation of $\hat{Q}_j$ as $\hat{Q}_j([\overline{s_{\mathcal{N}_i^\kappa}}]_{\mathcal{N}_j^\beta}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{\mathcal{N}_j^\beta})$. This gives rise to (11). The rationale of calculating

Q^i in this way is that, as shown later in the proof, we have for any s, a_{-i}, a_i, a_i'

$$Q^i(a_i, a_{\mathcal{N}_i^\kappa/i}, s_{\mathcal{N}_i^\kappa}) - Q^i(a_i', a_{\mathcal{N}_i^\kappa/i}, s_{\mathcal{N}_i^\kappa}) \approx \hat{Q}^m(a_i, a_{-i}, s) - \hat{Q}^m(a_i', a_{-i}, s)$$

where $\hat{Q}^m(s,a) = \frac{1}{n}\sum_{j\in\mathcal{N}} \hat{Q}_j^m(s_{\mathcal{N}_j^\beta}, a_{\mathcal{N}_j^\beta})$, the average of the truncated local Q functions for iteration m. In other words, Q^i is a good estimate of $\hat{Q}^m$ (up to a function that does not depend on a_i), which in turn is a good estimate of the true Q function $Q^{\hat{\zeta}^m}$. Therefore, we can then conduct a multiplicative weights policy update based on Q^i, which we discuss next.

Multiplicative weights policy update (12). The basic idea of (12) is for each agent i to update its policy using the multiplicative weights algorithm, with Q^i as the score for each action a_i. Note that Q^i not only depends on a_i, but also $a_{\mathcal{N}_i^\kappa/i}$, so when conducting the update, we take the expectation of a_j using the current policy of agent j, $\hat{\pi}_j^p$. When doing so, $\hat{\pi}_j^p$ depends on states that are outside $\mathcal{N}_i^\kappa$, and for these states, we set it to a default value (the same as the default value used in the calculation of Q^i). We show later in the proof (Lemma 8) that the update (12) is approximately solving the maximization in the Bellman optimal operator (7).

Computation and communication. The major computational burden of LPI lies in the steps that compute the truncated local Q functions and the κ-hop polices. The complexity of these scales with the largest state-action space size of the κ and β-hop neighborhood in the network. Therefore, our algorithm can avoid the exponential computation burden associated with the exponentially large state and action spaces.

In terms of communication, each agent only needs local communication with agents in the $\max(\beta,\kappa)$-hop neighborhood, as opposed to centralized communication. Specifically, in Line 5 each agent i needs to receive information on the states and actions of agents in the β-hop neighborhood $s_{\mathcal{N}_i^\beta}(t), a_{\mathcal{N}_i^\beta}(t)$. In Line 9 (eq. (12)), each agent i needs to know $\{\hat{\pi}_j^p\}_{j\in\mathcal{N}_i^\kappa}$ as well as the truncated Q-functions $\{\hat{Q}_j^m\}_{j\in\mathcal{N}_i^\kappa}$ for the agents in the κ-hop neighborhood.

4 CONVERGENCE ANALYSIS

This section provides a convergence guarantee for LPI. We first present our requirements for the `PolicyEvaluation` subroutine in Definition 7. Then, under these requirements, we present our main convergence result for LPI in Theorem 2. Finally, in Section 4.2, we provide a specific policy evaluation subroutine, Localized TD(0), which is a variation of the approach proposed in [21], and show that it can meet the Definition 7.

4.1 Convergence of Localized Policy Iteration

Our requirement for the `PolicyEvaluation` subroutine is stated formally in Definition 7. Intuitively, the policy evaluation process involves two parts of errors: a bias is introduced because we only estimate the truncated local Q functions (cf. (9)) using information collected within the β-hop neighborhood rather than global network. Such a truncation introduces a bias that is on the order of $O(\nu'/(\beta+1)^\mu)$ because of the (ν',μ)-decay property of the true local Q function (Proposition 2). In addition to the bias caused by the truncation, a stochastic error also exists due to the inherent stochasticity of the observed samples. Under proper assumptions, the stochastic error decays to zero as the number of samples T increases. Since the bias part of the error is 'inevitable', we require the policy evaluation algorithm to achieve an approximation error that is in the same order of the bias with high probability after a finite number of iterations, formally stated in Definition 7.

Definition 7. *A policy evaluation algorithm is said to be (σ,ν',μ)-exact if the following condition holds: For any σ regular policy $\hat{\zeta}$ whose corresponding local Q functions $\{Q_i^{\hat{\zeta}}\}$ is (ν',μ) decay, and $\delta\in(0,1)$, there exists a function $c_{pe}(\delta,\sigma,\nu',\mu)$, such that for any $\beta\in\mathbb{N}$, there exists a $T_{eval}(\delta,\sigma,\nu',\mu,\beta)\in$*

$\mathbb{N}$ *such that when the length of the trajectory supplied to the subroutine* $T \geq T_{eval}(\delta, \sigma, \nu', \mu, \beta)$ *steps, with probability at least* $1-\delta$*, the following inequality holds*

$$\sup_{(s,a)\in\mathcal{S}\times\mathcal{A}} \left| Q_i^{\hat{\zeta}}(s,a) - \hat{Q}_i(s_{\mathcal{N}_i^\beta}, a_{\mathcal{N}_i^\beta}) \right| \leq \frac{c_{pe}(\delta,\sigma,\nu',\mu)}{(\beta+1)^\mu}, \forall i \in \mathcal{N},$$

where $Q_i^{\hat{\zeta}}$ *is the true local Q-functions under policy* $\hat{\zeta}$*, and* $\hat{Q}_i$ *is the output of the policy evaluation subroutine. When the context is clear, we may drop the parenthesis of* c_{pe}, T_{eval}*.*

We will provide a concrete algorithm that satisfies Definition 7 in Section 4.2. For now, assuming `PolicyEvaluation` satisfies Definition 7, we can now state our main convergence result for Algorithm 1.

Theorem 2. *Suppose* $\gamma < 0.8$, $\tau \geq 40\bar{r}\frac{4-3\gamma}{4-5\gamma}A_{\max}^2 e$, *and* $\sum_{j\in\mathcal{N}_i} C_{ij} < 1/2$ *for all* $i \in \{1, 2, \cdots, n\}$*. Let* $\nu' = \frac{4-3\gamma}{4-5\gamma}\bar{r}, \mu = \min\left\{\log_2 \frac{\tau(4-5\gamma)}{40\bar{r}A_{\max}^2 e(4-3\gamma)}, \log_2 \frac{1}{2\sup_{i\in\mathcal{N}}\sum_{j\in\mathcal{N}_i} C_{ij}}\right\}$*, and* $\tilde{\sigma} = \frac{2\nu'}{\tau n}$*. Suppose the* `PolicyEvaluation` *subroutine in Algorithm 1 is* $(\tilde{\sigma}, \nu', \mu)$*-exact with constants* $c_{pe}(\cdot), T_{eval}(\cdot)$ *(cf. Definition 7). Fixing any* κ*, take the algorithm parameters as* $\beta = \frac{\kappa+1}{2}(\frac{2f(\kappa)c_{pe}}{\nu'})^{\frac{1}{\mu}}$, $\eta = \frac{1}{\tau}$, *and* $p_{\max} \geq -\log_2 \frac{4+c_{pe}/(2^\mu\nu')}{3(\kappa/2+1)^\mu}$ *(where we recall* $f(\kappa)$ *is the size of the largest* κ*-hop neighborhood). Then, given* $M \in \mathbb{N}$ *and* $\delta \in (0,1)$*, when the trajectory input to the* `PolicyEvaluation` *subroutine* $T \geq T_{eval}(\delta/M, \tilde{\sigma}, \nu', \mu, \beta)$*, we have with probability at least* $1-\delta$*,*

$$J(\zeta^*) - J(\hat{\zeta}^M) \leq \gamma^M \|V^{\hat{\zeta}^0} - V^*\|_\infty + \frac{3(4-3\gamma)\bar{r}}{(1-\gamma)^2(4-5\gamma)}\left(4 + \frac{c_{pe}(4-5\gamma)}{2^\mu(4-3\gamma)\bar{r}}\right)\frac{1}{(\kappa/2+1)^\mu}.$$

As shown in Theorem 2, LPI converges geometrically in M, with a steady state error on the order of $O(\frac{1}{(\kappa+1)^\mu})$, i.e., the steady state error decays polynomially in κ. This steady state error is a result of both the policy evaluation error in Definition 7, and the fact that we are using a truncated κ-hop policy as opposed to a global policy.

Sample complexity. As we discussed in Section 3.2, the `PolicyEvaluation` subroutine can be chosen to be any algorithm that satisfies Definition 7 and the sample complexity of LPI depends on the specific choice. In Section 4.2, we describe a specific `PolicyEvaluation` subroutine, β-hop Localized TD(0), that satisfies the requirement. Under this specific subroutine, we derive the sample complexity of LPI in the following corollary.

Corollary 3. *Let the assumptions of Theorem 2 and Theorem 5 hold. For any* $\varepsilon > 0$ *and* $\delta \in (0,1)$*, when using* β*-hop Localized TD(0) as the* `PolicyEvaluation` *subroutine, the sample complexity of Algorithm 1 to achieve* ε*-optimality is in the order of*

$$\Theta\left(\frac{1}{\varepsilon^2}\frac{f(\kappa)^2\log(1/\varepsilon)\log[f(\beta)\log(1/\varepsilon)]}{\xi(\tilde{\sigma},\beta)^2}\right), \tag{14}$$

where $\xi(\tilde{\sigma},\beta)$ *is defined in Lemma 4 and means the smallest probability the state-action pairs in* β*-hop neighborhoods are visited.*

Based on the convergence result in Theorem 2, to find an ε-optimal policy, we only need to set $M = \Theta(\log(1/\varepsilon))$ and $\kappa = \Theta((1/\varepsilon)^{1/\mu})$, and correspondingly $\beta = \Theta(\kappa(f(\kappa))^{1/\mu})$. According to (15) that we show later in Theorem 5, the iteration complexity of β-hop Localized TD(0) is $T_{eval} = \Theta(\frac{(\beta+1)^{2\mu}\log(f(\beta)M/\delta)}{\xi(\tilde{\sigma},\beta)^2}) = \Theta(\frac{(1/\varepsilon)^2 f(\kappa)^2\log[f(\beta)\log(1/\varepsilon)]}{\xi(\tilde{\sigma},\beta)^2})$. Thus the sample complexity of LPI follows the simple calculation MT_{eval} and is given in (14). In all the $\Theta(\cdot)$ notations above, we only keep the dependence on the network size n and the ε parameter, which involves ε itself and by extension, the κ, β, M parameters.

In this sample complexity bound, we have the $\frac{1}{\epsilon^2}$ factor that is standard in the RL literature. In addition, instead of depending on the global-state action space size (exponential in n), we have a square dependence on $\frac{1}{\xi(\tilde{\sigma},\beta)}$, where $\xi(\tilde{\sigma},\beta)$ is defined in Lemma 4 and means the smallest probability the state-action pairs in β-hop neighborhoods are visited. As such, $\frac{1}{\xi(\tilde{\sigma},\beta)}$ scales with the largest state action space size of the β-hop neighborhoods, as opposed to the entire network. Therefore, LPI is much more scalable and implementable than other methods that enjoy global convergence such as centralized tabular RL methods [4]. This advantage is especially pronounced in sparse networks where β-hop neighborhoods are much smaller than the entire network. Lastly, beyond the $\frac{1}{\epsilon^2}$ factor and $\frac{1}{(\xi(\tilde{\sigma},\beta))^2}$ factor and a few log factors, our bound also contains a factor $f(\kappa)^2$, where we recall $f(\kappa)$ is the size of the largest κ-hop neighborhood. This factor reflects the complexity of the κ-hop policy we try to learn.

We end this section with a comparison of our results with that in [21, 28–30, 42]. The foremost difference of our work from [21, 28, 29] is our guarantee on the global convergence, i.e. (approximately) converging to the global optimal policy, while [21, 28, 29] only offer local convergence, i.e. converging to a local minimum of the policy optimization problem. Shin et al. [30], Zhang et al. [42] show a (quasi-)exponential decay property of the optimal controller in the linear quadrati control problem which is similar to our Theorem 1. Apart from studying a different problem from ours, [30, 42] do not provide a learning algorithm that converges to the global optimal policy. In contrast, this work proposes a learning algorithm with provable global convergence guarantee.

4.2 Policy Evaluation via β-hop Localized TD(0)

Algorithm 2 β-hop Localized TD(0) (Agent i)

Require: Parameter β, a sequence of β-hop state, β-hop action, and local reward tuples $\{s_{\mathcal{N}_i^\beta}(t), a_{\mathcal{N}_i^\beta}(t), r_i(t)\}$, local policy $\hat{\zeta}_i^m$, and a sequence of learning rate $\{\alpha_t\}$.

1: Initialize $\hat{Q}_i^0$ to be all zero vector.

2: Set $\hat{r}_i(t) \leftarrow r_i(t) - n\tau\mathbb{E}_{a_i\sim\hat{\zeta}_i^m(\cdot|s_{\mathcal{N}_i^\kappa}(t))} \log \hat{\zeta}_i^m(a_i \mid s_{\mathcal{N}_i^\kappa}(t))$.

3: **for** $t = 1, \cdots, T$ **do**

4: Update the local estimation $\hat{Q}_i$ with step size α_{t-1},

$$\hat{Q}_i^t\big(s_{\mathcal{N}_i^\beta}(t-1), a_{\mathcal{N}_i^\beta}(t-1)\big) =$$
$$(1-\alpha_{t-1})\hat{Q}_i^{t-1}\big(s_{\mathcal{N}_i^\beta}(t-1), a_{\mathcal{N}_i^\beta}(t-1)\big) + \alpha_{t-1}\big(\hat{r}_i(t) + \gamma\hat{Q}_i^{t-1}\big(s_{\mathcal{N}_i^\beta}(t), a_{\mathcal{N}_i^\beta}(t)\big)\big),$$
$$\hat{Q}_i^t\big(s_{\mathcal{N}_i^\beta}, a_{\mathcal{N}_i^\beta}\big) = \hat{Q}_i^{t-1}\big(s_{\mathcal{N}_i^\beta}, a_{\mathcal{N}_i^\beta}\big) \text{ for } \big(s_{\mathcal{N}_i^\beta}, a_{\mathcal{N}_i^\beta}\big) \neq \big(s_{\mathcal{N}_i^\beta}(t-1), a_{\mathcal{N}_i^\beta}(t-1)\big).$$

5: **end for**

6: For every $(s_{\mathcal{N}_i^\beta}, a_{\mathcal{N}_i^\beta})$ pair, set $\hat{Q}_i^T(s_{\mathcal{N}_i^\beta}, a_{\mathcal{N}_i^\beta}) \leftarrow \hat{Q}_i^T(s_{\mathcal{N}_i^\beta}, a_{\mathcal{N}_i^\beta}) + n\tau\mathbb{E}_{a_i\sim\hat{\zeta}_i^m(\cdot|s_{\mathcal{N}_i^\kappa})} \log \hat{\zeta}_i^m(a_i \mid s_{\mathcal{N}_i^\kappa})$ and return $\hat{Q}_i^T(\cdot)$

To provide a concrete example of the `PolicyEvaluation` subroutine, we introduce and analyze Localized TD(0) in this subsection. The pseudo-code of Localized TD(0) is given in Algorithm 2. It is a variation of a policy evaluation subroutine used in [21], where each agent conducts temporal difference learning using the states and actions of its local β-hop neighborhood. Note that compared to [21], we also add entropy regularization in the update (see Line 4 of Algorithm 2).

In this section, we show that Algorithm 2 satisfies Definition 7 by utilizing the convergence results in [21]. We first state the assumption needed.

Assumption 1. *There exists a policy $\zeta_b = (\zeta_{b,1}, \zeta_{b,2}, \cdots, \zeta_{b,n}) \in \Delta_{policy}$ such that the Markov chain $\{s(t)\}_{t\geq 0}$ induced by ζ_b is irreducible and aperiodic.*

Based on Assumption 1, we show that all σ-regular policies are sufficiently explorative.

Lemma 4. *Define $\mathcal{Z} := \mathcal{S} \times \mathcal{A}$. Under Assumption 1, for any σ-regular policy $\hat{\zeta}$, the induced Markov chain $\{z(t) := (s(t), a(t))\}$ is irreducible and aperiodic, hence admits a unique stationary distribution $d^{\hat{\zeta}} \in \Delta_{\mathcal{Z}}$ with strictly positive components. Further, there exists positive constants $K_1(\sigma)$ and $K_2(\sigma) \geq 1$ depending on σ such that*

$$\forall z' \in \mathcal{Z}, \forall t \geq 0, \sup_{\mathcal{K}\subseteq\mathcal{Z}} \left| \sum_{z\in\mathcal{K}} d_z^{\hat{\zeta}} - \sum_{z\in\mathcal{K}} \mathbb{P}(z(t) = z \mid z(0) = z') \right| \leq K_1 e^{-t/K_2},$$

Further, given any $\beta \in \mathbb{N}$, for each agent i and $z' \in \mathcal{Z}_{\mathcal{N}_i^\beta}$, define $d_{i,\beta}^{\hat{\zeta}}(z') = \sum_{z\in\mathcal{Z}: z_{\mathcal{N}_i^\beta} = z'} d^{\hat{\zeta}}(z)$, which is the marginal stationary distribution for the state-actions in $\mathcal{N}_i^\beta$, and define $\xi(\sigma, \beta) := \inf_{i\in\mathcal{N}, z'\in\mathcal{Z}_{\mathcal{N}_i^\beta}} d^{\hat{\zeta}}(z')$ as the minimum probability in this distribution.

Lemma 4 is an immediate implication of a result on the sufficient exploration of non-deterministic policies stated in Appendix F. Note that the only assumption we need here is that there exists a policy ζ_b such that the induced Markov chain $\{S_k\}$ is irreducible and aperiodic (Assumption 1). This is in contrast with the analysis in existing literature, where Lemma 4 is directly assumed to hold [29]. Therefore, our proof of Lemma 4 is of independent interest in the literature of policy evaluation.

Lemma 4 enables us to apply Theorem 3.2 in [21] to show that β-hop Localized TD(0) is a (σ, ν', μ)-exact policy evaluation algorithm.

Theorem 5. *Suppose the policy $\hat{\zeta}$ we want to evaluate is σ regular and under the policy, the local Q-functions satisfy the (ν', μ)-decay property. Suppose $K_1(\sigma), K_2(\sigma), \xi(\sigma, \beta)$ are the constants in Lemma 4. Recall that the local reward at every time step is upper bounded by $\bar{r}$. Let the step size of Algorithm 2 be $\alpha_t = \frac{H}{t+t_0}$ with $t_0 = \max(4H, 2K_2 \log T)$, and $H = \frac{2}{(1-\gamma)\xi(\sigma,\beta)}$. Then, Algorithm 2 is (σ, ν', μ)-exact with constant $c_{pe}(\delta, \sigma, \nu', \mu) = \frac{2\nu'}{1-\gamma}$ and $T_{eval}(\delta, \sigma, \nu', \mu, \beta)$ is upper bounded by*

$$\tilde{O}\left(\frac{(\beta+1)^{2\mu}(\bar{r} + \tau \log A_{\max})^2 K_2(\sigma) \log(f(\beta)K_2(\sigma)/\delta)}{(\nu')^2(1-\gamma)^4(\xi(\sigma,\beta))^2} + \frac{(\beta+1)^{\mu}(\bar{r} + \tau \log A_{\max})(K_1(\sigma)+1)(K_2(\sigma)+1)}{\nu'(1-\gamma)^2(\xi(\sigma,\beta))^2} \right), \tag{15}$$

where we recall $f(\beta) = \sup_{i\in\mathcal{N}} \left|\mathcal{N}_i^\beta\right|$ is the size of the largest β-hop neighborhood.

5 CONVERGENCE PROOF FOR LOCALIZED POLICY ITERATION

We now prove our main results for the convergence of LPI. The key technical idea underlying our analysis is a novel closure property for a class of policies with spatially decaying properties. We introduce this property first, and then apply it to prove the results in the previous sections.

5.1 Key Idea: Closure of Decay Policy Class under Policy Iteration

We first present a prototype algorithm in Algorithm 3, which is an exact policy iteration algorithm that runs exact policy evaluation followed by exact policy improvement to update the global policy. While this algorithm is not practical, it is useful in developing a generic framework to analyze the convergence of LPI (Algorithm 1). For that purpose, in this subsection we prove an important "closure" property of the prototype algorithm (Algorithm 3): when starting at an initial policy in

a decay policy class $\Delta_{\nu,\mu,\sigma}$ (defined in Definition 2), the policy at every iteration of the prototype algorithm remains in the same $\Delta_{\nu,\mu,\sigma}$ class. This property provides the foundation of our proof for both Theorem 1 and Theorem 2.

Algorithm 3 Exact Policy Iteration (Prototype)

1: **Input:** initial global policy ζ_0.
2: **for** $m = 0, 1, 2, \cdots, M-1$ **do**
3: Calculate the Q-function for policy ζ^m as Q^{ζ^m}
4: Updates the global policy $\zeta^{m+1} = (\zeta_1^{m+1}, \ldots, \zeta_n^{m+1})$, where

$$\{\zeta_i^{m+1}(\cdot|s)\}_{i\in\mathcal{N}} = \underset{\pi_1(\cdot|s)\in\Delta_{\mathcal{A}_1},\ldots,\pi_n(\cdot|s)\in\Delta_{\mathcal{A}_n}}{\arg\max} \mathbb{E}_{a_1\sim\pi_1(\cdot|s),\ldots,a_n\sim\pi_n(\cdot|s)}\left[Q^{\zeta^m}(s,a) - \sum_{i=1}^{n}\tau\log\pi_i(a_i|s)\right].$$

5: **end for**
6: **Output:** ζ^M

Closure of $\Delta_{\nu,\mu,\sigma}$ under Algorithm 3. In Theorem 6, we formally establish the property that when starting with $\zeta^0 \in \Delta_{\nu,\mu,\sigma}$, the iterates in Algorithm 3 remain in the policy class $\Delta_{\nu,\mu,\sigma}$. The proof of Theorem 6 can be found in Appendix B.

Theorem 6. *Suppose $\gamma < 0.8$ and $\tau \geq 6\bar{r}\frac{4-3\gamma}{4-5\gamma}A_{\max}^2 e$, and $\forall i, \sum_{j\in\mathcal{N}_i} C_{ij} < 1/2$. Define*

$$\nu = 1/2, \quad \mu = \min\left\{\log_2\frac{\tau(4-5\gamma)}{6\bar{r}A_{\max}^2 e(4-3\gamma)}, \log_2\frac{1}{2\sup_{i\in\mathcal{N}}\sum_{j\in\mathcal{N}_i}C_{ij}}\right\}, \quad \sigma = \frac{\bar{r}(4-3\gamma)}{(4-5\gamma)n\tau}.$$

Then, the following holds.

(a) When initial policy satisfies $\zeta^0 \in \Delta_{\nu,\mu,\sigma}$, all iterates ζ^m in Algorithm 3 will be in $\Delta_{\nu,\mu,\sigma}$.
(b) Q^{ζ^m} is (ν',μ)-decay with $\nu' = \bar{r}\frac{4-3\gamma}{4-5\gamma}$.
(c) V^{ζ^m} converges to the unique fixed point V^* of $\mathcal{T}$ with geometric rate, i.e. $\|V^{\zeta^m} - V^*\|_\infty \leq \gamma^m\|V^{\zeta^0} - V^*\|_\infty$.

The results in Theorem 6 demonstrate that the defined policy class $\Delta_{\nu,\mu,\sigma}$ is closed under the policy iteration operator in Algorithm 3. This key observation provides a pathway for proving our main theoretical results, as we explain below.

Pathway to prove Theorem 1. Based on the closure of the policy class $\Delta_{\nu,\mu,\sigma}$, the optimal policy is also in $\Delta_{\nu,\mu,\sigma}$. We show that if we "truncate" the optimal policy to a κ-hop policy, the performance loss is on the order of $O(\frac{1}{(\kappa+1)^\mu})$. This directly leads to a proof of Theorem 1, which we detail in Section 5.2.

Pathway to prove Theorem 2. A second consequence of Theorem 6 is that we can view Algorithm 1 as an inexact version of the prototype algorithm (Algorithm 3), which reduces the complexity in implementation, communication, and training. Specifically, Algorithm 1 introduces the following types of errors:

- *Error caused by truncated Q-functions*: In Algorithm 1, we only learn β-hop truncated versions of the Q-functions. By Theorem 6, the true Q-functions are (ν',μ)-decay, and as a result, the truncation causes an error on the order of $O(\frac{1}{(\beta+1)^\mu}) = O(\frac{1}{(\kappa+1)^\mu})$ (Noticing $\kappa \leq \beta$).
- *Error caused by truncated policies*: We only learn κ-hop truncated policies, which also causes an $O(\frac{1}{(\kappa+1)^\mu})$ error.

- *Error caused by inexact policy evaluation*: Since we use finite samples to estimate the truncated Q-functions, this causes statistical errors depending on the sample size. However, with high probability, this error diminishes to 0 as the number of samples T increases.
- *Error caused by inexact policy improvement*: In Algorithm 1, we cannot directly implement the arg max procedure as in Algorithm 3. Instead, we use an iterative procedure (12), which we later show corresponds to the mirror descent algorithm for solving the arg max. This causes an optimization error, which diminishes to 0 as the number of iterations $p_{\max}$ for (12) increases.

As discussed above, the first two types of errors can be bounded by the order of $O(\frac{1}{(\kappa+1)^\mu})$ which depends on the size of the neighborhood in our localized approximation of the policy and the Q functions, while the last two types of errors can be made arbitrarily small as long as the algorithm is run for sufficiently many steps. This eventually leads to a proof of Theorem 2, which we detail in Section 5.3.

5.2 Proof of Theorem 1: Near-optimality of κ-hop Policies

As discussed in Section 5.1, the optimal policy ζ^* is a σ-regular (ν, μ)-decay policy with the values of ν, μ, σ given by the condition in Theorem 6. To prove Theorem 1, we simply truncate the optimal policy to the κ-hop neighborhood. Specifically, the "truncated" policy is defined as

$$\hat{\zeta}_i^*(\cdot|s_{\mathcal{N}_i^\kappa}) = \zeta_i^*(\cdot|s_{\mathcal{N}_i^\kappa}, s^{\text{default}}_{\mathcal{N}_{-i}^\kappa}), \forall i. \tag{16}$$

For any $i \in \mathcal{N}$, $s \in \mathcal{S}$, for the simplicity of notations, we re-order the set $\mathcal{N}$ as follows

$$\{l_1, \ldots, l_{|\mathcal{N}_{-i}^\kappa|}, l_{|\mathcal{N}_{-i}^\kappa|+1}, \ldots, l_n\},$$

where $\{l_1, \ldots, l_{|\mathcal{N}_{-i}^\kappa|}\} = \mathcal{N}_{-i}^\kappa$ and $\{l_{|\mathcal{N}_{-i}^\kappa|+1}, \ldots, l_n\} = \mathcal{N}_i^\kappa$. We also use $\mathcal{N}_{[l_j]}$ to denote $\{l_1, \ldots, l_j\}$ and $\mathcal{N}/\mathcal{N}_{[l_j]}$ to denote $\{l_{j+1}, \ldots, l_n\}$. Then, using the triangle inequality we have

$$\text{TV}(\hat{\zeta}_i^*(\cdot|s_{\mathcal{N}_i^\kappa}), \zeta_i^*(\cdot|s)) \le \sum_{j=|\mathcal{N}_{-i}^\kappa|}^{1} \text{TV}(\zeta_i^*(\cdot|s^{\text{default}}_{\mathcal{N}_{[l_j]}}, s_{\mathcal{N}/\mathcal{N}_{[l_j]}}), \zeta_i^*(\cdot|s^{\text{default}}_{\mathcal{N}_{[l_{j-1}]}}, s_{\mathcal{N}/\mathcal{N}_{[l_{j-1}]}})), \tag{17}$$

where we define $\mathcal{N}_{l_0} = \Phi$ to be the empty set, which implies for $j = 1$, $\mathcal{N}/\mathcal{N}_{[l_{j-1}]} = \mathcal{N}$ and $s_{\mathcal{N}/\mathcal{N}_{[l_{j-1}]}} = s$ is the global state. By the definition of the interaction matrix in Definition 2, we immediately have

$$\begin{aligned}
\sup_{i\in\mathcal{N}} \sup_{s\in\mathcal{S}} \text{TV}(\hat{\zeta}_i^*(\cdot|s_{\mathcal{N}_i^\kappa}), \zeta_i^*(\cdot|s)) &\le \sup_{i\in\mathcal{N}} \sum_{j\in\mathcal{N}_{-i}^\kappa} Z_{ij}^{\zeta^*} \\
&\le \frac{1}{(\kappa+1)^\mu} \sup_{i\in\mathcal{N}} \sum_{j\in\mathcal{N}_{-i}^\kappa} (\text{dist}(i,j)+1)^\mu Z_{ij}^{\zeta^*} \\
&\le \frac{\nu}{(\kappa+1)^\mu},
\end{aligned} \tag{18}$$

where the second inequality is due to (8) and the last inequality holds because ζ^* is in the (ν, μ)-decay class. This further indicates $\hat{\zeta}_i^*$ is $O(\frac{1}{(\kappa+1)^\mu})$ close to ζ_i^*.

To proceed, we introduce the following performance difference bound that bounds the value functions of two policies by the TV distance between the two policies:

Lemma 7. *Suppose $\zeta, \tilde{\zeta} \in \Delta_{\text{policy}}$ are σ-regular, and further, $Q^{\tilde{\zeta}}$ is (ν', μ)-decay. Then, we have*

$$\|V^\zeta - V^{\tilde{\zeta}}\|_\infty \le \frac{1}{1-\gamma}\left(\tau\sigma + \frac{\nu'}{n}\right) \sum_{i=1}^{n} \sup_{s\in\mathcal{S}} \text{TV}(\tilde{\zeta}_i(\cdot|s), \zeta_i(\cdot|s)). \tag{19}$$

With the above Lemma 7, we have

$$J(\zeta^*) - J(\hat{\zeta}^*) = \mathbb{E}_{s\sim\rho}\left(V^{\zeta^*}(s) - V^{\hat{\zeta}^*}(s)\right) \le \|V^{\zeta^*} - V^{\hat{\zeta}^*}\|_\infty$$
$$\le \frac{1}{1-\gamma}\left(\tau\sigma + \frac{\nu'}{n}\right)\sum_{i=1}^{n}\sup_{s\in\mathcal{S}}\mathrm{TV}(\zeta_i^*(\cdot|s), \hat{\zeta}_i^*(\cdot|s)) \le \frac{1}{1-\gamma}(n\tau\sigma + \nu')\frac{\nu}{(\kappa+1)^\mu}$$
$$\le \frac{\bar{r}(4-3\gamma)}{(1-\gamma)(4-5\gamma)}\frac{1}{(\kappa+1)^\mu},$$

where in the last inequality, we have plugged in $n\tau\sigma = \nu' = \bar{r}\frac{4-3\gamma}{4-5\gamma}$ and $\nu = 1/2$. This demonstrates that the κ-hop policy $\hat{\zeta}^*$ is $O(\frac{1}{\kappa^\mu})$-optimal, and setting $\hat{\zeta}^{*,\kappa}$ as $\hat{\zeta}^*$ concludes the proof of Theorem 1.

5.3 Proof of Theorem 2: Convergence Analysis of Localized Policy Improvement

In this section, we prove the convergence of LPI. Our proof uses an induction argument, and the induction assumption is that, at the m-th iteration of Algorithm 1, the policy $\hat{\zeta}^m$ is $\tilde{\sigma}$ regular and its local Q-functions $\{Q_i^{\hat{\zeta}^m}\}_{i=1}^n$ is (ν', μ)-decay, with constants $\tilde{\sigma}, \nu', \mu$ given in Theorem 2's statement. This induction assumption is clearly true for $m = 0$ because of Lemma 9 and the fact that $\hat{\zeta}^0$ is a uniform policy.

Now under this induction assumption, we first show that the local Q functions $\hat{Q}_i^m$ returned by the policy evaluation step (i.e., the `PolicyEvaluation` subroutine in Line 7) of Algorithm 1 is a good approximation of the true local Q-functions $Q_i^{\hat{\zeta}^m}$ of the policy $\hat{\zeta}^m$.

More specifically, recall that we assumed the policy evaluation step in Algorithm 1 is $(\tilde{\sigma}, \nu', \mu)$-exact as defined in Definition 7. This means for any given $\delta \in (0,1)$ and $\beta \in \mathbb{N}$, there exists constants $T_{eval}(\frac{\delta}{M}, \tilde{\sigma}, \nu', \mu, \beta), c_{pe}(\frac{\delta}{M}, \tilde{\sigma}, \nu', \mu)$ such that when the input trajectory to `PolicyEvaluation` has length $T \ge T_{eval}(\frac{\delta}{M}, \tilde{\sigma}, \nu', \mu, \beta)$ steps, its output $\hat{Q}_i^m$ satisfies

$$\forall i, \sup_{(s,a)\in\mathcal{S}\times\mathcal{A}}\left|Q_i^{\hat{\zeta}^m}(s,a) - \hat{Q}_i^m(s_{\mathcal{N}_i^\beta}, a_{\mathcal{N}_i^\beta})\right| \le \frac{c_{pe}(\frac{\delta}{M}, \tilde{\sigma}, \nu', \mu)}{(\beta+1)^\mu} := \epsilon \tag{20}$$

with probability at least $1 - \frac{\delta}{M}$. In the rest of the proof, we write $c_{pe}(\frac{\delta}{M}, \tilde{\sigma}, \nu', \mu)$ as c_{pe} for simplicity.

Given (20), we next show that the policy improvement step (Line 11) of Algorithm 1 returns a $\tilde{\sigma}$ regular policy $\hat{\zeta}^{m+1}$ that is $O(\frac{1}{\kappa^\mu})$ close to the policy given by the optimal Bellman operator, and further $\{Q_i^{\hat{\zeta}^{m+1}}\}$ is (ν', μ) decay. More precisely, we present the following Lemma 8, the full proof of which can be found in Appendix E.

Lemma 8. *Consider the settings of Theorem 2. Suppose at iteration m, the local Q-functions $\{Q_i^{\hat{\zeta}^m}\}_{i=1}^n$ is (ν', μ)-decay and (20) is true. Then, the policy improvement part (Line 11) will return a policy $\hat{\zeta}^{m+1}$ that satisfies*

$$\mathcal{T}V^{\hat{\zeta}^m} - V^{\hat{\zeta}^{m+1}} \le \frac{3(4-3\gamma)\bar{r}}{(1-\gamma)(4-5\gamma)}\frac{4 + c_{pe}(4-5\gamma)/(2^\mu(4-3\gamma)\bar{r})}{(\kappa/2+1)^\mu}\mathbf{1} := \frac{c_{pi}}{(\kappa/2+1)^\mu}\mathbf{1}, \tag{21}$$

where $\mathcal{T}$ is the global Bellman optimal operator with entropy regularization defined in (7). Further, $\hat{\zeta}^{m+1}$ is $\tilde{\sigma}$ regular and $Q^{\hat{\zeta}^{m+1}}$ is (ν', μ)-decay.

Note that Lemma 8 implies that $\hat{\zeta}^{m+1}$ is $\tilde{\sigma}$-regular and $Q^{\hat{\zeta}^{m+1}}$ is (ν', μ)-decay. Therefore, conditioned on the induction assumption for m, the induction assumption holds for $m+1$ with probability at least $1 - \frac{\delta}{M}$. As a consequence, using a union bound, we must have with probability at least $1 - \delta$,

for all $m = 0, \ldots, M-1$, the induction assumption, and by extension (21), is true. We now condition on this event to show the final convergence bound.

For any $m = 0, \ldots, M-1$, by (21),

$$0 \leq V^* - V^{\hat{\zeta}^{m+1}} = V^* - \mathcal{T}V^{\hat{\zeta}^m} + \mathcal{T}V^{\hat{\zeta}^m} - V^{\hat{\zeta}^{m+1}} \leq \|\mathcal{T}V^{\hat{\zeta}^m} - V^*\|_\infty \mathbf{1} + \frac{c_{pi}}{(\kappa/2+1)^\mu}\mathbf{1}.$$

Taking the infinity norm, we get

$$\begin{aligned}\|V^* - V^{\hat{\zeta}^{m+1}}\|_\infty &\leq \|\mathcal{T}V^{\hat{\zeta}^m} - V^*\|_\infty + \frac{c_{pi}}{(\kappa/2+1)^\mu} \leq \|\mathcal{T}V^{\hat{\zeta}^m} - \mathcal{T}V^*\|_\infty + \frac{c_{pi}}{(\kappa/2+1)^\mu} \\ &\leq \gamma\|V^{\hat{\zeta}^m} - V^*\|_\infty + \frac{c_{pi}}{(\kappa/2+1)^\mu} \leq \gamma^{m+1}\|V^{\hat{\zeta}^0} - V^*\|_\infty + \frac{1}{1-\gamma}\frac{c_{pi}}{(\kappa/2+1)^\mu},\end{aligned}$$

where the third inequality is due to Proposition 1. The desired convergence bound follows by noticing $J(\zeta^*) - J(\hat{\zeta}^M) \leq \|V^* - V^{\hat{\zeta}^M}\|_\infty$.

6 EXPERIMENTS

Though the main focus of this paper is the theoretical convergence of the proposed localized algorithm LPI to the global optimal policy, we also provide some preliminary experiments to demonstrate its empirical advantages in practice.

6.1 Experimental Setup

We first describe an example of networked MARL problems, where n agents cooperatively work together to control the spread of a global process, e.g., information leakage or an infectious disease.

In particular, we consider n agents that are connected through a line graph. For each agent $i \in [n]$, its state is a 2-dimensional vector $s_i = (s_i^1, s_i^2)$, where s_i^1 and s_i^2 take binary values. Each agent has a binary action space, $a_i \in \{0, 1\}$. The value of s_i^1 in state s_i obeys the following transition rule that is not affected by the action:

$$\bar{s}_i^1(t+1) = \begin{cases} 1 & \text{if one of } s_{i-1}^1(t), s_i^1(t), s_{i+1}^1(t) = 1, \\ 0 & \text{otherwise.} \end{cases}$$

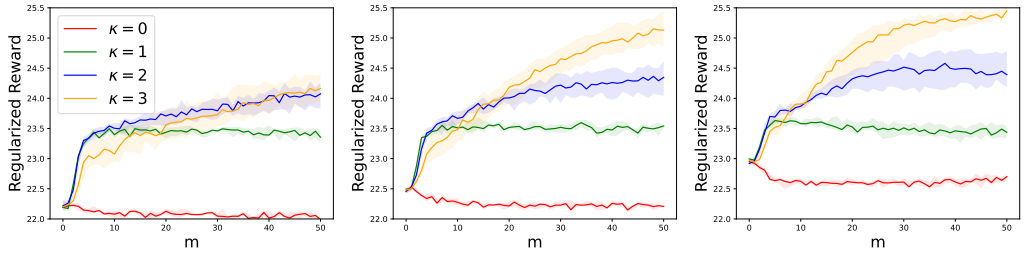
$$P(s_i^1(t+1) = 1 | \bar{s}_i^1(t+1)) = \begin{cases} 1-p_1 & \text{if } \bar{s}_i^1(t+1) = 1, \\ 0 & \text{otherwise,} \end{cases}$$

where $p_1 \in (0, 1)$. To interpret this, each agent first deterministically enters an intermediate state $\bar{s}_i^1(t+1)$ based on its own and neighbors' states in the previous time step. The true $s_i^1(t+1)$ of agent i will be activated (set to 1) with probability $1 - p_1$ or 0 according to the intermediate state. Similarly, the value of s_i^2 in state s_i obeys the following transition rule determined by the action:

$$P\big(\bar{s}_i^2(t+1) = 1 | s_i^2(t), a_i(t)\big) = \begin{cases} 1 & \text{if } s_i^2(t) = 1, \\ p_{\text{eff}} & \text{if } s_i^2(t) = 0, a_i(t) = 1, \\ 0 & \text{otherwise.} \end{cases}$$

$$P\big(s_i^2(t+1) = 1 | \bar{s}_i^2(t+1)\big) = \begin{cases} 1-p_2 & \text{if } \bar{s}_i^2(t+1) = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Here $\bar{s}_i^2(t+1)$ is an intermediate state, which will be activated determnistically if the agent's previous state is already activated, i.e, $s_i^2(t) = 1$, or activated with probability p_{eff} if $s_i^2(t) = 0$ but the agent takes action 1. And the true $s_i^2(t+1)$ remains activated with probability $1 - p_2$.



(a) $\tau = 0.02$ (b) $\tau = 0.05$ (c) $\tau = 0.1$

Fig. 1. Comparison of the performance of LPI for different κ under different levels of entropy regularization.

This model defines a simple form of propagating/spreading dynamics on a graph, with the agents' actions providing a mechanism to control the spread. We decompose the reward into two terms, one depending only on the state $s_i = (s_i^1, s_i^2)$ and the other depending only on the action, $r_i(s_i, a_i) = r_i^s(s_i^1, s_i^2) + r_i^a(a_i)$, where the state reward $r_i^s(\cdot)$ and action reward $r_i^a(\cdot)$ are defined as

$$r_i^s(s_i^a, s_i^b) = \begin{cases} 0 & \text{if } s_i^1 = 1, s_i^2 = 0, \\ 1 & \text{otherwise.} \end{cases} \qquad r_i^a(a_i) = \begin{cases} 1 & \text{if } a_i = 0, \\ 1 - c & \text{if } a_i = 1. \end{cases}$$

It is easy to verify that states $(s_i^1, s_i^2) = (0, 0), (1, 1), (0, 1)$ have a high state reward r_i^s, and $(s_i^1, s_i^2) = (1, 0)$ has a low reward (or incurs a penalty). This corresponds to an agent incurring a penalty if a protection action is not taken before the dynamics reaches the agent. Furthermore, the protection action itself incurs a cost c in the action reward r_i^a.

Intuitively, if an agent does not observe any of its κ-hop neighbors to have $s^1(t) = 1$, it should take action $a(t) = 0$ to get the largest action reward. However, if some of its neighboring agents have $s^1(t) = 1$ and the cost c of taking action 1 is relatively small, it should take action 1 to increase s^2 in order to get a reward $r^s = 1$ even if its s^1 changes to 1 at some time. If the probability p_{eff} that action $a(t) = 1$ being effective is smaller, the agent should start taking action $a = 1$ earlier, when its nearest agent has $s^1(t) = 1$ is at a larger distance, to make sure that it has a high probability to have $s^2 = 1$ when s^1 propagates to its position. In this sense, the optimal policy should depend on not only the local state but the states of other agents as well, and we can expect a large performance difference between localized policy with small κ and the global optimal policy.

Parameter Settings. In our experiments, we use a variation of β-hop Localized TD(0) which has a constant step size of 0.1. We set $\beta = \kappa$ for simplicity. We set the environment parameters to be $p_1 = 0.6$, $p_2 = 0.7$, $c = 0.3$ and $p_{\text{eff}} = 0.4$. We set the training parameters to be $\gamma = 0.95$ and $\eta = 0.05$. We train LPI for 50 outer loops and in the inner loop we perform policy update (Line 9 of Algorithm 1) for $p_{\max} = 10$ steps. For each experiment, we repeat it for 10 times and plot the median in a solid line and 25/75 percent performance values in the shaded area.

6.2 Experimental Results

In Figure 1, we compare the performance of LPI with different choices of κ, which represents how much local information we used in the local policy learning on each agent. We also present different levels of entropy regularization in each subfigure of Figure 1 by setting $\tau = 0.02, 0.05$ and 0.1. We set $n = 8$ in the line graph. It can be seen that a larger κ leads to a better total regularized reward but a lower convergence rate, which is consistent with our theoretical findings. Note that the theoretic

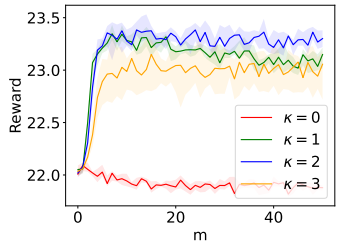
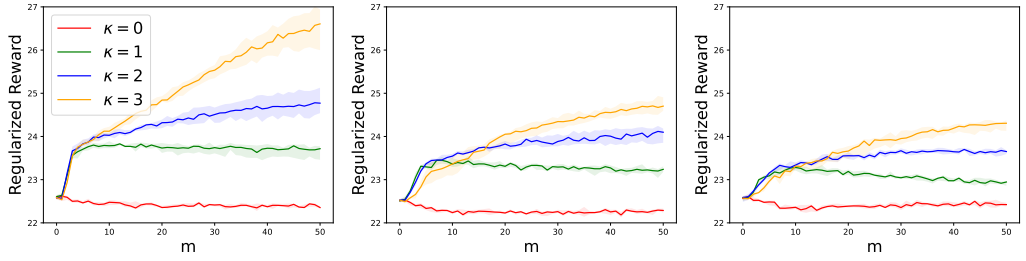


Fig. 2. Comparison of unregularized rewards under different κ when $\tau = 0.05$.



(a) $n = 5$ (b) $n = 10$ (c) $n = 15$

Fig. 3. Comparison of the performance of LPI for different κ in the settings with different network sizes.

results in this paper are for the regularized rewards. For unregularized rewards, we provide Figure 2 which shows the unregularized rewards under different κ when running our algorithm (fixing $\tau = 0.05$). Figure 2 shows when increasing κ from 0 to 2, the unregularized rewards increase, still consistent with our theoretic results. However, $\kappa = 3$ is slightly worse than $\kappa = 1$, and this may be due to our algorithm and guarantee are only tailored to the regularized case. How to better handle the unregularized case remains an interesting future direction.

In Figure 3, we illustrate that our algorithm works for various network sizes by comparing the training curves among different network sizes $n = 5, 10$ and 15 while keeping τ as a constant. In each plot in Figure 3, larger κ always leads to a better total regularized reward. Also, by comparing these three plots, we can see similar convergence rates for these three different network sizes.

We note that some of our experimental settings violate the theoretical requirements on τ, η, and γ in our theorems, whereas LPI still performs well. Therefore, in practice, our algorithm may be more widely applicable than what is suggested by our theoretical analysis, which could be conservative due to proof techniques.

7 CONCLUSION AND FUTURE WORK

This paper studies a cooperative MARL problem in networked systems. We provided the first theoretical guarantee on the performance gap between localized polices that make decisions only based on the information within a neighborhood of each agent and the optimal centralized policy. We achieved this by showing that a κ-hop localized policy where each agent chooses actions based on the states of its κ-hop neighbors is nearly globally optimal. Our analysis relies on a novel characterization of a class of spatial decaying policies. Further, we proposed a specific algorithm, Localized Policy Iteration (LPI), that learns a policy in this class. LPI is communication-efficient

and scales to large networked systems due to its localized implementation. We further provided a finite-sample analysis of LPI showing that it converges to the globally optimal centralized policy and achieves ε-optimality using only local information of the $\kappa = \Theta(\text{poly}(1/\varepsilon))$-hop neighborhood of each agent. It is worth noting that LPI is the first localized algorithm that converges to the centralized globally optimal policy in the networked MARL setting. Finally, we conducted numerical experiments on a MARL problem where multiple agents on a graph cooperate together to control the spreading dynamics of a global process such as information leakage or disease transmission.

There are many interesting questions motivated by this work. For example, it remains an open problem whether the polynomial dependence of the optimality gap on the neighborhood size κ can be improved to a (quasi-)exponential decaying gap, which has been shown in the LQC setting by Shin et al. [30], Zhang et al. [42]. It is also interesting to explore whether our analysis extends to the competitive MARL setting, where the objectives of agents are different from each other. Another future direction is to consider continuous state and action space for each agent. We note that the bounds in Theorem 1 and Theorem 2 depend on $A_{\max}$ (the action space size), which will degenerate as $A_{\max} \to \infty$. Despite this, we conjecture similar results as Theorem 1 and Theorem 2 will still hold in the continuous state/action space case but requires different analysis techniques.

ACKNOWLEDGMENTS

Guannan Qu is supported by NSF Grant EPCN-2154171 and C3 AI Institute. Pan Xu is supported by the startup funding at the Department of Biostatistics and Bioinformatics at Duke University. Yiheng Lin is supported by PIMCO Graduate Fellowship. Zaiwei Chen is supported by PIMCO Postdoc Fellowship and Simoudis Discovery Prize. Adam Wierman is supported by NSF Grants CNS-2146814, CPS-2136197, CNS-2106403, NGSDI-2105648, with additional support from Amazon AWS.

REFERENCES

[1] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. 2020. Optimality and approximation with policy gradient methods in markov decision processes. In *Conference on Learning Theory*. PMLR, 64–66.

[2] Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. 2021. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *Journal of Machine Learning Research* 22, 98 (2021), 1–76.

[3] Bassam Bamieh, Fernando Paganini, and Munther A Dahleh. 2002. Distributed control of spatially invariant systems. *IEEE Transactions on automatic control* 47, 7 (2002), 1091–1107.

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

[5] Jalaj Bhandari and Daniel Russo. 2019. Global optimality guarantees for policy gradient methods. *arXiv preprint arXiv:1906.01786* (2019).

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

[7] Semih Cayci, Niao He, and R Srikant. 2021. Linear convergence of entropy-regularized natural policy gradient with linear function approximation. *Preprint arXiv:2106.04096* (2021).

[8] Shicong Cen, Chen Cheng, Yuxin Chen, Yuting Wei, and Yuejie Chi. 2021. Fast global convergence of natural policy gradient methods with entropy regularization. *Operations Research* (2021).

[9] Xin Chen, Guannan Qu, Yujie Tang, Steven Low, and Na Li. 2022. Reinforcement learning for selective key applications in power systems: Recent advances and future challenges. *IEEE Transactions on Smart Grid* (2022).

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

[11] Dongsheng Ding, Chen-Yu Wei, Kaiqing Zhang, and Mihailo Jovanovic. 2022. Independent policy gradient for large-scale markov potential games: Sharper rates, function approximation, and game-agnostic convergence. In *International Conference on Machine Learning*. PMLR, 5166–5220.

[12] Thinh Doan, Siva Maguluri, and Justin Romberg. 2019. Finite-time analysis of distributed TD (0) with linear function approximation on multi-agent reinforcement learning. In *International Conference on Machine Learning*. PMLR, 1626–1635.

[13] Jakob Foerster, Ioannis Alexandros Assael, Nando De Freitas, and Shimon Whiteson. 2016. Learning to communicate with deep multi-agent reinforcement learning. *Advances in neural information processing systems* 29 (2016).

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

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

[16] Carlos Guestrin, Daphne Koller, Ronald Parr, and Shobha Venkataraman. 2003. Efficient solution algorithms for factored MDPs. *Journal of Artificial Intelligence Research* 19 (2003), 399–468.

[17] Chi Jin, Zeyuan Allen-Zhu, Sebastien Bubeck, and Michael I. Jordan. 2018. Is Q-learning Provably Efficient? *arXiv:1807.03765 [cs, math, stat]* (July 2018). http://arxiv.org/abs/1807.03765 arXiv: 1807.03765.

[18] Michael Kearns and Daphne Koller. 1999. Efficient reinforcement learning in factored MDPs. In *IJCAI*, Vol. 16. 740–747.

[19] Stefanos Leonardos, Will Overman, Ioannis Panageas, and Georgios Piliouras. 2021. Global convergence of multi-agent policy gradient in markov potential games. *arXiv preprint arXiv:2106.01969* (2021).

[20] David A Levin and Yuval Peres. 2017. *Markov chains and mixing times*. Vol. 107. American Mathematical Soc.

[21] Yiheng Lin, Guannan Qu, Longbo Huang, and Adam Wierman. 2021. Multi-agent reinforcement learning in stochastic networked systems. *Advances in Neural Information Processing Systems* 34 (2021), 7825–7837.

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

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

[24] Jincheng Mei, Chenjun Xiao, Csaba Szepesvari, and Dale Schuurmans. 2020. On the global convergence rates of softmax policy gradient methods. In *International Conference on Machine Learning*. PMLR, 6820–6829.

[25] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*. PMLR, 1928–1937.

[26] 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. 2015. Human-level control through deep reinforcement learning. *Nature* 518, 7540 (2015), 529.

[27] Nader Motee and Ali Jadbabaie. 2008. Optimal control of spatially distributed systems. *IEEE Trans. Automat. Control* 53, 7 (2008), 1616–1629.

[28] Guannan Qu, Yiheng Lin, Adam Wierman, and Na Li. 2020. Scalable multi-agent reinforcement learning for networked systems with average reward. *Advances in Neural Information Processing Systems* 33 (2020), 2074–2086.

[29] Guannan Qu, Adam Wierman, and Na Li. 2020. Scalable reinforcement learning of localized policies for multi-agent networked systems. In *Learning for Dynamics and Control*. PMLR, 256–266.

[30] Sungho Shin, Yiheng Lin, Guannan Qu, Adam Wierman, and Mihai Anitescu. 2022. Near-Optimal Distributed Linear-Quadratic Regulator for Networked Systems. *arXiv preprint arXiv:2204.05551* (2022).

[31] 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. 2016. Mastering the game of Go with deep neural networks and tree search. *nature* 529, 7587 (2016), 484.

[32] Sainbayar Sukhbaatar, Rob Fergus, et al. 2016. Learning multiagent communication with backpropagation. *Advances in neural information processing systems* 29 (2016).

[33] Wesley Suttle, Zhuoran Yang, Kaiqing Zhang, Zhaoran Wang, Tamer Başar, and Ji Liu. 2020. A multi-agent off-policy actor-critic algorithm for distributed reinforcement learning. *IFAC-PapersOnLine* 53, 2 (2020), 1549–1554.

[34] Richard S Sutton and Andrew G Barto. 2018. *Reinforcement learning: An introduction*. MIT press.

[35] Ming Tan. 1993. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*. 330–337.

[36] Nino Vieillard, Tadashi Kozuno, Bruno Scherrer, Olivier Pietquin, Rémi Munos, and Matthieu Geist. 2020. Leverage the average: an analysis of KL regularization in reinforcement learning. *Advances in Neural Information Processing Systems* 33 (2020), 12163–12174.

[37] Neil Walton and Kuang Xu. 2021. Learning and information in stochastic networks and queues. In *Tutorials in Operations Research: Emerging Optimization Methods and Modeling Techniques with Applications*. INFORMS, 161–198.

[38] Ronald J Williams and Jing Peng. 1991. Function optimization using connectionist reinforcement learning algorithms. *Connection Science* 3, 3 (1991), 241–268.

[39] Yaodong Yang, Rui Luo, Minne Li, Ming Zhou, Weinan Zhang, and Jun Wang. 2018. Mean field multi-agent reinforcement learning. In *International conference on machine learning*. PMLR, 5571–5580.

[40] Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. 2021. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of Reinforcement Learning and Control* (2021), 321–384.

[41] Kaiqing Zhang, Zhuoran Yang, Han Liu, Tong Zhang, and Tamer Basar. 2018. Fully decentralized multi-agent reinforcement learning with networked agents. In *International Conference on Machine Learning*. PMLR, 5872–5881.
[42] Runyu Zhang, Weiyu Li, and Na Li. 2022. On the Optimal Control of Network LQR with Spatially-Exponential Decaying Structure. *arXiv preprint arXiv:2209.14376* (2022).
[43] Shangtong Zhang, Remi Tachet, and Romain Laroche. 2021. Global optimality and finite sample analysis of softmax off-policy actor critic under state distribution mismatch. *Preprint arXiv:2111.02997* (2021).

A NOTATION

The notations used in this paper are summarized in Table 1.

Table 1. Notations

Symbol	Definition
$\lvert\mathcal{S}\rvert$	The cardinality of state space $\mathcal{S}$. Similar notations apply to $\mathcal{A}$, $\mathcal{S}_i$, and $\mathcal{A}_i$.
$A_{\max}$	The maximum cardinality for local action spaces, i.e., $A_{\max} = \arg\max_i \lvert\mathcal{A}_i\rvert$.
$[n]$	Index set $\{1, 2, \ldots, n\}$.
β, κ.	The neighborhood sizes of the approximation for policy evaluation and policy improvement respectively in Algorithm 1.
$\mathcal{N}_i$.	The set including node i and all its neighbors on the graph.
$\mathcal{N}_i^\kappa$	The set including node i and all its κ-hop neighbors on the graph.
$\mathcal{N}_{-i}$	The complement set of $\mathcal{N}_i$, i.e., $\mathcal{N}/\mathcal{N}_i$. Similarly, $\mathcal{N}_{-i}^\kappa = \mathcal{N}/\mathcal{N}_i^\kappa$.
$f(\kappa)$	The size of the largest κ-hop neighborhood, $f(\kappa) = \sup_{i\in\mathcal{N}} \lvert\mathcal{N}_i^\kappa\rvert$.
$s_{\mathcal{N}_i^\kappa}$	The states of agents in $\mathcal{N}_i^\kappa$. Similarly for $s_{\mathcal{N}_{-i}^\kappa}$ and other subscripts.
$\overline{s_{\mathcal{J}}}$	The extension of a local state $s_{\mathcal{J}}$ that are supported on a set of indices $\mathcal{J}$ to a global state. See Definition 6.
ρ	The initial state distribution.
τ	The entropy regularization parameter.
$\Delta_{\nu,\mu,\sigma}$	The set of (ν, μ)-decay and σ-regular policy. See Definition 2 and Definition 3.
$\Delta_{\mathcal{A}_i}$	the probability simplex over the set $\mathcal{A}_i$.
$\Delta_{\mathcal{A}_i\lvert\mathcal{S}}$	The space of distributions over $\mathcal{A}_i$ that can depend on elements in $\mathcal{S}$. This is the space agent i's policy lies in.
Δ_{policy}	The space of all possible joint policies, cf. Section 2.1.
$\zeta, \pi, \hat{\zeta}, \hat{\pi}$	Letters reserved for policies. Letter π is used for policies in the inner-loop multiplicative weight updates. The hat indicates the policy is a κ-hop policy.
V^ζ, Q^ζ	Value and Q functions for a given policy ζ, cf. (2) and (3).
V_i^ζ, Q_i^ζ	Local value and Q functions of agent i for a given policy ζ, cf. (5) and (6).
$\mathbf{Q}$	Vector $(Q_1, \ldots, Q_n)$ that denotes the complete profile of all local Q functions.
C	The matrix that characterizes the total variation of transition probabilities. See (10).
Z^ζ	The interaction matrix of a policy $\zeta = (\zeta_1, \ldots, \zeta_n)$. See Definition 2.
$Z^{\mathbf{Q}}$	The interaction matrix of the local Q functions $\mathbf{Q} = \{Q_i\}_i^n$. See Definition 4.
H^Q	The second-order interaction matrix of a Q function. See Definition 8.
$O(\cdot)$	We write $f(n) = O(g(n))$ if there is a constant c such that $f(n) \le cg(n)$ for all large enough n.
$\Omega(\cdot)$	We write $f(n) = \Omega(g(n))$ if there is a constant c such that $f(n) \ge cg(n)$ for all large enough n.
$\Theta(\cdot)$	We write $f(n) = \Theta(g(n))$ if both $f(n) = O(g(n))$ and $f(n) = \Omega(g(n))$.

B PROOF OF THEOREM 6: CLOSURE OF POLICY CLASS $\Delta_{\nu,\mu,\sigma}$ UNDER POLICY IMPROVEMENT

To prove the closure of policy class $\Delta_{\nu,\mu,\sigma}$ under policy improvement (i.e., Theorem 6), our first step is to show that for a policy $\zeta \in \Delta_{\nu,\mu,\sigma}$, its local Q functions $\mathbf{Q}^\zeta = \{Q_i^\zeta\}_{i\in\mathcal{N}}$ will also be (ν', μ)-decay for some ν' (Lemma 9). Secondly, we show that if the local Q functions are (ν', μ)-decay, conducting policy improvement with respect to it will lead to a (ν'', μ) decay policy (Lemma 10), with $\nu'' \le \nu$

(under the parameter settings of Theorem 6). Combining the two lemmas will lead to the closure property, i.e. part (a) and (b) of Theorem 6. The proofs of Lemma 9 and Lemma 10 can be found in Appendix C.

We first present Lemma 9 below, showing decaying policies have decaying local Q functions.

LEMMA 9. *For a policy $\zeta = (\zeta_1, \ldots, \zeta_n) \in \Delta_{\nu,\mu,\sigma}$, the interaction matrix of its corresponding local Q-functions $\mathbf{Q}^\zeta = \{Q_i^\zeta\}_{i\in\mathcal{N}}$ is given by*

$$Z^{\mathbf{Q}^\zeta} \leq \bar{r}I + \gamma(\bar{r} + n\tau\sigma)\sum_{t=0}^{\infty}\gamma^t Z^\zeta (C + Z^\zeta)^t C,$$

where Z^ζ is the interaction matrix for policy ζ and $C = [C_{ij}]_{i,j=1,\ldots,n}$ is defined in (10), and the "$\leq$" above is element-wise. Furthermore, if we assume $\forall i, \sum_{j\in\mathcal{N}_i} 2^\mu C_{ij} \leq 1/2$ and $\nu \leq 1/2$, then $\mathbf{Q}^\zeta$ is (ν', μ)-decay with $\nu' = \bar{r} + \frac{\gamma(\bar{r}+n\tau\sigma)}{4(1-\gamma)}$.

We next present Lemma 10, showing that given local Q functions that are (ν', μ)-decay, the policy improvement step will result in a (ν'', μ)-decay policy for some ν'' given below.

LEMMA 10. *Consider Q-functions that are (ν', μ)-decay. For $\forall s$, let $\zeta(\cdot|s) = (\zeta_1(\cdot|s), \ldots, \zeta_n(\cdot|s))$ be the solution to the following policy improvement step:*

$$\max_{\zeta_1(\cdot|s)\in\Delta_{\mathcal{A}_1},\ldots,\zeta_n(\cdot|s)\in\Delta_{\mathcal{A}_n}} \mathbb{E}_{a_1\sim\zeta_1(\cdot|s),\ldots,a_n\sim\zeta_n(\cdot|s)}\left[Q(s,a) - \sum_{i=1}^{n}\tau\log\zeta_i(a_i|s)\right]. \tag{22}$$

Then, when $\tau \geq 3\times 2^{\mu+1}\nu' A_{\max}^2 e$, ζ is a $\sigma' = \frac{\nu'}{n\tau}$ regular and (ν'', μ)-decay policy, where $\nu'' = \frac{2^{\mu+1}\nu' A_{\max}^2 e^{\frac{\nu'}{n\tau}}}{\tau - 2^{\mu+1}\nu' A_{\max}^2 e^{\frac{\nu'}{n\tau}}}$.

We now prove parts (a) (b) of Theorem 6 by induction. Consider $\zeta^m \in \Delta_{\nu,\mu,\sigma}$. Under the parameter settings in Theorem 6, we have $2^\mu \sup_i \sum_{j\in\mathcal{N}_i} C_{ij} \leq 1/2$, and $\nu = 1/2$. As a result, we can apply Lemma 9 to ζ^m and get $\mathbf{Q}^{\zeta^m}$ is (ν', μ) decay, where we have utilized the easy to check fact that $\bar{r} + \frac{\gamma(\bar{r}+n\tau\sigma)}{4(1-\gamma)} = \nu'$ (using $\sigma = \frac{\nu'}{n\tau}$). Then, notice the parameter settings in Theorem 6 leads to $\tau \geq 2^\mu 6\frac{4-3\gamma}{4-5\gamma}\bar{r}A_{\max}^2 e \geq 3\times 2^{\mu+1}\nu' A_{\max}^2 e^{\frac{\nu'}{n\tau}}$. Therefore, we can now apply Lemma 10 to Q^{ζ^m} to show that ζ^{m+1} is σ regular and (ν'', μ) decay with $\nu'' \leq 1/2 = \nu$. As a result, $\zeta^{m+1} \in \Delta_{\nu,\mu,\sigma}$ and the induction is finished.

This concludes the proof for part of (a) and (b) of Theorem 6. For part (c), note that

$$0 \leq V^* - V^{\zeta^{m+1}} = V^* - \mathcal{T}V^{\zeta^m} + \mathcal{T}V^{\zeta^m} - V^{\zeta^{m+1}} \leq \|\mathcal{T}V^{\zeta^m} - V^*\|_\infty \mathbf{1}.$$

Taking the infinity norm, we get

$$\begin{aligned}
\|V^* - V^{\zeta^{m+1}}\|_\infty &\leq \|\mathcal{T}V^{\zeta^m} - V^*\|_\infty \\
&\leq \|\mathcal{T}V^{\zeta^m} - \mathcal{T}V^*\|_\infty \\
&\leq \gamma\|V^{\zeta^m} - V^*\|_\infty \\
&\leq \gamma^{m+1}\|V^{\zeta^0} - V^*\|_\infty,
\end{aligned}$$

where the third inequality is due to Proposition 1.

C PROOF OF LEMMAS IN APPENDIX B

C.1 Proof of Lemma 9

Before we present the proof of Lemma 9, we lay out some useful lemmas that will be frequently used in our analysis.

Firstly, we use the following lemma regarding the summation and product of decaying matrices, the proof of which can be found in Appendix D.3.

LEMMA 11. *Let $A \in \mathbb{R}^{n\times n}$ be a (ν,μ)-decay matrix and $A' \in \mathbb{R}^{n\times n}$ be a (ν',μ)-decay matrix. Then we have the following results.*

(a) $cA + c'A'$ is a $(c\nu + c'\nu',\mu)$-decay matrix, where $c, c' \geq 0$ are arbitrary constants.
(b) AA' is a $(\nu\nu',\mu)$-decay matrix.

Next, we will use the Lipschitz property of the entropy function, the proof of which is deferred to Appendix D.4.

LEMMA 12. *Let d, d' be two σ-regular distributions over the same set $\{1, 2, \cdots, M\}$. Then we have $|H(d) - H(d')| \leq \sigma \mathrm{TV}(d, d')$, where $H(\cdot)$ is the entropy function.*

Our proof also uses the following two lemmas, whose proofs can be found in [28].

LEMMA 13 (LEMMA 5 IN QU ET AL. [28]). *Let $f(\cdot)$ be a function that maps $\prod_{i\in V} \mathcal{Z}_i$ to $\mathbb{R}$, where $\mathcal{Z}_i$ is a finite set for all i. Let P_i and $\tilde{P}_i$ be two distributions on $\mathcal{Z}_i$, and further let P (and $\tilde{P}$) be the product distribution of P_i (and $\tilde{P}_i$). Then we have*

$$\left|\mathbb{E}_{z\sim P} f(z) - \mathbb{E}_{z\sim \tilde{P}} f(z)\right| \leq \sum_{i\in V} \mathrm{TV}(P_i, \tilde{P}_i)\delta_i(f),$$

where $\delta_i(f) = \sup_{z_{V/i}} \sup_{z_i, z_i'} |f(z_i, z_{V/i}) - f(z_i', z_{V/i})|$.

LEMMA 14 (LEMMA 3,4 IN QU ET AL. [28]). *Consider a Markov Chain with state $z = (z_1, \ldots, z_n) \in \mathcal{Z} = \mathcal{Z}_1 \times \cdots \times \mathcal{Z}_n$, where each $\mathcal{Z}_i$ is some finite set. Suppose its transition probability factorizes as*

$$P(z(t+1)|z(t)) = \prod_{i=1}^{n} P_i(z_i(t+1)|z(t))$$

and further, define C^z to be the following matrix,

$$C^z_{ij} = \sup_{z_{-j}} \sup_{z_j, z_j'} \mathrm{TV}(P_i(\cdot|z_j, z_{-j}), P_i(\cdot|z_j', z_{-j}))$$

Then, fixing a pair of (i, j) for any $z = (z_j, z_{-j})$, $z' = (z_j', z_{-j})$, the following holds.

(a) We have,

$$\mathrm{TV}(\pi_{t,i}, \pi'_{t,i}) \leq [(C^z)^t e_j]_i,$$

where $\pi_{t,i}$ is the distribution of $z_i(t)$ given $z(0) = z$, and $\pi'_{t,i}$ is the distribution of $z_i(t)$ given $z(0) = z'$, and e_j is the indicator vector in which the j'th entry is 1 and all other entries are 0.

(b) For any function $f : \mathbb{R}^{\mathcal{Z}} \to \mathbb{R}$, we have

$$\left|\mathbb{E}_{z\sim\pi_t} f(z) - \mathbb{E}_{z\sim\pi'_t} f(z)\right| \leq \sum_{\ell=1}^{n} [(C^z)^t e_j]_\ell \delta_\ell(f)$$

where $\delta_\ell(f) = \sup_{z_{-\ell}} \sup_{z_\ell, z_\ell'} |f(z_\ell, z_{-\ell}) - f(z_\ell', z_{-\ell})|$ and π_t is the distribution of $z(t)$ given $z(0) = z$, and π'_t is the distribution of $z(t)$ given $z(0) = z'$

We start our proof by showing the following Lemma 15, where we treat the Markov Chain in Lemma 14 as the Markov Chain induced by policy ζ and bound the C^z matrix. The proof of Lemma 15 is deferred to Appendix D.5.

Lemma 15. *In the settings of Lemma 14, we set the Markov Chain as the induced Markov Chain of our MDP when using policy ζ, and we treat $z_i = s_i$ and $z = s$. The transition probabilities of this induced Markov Chain is given by,*

$$P(z(t+1)|z(t)) = \prod_{i=1}^{n} \underbrace{\sum_{a_i(t)\in\mathcal{A}_i} \left[P_i(s_i(t+1)|s_{\mathcal{N}_i}(t), a_i(t))\zeta_i(a_i(t)|s(t))\right]}_{:=P_i^{\zeta}(s_i(t+1)|s(t))}$$

the resulting Markov chain's interaction strength parameter $C^z = [C^z_{ij}]_{i,j=1,\dots,n}$ (as defined in Lemma 14) satisfies

$$C^z_{ij} = \sup_{s_{-j}} \sup_{s_j, s_j'} \mathrm{TV}(P_i^{\zeta}(\cdot|s_j, s_{-j}), P_i^{\zeta}(\cdot|s_j', s_{-j})) \le Z_{ij}^{\zeta} + C_{ij},$$

where $[Z_{ij}^{\zeta}]_{i,j=1,\dots,n}$ is the interaction matrix of policy ζ (cf. Definition 2), and $C = [C_{ij}]_{i,j=1,\dots,n}$ is defined in (10).

Now we start proving Lemma 9.

Proof of Lemma 9. Given the relationship between Q_i^{ζ} and V_i^{ζ} in (6), it is easy to compute the interaction matrix for the local value functions $\mathbf{V}^{\zeta} = \{V_i^{\zeta}\}_{i\in\mathcal{N}}$, which we define as below:

$$Z_{ij}^{\mathbf{V}^{\zeta}} = \sup_{s_{-j}} \sup_{s_j, s_j'} |V_i^{\zeta}(s_j, s_{-j}) - V_i^{\zeta}(s_j', s_{-j})|. \tag{23}$$

Fixing i, j, s_j, s_j', s_{-j}, define π_t (or π_t') to be the distribution of $s(t)$ given $s(0) = (s_j, s_{-j})$ (or $s(0) = (s_j', s_{-j})$). Let $r_i^{\zeta}(s) = \mathbb{E}_{a_i\sim\zeta_i(\cdot|s)}[r_i(s_i, a_i)]$. Also, let $H_i^{\zeta}(s) = -\sum_{a_i\in\mathcal{A}_i}\zeta_i(a_i|s)\log\zeta_i(a_i|s)$ be the entropy of local policy $\zeta_i(\cdot|s)$ as a function of s. Then, by (5), we have

$$\begin{aligned}
&|V_i^{\zeta}(s_j, s_{-j}) - V_i^{\zeta}(s_j', s_{-j})| \\
&= \left|\mathbb{E}_{a(t)\sim\zeta(\cdot|s(t))}\left[\sum_{t=0}^{\infty}\gamma^t\left[r_i(s_i(t), a_i(t)) - n\tau\log(\zeta_i(a_i(t)|s(t)))\right]\middle| s(0) = (s_j, s_{-j})\right]\right. \\
&\quad \left. - \mathbb{E}_{a(t)\sim\zeta(\cdot|s(t))}\left[\sum_{t=0}^{\infty}\gamma^t\left[r_i(s_i(t), a_i(t)) - n\tau\log(\zeta_i(a_i(t)|s(t)))\right]\middle| s(0) = (s_j', s_{-j})\right]\right| \\
&\le \sum_{t=0}^{\infty}\gamma^t\left|\mathbb{E}_{s\sim\pi_t}[r_i^{\zeta}(s) + n\tau H_i^{\zeta}(s)] - \mathbb{E}_{s\sim\pi_t'}[r_i^{\zeta}(s) + n\tau H_i^{\zeta}(s)]\right| \\
&\le \sum_{t=0}^{\infty}\gamma^t\sum_{\ell=1}^{n}[(C^z)^t e_j]_{\ell}\delta_{\ell}(r_i^{\zeta}(\cdot) + n\tau H_i^{\zeta}(\cdot)),
\end{aligned}$$

where in the last inequality we have used Statement (b) of Lemma 14. Now let's compute $\delta_{\ell}(r_i^{\zeta}(\cdot) + n\tau H_i^{\zeta}(\cdot))$. For any $s_{\ell}, s_{\ell}', s_{-\ell}$, we have by definition that

$$\begin{aligned}
|r_i^{\zeta}(s_{\ell}, s_{-\ell}) - r_i^{\zeta}(s_{\ell}', s_{-\ell})| &= |\mathbb{E}_{a_i\sim\zeta_i(\cdot|s_{\ell}, s_{-\ell})}r_i(s_i, a_i) - \mathbb{E}_{a_i\sim\zeta_i(\cdot|s_{\ell}', s_{-\ell})}r_i(s_i, a_i)| \\
&\le \mathrm{TV}(\zeta_i(\cdot|s_{\ell}, s_{-\ell}), \zeta_i(\cdot|s_{\ell}', s_{-\ell}))\bar{r}
\end{aligned}$$

$$\leq Z^{\zeta}_{i\ell}\bar{r},$$

where Z^{ζ} is the interaction matrix of global policy ζ. Similarly, we have by Lemma 12,

$$|H^{\zeta}_i(s_\ell, s_{-\ell}) - H^{\zeta}_i(s'_\ell, s_{-\ell})| \leq \sigma \mathrm{TV}(\zeta_i(\cdot|s_\ell, s_{-\ell}), \zeta_i(\cdot|s'_\ell, s_{-\ell})) \leq \sigma Z^{\zeta}_{i\ell},$$

which immediately implies that

$$\delta_\ell(r^{\zeta}_i(\cdot) + n\tau H^{\zeta}_i(\cdot)) \leq Z^{\zeta}_{i\ell}(\bar{r} + n\tau\sigma).$$

As a result,

$$|V^{\zeta}_i(s_j, s_{-j}) - V^{\zeta}_i(s'_j, s_{-j})| \leq \sum_{t=0}^{\infty}\gamma^t \sum_{\ell=1}^{n}[(C^z)^t e_j]_\ell Z^{\zeta}_{i\ell}(\bar{r} + n\tau\sigma) = (\bar{r} + n\tau\sigma)\sum_{t=0}^{\infty}\gamma^t [Z^{\zeta}(C^z)^t]_{ij}.$$

By the definition in (23), we further have

$$Z^{\mathbf{V}^{\zeta}} \leq (\bar{r} + n\tau\sigma)\sum_{t=0}^{\infty}\gamma^t Z^{\zeta}(C^z)^t. \tag{24}$$

It remains to convert this result to the interaction matrix of the local Q-functions $\mathbf{Q}^{\zeta} = \{Q^{\zeta}_i\}_{i\in\mathcal{N}}$. Recall that,

$$Q^{\zeta}_i(s, a) = r_i(s_i, a_i) + \gamma \mathbb{E}_{\bar{s}\sim P(\cdot|s,a)}\left[V^{\zeta}_i(\bar{s})\right].$$

Therefore, we have

$$\begin{aligned}
&|Q^{\zeta}_i(s_j, a_j, s_{-j}, a_{-j}) - Q^{\zeta}_i(s'_j, a'_j, s_{-j}, a_{-j})| \\
&\leq \bar{r}\mathbf{1}(j = i) + \gamma\left|\mathbb{E}_{\bar{s}\sim P(\cdot|s_j,a_j,s_{-j},a_{-j})}V^{\zeta}_i(\bar{s}) - \mathbb{E}_{\bar{s}\sim P(\cdot|s'_j,a'_j,s_{-j},a_{-j})}V^{\zeta}_i(\bar{s})\right| \\
&\leq \bar{r}\mathbf{1}(j = i) + \gamma\sum_{\ell\in N_j} C_{\ell j} Z^{\mathbf{V}^{\zeta}}_{i\ell},
\end{aligned}$$

where in the second inequality we have used Lemma 13 and the definition of $C_{\ell,j}$. So by Definition 4 and (24) we further have

$$Z^{\mathbf{Q}^{\zeta}} \leq \bar{r}I + \gamma Z^{\mathbf{V}^{\zeta}} C = \bar{r}I + \gamma(\bar{r} + n\tau\sigma)\sum_{t=0}^{\infty}\gamma^t Z^{\zeta}(C^z)^t C. \tag{25}$$

This gives rises to the upper bound on $Z^{\mathbf{Q}^{\zeta}}$ in Lemma 9.

Further, consider the scenario where we assume Z^{ζ} is (ν, μ)-decay, $\nu \leq \frac{1}{2}$ and $\forall i, \sum_{j\in\mathcal{N}_i} 2^{\mu} C_{ij} \leq 1/2$. This means C is $(1/2, \mu)$-decay since by (10), C is a sparse matrix and $C_{ij} = 0$ when $\mathrm{dist}(i, j) > 1$. Combining these with the fact that $C^z_{ij} \leq Z^{\zeta}_{ij} + C_{ij}$ (Lemma 15), we can easily prove that C^z is $(\nu + 1/2, \mu)$-decay by Lemma 11. Also note that I is $(1, \mu)$-decay for any μ. Then applying Lemma 11 to (25), we can show that $Z^{\mathbf{Q}^{\zeta}}$ is (ν', μ)-decay, where

$$\nu' = \bar{r} + \gamma(\bar{r} + n\tau\sigma)\sum_{t=0}^{\infty}\gamma^t \nu(\nu + 1/2)^t \frac{1}{2} = \bar{r} + \nu(\bar{r} + n\tau\sigma)\frac{\frac{1}{2}\gamma}{1 - \gamma(\nu + 1/2)} \leq \bar{r} + \frac{\gamma(\bar{r} + n\tau\sigma)}{4(1 - \gamma)}, \tag{26}$$

where the inequality uses $\nu \leq 1/2$. This completes the proof. □

C.2 Proof of Lemma 10

We first define the notion of the second order interaction matrix for the Q function.

Definition 8. *Consider a Q-function $Q(s,a)$ defined on the global state-action pair. The second order interaction matrix $H^Q = [H^Q_{ij}]_{i,j=1,\dots,n}$ is defined by*

$$H^Q_{ij} = \sup_{z_i,z_j,z_i',z_j'} \sup_{z_{-(i,j)}} \left|[Q(z_i,z_j,z_{-(i,j)}) - Q(z_i',z_j,z_{-(i,j)})] - [Q(z_i,z_j',z_{-(i,j)}) - Q(z_i',z_j',z_{-(i,j)})]\right|,$$

where $-(i,j)$ denotes the set of nodes outside of i and j, and here for notational simplicity, we use z to represent state-action pairs, e.g. we use z_i to represent (s_i,a_i) and similarly $z_{-(i,j)} = (s_{-(i,j)}, a_{-(i,j)})$.

Our first result is Lemma 16 on the decay property of the second order interaction matrix of the global Q function of a given policy. The proof is deferred to Appendix D.6

Lemma 16. *Suppose the local Q functions $\mathbf{Q} = \{Q_i\}_{i\in\mathcal{N}}$ is (ν',μ)-decay. Then, for the corresponding global Q function $Q = \frac{1}{n}\sum_{i\in\mathcal{N}} Q_i$ its second order interaction matrix H^Q is $(2^{\mu+1}\nu',\mu)$-decay.*

We will also frequently use the following auxilliary result, the proof of which is deferred to Appendix D.7.

Lemma 17. *For two distributions d, d' on $\{1,2,\dots,m\}$, suppose for all i,j, we have $|\log\frac{d_i}{d_j} - \log\frac{d_i'}{d_j'}| \le \epsilon$, and $\max(|\log\frac{d_i}{d_j}|, |\log\frac{d_i'}{d_j'}|) \le c$. Then, $\mathrm{TV}(d,d') \le \frac{1}{2}m^2e^c(e^\epsilon - 1)$.*

Now we prove Lemma 10. For each state s we consider the following entropy regularized optimization problem over $\Delta_{A_1}\times\cdots\Delta_{A_n}$,

$$(\zeta_1(\cdot|s),\dots,\zeta_n(\cdot|s)) \leftarrow \mathop{\arg\max}_{\pi_1(\cdot|s),\dots,\pi_n(\cdot|s)} \mathbb{E}_{a_i\sim\pi_i(\cdot|s)}[Q(s,a_1,\dots,a_n)] + \tau\sum_{i=1}^n H(\pi_i(\cdot|s)), \tag{27}$$

where H is the entropy function. When the context is clear, we also use $\pi_i(s)$ to denote the distribution $\pi_i(\cdot|s)$. We also will aggregate the subscripts, e.g. $\pi(s)$ will denote the collection of $(\pi_i(s))_{i=1}^n$; $\pi_{-i}(s)$ denotes $(\pi_j(s))_{j\neq i}$. We consider the following multiplicative weights algorithm (where p is the iteration counter):

$$\pi_i^{p+1}(a_i|s) \propto \pi_i^p(a_i|s)^{1-\eta\tau}\exp(\eta\mathbb{E}_{a_{-i}\sim\pi_{-i}^p(s)}Q(s,a_{-i},a_i)), i=1,2,\dots,n, \tag{28}$$

and we will set the initial distribution ($p=0$) $\pi_i^0(\cdot|s)$ to be the uniform distribution over $\mathcal{A}_i$. The convergence of algorithm (28) is shown in the following Lemma 18 whose proof can be found in Appendix D.8.

Lemma 18. *When $\tau > 2\times 2^{\mu+1}\nu' A_{\max}^2 e^{\nu'/n\tau}$ and $\eta = \frac{1}{\tau}$, for each s, optimization problem (27) has a unique solution $(\zeta_1(\cdot|s),\dots,\zeta_n(\cdot|s))$ and the algorithm (28) will converge to it with a geometric convergence rate:*

$$\sup_{i\in\mathcal{N}} \mathrm{TV}(\pi_i^p(\cdot|s),\zeta_i(\cdot|s)) \le \left(\frac{1}{\tau}2^{\mu+1}\nu' A_{\max}^2 e^{\nu'/n\tau}\right)^p.$$

Proof of Lemma 10. Recall that Lemma 10 says that the ζ obtained by (22) is σ'-regular and (ν'',μ)-decay. We show these two properties now.

Proof of σ'-regular. By considering (28), we have for each i and a_i

$$\log\pi_i^{p+1}(a_i|s) = (1-\eta\tau)\log\pi_i^p(a_i|s) + \eta\mathbb{E}_{a_{-i}\sim\pi_{-i}^p(s)}[Q(s,a_{-i},a_i)] + c(s), \tag{29}$$

where $c(s)$ is a normalization constant that only depends on s and $\eta = \frac{\tau}{n}$. Therefore, for any 2 action pairs $a_i, \tilde{a}_i$, we have

$$\begin{aligned}
&\overbrace{\log \pi_i^{p+1}(a_i|s) - \log \pi_i^{p+1}(\tilde{a}_i|s)}^{\xi_i^{p+1}(s)} \\
&= \underbrace{(1-\eta\tau)(\log \pi_i^{p}(a_i|s) - \log \pi_i^{p}(\tilde{a}_i|s))}_{:=\xi_i^{p}(s)} + \eta \mathbb{E}_{a_{-i}\sim\pi_{-i}^p(s)}[Q(s, a_{-i}, a_i) - Q(s, a_{-i}, \tilde{a}_i)],
\end{aligned}$$

where for notational simplicity, we denote $\log \pi_i^{p+1}(a_i|s) - \log \pi_i^{p+1}(\tilde{a}_i|s) = \xi_i^{p+1}(s)$ for now (where we have fixed an action pair $(a_i, \tilde{a}_i)$. Note that,

$$|Q(s, a_{-i}, a_i) - Q(s, a_{-i}, \tilde{a}_i)| \le \frac{1}{n}\sum_{\ell=1}^{n} |Q_\ell(s, a_{-i}, a_i) - Q_\ell(s, a_{-i}, \tilde{a}_i)| \le \frac{1}{n}\sum_{\ell=1}^{n} Z_{\ell i}^Q \le \frac{\nu'}{n}.$$

Next, we show that $|\xi_i^p(s)| \le \sigma'$ by induction. The statement is clearly true for $p = 0$ as $\pi_i^0(\cdot|s)$ is an uniform distribution. Suppose the statement is true for p, we have

$$|\xi_i^{p+1}(s)| \le (1-\eta\tau)\sigma' + \eta\frac{\nu'}{n} \le \sigma', \tag{30}$$

where we have used the fact that $\sigma' = \frac{\nu'}{n\tau}$.

Proof of (ν'', μ)-decay. The bulk of the proof is to show that for any p, $\pi^p = (\pi_1^p, \ldots, \pi_n^p)$ is a (ν'', μ)-decay policy under the assumptions of Lemma 10, in other words, the interaction matrix Z^{π^k} is a (ν'', μ)-decay matrix. Now we fix a j and consider (28) for two values of s: $s = (s_j, s_{-j})$ and $s' = (s'_j, s_{-j})$. The respective trajectories are $(\pi_i^p(\cdot|s))_{i=0,\ldots,n}$ and $(\pi_i^p(\cdot|s'))_{i=0,\ldots,n}$ where p is the iteration counter, and the two have the same initializer, that is $\pi_i^0(\cdot|s) = \pi_i^0(\cdot|s')$ as we initialize both with the uniform distribution. Then, consider (29) for s and s', we get

$$\begin{aligned}
&\xi_i^{p+1}(s) - \xi_i^{p+1}(s') \\
&= (1-\eta\tau)(\xi_i^p(s) - \xi_i^p(s')) \\
&\quad + \eta\mathbb{E}_{a_{-i}\sim\pi_{-i}^p(s)}[Q(s, a_{-i}, a_i) - Q(s, a_{-i}, \tilde{a}_i)] - \eta\mathbb{E}_{a_{-i}\sim\pi_{-i}^p(s')}[Q(s', a_{-i}, a_i) - Q(s', a_{-i}, \tilde{a}_i)] \\
&= (1-\eta\tau)(\xi_i^p(s) - \xi_i^p(s')) \\
&\quad + \underbrace{\eta\mathbb{E}_{a_{-i}\sim\pi_{-i}^p(s)}[Q(s, a_{-i}, a_i) - Q(s, a_{-i}, \tilde{a}_i)] - \eta\mathbb{E}_{a_{-i}\sim\pi_{-i}^p(s')}[Q(s, a_{-i}, a_i) - Q(s, a_{-i}, \tilde{a}_i)]}_{:=I_1} \\
&\quad + \underbrace{\eta\mathbb{E}_{a_{-i}\sim\pi_{-i}^p(s')}\Big([Q(s, a_{-i}, a_i) - Q(s, a_{-i}, \tilde{a}_i)] - [Q(s', a_{-i}, a_i) - Q(s', a_{-i}, \tilde{a}_i)]\Big)}_{:=I_2}.
\end{aligned}$$

The absolute values of I_1, I_2 can be bounded as follows:

$$|I_1| \le \eta \sum_{\ell\neq i} \mathrm{TV}(\pi_\ell^p(s), \pi_\ell^p(s'))H_{i\ell}^Q,$$

where we have used Lemma 13 and the definition of second order interaction matrix H^Q. Similarly, for term I_2, we have

$$|I_2| \le \eta H_{ij}^Q.$$

This shows that,

$$|\xi_i^{p+1}(s) - \xi_i^{p+1}(s')| \le (1-\eta\tau)|\xi_i^p(s) - \xi_i^p(s')| + \eta \sum_{\ell \neq i} \mathrm{TV}(\pi_\ell^p(s), \pi_\ell^p(s')) H_{i\ell}^Q + \eta H_{ij}^Q.$$

Denote vector $H_{:j}^Q$ as the j'th column of H^Q, $H_{i:}^Q$ as the i'th row, and $H_{jj}^Q = 0$ for any $i, j \in [n]$. Let vector v^p denote $v_i^p = \mathrm{TV}(\pi_i^p(s), \pi_i^p(s'))$. Then it holds that

$$\begin{aligned}|\xi_i^{p+1}(s) - \xi_i^{p+1}(s')| &\le (1-\eta\tau)|\xi_i^p(s) - \xi_i^p(s')| + \eta H_{i:}^Q v^p + \eta H_{ij}^Q \\ &\le (1-\eta\tau)^{p+1}|\xi_i^0(s) - \xi_i^0(s')| + \sum_{k=0}^{p} \eta(1-\eta\tau)^{p-k}(H_{i:}^Q v^k + H_{ij}^Q).\end{aligned}$$

Recall the definition $\xi_i^{p+1}(s) = \log \pi_i^{p+1}(a_i|s) - \log \pi_i^{p+1}(\tilde{a}_i|s)$ and that $|\xi_i^{p+1}(s)| \le \nu'/\tau$ by (30). By Lemma 17, we obtain the following bound on $v_i^{p+1} = \mathrm{TV}(\pi_i^{p+1}(s), \pi_i^{p+1}(s'))$:

$$v_i^{p+1} \le \frac{|\mathcal{A}_i|^2 e^{\sigma'}}{2}\left(\exp\left((1-\eta\tau)^{p+1}|\xi^0(s) - \xi^0(s')| + \sum_{k=0}^{p} \eta(1-\eta\tau)^{p-k}(H_{i:}^Q v^k + H_{ij}^Q)\right) - 1\right).$$

where $|\mathcal{A}_i|$ is the cardinality of $\mathcal{A}_i$. Now we choose $\eta = 1/\tau$ to simplify the presentation. Note that in general $\eta \le 1/\tau$ works and the proof will be more involved. Then we obtain

$$v_i^{p+1} \le |\mathcal{A}_i|^2 e^{\sigma'}/2\big(\exp\left(\eta(H_{i:}^Q v^p + H_{ij}^Q)\right) - 1\big).$$

Denote $c_{\mathrm{TV}} = A_{\max}^2 e^{\sigma'}/2$. For all $i, j \in [n]$ and $p = 0, 1, \ldots$, we have that $v_i^p \le 1$ and that $H_{ij}^Q \le \sum_k H_{ik}^Q \le \sum_k H_{ik}^Q (\mathrm{dist}(i,k)+1)^\mu \le 2^{\mu+1}\nu'$ by Lemma 16.

Then it holds that $\eta(\sum_k H_{ik}^Q v_k^p + H_{ij}^Q) \le 1/2$ which is due to our choice of τ parameter satisfies $\tau = \frac{1}{\eta} \ge \nu' 2^{\mu+3}$, which further implies

$$v_i^{p+1} \le c_{\mathrm{TV}}\left(\exp\left(\eta\left(\sum_k H_{ik}^Q v_k^p + H_{ij}^Q\right)\right) - 1\right) \le 2c_{\mathrm{TV}}\eta\left(\sum_k H_{ik}^Q v_k^p + H_{ij}^Q\right).$$

where the second inequality is due to the fact that for $0 < x < c < 1$, we have $e^x \le 1 + x/(1-c)$. Stacking all v_i^{p+1} together yields

$$v^{p+1} \le 2c_{\mathrm{TV}}\eta(H^Q v^p + H_{:j}^Q),$$

where the absolute values and $\le$ are interpreted element-wise. By Definition 2, we have $Z_{ij}^{\pi^p} = \sup_{s_{-j}} \sup_{s_j, s_j'} v_i^p$. Note that for a given Q function, H^Q is fixed and independent of states or actions.

Since the above inequality holds for any s and s' that only differ in the j-th entry, $\forall j \in [n]$, it immediately implies

$$Z^{\pi^{p+1}} \le 2c_{\mathrm{TV}}\eta(H^Q Z^{\pi^p} + H^Q).$$

By Lemma 16 H^Q is $(2^{\mu+1}\nu', \mu)$-decay. Denote $\nu_H = 2^{\mu+1}\nu'$. We will prove that Z^{π_p} is (ν_p, μ)-decay by induction, where ν_p is to be determined later. First, π_0 is a uniform policy and thus $\nu_0 = 0$ for Z^{π_0}. By Lemma 11, it holds that

$$\nu_{p+1} \le 2c_{\mathrm{TV}}\eta(\nu_H\nu_p + \nu_H) \le (2c_{\mathrm{TV}}\eta\nu_H)^{p+1}\nu_0 + \sum_{k=1}^{p+1}(2c_{\mathrm{TV}}\eta\nu_H)^k \le \frac{2c_{\mathrm{TV}}\eta\nu_H}{1 - 2c_{\mathrm{TV}}\eta\nu_H},$$

which implies that Z^{π^p} is $(\frac{2^{\mu+2}\nu' c_{\mathrm{TV}}}{\tau - 2^{\mu+2}\nu' c_{\mathrm{TV}}}, \mu)$-decay for all p. Further by Lemma 18, let $p \to \infty$, we have that Z^{ζ^*} is (ν'', μ)-decay and thus $\zeta^* \in \Delta_{\nu'',\mu,\sigma'}$, where $\nu'' = \frac{2^{\mu+1}\nu' A_{\max}^2 e^{\frac{\nu'}{n\tau}}}{\tau - 2^{\mu+1}\nu' A_{\max}^2 e^{\frac{\nu'}{n\tau}}}$. This completes the proof of Lemma 10.

□

Based on the proof of Lemma 10, it is easy to see that the intermediate iterates in the iterative algorithm in (28) are also (ν'', μ)-decay policies. We wrote this intermediate result below as a Corollary and we will use it in other parts of the proof.

Corollary 19. *We have that under the same setting of Lemma 10 and under the multiplicative algorithm (28), π^p is $\left(\frac{2^{\mu+1}\nu' A_{\max}^2 e^{\frac{\nu'}{n\tau}}}{\tau - 2^{\mu+1}\nu' A_{\max}^2 e^{\frac{\nu'}{n\tau}}}, \mu\right)$-decay.*

D PROOFS OF SUPPORTING LEMMAS

In this section, we provide the proofs of the lemmas we used in proving the main results.

D.1 Proof of Proposition 1

Since $|\sup_x f(x) - \sup_{x'} g(x')| \le \sup_x |f(x) - g(x)|$, we have for any $V_1, V_2 \in \mathbb{R}^{|\mathcal{S}|}$ and $s \in \mathcal{S}$ that

$$\begin{aligned}
|[\mathcal{T}(V_1)](s) - [\mathcal{T}(V_2)](s)| &\le \gamma \sup_{\zeta \in \Delta_{\mathrm{policy}}} |\mathbb{E}_\zeta \left[V_1(s(1)) - V_2(s(1)) \mid s(0) = s\right]| \\
&\le \gamma \sup_{\zeta \in \Delta_{\mathrm{policy}}} \mathbb{E}_\zeta \left[|V_1(s(1)) - V_2(s(1))| \mid s(0) = s\right] \\
&\le \gamma \|V_1 - V_2\|_\infty.
\end{aligned}$$

Since the previous inequality holds for all $s \in \mathcal{S}$, we have the desired contraction property:

$$\|\mathcal{T}(V_1) - \mathcal{T}(V_2)\|_\infty \le \gamma \|V_1 - V_2\|_\infty.$$

We next show that if a policy ζ^* is such that $V^{\zeta^*} = \mathcal{T}(V^{\zeta^*})$, then ζ^* is an optimal policy. For any policy $\zeta \in \Delta_{\mathrm{policy}}$, we have for all $s \in \mathcal{S}$ that

$$\begin{aligned}
V^{\zeta^*}(s) &= (\mathcal{T}V^{\zeta^*})(s) \\
&\ge \mathbb{E}_\zeta \left[r(s(0), a(0)) - \tau \log \zeta(a(0)|s(0)) + \gamma V^{\zeta^*}(s(1)) \,\middle|\, s(0) = s\right] \\
&= \mathbb{E}_\zeta \left[r(s(0), a(0)) - \tau \log \zeta(a(0)|s(0)) \mid s(0) = s\right] + \gamma \mathbb{E}_\zeta \left[V^{\zeta^*}(s(1)) \,\middle|\, s(0) = s\right] \\
&\ge \mathbb{E}_\zeta \left[r(s(0), a(0)) - \tau \log \zeta(a(0)|s(0)) \mid s(0) = s\right] \\
&\quad + \gamma \mathbb{E}_\zeta \left[r(s(1), a(1)) - \tau \log \zeta(a(1)|s(1)) \mid s(0) = s\right] + \gamma^2 \mathbb{E}_\zeta \left[V^{\zeta^*}(s(2)) \,\middle|\, s(0) = s\right] \\
&\ge \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_\zeta \left[r(s(t), a(t)) - \tau \log \zeta(a(t)|s(t)) \mid s(0) = s\right] \\
&= V^\zeta(s).
\end{aligned}$$

As a result, we have for all $\zeta \in \Delta_{\mathrm{policy}}$ that

$$J(\zeta^*) = \sum_{s \in \mathcal{S}} \rho(s) V^{\zeta^*}(s) \ge \sum_{s \in \mathcal{S}} \rho(s) V^\zeta(s) = J(\zeta),$$

hence ζ^* is an optimal policy.

D.2 Proof of Lemma 7

For each $i \in [n]$, we have

$$
\begin{aligned}
&|V_i^{\zeta}(s) - V_i^{\tilde{\zeta}}(s)| \\
&= \Big|\mathbb{E}_{a\sim\zeta(\cdot|s)}[r_i(s_i,a_i) - n\tau\log\zeta_i(a_i|s) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}V_i^{\zeta}(s')] \\
&\quad - \mathbb{E}_{a\sim\tilde{\zeta}(\cdot|s)}[r_i(s_i,a_i) - n\tau\log\tilde{\zeta}_i(a_i|s) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}V_i^{\tilde{\zeta}}(s')]\Big| \\
&\le \underbrace{\Big|\mathbb{E}_{a\sim\zeta(\cdot|s)}[r_i(s_i,a_i) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}V_i^{\zeta}(s')] - \mathbb{E}_{a\sim\tilde{\zeta}(\cdot|s)}[r_i(s_i,a_i) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}V_i^{\zeta}(s')]\Big|}_{I_1} \\
&\quad + \underbrace{\Big|\mathbb{E}_{a\sim\tilde{\zeta}(\cdot|s)}[r_i(s_i,a_i) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}V_i^{\zeta}(s')] - \mathbb{E}_{a\sim\tilde{\zeta}(\cdot|s)}[r_i(s_i,a_i) + \gamma\mathbb{E}_{s'\sim P(\cdot|s,a)}V_i^{\tilde{\zeta}}(s')]\Big|}_{I_2} \\
&\quad + \underbrace{n\tau|H(\zeta_i(\cdot|s)) - H(\tilde{\zeta}_i(\cdot|s))|}_{I_3} \\
&\le \gamma\|V_i^{\zeta} - V_i^{\tilde{\zeta}}\|_\infty + n\tau\sigma\mathrm{TV}(\tilde{\zeta}_i(\cdot|s),\zeta_i(\cdot|s)) + \sum_{j=1}^{n}\mathrm{TV}(\zeta_j(\cdot|s),\tilde{\zeta}_j(\cdot|s))Z_{ij}^{Q^{\tilde{\zeta}}},
\end{aligned}
$$

where in the last inequality, we have used Lemma 13 (for term I_1), the definition of $Z^{Q^{\tilde{\zeta}}}$ in Definition 4 (for term I_2), and Lemma 12 (for term I_3) respectively. Therefore, we have

$$
\|V_i^{\zeta} - V_i^{\tilde{\zeta}}\|_\infty \le \frac{1}{1-\gamma}\left(n\tau\sigma\sup_{s\in\mathcal{S}}\mathrm{TV}(\tilde{\zeta}_i(\cdot|s),\zeta_i(\cdot|s)) + \sum_{j=1}^{n}\sup_{s\in\mathcal{S}}\mathrm{TV}(\zeta_j(\cdot|s),\tilde{\zeta}_j(\cdot|s))Z_{ij}^{Q^{\tilde{\zeta}}}\right).
$$

Taking the average over i, we have,

$$
\begin{aligned}
\|V^{\zeta} - V^{\tilde{\zeta}}\|_\infty &\le \frac{1}{1-\gamma}\left(\tau\sigma\sum_{i=1}^{n}\sup_{s\in\mathcal{S}}\mathrm{TV}(\tilde{\zeta}_i(\cdot|s),\zeta_i(\cdot|s)) + \frac{1}{n}\sum_{i=1}^{n}\sum_{j=1}^{n}\sup_{s\in\mathcal{S}}\mathrm{TV}(\zeta_j(\cdot|s),\tilde{\zeta}_j(\cdot|s))Z_{ij}^{Q^{\tilde{\zeta}}}\right) \\
&\le \frac{1}{1-\gamma}\left(\tau\sigma + \frac{\nu'}{n}\right)\sum_{i=1}^{n}\sup_{s\in\mathcal{S}}\mathrm{TV}(\tilde{\zeta}_i(\cdot|s),\zeta_i(\cdot|s)),
\end{aligned}
$$

where in the last inequality, we have used $\sum_{i\in\mathcal{N}} Z_{ij}^{Q^{\tilde{\zeta}}} \le \nu'$.

D.3 Proof of Lemma 11

Due to the symmetry in Definition 1, we only show the (ν,μ)-decay property of a matrix for each row sums. The proof for column sums follows the same process.

(a) For any $c, c' \ge 0$, it holds that

$$
\sum_{j=1}^{n}(cA_{ij} + c'A'_{ij})(\mathrm{dist}(i,j)+1)^{\mu} \le c\nu + c'\nu'.
$$

(b) Since $1 \le \mathrm{dist}(i,j)+1 \le (\mathrm{dist}(i,k)+1)(\mathrm{dist}(k,j)+1)$ for all $i,j,k \in [n]$, we have

$$
\sum_{j}[AA']_{ij}(\mathrm{dist}(i,j)+1)^{\mu} = \sum_{j}\sum_{k}a_{ik}a'_{kj}(\mathrm{dist}(i,j)+1)^{\mu}
$$

$$\begin{aligned}
&\leq \sum_j \sum_k a_{ik} a'_{kj} (\mathrm{dist}(i,k)+1)^{\mu} (\mathrm{dist}(k,j)+1)^{\mu} \\
&= \sum_k a_{ik} (\mathrm{dist}(i,k)+1)^{\mu} \sum_j a'_{kj} (\mathrm{dist}(k,j)+1)^{\mu} \\
&\leq \nu\nu'.
\end{aligned}$$

D.4 Proof of Lemma 12

Define $h(t) = H(d + t(d' - d))$ for all $t \in [0,1]$. Then we have

$$\begin{aligned}
|H(d) - H(d')| &= |h(1) - h(0)| \\
&= \left| \int_0^1 h'(t) dt \right| \\
&= \left| \int_0^1 \langle \nabla H(d + t(d' - d)), d' - d \rangle dt \right|.
\end{aligned}$$

For simplicity of notation, denote $d'' = d + t(d' - d)$. Since the set of all σ-regular distributions over $\{1, 2, \cdots, M\}$ is a polytope (hence convex), the distribution d'' as a convex combination between d and d' is also σ-regular. Let $d''_{\max} = \max_m d''_m$ and $d''_{\min} = \min_m d''_m$. Then, we have,

$$\begin{aligned}
|\langle \nabla H(d''), d' - d \rangle| &= \left| \sum_{m=1}^M (1 + \log d''_m)(d_m - d'_m) \right| \\
&= \left| \sum_{m=1}^M \left(- \frac{\log d''_{\max} + \log d''_{\min}}{2} + \log d''_m \right)(d_m - d'_m) \right| \qquad (31) \\
&\leq \sum_{m=1}^M \left| - \frac{\log d''_{\max} + \log d''_{\min}}{2} + \log d''_m \right| \left| d_m - d'_m \right| \\
&\leq \sum_{m=1}^M \left| \frac{\log d''_{\max} - \log d''_{\min}}{2} \right| \left| d_m - d'_m \right| \qquad (32) \\
&\leq \sigma \mathrm{TV}(d, d'),
\end{aligned}$$

where Eq. (31) is due to the fact that $\sum_{m=1}^M c(d_m - d'_m) = 0$ for any constant $c \in \mathbb{R}$, and Eq. (32) is due to the fact that

$$-(\log d''_{\max} - \log d''_{\min})/2 \leq -(\log d''_{\max} + \log d''_{\min})/2 + \log d''_m \leq (\log d''_{\max} - \log d''_{\min})/2.$$

D.5 Proof of Lemma 15

To calculate C^z_{ij}, we consider the following two cases.

Scenario 1: $j \notin \mathcal{N}_i$. The probability of the next state $\bar{s}_i$ given the current state (s_j, s_{-j}) is

$$\sum_{a_i \in \mathcal{A}_i} P_i(\bar{s}_i | s_{\mathcal{N}_i}, a_i) \zeta_i(a_i | s_j, s_{-j}) = P_i^{\zeta}(\bar{s}_i | s_j, s_{-j}).$$

Therefore, when s_j is changed to s'_j, the TV distance is,

$$\begin{aligned}
&\mathrm{TV}(P_i^{\zeta}(\cdot | s_j, s_{-j}), P_i^{\zeta}(\cdot | s'_j, s_{-j})) \\
&= \frac{1}{2} \sum_{\bar{s}_i} \left| \sum_{a_i \in \mathcal{A}_i} P_i(\bar{s}_i | s_{\mathcal{N}_i}, a_i) \zeta_i(a_i | s_j, s_{-j}) - \sum_{a_i \in \mathcal{A}_i} P_i(\bar{s}_i | s_{\mathcal{N}_i}, a_i) \zeta_i(a_i | s'_j, s_{-j}) \right|
\end{aligned}$$

$$
\begin{aligned}
&\leq \frac{1}{2}\sum_{a_i}\left(\sum_{\bar{s}_i} P_i(\bar{s}_i|s_{\mathcal{N}_i}, a_i)\right)\left|\zeta_i(a_i|s_j, s_{-j}) - \zeta_i(a_i|s'_j, s_{-j})\right| \\
&= \mathrm{TV}(\zeta_i(\cdot|s_j, s_{-j}), \zeta_i(\cdot|s'_j, s_{-j})) \\
&\leq Z^{\zeta}_{ij},
\end{aligned}
$$

where in the last inequality we have used the definition of the interaction matrix Z^{ζ}_{ij} of policy ζ in Definition 2.

Scenario 2: $j \in \mathcal{N}_i$: When s_j is changed to s'_j, we have

$$
\begin{aligned}
&\mathrm{TV}(P^{\zeta}_i(\cdot|s_j, s_{-j}), P^{\zeta}_i(\cdot|s'_j, s_{-j})) \\
&= \frac{1}{2}\sum_{\bar{s}_i}\Big|\sum_{a_i\in\mathcal{A}_i} P_i(\bar{s}_i|s_{\mathcal{N}_i/j}, s_j, a_i)\zeta_i(a_i|s_j, s_{-j}) - \sum_{a_i\in\mathcal{A}_i} P_i(\bar{s}_i|s_{\mathcal{N}_i/j}, s'_j, a_i)\zeta_i(a_i|s'_j, s_{-j})\Big| \\
&\leq \frac{1}{2}\sum_{\bar{s}_i}\sum_{a_i\in\mathcal{A}_i}\Big|P_i(\bar{s}_i|s_{\mathcal{N}_i/j}, s_j, a_i)\zeta_i(a_i|s_j, s_{-j}) - P_i(\bar{s}_i|s_{\mathcal{N}_i/j}, s'_j, a_i)\zeta_i(a_i|s'_j, s_{-j})\Big| \\
&\leq \frac{1}{2}\sum_{\bar{s}_i}\sum_{a_i\in\mathcal{A}_i}\Big|P_i(\bar{s}_i|s_{\mathcal{N}_i/j}, s_j, a_i)\zeta_i(a_i|s_j, s_{-j}) - P_i(\bar{s}_i|s_{\mathcal{N}_i/j}, s_j, a_i)\zeta_i(a_i|s'_j, s_{-j})\Big| \\
&\quad + \frac{1}{2}\sum_{\bar{s}_i}\sum_{a_i\in\mathcal{A}_i}\Big|P_i(\bar{s}_i|s_{\mathcal{N}_i/j}, s_j, a_i)\zeta_i(a_i|s'_j, s_{-j}) - P_i(\bar{s}_i|s_{\mathcal{N}_i/j}, s'_j, a_i)\zeta_i(a_i|s'_j, s_{-j})\Big| \\
&\leq \frac{1}{2}\sum_{a_i\in\mathcal{A}_i}\sum_{\bar{s}_i} P_i(\bar{s}_i|s_{\mathcal{N}_i/j}, s_j, a_i)\Big|\zeta_i(a_i|s_j, s_{-j}) - \zeta_i(a_i|s'_j, s_{-j})\Big| \\
&\quad + \frac{1}{2}\sum_{a_i\in\mathcal{A}_i}\zeta_i(a_i|s'_j, s_{-j})\sum_{\bar{s}_i}\Big|P_i(\bar{s}_i|s_{\mathcal{N}_i/j}, s_j, a_i) - P_i(\bar{s}_i|s_{\mathcal{N}_i/j}, s'_j, a_i)\Big| \\
&= \mathrm{TV}(\zeta_i(\cdot|s_j, s_{-j}), \zeta_i(\cdot|s'_j, s_{-j})) + \sum_{a_i\in\mathcal{A}_i}\zeta_i(a_i|s'_j, s_{-j})\mathrm{TV}(P_i(\cdot|s_{\mathcal{N}_i/j}, s_j, a_i), P_i(\cdot|s_{\mathcal{N}_i/j}, s'_j, a_i)) \\
&\leq Z^{\zeta}_{ij} + C_{ij}.
\end{aligned}
$$

D.6 Proof of Lemma 16

First notice the following,

$$
\begin{aligned}
H^{Q}_{ij} \leq \sup_{z_i, z_j, z'_i, z'_j}\sup_{z_{-(i,j)}}\frac{1}{n}\sum_{\ell=1}^{n}\Big|&[Q_\ell(z_i, z_j, z_{-(i,j)}) - Q_\ell(z'_i, z_j, z_{-(i,j)})] \\
&- [Q_\ell(z_i, z'_j, z_{-(i,j)}) - Q_\ell(z'_i, z'_j, z_{-(i,j)})].\Big|
\end{aligned}
$$

Clearly, by the definition of (ν', μ)-decay of Q function, we have

$$
\begin{aligned}
&\Big|[Q_\ell(z_i, z_j, z_{-(i,j)}) - Q_\ell(z'_i, z_j, z_{-(i,j)})] - [Q_\ell(z_i, z'_j, z_{-(i,j)}) - Q_\ell(z'_i, z'_j, z_{-(i,j)})]\Big| \\
&\leq |Q_\ell(z_i, z_j, z_{-(i,j)}) - Q_\ell(z'_i, z_j, z_{-(i,j)})| + |Q_\ell(z_i, z'_j, z_{-(i,j)}) - Q_\ell(z'_i, z'_j, z_{-(i,j)})| \\
&\leq 2Z^{Q}_{\ell i}.
\end{aligned}
$$

Using a symmetric argument, we also have,

$$
\Big|[Q_\ell(z_i, z_j, z_{-(i,j)}) - Q_\ell(z'_i, z_j, z_{-(i,j)})] - [Q_\ell(z_i, z'_j, z_{-(i,j)}) - Q_\ell(z'_i, z'_j, z_{-(i,j)})]\Big|
$$

$$= \left| [Q_\ell(z_i, z_j, z_{-(i,j)}) - Q_\ell(z_i, z'_j, z_{-(i,j)})] - [Q_\ell(z'_i, z_j, z_{-(i,j)}) - Q_\ell(z'_i, z'_j, z_{-(i,j)})] \right|$$
$$\leq 2Z^Q_{\ell j}.$$

As a result,

$$H^Q_{ij} \leq \frac{1}{n} \sum_{\ell=1}^{n} 2 \min(Z^Q_{\ell j}, Z^Q_{\ell i}).$$

As such, we have,

$$\begin{aligned}
&\sum_{j=1}^{n} H^Q_{ij} (\mathrm{dist}(i,j)+1)^\mu \\
&\leq \sum_{j=1}^{n} \frac{1}{n} \sum_{\ell=1}^{n} 2 \min(Z^Q_{\ell j}, Z^Q_{\ell i}) (\mathrm{dist}(i,\ell) + 1 + \mathrm{dist}(j,\ell)+1)^\mu \\
&\leq \sum_{j=1}^{n} \frac{1}{n} \sum_{\ell=1}^{n} 2 \min(Z^Q_{\ell j}, Z^Q_{\ell i}) 2^{\mu-1} \left[(\mathrm{dist}(i,\ell)+1)^\mu + (\mathrm{dist}(j,\ell)+1)^\mu \right] \\
&\leq \sum_{j=1}^{n} \frac{1}{n} \sum_{\ell=1}^{n} 2 Z^Q_{\ell i} 2^{\mu-1} (\mathrm{dist}(i,\ell)+1)^\mu + \sum_{j=1}^{n} \frac{1}{n} \sum_{\ell=1}^{n} 2 Z^Q_{\ell j} 2^{\mu-1} (\mathrm{dist}(j,\ell)+1)^\mu \\
&\leq 2^\mu \nu' + 2^\mu \nu' = 2^{\mu+1} \nu'.
\end{aligned}$$

Similarly, we have,

$$\begin{aligned}
&\sum_{i=1}^{n} H^Q_{ij} (\mathrm{dist}(i,j)+1)^\mu \\
&\leq \sum_{i=1}^{n} \frac{1}{n} \sum_{\ell=1}^{n} 2 \min(Z^Q_{\ell j}, Z^Q_{\ell i}) (\mathrm{dist}(i,\ell) + 1 + \mathrm{dist}(j,\ell)+1)^\mu \\
&\leq \sum_{i=1}^{n} \frac{1}{n} \sum_{\ell=1}^{n} 2 \min(Z^Q_{\ell j}, Z^Q_{\ell i}) 2^{\mu-1} \left[(\mathrm{dist}(i,\ell)+1)^\mu + (\mathrm{dist}(j,\ell)+1)^\mu \right] \\
&\leq \frac{2^\mu}{n} \sum_{\ell=1}^{n} \sum_{i=1}^{n} Z^Q_{\ell i} (\mathrm{dist}(\ell,i)+1)^\mu + \frac{2^\mu}{n} \sum_{i=1}^{n} \sum_{\ell=1}^{n} Z^Q_{\ell j} (\mathrm{dist}(\ell,j)+1)^\mu \\
&\leq 2^\mu \nu' + 2^\mu \nu' = 2^{\mu+1} \nu',
\end{aligned}$$

which completes the proof.

D.7 Proof of Lemma 17

We have, $\frac{d_i/d_j}{d'_i/d'_j} = \exp\left(\log \frac{d_i/d_j}{d'_i/d'_j}\right) \in [e^{-\epsilon}, e^{\epsilon}]$. Then, $|d_i/d_j - d'_i/d'_j| \leq |d'_i/d'_j| |\frac{d_i/d_j}{d'_i/d'_j} - 1| \leq e^c \max(e^\epsilon - 1, 1 - e^{-\epsilon}) = e^c(e^\epsilon - 1)$, since $e^\epsilon - 1 \geq 1 - e^{-\epsilon}$ for any ϵ. Based on these observations, we have

$$|d_i - d'_i| = \left| \frac{d_i}{\sum_j d_j} - \frac{d'_i}{\sum_j d'_j} \right| = \left| \frac{1}{\sum_j d_j/d_i} - \frac{1}{\sum_j d'_j/d'_i} \right| \leq \sum_j |d_j/d_i - d'_j/d'_i| \leq m e^c (e^\epsilon - 1),$$

where we used the fact that $\sum_j d_j/d_i \geq 1$. This concludes the proof.

D.8 Proof of Lemma 18

In this subsection, we prove the convergence of the proposed algorithm in (28). This guarantees the existence of a limiting policy, which is the solution of the entropy regularized optimization problem in (22).

Derivation of the multiplicative weights. We derive the multiplicative weights approach using the mirror descent perspective. For this paragraph, we drop the dependence on s as it is fixed. We denote $\pi = (\pi_1, \pi_2, \ldots, \pi_n) \in \Delta_{\mathcal{A}_1} \times \cdots \times \Delta_{\mathcal{A}_n}$. We consider the following $h(\pi) = \sum_{i=1}^n h_i(\pi_i) := \sum_{i=1}^n \sum_{a_i \in \mathcal{A}_i} \pi_i(a_i) \log \pi_i(a_i)$. Its corresponding Bregman divergence is

$$\begin{aligned} D_h(\pi' || \pi) &= h(\pi') - h(\pi) - \langle \nabla h(\pi), \pi' - \pi \rangle \\ &= \sum_{i=1}^n \left(h_i(\pi_i') - h_i(\pi_i) - \langle \nabla h_i(\pi_i), \pi_i' - \pi_i \rangle \right) \\ &= \sum_{i=1}^n \sum_{a_i \in \mathcal{A}_i} \pi_i'(a_i) \log \frac{\pi_i'(a_i)}{\pi_i(a_i)} \\ &= \sum_{i=1}^n D_{KL}(\pi_i' || \pi_i). \end{aligned}$$

Let the objective function in (22) be $F(\pi)$. Then mirror descent update rule given the Bregman divergence is then,

$$\pi^{p+1} = \underset{\pi \in \Delta_{\mathcal{A}_1} \times \cdots \times \Delta_{\mathcal{A}_n}}{\arg\min} \; (\eta \langle -\nabla F(\pi^p), \pi \rangle + D_h(\pi || \pi^p)).$$

Given the additive structure of the Bregman divergence, we have the above rule can be written component wise as

$$\pi_i^{p+1} = \underset{\pi_i \in \Delta_{\mathcal{A}_i}}{\arg\min} (\eta \langle -\nabla_{\pi_i} F(\pi^p), \pi_i \rangle + D_{KL}(\pi_i || \pi_i^p)).$$

Therefore, π_i^{p+1} will satisfy

$$-\eta [\mathbb{E}_{a_{-i} \sim \pi_{-i}^p} Q(s, a_{-i}, a_i) - \tau(\log \pi_i^p(a_i) + 1)] + 1 + \log \pi_i^{p+1}(a_i) - \log \pi_i^p(a_i) + \lambda = 0,$$

where λ is a Lagrangian multiplier that accounts for the constraint that $\sum_{a_i} \pi_i^{p+1}(a_i) = 1$.

As a result,

$$\pi_i^{p+1}(a_i) \propto \pi_i^p(a_i)^{1-\eta\tau} \exp(\eta \mathbb{E}_{a_{-i} \sim \pi_{-i}^p} Q(s, a_{-i}, a_i)).$$

which recovers the algorithm in (28). Using this interpretation, we prove the convergence of the algorithm (28).

Proof of Lemma 18. Fixing s, we consider two runs of the multiplicative weight algorithm with different initializers π^0 and $\bar{\pi}^0$ respectively. Their respective trajectories are denoted as π^p and $\bar{\pi}^p$. The update rule in (28) suggests that for each ℓ and a_ℓ, it holds that

$$\log \pi_\ell^{p+1}(a_\ell | s) = (1 - \eta\tau) \log \pi_\ell^p(a_\ell | s) + \eta \mathbb{E}_{a_{-\ell} \sim \pi_{-\ell}^p(s)} [Q(s, a_{-\ell}, a_\ell)] + c(s),$$

where $c(s)$ is a normalization constant that only depends on s. Therefore, for any two action pairs $a_\ell, \tilde{a}_\ell$, we have

$$\overbrace{\log \pi_\ell^{p+1}(a_\ell | s) - \log \pi_\ell^{p+1}(\tilde{a}_\ell | s)}^{\xi_\ell^{p+1}}$$

$$= (1-\eta\tau)\underbrace{(\log \pi_\ell^p(a_\ell|s) - \log \pi_\ell^p(\tilde{a}_\ell|s))}_{\xi_\ell^p} + \eta \mathbb{E}_{a_{-\ell}\sim\pi_{-\ell}^p(s)}[Q(s,a_{-\ell},a_\ell) - Q(s,a_{-\ell},\tilde{a}_\ell)],$$

where for notational simplicity, we denote $\log \pi_\ell^{p+1}(a_\ell|s) - \log \pi_\ell^{p+1}(\tilde{a}_\ell|s) = \xi_\ell^{p+1}$ for now (where we have fixed an action pair $a_\ell, \tilde{a}_\ell$). Similar to (30), we have $|\xi_\ell^p| \le \nu'/\tau$ for all p, ℓ.

Do the same for the other trajectory starting with a different initialization policy and define $\bar{\xi}_\ell^{p+1}$ similarly. We have

$$\begin{aligned}
&\xi_\ell^{p+1} - \bar{\xi}_\ell^{p+1} \\
&= (1-\eta\tau)(\xi_\ell^p - \bar{\xi}_\ell^p) \\
&\quad + \eta \mathbb{E}_{a_{-\ell}\sim\pi_{-\ell}^p(s)}[Q(s,a_{-\ell},a_\ell) - Q(s,a_{-\ell},\tilde{a}_\ell)] - \eta \mathbb{E}_{a_{-\ell}\sim\bar{\pi}_{-\ell}^p(s)}[Q(s,a_{-\ell},a_\ell) - Q(s,a_{-\ell},\tilde{a}_\ell)].
\end{aligned}$$

Using Lemma 13, we have that,

$$|\xi_\ell^{p+1} - \bar{\xi}_\ell^{p+1}| \le (1-\eta\tau)|\xi_\ell^p - \bar{\xi}_\ell^p| + \eta \sum_{j\neq\ell} \mathrm{TV}(\pi_j^p(s), \bar{\pi}_j^p(s)) H_{\ell j}^Q.$$

Denote $H_{\ell:}^Q$ as the ℓ-th row of H^Q. Let vector $v^p \in \mathbb{R}^n$ be defined as $v_\ell^p = \mathrm{TV}(\pi_\ell^p(s), \bar{\pi}_\ell^p(s))$ $(\ell = 1, 2, \ldots, n)$. Then the above update can be written as

$$\begin{aligned}
|\xi_\ell^{p+1} - \bar{\xi}_\ell^{p+1}| &\le (1-\eta\tau)|\xi_\ell^p - \bar{\xi}_\ell^p| + \eta H_{\ell:}^Q v^p \\
&\le (1-\eta\tau)^{p+1}|\xi_\ell^0 - \bar{\xi}_\ell^0| + \sum_{k=0}^{p} \eta(1-\eta\tau)^{p-k} H_{\ell:}^Q v^k.
\end{aligned}$$

Now we use $\eta = 1/\tau$ and obtain

$$|\xi_\ell^{p+1} - \bar{\xi}_\ell^{p+1}| \le \frac{1}{\tau} H_{\ell:}^Q v^p.$$

By the definition of ξ_ℓ^{p+1} and the regularity of π_ℓ^{p+1} it holds that

$$|\xi_\ell^{p+1}| = |\log \pi_\ell^{p+1}(a_\ell|s) - \log \pi_\ell^{p+1}(\tilde{a}_\ell|s)| \le \nu'/\tau. \tag{33}$$

By Lemma 17 we have

$$v_\ell^{p+1} \le 1/2 A_{\max}^2 e^{\nu'/n\tau}\big(\exp\big(\frac{1}{\tau} H_{\ell:}^Q v^p\big) - 1\big) = c_{\mathrm{TV}}\big(\exp(\eta H_{\ell:}^Q v^p) - 1\big),$$

where $c_{\mathrm{TV}} = A_{\max}^2 e^{\nu'/n\tau}/2$. Similar to the previous proof, we have $H_{\ell:}^Q v^p \le 2^{\mu+1}\nu'$ due to Lemma 16 and the fact that $v_\ell^p \le 1$ for all ℓ, p. Therefore, when $\tau \ge \nu' 2^{\mu+2}$, we have

$$v^{p+1} \le 2c_{\mathrm{TV}}(\eta H^Q v^p) \le (2\frac{1}{\tau} c_{\mathrm{TV}} H^Q)^{p+1} v^0, \tag{34}$$

where in the first inequality we stack all v_ℓ^{p+1} together. When taking the infinity norm on both sides, we have

$$\|v^{p+1}\|_\infty \le (2\frac{1}{\tau} c_{\mathrm{TV}})^{p+1} \|H^Q\|_\infty^{p+1} \|v^0\|_\infty \le (2\frac{1}{\tau} c_{\mathrm{TV}} 2^{\mu+1}\nu')^{p+1} = (\frac{1}{\tau} 2^{\mu+1}\nu' A_{\max}^2 e^{\nu'/n\tau})^{p+1} \tag{35}$$

Therefore, when $\tau > 2^{\mu+1}\nu' A_{\max}^2 e^{\nu'/n\tau}$, the right hand side of the above inequality converges to zero when $p \to \infty$.

The above argument shows the optimization problem (27) has at most one stationary point in the manifold Δ_{policy} because otherwise, the algorithm (28) starting from two different stationary points will not converge together. Also, the global maximizer of the optimization problem (27), $(\zeta_1(\cdot|s), \ldots, \zeta_n(\cdot|s))$, must be one stationary point, and this stationary point must be unique.

As a result, for the two trajectories considered in (35), if we select one to be the trajectory that always sits at the unique stationary point, and the other is from an arbitrary starting point, we have convergence the algorithm (28) will converge in the following sense:

$$\sup_{i\in\mathcal{N}} \mathrm{TV}(\pi_i^p(\cdot|s), \zeta_i(\cdot|s)) \le \left(\frac{1}{\tau} 2^{\mu+1}\nu' A_{\max}^2 e^{\nu'/n\tau}\right)^p. \tag{36}$$

This completes the proof.

□

E PROOF OF LEMMA 8: ANALYSIS OF POLICY IMPROVEMENT ERROR

Since Lemma 8 only concerns a single outer loop step m, for notational convenience we drop the dependence on m and restate a slightly different version of the lemma as follows.

Lemma 20. *Suppose Q-functions $\{Q_i\}_{i=1}^n$ are (ν', μ)-decay. Further, suppose its truncated estimates $\{\hat{Q}_i\}_{i=1}^n$ satisfies*

$$\sup_{s\in\mathcal{S}, a\in\mathcal{A}} |Q_i(s,a) - \hat{Q}_i(s_{\mathcal{N}_i^\beta}, a_{\mathcal{N}_i^\beta})| \le \epsilon = \frac{c_{pe}}{(\beta+1)^\mu}.$$

Consider the following updates:

$$\hat{\pi}_i^{p+1}(a_i|s_{\mathcal{N}_i^\kappa}) \propto \hat{\pi}_i^p(a_i|s_{\mathcal{N}_i^\kappa})^{1-\eta\tau} \exp\left(\eta \mathbb{E}_{a_j\sim\hat{\pi}_j^p([\overline{s_{\mathcal{N}_i^\kappa}}]_{\mathcal{N}_j^\kappa}), j\in\mathcal{N}_i^\kappa/\{i\}}\left[\frac{1}{n}\sum_{j\in\mathcal{N}_i^\kappa} \hat{Q}_j([\overline{s_{\mathcal{N}_i^\kappa}}]_{\mathcal{N}_j^\beta}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{\mathcal{N}_j^\beta\setminus i}, a_i)\right]\right). \tag{37}$$

Let $\zeta = (\zeta_1, \ldots, \zeta_n)$ be the result of the policy improvement w.r.t. $Q = 1/n\sum_i Q_i$, i.e.

$$(\zeta_1(\cdot|s), \ldots, \zeta_n(\cdot|s)) \leftarrow \underset{\pi_1(\cdot|s),\ldots,\pi_n(\cdot|s)}{\arg\max} \mathbb{E}_{a_i\sim\pi_i(\cdot|s)}[Q(s,a_1,\ldots,a_n)] + \tau\sum_{i=1}^n H(\pi_i(\cdot|s)). \tag{38}$$

Let $\tilde{\sigma}' = \frac{1}{\tau}(\frac{2f(\kappa)c_{pe}}{n(\beta+1)^\mu} + \frac{\nu'}{n})$. Assume $\tau \ge \max(6A_{\max}^2 e^{\tilde{\sigma}'} 2^{\mu+1}\nu', 2(2^{\mu+4}\nu' + 2\frac{c_{pe}}{(\beta+1)^\mu}), 4(2^{\mu+1}\nu' + f(\kappa)(\kappa+1)^\mu \frac{c_{pe}}{(\beta+1)^\mu}))$ and $\eta = \frac{1}{\tau}$. Then we have

(a) $\forall p$, $\hat{\pi}_i^p$ is $\tilde{\sigma}'$-regular.

(b) $\forall p$, $\hat{\pi}_i^p$ is $(\tilde{\nu}, \mu)$-decay where $\tilde{\nu} = \frac{A_{\max}^2 e^{\tilde{\sigma}'}\tilde{\nu}_H}{\tau - A_{\max}^2 e^{\tilde{\sigma}'}\tilde{\nu}_H}$ and $\tilde{\nu}_H = 2^{\mu+1}\nu' + f(\kappa)(\kappa+1)^\mu \frac{c_{pe}}{(\beta+1)^\mu}$.

(c) When $p \ge -\log_2 \frac{4+\frac{c_{pe}}{2^\mu\nu'}}{3(\kappa/2+1)^\mu}$, we have

$$\sup_{s\in\mathcal{S}} \mathrm{TV}(\hat{\pi}_i^p(\cdot|s_{\mathcal{N}_i^\kappa}), \zeta_i(\cdot|s)) \le \frac{4+\frac{c_{pe}}{2^\mu\nu'}}{(\kappa/2+1)^\mu},$$

and

$$\mathcal{T}V - V^{\hat{\pi}^p} \le \frac{1}{1-\gamma}\left(\frac{2f(\kappa)c_{pe}}{(\beta+1)^\mu} + 2\nu'\right)\frac{4+\frac{c_{pe}}{2^\mu\nu'}}{(\kappa/2+1)^\mu}\mathbf{1}. \tag{39}$$

Based on Lemma 20, we now prove Lemma 8 as follows. Recall the parameter settings in Theorem 2, which implies

$$\tau \ge 40\times 2^\mu \bar{r}\frac{4-3\gamma}{4-5\gamma}A_{\max}^2 e \ge 10A_{\max}^2 e^{\frac{2\nu'}{\tau n}} 2^{\mu+2}\nu' = 10A_{\max}^2 e^{\tilde{\sigma}} 2^{\mu+2}\nu'.$$

Further, the parameter $\beta = \frac{\kappa+1}{2}(\frac{2f(\kappa)c_{pe}}{\nu'})^{\frac{1}{\mu}}$ ensures that $f(\kappa)(\kappa+1)^\mu \frac{c_{pe}}{(\beta+1)^\mu} \le 2^\mu\nu'$ and $\tilde{\sigma}' \le \frac{2\nu'}{n\tau} = \tilde{\sigma}$. As a result, one can easily verify that the lower bound assumption on τ in Lemma 20 holds.

Therefore, by setting the Q-functions in Lemma 20 as $\{Q_i^{\hat{\zeta}^m}\}_{i\in\mathcal{N}}$, we obtain that the returned policy $\hat{\zeta}^{m+1}$ is $\tilde{\sigma}$ regular and $(\tilde{\nu},\mu)$-decay by part (a) and part (b) of Lemma 20. Further, we can check that $\tilde{\nu}_H \le 2^{\mu+2}\nu'$ and $\tilde{\nu} \le \frac{1}{9}$. Therefore, $\hat{\zeta}^{m+1} \in \Delta_{\frac{1}{9},\mu,\tilde{\sigma}}$ and we can apply Lemma 9 (more specifically, Equation (26)) and get that the local Q-functions $\{Q_i^{\hat{\zeta}^{m+1}}\}_{i\in\mathcal{N}}$ of the output policy $\hat{\zeta}^{m+1}$ satisfies (ν'_{next},μ)-decay, where $\nu'_{\text{next}} = \bar{r} + \frac{\gamma(\bar{r}+n\tau\tilde{\sigma})}{8(1-\gamma)}$, and we can easily check $\nu'_{\text{next}} \le \nu'$. As a result, the local Q-functions $\{Q_i^{\hat{\zeta}^{m+1}}\}_{i\in\mathcal{N}}$ satisfies (ν',μ) decay property.

Also, by part (c) of Lemma 20, we have

$$\mathcal{T}V - V^{\hat{\zeta}^{m+1}} \le \frac{3\nu'}{1-\gamma}\frac{4+\frac{c_{pe}}{2^\mu\nu'}}{(\kappa/2+1)^\mu}\mathbf{1} = \frac{3(4-3\gamma)\bar{r}}{(1-\gamma)(4-5\gamma)}\frac{4+\frac{c_{pe}(4-5\gamma)}{2^\mu(4-3\gamma)\bar{r}}}{(\kappa/2+1)^\mu}\mathbf{1} := \frac{c_{pe}}{(\kappa/2+1)^\mu}\mathbf{1}.$$

This concludes the proof of Lemma 8. We will next prove the three parts of Lemma 20 in the following three subsections.

E.1 Proof of Part (a)

Recall that in (37), $\hat{\pi}_i^p$ is defined and updated only on the local state $s_{\mathcal{N}_i^\kappa}$. For each i, we define $\tilde{\pi}_i^p(\cdot|s) = \tilde{\pi}_i^p(\cdot|s_{\mathcal{N}_i^\kappa}, s_{\mathcal{N}_{-i}^\kappa}) := \hat{\pi}_i^p(\cdot|s_{\mathcal{N}_i^\kappa})$ which is an extended version of $\hat{\pi}_i^p$ with nominal (but not actual) dependence on $s_{\mathcal{N}_{-i}^\kappa}$. Then, $\tilde{\pi}_i^p(\cdot|s)$ obeys,

$$\tilde{\pi}_i^{p+1}(a_i|s) \propto \tilde{\pi}_i^p(a_i|s)^{1-\eta\tau}\exp\left(\eta\mathbb{E}_{a_j\sim\tilde{\pi}_j^p(\cdot|\overline{s_{\mathcal{N}_i^\kappa}}),\forall j\neq i}\left[\frac{1}{n}\sum_{j\in\mathcal{N}_i^\kappa}\hat{Q}_j([\overline{s_{\mathcal{N}_i^\kappa}}]_{\mathcal{N}_j^\beta},[\overline{a_{\mathcal{N}_i^\kappa}}]_{\mathcal{N}_j^\beta\setminus i},a_i)\right]\right), \tag{40}$$

which is equivalent to (37). We also use the following abbreviation notation $\tilde{Q}^i(s,a_{-i},a_i) = \frac{1}{n}\sum_{j\in\mathcal{N}_i^\kappa}\hat{Q}_j([\overline{s_{\mathcal{N}_i^\kappa}}]_{\mathcal{N}_j^\beta},[\overline{a_{\mathcal{N}_i^\kappa}}]_{\mathcal{N}_j^\beta\setminus i},a_i)$ which again has nominal (but not actual) dependence on $(s_{\mathcal{N}_{-i}^\kappa}, a_{\mathcal{N}_{-i}^\kappa})$. Then the update rule (37) can be further written as for all i and $s\in\mathcal{S}$

$$\tilde{\pi}_i^{p+1}(a_i|s) \propto \tilde{\pi}_i^p(a_i|s)^{1-\eta\tau}\exp\left(\eta\mathbb{E}_{a_j\sim\tilde{\pi}_j^p(\cdot|\overline{s_{\mathcal{N}_i^\kappa}}),j\neq i}[\tilde{Q}^i(s,a_{-i},a_i)]\right). \tag{41}$$

Now consider a given agent i and fix two actions a_i, a_i'. We define $\tilde{\xi}_i^{p+1}(s) = \log\tilde{\pi}_i^{p+1}(a_i|s) - \log\tilde{\pi}_i^{p+1}(a_i'|s)$ which according to (41), follows the following update,

$$\tilde{\xi}_i^{p+1}(s) = (1-\eta\tau)\tilde{\xi}_i^p(s) + \eta\mathbb{E}_{a_j\sim\tilde{\pi}_j^p(\cdot|\overline{s_{\mathcal{N}_i^\kappa}}),j\neq i}[\tilde{Q}^i(s,a_{-i},a_i) - \tilde{Q}^i(s,a_{-i},a_i')]. \tag{42}$$

We now give an upper bound for $|\tilde{Q}^i(s,a_{-i},a_i) - \tilde{Q}^i(s,a_{-i},a_i')|$ as follows:

$$\begin{aligned}
|\tilde{Q}^i(s,a_{-i},a_i) - \tilde{Q}^i(s,a_{-i},a_i')| &\le \frac{1}{n}\sum_{j\in\mathcal{N}_i^\kappa}|\hat{Q}_j([\overline{s_{\mathcal{N}_i^\kappa}}]_{\mathcal{N}_j^\beta},[\overline{a_{\mathcal{N}_i^\kappa}}]_{\mathcal{N}_j^\beta\setminus i},a_i) - \hat{Q}_j([\overline{s_{\mathcal{N}_i^\kappa}}]_{\mathcal{N}_j^\beta},[\overline{a_{\mathcal{N}_i^\kappa}}]_{\mathcal{N}_j^\beta\setminus i},a_i')| \\
&= \frac{1}{n}\sum_{j\in\mathcal{N}_i^\kappa}\Big(\Big|\hat{Q}_j([\overline{s_{\mathcal{N}_i^\kappa}}]_{\mathcal{N}_j^\beta},[\overline{a_{\mathcal{N}_i^\kappa}}]_{\mathcal{N}_j^\beta\setminus i},a_i) - Q_j(\overline{s_{\mathcal{N}_i^\kappa}},[\overline{a_{\mathcal{N}_i^\kappa}}]_{-i},a_i)\Big| \\
&\quad + \Big|Q_j(\overline{s_{\mathcal{N}_i^\kappa}},[\overline{a_{\mathcal{N}_i^\kappa}}]_{-i},a_i') - \hat{Q}_j([\overline{s_{\mathcal{N}_i^\kappa}}]_{\mathcal{N}_j^\beta},[\overline{a_{\mathcal{N}_i^\kappa}}]_{\mathcal{N}_j^\beta\setminus i},a_i')\Big| \\
&\quad + \Big|Q_j(\overline{s_{\mathcal{N}_i^\kappa}},[\overline{a_{\mathcal{N}_i^\kappa}}]_{-i},a_i) - Q_j(\overline{s_{\mathcal{N}_i^\kappa}},[\overline{a_{\mathcal{N}_i^\kappa}}]_{-i},a_i')\Big|\Big) \\
&\le \frac{1}{n}\sum_{j\in\mathcal{N}_i^\kappa}(2\epsilon + Z_{ji}^Q) \\
&\le 2\frac{|\mathcal{N}_i^\kappa|}{n}\epsilon + \frac{\nu'}{n},
\end{aligned} \tag{43}$$

where the second inequality is due to Definition 4 and (20). As a result, we have

$$|\tilde{\xi}_i^{p+1}(s)| = |(1-\eta\tau)\tilde{\xi}_i^p(s) + \eta\mathbb{E}_{a_j\sim\tilde{\pi}_j^p(\cdot|\overline{\mathcal{N}_i^\kappa}),j\neq i}[\tilde{Q}^i(s,a_{-i},a_i) - \tilde{Q}^i(s,a_{-i},\tilde{a}_i)]|$$
$$\leq (1-\eta\tau)|\tilde{\xi}_i^p(s)| + \eta\left(2\frac{|\mathcal{N}_i^\kappa|}{n}\epsilon + \frac{\nu'}{n}\right).$$

Considering the fact that we start from a uniform policy, we have $\forall p$, $|\tilde{\xi}_i^p(s)| \leq \frac{1}{\tau}(2\frac{|\mathcal{N}_i^\kappa|}{n}\epsilon + \frac{\nu'}{n}) \leq \tilde{\sigma}' = \frac{1}{\tau}(\frac{2f(\kappa)c_{pe}}{n(\beta+1)^\mu} + \frac{\nu'}{n})$, where we have used $\epsilon = \frac{c_{pe}}{(\beta+1)^\mu}$.

E.2 Proof of Part (b)

The proof is a similar but slightly more complicated version of the proof of Lemma 10. Similar to the steps in the proof of Lemma 10, we fix $i, j \in \mathcal{N}$, and let $s = (s_j, s_{-j})$ and $s' = (s'_j, s_{-j})$. Then, we calculate,

$$\begin{aligned}
&\tilde{\xi}_i^{p+1}(s) - \tilde{\xi}_i^{p+1}(s') \\
&= (1-\eta\tau)(\tilde{\xi}_i^p(s) - \tilde{\xi}_i^p(s')) + \eta\mathbb{E}_{a_\ell\sim\tilde{\pi}_\ell^p(\cdot|\overline{s_{\mathcal{N}_i^\kappa}}),\ell\neq i}[\tilde{Q}^i(s,a_{-i},a_i) - \tilde{Q}^i(s,a_{-i},\tilde{a}_i)] \\
&\quad - \eta\mathbb{E}_{a_\ell\sim\tilde{\pi}_\ell^p(\cdot|\overline{s'_{\mathcal{N}_i^\kappa}}),\ell\neq i}[\tilde{Q}^i(s',a_{-i},a_i) - \tilde{Q}^i(s',a_{-i},\tilde{a}_i)] \\
&= (1-\eta\tau)(\tilde{\xi}_i^p(s) - \tilde{\xi}_i^p(s')) \\
&+ \eta\mathbb{E}_{a_\ell\sim\tilde{\pi}_\ell^p(\cdot|\overline{s_{\mathcal{N}_i^\kappa}}),\ell\neq i}[\tilde{Q}^i(s,a_{-i},a_i) - \tilde{Q}^i(s,a_{-i},\tilde{a}_i)] - \eta\mathbb{E}_{a_\ell\sim\tilde{\pi}_\ell^p(\cdot|\overline{s'_{\mathcal{N}_i^\kappa}}),\ell\neq i}[\tilde{Q}^i(s,a_{-i},a_i) - \tilde{Q}^i(s,a_{-i},\tilde{a}_i)] \\
&+ \eta\mathbb{E}_{a_\ell\sim\tilde{\pi}_\ell^p(\cdot|\overline{s'_{\mathcal{N}_i^\kappa}}),\ell\neq i}[\tilde{Q}^i(s,a_{-i},a_i) - \tilde{Q}^i(s,a_{-i},\tilde{a}_i)] - \eta\mathbb{E}_{a_\ell\sim\tilde{\pi}_\ell^p(\cdot|\overline{s'_{\mathcal{N}_i^\kappa}}),\ell\neq i}[\tilde{Q}^i(s',a_{-i},a_i) - \tilde{Q}^i(s',a_{-i},\tilde{a}_i)] \\
&\leq (1-\eta\tau)(\tilde{\xi}_i^p(s) - \tilde{\xi}_i^p(s')) + \eta\sum_{\ell\neq i}\mathrm{TV}(\pi_\ell^p(\cdot|\overline{s_{\mathcal{N}_i^\kappa}}), \pi_\ell^p(\cdot|\overline{s'_{\mathcal{N}_i^\kappa}}))H_{i\ell}^{\tilde{Q}^i} + \eta H_{ij}^{\tilde{Q}^i}.
\end{aligned}$$

Recall the notation $Z_{\ell j}^{\tilde{\pi}^p} = \sup_{s_j,s'_j,s_{-j}}\mathrm{TV}(\tilde{\pi}_\ell^p(\cdot|s_j,s_{-j}), \tilde{\pi}_\ell^p(\cdot|s'_j,s_{-j}))$. We then have that,

$$\begin{aligned}
\tilde{\xi}_i^{p+1}(s) - \tilde{\xi}_i^{p+1}(s') &\leq (1-\eta\tau)(\tilde{\xi}_i^p(s) - \tilde{\xi}_i^p(s')) + \eta\sum_{\ell\neq i}Z_{\ell j}^{\tilde{\pi}^p}H_{i\ell}^{\tilde{Q}^i} + \eta H_{ij}^{\tilde{Q}^i} \\
&\leq (1-\eta\tau)(\tilde{\xi}_i^p(s) - \tilde{\xi}_i^p(s')) + \eta H_{i:}^{\tilde{Q}^i}Z_{:j}^{\tilde{\pi}^p} + \eta H_{ij}^{\tilde{Q}^i} \\
&\leq (1-\eta\tau)^{p+1}(\tilde{\xi}_i^0(s) - \tilde{\xi}_i^0(s')) + \sum_{k=0}^{p}\eta(1-\eta\tau)^{p-k}(H_{i:}^{\tilde{Q}^i}Z_{:j}^{\tilde{\pi}^k} + H_{ij}^{\tilde{Q}^i}).
\end{aligned}$$

Recalling $\eta = \frac{1}{\tau}$, we have,

$$\tilde{\xi}_i^{p+1}(s) - \tilde{\xi}_i^{p+1}(s') \leq \frac{1}{\tau}(H_{i:}^{\tilde{Q}^i}Z_{:j}^{\tilde{\pi}^p} + H_{ij}^{\tilde{Q}^i}). \tag{44}$$

Since $\tilde{\pi}_i^p$ is $\tilde{\sigma}'$-regular, we can use Lemma 17 to obtain

$$\mathrm{TV}(\tilde{\pi}_i^{p+1}(\cdot|s), \tilde{\pi}_i^{p+1}(\cdot|s')) \leq \frac{A_{\max}^2 e^{\tilde{\sigma}'}}{2}\left(\exp\left(\frac{1}{\tau}(H_{i:}^{\tilde{Q}^i}Z_{:j}^{\tilde{\pi}^p} + H_{ij}^{\tilde{Q}^i})\right) - 1\right). \tag{45}$$

We have the following Lemma regarding $H^{\tilde{Q}^i}$.

Lemma 21. *$H^{\tilde{Q}^i}$ satisfies,*

$$\sum_{\ell\in\mathcal{N}}H_{i\ell}^{\tilde{Q}^i}(\mathrm{dist}(i,\ell)+1)^\mu \leq 2^{\mu+1}\nu' + f(\kappa)(\kappa+1)^\mu\frac{c_{pe}}{(\beta+1)^\mu} := \tilde{\nu}_H.$$

Therefore, $\frac{1}{\tau}(H^{\tilde{Q}^i}_{i:} Z^{\tilde{\pi}^p}_{:j} + H^{\tilde{Q}^i}_{ij}) \le \frac{1}{\tau}(2\tilde{v}_H) \le \frac{1}{2}$, which can be satisfied when $\tau \ge 4\tilde{v}_H$. As a result, we can simplify the above equation to

$$\mathrm{TV}(\tilde{\pi}_i^{p+1}(\cdot|s), \tilde{\pi}_i^{p+1}(\cdot|s')) \le A_{\max}^2 e^{\tilde{\sigma}'} \frac{1}{\tau}(H^{\tilde{Q}^i}_{i:} Z^{\tilde{\pi}^p}_{:j} + H^{\tilde{Q}^i}_{ij}).$$

In the above equation, taking the sup over s_{-j}, s_j, s'_j, the left hand side would become $Z^{\tilde{\pi}^{p+1}}_{ij}$. As a result, if we repeat the above equation for all i, j pairs, we get,

$$Z^{\tilde{\pi}^{p+1}} \le |A_{\max}|^2 e^{\tilde{\sigma}'} \frac{1}{\tau}(\tilde{H} Z^{\tilde{\pi}^p} + \tilde{H}), \tag{46}$$

where $\tilde{H}$ is a matrix whose ith row is $H^{\tilde{Q}^i}_{i,:}$. Suppose $Z^{\tilde{\pi}^p}$ is $(\tilde{v}_p, \mu)$-decay. Clearly we have $\tilde{v}_p = 0$, and by (46), we have that

$$\tilde{v}_{p+1} \le A_{\max}^2 e^{\tilde{\sigma}'} \frac{1}{\tau}(\tilde{v}_H \tilde{v}_p + \tilde{v}_H) \le \frac{A_{\max}^2 e^{\tilde{\sigma}'} \tilde{v}_H}{\tau - A_{\max}^2 e^{\tilde{\sigma}'} \tilde{v}_H}. \tag{47}$$

Now it remains to prove Lemma 21 to finish the proof of Part (b) of Lemma 20.

Proof of Lemma 21. we give an upper bound for the term $|[\tilde{Q}^i(z_j, z_k, z_{-(j,k)}) - \tilde{Q}^i(z'_j, z_k, z_{-(j,k)})] - [\tilde{Q}^i(z_j, z'_k, z_{-(j,k)}) - \tilde{Q}^i(z'_j, z'_k, z_{-(j,k)})]|$. For $j, k \in \mathcal{N}_i^\kappa$, consider $\tilde{Q}^i(z_j, z_k, z_{-(j,k)}) - \tilde{Q}^i(z'_j, z_k, z_{-(j,k)})$, we have that

$$\begin{aligned}
&\tilde{Q}^i(z_j, z_k, z_{-(j,k)}) - \tilde{Q}^i(z'_j, z_k, z_{-(j,k)}) - [\tilde{Q}^i(z_j, z'_k, z_{-(j,k)}) - \tilde{Q}^i(z'_j, z'_k, z_{-(j,k)})] \\
&= \frac{1}{n} \sum_{\ell \in \mathcal{N}_i^\kappa \cap \mathcal{N}_j^\beta \cap \mathcal{N}_k^\beta} \Big[\hat{Q}_\ell(z_j, z_k, [\overline{z_{\mathcal{N}_i^\kappa}}]_{\mathcal{N}_\ell^\beta \setminus (j,k)}) - \hat{Q}_\ell(z'_j, z_k, [\overline{z_{\mathcal{N}_i^\kappa}}]_{\mathcal{N}_\ell^\beta \setminus (j,k)}) \\
&\quad - \hat{Q}_\ell(z_j, z'_k, [\overline{z_{\mathcal{N}_i^\kappa}}]_{\mathcal{N}_\ell^\beta \setminus (j,k)}) + \hat{Q}_\ell(z'_j, z'_k, [\overline{z_{\mathcal{N}_i^\kappa}}]_{\mathcal{N}_\ell^\beta \setminus (j,k)}) \Big].
\end{aligned}$$

Notice that

$$|\hat{Q}_\ell(z_j, z_k, [\overline{z_{\mathcal{N}_i^\kappa}}]_{\mathcal{N}_\ell^\beta \setminus (j,k)}) - Q_\ell(z_j, z_k, [\overline{z_{\mathcal{N}_i^\kappa}}]_{-(j,k)})| \le \epsilon. \tag{48}$$

We can use triangle inequality to obtain

$$\begin{aligned}
&\left|\left[\tilde{Q}^i(z_j, z_k, z_{-(j,k)}) - \tilde{Q}^i(z'_j, z_k, z_{-(j,k)})\right] - \left[\tilde{Q}^i(z_j, z'_k, z_{-(j,k)}) - \tilde{Q}^i(z'_j, z'_k, z_{-(j,k)})\right]\right| \\
&\le 4\frac{|\mathcal{N}_i^\kappa|}{n}\epsilon + \frac{1}{n} \sum_{\ell \in \mathcal{N}_i^\kappa \cap \mathcal{N}_j^\beta \cap \mathcal{N}_k^\beta} \Big| Q_\ell(z_j, z_k, [\overline{z_{\mathcal{N}_i^\kappa}}]_{-(j,k)}) - Q_\ell(z'_j, z_k, [\overline{z_{\mathcal{N}_i^\kappa}}]_{-(j,k)}) \\
&\quad - Q_\ell(z_j, z'_k, [\overline{z_{\mathcal{N}_i^\kappa}}]_{-(j,k)}) + Q_\ell(z'_j, z'_k, [\overline{z_{\mathcal{N}_i^\kappa}}]_{-(j,k)}) \Big|
\end{aligned}$$

Thus, we have

$$H^{\tilde{Q}^i}_{jk} \le H^Q_{jk} + 4\frac{|\mathcal{N}_i^\kappa|}{n}\epsilon. \tag{49}$$

For the decay property, note that $\forall j \notin \mathcal{N}_i^\kappa, H^{\tilde{Q}^i}_{ij} = 0$, we have

$$\begin{aligned}
\sum_{j \in \mathcal{N}} H^{\tilde{Q}^i}_{ij} (\mathrm{dist}(i,j)+1)^\mu &\le \sum_{j \in \mathcal{N}_i^\kappa} (H^Q_{ij} + 4\frac{|\mathcal{N}_i^\kappa|}{n}\epsilon)(\mathrm{dist}(i,j)+1)^\mu \\
&\le 2^{\mu+1}\nu' + |\mathcal{N}_i^\kappa|\epsilon(\kappa+1)^\mu,
\end{aligned} \tag{50}$$

which completes the proof. □

E.3 Proof of Part (c)

Recall that in Lemma 18, we showed that the following update will converge to ζ, the policy resulting from the exact policy improvement w.r.t. Q (with appropriate choices of η and τ).

$$\pi_i^{p+1}(a_i|s) \propto \pi_i^p(a_i|s)^{1-\eta\tau} \exp\left(\eta \mathbb{E}_{a_j \sim \pi_j^p(\cdot|s), j\neq i}[Q(s, a_{-i}, a_i)]\right), \forall i \in \mathcal{N}, s \in \mathcal{S}. \tag{51}$$

For two arbitrary actions a_i, a_i' we define $\xi_i^{p+1}(s) = \log \pi_i^{p+1}(a_i|s) - \log \pi_i^{p+1}(a_i'|s)$. The central step of our proof is to track the difference between our algorithm (41) and the algorithm (51). To this end, we will provide a recursive bound on

$$v_i^p = \sup_{s\in\mathcal{S}} \mathrm{TV}(\pi_i^p(\cdot|s), \tilde{\pi}_i^p(\cdot|s)). \tag{52}$$

Recall (42) as follows.

$$\tilde{\xi}_i^{p+1}(s) = (1-\eta\tau)\tilde{\xi}_i^p(s) + \eta \mathbb{E}_{a_j\sim\tilde{\pi}_j^p(\cdot|\overline{s_{\mathcal{N}_i^\kappa}}), j\neq i}[\tilde{Q}^i(s, a_{-i}, a_i) - \tilde{Q}^i(s, a_{-i}, a_i')].$$

Now let's track the difference of our algorithm and the prototype algorithm. For all i, and for all s, we have

$$\begin{aligned}
\xi_i^{p+1}(s) - \tilde{\xi}_i^{p+1}(s) &= (1-\eta\tau)(\xi_i^p(s) - \tilde{\xi}_i^p(s)) + \eta \mathbb{E}_{a_{-i}\sim\pi_{-i}^p(\cdot|s)}[Q(s, a_{-i}, a_i) - Q(s, a_{-i}, a_i')] \\
&\quad - \eta \mathbb{E}_{a_j\sim\tilde{\pi}_j^p(\cdot|\overline{s_{\mathcal{N}_i^\kappa}}), j\neq i}[\tilde{Q}^i(s, a_{-i}, a_i) - \tilde{Q}^i(s, a_{-i}, a_i')] \\
&= (1-\eta\tau)(\xi_i^p(s) - \tilde{\xi}_i^p(s)) + \eta(E_0 + E_1 + E_2 + E_3),
\end{aligned} \tag{53}$$

where we decompose the last two terms into the following four quantities:

$$\begin{aligned}
E_0 &= \mathbb{E}_{a_{-i}\sim\pi_{-i}^p(\cdot|s)}[Q(s, a_{-i}, a_i) - Q(s, a_{-i}, a_i')] - \mathbb{E}_{a_j\sim\tilde{\pi}_j^p(\cdot|s), j\neq i}[Q(s, a_{-i}, a_i) - Q(s, a_{-i}, a_i')] \\
E_1 &= \mathbb{E}_{a_j\sim\tilde{\pi}_j^p(\cdot|s), j\neq i}[Q(s, a_{-i}, a_i) - Q(s, a_{-i}, a_i')] - \mathbb{E}_{a_j\sim\tilde{\pi}_j^p(\cdot|\overline{s_{\mathcal{N}_i^\kappa}}), j\neq i}[Q(s, a_{-i}, a_i) - Q(s, a_{-i}, a_i')] \\
E_2 &= \mathbb{E}_{a_j\sim\tilde{\pi}_j^p(\cdot|\overline{s_{\mathcal{N}_i^\kappa}}), j\neq i}[Q(s, a_{-i}, a_i) - Q(s, a_{-i}, a_i')] \\
&\quad - \mathbb{E}_{a_j\sim\tilde{\pi}_j^p(\cdot|\overline{s_{\mathcal{N}_i^\kappa}}), j\neq i}[Q(\overline{s_{\mathcal{N}_i^\kappa}}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{-i}, a_i) - Q(\overline{s_{\mathcal{N}_i^\kappa}}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{-i}, a_i')] \\
E_3 &= \mathbb{E}_{a_j\sim\tilde{\pi}_j^p(\cdot|\overline{s_{\mathcal{N}_i^\kappa}}), j\neq i}[Q(\overline{s_{\mathcal{N}_i^\kappa}}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{-i}, a_i) - Q(\overline{s_{\mathcal{N}_i^\kappa}}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{-i}, a_i')] \\
&\quad - \mathbb{E}_{a_j\sim\tilde{\pi}_j^p(\cdot|\overline{s_{\mathcal{N}_i^\kappa}}), j\neq i}[\tilde{Q}^i(s, a_{-i}, a_i) - \tilde{Q}^i(s, a_{-i}, a_i')].
\end{aligned}$$

First note that E_0 is the difference caused by sampling from $\pi_j^p(\cdot|s)$ and $\tilde{\pi}_j^p(\cdot|s)$ while E_1 is the difference induced by sampling from $\tilde{\pi}_j^p(\cdot|s)$ and $\tilde{\pi}_j^p(\cdot|\overline{s_{\mathcal{N}_i^\kappa}})$. Second, term E_2 is the error between querying Q functions at s and a_{-i} and at $\overline{s_{\mathcal{N}_i^\kappa}}$ and $[\overline{a_{\mathcal{N}_i^\kappa}}]_{-i}$. Finally, E_3 accounts for the difference in Q and $\tilde{Q}$. In what follows we bound these quantities one by one.

Bound on E_0. Lemma 13 can be directly applied to bound E_0, which leads to

$$|E_0| \leq \sum_{\ell\neq i} \mathrm{TV}(\pi_\ell^p(\cdot|s), \tilde{\pi}_\ell^p(\cdot|s)) H_{i\ell}^Q \leq \sum_{\ell\neq i} H_{i\ell}^Q v_\ell^p, \tag{54}$$

where in the last inequality we have used the definition of v_ℓ^p in (52).

Bound on E_1. By Lemma 13, E_1 can be bounded by

$$\begin{aligned}
|E_1| &\leq \sum_{j\neq i} \mathrm{TV}(\tilde{\pi}_j^p(\cdot|s), \tilde{\pi}_j^p(\cdot|\overline{s_{\mathcal{N}_i^\kappa}})) H_{ij}^Q \\
&\leq \sum_{j\neq i} \left(\mathrm{TV}(\tilde{\pi}_j^p(\cdot|s), \pi_j^p(\cdot|s)) + \mathrm{TV}(\pi_j^p(\cdot|s), \pi_j^p(\cdot|\overline{s_{\mathcal{N}_i^\kappa}})) + \mathrm{TV}(\pi_j^p(\cdot|\overline{s_{\mathcal{N}_i^\kappa}}), \tilde{\pi}_j^p(\cdot|\overline{s_{\mathcal{N}_i^\kappa}}))\right) H_{ij}^Q
\end{aligned}$$

$$
\begin{aligned}
&\le 2\sum_{j\neq i} v_j^p H_{ij}^Q + \sum_{j\neq i} \mathrm{TV}(\pi_j^p(\cdot|s), \pi_j^p(\cdot|\overline{s_{\mathcal{N}_i^\kappa}}))H_{ij}^Q \\
&\le 2\sum_{j\neq i} v_j^p H_{ij}^Q + \sum_{j\notin \mathcal{N}_i^{\kappa/2}} H_{ij}^Q + \sum_{j\in\mathcal{N}_i^{\kappa/2}, j\neq i} \sum_{\ell\notin\mathcal{N}_i^\kappa} Z_{j\ell}^{\pi^p} H_{ij}^Q \\
&\le 2\sum_{j\neq i} v_j^p H_{ij}^Q + \sum_{j\notin \mathcal{N}_i^{\kappa/2}} H_{ij}^Q + \sum_{j\in\mathcal{N}_i^{\kappa/2}, j\neq i} \sum_{\ell:\mathrm{dist}(\ell,j)>\frac{\kappa}{2}} Z_{j\ell}^{\pi^p} H_{ij}^Q \\
&\le 2\sum_{j\neq i} v_j^p H_{ij}^Q + \frac{2^{\mu+1}\nu' + 2^{\mu+1}\nu' \frac{2^{\mu+1}\nu' A_{\max}^2 e^{\frac{\nu'}{n\tau}}}{\tau - 2^{\mu+1}\nu' A_{\max}^2 e^{\frac{\nu'}{n\tau}}}}{(\kappa/2+1)^\mu} \\
&\le 2\sum_{j\neq i} v_j^p H_{ij}^Q + \frac{2^{\mu+2}\nu'}{(\frac{\kappa}{2}+1)^\mu},
\end{aligned}
$$

where in the fourth inequality we used similar augments like we did in (17) and (18), and in the last two inequalities we have used Lemma 16 and Corollary 19 and the conditions on $\tau > 2^{\mu+2}\nu' A_{\max}^2 e^{\frac{\nu'}{n\tau}}$.

Bound on E_2. To bound E_2, we only have to bound

$$
\begin{aligned}
&\left|Q(s, a_{-i}, a_i) - Q(s, a_{-i}, a_i') - Q(\overline{s_{\mathcal{N}_i^\kappa}}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{-i}, a_i) + Q(\overline{s_{\mathcal{N}_i^\kappa}}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{-i}, a_i')\right| \\
&\le \frac{1}{n}\sum_{j=1}^n \left|Q_j(s, a_{-i}, a_i) - Q_j(s, a_{-i}, a_i') - Q_j(\overline{s_{\mathcal{N}_i^\kappa}}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{-i}, a_i) + Q_j(\overline{s_{\mathcal{N}_i^\kappa}}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{-i}, a_i')\right|,
\end{aligned}
$$

which further implies that we only have to upper bound each term inside the above summation. To this end, we introduce the following lemma.

Lemma 22. *Under the scenario of Lemma 8, we have the following bound:*

$$
|Q_j(s, a_{-i}, a_i) - Q_j(s, a_{-i}, a_i') - Q_j(\overline{s_{\mathcal{N}_i^\kappa}}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{-i}, a_i) + Q_j(\overline{s_{\mathcal{N}_i^\kappa}}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{-i}, a_i')| \le \frac{2\nu'}{(\kappa/2+1)^\mu}.
$$

Proof. Indeed, we have two ways to upper bound this term, the first way is

$$
\begin{aligned}
&|Q_j(s, a_{-i}, a_i) - Q_j(s, a_{-i}, a_i') - Q_j(\overline{s_{\mathcal{N}_i^\kappa}}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{-i}, a_i) + Q_j(\overline{s_{\mathcal{N}_i^\kappa}}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{-i}, a_i')| \\
&\le |Q_j(s, a_{-i}, a_i) - Q_j(s, a_{-i}, a_i')| + |Q_j(\overline{s_{\mathcal{N}_i^\kappa}}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{-i}, a_i) - Q_j(\overline{s_{\mathcal{N}_i^\kappa}}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{-i}, a_i')| \\
&\le 2\frac{\nu'}{(\mathrm{dist}(i,j)+1)^\mu},
\end{aligned}
$$

where in the last inequality, we have used the (ν', μ)-decay property of $\{Q_j\}$. Also we have another way for this bound, that is,

$$
\begin{aligned}
&|Q_j(s, a_{-i}, a_i) - Q_j(s, a_{-i}, a_i') - Q_j(\overline{s_{\mathcal{N}_i^\kappa}}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{-i}, a_i) + Q_j(\overline{s_{\mathcal{N}_i^\kappa}}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{-i}, a_i')| \\
&\le |Q_j(s, a_{-i}, a_i) - Q_j(\overline{s_{\mathcal{N}_i^\kappa}}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{-i}, a_i)| + |Q_j(s, a_{-i}, a_i') - Q_j(\overline{s_{\mathcal{N}_i^\kappa}}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{-i}, a_i')| \\
&\le 2\frac{\nu'}{\max(1, \kappa - \mathrm{dist}(i,j)+1)^\mu}.
\end{aligned}
$$

Thus we have

$$
\begin{aligned}
&|Q_j(s, a_{-i}, a_i) - Q_j(s, a_{-i}, a_i') - Q_j(\overline{s_{\mathcal{N}_i^\kappa}}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{-i}, a_i) + Q_j(\overline{s_{\mathcal{N}_i^\kappa}}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{-i}, a_i')| \\
&\le 2\min\left\{\frac{\nu'}{(\mathrm{dist}(i,j)+1)^\mu}, \frac{\nu'}{\max(1, \kappa - \mathrm{dist}(i,j)+1)^\mu}\right\}
\end{aligned}
$$

$$\leq \frac{2\nu'}{(\kappa/2+1)^{\mu}},$$

where the last inequality holds since either $\text{dist}(i,j)$ or $\kappa - \text{dist}(i,j)$ will be larger than $\kappa/2$. □

By Lemma 22, we can derive an upper bound for E_2 as follows.

$$|E_2| \leq \frac{2}{n}\sum_{j=1}^{n}\frac{\nu'}{(\kappa/2+1)^{\mu}} = \frac{2\nu'}{(\kappa/2+1)^{\mu}}. \tag{55}$$

Bound on E_3. Similar to the proof we presented in bounding term E_2, to upper bound E_3, we only need to bound the following term.

$$\begin{aligned}
&\left|Q(\overline{s_{\mathcal{N}_i^\kappa}}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{-i}, a_i) - Q(\overline{s_{\mathcal{N}_i^\kappa}}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{-i}, a_i') - \tilde{Q}(s, a_{-i}, a_i) + \tilde{Q}(s, a_{-i}, a_i')\right| \\
&= \frac{1}{n}\Bigg| \sum_{j\in\mathcal{N}_i^\kappa}\left[Q_j(\overline{s_{\mathcal{N}_i^\kappa}}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{-i}, a_i) - \hat{Q}_j([\overline{s_{\mathcal{N}_i^\kappa}}]_{\mathcal{N}_j^\beta}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{\mathcal{N}_j^\beta\setminus i}, a_i)\right] + \sum_{j\notin\mathcal{N}_i^\kappa} Q_j(\overline{s_{\mathcal{N}_i^\kappa}}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{-i}, a_i) \\
&\quad - \sum_{j\in\mathcal{N}_i^\kappa}\left[Q_j(\overline{s_{\mathcal{N}_i^\kappa}}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{-i}, a_i') - \hat{Q}_j([\overline{s_{\mathcal{N}_i^\kappa}}]_{\mathcal{N}_j^\beta}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{\mathcal{N}_j^\beta\setminus i}, a_i')\right] - \sum_{j\notin\mathcal{N}_i^\kappa} Q_j(\overline{s_{\mathcal{N}_i^\kappa}}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{-i}, a_i')\Bigg| \\
&\leq \frac{1}{n}\sum_{j\in\mathcal{N}_i^\kappa}\left|Q_j(\overline{s_{\mathcal{N}_i^\kappa}}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{-i}, a_i) - \hat{Q}_j([\overline{s_{\mathcal{N}_i^\kappa}}]_{\mathcal{N}_j^\beta}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{\mathcal{N}_j^\beta\setminus i}, a_i)\right| \\
&\quad + \frac{1}{n}\sum_{j\in\mathcal{N}_i^\kappa}\left|Q_j(\overline{s_{\mathcal{N}_i^\kappa}}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{-i}, a_i') - \hat{Q}_j([\overline{s_{\mathcal{N}_i^\kappa}}]_{\mathcal{N}_j^\beta}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{\mathcal{N}_j^\beta\setminus i}, a_i')\right| \\
&\quad + \frac{1}{n}\sum_{j\notin\mathcal{N}_i^\kappa}\left|Q_j(\overline{s_{\mathcal{N}_i^\kappa}}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{-i}, a_i) - Q_j(\overline{s_{\mathcal{N}_i^\kappa}}, [\overline{a_{\mathcal{N}_i^\kappa}}]_{-i}, a_i')\right| \\
&\leq 2\frac{|\mathcal{N}_i^\kappa|}{n}\epsilon + \frac{\nu'}{n(\kappa+1)^{\mu}},
\end{aligned} \tag{56}$$

where in the last step we used the error bound on $\hat{Q}_j$ in the condition of Lemma 20, and the property of (μ,ν)-decay matrices presented in (8).

Putting terms E_0, E_1, E_2, E_3 together. Combining the bounds in the previous steps, we have,

$$\begin{aligned}
|\xi_i^{p+1}(s) - \tilde{\xi}_i^{p+1}(s)| &\leq (1-\eta\tau)|\xi_i^p(s) - \tilde{\xi}_i^p(s)| + \eta(|E_0| + |E_1| + |E_2| + |E_3|) \\
&\leq (1-\eta\tau)|\xi_i^p(s) - \tilde{\xi}_i^p(s)| + \eta\left(3\sum_{\ell\neq i} v_\ell^p H_{i\ell}^Q + \frac{2^{\mu+2}\nu' + 2\nu'}{(\kappa/2+1)^{\mu}} + \frac{\nu'/n}{(\kappa+1)^{\mu}} + 2|\mathcal{N}_i^\kappa|\epsilon/n\right) \\
&\leq (1-\eta\tau)|\xi_i^p(s) - \tilde{\xi}_i^p(s)| + 3\eta\sum_{\ell\neq i} v_\ell^p H_{i\ell}^Q + \eta\left(\frac{3\nu' + 2^{\mu+2}\nu'}{(\kappa/2+1)^{\mu}} + 2\epsilon\right).
\end{aligned} \tag{57}$$

Plugging $\eta = \frac{1}{\tau}$ into the above inequality, we have

$$|\xi_i^{p+1}(s) - \tilde{\xi}_i^{p+1}(s)| \leq \frac{3}{\tau}\sum_{\ell\neq i} v_\ell^p H_{i\ell}^Q + \frac{1}{\tau}\left(\frac{3\nu' + 2^{\mu+2}\nu'}{(\kappa/2+1)^{\mu}} + 2\epsilon\right).$$

By Lemma 17, we have

$$v_i^{p+1} \leq \frac{1}{2}A_{\max}^2 e^{\tilde{\sigma}'}\left(\exp\left(\frac{3}{\tau}\sum_{\ell\neq i} v_\ell^p H_{i\ell}^Q + \frac{1}{\tau}\left(\frac{3\nu' + 2^{\mu+2}\nu'}{(\kappa/2+1)^{\mu}} + 2\epsilon\right)\right) - 1\right).$$

Recall our definition of $\epsilon = \frac{c_{pe}}{(\beta+1)^\mu}$ in (20) and the choice of $\tau > 2(2^{\mu+4}\nu' + 2\frac{c_{pe}}{(\beta+1)^\mu})$ in Lemma 20. With these results, we can easily show that the quantity inside the $\exp(\cdot)$ can be upper bounded by

$$\frac{3}{\tau}\sum_{\ell\neq i} v_\ell^p H_{i\ell}^Q + \frac{1}{\tau}\left(\frac{3\nu' + 2^{\mu+2}\nu'}{(\kappa/2+1)^\mu} + 2\epsilon\right) \leq \frac{1}{\tau}\left(3\cdot 2^{\mu+1}\nu' + 3\nu' + 2^{\mu+2}\nu' + 2c_{pe}\frac{1}{(\beta+1)^\mu}\right)$$
$$\leq \frac{1}{2}. \tag{58}$$

By the simple fact that $e^x \leq 1 + 2x$ for $x \leq 1/2$, we have

$$v_i^{p+1} \leq A_{\max}^2 e^{\tilde{\sigma}'}\left(3\frac{1}{\tau}\sum_{\ell\neq i} v_\ell^p H_{i\ell}^Q + \frac{1}{\tau}\left(\frac{2^{\mu+3}\nu' + 2c_{pe}}{(\kappa/2+1)^\mu}\right)\right).$$

If we stack all v_i^p into a vector $v^p = [v_i^p]_{i\in\mathcal{N}}$, the above inequality implies

$$v^{p+1} \leq A_{\max}^2 e^{\tilde{\sigma}'}\left(3\frac{1}{\tau}H^Q v^p + \frac{1}{\tau}\left(\frac{2^{\mu+3}\nu' + 2c_{pe}}{(\kappa/2+1)^\mu}\right)\mathbf{1}\right),$$

and it immediately follows that

$$\|v^{p+1}\|_\infty \leq A_{\max}^2 e^{\tilde{\sigma}'}\frac{3}{\tau}\|H^Q\|_\infty\|v^p\|_\infty + A_{\max}^2 e^{\tilde{\sigma}'}\frac{1}{\tau}\frac{2^{\mu+3}\nu' + 2c_{pe}}{(\kappa/2+1)^\mu}.$$

Therefore, when $\tau > 6A_{\max}^2 e^{\tilde{\sigma}'}2^{\mu+1}\nu'$, we have $2A_{\max}^2 e^{\tilde{\sigma}'}\|H^Q\|_\infty/\tau \leq 1/2$. And thus for all p, it holds that

$$\|v^p\|_\infty \leq 2A_{\max}^2 e^{\tilde{\sigma}'}\frac{1}{\tau}\frac{2^{\mu+3}\nu' + 2c_{pe}}{(\kappa/2+1)^\mu} \leq \frac{4 + \frac{c_{pe}}{2^\mu\nu'}}{3(\kappa/2+1)^\mu}.$$

Further, using Lemma 18, when $\tau > 2^{\mu+2}\nu' A_{\max}^2 e^{\nu'/n\tau}$, we have for each $i \in [n]$ that

$$\sup_{s\in\mathcal{S}} \mathrm{TV}(\pi_i^p(\cdot|s), \zeta_i(\cdot|s)) \leq \frac{1}{2^p}.$$

By triangle inequality, we have

$$\sup_{s\in\mathcal{S}} \mathrm{TV}(\hat{\pi}_i^p(\cdot|s_{\mathcal{N}_i^\kappa}), \zeta_i(\cdot|s)) \leq \frac{1}{2^p} + \frac{4 + \frac{c_{pe}}{2^\mu\nu'}}{3(\kappa/2+1)^\mu}.$$

Therefore, when $p \geq -\log_2 \frac{4+\frac{c_{pe}}{2^\mu\nu'}}{3(\kappa/2+1)^\mu}$, we have

$$\sup_{s\in\mathcal{S}} \mathrm{TV}(\hat{\pi}_i^p(\cdot|s_{\mathcal{N}_i^\kappa}), \zeta_i(\cdot|s)) \leq \frac{4 + \frac{c_{pe}}{2^\mu\nu'}}{(\kappa/2+1)^\mu}. \tag{59}$$

This completes the proof of the first statement in **Part (c)** of Lemma 20. Next we prove the second statement. Since we now have a total variation bound (59), we can use Lemma 7 to bound the value function difference between the policy ζ_i obtained by directly applying the Bellman optimal operator and the policy $\hat{\pi}_i$ carried out by our policy iteration process. Specifically, we have that

$$\begin{aligned}
\mathcal{T}V - V^{\hat{\pi}^p} &\leq V^\zeta - V^{\hat{\pi}^p} \\
&\leq \|V^\zeta - V^{\hat{\pi}^p}\|_\infty \mathbf{1} \\
&\leq \frac{1}{1-\gamma}(\tau\tilde{\sigma}' + \frac{\nu'}{n})\sum_{i=1}^n \sup_{s\in\mathcal{S}} \mathrm{TV}(\zeta_i(\cdot|s), \hat{\pi}_i(\cdot|s_{\mathcal{N}_i^\kappa}))\mathbf{1}
\end{aligned}$$

$$\leq \frac{1}{1-\gamma}(n\tau\tilde{\sigma}' + \nu')\frac{4+\frac{c_{pe}}{2^\mu \nu'}}{(\kappa/2+1)^\mu}\mathbf{1}$$
$$= \frac{1}{1-\gamma}(\frac{2f(\kappa)c_{pe}}{(\beta+1)^\mu} + 2\nu')\frac{4+\frac{c_{pe}}{2^\mu \nu'}}{(\kappa/2+1)^\mu}\mathbf{1}.$$

where the third inequality is due to **Part (b)** of Lemma 20 and Lemma 7. Now that we obtain the upper bound for each step of our outer loop, which completes the proof of part (c).

F RESULTS ON SUFFICIENT EXPLORATION

In this section, we provide a general result on the sufficient exploration for non-deterministic policies. Given $\delta > 0$, we define $\Pi_\delta = \{\zeta \in \Delta_{\text{policy}} \mid \min_{s,a} \zeta(a|s) \geq \delta\}$. For a σ-regular policy ζ, since $\zeta_i(a_i|s)/\zeta_i(\tilde{a}_i|s) \leq e^\sigma$ for all $i = 1, 2\cdots, n$, $s \in \mathcal{S}$, and $a_i, \tilde{a}_i \in \mathcal{A}_i$, we have

$$\min_{a_i \in \mathcal{A}_i} \zeta_i(a_i|s) \geq \frac{\max_{\tilde{a}_i \in \mathcal{A}_i} \zeta_i(\tilde{a}_i|s)}{e^\sigma} \geq \frac{1}{|\mathcal{A}_i|e^\sigma}.$$

It follows that

$$\min_{a \in \mathcal{A}} \zeta(a|s) \geq \frac{1}{e^{n\sigma}\prod_{i=1}^n |\mathcal{A}_i|}.$$

Therefore, all σ-regular policies are contained in Π_δ with $\delta = 1/(e^{n\sigma}\prod_{i=1}^n |\mathcal{A}_i|)$.

Proposition 3. *Suppose that Assumption 1 is satisfied. Then we have the following results for any $\zeta \in \Pi_\delta$.*

(1) The Markov chain $\{S_k\}$ induced by ζ is irreducible and aperiodic.
(2) Let μ^ζ be the unique stationary distribution of the Markov chain $\{S_k\}$ induced by ζ, then we have $\mu_{\min} := \inf_{\zeta \in \Pi_\delta} \min_{s \in \mathcal{S}} \mu^\zeta(s) > 0$.
(3) There exist constants $C > 0$ and $\rho \in (0,1)$ such that

$$\sup_{\zeta \in \Pi_\delta} TV\Big(P_\zeta^t(\cdot \mid s(0) = s), \mu_\zeta(\cdot)\Big) \leq C\rho^t, \quad \forall\, t \geq 0.$$

Proposition 3 states that, as long as there is one policy $\zeta_b \in \Pi_\delta$ that is explorative, then all policies in Π_δ are uniformly explorative. Note that Lemma 4 is a direct consequence of Proposition 3. Therefore, we only need to prove Proposition 3.

Proof of Proposition 3. **Proof of Part (1).** Let $\zeta \in \Pi_\delta$ be arbitrary, and let P_{ζ_b} and P_ζ be the transition probability matrices under ζ_b and ζ, respectively. For any $s, s' \in \mathcal{S}$ and $t \geq 1$, we have

$$\begin{aligned}
P_{\zeta_b}^t(s,s') &= \sum_{s_0} P_{\zeta_b}^{t-1}(s,s_0)P_{\zeta_b}(s_0,s') \\
&= \sum_{s_0} P_{\zeta_b}^{t-1}(s,s_0)\sum_{a\in\mathcal{A}} \zeta_b(a|s_0)P_a(s_0,s') \\
&= \sum_{s_0} P_{\zeta_b}^{t-1}(s,s_0)\sum_{a\in\mathcal{A}} \frac{\zeta_b(a|s_0)}{\zeta(a|s_0)}\zeta(a|s_0)P_a(s_0,s') \\
&\leq \frac{1}{\delta}\sum_{s_0} P_{\zeta_b}^{t-1}(s,s_0)\sum_{a\in\mathcal{A}} \zeta(a|s_0)P_a(s_0,s') \\
&\leq \frac{1}{\delta}\sum_{s_0} P_{\zeta_b}^{t-1}(s,s_0)P_\zeta(s_0,s') \\
&= \frac{1}{\delta}[P_{\zeta_b}^{t-1}P_\zeta](s,s').
\end{aligned}$$

Since the previous inequality holds for all s and s', we in fact have $\delta P^t_{\zeta_b} \le P^{t-1}_{\zeta_b} P_\zeta$. Repeatedly using the previous inequality and we obtain

$$(\delta)^t P^t_{\zeta_b} \le P^t_\zeta,$$

for all $t \ge 1$. When P_{ζ_b} is irreducible and aperiodic, the previous inequality implies that P_ζ is irreducible and aperiodic.

Proof of Part (2). To establish the result, we need the following sequence of lemmas.

Lemma 23. *The mapping from a policy $\zeta \in \Pi_\delta$ to its corresponding transition probability matrix P_ζ is linear and is 1-Lipschitz continuous with respect to the ℓ_∞-norm.*

Proof of Lemma 23. The linearity is straightforward, we here only compute the Lipschitz constant. Let $\zeta_1, \zeta_2 \in \Pi_\delta$ be two policies, and let P_{ζ_1} and P_{ζ_2} be the transition probability matrices induced by ζ_1 and ζ_2, respectively. We will view ζ_1 and ζ_2 as $|\mathcal{S}|$ by $|\mathcal{A}|$ matrices. Using the definition of induced matrix norm and we have

$$\begin{aligned}
\|P_{\zeta_1} - P_{\zeta_2}\|_\infty &= \max_{s\in\mathcal{S}} \sum_{s'\in\mathcal{S}} |P_{\zeta_1}(s,s') - P_{\zeta_2}(s,s')| \\
&= \max_{s\in\mathcal{S}} \sum_{s'\in\mathcal{S}} \left| \sum_{a\in\mathcal{A}} (\zeta_1(a|s) - \zeta_2(a|s)) P_a(s,s') \right| \\
&\le \max_{s\in\mathcal{S}} \sum_{s'\in\mathcal{S}} \sum_{a\in\mathcal{A}} |\zeta_1(a|s) - \zeta_2(a|s)| P_a(s,s') \\
&= \max_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} \sum_{s'\in\mathcal{S}} |\zeta_1(a|s) - \zeta_2(a|s)| P_a(s,s') \\
&= \max_{s\in\mathcal{S}} \sum_{a\in\mathcal{A}} |\zeta_1(a|s) - \zeta_2(a|s)| \\
&= \|\zeta_1 - \zeta_2\|_\infty.
\end{aligned}$$

This completes the proof. □

Since Π_δ is a convex compact set and the mapping from $\zeta \in \Pi_\delta$ to the transition probability P_ζ is linear (hence continuous), the image space $P_{\Pi_\delta} = \{P_\zeta \mid \zeta \in \Pi_\delta\}$ is also convex and compact.

Lemma 24. *Let $P_1, P_2 \in P_{\Pi_\delta}$, and let μ_1 and μ_2 be their corresponding unique stationary distributions, respectively. Then there exists L (which depends on P_1) such that*

$$\|\mu_1 - \mu_2\|_2 \le L \|P_1 - P_2\|_2.$$

As a result, the mapping from an irreducible and aperiodic stochastic matrix $P \in P_{\Pi_\delta}$ to its unique stationary distribution is continuous.

Proof of Lemma 24. Since P_1 is an irreducible and aperiodic stochastic matrix, there exists $C_1 > 0$ and $\rho_1 \in (0,1)$ such that

$$\|P_1^t - M_1\|_2 \le C_1 \rho_1^t,$$

for all $t \ge 0$, where M_1 is a matrix with $|\mathcal{S}|$ rows, each of which is the vector μ_1 [20, Theorem 4.9]. By definition we have $\mu_1^\top P_1^t = \mu_1^\top$ and $\mu_2^\top P_2^t = \mu_2^\top$ for all $t \ge 0$. It follows that

$$\begin{aligned}
\mu_1 - \mu_2 &= (P_1^t)^\top (\mu_1 - \mu_2) + (P_1^t - P_2^t)^\top \mu_2 \\
&= (P_1^t - M_1)^\top (\mu_1 - \mu_2) + M_1^\top (\mu_1 - \mu_2) + (P_1^t - P_2^t)^\top \mu_2. \qquad (60)
\end{aligned}$$

Applying $\|\cdot\|_2$ on both sides of Eq. (60) and then using triangle inequality, and we have

$$\begin{aligned}\|\mu_1-\mu_2\|_2 &\le \|(P_1^t-M_1)^\top(\mu_1-\mu_2)\|_2+\|M_1^\top(\mu_1-\mu_2)\|_2+\|(P_1^t-P_2^t)^\top\mu_2\|_2\\ &\le \|(P_1^t-M_1)^\top\|_2\|\mu_1-\mu_2\|_2+\|M_1^\top(\mu_1-\mu_2)\|_2+\|(P_1^t-P_2^t)^\top\|_2\|\mu_2\|_2\\ &\le C_1\rho_1^t\|\mu_1-\mu_2\|_2+\|M_1^\top(\mu_1-\mu_2)\|_2+\|P_1^t-P_2^t\|_2.\end{aligned}$$

For the term $\|M_1^\top(\mu_1-\mu_2)\|_2$, we have by definition of M_1 that

$$\|M_1^\top(\mu_1-\mu_2)\|_2=\|\mu_1-\mu_1\|_2=0.$$

It follows that

$$\|\mu_1-\mu_2\|_2\le C_1\rho_1^t\|\mu_1-\mu_2\|_2+\|P_1^t-P_2^t\|_2.$$

When $t\ge\frac{\log(2C_1)}{\log(1/\rho_1)}$ (which implies $C_1\rho_1^t\le\frac{1}{2}$), we have from the previous inequality that

$$\|\mu_1-\mu_2\|_2\le 2\|P_1^t-P_2^t\|_2.$$

We next bound $\|P_1^t-P_2^t\|_2$ in terms of $\|P_1-P_2\|_2$. Observe that

$$\begin{aligned}P_1^t-P_2^t&=P_1^t+\sum_{i=1}^{t-1}(-P_1^{t-i}P_2^i+P_1^{t-i}P_2^i)-P_2^t\\ &=\sum_{i=1}^{t}(P_1^{k-i+1}P_2^{i-1}-P_1^{t-i}P_2^i)\\ &=\sum_{i=1}^{t}P_1^{t-i}(P_1-P_2)P_2^{i-1}.\end{aligned}$$

Therefore, we have

$$\|P_1^t-P_2^t\|_2\le\sum_{i=1}^{t}\|P_1^{t-i}\|_2\|P_1-P_2\|_2\|P_2^{i-1}\|_2\le tN_{\max}^2\|P_1-P_2\|_2,$$

Here $N_{\max}:=\max\{\|P\|_2 : P\in P_{\Pi_\delta}\}$, which is well-defined and finite since P_{Π_δ} is a compact set and the ℓ_2-norm is a continuous function. Finally, we obtain

$$\|\mu_1-\mu_2\|_2\le 2tN_{\max}^2\|P_1-P_2\|_2,$$

when $t\ge\frac{\log(2C_1)}{\log(1/\rho_1)}$. Note that while $N_{\max}$ is independent of P_1 and P_2, the factor t depends on C_1 and ρ_1, both of which are functions of the stochastic matrix P_1. □

Lemma 25. *For any $\mu\in\mathbb{R}^{|\mathcal{S}|}$, the mapping from μ to its minimum component is Lipschitz continuous.*

Proof of Lemma 25. For any $\mu_1,\mu_2\in\mathbb{R}^{|\mathcal{S}|}$, we have

$$\left|\min_{s\in\mathcal{S}}\mu_1(s)-\min_{s\in\mathcal{S}}\mu_2(s)\right|\le\max_{s\in\mathcal{S}}|\mu_1(s)-\mu_2(s)|=\|\mu_1-\mu_2\|_\infty.$$

□

Combining Lemma 23, Lemma 24, and Lemma 25 and we conclude that the mapping from $\zeta\in\Pi_\delta$ to the minimum component of the stationary distribution μ_ζ of the induced Markov chain $\{S_k\}$ is continuous. In addition, since Π_δ is compact, we have by Weierstrass extreme value theorem that

$$\mu_{\min}=\inf_{\zeta\in\Pi_\delta}\min_{s\in\mathcal{S}}\mu_\zeta(s)=\min_{\zeta\in\Pi_\delta}\min_{s\in\mathcal{S}}\mu_\zeta(s)>0.$$

Proof of Part (3). Given the lemmas presented in the proof of Proposition Proposition 3 (2), this part directly follows from existing results in the literature, see for example [43, Lemma 1]. The idea is to mimick the proof of the ergodic theorem [20, Theorem 4.9] for irreducible and aperiodic Markov chains and perform a refined analysis. □

Received October 2022; revised December 2022; accepted January 2023