Convergence Rates for Localized Actor-Critic in Networked Markov Potential Games (Supplementary material)

Zhaoyi Zhou[1] **Zaiwei Chen**[2] **Yiheng Lin**[2] **Adam Wierman**[2]

[1]Institute for Interdisciplinary Information Sciences, Tsinghua University
[2]The Computing + Mathematical Sciences (CMS) Department, California Institute of Technology

Abstract

We introduce a class of networked Markov potential games where agents are associated with nodes in a network. Each agent has its own local potential function, and the reward of each agent depends only on the states and actions of agents within a neighborhood. In this context, we propose a localized actor-critic algorithm. The algorithm is scalable since each agent uses only local information and does not need access to the global state. Further, the algorithm overcomes the curse of dimensionality through the use of function approximation. Our main results provide finite-sample guarantees up to a localization error and a function approximation error. Specifically, we achieve an $\tilde{\mathcal{O}}(\tilde{\epsilon}^{-4})$ sample complexity measured by the averaged Nash regret. This is the first finite-sample bound for multi-agent competitive games that does not depend on the number of agents.

1 INTRODUCTION

Large-scale systems where agents interact competitively with each other have received significant attention recently, motivated by applications in power systems [Shi et al., 2022], EV charging [Lee et al., 2022], and board games [Silver et al., 2017], etc. Controlling such systems can be challenging due to the scale of the system, uncertainty about the model, communication constraints, and the interaction between agents. Inspired by the recent success of reinforcement learning (RL), there is an increasing interest in applying RL methods to environments with multi-agent interactions. However, in multi-agent RL (MARL), the analysis of the system behavior becomes challenging due to the time-varying nature of the environment faced by each agent, which results from the (time-varying) competitive decisions of other agents. As a result, the theoretical analy-

sis of MARL, especially in the competitive setting, is still limited especially when it comes to large-scale systems.

Results on MARL in competitive settings to this point have tended to focus on games with a small number of players, e.g., 2-player zero-sum stochastic games [Littman, 1994], or games with special structure, e.g., Markov potential games (MPGs) [Fox et al., 2022]. MPGs in particular provide a setting in which the challenges of large-scale systems can be studied. The intuition behind an MPG parallels that of classical (one-shot) potential games. Specifically, the existence of a potential function guarantees that agents can converge to a global equilibrium even when using greedy localized updates. MPGs have wide-ranging applications including variants of congestion games [Leonardos et al., 2022, Fox et al., 2022], medium access control [Macua et al., 2018], and the stochastic lake game [Dechert and O'Donnell, 2006]. However, existing theoretical results for MPGs rely on the assumption that a centralized global state exists and can be observed by each individual agent. Such an assumption rules out applications in many large-scale systems including transportation networks [Zhang and Pavone, 2016] and social networks [Chakrabarti et al., 2008], where the global state space can be exponentially large in the number of agents and/or each agent can only observe its own local state.

A promising approach for the design of scalable and local MARL algorithms in competitive settings is to exploit the networked structure of practical applications to design algorithms with sample complexity that only depends on *local* properties of the network instead of the *global* state. This approach has recently been successful in the case of cooperative MARL. For example, Qu et al. [2020], Lin et al. [2021], Zhang et al. [2022c] provides a scalable localized algorithm with a sample complexity that does not depend on the number of agents. However, to this point, local algorithms that exploit network structure do not exist in the competitive MARL setting. Thus, we ask: *Can we design a scalable and local algorithm with finite-time bounds for networked MARL with competitive agents?*

1.1 MAIN CONTRIBUTIONS

We address the question above by introducing a class of networked Markov potential games (NMPGs) as the networked counterpart of classical MPGs. Importantly, NMPGs represent a broader class of games than MPGs, and draw focus to algorithm design that uses only local information.

We design a localized actor-critic algorithm that is a combination of independent policy gradient and localized TD(λ) with linear function approximation. Notably, our algorithm is *model-free*, uses only *local information*, and successfully incorporates *function approximation*. This avoids both the need for communication of the global state and the so-called "curse of dimensionality" in MARL.

Our main results provide a finite-sample bound on the averaged Nash regret for our proposed algorithm, which implies an $\tilde{\mathcal{O}}(\tilde{\epsilon}^{-4})$ sample complexity (where $\tilde{\epsilon}$ is the accuracy) up to an approximation error of using local information and a function approximation error. To our knowledge, we are the first to develop a localized algorithm in competitive MARL settings with provable performance guarantees that do not depend on the number of agents.

Our results are enabled by a novel analysis of the critic in our localized actor-critic framework. In particular, we propose a localized cost evaluation problem, a new MARL setting to investigate the performance of a local algorithm under a fixed policy. As a critical part of the proof, we propose a novel concept called a "sub-chain" that connects local algorithms to their global counterparts, enabling performance bounds via bounds on the gap between the two.

1.2 RELATED WORK

Markov Potential Games. Our work adds to the literature on MPGs in MARL. Analytic results for non-cooperative MARL are challenging to obtain because agents learn in a non-stationary environment as other agents update their policies. As a result, existing analysis has focused on special cases like 2-player stochastic games [Littman, 1994], adversarial team Markov games [Kalogiannis et al., 2022], and MPGs [Fox et al., 2022]. The case of MPGs has received considerable attention recently because the potential games are broadly applicable [Leonardos et al., 2022] and the existence of potential functions enables provable guarantees [Zhang et al., 2022b, Ding et al., 2022, Fox et al., 2022, Zhang et al., 2022a]. While these papers provide algorithms with provable convergence guarantees, they assume that all agents share a common global state and can observe the global state to decide local actions. An important open question is understanding how to learn in settings where global information is not available. Our work studies the MARL setting where each agent has its own local state and can only decide local actions based on the local states.

MARL in Networked Systems. The Markov decision process (MDP) model we study is inspired by a series of works on Networked MARL [Qu et al., 2020, Lin et al., 2021, Zhang et al., 2022c], where RL agents are located on a network. In such models, the local state transition of an agent is affected by its own local state/action and its direct neighbors' local states. Networked MARL is applicable to a wide range of applications, including communication networks [Vogels et al., 2003], social networks [Chakrabarti et al., 2008], and traffic networks [Zhang and Pavone, 2016]. Compared with general MARL, the additional structure of networked MARL enables us to establish a critical exponential decay property on the local Q-functions, which leads to the design of localized actor-critic algorithms [Qu et al., 2020, Lin et al., 2021]. All prior works on networked MARL study the case when agents cooperatively maximize the sum of all local rewards. In contrast, our work studies a non-cooperative NMPG in which each agent has its own objective.

Another approach to study MARL problems is to use mean-field control (MFC) [Gu et al., 2021, Mondal et al., 2022a,b]. The major difference between the mean-field setting and our setting is that mean-field MARL focuses on homogeneous agents, while we allow each agent to have different transition probabilities and local policies.

Finite-Sample Analysis of TD-Learning Variants. TD-learning and its variants are widely used for policy evaluation in RL, which plays a critical role in most policy-space algorithms. The asymptotic analysis of TD-learning dates back to Tsitsiklis [1994], Jaakkola et al. [1994], while finite-sample convergence bounds have received attention in the last decade. In TD-learning, function approximation is a useful technique to reduce the dimension of learning parameters at the cost of incurring an approximation error that depends on the function class. Recently, many breakthroughs are made on finite-sample error bounds for TD-learning with function approximation [Bhandari et al., 2018, Srikant and Ying, 2019, Dalal et al., 2018, Yu and Bertsekas, 2009]. Meanwhile, in multi-agent settings, localized TD-learning is crucial for limiting communication and the need for global information [Lin et al., 2021]. Our work provides a novel finite-sample error bound for localized TD-learning with function approximation.

2 PROBLEM DESCRIPTION

Network Structure. We study MARL in the context of networked multi-agent Markov games. Specifically, we consider a setting with n agents that are associated with an undirected graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$, where $\mathcal{N} = \{1, 2, \ldots, n\}$ is the set of nodes and $\mathcal{E} \subseteq \mathcal{N} \times \mathcal{N}$ is the set of edges. We denote by $\text{dist}(i, j)$ the graph distance between agents i and j. The local state space and local action space of agent i are denoted by $\mathcal{S}_i$ and $\mathcal{A}_i$, respectively, which are both finite sets. The

global state is denoted as $s = (s_1, \ldots, s_n) \in \mathcal{S} := \prod_{i=1}^{n} \mathcal{S}_i$ and the global action is defined similarly. For any subset $I \subseteq \mathcal{N}$, we use s_I to denote the joint state of the agents in I and use $\mathcal{S}_I := \prod_{i \in I} \mathcal{S}_i$ to denote the joint state space of agents in I. Similarly, we define a_I and $\mathcal{A}_I$ as the joint action and joint action space of the agents in I. Denote $\mu \in \Delta(\mathcal{S})$ as the initial state distribution, where $\Delta(\mathcal{S})$ denotes the $|\mathcal{S}|$-dimensional probability simplex.

Transition Probabilities. At time $t \geq 0$, given current state $s(t)$ and action $a(t)$, for each agent $i \in \mathcal{N}$, its successor state $s_i(t+1)$ is independently generated according to the following transition probability, which is only dependent on its neighbors' states and its own action:

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

where $\mathcal{N}_i = \{i\} \cup \{j \in \mathcal{N} \mid (i,j) \in \mathcal{E}\}$ denotes the neighborhood of i, including i itself. In addition, given an arbitrary integer $\kappa \geq 0$, we use N_i^κ to denote the κ-hop neighborhood of i, i.e., $N_i^\kappa = \{i\} \cup \{j \in \mathcal{N} \mid \text{dist}(i,j) \leq \kappa\}$, and use $-N_i^\kappa = \mathcal{N}/N_i^\kappa$ to denote the set of agents that are not in N_i^κ. We use $U_i^\kappa = N_i^\kappa/\{i\}$ to denote the agents in the κ-hop neighborhood of i, excluding i itself.

Remark 2.1. We require that each agent's transition probability depends only on the states of its neighbors and its own action, which is common in networked MARL literature [Qu et al., 2020, Zhang et al., 2022c]. Intuitively, it implies that the impact from far-away agents on the network is "negligible", which eventually leads to the exponential decay property (cf. Lemma 4.1).

Reward Function. Each agent $i \in \mathcal{N}$ is associated with a deterministic reward function $r_i : \mathcal{S} \times \mathcal{A} \mapsto [0,1]$. The interval $[0,1]$ is chosen without loss of generality over the set of bounded reward functions. In general, agent i's reward depends on the global state and the global action. Due to the network structure, we assume that there exists a non-negative integer κ_r such that the reward function of each agent depends only on the states and the actions of other agents within its κ_r-hop neighborhood, i.e., $r_i(s,a) = r_i(s_{\mathcal{N}_i^{\kappa_r}}, a_{\mathcal{N}_i^{\kappa_r}})$ for all i. This makes intuitive sense as we expect the dependence between two agents to weaken as their graph distance grows.

Policy. In this work, we consider stationary policies [Zhang et al., 2021]. Specifically, each agent $i \in \mathcal{N}$ is associated with a localized policy $\xi_i : \mathcal{S}_i \mapsto \Delta(\mathcal{A}_i)$. Given a subset $I \subseteq \mathcal{N}$, we define $\xi_I : \mathcal{S}_I \mapsto \Delta(\mathcal{A}_I)$ as the joint policy of agents in I. Note that $\xi_I(a_I \mid s_I) = \prod_{i \in I} \xi_i(a_i \mid s_i)$. We use Ξ_i to denote agent i's local policy space, and Ξ_I to denote the joint policy space of agents in I. When $I = \mathcal{N}$, we omit the subscript and just write ξ for $\xi_{\mathcal{N}}$ (and Ξ for $\Xi_{\mathcal{N}}$). Throughout, we also use $\xi = (\xi_1, \xi_2, \cdots, \xi_n)$ to highlight the local policy components. In this work, we will frequently

work with softmax policies, which are defined as

$$\xi_i^{\theta_i}(a_i|s_i) = \frac{\exp(\theta_{i,s_i,a_i})}{\sum_{a_i' \in \mathcal{A}_i} \exp(\theta_{i,s_i,a_i'})}, \ \forall \, i, s_i, a_i, \quad (1)$$

where $\xi_i^{\theta_i}$ stands for agent i's local policy parametrized by the weight vector $\theta_i \in \mathbb{R}^{|\mathcal{S}_i||\mathcal{A}_i|}$. We denote $\theta = (\theta_1, \theta_2, \cdots, \theta_n)$ as the parameter of a global policy ξ^θ.

Value Function. Given a global policy ξ and an agent i, we define agent i's Q-function $Q_i^\xi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ as

$$Q_i^\xi(s,a) = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_\xi \left[r_i(s(t), a(t)) \mid s(0) = s, a(0) = a \right]$$

for all (s,a), where $\gamma \in (0,1)$ is the discount factor, and $\mathbb{E}_\xi[\cdot]$ is taken w.r.t. the randomness in the (stochastic) policy ξ and the transition probabilities. With Q_i^ξ defined above, the averaged Q-function $\overline{Q}_i^\xi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}_i|}$ and the value function $V_i^\xi \in \mathbb{R}^{|\mathcal{S}|}$ of agent i are defined as $\overline{Q}_i^\xi(s,a_i) = \mathbb{E}_{a_{-i} \sim \xi_{-i}(\cdot|s_{-i})}[Q_i^\xi(s, a_i, a_{-i})]$ for all (s, a_i) and $V_i^\xi(s) = \mathbb{E}_{a_i \sim \xi_i(\cdot|s_i)}[\overline{Q}_i^\xi(s,a_i)]$ for all s, where we use s_{-i}, a_{-i}, and ξ_{-i} to denote the joint state, the joint action, and the joint policy of the agents in $\mathcal{N}/\{i\}$, respectively. With the initial state distribution μ, we define $J_i(\xi) = \mathbb{E}_{s \sim \mu}[V_i^\xi(s)]$. Finally, we define the advantage function of agent i as $A_i^\xi(s,a) = Q_i^\xi(s,a) - V_i^\xi(s)$ for all (s,a), and the averaged advantage function of agent i as $\overline{A}_i^\xi(s,a_i) = \overline{Q}_i^\xi(s,a_i) - V_i^\xi(s)$ for all (s,a_i). When the policy uses softmax parameterization with parameter θ, we may abuse the policy parameter θ to represent the policy ξ for simplicity. For example, we may write $J_i(\theta)$ for $J_i(\xi^\theta)$.

Discounted State Visitation Distribution. Given a policy ξ and an initial state s', we define the *discounted state visitation distribution* as $d_{s'}^\xi(s) = (1-\gamma) \sum_{t=0}^{\infty} \gamma^t \text{Pr}^\xi[s(t) = s \mid s(0) = s']$ for all $s \in \mathcal{S}$, where $\text{Pr}^\xi[s(t) = s \mid s(0) = s']$ denotes the probability that $s(t) = s$ given that the initial state is s' and the global policy is ξ. We use $d^\xi(s) := \mathbb{E}_{s' \sim \mu}[d_{s'}^\xi(s)]$ to represent the discounted state visitation distribution when the initial state distribution is μ.

3 NETWORKED MPGS

Our focus is a class of networked multi-agent Markov games that we named NMPGs, which is defined in the following.

Definition 3.1. *A multi-agent Markov game is called a κ_G-NMPG (where κ_G is a non-negative integer) if there exists a set of local potential functions $\{\Phi_i\}_{i \in \mathcal{N}}$, where $\Phi_i : \Xi \to \mathbb{R}$ for all $i \in \mathcal{N}$, such that the following equality holds for any $i \in \mathcal{N}$, $j \in \mathcal{N}_i^{\kappa_G}$, $\xi_j, \xi_j' \in \Xi_j$, and $\xi_{-j} \in \Xi_{-j}$:*

$$J_j(\xi_j', \xi_{-j}) - J_j(\xi_j, \xi_{-j}) = \Phi_i(\xi_j', \xi_{-j}) - \Phi_i(\xi_j, \xi_{-j}). \quad (2)$$

Definition 3.1 states that when agent j changes its local policy, the change in its objective function $J_j(\cdot, \xi_{-j})$ can be measured by the change of local potential functions from any agent in its κ_G-hop neighborhood. The non-negative integer κ_G is determined by the networked MPG setting and reflects the extent to which the networked MPG is relaxed from an MPG. Recall that in the definition of a standard MPG [Leonardos et al., 2022], there exists a (global) potential function Φ such that Eq. (2) holds with Φ_i being replaced by Φ for all i. Therefore, an MPG is always an NMPG (by choosing $\Phi_i = \Phi$ for all i), and hence NMPG represents a strictly broader class of games. More discussions are given in Appendix F.1, and a concrete example of an NMPG is presented in Section 3.1.

Due to the boundedness of the reward function and Eq. (2), the local potential functions are uniformly bounded from above and below, i.e., there exist $\Phi_{\min}, \Phi_{\max} > 0$ such that $\Phi_i(\xi) \in [\Phi_{\min}, \Phi_{\max}]$ for all $i \in \mathcal{N}$ and $\xi \in \Xi$. See Appendix F.6 for more details.

Unlike in single-agent RL or cooperative MARL, the optimal policy is not well-defined in the competitive setting, and thus our goal is to design algorithms that learn Nash equilibria of NMPGs. We next introduce the concepts of Nash equilibrium, Nash gap, and averaged Nash regret.

Definition 3.2. *A global policy ξ is a Nash equilibrium if $J_i(\xi_i, \xi_{-i}) \geq J_i(\xi_i', \xi_{-i})$ for all $\xi_i' \in \Xi_i$ and $i \in \mathcal{N}$.*

To measure the performance of a policy by its "distance" to a Nash equilibrium, we use the Nash gap.

Definition 3.3. *Given a global policy ξ, agent i's Nash gap and the global Nash gap are defined as*

$$NE\text{-}Gap_i(\xi) := \max_{\xi_i'} J_i(\xi_i', \xi_{-i}) - J_i(\xi_i, \xi_{-i}),$$

$$NE\text{-}Gap(\xi) := \max_{i \in \mathcal{N}} NE\text{-}Gap_i(\xi).$$

With NE-Gap$(\cdot)$ defined above, given $\hat{\epsilon} > 0$, we say that a policy ξ is an $\hat{\epsilon}$-approximate Nash equilibrium if NE-Gap$(\xi) \leq \hat{\epsilon}$. When using a softmax policy with parameter θ, we may abuse the notation to denote NE-Gap$_i(\theta)$ for NE-Gap$_i(\xi^\theta)$ and also NE-Gap(θ) for NE-Gap(ξ^θ).

While Definition 3.3 enables us to measure the performance of a single policy, in MARL, most algorithms iterate over a sequence of policies. To measure the performance of a sequence of policies, we use the averaged Nash regret, which is defined in the following.

Definition 3.4. *Given a sequence of M policies $\{\xi(0), \xi(1), \ldots, \xi(M-1)\}$, the averaged Nash regret of agent i and the global averaged Nash regret are defined as*

$$Avg\text{-}Nash\text{-}Regret_i(M) = \frac{1}{M} \sum_{m=0}^{M-1} NE\text{-}Gap_i(\xi(m)),$$

$$Avg\text{-}Nash\text{-}Regret(M) = \max_{i \in \mathcal{N}} Avg\text{-}Nash\text{-}Regret_i(M).$$

Note that a similar concept called "Nash Regret" was previously introduced in Ding et al. [2022], and is defined as

$$\text{Nash-Regret}(M) = \frac{1}{M} \sum_{m=0}^{M-1} \max_{i \in \mathcal{N}} \text{NE-Gap}_i(\xi(m)). \quad (3)$$

By using Jensen's inequality and the fact that the maximum of a set of positive real numbers is less than the summation, we easily have Avg-Nash-Regret$(M) = \Theta(\text{Nash-Regret}(M))$. See Appendix F.4 for the proof. As a result, Avg-Nash-Regret(M) and Nash-Regret(M) have the same rate of convergence (up to a multiplicative constant that depends on the number of agents).

3.1 AN EXAMPLE OF NMPGS

To illustrate the model, we present an extension of classical congestion games [Roughgarden and Tardos, 2004] and distributed welfare games [Marden and Wierman, 2013]. In this example, n agents are located on a traffic network $\mathcal{T} = (\mathcal{V}, \zeta)$, where $\mathcal{V}$ denotes the set of nodes and ζ denotes the set of directed edges with self-loops[1]. The objective of each agent i is to commute from its start node h_i to its destination d_i. In this example, the local state $s_i(t)$ of agent i at time t is its current location (a node $v \in \mathcal{V}$). By choosing a directed edge $(v, u) \in \zeta$ as its local action $a_i(t)$ at time t, agent i will transit to state $s_i(t + 1) = u$ at time $t + 1$. Without the loss of generality, we assume an agent will stay at the same node after it arrives at its destination.

The reward of agent i is defined as $r_i(t) = 0$ if $s_i(t) = d_i$, $r_i(t) = -\bar{\epsilon}$ if $s_i(t + 1) = s_i(t)$, and $r_i(t) = -\bar{\epsilon} - N(a_i(t), t)$ otherwise, where $\bar{\epsilon} > 0$ is a constant and $N(e, t)$ denote the number of agents that chooses edge e at time t. The reward is designed so that the agent incurs a time cost of $\bar{\epsilon}$ for every step spent on its trip and a congestion cost of $N(a_i(t), t)$ depending on the traffic on the edge it travels through. The congestion cost is avoided if the agent chooses to wait at its current location (i.e., $s_i(t+1) = s_i(t)$). Each agent's goal is to maximize its expected discounted cumulative reward $\mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t r_i(t)\right]$.

To see that this congestion game fits in our NMPG framework, consider the following communication network $\mathcal{G}$: agents i and j are neighbors if and only if there exists a global policy ξ such that $\sum_{t=0}^{\infty} \Pr(s_i(t) = s_j(t), s_i(t) \neq d_i, s_j(t) \neq d_j) > 0$. Under this communication network, the transition kernel is completely local because the next state of any agent i is decided completely locally and the local reward of agent i is a function that depends on the 1-hop local states and actions $(s_{\mathcal{N}_i^1}, a_{\mathcal{N}_i^1})$. We provide more discussion of this example and numerical simulations using it in Appendix A.

---

[1]Note that the traffic network and the communication network $\mathcal{G}$ may be different.

4 ALGORITHM DESIGN

We now present a novel algorithm for solving NMPGs. Our approach uses a combination of independent policy gradient (IPG) with localized TD-learning to form a localized actor-critic framework.

4.1 ACTOR: INDEPENDENT POLICY GRADIENT

Suppose that the agents have complete knowledge about the underlying model (e.g., reward function and transition dynamics). Then a popular approach for solving MPGs is to use IPG, which is presented in Algorithm 1 [Leonardos et al., 2022, Zhang et al., 2022b, Ding et al., 2022, Fox et al., 2022, Zhang et al., 2022a].

Algorithm 1 Independent Policy Gradient

1: **Input:** Initialization $\theta_i(0) = 0, \forall i \in \mathcal{N}$.
2: **for** $m = 0, 1, 2, \cdots, M - 1$ **do**
3:    $\theta_i(m + 1) = \theta_i(m) + \beta \nabla_{\theta_i} J_i(\theta(m))$ for all $i \in \mathcal{N}$
4: **end for**

In each round of Algorithm 1, each agent simultaneously updates its policy by implementing gradient ascent (in the policy space) w.r.t. their own objective function (cf. Algorithm 1 Line 3). Notably, to carry out Algorithm 1, each agent only needs to know its own policy. While Algorithm 1 is promising, it is not a model-free algorithm as computing the gradient requires knowledge of the underlying MDP model. This motivates the design of a critic to help estimate the gradient.

4.2 CRITIC: LOCALIZED TD(λ) WITH LINEAR FUNCTION APPROXIMATION

To motivate the design of the critic, we first present an explicit expression of the policy gradient of agent i [Sutton et al., 1999]:

$$\nabla_{\theta_i} J_i(\theta) = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\xi^\theta} \left[\nabla_{\theta_i} \log \xi_i^{\theta_i}(a_i(t)|s_i(t)) \right.$$
$$\left. \times \overline{Q}_i^\theta(s(t), a_i(t)) \right]. \tag{4}$$

Similar versions of policy gradient theorems under different multi-agent settings were previously developed in Zhang et al. [2022a], Mao et al. [2022]. For completeness, we present a proof of Eq. (4) in Appendix F.2.

In view of Eq. (4), to estimate $\nabla_{\theta_i} J_i(\theta)$, the key is to construct an estimate of the averaged Q-function $\overline{Q}_i^\theta$. However, directly estimating the averaged Q-function of agent i requires information about the global state, incurring long-distance communication. To localize the algorithm, we introduce a hyper-parameter $\kappa_c \in \mathbb{N}$, and for each agent, we

learn an approximation of the averaged Q-function (which we refer to as the κ_c-truncated averaged Q-function) using only information in its κ_c-hop neighborhood.

Truncated Averaged Q-functions. Given the non-negative integer κ_c, agent $i \in \mathcal{N}$, and a global policy parameter θ, we define $\mathcal{Q}_i^{\theta, \kappa_c}$ as the class of κ_c-truncated averaged Q-functions w.r.t. $\overline{Q}_i^\theta$. Specifically,

$$\mathcal{Q}_i^{\theta, \kappa_c} = \left\{ \overline{Q}_i^{\theta, \kappa_c} \in \mathbb{R}^{|\mathcal{S}_{N_i^{\kappa_c}}||\mathcal{A}_i|} \ \middle| \ \exists u_i \in \Delta(\mathcal{S}_{-N_i^{\kappa_c}}) \text{ s.t.} \right.$$
$$\overline{Q}_i^{\theta, \kappa_c}(s_{N_i^{\kappa_c}}, a_i) = \mathbb{E}_{s_{-N_i^{\kappa_c}} \sim u_i} \left[\overline{Q}_i^\theta(s_{N_i^{\kappa_c}}, s_{-N_i^{\kappa_c}}, a_i) \right],$$
$$\left. \forall (s_{N_i^{\kappa_c}}, a_i) \in \mathcal{S}_{N_i^{\kappa_c}} \times \mathcal{A}_i \right\}.$$

Note that when $\kappa_c \geq \max_{i,j} \text{dist}(i, j)$, there is essentially no truncation, i.e., any element in $\mathcal{Q}_i^{\theta, \kappa_c}$ is equal to $\overline{Q}_i^\theta$. When $\kappa_c < \max_{i,j} \text{dist}(i, j)$, we have the following *exponential-decay property*. See Appendix F.3 for the proof.

Lemma 4.1. *For any $\kappa_c \in \mathbb{N}$, agent i, and global policy parameter θ, it holds that*

$$\sup_{\overline{Q}_i^{\theta, \kappa_c} \in \mathcal{Q}_i^{\theta, \kappa_c}} \max_{s, a_i} \left| \overline{Q}_i^{\theta, \kappa_c}(s_{N_i^{\kappa_c}}, a_i) - \overline{Q}_i^\theta(s, a_i) \right|$$
$$\leq \frac{2 \min \left(\gamma^{\kappa_c - \kappa_r + 1}, 1 \right)}{1 - \gamma}. \tag{5}$$

In view of Lemma 4.1, the κ_c-truncated averaged Q-function approximates the averaged Q-function (at a geometric rate) as κ_c increases. Therefore, it is enough for the critic to estimate an arbitrary κ_c-truncated averaged Q-function within the class $\mathcal{Q}_i^{\theta, \kappa_c}$. It is worth noting that the use of truncated Q-functions and the exponential-decay property have been widely exploited in the cooperative MARL literature for communication and dimension reduction in recent years [Qu et al., 2020, Gu et al., 2022, Lin et al., 2021]. In this work, we show how to use such an approach in a non-cooperative setting for the first time.

Linear Function Approximation. While using the κ_c-truncated Q-functions enables us to overcome the computational bottleneck as the number of agents increases, there is still the challenge due to the curse of dimensionality. To further reduce the parameter dimension, we use linear function approximation. To be specific, for each $i \in \mathcal{N}$, let $\phi_i : \mathcal{S}_{N_i^{\kappa_c}} \times \mathcal{A}_i \rightarrow \mathbb{R}^{d_i}$ be a feature mapping of agent i. Then, with weight vector $w_i \in \mathbb{R}^{d_i}$, we consider approximating the κ-truncated Q-functions using $\hat{Q}_i(s_{N_i^{\kappa_c}}, a_i, w_i) = \langle \phi_i(s_{N_i^{\kappa_c}}, a_i), w_i \rangle$ for all $(s_{N_i^{\kappa_c}}, a_i)$. Let $\tilde{\phi}_i(s, a_i) = \phi_i(s_{N_i^{\kappa_c}}, a_i)$ for any $i \in \mathcal{N}$, $s \in \mathcal{S}$, and $a_i \in \mathcal{A}_i$. That is, given an agent i, for each pair (s, a_i) of global state and local action, we look at the states of agents in agent i's κ_c-hop neighborhood (i.e., $s_{N_i^{\kappa_c}}$) and agent i's action (i.e., a_i) and assign the vector $\phi(s_{N_i^{\kappa_c}}, a_i)$ to $\tilde{\phi}_i(s, a_i)$. Then agent i's feature matrix Ω_i is defined

to be an $|\mathcal{S}||\mathcal{A}_i| \times d_i$ matrix with its (s, a_i)-th row being $\tilde{\phi}_i^\top(s, a_i)$, where $(s, a_i) \in \mathcal{S} \times \mathcal{A}_i$.

We propose a novel policy evaluation algorithm called localized TD(λ) with linear function approximation, which is presented in Algorithm 2. The algorithm can be viewed as an extension of the classical TD(λ) with linear function approximation [Tsitsiklis and Van Roy, 1997] to the case where we estimate the κ_c-truncated averaged Q-functions using local information.

Algorithm 2 Localized TD(λ) with Linear Function Approximation

---

1: **Input**: Target policy ξ^θ, positive integers K and $\kappa_c \geq \kappa_r$, initializations $w_i(0) = 0$ for all i, step size $\alpha > 0$, $\lambda \in [0, 1)$, and $\epsilon > 0$.
2: Construct ϵ-exploration policy $\hat{\xi}_i(a_i|s_i) = (1 - \epsilon)\xi_i^{\theta_i}(a_i|s_i) + \epsilon/|\mathcal{A}_i|$, for all i, a_i, and s_i.
3: The agents use the joint policy $\hat{\xi} = (\hat{\xi}_1, \hat{\xi}_2, \cdots, \hat{\xi}_n)$ to collect a sequence of samples $\tau = \{(s(t), a(t), r(t))\}_{0 \leq t \leq K-1} \cup \{s(K)\}$
4: **for** $i = 1, 2, \cdots, n$ **do**
5: $\quad \tau|_{(i,\kappa_c)} := \{(s_{\mathcal{N}_i^{\kappa_c}}(t), a_i(t), r_i(t))\}_{0 \leq t \leq K-1} \cup \{s_{\mathcal{N}_i^{\kappa_c}}(K)\}$
6: $\quad$ **for** $t = 0, 1, \cdots, K-1$ **do**
7: $\quad\quad \delta_i(t) = \phi_i(s_{\mathcal{N}_i^{\kappa_c}}(t), a_i(t))^\top w_i(t) - r_i(t) - \gamma\phi_i(s_{\mathcal{N}_i^{\kappa_c}}(t+1), a_i(t+1))^\top w_i(t)$
8: $\quad\quad w_i(t+1) = w_i(t) - \alpha\delta_i(t)\zeta_i^{\kappa_c}(t)$
9: $\quad\quad \zeta_i^{\kappa_c}(t+1) = (\gamma\lambda)\zeta_i^{\kappa_c}(t) + \phi_i(s_{\mathcal{N}_i^{\kappa_c}}(t+1), a_i(t+1))$
10: $\quad$ **end for**
11: **end for**
12: **Return** $\{w_i(K)\}_{i \in \mathcal{N}}$.

Note from Algorithm 2 Line 2 that we use ϵ-exploration policies to ensure exploration in localized TD(λ). Denote the set of all ϵ-exploration policies by Ξ^ϵ. Importantly, agent i requires only the states and the actions of the agents in its κ_c-hop neighborhood to carry out the algorithm, where κ_c can be viewed as a tunable parameter that trades off the communication effort and the accuracy. In particular, the larger κ_c is, the closer the κ_c-truncated averaged Q-function is to the true averaged Q-function, albeit at a cost of requiring more communication among agents.

4.3 LOCALIZED ACTOR-CRITIC

Combining IPG with localized TD(λ), we arrive at a localized actor-critic algorithm for solving NMPGs, which is presented in Algorithm 3.

The algorithm consists of three major steps. First, in Algorithm 3 Line 3, each agent calls localized TD(λ) with linear function approximation for policy evaluation and outputs a

weight vector w_i^m for all $i \in \mathcal{N}$. Then, in Algorithm 3 Lines $4 - 8$, each agent uses the averaged Q-function estimate to iteratively construct an estimate of the independent policy gradient. Specifically, since the independent policy gradient is an expected discounted sum of the averaged Q-functions (cf. Eq. (4)), we essentially construct an estimator $\Delta_i^T(m)$ (cf. Algorithm 3 Line 8) of it by taking average of total T samples $\{\eta_i^t(m)\}_{0 \leq t \leq T-1}$ (cf. Algorithm 3 Line 6). Finally, in Algorithm 3 Line 9, using the estimated gradient, each agent implements an approximate version of the IPG algorithm presented in Algorithm 1.

Compared with Algorithm 1, Algorithm 3 has the following strengths: (1) the algorithm is model-free, (2) due to the use of truncated Q-functions, each agent only requires information from its κ_c-hop neighborhood to carry out the algorithm, which eliminates long-distance communication along the network, and (3) the algorithm, to some extent, overcomes the curse of dimensionality thanks to the use of linear function approximation.

5 ALGORITHM ANALYSIS

We next present the main results of the paper. We formally state our assumptions in Section 5.1 and then present convergence bounds for Algorithms 1, 2, and 3 in Section 5.2. A proof sketch of our main theorems is given in Section 5.3.

5.1 ASSUMPTIONS

We make the following assumptions.

Assumption 5.1. *There exists a decreasing function $\nu : \mathbb{N} \to \mathbb{R}^+$ such that:*

$$\left| \Phi_i(\theta_{N_i^\kappa}, \theta'_{-N_i^\kappa}) - \Phi_i(\theta_{N_i^\kappa}, \theta_{-N_i^\kappa}) \right| \\ \leq \nu(\kappa) \max_{j \in -N_i^\kappa} \left\| \theta'_j - \theta_j \right\|, \, \forall \, \kappa \in \mathbb{N}, \quad (6)$$

where $\Phi_i(\theta)$ is the short-hand notation for $\Phi_i(\xi^\theta)$.

Assumption 5.1 captures the idea that, for each agent, its potential function is less impacted by the agents far away, and can be viewed as a generalization of the decay property of the Q-functions in the existing literature to the networked MPG setting [Qu et al., 2020, Lin et al., 2021, Zhang et al., 2022c]. In the extreme case where κ exceeds the diameter $\max_{i,j} \text{dist}(i, j)$ of the network, we have $\nu(\kappa) = 0$. Note that this assumption is automatically satisfied for our illustrative example in Section 3.1, where changing the policy of an agent will only affect its direct neighbors. In Appendix F.5, we show that this assumption is also satisfied when each local potential function admits a stage-wise representation [Zhang et al., 2022a].

Algorithm 3 Localized Actor-Critic

1: **Input**: Non-negative integers $M, T, K, H, \kappa_c \geq \kappa_r$, and a positive real number $\epsilon > 0$, initializations $\theta_i(0) = 0$ for all i, and $\Delta_i^0(m) = 0$ for all i and m.

2: **for** $m = 0, 1, 2, \cdots, M-1$ **do**

3:     All agents simultaneously execute localized TD(λ) with linear function approximation (with K iterations) to estimate their κ_c-truncated averaged Q-function $T_{\kappa_c}^i \overline{Q}_i^{\theta(m)}$, $i \in \mathcal{N}$, and output weight vectors $\{w_i^m\}_{i \in \mathcal{N}}$.   ▷ Critic Update

4:     **for** $t = 0, 1, \cdots, T-1$ **do**

5:         The agents use the joint policy $\xi^{\theta(m)} = (\xi_1^{\theta_1(m)}, \xi_2^{\theta_2(m)}, \cdots, \xi_n^{\theta_n(m)})$ to collect a sequence of samples $\{(s^t(k), a^t(k))\}_{0 \leq k \leq H-1}$

6:         $\eta_i^t(m) = \sum_{k=0}^{H-1} \gamma^k \nabla_{\theta_i} \log \xi_i^{\theta_i(m)}(a_i^t(k)|s_i^t(k)) \phi_i(s_{\mathcal{N}_i^{\kappa_c}}^t(k), a_i^t(k))^\top w_i^m$

7:         $\Delta_i^{t+1}(m) = \frac{t}{t+1} \Delta_i^t(m) + \frac{1}{t+1} \eta_i^t(m)$

8:     **end for**

9:     $\theta_i(m+1) = \theta_i(m) + \beta \Delta_i^T(m)$ ▷ Actor Update

10: **end for**

Assumption 5.2. *It holds that* $\inf_\theta \min_{s \in \mathcal{S}} d^\theta(s) > 0$, *where we recall that* d^θ *is the discounted state visitation distribution under a softmax policy* ξ^θ

Assumption 5.2 states that every state can be visited with positive probability under any policy, which easily holds when the initial state distribution $\mu(\cdot)$ is supported on the entire state space. This assumption is standard and has been used in, e.g., Zhang et al. [2022a], Agarwal et al. [2021], Mei et al. [2020]. Under Assumption 5.2, we define $D = 1/\inf_\theta \min_{s \in \mathcal{S}} d^\theta(s)$, which is finite.

Assumption 5.3. *There exists a joint policy* ξ *such that the Markov chain* $\{s(t)\}$ *induced by* ξ *is uniformly ergodic.*

Under Assumption 5.3, [Zhang et al., 2022c, Lemma 4] implies a uniform exploration property for the Markov chain $\{(s(t), a(t))\}$ induced by any policy with entries bounded away from zero, which includes ϵ-exploration policy. Therefore, for any $\hat{\xi} \in \Xi^\epsilon$, the Markov chain $\{(s(t), a(t))\}$ induced by $\hat{\xi}$ has a unique stationary distribution, denoted by $\overline{\pi}^{\hat{\xi}} \in \Delta(\mathcal{S} \times \mathcal{A})$, which satisfies $\pi_{\min} := \inf_{\hat{\xi} \in \Xi^\epsilon} \min_{i \in \mathcal{N}} \min_{s_{\mathcal{N}_i^{\kappa_c}}, a_i} \overline{\pi}^{\hat{\xi}}(s_{\mathcal{N}_i^{\kappa_c}}, a_i) > 0$.

While Assumption 5.2, to some extent, already ensures uniform exploration of our policy class, we further impose Assumption 5.3 to deal with the Markovian sampling in Algorithm 3. This type of assumption is standard in the existing literature even for the single-agent setting [Srikant and Ying, 2019, Tsitsiklis and Van Roy, 1997].

Assumption 5.4. *For all* $i \in \mathcal{N}$, *the feature mapping is normalized so that* $\max_{i,s,a_i} \|\tilde{\phi}_i(s, a_i)\| \leq 1$. *In addition, the feature matrix* Ω_i *(the row vectors of which are* $\{\tilde{\phi}_i^\top(s, a_i)\}_{(s,a_i) \in \mathcal{S} \times \mathcal{A}_i}$ *) has linearly independent columns.*

Assumption 5.4 is indeed without loss of generality because neither disregarding dependent features nor performing feature normalization changes the approximation power of the function class [Bertsekas and Tsitsiklis, 1996].

To state our last assumption, let $D^{\hat{\xi}} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}$ be the diagonal matrix with diagonal entries $\{\overline{\pi}^{\hat{\xi}}(s, a)\}_{(s,a) \in \mathcal{S} \times \mathcal{A}}$. Since $D^{\hat{\xi}}$ has strictly positive diagonal entries under Assumption 5.3 and the feature matrix Ω_i has linearly independent columns for all i, we have $\underline{\lambda} := \min_{i \in \mathcal{N}} \inf_{\hat{\xi} \in \Xi^\epsilon} \lambda_{\min}(\Omega_i D^{\hat{\xi}} \Omega_i) > 0$, where $\lambda_{\min}(\cdot)$ returns the smallest eigenvalue of a positive definite matrix. For any $i \in \mathcal{N}$ and $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, let $c_i(\theta) := \min_s \sum_{a_i^* \in \arg\max_{a_i} \overline{Q}_i^\theta(s, a_i)} \xi_i^{\theta_i}(a_i^*|s_i)$.

Assumption 5.5. $c := \inf_{m \geq 0} \min_{1 \leq i \leq N} c_i(\theta(m)) > 0$, *where* $\{\theta(m)\}_{m \geq 0}$ *are policy parameters encountered from the algorithm trajectory (cf. Algorithm 3).*

The inequality stated in Assumption 5.5 is called a non-uniform Łojasiewicz inequality [Zhang et al., 2022a, Mei et al., 2020], which is used to connect the NE-Gap with the gradient of the objective function through gradient domination. This assumption automatically holds in the existing literature when the policy gradient is exact [Zhang et al., 2022a]. However, for Algorithm 3, due to the more challenging model-free setup and the presence of noise in sampling, c is not necessarily strictly positive, which motivates Assumption 5.5 as a means for analytical tractability. Further relaxing this assumption is our immediate future direction. One approach for removing Assumption 5.5 is to regularize the problem (e.g., using log-barrier regularization like in Zhang et al. [2022b]), which prevents the policy generated by IPG from being deterministic, albeit at a cost of introducing an asymptotic bias due to regularization.

5.2 RESULTS

We are now ready to present our main results. We first present the averaged Nash-regret bound of the IPG algorithm (cf. Algorithm 1) as a warm-up, then we present the finite-sample bound of Algorithm 3, which involves a critic

error. Finally, we present a concise bound of the critic estimation error when using our localized TD(λ) with linear function approximation. Given an arbitrary integer κ, let $n(\kappa) := \max_{i \in \mathcal{N}} |N_i^\kappa|$ be the size of the largest κ-hop neighborhood.

Theorem 5.6. *Consider $\{\theta_i(m)\}_{0 \le m \le M-1}$ generated by Algorithm 1. Suppose that Assumptions 5.1, 5.2, and 5.5 are satisfied, and the step size $\beta = \frac{(1-\gamma)^3}{6n(\kappa_G)}$. Then,*

$$Avg\text{-}Nash\text{-}Regret(M)$$
$$\le \mathcal{O}\left(\frac{D}{c} \sqrt{\frac{\max_{j \in \mathcal{N}} |\mathcal{A}_j| n(\kappa_G)(\Phi_{\max} - \Phi_{\min})}{(1-\gamma)^3 M}} \right)$$
$$+ \mathcal{O}\left(\frac{D\sqrt{\max_{j \in \mathcal{N}} |\mathcal{A}_j| \nu(\kappa_G)}}{c(1-\gamma)} \right). \tag{7}$$

The first term on the right-hand side of Eq. (7) goes to zero at a rate of $\mathcal{O}(M^{-1/2})$, which matches with the existing convergence rate of IPG for solving MPGs [Zhang et al., 2022a]. Note that, unlike in existing results, the total number of agents n does not appear in the bound. Instead, we have $n(\kappa_G)$, which captures the impact of network structure. The second term on the right-hand side of Eq. (7) arises because of the relaxation from MPG to NMPG (see Definition 3.1), which decreases with κ_G, and vanishes when $\kappa_G \ge \max_{i,j} \text{dist}(i,j)$.

We next move on to study Algorithm 3.

Theorem 5.7. *Consider $\{\theta_i(m)\}_{0 \le m \le M-1}$ generated by Algorithm 3. Suppose that Assumptions 5.1 – 5.5 are satisfied, and $\beta = \frac{(1-\gamma)^3}{24n(\kappa_G)}$. Then,*

$$\mathbb{E}\left[Avg\text{-}Nash\text{-}Regret(M) \right]$$
$$\le \frac{\sqrt{\max_{j \in \mathcal{N}} |\mathcal{A}_j| D}}{c} \left\{ \mathcal{O}\left(\frac{\sqrt{n(\kappa_G)(\Phi_{\max} - \Phi_{\min})}}{(1-\gamma)^{1.5} M^{1/4}} \right) \right.$$
$$+ \mathcal{O}\left(\frac{\sqrt{\nu(\kappa_G)}}{1-\gamma} \right) + \mathcal{O}\left(\frac{\sqrt{n(\kappa_G)}[1 + (1-\gamma)\epsilon_{critic}]}{(1-\gamma)^2 M^{1/4}} \right)$$
$$+ \mathcal{O}\left(\frac{\sqrt{n(\kappa_G)}\epsilon_{critic}^{1/2}}{(1-\gamma)^{1.5}} \right) + \left. \mathcal{O}\left(\frac{\sqrt{n(\kappa_G)}\gamma^{H/2}}{(1-\gamma)^2} \right) \right\}, \tag{8}$$

where ϵ_{critic} stands for the critic estimation error in policy evaluation:

$$\epsilon_{critic} = \sup_{\theta, i} \mathbb{E}^{1/2}\left[\sup_{s, a_i} \left| \overline{Q}_i^\theta(s, a_i) - \phi_i(s_{\mathcal{N}_i^{\kappa_c}}, a_i)^\top w_i^\theta \right|^2 \right].$$

The first two terms on the right-hand side of Eq. (8) are analogous to the two terms on the right-hand side of the IPG error bounds presented in Theorem 5.6. The last 4 terms are approximation errors for the independent policy gradient, which (in the order as they appear in the bound) consist

of a localization error, an error incurred by using a finite sum (Algorithm 3 Line 6) to approximate an infinite sum (cf. Eq. (4)), a critic error, and an error incurred by using a finite average (Algorithm 3 Lines 4 − 8) to approximate an expectation (cf. Eq. (4)).

To establish an overall sample complexity bound of Algorithm 3, we need to specify how the critic error decays as a function of the number of iterations in localized TD(λ) with linear function approximation, which is presented in the following.

Theorem 5.8. *Consider $\{w_i(K)\}_{i \in \mathcal{N}}$ generated by Algorithm 2. Suppose that Assumption 5.3 is satisfied. Then, with appropriately chosen step size α (see Appendix D for the explicit requirements) and large enough K, we have*

$$\epsilon_{critic} \le \mathcal{O}(1 - (1-\gamma)\underline{\lambda}\alpha)^{\frac{K}{2}} + \mathcal{O}\left[\frac{\alpha \log(1/\alpha)}{(1-\gamma)\underline{\lambda}} \right]^{1/2}$$
$$+ \mathcal{O}\left(\frac{\epsilon_{app}}{\pi_{\min}(1-\gamma)} \right) + \mathcal{O}\left(\frac{\gamma^{\kappa_c - \kappa_r}}{1-\gamma} \right)$$
$$+ \mathcal{O}\left(\frac{n\epsilon}{(1-\gamma)^2} \right), \tag{9}$$

where ϵ_{app} stands for the function approximation error. See Appendix D for the explicit definition.

The first two terms on the right-hand side of Eq. (9) represent the convergence bias (which has geometric convergence rate) and the variance (which decreases with the step size α), and their behaviors agree with existing results on stochastic approximation [Srikant and Ying, 2019, Chen et al., 2022]. The third term arises from using linear function approximation and vanishes in the tabular setting where we use a complete basis. The fourth term represents the error between the averaged Q-function and the κ_c-truncated averaged Q-function, which is introduced to overcome the scalability issue when the number of agents increases. Note that the fourth term decays exponentially with the choice of κ_c, and vanishes when κ_c is greater than the diameter (i.e., $\max_{i,j} \text{dist}(i,j)$) of the network. The last term arises because of using ϵ-exploration behavior policies to ensure sufficient exploration.

Combining Theorem 5.7 and Theorem 5.8 leads to the following sample complexity bound.

Corollary 5.9. *To achieve $\mathbb{E}[Avg\text{-}Nash\text{-}Regret(M)] \le \tilde{\epsilon} + \mathcal{E}_{EX} + \mathcal{E}_{FA} + \mathcal{E}_{LO}$, the sample complexity is $\tilde{\mathcal{O}}(\tilde{\epsilon}^{-4})$, where $\mathcal{E}_{EX}$ stands for the induced error from exploration (cf. the last term on the right-hand side of Eq. (9)), $\mathcal{E}_{FA}$ stands for the function approximation error (cf. the third term on the right-hand side of Eq. (9)), and $\mathcal{E}_{LO}$ stands for the induced error from localization (cf. the summation of the second last term on the right-hand side of Eq. (9) and the third term on the right-hand side of Eq. (8)).*

In Corollary 5.9 The presence of $\mathcal{E}_{\mathrm{EX}} + \mathcal{E}_{\mathrm{FA}} + \mathcal{E}_{\mathrm{LO}}$ are due to the fundamental limit of the problem, such as the approximation power of function class, using truncated averaged Q-functions to approximate global averaged Q-functions, and using "soft" policies to ensure exploration.

In single-agent RL, popular algorithms such as Q-learning and natural actor-critic are known to achieve $\tilde{\mathcal{O}}(\tilde{\epsilon}^{-2})$ sample complexity [Qu and Wierman, 2020, Lan, 2022]. While we study the more challenging setting of using localized algorithms to solve MARL problems, it is an interesting direction to investigate whether there is a fundamental gap. In addition, while Localized Actor-Critic (cf. Algorithm 3) is an independent learning algorithm, our theoretical results require all agents to follow the same learning dynamics, which suggests some implicit coordination among the agents. Although this is common in the existing literature [Leonardos et al., 2022, Ding et al., 2022, Zhang et al., 2022a], developing completely independent learning dynamics is an interesting future direction.

5.3 PROOF SKETCH

Analysis of the Actor. At a high level, we use a Lyapunov approach to analyze the policy update, where the potential function is a natural choice of the Lyapunov function. The key is to bound $\Phi_i(\theta(m+1)) - \Phi_i(\theta(m))$, $i \in \mathcal{N}$, in each iteration using the gradient of objective function $J_i(\cdot)$, which is related to NE-Gap of agent i through the non-uniform Łojasiewicz inequality [Zhang et al., 2022a, Mei et al., 2020]. To exploit the network structure and to remove the raw dependence on the total number of agents in the NMPG setting, instead of directly bounding $\Phi_i(\theta(m+1)) - \Phi_i(\theta(m))$, we perform the following decomposition:

$$
\begin{aligned}
&\Phi_i(\theta(m+1)) - \Phi_i(\theta(m) \\
&= \underbrace{\left[\Phi_i(\theta_{N_i^{\kappa_G}}(m+1), \theta_{-N_i^{\kappa_G}}(m)) - \Phi_i(\theta(m)) \right]}_{(a)} \\
&\quad + \underbrace{\left[\Phi_i(\theta(m+1)) - \Phi_i(\theta_{N_i^{\kappa_G}}(m+1), \theta_{-N_i^{\kappa_G}}(m)) \right]}_{(b)}.
\end{aligned}
$$

The term (a) captures the policy change of the agents inside the κ_G-hop neighborhood of agent i, and the first step of bounding it is to use the smoothness property of the potential function, which is similar to that of Zhang et al. [2022a]. However, unlike existing analysis of IPG, we also need to bound the error in approximating the gradient, which can be decomposed into three error terms:

e_1: error due to estimating the averaged Q-function, which is exactly the critic error;

e_2: error due to the randomness in the trajectory sampling (see Algorithm 3 Lines $4 - 8$), which has zero mean;

e_3: error resulted from truncating the sample trajectory at

horizon H (see Algorithm 3 Lines 6), which decays exponentially with H.

Term (b) results from the policy change of agents outside the κ_G-hop neighborhood of agent i, and is a decreasing function of κ_G (cf. Assumption 5.1).

Analysis of the Critic. The critic is designed to perform policy evaluation of a softmax policy ξ^θ using localized TD(λ) with linear function approximation. Similar to Chen et al. [2022], Srikant and Ying [2019], we formulate localized TD(λ) as a stochastic approximation algorithm and again use a Lyapunov approach to establish the finite-sample bound of the difference between $w_i(K)$ and w_i^θ, where w_i^θ is the solution to a properly defined projected Bellman equation associated with agent i.

The challenge lies in bounding the difference between the Q-function associated with the weight vector w_i^θ (denoted by $Q(w_i^\theta)$) and the true averaged Q-function $\overline{Q}_i^\theta$ of policy ξ^θ, which we decompose into a function approximation error, an error due to using ϵ-exploration policy, and an error due to truncating the averaged Q-function at its κ_c-hop neighborhood, and bound them separately. To achieve that, we develop a novel approach involving the construction of a "sub-chain", which is an auxiliary Markov chain with state space $\mathcal{S}_{\mathcal{N}_i^{\kappa_c}} \times \mathcal{A}_i$. See Appendix D for more details.

6 CONCLUSION

We study MARL in the context of MPGs and introduce a networked structure that allows agents to learn equilibria using local information. In particular, we develop a localized actor-critic framework for minimizing the averaged Nash regret of NMPGs. Importantly, the algorithm is scalable and uses function approximation. We provide finite-sample convergence bounds to theoretically support our proposed algorithm and conduct numerical simulations to demonstrate its empirical effectiveness.

An immediate future direction is to investigate whether there is a fundamental gap in the convergence rates between localized MARL algorithms and single-agent RL algorithms. It is also interesting to see if localized algorithms (with provable guarantees) can be designed to solve other classes of games beyond NMPGs.

References

Alekh Agarwal, Sham M Kakade, Jason D Lee, and Gaurav Mahajan. On the theory of policy gradient methods: Optimality, approximation, and distribution shift. *The Journal of Machine Learning Research*, 22(1):4431–4506, 2021.

Dimitri Bertsekas and John N Tsitsiklis. *Neuro-dynamic programming*. Athena Scientific, 1996.

Jalaj Bhandari, Daniel Russo, and Raghav Singal. A Finite Time Analysis of Temporal Difference Learning With Linear Function Approximation. In *Conference On Learning Theory*, pages 1691–1692, 2018.

Justin A Boyan. Least-Squares Temporal Difference Learning. In *Proceedings of the Sixteenth International Conference on Machine Learning*, pages 49–56, 1999.

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

Zaiwei Chen, Sheng Zhang, Thinh T Doan, John-Paul Clarke, and Siva Theja Maguluri. Finite-sample analysis of nonlinear stochastic approximation with applications in reinforcement learning. *Automatica*, 146:110623, 2022.

Gal Dalal, Gugan Thoppe, Balázs Szörényi, and Shie Mannor. Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning. In *Conference On Learning Theory*, pages 1199–1233. PMLR, 2018.

W Davis Dechert and SI O'Donnell. The stochastic lake game: A numerical solution. *Journal of Economic Dynamics and Control*, 30(9-10):1569–1587, 2006.

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

Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

Roy Fox, Stephen M Mcaleer, Will Overman, and Ioannis Panageas. Independent natural policy gradient always converges in markov potential games. In *International Conference on Artificial Intelligence and Statistics*, pages 4414–4425. PMLR, 2022.

Bolin Gao and Lacra Pavel. On the properties of the softmax function with application in game theory and reinforcement learning. *Preprint arXiv:1704.00805*, 2017.

Haotian Gu, Xin Guo, Xiaoli Wei, and Renyuan Xu. Mean-field controls with q-learning for cooperative marl: convergence and complexity analysis. *SIAM Journal on Mathematics of Data Science*, 3(4):1168–1196, 2021.

Haotian Gu, Xin Guo, Xiaoli Wei, and Renyuan Xu. Mean-Field Multi-Agent Reinforcement Learning: A Decentralized Network Approach. *Available at SSRN 3900139*, 2022.

Tommi Jaakkola, Michael I Jordan, and Satinder P Singh. Convergence of stochastic iterative dynamic programming algorithms. In *Advances in neural information processing systems*, pages 703–710, 1994.

Fivos Kalogiannis, Ioannis Anagnostides, Ioannis Panageas, Emmanouil-Vasileios Vlatakis-Gkaragkounis, Vaggos Chatziafratis, and Stelios Stavroulakis. Efficiently computing nash equilibria in adversarial team markov games. *Preprint arXiv:2208.02204*, 2022.

Guanghui Lan. Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical programming*, pages 1–48, 2022.

Zachary J Lee, Tongxin Li, Steven H Low, and Sunash B Sharma. Systems and methods for adaptive ev charging, July 5 2022. US Patent 11,376,981.

Stefanos Leonardos, Will Overman, Ioannis Panageas, and Georgios Piliouras. Global convergence of multi-agent policy gradient in markov potential games. In *International Conference on Learning Representations*, 2022.

David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

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

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

Sergio Valcarcel Macua, Javier Zazo, and Santiago Zazo. Learning parametric closed-loop policies for markov potential games. In *International Conference on Learning Representations*, 2018.

Weichao Mao, Lin Yang, Kaiqing Zhang, and Tamer Basar. On improving model-free algorithms for decentralized multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 15007–15049. PMLR, 2022.

Jason R Marden and Adam Wierman. Distributed welfare games. *Operations Research*, 61(1):155–168, 2013.

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

Washim Uddin Mondal, Vaneet Aggarwal, and Satish Ukkusuri. On the near-optimality of local policies in large cooperative multi-agent reinforcement learning. *Transactions on Machine Learning Research*, 2022a. ISSN 2835-8856.

Washim Uddin Mondal, Vaneet Aggarwal, and Satish V Ukkusuri. Can mean field control (mfc) approximate cooperative multi agent reinforcement learning (marl) with non-uniform interaction? In *Uncertainty in Artificial Intelligence*, pages 1371–1380. PMLR, 2022b.

Guannan Qu and Adam Wierman. Finite-time analysis of asynchronous stochastic approximation and Q-learning. In *Conference on Learning Theory*, pages 3185–3205. PMLR, 2020.

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

Tim Roughgarden and Éva Tardos. Bounding the inefficiency of equilibria in nonatomic congestion games. *Games and economic behavior*, 47(2):389–403, 2004.

Yuanyuan Shi, Guannan Qu, Steven Low, Anima Anandkumar, and Adam Wierman. Stability constrained reinforcement learning for real-time voltage control. In *2022 American Control Conference (ACC)*, pages 2715–2721. IEEE, 2022.

David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

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

Richard S Sutton, David McAllester, Satinder Singh, and Yishay Mansour. Policy gradient methods for reinforcement learning with function approximation. *Advances in neural information processing systems*, 12, 1999.

John N Tsitsiklis. Asynchronous stochastic approximation and Q-learning. *Machine learning*, 16:185–202, 1994.

John N Tsitsiklis and Benjamin Van Roy. An Analysis of Temporal-Difference Learning with Function Approximation. *IEEE Transactions on Automatic Control*, 42(5): 674–690, 1997.

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

Huizhen Yu and Dimitri P Bertsekas. Convergence results for some temporal difference methods based on least squares. *IEEE Transactions on Automatic Control*, 54(7): 1515–1531, 2009.

Kaiqing Zhang, Zhuoran Yang, and Tamer Başar. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Handbook of reinforcement learning and control*, pages 321–384, 2021.

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

Runyu Zhang, Jincheng Mei, Bo Dai, Dale Schuurmans, and Na Li. On the global convergence rates of decentralized softmax gradient play in markov potential games. In *Advances in Neural Information Processing Systems*, 2022a.

Runyu Zhang, Zhaolin Ren, and Na Li. Gradient play in stochastic games: Stationary points and local geometry. In *25th International Symposium on Mathematical Theory of Networks and Systems (MTNS 2022)*, 2022b.

Yizhou Zhang, Guannan Qu, Pan Xu, Yiheng Lin, Zaiwei Chen, and Adam Wierman. Global Convergence of Localized Policy Iteration in Networked Multi-Agent Reinforcement Learning. *Preprint arXiv:2211.17116*, 2022c.

A MARKOV CONGESTION GAME EXAMPLE

In this section, we provide the detailed settings and proofs of the Markov congestion game example in Section 3.1. Since our Markov congestion game is an extension of the classic potential game, it is natural to conjecture that it is a MPG because every single time step is a one-shot potential game. However, this intuition does not hold in general (see counterexamples in Leonardos et al. [2022]). To show that the Markov congestion game is actually a MPG, we need the critical observation that the transition probability of each agent in this congestion game is completely local, i.e., the next local state of agent is determined by its current local state and local action. In Theorem A.1, we show that completely local transition probability is a sufficient condition for a networked Markov game that is potential at every single step to be an NMPG.

Theorem A.1. *If the networked MDP satisfies that for an arbitrary fixed global state, the one-round game is a potential game, i.e., there exists a global potential function $\phi : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ such that*

$$r_i(s_i, a_i, s_{-i}, a_{-i}) - r_i(s_i', a_i', s_{-i}, a_{-i}) = \phi(s_i, a_i, s_{-i}, a_{-i}) - \phi(s_i', a_i', s_{-i}, a_{-i}),$$

and the transition probability of each agent is completely local, i.e., $s_i(t+1) \sim P_i(\cdot \mid s_i(t), a_i(t))$, then the networked MDP is an Markov potential game, i.e., there exists a potential function $\Phi : \Pi \times \mathcal{S} \to \mathbb{R}$ such that

$$V_i^{\xi_i, \xi_{-i}}(s) - V_i^{\xi_i', \xi_{-i}}(s) = \Phi^{\xi_i, \xi_{-i}}(s) - \Phi^{\xi_i', \xi_{-i}}(s).$$

We defer the proof of Theorem A.1 to Appendix A.2. Note that the Markov congestion game satisfies the assumptions of Theorem A.1 because at each time step t, we can set the potential function ϕ as

$$\phi(s, a) = -\frac{1}{2} \sum_{e \in \zeta} N(e, t) \left(N(e, t) - 1 \right),$$

where $N(e, t)$ (the number of agents that choose edge e) is decided by the global state/action pair (s, a). Since MPG is a special case of NMPG, we know the Markov congestion game is an NMPG.

A.1 SIMULATION RESULTS

In the numerical simulation, we consider a problem instance with 12 agents moving from 4 different start nodes to the same destination node on an acyclic graph (see Figure 1). Specifically, for $i \in \{1, 2, \ldots, 12\}$, agent i travels from start node $b_{\lceil i/3 \rceil}$ to the destination d. The local state space of each agent i is the set of all possible locations i can visit. For example, the local state space of agent 4 (starts from b_2 and goes to d) contains 4 locations $\{b_2, c_1, c_2, d\}$. Since the maximum out degree of each node in this example is 2, the local action space of agent i contains 3 actions $\{0, 1, 2\}$, where 0 means "wait for one step at the current node" and 1(2) means "go through the first(second) out edge". When a node only has one out edge (e.g., node c_1), local action 2 is treated as action 1. We set the time elapse cost $\epsilon = 0.5$ in this simulation.
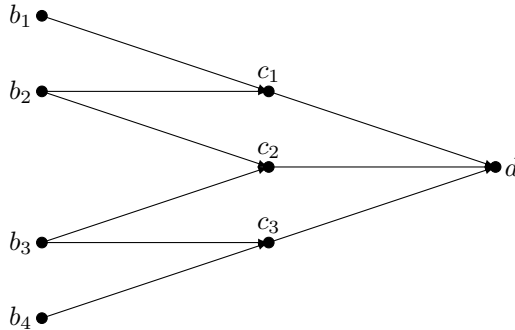

Figure 1: Illustration of the simulation setting. On this directed acyclic graph, each agent $i \in \{1, 2, \ldots, 12\}$ travels from start node $b_{\lceil i/3 \rceil}$ to the same destination d. The policy of each agent is a mapping from its local state (its current location) to the distributions of local actions (pass through an outgoing edge or wait for one time step).

We simulate our Localized Actor-Critic algorithm (Algorithm 1). For the policy evaluation subroutine, we adopt localized TD(0) with linear function approximation and $\kappa_c = 1$. For each agent i, we use the one-hot encoding of 1-hop state $s_{\mathcal{N}_i^1}$
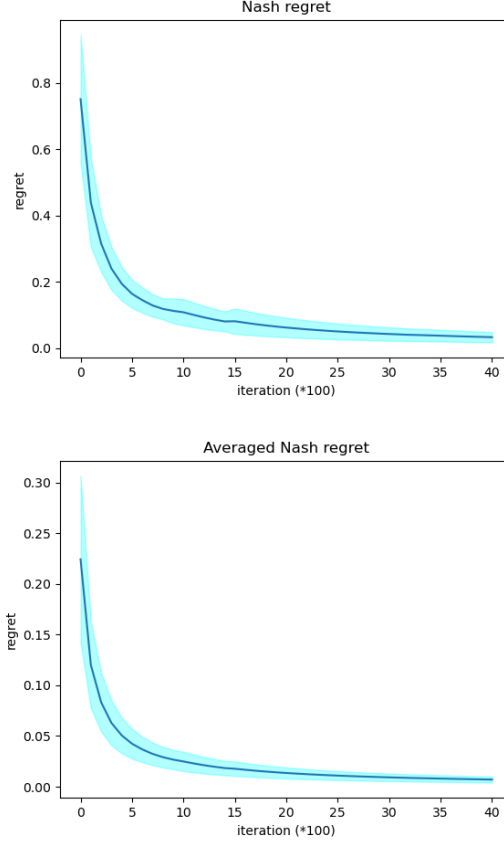
Figure 2: Nash regret (left) and averaged Nash regret (right) of Localized Actor-Critic (Algorithm 1). The blue curve represents the mean and the cyan area represents the standard deviation over 10 random initial policies.

and local action a_i as the feature vector $\phi_i(s_{\mathcal{N}_i^1}, a_i)$. During training, we evaluate the Nash gap NE-Gap$_i(\xi)$ by fixing the local policies of all agents except i and use policy gradient to learn the best response of agent i under other agents' policy profile. We list the detailed choice of hyperparameters in Appendix A.3 and present the code in `https://github.com/yihenglin97/Networked_MPG_Traffic_Game`. We plot the Nash regret curve and the averaged Nash regret curve in Figure 1. The results show both the Nash regret and averaged Nash regret converge to zero as learning proceeds.

A.2 PROOF OF THEOREM A.1

Under the completely local transition probability, the distribution of local state of agent i at time t only depends the initial state $s_i(0)$. We denote this local distribution as $d_i^{\xi_i}[t](\cdot \mid s_i(0))$ and the joint distribution as $d_{\mathcal{N}}^{\xi}[t](\cdot \mid s(0))$. We define the potential function as

$$\Phi^{\xi}(s) = \sum_{t=0}^{\infty} \gamma^t \sum_{s' \in \mathcal{S}, a' \in \mathcal{A}} d_{\mathcal{N}}^{\xi}[t]((s', a') \mid s(0) = s) \cdot \phi(s', a').$$

The potential function can be rewritten as

$$
\Phi^{\xi_i,\xi_{-i}}(s)
$$

$$
= \sum_{t=0}^{\infty} \gamma^t \sum_{s',a'} d_{-i}^{\xi_{-i}}[t]((s'_{-i},a'_{-i}) \mid s_{-i}(0)=s_{-i}) \cdot d_i^{\xi_i}[t]((s'_i,a'_i) \mid s_i(0)=s_i) \cdot \phi(s',a'_i,a'_{-i})
$$

$$
= \sum_{t=0}^{\infty} \gamma^t \sum_{s'_{-i},a'_{-i}} d_{-i}^{\xi_{-i}}[t]((s'_{-i},a'_{-i}) \mid s_{-i}(0)=s_{-i})
$$

$$
\times \sum_{s'_i,a'_i} d_i^{\xi_i}[t]((s'_i,a'_i) \mid s_i(0)=s_i) \cdot \phi(s',a'_i,a'_{-i}).
$$

Note that

$$
\sum_{s'_i,a'_i} \left(d_i^{\xi_i}[t]((s'_i,a'_i) \mid s_i(0)=s_i) - d_i^{\xi'_i}[t]((s'_i,a'_i) \mid s_i(0)=s_i) \right) \cdot \phi(s',a'_i,a'_{-i})
$$

$$
= \sum_{s'_i,a'_i} \left(d_i^{\xi_i}[t]((s'_i,a'_i) \mid s_i(0)=s_i) - d_i^{\xi'_i}[t]((s'_i,a'_i) \mid s_i(0)=s_i) \right) \cdot r_i(s',a'_i,a'_{-i}).
$$

Therefore, we have

$$
\Phi^{\xi_i,\xi_{-i}}(s) - \Phi^{\xi'_i,\xi_{-i}}(s) = V_i^{\xi_i,\xi_{-i}}(s) - V_i^{\xi'_i,\xi_{-i}}(s).
$$

A.3 CHOICE OF HYPERPARAMETERS

In numerical experiments, we use the following hyperparameters: Localized TD(λ) with Linear Function Approximation (Algorithm 2): $\lambda = 0, K = 10, \alpha = 0.001, \epsilon = 0$. Localized Actor-Critic (Algorithm 3): $M = 4000, T = 1, H = 15, K = 10, \kappa_c = 1, \beta = 0.001$. We construct the feature $\phi_i(s_{\mathcal{N}_i^{\kappa_c}}, a_i)$ by concatenating the one-hot encodings for $s_j, j \in \mathcal{N}_i^{\kappa_c}$ and a_i.

B PROOF OF THEOREM 5.6

For any $\kappa \leq \kappa_G$ and $\beta \leq 1/L(\kappa)$, we have by Lemma B.3 that

$$
\Phi_i(\theta(M)) - \Phi_i(\theta(0)) = \sum_{m=0}^{M-1} [\Phi_i(\theta(m+1)) - \Phi_i(\theta(m))]
$$

$$
\geq \sum_{m=0}^{M-1} \left[\frac{\beta}{2} \left\| \nabla_{\theta_{N_i^\kappa}} \Phi_i(\theta(m)) \right\|^2 - \frac{\sqrt{2}\nu(\kappa)\beta}{(1-\gamma)^2} \right].
$$

It follows that

$$
\frac{1}{M} \sum_{m=0}^{M-1} \left\| \nabla_{\theta_{N_i^\kappa}} \Phi_i(\theta(m)) \right\|^2 \leq \frac{2}{M\beta}[\Phi_i(\theta(M)) - \Phi_i(\theta(0))] + \frac{2\sqrt{2}\nu(\kappa)}{M(1-\gamma)^2}
$$

$$
\leq \frac{2(\Phi_{\max} - \Phi_{\min})}{M\beta} + \frac{2\sqrt{2}\nu(\kappa)}{(1-\gamma)^2},
$$

where the last line follows from Lemma F.7. Note that

$$
\left\| \nabla_{\theta_{N_i^\kappa}} \Phi_i(\theta(m)) \right\|^2 = \left\| \nabla_{\theta_{N_i^\kappa}} J_i(\theta(m)) \right\|^2
$$

$$
\geq \left\| \nabla_{\theta_i} J_i(\theta(m)) \right\|^2
$$

$$
\geq \frac{c(\theta(m))^2}{\max_{j \in \mathcal{N}} |\mathcal{A}_j| D(\theta(m))^2} \text{NE-Gap}_i(\theta(m))^2 \qquad \text{(Lemma B.4)}
$$

$$
\geq \frac{c^2 \text{NE-Gap}_i(\theta(m))^2}{\max_{j \in \mathcal{N}} |\mathcal{A}_j| D^2}. \qquad (c = \inf_\theta \min_i c_i(\theta) \text{ and } D = \sup_\theta D(\theta))
$$

Therefore, we have

$$\frac{1}{M}\sum_{m=0}^{M-1}\text{NE-Gap}_i(\theta(m))^2 \leq \frac{\max_{j\in\mathcal{N}}|\mathcal{A}_j|D^2}{c^2}\left(\frac{2(\Phi_{\max}-\Phi_{\min})}{M\beta} + \frac{2\sqrt{2}\nu(\kappa)}{(1-\gamma)^2}\right).$$

It then follows from the definition of Avg-Nash-Regret$_i(M)$ that

$$\begin{aligned}
\text{Avg-Nash-Regret}_i(M) &= \frac{1}{M}\sum_{m=0}^{M-1}\text{NE-Gap}_i(\theta(m)) \\
&\leq \left(\frac{1}{M}\sum_{m=0}^{M-1}\text{NE-Gap}_i(\theta(m))^2\right)^{1/2} \qquad \text{(Jensen's inequality)} \\
&\leq \frac{\max_{j\in\mathcal{N}}|\mathcal{A}_j|^{1/2}D}{c}\left(\frac{2(\Phi_{\max}-\Phi_{\min})}{M\beta} + \frac{2\sqrt{2}\nu(\kappa)}{(1-\gamma)^2}\right)^{1/2} \\
&\leq \frac{2\max_{j\in\mathcal{N}}|\mathcal{A}_j|^{1/2}D}{c}\left(\frac{(\Phi_{\max}-\Phi_{\min})^{1/2}}{M^{1/2}\beta^{1/2}} + \frac{\nu(\kappa)^{1/2}}{(1-\gamma)}\right),
\end{aligned}$$

where the last line follows from $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$ for any $a, b \geq 0$. Since the RHS of the previous inequality is not a function of i, choosing $\kappa = \kappa_G$ and $\beta = 1/L(\kappa_G) = \frac{(1-\gamma)^3}{6n(\kappa_G)}$, and we complete the proof.

B.1 SUPPORTING LEMMAS

Lemma B.1. *The following inequality holds for all θ, θ', and $i \in \mathcal{N}$:*

$$\left\|\nabla_{\theta_i}J_i(\theta) - \nabla_{\theta_i'}J_i(\theta')\right\|_1 \leq \frac{6}{(1-\gamma)^3}\sum_{j\in\mathcal{N}}\left\|\theta_j - \theta_j'\right\|.$$

Proof of Lemma B.1. Using Lemma F.12, and we have

$$\begin{aligned}
&\left\|\nabla_{\theta_i}J_i(\theta) - \nabla_{\theta_i'}J_i(\theta')\right\|_1 \\
&= \frac{1}{1-\gamma}\sum_{s_i,a_i}\left|\sum_{s_{-i}}\left(d^\theta(s_i,s_{-i})\xi_i^{\theta_i}(a_i|s_i)\overline{A}_i^\theta(s,a_i) - d^{\theta'}(s_i,s_{-i})\xi_i^{\theta_i'}(a_i|s_i)\overline{A}_i^{\theta'}(s,a_i)\right)\right| \\
&= \frac{1}{1-\gamma}\sum_{s_i,a_i}\left|\sum_{s_{-i},a_{-i}}\left(d^\theta(s)\xi^\theta(a|s)A_i^\theta(s,a) - d^{\theta'}(s)\xi^{\theta'}(a|s)A_i^{\theta'}(s,a)\right)\right| \\
&\leq \frac{1}{1-\gamma}\sum_{s,a}\left|d^\theta(s)\xi^\theta(a|s)A_i^\theta(s,a) - d^{\theta'}(s)\xi^{\theta'}(a|s)A_i^{\theta'}(s,a)\right| \\
&\leq \frac{1}{1-\gamma}\left(\sum_{s,a}\left|d^\theta(s)\xi^\theta(a|s) - d^{\theta'}(s)\xi^{\theta'}(a|s)\right|\left|A_i^\theta(s,a)\right| \right. \\
&\qquad\qquad \left. + \sum_{s,a}d^{\theta'}(s)\xi^{\theta'}(a|s)\left|A_i^\theta(s,a) - A_i^{\theta'}(s,a)\right|\right) \\
&\leq \frac{1}{1-\gamma}\left(\sum_{s,a}\frac{1}{1-\gamma}\left|d^\theta(s)\xi^\theta(a|s) - d^{\theta'}(s)\xi^{\theta'}(a|s)\right| + \max_{s,a}\left|A_i^\theta(s,a) - A_i^{\theta'}(s,a)\right|\right).
\end{aligned}$$

Lemma 32 and Corollary 35 of Zhang et al. [2022a] imply

$$\left|A_i^\theta(s,a) - A_i^{\theta'}(s,a)\right| \leq \frac{2}{(1-\gamma)^2}\max_s\sum_{a\in\mathcal{A}}\left|\xi^\theta(a|s) - \xi^{\theta'}(a|s)\right|$$

$$\frac{1}{1-\gamma}\sum_{s,a}\left|d^\theta(s)\xi^\theta(a|s)\right| \leq \frac{1}{(1-\gamma)^2}\max_s\sum_{a\in\mathcal{A}}\left|\xi^\theta(a|s) - \xi^{\theta'}(a|s)\right|.$$

Therefore, we have

$$\|\nabla_{\theta_i} J_i(\theta) - \nabla_{\theta_i} J_i(\theta')\|_1 \leq \frac{3}{(1-\gamma)^3} \max_s \sum_{a \in \mathcal{A}} \left| \xi^\theta(a|s) - \xi^{\theta'}(a|s) \right|$$

$$= \frac{3}{(1-\gamma)^3} \max_s \sum_{i \in \mathcal{N}} \sum_{a_i \in \mathcal{A}_i} \left| \xi_i^{\theta_i}(a_i|s_i) - \xi_i^{\theta_i'}(a_i|s_i) \right| \qquad \text{(Lemma F.15)}$$

$$\leq \frac{6}{(1-\gamma)^3} \sum_{i \in \mathcal{N}} \|\theta_i - \theta_i'\|.$$

where the last line follows from [Zhang et al., 2022a, Corollary 37]. □

Lemma B.2. *The following inequality holds for all $\kappa \leq \kappa_G$, $\theta = (\theta_{N_i^\kappa}, \theta_{-N_i^\kappa})$, and $\theta' = (\theta'_{N_i^\kappa}, \theta_{-N_i^\kappa})$:*

$$\left\| \nabla_{\theta_{N_i^\kappa}} \Phi_i(\theta) - \nabla_{\theta_{N_i^\kappa}} \Phi_i(\theta') \right\| \leq L(\kappa) \left\| \theta_{N_i^\kappa} - \theta'_{N_i^\kappa} \right\|,$$

where $L(\kappa) = \frac{6n(\kappa)}{(1-\gamma)^3}$.

Proof of Lemma B.2. Using the definition of NMPG (cf. Definition 3.1) and we have

$$\left\| \nabla_{\theta_{N_i^\kappa}} \Phi_i(\theta_{N_i^\kappa}, \theta_{-N_i^\kappa}) - \nabla_{\theta_{N_i^\kappa}} \Phi_i(\theta'_{N_i^\kappa}, \theta_{-N_i^\kappa}) \right\|^2$$

$$= \sum_{j \in N_i^\kappa} \left\| \nabla_{\theta_j} J_j(\theta_{N_i^\kappa}, \theta_{-N_i^\kappa}) - \nabla_{\theta_j} J_j(\theta'_{N_i^\kappa}, \theta_{-N_i^\kappa}) \right\|^2$$

$$\leq \sum_{j \in N_i^\kappa} \left\| \nabla_{\theta_j} J_j(\theta_{N_i^\kappa}, \theta_{-N_i^\kappa}) - \nabla_{\theta_j} J_j(\theta'_{N_i^\kappa}, \theta_{-N_i^\kappa}) \right\|_1^2$$

$$\leq \sum_{j \in N_i^\kappa} \left(\frac{6}{(1-\gamma)^3} \sum_{j' \in N_i^\kappa} \|\theta_{j'} - \theta'_{j'}\| \right)^2 \qquad \text{(Lemma B.1)}$$

$$= \frac{36 |N_i^\kappa|}{(1-\gamma)^6} \left(\sum_{j' \in N_i^\kappa} \|\theta_{j'} - \theta'_{j'}\| \right)^2$$

$$\leq \frac{36 |N_i^\kappa|^2}{(1-\gamma)^6} \sum_{j' \in N_i^\kappa} \|\theta_{j'} - \theta'_{j'}\|^2 \qquad \text{(Cauchy-Schwarz inequality)}$$

$$\leq \frac{36 n(\kappa)^2}{(1-\gamma)^6} \left\| \theta_{N_i^\kappa} - \theta'_{N_i^\kappa} \right\|^2,$$

where the last line follows from the definition of $n(\kappa)$. Taking square root on both sides of the previous inequality and we have the desired result. □

Lemma B.3. *Consider $\{\theta_i(m)\}_{0 \leq m \leq M}$ generated by Algorithm 1. Suppose that $\kappa \leq \kappa_G$ and $\beta \leq \frac{1}{L(\kappa)}$, where $L(\kappa)$ is defined in Lemma B.2. Then we have for any $i \in \mathcal{N}$ and $m \geq 0$ that*

$$\Phi_i(\theta(m+1)) - \Phi_i(\theta(m)) \geq \frac{\beta}{2} \left\| \nabla_{\theta_{N_i^\kappa}} \Phi_i(\theta(m)) \right\|^2 - \frac{\sqrt{2}\nu(\kappa)\beta}{(1-\gamma)^2}.$$

Proof of Lemma B.3. For any $i \in \mathcal{N}$ and $m \geq 0$, we have

$$\Phi_i(\theta(m+1)) - \Phi_i(\theta(m)$$
$$= \left[\Phi_i(\theta(m+1)) - \Phi_i(\theta_{N_i^\kappa}(m+1), \theta_{-N_i^\kappa}(m)) \right]$$
$$+ \left[\Phi_i(\theta_{N_i^\kappa}(m+1), \theta_{-N_i^\kappa}(m)) - \Phi_i(\theta(m)) \right]. \qquad (10)$$

To bound the first term on the RHS of Eq. (10), using Assumption 5.1 and the update equation in Algorithm 1 Line 3 and we have

$$\Phi_i(\theta(m+1)) - \Phi_i(\theta_{N_i^\kappa}(m+1), \theta_{-N_i^\kappa}(m)) \geq -\nu(\kappa) \max_{j \in -N_i^\kappa} \|\theta_j(m+1) - \theta_j(m)\|$$

$$= -\nu(\kappa)\beta \max_{j \in -N_i^\kappa} \left\|\nabla_{\theta_j} J_j(\theta(m))\right\|$$

$$\geq -\frac{\sqrt{2}\nu(\kappa)\beta}{(1-\gamma)^2},$$

where the last line follows from Lemma F.11.

To bound the second term on the RHS of the previous inequality, note that the smoothness property (cf. Lemma B.2) implies that

$$\Phi_i(\theta_{N_i^\kappa}(m+1), \theta_{-N_i^\kappa}(m)) - \Phi_i(\theta(m))$$

$$\geq \langle \nabla_{\theta_{N_i^\kappa}} \Phi_i(\theta(m)), \theta_{N_i^\kappa}(m+1) - \theta_{N_i^\kappa}(m) \rangle - \frac{L(\kappa)}{2} \left\|\theta_{N_i^\kappa}(m+1) - \theta_{N_i^\kappa}(m)\right\|^2$$

$$= \beta \langle \nabla_{\theta_{N_i^\kappa}} \Phi_i(\theta(m)), \nabla_{\theta_{N_i^\kappa}} J_i(\theta(m)) \rangle - \frac{L(\kappa)\beta^2}{2} \left\|\nabla_{\theta_j} J_i(\theta(m))\right\|^2$$

$$= \left(\beta - \frac{L(\kappa)\beta^2}{2}\right) \left\|\nabla_{\theta_{N_i^\kappa}} \Phi_i(\theta(m))\right\|^2 \qquad \text{(Lemma F.16)}$$

$$\geq \frac{\beta}{2} \left\|\nabla_{\theta_{N_i^\kappa}} \Phi_i(\theta(m))\right\|^2,$$

where the last line follows from $\beta \leq 1/L(\kappa)$.

Using the previous two bounds in Eq. (10) and we have

$$\Phi_i(\theta(m+1)) - \Phi_i(\theta(m) \geq \frac{\beta}{2} \left\|\nabla_{\theta_{N_i^\kappa}} \Phi_i(\theta(m))\right\|^2 - \frac{\sqrt{2}\nu(\kappa)\beta}{(1-\gamma)^2},$$

which completes the proof. $\qquad \square$

Define $D(\theta) = 1/\min_s d^\theta(s)$ and

$$c_i(\theta) = \min_s \sum_{a_i^* \in \arg\max_{a_i} \overline{Q}_i^\theta(s, a_i)} \xi_i^{\theta_i}(a_i^*|s_i), \quad \forall\, i \in \mathcal{N}.$$

Lemma B.4 (Lemma 1 of Zhang et al. [2022a]). *It holds for all $i \in \mathcal{N}$ that*

$$NE\text{-}Gap_i(\theta) \leq \frac{\sqrt{|\mathcal{A}_i|}D(\theta)}{c(\theta)} \left\|\nabla_{\theta_i} J_i(\theta)\right\|.$$

C   ANALYSIS OF CRITIC

In this section, we generalize the localized TD(λ) algorithm (cf. Algorithm 2) to *localized stochastic approximation* and analyze its performance. The localized stochastic approximation problem can be of independent interest.

C.1   LOCALIZED STOCHASTIC APPROXIMATION

To make this section self-contained, we first introduce the problem setting of localized stochastic approximation, and then propose the generalized TD(λ) algorithm. After that, we state assumptions and the main result of the localized stochastic approximation problem.

A localized stochastic approximation problem consists of an infinite-horizon, multi-agent Markov chain $\mathcal{M} = (\mathcal{N}, \mathcal{E}, \mathcal{Z}, \mathbb{P}, \tilde{r}, \gamma, \mu')$ and a fixed agent $i_0 \in \mathcal{N}$. Here $\mathcal{N} = \{1, 2, \cdots, n\}$ is the set of agents, associated with an undirected graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$. $\mathcal{Z} = \prod_{i \in \mathcal{N}} \mathcal{Z}_i$ is the global state space, where $\mathcal{Z}_i$ is the local state space of agent i.

At time $t \geq 0$, given current state $z(t) \in \mathcal{Z}$, for each $i \in \mathcal{N}$, the next individual state $z_i(t+1)$ is independently generated and is only dependent on its neighbors' states and its own action:

$$\mathbb{P}(z(t+1) \mid z(t)) = \prod_{i=1}^{n} \mathbb{P}_i(z_i(t+1) \mid z_{N_i}(t)). \tag{11}$$

γ is the discount factor. μ' is the initial state distribution. $\tilde{r} : \mathcal{Z}_{N_{i_0}^{\kappa_r}} \rightarrow [0,1]$ is the reward function.

Before the learning stage, the learner can observe partial information of a trajectory sampled from the Markov chain:

$$\tau_{\kappa_c} = (z_{N_{i_0}^{\kappa_c}}(0), \tilde{r}(0), \cdots, z_{N_{i_0}^{\kappa_c}}(K)).$$

Here K is the horizon of the sampled trajectory, $\kappa_c > \kappa_r$ measures the observability of the learner.

The goal is to estimate agent i_0's cost function $C(z) = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}[\tilde{r}(z_{N_{i_0}^{\kappa_r}}(t))|z(0) = z]$.

The localized stochastic approximation problem has the following two applications in localized stochastic approximation, depending on the choice of z_i's.

- Estimate local Q-function $Q_i(s,a)$. Let $z_j = (s_j, a_j), \forall j \in \mathcal{N}$.
- Estimate averaged Q-function $\bar{Q}_i(s, a_i)$. Let $z_j = s_j, \forall j \neq k, z_{i_0} = (s_{i_0}, a_{i_0})$. Notice that $\{s(t), a_{i_0}(t)\}$ forms an induced Markov chain of agent i_0's averaged MDP [Zhang et al., 2022b].

Additional notations. We introduce some other notations that will be used in this section.

We denote by $\pi_t \in \Delta(\mathcal{Z})$ the state distribution at time t and denote by $\bar{\pi} \in \Delta(\mathcal{Z})$ the stationary state distribution.

We use the notation $\pi_{\kappa_c, t}(z_{N_{i_0}^{\kappa_c}})$ to represent the marginal probability of state $z_{N_{i_0}^{\kappa_c}}$ at time t and use $\bar{\pi}_{\kappa_c}(z_{N_{i_0}^{\kappa_c}})$ to represent the marginal probability of state $z_{N_{i_0}^{\kappa_c}}$ under stationary distribution.

For any set $\mathcal{X}$ and two distributions $\pi_1, \pi_2 \in \Delta(\mathcal{X})$, define

$$\mathrm{TV}(\pi_1, \pi_2) = \max_{A \subseteq X} |\pi_1(A) - \pi_2(A)|. \tag{12}$$

to be the total variation distance between π_1 and π_2. The total variation distance has the following property [Levin and Peres, 2017]:

$$\mathrm{TV}(\pi_1, \pi_2) = \frac{1}{2} \|\pi_1 - \pi_2\|_1 \leq 1. \tag{13}$$

Given a Markov chain with state space $\mathcal{X} = \mathcal{X}_1 \times \cdots \times \mathcal{X}_l$ and transition probability Γ, for any set $I \subseteq \mathcal{U} = \{1, 2, \cdots, l\}$ and any $x_I' \in \mathcal{X}_I$, any $x \in \mathcal{X}$, let

$$\Gamma_I(x_I'|x) := \sum_{x_{\mathcal{U}/I}'} \Gamma(x_I', x_{\mathcal{U}/I}'|x) \tag{14}$$

be the marginal transition probability of x_I' given previous global state x.

C.2 GENERALIZED TD(λ) ALGORITHM

Now we design a generalized version of Algorithm 2 to make it compatible with more classical policy evaluation methods. We consider approximating the cost function C with function class $\hat{C} : \mathcal{Z}_{N_{i_0}^{\kappa_c}} \times \mathbb{R}^d \rightarrow \mathbb{R}$, where $\kappa_c > \kappa_r$. That is, $C(z)$ is approximated by $\hat{C}(z_{N_{i_0}^{\kappa_c}}, w)$, where $w \in \mathbb{R}^d$ is the parameter. Notice that we allow arbitrary function approximation class, and $\hat{C}$ only depends on states of agents in κ_c-hop neighborhood.

Furthermore, we denote by $\psi : \mathcal{Z}_{N_{i_0}^{\kappa_c}} \rightarrow \mathbb{R}^d$ the feature vector. Assume $\left\|\psi(z_{N_{i_0}^{\kappa_c}})\right\| \leq 1$ without loss of generality. We introduce the feature vector for compatibility with linear function approximation. Nevertheless, we emphasize that our algorithm still allows general function approximation class by choosing $\lambda = 0$. See Algorithm 4 for the complete algorithm.

We point out that Algorithm 4 reduces to Algorithm 2 if we choose $t_0 = 0$ and let F be the temporal difference $\delta_{i_0}(t)$ in Algorithm 2. Besides, many other classical single agent policy evaluation algorithm, such as LSTD [Boyan, 1999], are a special case of Algorithm 4.

Algorithm 4 Generalized TD(λ)

1: **Input**: τ_{κ_c}
2: **Parameter**: $\lambda \in [0, 1)$, $t_0 \geq 0$, function $F : (\mathcal{Z}_{N_{i_0}^{\kappa_c}})^{t_0+2} \times \mathbb{R}^d \to \mathbb{R}$.
   Initialization: $w(0) := 0$, $\zeta(0) := \psi(z_{N_{i_0}^{\kappa_c}}(0))$.
3: **for** $t = t_0, t_0 + 1, \cdots, K - 1$ **do**
4:    $X(t) := (z_{N_{i_0}^{\kappa_c}}(t - t_0), \cdots, z_{N_{i_0}^{\kappa_c}}(t + 1))$.
5:    Update parameter $w(t + 1) = w(t) + \alpha F(X(t), w(t))\zeta(t)$.
6:    Update eligibility vector $\zeta(t + 1) = \lambda\zeta(t) + \psi(z_{N_{i_0}^{\kappa_c}}(t + 1))$.
7: **end for**
8: **Return** $w(K)$

C.3 CONVERGENCE RESULT

To make the section self-contained, we restate assumptions needed for localized stochastic approximation. We point out that all assumptions below can be satisfied by localized TD(λ) with linear function approximation under the assumptions in the main text.

Assumption C.1. $\mathbb{P}$ *is aperiodic and irreducible.*

Assumption C.1 guarantees the existence and uniqueness of stationary distribution. In addition, Assumption C.1 ensures that there exists $c' > 1$ and $\rho' \in (0, 1)$ such that

$$\text{TV}(\pi_t, \overline{\pi}) \leq c'\rho'^t. \tag{15}$$

.

We further define the stationary distribution of $\zeta(t)$ [Tsitsiklis and Van Roy, 1997]. Consider a stationary Markov process $\{z(t)\}$ ($-\infty < t < \infty$), in which the state distribution at each time t is the stationary distribution. Let

$$\overline{\zeta}(t) = \sum_{k=-\infty}^{t} \lambda^{t-k}\psi(z_{N_{i_0}^{\kappa_c}}(k)),$$

where $\{z(k)\}$ is sampled from the stationary Markov process. $\overline{\zeta}(t)$ is well-defined, and its distribution is invariant of t. Thus we can use the distribution of $\overline{\zeta}(t)$ under stationary Markov process as the stationary distribution of $\zeta(t)$. We use $\overline{\mathbb{E}}[\cdot]$ to represent the expected value of a formula, given that $\{z(t)\}$ follows the stationary Markov process and $\zeta(t)$ is sampled from the defined stationary distribution.

Let $G(\zeta(t), X, w) = \zeta(t)F(X, w)$ and denote

$$\overline{G}(w) := \overline{\mathbb{E}}[\zeta(t)F(X, w)].$$

Assumption C.2.   *1. There exists $L_1 > 1$ such that*

$$|F(x, w_1) - F(x, w_2)| \leq L_1 \|w_1 - w_2\|, \forall w_1, w_2, x$$
$$|F(x, 0)| \leq L_1, \forall x$$

2. $\overline{G}(w)$ has a unique zero w^*. In addition, there exists $c_0 > 0$ such that

$$(w - w^*)^\top \overline{G}(w) \leq -c_0\|w - w^*\|^2, \forall w \in \mathbb{R}^d.$$

Point 1. ensures that our updating term is Lipschitz, while point 2. guarantees the existence of negative drift, which is crucial in Lyapunov analysis. Both assumptions are standard in non-linear stochastic approximation problems [Chen et al., 2022].

Assumption C.3. $\hat{C}(z_{N_{i_0}^{\kappa_c}}, w)$ *is L_2-Lipschitz with respect to w, i.e.,*

$$\left|\hat{C}(z_{\mathcal{N}_i^\kappa}, w_1) - \hat{C}(z_{\mathcal{N}_i^\kappa}, w_2)\right| \leq L_2\|w_1 - w_2\|$$

for all $z_{\mathcal{N}_i^\kappa} \in \mathcal{Z}_{\mathcal{N}_i^\kappa}, w_1, w_2 \in \mathbb{R}^d$.

Given the assumptions above, we can show a geometric mixing rate of $G(\zeta(t), X(t), w)$ to $\overline{G}(w)$ w.r.t. t. To be concrete, there exists $c_g(c', \rho', \lambda, t_0) > 0$ and $\rho_g(\rho', \lambda) \in (0, 1)$, such that for any $t \geq t_0$, we have

$$\left\| \mathbb{E}[G(\zeta(t), X(t), w)] - \overline{G}(w) \right\| \leq L_1(\|w\| + 1)c_g(c', \rho', \lambda, t_0)[\rho_g(\rho', \lambda)]^t. \tag{16}$$

This can be viewed as a generalized result of [Bertsekas and Tsitsiklis, 1996, Lemma 6.7], and we defer the proof to Appendix C.5.2. In order to state conditions on stepsize, we introduce the concept of mixing time of function G.

Definition C.4 (Mixing time of function G). *The mixing time of function G with precision δ, for any $\delta > 0$, is defined as*

$$t'_\delta := \min\left\{ t \geq 1 \,\big|\, \left\| \mathbb{E}[G(\zeta_t, X(t), w)] - \bar{G}(w) \right\| \leq \delta L_1(\|w\| + 1) \right\}.$$

Eq. (15) ensures that

$$t'_\delta = O(\log(1/\delta))$$

for any $\delta > 0$, so $\lim_{\delta \to 0} \delta t'_\delta = 0$.

The performance of Algorithm 4 is related to its globalized version on state space $\mathcal{Z}_{N_{i_0}^{\kappa_c}}$. To illustrate this idea, we define the concept "sub-chain".

Construct transition probability $\overline{\mathbb{P}}$ on state space $\mathcal{Z}_{N_{i_0}^{\kappa_c}}$ satisfying

$$\overline{\mathbb{P}}(z'_{N_{i_0}^{\kappa_c}} | z_{N_{i_0}^{\kappa_c}}) = \sum_{z_{-N_{i_0}^{\kappa_c}}} \frac{\overline{\pi}(z_{N_{i_0}^{\kappa_c}}, z_{-N_{i_0}^{\kappa_c}})}{\overline{\pi}_{N_{i_0}^{\kappa_c}}(z_{N_{i_0}^{\kappa_c}})} \mathbb{P}_{N_{i_0}^{\kappa_c}}(z'_{N_{i_0}^{\kappa_c}} | (z_{N_{i_0}^{\kappa_c}}, z_{-N_{i_0}^{\kappa_c}})). \tag{17}$$

Here

$$\mathbb{P}_{N_{i_0}^{\kappa_c}}(z'_{N_{i_0}^{\kappa_c}} | (z_{N_{i_0}^{\kappa_c}}, z_{-N_{i_0}^{\kappa_c}})) = \sum_{z'_{-N_{i_0}^{\kappa_c}}} \mathbb{P}(z'_{N_{i_0}^{\kappa_c}}, z'_{-N_{i_0}^{\kappa_c}} | (z_{N_{i_0}^{\kappa_c}}, z_{-N_{i_0}^{\kappa_c}})).$$

Notice that $\overline{\pi}_{N_{i_0}^{\kappa_c}}(z_{N_{i_0}^{\kappa_c}}) > 0$ due to Assumption C.1.

Definition C.5. *Let $\mathcal{E}_{N_{i_0}^{\kappa_c}} \subseteq N_{i_0}^{\kappa_c} \times N_{i_0}^{\kappa_c}$ denote the edges with two end points in $N_{i_0}^{\kappa_c}$. Let $\mu'_{N_{i_0}^{\kappa_c}}$ denote the marginal initial state distribution of state space $\mathcal{Z}_{N_{i_0}^{\kappa_c}}$.*

Then Markov chain $\mathcal{M}_{N_{i_0}^{\kappa_c}} = (N_{i_0}^{\kappa_c}, \mathcal{E}_{N_{i_0}^{\kappa_c}}, \mathcal{Z}_{N_{i_0}^{\kappa_c}}, \overline{\mathbb{P}}, \tilde{r}, \gamma, \mu'_{N_{i_0}^{\kappa_c}})$ is called the sub-chain of Markov chain $(\mathcal{N}, \mathcal{E}, \mathcal{Z}, \mathbb{P}, \tilde{r}, \gamma, \mu')$ with respect to agents $N_{i_0}^{\kappa_c}$.

We denote by $\tilde{C} : \mathcal{Z}_{N_{i_0}^{\kappa_c}} \to \mathbb{R}$ the sub-chain's cost function.

The following concept is critical for the reduction of a localized algorithm to its globalized version.

Definition C.6. *The reduction error is defined as*

$$\epsilon_{red} := \sup_{z_{N_{i_0}^{\kappa_c}} \in \mathcal{Z}_{N_{i_0}^{\kappa_c}}} \left| \hat{C}(z_{N_{i_0}^{\kappa_c}}, w^*) - \tilde{C}(z_{N_{i_0}^{\kappa_c}}) \right|. \tag{18}$$

The reduction error is dependent on the actual algorithm used, including update rule, choice of function approximation class.

We now present our main theorem of localized stochastic approximation. The proof sketch is given in subsection C.4, and the detailed proofs are deferred to section C.5.

Theorem C.7. *Suppose Assumptions C.1, C.2 and C.3 are satisfied. Choose stepsize α satisfying $\alpha t'_\alpha \leq \min\left\{ \frac{1}{4L'_1}, \frac{c_0}{114L_1'^2} \right\}$, where $L'_1 = \frac{L_1}{1-\lambda}$. Then we have for all $K \geq t'_\alpha$*

$$\mathbb{E}\left[\sup_z \left| \hat{C}(z_{N_{i_0}^{\kappa_c}}, w(K)) - C(z) \right| \right] \leq 3 \left[L_2^2 \left(c_1(1 - c_0\alpha)^{K - t'_\alpha} + c_2 \frac{\alpha t'_\alpha}{c_0} \right) \right.$$
$$\left. + \epsilon_{red}^2 + \left(\frac{\gamma^{\kappa_c - \kappa_r + 1}}{1 - \gamma} \right)^2 \right].$$

Here $c_1 = (\|w(0)\| + \|w(0) - w^*\| + 1)^2$, $c_2 = 114L_1'^2(\|w^*\| + 1)^2$, ϵ_{red} is defined in Definition C.6.

Notice that $t'_\delta = O(\log \frac{1}{\delta})$. So $\alpha t'_\alpha \to 0$ when $\alpha \to 0$. Thus the conditions for α in Theorem C.7 can be satisfied when α is small enough.

C.4 PROOF OF THEOREM C.7

Key idea: A Reduction Approach. The key to analyzing the localized stochastic approximation algorithm is to reduce it to a globalized policy evaluation algorithm on the sub-chain with respect to agents $N_{i_0}^{\kappa_c}$. To be specific, Algorithm 4 itself can be regarded as a globalized policy evaluation algorithm for the sub-chain $\mathcal{M}_{N_{i_0}^{\kappa_c}}$.

The most important part of the proof is that the sub-chain has the following properties:

1. The "sub-chain" is an aperiodic and irreducible Markov chain.
2. The local transition probabilities of agents in $N_{i_0}^{\kappa_c-1}$ are the same for the sub-chain and the original chain. The local transition probabilities in the sub-chain do not need to be independent.
3. The stationary distribution of sub-chain equals the marginal stationary distribution of $\mathcal{Z}_{N_{i_0}^{\kappa_c}}$ in the original chain.

We formulate the properties as the three lemmas below. The proofs are deferred to Appendix C.5.1.

Lemma C.8. $\overline{\mathbb{P}}$ *is aperiodic and irreducible.*

Lemma C.9. *We have for any agent $i \in N_{i_0}^{\kappa_c-1}$ that*

$$\overline{\mathbb{P}}_i(z'_i | z_{N_i^{\kappa_c}}) = \mathbb{P}_i(z'_i | z_{N_i}), \forall z'_i \in \mathcal{Z}_i, z_{N_{i_0}^{\kappa_c}} \in \mathcal{Z}_{N_{i_0}^{\kappa_c}}. \tag{19}$$

Notice that $N_i \subseteq N_{i_0}^{\kappa_c}$ when $i \in N_{i_0}^{\kappa_c-1}$.

Lemma C.10. *Consider the stationary Markov chain $(\mathcal{N}, \mathcal{E}, \mathcal{Z}, \mathbb{P}, \tilde{r}, \gamma, \overline{\pi})$. That is, the initial state distribution $\mu' = \overline{\pi}$ is the stationary distribution.*

Then for any $t \in \mathbb{N}$, $z_{N_{i_0}^{\kappa_c}}, z'_{N_{i_0}^{\kappa_c}} \in \mathcal{Z}_{N_{i_0}^{\kappa_c}}$, we have

$$\Pr\left[z_{N_{i_0}^{\kappa_c}}(t+1) = z'_{N_{i_0}^{\kappa_c}} | z_{N_{i_0}^{\kappa_c}}(t) = z_{N_{i_0}^{\kappa_c}}\right] = \overline{\mathbb{P}}(z'_{N_{i_0}^{\kappa_c}} | z_{N_{i_0}^{\kappa_c}}).$$

Now we can discuss the proof of Theorem C.7. We can think that Algorithm 4 is executed on the local chain $\mathcal{M}_{N_{i_0}^{\kappa_c}}$, with cost function denoted by $\tilde{C}$. Then w^* can be regarded as the fix point of parameter update in a globalized algorithm. In this

regard, we can do the following error decomposition.

$$\mathbb{E}\left[\sup_z \left|\hat{C}(z_{N_{i_0}^{\kappa_c}}, w(t)) - C(z)\right|^2\right]$$

$$\leq \mathbb{E}\left[\sup_z \left(\left|\hat{C}(z_{N_{i_0}^{\kappa_c}}, w(t)) - \hat{C}(z_{N_{i_0}^{\kappa_c}}, w^*)\right| + \left|\hat{C}(z_{N_{i_0}^{\kappa_c}}, w^*) - \tilde{C}(z_{N_{i_0}^{\kappa_c}})\right|\right.\right.$$

$$\left.\left. + \left|\tilde{C}(z_{N_{i_0}^{\kappa_c}}) - C(z)\right|\right)^2\right]$$

$$\overset{(i)}{\leq} 3\left\{\mathbb{E}\left[\sup_z \left|\hat{C}(z_{N_{i_0}^{\kappa_c}}, w(t)) - \hat{C}(z_{N_{i_0}^{\kappa_c}}, w^*)\right|^2\right] + \mathbb{E}\left[\sup_z \left|\hat{C}(z_{N_{i_0}^{\kappa_c}}, w^*) - \tilde{C}(z_{N_{i_0}^{\kappa_c}})\right|^2\right]\right.$$

$$\left. + \mathbb{E}\left[\sup_z \left|\tilde{C}(z_{N_{i_0}^{\kappa_c}}) - C(z)\right|^2\right]\right\}$$

$$= 3\left\{\underbrace{\mathbb{E}\left[\sup_{z_{N_{i_0}^{\kappa_c}}} \left|\hat{C}(z_{N_{i_0}^{\kappa_c}}, w(t)) - \hat{C}(z_{N_{i_0}^{\kappa_c}}, w^*)\right|^2\right]}_{(a)} + \underbrace{\sup_{z_{N_{i_0}^{\kappa_c}}} \left|\hat{C}(z_{N_{i_0}^{\kappa_c}}, w^*) - \tilde{C}(z_{N_{i_0}^{\kappa_c}})\right|^2}_{(b)}\right.$$

$$\left. + \underbrace{\sup_z \left|\tilde{C}(z_{N_{i_0}^{\kappa_c}}) - C(z)\right|^2}_{(c)}\right\}. \tag{20}$$

Here (i) is by Cauchy-Schwarz inequality.

To interpret the error terms, (a) is related to the convergence of $w(t)$ to the fix point w^*, which can be viewed as a globalized non-linear stochastic approximation problem [Chen et al., 2022]. (b) is the inherent property of the globalized algorithm and the Markov chain, which is defined as ϵ_{red} in Definition C.6. We point out that for some algorithms, such as TD(λ) with linear function approximation, ϵ_{red} can be bounded by the function approximation error. (c) is the difference of the cost function in the sub-chain and in the original chain, which originates from the use of a localized algorithm and decays exponentially with κ_c.

Bound (a). This can be done in two steps. The first step is to analyze the convergence rate of $w(t)$ to the stationary point w^*, which can be viewed as a stochastic approximation problem. Thus we can adopt Lyapunov approach to bound the convergence rate, which is a standard method in stochastic approximation [Srikant and Ying, 2019, Chen et al., 2022].

The second step is to combine the Lipschitz assumption of $\hat{C}$ (Assumption C.3) with convergence result of $w(t)$ and derive the bound of (a).

With the proof idea above, we proceed to bound the convergence rate of $w(t)$. In order to utilize the mixing of function $G(\zeta(t), X(t), w(t))$, we take expectation conditioned on $X(t - t_\alpha')$, $\zeta(t - t_\alpha')$ and $w(t - t_\alpha')$. For convenience of notation, we use $\mathbb{E}_\alpha[\cdot] = \mathbb{E}[\cdot|X(t - t_\alpha'), \zeta(t - t_\alpha'), w(t - t_\alpha')]$.

For any $t \geq t_\alpha'$, we have

$$\mathbb{E}_\alpha\left[\|w(t+1) - w^*\|^2\right] - \mathbb{E}_\alpha\left[\|w(t) - w^*\|^2\right]$$

$$= 2\mathbb{E}_\alpha\left[(w(t) - w^*)^\top(w(t+1) - w(t))\right] + \mathbb{E}_\alpha\left[(w(t+1) - w(t))^2\right]$$

$$= 2\alpha \underbrace{\mathbb{E}_\alpha\left[(w(t) - w^*)^\top \overline{G}(w(t))\right]}_{(a_1)}$$

$$+ 2\alpha \underbrace{\mathbb{E}_\alpha\left[(w(t) - w^*)^\top(G(\zeta(t), X(t), w(t)) - \overline{G}(w(t)))\right]}_{(a_2)}$$

$$+ \alpha^2 \underbrace{\mathbb{E}_\alpha\left[\|G(\zeta(t), X(t), w(t))\|^2\right]}_{(a_3)}. \tag{21}$$

Term (a_1) corresponds to the negative drift, and Assumption C.2 indicates that

$$(a_1) \leq -c_0 \mathbb{E}_\alpha[\|w(t) - w^*\|^2],$$

Before we bound (a_2) and (a_3), we show that function G is Lipschitz. See Appendix C.5.3 for the proof.

Lemma C.11. *Let* $L_1' = \frac{L_1}{1-\lambda}$. *Then we have for all* t, x,

$$\|G(\zeta(t), x, w_1) - G(\zeta(t), x, w_2)\| \leq L_1' \|w_1 - w_2\|, \forall w_1, w_2$$
$$\|G(\zeta(t), x, 0)\| \leq L_1'.$$

Besides, Lemma C.11 also implies for any t, x, w, w_1, w_2 that

$$\left\|\overline{G}(w_1) - \overline{G}(w_2)\right\| \leq \overline{\mathbb{E}}\left[\|G(\zeta(t), X(t), w_1) - G(\zeta(t), X(t), w_2)\|\right] \leq L_1' \|w_1 - w_2\|$$
$$\|G(\zeta(t), x, w)\| \leq \|G(\zeta(t), x, w) - G(\zeta(t), x, 0)\| + \|G(\zeta(t), x, 0)\| \leq L_1'(\|w\| + 1)$$
$$\left\|\overline{G}(w)\right\| \leq \overline{\mathbb{E}}[\|G(\zeta(t), X(t), w)\|] \leq L_1'(\|w\| + 1).$$

We apply Lemma C.11 and bound term (a_3) as

$$\begin{aligned}
(a_3) =& \mathbb{E}_\alpha\left[\|G(\zeta(t), X(t), w(t))\|^2\right] \\
\leq & L_1'^2 \mathbb{E}_\alpha\left[(\|w(t)\| + 1)^2\right] \\
\leq & L_1'^2 \mathbb{E}_\alpha\left[(\|w(t) - w^*\| + \|w^*\| + 1)^2\right] \\
\leq & 2L_1'^2 \mathbb{E}_\alpha\left[\|w(t) - w^*\|^2 + (\|w^*\| + 1)^2\right].
\end{aligned}$$

Finally we bound (a_2). Before that, we need to control the difference of $w(t_1)$ and $w(t_2)$ for any t_1, t_2 under certain conditions of stepsize. This can be formulated as the lemma below, with proof deferred to Appendix C.5.4.

Lemma C.12. *For any* $t_1 > t_2 \geq 0$, *if* $\alpha(t_1 - t_2) \leq \frac{1}{4L_1'}$, *then we have*

$$\begin{aligned}
\|w(t_1) - w(t_2)\| &\leq 2L_1' \alpha(t_1 - t_2)(\|w(t_2)\| + 1) \\
&\leq 4L_1' \alpha(t_1 - t_2)(\|w(t_1)\| + 1).
\end{aligned}$$

With Lemma C.12, we are able to control (a_2). The proof is given in Appendix C.5.5.

Lemma C.13. *Suppose* $\alpha t_\alpha' \leq \frac{1}{4L_1'}$. *Then the following inequality holds for all* $t \geq t_\alpha'$:

$$(a_2) \leq 56 L_1'^2 \alpha t_\alpha' \mathbb{E}_\alpha\left[\|w(t) - w^*\|^2 + (\|w^*\| + 1)^2\right].$$

Now we apply the upper bounds for (a_1), (a_2) and (a_3) to Eq. (21), and we get

$$\begin{aligned}
& \mathbb{E}_\alpha\left[\|w(t+1) - w^*\|^2\right] - \mathbb{E}_\alpha\left[\|w(t) - w^*\|^2\right] \\
\leq & 2\alpha(a_1) + 2\alpha(a_2) + \alpha^2(a_3) \\
\leq & -2c_0 \alpha \mathbb{E}_\alpha[\|w(t) - w^*\|^2] \\
& + 112 L_1'^2 \alpha^2 t_\alpha' \mathbb{E}_\alpha\left[\|w(t) - w^*\|^2 + (\|w^*\| + 1)^2\right] \\
& + 2L_1'^2 \alpha^2 \mathbb{E}_\alpha\left[\|w(t) - w^*\|^2 + (\|w^*\| + 1)^2\right] \\
\leq & -2c_0 \mathbb{E}_\alpha[\|w(t) - w^*\|^2] \\
& + 114 L_1'^2 \alpha^2 t_\alpha' \mathbb{E}_\alpha\left[\|w(t) - w^*\|^2 + (\|w^*\| + 1)^2\right].
\end{aligned}$$

Rearranging the terms, we derive that

$$\mathbb{E}_\alpha \left[\|w(t+1) - w^*\|^2 \right]$$

$$\leq \left(1 - 2c_0\alpha + 114L_1'^2\alpha^2 t_\alpha'\right) \mathbb{E}_\alpha \left[\|w(t) - w^*\|^2 \right] + 114L_1'^2\alpha^2 t_\alpha'(\|w^*\| + 1)^2$$

$$\leq (1 - c_0\alpha)\mathbb{E}_\alpha \left[\|w(t) - w^*\|^2 \right] + 114L_1'^2\alpha^2 t_\alpha'(\|w^*\| + 1)^2,$$

where the last inequality is due to $\alpha t_\alpha' \leq \frac{c_0}{114L_1'^2}$. Taking total expectation on both sides, we get

$$\mathbb{E}\left[\|w(t+1) - w^*\|^2 \right] \leq (1 - c_0\alpha)\mathbb{E}\left[\|w(t) - w^*\|^2 \right] + 114L_1'^2\alpha^2 t_\alpha'(\|w^*\| + 1)^2.$$

Repeatedly use the above inequality starting from t_α', and we have

$$\mathbb{E}\left[\|w(t) - w^*\|^2 \right] \leq (1 - c_0\alpha)^{t - t_\alpha'}\mathbb{E}\left[\|w(t_\alpha') - w^*\|^2 \right]$$

$$+ 114L_1'^2\alpha^2 t_\alpha'(\|w^*\| + 1)^2 \sum_{k=0}^{t - t_\alpha' - 1} (1 - c_0\alpha)^k$$

$$\leq (1 - c_0\alpha)^{t - t_\alpha'}\mathbb{E}\left[\|w(t_\alpha') - w^*\|^2 \right] + \frac{114L_1'^2\alpha t_\alpha'(\|w^*\| + 1)^2}{c_0}.$$

We can bound $\mathbb{E}\left[\|w(t_\alpha') - w^*\|^2 \right]$ by

$$\mathbb{E}\left[\|w(t_\alpha') - w^*\|^2 \right] \leq \mathbb{E}\left[\left(\|w(t_\alpha') - w(0)\| + \|w(0) - w^*\|^2 \right) \right]$$

$$\overset{(i)}{\leq} \mathbb{E}\left[(\|w(0)\| + \|w(0) - w^*\| + 1)^2 \right]$$

$$= c_1,$$

where (i) is by Lemma C.12, with $t_1 = t_\alpha'$ and $t_2 = 0$, and by $\alpha t_\alpha' \leq \frac{1}{4L_1'}$.

Noticing that $c_2 = 114L_1'^2(\|w^*\| + 1)^2$, we substitute t with K and get

$$\mathbb{E}\left[\|w(K) - w^*\|^2 \right] \leq c_1(1 - c_0\alpha)^{K - t_\alpha'} + c_2\frac{\alpha t_\alpha'}{c_0}, \ \forall K \geq t_\alpha'. \tag{22}$$

For the second step, we combine Assumption C.3 with Eq. (22) and get the following result, the proof of which is in Appendix C.5.6.

Lemma C.14. *For $K > t_\alpha'$, we have*

$$(a) \leq L_2^2\left(c_1(1 - \alpha c_0)^{K - t_\alpha'} + c_2\frac{\alpha t_\alpha'}{c_0} \right).$$

Bound (b). We have $(b) = \epsilon_{red}^2$ by Definition C.6.

Bound (c). This can be derived by the exponential decay property of the cost function. See Appendix C.5.7 for the proof.

Lemma C.15. *The cost function of the sub-chain and the original chain differs by*

$$\sup_z \left| \tilde{C}(z_{N_{i_0}^{\kappa_c}}) - C(z) \right| \leq \frac{1}{1 - \gamma}\gamma^{\kappa_c - \kappa_r + 1}.$$

Thus $(c) \leq \left(\frac{\gamma^{\kappa_c - \kappa_r + 1}}{1 - \gamma} \right)^2$.

Eventually. we combine bounds for (a), (b), (c) (Lemma C.14, Definition C.6 and Lemma C.15) and plug into Eq. (20). Then we complete the proof.

C.5 PROOFS OF TECHNICAL LEMMAS IN APPENDIX C

C.5.1 Properties of Sub-Chain

We give the proofs of Lemma C.8, C.9, C.10.

Proof of Lemma C.8. For any Markov chain with transition probability Γ on some state space $\mathcal{X}$, we write $\Gamma^k(x'|x) = \Pr[x(t+k) = x'|x(t) = x]$, for any $t \in \mathbb{N}$, $x, x' \in \mathcal{X}$.

For any $z_{N_{i_0}^{\kappa_c}}, z'_{N_{i_0}^{\kappa_c}} \in \mathcal{Z}_{N_{i_0}^{\kappa_c}}$, we randomly pick $z_{-N_{i_0}^{\kappa_c}}, z'_{-N_{i_0}^{\kappa_c}} \in \mathcal{Z}_{-N_{i_0}^{\kappa_c}}$ and let $z = (z_{N_{i_0}^{\kappa_c}}, z_{-N_{i_0}^{\kappa_c}})$, $z' = (z'_{N_{i_0}^{\kappa_c}}, z'_{-N_{i_0}^{\kappa_c}})$. Since $\mathbb{P}$ is irreducible, there exists $i_0 > 0$, such that $\mathbb{P}^k(z'|z) > 0$, so $\mathbb{P}^k_{N_{i_0}^{\kappa_c}}(z'_{N_{i_0}^{\kappa_c}}|z) > 0$. $\mathbb{P}^k_{N_{i_0}^{\kappa_c}}(z'_{N_{i_0}^{\kappa_c}}|z)$ represents the marginal probability of $z'_{N_{i_0}^{\kappa_c}}$ given previous state z. By the interpretation of $\overline{\mathbb{P}}$ given in Lemma C.10,

$$
\begin{aligned}
\overline{\mathbb{P}}^k(z'_{N_{i_0}^{\kappa_c}}|z_{N_{i_0}^{\kappa_c}}) &= \sum_{\hat{z}_{-N_{i_0}^{\kappa_c}} \in \mathcal{Z}_{-N_{i_0}^{\kappa_c}}} \frac{\overline{\pi}(z_{N_{i_0}^{\kappa_c}}, \hat{z}_{-N_{i_0}^{\kappa_c}})}{\overline{\pi}_{N_{i_0}^{\kappa_c}}(z_{N_{i_0}^{\kappa_c}})} \mathbb{P}^k_{N_{i_0}^{\kappa_c}}(z'_{N_{i_0}^{\kappa_c}}|z_{N_{i_0}^{\kappa_c}}, \hat{z}_{-N_{i_0}^{\kappa_c}}) \\
&\geq \frac{\overline{\pi}(z)}{\overline{\pi}_{N_{i_0}^{\kappa_c}}(z_{N_{i_0}^{\kappa_c}})} \mathbb{P}^k_{N_{i_0}^{\kappa_c}}(z'_{N_{i_0}^{\kappa_c}}|z) \\
&> 0.
\end{aligned}
$$

Therefore, $\overline{\mathbb{P}}$ is irreducible.

To show that $\overline{\mathbb{P}}$ is aperiodic, we assume that $\overline{\mathbb{P}}$ has period $T \geq 2$. Then for any i_0 not divisible by T and any $z_{N_{i_0}^{\kappa_c}}$, $\overline{\mathbb{P}}^k(z_{N_{i_0}^{\kappa_c}}|z_{N_{i_0}^{\kappa_c}}) = 0$. For any $z \in \mathcal{Z}$, since for i_0 not divisible by T,

$$
0 = \overline{\mathbb{P}}^k(z_{N_{i_0}^{\kappa_c}}|z_{N_{i_0}^{\kappa_c}}) = \sum_{\hat{z}_{-N_{i_0}^{\kappa_c}} \in \mathcal{Z}_{-N_{i_0}^{\kappa_c}}} \frac{\overline{\pi}(z_{N_{i_0}^{\kappa_c}}, \hat{z}_{-N_{i_0}^{\kappa_c}})}{\pi^*_{N_{i_0}^{\kappa_c}}(z_{N_{i_0}^{\kappa_c}})} \mathbb{P}^k_{N_{i_0}^{\kappa_c}}(z_{N_{i_0}^{\kappa_c}}|(z_{N_{i_0}^{\kappa_c}}, \hat{z}_{-N_{i_0}^{\kappa_c}})),
$$

we have $\mathbb{P}^k_{N_{i_0}^{\kappa_c}}(z_{N_{i_0}^{\kappa_c}}|(z_{N_{i_0}^{\kappa_c}}, \hat{z}_{-N_{i_0}^{\kappa_c}})) = 0$ for any $\hat{z}_{-N_{i_0}^{\kappa_c}} \in \mathcal{Z}_{-N_{i_0}^{\kappa_c}}$. In particular, $\overline{\mathbb{P}}^k_{N_{i_0}^{\kappa_c}}(z_{N_{i_0}^{\kappa_c}}|z) = 0$ and thus $\overline{\mathbb{P}}^k(z|z) = 0$. This implies that the period of state z is at least $T \geq 2$, which contradicts the assumption that $\mathbb{P}$ is aperiodic. Hence $\overline{\mathbb{P}}$ is aperiodic.

In conclusion, $\overline{\mathbb{P}}$ is irreducible and aperiodic. $\qquad\square$

Proof of Lemma C.9. For simplicity, let $I = N_{i_0}^{\kappa_c}/\{i\}$. Then we have

$$
\begin{aligned}
\overline{\mathbb{P}}_i(z'_i|z_{N_{i_0}^{\kappa_c}}) &= \sum_{z'_I} \overline{\mathbb{P}}(z'_i, z'_I|z_{N_{i_0}^{\kappa_c}}) \\
&= \sum_{z'_I} \sum_{z_{-N_{i_0}^{\kappa_c}}} \frac{\overline{\pi}(z_{N_{i_0}^{\kappa_c}}, z_{-N_{i_0}^{\kappa_c}})}{\overline{\pi}_{\kappa_c}(z_{N_{i_0}^{\kappa_c}})} \mathbb{P}_{N_{i_0}^{\kappa_c}}(z'_{N_{i_0}^{\kappa_c}}|(z_{N_{i_0}^{\kappa_c}}, z_{-N_{i_0}^{\kappa_c}})) \\
&= \sum_{z_{-N_{i_0}^{\kappa_c}}} \frac{\overline{\pi}(z_{N_{i_0}^{\kappa_c}}, z_{-N_{i_0}^{\kappa_c}})}{\overline{\pi}_{\kappa_c}(z_{N_{i_0}^{\kappa_c}})} \sum_{z'_I} \mathbb{P}_{N_{i_0}^{\kappa_c}}(z'_{N_{i_0}^{\kappa_c}}|(z_{N_{i_0}^{\kappa_c}}, z_{-N_{i_0}^{\kappa_c}})) \\
&\overset{(i)}{=} \sum_{z_{-N_{i_0}^{\kappa_c}}} \frac{\overline{\pi}(z_{N_{i_0}^{\kappa_c}}, z_{-N_{i_0}^{\kappa_c}})}{\overline{\pi}_{\kappa_c}(z_{N_{i_0}^{\kappa_c}})} \mathbb{P}_i(z'_i|z_{N_i}) \\
&\overset{(ii)}{=} \mathbb{P}_i(z'_i|z_{N_i}) \sum_{z_{-N_{i_0}^{\kappa_c}}} \frac{\overline{\pi}(z_{N_{i_0}^{\kappa_c}}, z_{-N_{i_0}^{\kappa_c}})}{\overline{\pi}_{\kappa_c}(z_{N_{i_0}^{\kappa_c}})} \\
&= \mathbb{P}_i(z'_i|z_{N_i}).
\end{aligned}
$$

Here (i) uses the fact that the local transition probability $\mathbb{P}_i$ only depends on the states of agents in N_i, and (ii) is because $N_i \cap -N_{i_0}^{\kappa_c} = \varnothing$. $\qquad\square$

Proof of Lemma C.10. Since we consider the stationary Markov chain, the state distribution at any time t is equal to the stationary state distribution $\overline{\pi}$, so we have

$$\Pr\left[z_{N_{i_0}^{\kappa_c}}(t+1) = z'_{N_{i_0}^{\kappa_c}} | z_{N_{i_0}^{\kappa_c}}(t) = z_{N_{i_0}^{\kappa_c}}\right]$$

$$= \frac{\Pr\left[z_{N_{i_0}^{\kappa_c}}(t+1) = z'_{N_{i_0}^{\kappa_c}}, z_{N_{i_0}^{\kappa_c}}(t) = z_{N_{i_0}^{\kappa_c}}\right]}{\Pr\left[z_{N_{i_0}^{\kappa_c}}(t) = z_{N_{i_0}^{\kappa_c}}\right]}$$

$$= \frac{\sum_{z_{-N_{i_0}^{\kappa_c}}} \Pr\left[z_{N_{i_0}^{\kappa_c}}(t+1) = z'_{N_{i_0}^{\kappa_c}}, z(t) = (z_{N_{i_0}^{\kappa_c}}, z_{-N_{i_0}^{\kappa_c}})\right]}{\overline{\pi}_{N_{i_0}^{\kappa_c}}(z_{N_{i_0}^{\kappa_c}})}$$

$$= \frac{\sum_{z_{-N_{i_0}^{\kappa_c}}} \overline{\pi}(z_{N_{i_0}^{\kappa_c}}, z_{-N_{i_0}^{\kappa_c}}) \mathbb{P}_{N_{i_0}^{\kappa_c}}(z'_{N_{i_0}^{\kappa_c}} | (z_{N_{i_0}^{\kappa_c}}, z_{-N_{i_0}^{\kappa_c}}))}{\overline{\pi}_{N_{i_0}^{\kappa_c}}(z_{N_{i_0}^{\kappa_c}})}$$

$$= \overline{\mathbb{P}}(z'_{N_{i_0}^{\kappa_c}} | z_{N_{i_0}^{\kappa_c}}).$$

$\qquad\square$

C.5.2 Geometric mixing of the function G

We prove the geometric mixing property of function G (cf. Eq. (16)), which can be formalized as the lemma below:

Lemma C.16. *There exists $c_g(c', \rho', \lambda, t_0) > 0$ and $\rho_g(\rho', \lambda) \in (0,1) \in (0,1)$, such that for any $t \geq t_0$, we have*

$$\left\| \mathbb{E}[G(\zeta(t), X(t), w)] - \overline{G}(w) \right\| \leq L_1(\|w\| + 1) c_g(c', \rho', \lambda, t_0) [\rho_g(\rho', \lambda)]^t.$$

To prove Lemma C.16, we need some auxiliary results. Let $Y^{t_1}(t) = (z(t - t_1), \cdots, z(t+1)) \in \mathcal{Y}^{t_1} = \mathcal{Z}^{t_1+2}$. Denote by $\pi_{Y,t_1,t} \in \Delta(\mathcal{Y}^{t_1})$ the distribution of $Y^{t_1}(t)$ and by $\overline{\pi}_{Y,t_1} \in \Delta(\mathcal{Y}^{t_1})$ the corresponding stationary distribution.

$\pi_{Y,t_1,t}$ and $\overline{\pi}_{Y,t_1}$ can be computed by

$$\pi_{Y,t_1,t}(z_{t-t_1}, \cdots, z_{t+1}) = \pi_{t-t_1}(z_{t-t_1}) \prod_{i=t-t_1}^{t} \mathbb{P}(z_{i+1}|z_i) \tag{23}$$

$$\overline{\pi}_Y(z_0, \cdots, z_{t_1+2}) = \overline{\pi}(z_0) \prod_{i=0}^{t+1} \mathbb{P}(z_{i+1}|z_i). \tag{24}$$

The following lemma states the convergence rate of $\pi_{Y,t_1,t}$.

Lemma C.17. *For any $t \geq t_1$, $\mathrm{TV}(\pi_{Y,t_1,t}, \overline{\pi}_{Y,t_1}) \leq c' \rho'^{t-t_1}$.*

Proof of Lemma C.17. We have by Eqs. (23) and (24)

$$\mathrm{TV}(\pi_{Y,t_1,t}, \overline{\pi}_{Y,t_1}) = \frac{1}{2} \sum_{z_{t-t_1}, \cdots, z_{t+1}} |\pi_{Y,t_1,t}(z_{t-t_1}, \cdots, z_{t+1}) - \overline{\pi}_Y(z_{t-t_1}, \cdots, z_{t+1})|$$

$$= \frac{1}{2} \sum_{z_{t-t_1}, \cdots, z_{t+1}} |\pi_{t-t_1}(z_{t-t_1}) - \overline{\pi}(z_{t-t_1})| \prod_{i=t-t_1}^{t} [\mathbb{P}(z_{i+1}|z_i)]|$$

$$= \frac{1}{2} \sum_{z_{t-t_1}} |\pi_{t-t_1}(z_{t-t_1}) - \overline{\pi}(z_{t-t_1})|$$

$$= \mathrm{TV}(\pi_{t-t_1}, \overline{\pi})$$

$$\leq c' \rho'^{t-t_1}.$$

$\square$

Lemma C.18. *For $t \geq m$, any w and any $X(t) = (z_{N_{i_0}^{\kappa_c}}(t - t_0), \cdots, z_{N_{i_0}^{\kappa_c}}(t + 1))$, we have*

$$\left| \mathbb{E}[\psi(z(t - m))F(X(t), w)] - \overline{\mathbb{E}}[\psi(z(t - m))F(X(t), w)] \right|$$
$$\leq 2L_1(\|w\| + 1)c'\rho'^{t - \max\{m, t_0\}}.$$

Proof of Lemma C.18. Let set $\mathcal{S}^{t_1} = \{(z(t - t_1), \cdots, z(t + 1)) \in \mathcal{Y}^{t_1} \mid (z_{N_{i_0}^{\kappa_c}}(t - t_1), \cdots, z_{N_{i_0}^{\kappa_c}}(t + 1) = X(t))\}$, for any $t_1 \in \mathbb{N}$.

When $m \geq t_0$, we have

$$\left\| \mathbb{E}[\psi(z(t - m))F(X(t), w)] - \overline{\mathbb{E}}[\psi(z(t - m))F(X(t), w)] \right\|$$
$$= \left\| \sum_{Y^m \in \mathcal{S}^m} \psi(z(t - m))F(X(t), w)(\pi_{Y,m,t}(Y^m) - \overline{\pi}_{Y,m}(Y^m)) \right\|$$
$$\leq \sum_{Y^m \in \mathcal{S}^m} \|\psi(z(t - m))\| \, |F(X(t), w)| \, |\pi_{Y,m,t}(Y^m) - \overline{\pi}_{Y,m}(Y^m))|$$
$$\leq L_1(\|w\| + 1) \sum_{Y^m \in \mathcal{S}^m} |\pi_{Y,m,t}(Y^m) - \overline{\pi}_{Y,m}(Y^m))|$$
$$= 2L_1(\|w\| + 1)\text{TV}(\pi_{Y,m,t}, \overline{\pi}_{Y,m})$$
$$\leq 2L_1(\|w\| + 1)c'\rho'^{t - m}.$$

Here the last inequality is due to Lemma C.17.

We can similarly prove the case for $m < t_0$. In that case, $z(t - m)$ is included in $(z(t - t_0), \cdots, z(t + 1))$, so we just need to replace m with t_0 in the proof above. $\square$

Now we can bound the convergence rate of function G.

Proof of Lemma C.16. We have by definition of function G and $\overline{G}$ that

$$\left\| \mathbb{E}[G(\zeta(t), X(t), w)] - \overline{G}(w) \right\|$$
$$= \left\| \sum_{m=0}^{t} \lambda^m \mathbb{E}[\psi(z(t - m))F(X(t), w)] - \sum_{m=0}^{\infty} \lambda^m \overline{\mathbb{E}}[\psi(z(t - m))F(X(t), w)] \right\|$$
$$\leq \sum_{m=0}^{t} \lambda^m \left\| \mathbb{E}[\psi(z(t - m))F(X(t), w)] - \overline{\mathbb{E}}[\psi(z(t - m))F(X(t), w)] \right\|$$
$$+ \sum_{m=t+1}^{\infty} \lambda^m \left\| \overline{\mathbb{E}}[\psi(z(t - m))F(X(t), w)] \right\|.$$

Notice that Assumption C.2 indicates that

$$\left\| \overline{\mathbb{E}}[\psi(z(t - m))F(X(t), w)] \right\| \leq L_1(\|w\| + 1),$$

and Lemma C.18 implies that

$$\left\| \mathbb{E}[\psi(z(t - m))F(X(t), w)] - \overline{\mathbb{E}}[\psi(z(t - m))F(X(t), w)] \right\|$$
$$\leq 2L_1(\|w\| + 1)c'\rho'^{t - \max\{t_0, m\}}.$$

Plug in the two bounds back, and we have

$$\|\mathbb{E}[G(\zeta(t), X(t), w)] - \overline{G}(w)\|$$

$$\leq \sum_{m=0}^{t_0-1} \lambda^m 2L_1(\|w\|+1)c'\rho'^{t-t_0} + \sum_{m=t_0}^{t} \lambda^m 2L_1(\|w\|+1)c'\rho'^{t-m} + \sum_{m=t+1}^{\infty} \lambda^m L_1(\|w\|+1)$$

$$= L_1(\|w\|+1) \left[2c' \underbrace{\sum_{m=0}^{t_0-1} \lambda^m \rho'^{t-t_0}}_{(a)} + 2c' \underbrace{\sum_{m=t_0}^{t} \lambda^m \rho'^{t-m}}_{(b)} + \underbrace{\sum_{m=t+1}^{\infty} \lambda^m}_{(c)} \right]. \tag{25}$$

where we use Assumption C.2 as well as Lemma C.18. Obviously, $(a) = O(\rho'^t)$ and $(c) = O(\lambda^t)$. For (b), there are three cases:

- $\lambda < \rho'$, then $(b) = O(\rho'^t)$.
- $\lambda = \rho'$, then $(b) = (t - t_0 + 1)\rho'^t$, which also decays exponentially fast with t, with any decay rate less than ρ'.
- $\lambda > \rho'$, then $(b) = \sum_{m'=0}^{t-t_0} \lambda^{t-m'} \rho'^m = O(\lambda^t)$.

In conclusion, (a), (b), and (c) all decays exponentiallly fast with t, with decay rates depending only on λ or ρ'. Let $\rho'(\rho', \lambda)$ be the maximum value among the three decay rates, then there exists some $c_g(c', \rho', \lambda, t_0) > 0$, such that

$$2c' \sum_{m=0}^{t_0-1} \lambda^m \rho'^{t-t_0} + 2c' \sum_{m=t_0}^{t} \lambda^m \rho'^{t-m} + \sum_{m=t+1}^{\infty} \lambda^m \leq c_g(c', \rho', \lambda, t_0)[\rho_g(\rho', \lambda)]^t.$$

Plug into Eq. (25), and we complete the proof. □

C.5.3 Proof of Lemma C.11

The key is to notice that

$$\|\zeta(t)\| \leq \sum_{k=0}^{t} \lambda^{t-k} \|\psi(z(k))\|$$

$$\leq \sum_{k=0}^{t} \lambda^{t-k}$$

$$< \frac{1}{1-\lambda}.$$

Then by Assumption C.2, we get

$$\|G(\zeta(t), x, w_1) - G(\zeta(t), x, w_2)\|$$
$$= \|\zeta(t)[F(x, w_1) - F(x, w_2)]\|$$
$$\leq \|\zeta(t)\| \, |F(x, w_1) - F(x, w_2)|$$
$$\leq \frac{1}{1-\lambda} \cdot L_1 \|w_1 - w_2\|$$
$$= L_1' \|w_1 - w_2\|.$$

Furthermore, we have

$$\|G(\zeta(t), x, 0)\|$$
$$= \|\zeta(t)F(x, 0)\|$$
$$\leq \|\zeta(t)\| \, |F(x, 0)|$$
$$\leq \frac{L_1}{1-\lambda}$$
$$= L_1'.$$

C.5.4 Proof of Lemma C.12

By Lemma C.11, we have for ant t that

$$
\begin{aligned}
\|w(t)\| - \|w(t-1)\| &\leq \|w(t) - w(t-1)\| \\
&= \alpha \|G(\zeta_{t-1}, X(t-1), w(t-1)\| \\
&\leq \alpha L'_1(\|w(t-1)\| + 1).
\end{aligned}
$$

So $\|w(t)\| + 1 \leq (1 + \alpha L'_1)(\|w(t-1)\| + 1)$. Then we have for any $t \geq t_2$,

$$
\|w(t)\| + 1 \leq (1 + \alpha L'_1)^{t-t_2}(\|w(t_2)\| + 1).
$$

Therefore, we get

$$
\begin{aligned}
&\|w(t_1) - w(t_2)\| \\
&\leq \sum_{t'=t_2}^{t_1-1} \|w(t'+1) - w(t')\| \\
&\leq \sum_{t'=t_2}^{t_1-1} \alpha L'_1(\|w(t')\| + 1) \\
&\leq \sum_{t'=t_2}^{t_1-1} \alpha L'_1(1 + \alpha L'_1)^{t'-t_2}(\|w(t_2)\| + 1) \\
&= \left[(1 + \alpha L'_1)^{t_1-t_2} - 1 \right] (\|w(t_2)\| + 1).
\end{aligned}
$$

Notice that $(1 + x)^p \leq 1 + 2px$ for any $p > 0$ and $x \in [0, \frac{1}{2p}]$. So for $\alpha(t_1 - t_2) \leq \frac{1}{2L'_1}$, we have

$$
(1 + \alpha L'_1)^{t_1-t_2} \leq 1 + 2L'_1 \alpha(t_1 - t_2).
$$

Therefore, we have $\|w(t_1) - w(t_2)\| \leq 2L'_1 \alpha(t_1 - t_2)(\|w(t_2)\| + 1)$.
Since $\alpha(t_1 - t_2) \leq \frac{1}{4L'}$, we further have

$$
\begin{aligned}
&\|w(t_1) - w(t_2)\| \\
&\leq 2L'_1 \alpha(t_1 - t_2)(\|w(t_2)\| + 1) \\
&\leq 2L'_1 \alpha(t_1 - t_2)(\|w(t_1) - w(t_2)\| + \|w(t_1)\| + 1) \\
&\leq \frac{1}{2} \|w(t_1) - w(t_2)\| + 2L'_1 \alpha(t_1 - t_2)(\|w(t_1)\| + 1).
\end{aligned}
$$

Rearrange the terms, and we have $\|w(t_1) - w(t_2)\| \leq 4L'_1 \alpha(t_1 - t_2)(\|w(t_1)\| + 1)$.

C.5.5 Proof of Lemma C.13

We can decompose (a_2) as

$$
\begin{aligned}
(a_2) =& \mathbb{E}_\alpha \left[(w(t) - w^*)^\top \left(G(\zeta(t), X(t), w(t)) - \overline{G}(w(t)) \right) \right] \\
=& \mathbb{E}_\alpha \left[(w(t) - w(t - t'_\alpha)^\top \left(G(\zeta(t), X(t), w(t)) - \overline{G}(w(t)) \right) \right] && (b_1) \\
&+ (w(t - t'_\alpha) - w^*)^\top \mathbb{E}_\alpha \left[G(\zeta(t), X(t), w(t - t'_\alpha)) - \overline{G}(w(t - t'_\alpha)) \right] && (b_2) \\
&+ \mathbb{E}_\alpha \left[(w(t - t'_\alpha) - w^*)^\top \left(G(\zeta(t), X(t), w(t)) - G(\zeta(t), X(t), w(t - t'_\alpha)) \right) \right] && (b_3) \\
&+ \mathbb{E}_\alpha \left[(w(t - t'_\alpha) - w^*)^\top \left(\overline{G}(w(t - t'_\alpha)) - \overline{G}(w(t)) \right) \right]. && (b_4)
\end{aligned}
$$

For term (b_1), we apply Lemma C.12 with $t_1 = t$ and $t_2 = t - t'_\alpha$. Thus we get

$$
\begin{aligned}
(b_1) &\leq \mathbb{E}_\alpha \left[\|w(t) - w(t - t'_\alpha)\| \, \|G(\zeta(t), X(t), w(t) - \overline{G}(w(t))\| \right] \\
&\leq \mathbb{E}_\alpha \left[\|w(t) - w(t - t'_\alpha)\| \left(\|G(\zeta(t), X(t), w(t)\| + \|\overline{G}(w(t))\| \right) \right] \\
&\leq 4L'_1 \alpha t'_\alpha \mathbb{E}_\alpha \left[(\|w(t)\| + 1) \left(\|G(\zeta(t), X(t), w(t)\| + \|\overline{G}(w(t))\| \right) \right] \quad \text{(Lemma C.12)} \\
&\leq 8L'^2_1 \alpha t'_\alpha \mathbb{E}_\alpha \left[(\|w(t)\| + 1)^2 \right] \\
&\leq 8L'^2_1 \alpha t'_\alpha \mathbb{E}_\alpha \left[(\|w(t) - w^*\| + \|w^*\| + 1)^2 \right] \\
&\leq 16L'^2_1 \alpha t'_\alpha \mathbb{E}_\alpha \left[\|w(t) - w^*\|^2 + (\|w^*\| + 1)^2 \right] \\
&= 16L'^2_1 \alpha t'_\alpha \left(\mathbb{E}_\alpha \left[\|w(t) - w^*\|^2 \right] + (\|w^*\| + 1)^2 \right). \quad (26)
\end{aligned}
$$

Now we bound (b_2). We have by Definition C.4 that

$$
\begin{aligned}
(b_2) &\leq \|w(t - t'_\alpha) - w^*\| \, \|\mathbb{E}_\alpha \left[(G(\zeta(t), X(t), w(t - t'_\alpha)) - \overline{G}(w(t - t'_\alpha)) \right] \| \\
&\leq L_1 \alpha (\|w(t - t'_\alpha)\| + 1) \|w(t - t'_\alpha) - w^*\| \\
&= L_1 \alpha \mathbb{E}_\alpha \left[(\|w(t - t'_\alpha)\| + 1) \|w(t - t'_\alpha) - w^*\| \right].
\end{aligned}
$$

We further bound (b_2). By Lemma C.12 and the fact that $\alpha t'_\alpha \leq \frac{1}{4L'_1}$, we have

$$
\|w(t) - w(t - t'_\alpha)\| \leq 4L'_1 \alpha t'_\alpha (\|w(t)\| + 1) \leq \|w(t)\| + 1. \quad (27)
$$

Thus we have

$$
\begin{aligned}
&(\|w(t - t'_\alpha)\| + 1) \|w(t - t'_\alpha) - w^*\| \\
\leq &(\|w(t) - w(t - t'_\alpha)\| + \|w(t) - w^*\| + \|w^*\| + 1)(\|w(t) - w(t - t'_\alpha)\| + \|w(t) - w^*\|) \\
\leq &(\|w(t)\| + \|w(t) - w^*\| + \|w^*\| + 2)(\|w(t)\| + \|w(t) - w^*\| + 1) \\
\leq &(2 \|w(t) - w^*\| + 2 \|w^*\| + 2)(2 \|w(t) - w^*\| + \|w^*\| + 1) \\
\leq &4(\|w(t) - w^*\| + \|w^*\| + 1)^2 \\
\leq &8 \left[\|w(t) - w^*\|^2 + (\|w^*\| + 1)^2 \right].
\end{aligned}
$$

So we can bound (b_2) as

$$
(b_2) \leq 8L_1 \alpha \mathbb{E}_\alpha \left[\|w(t) - w^*\|^2 + (\|w^*\| + 1)^2 \right]. \quad (28)
$$

Finally we bound the sum of (b_3) and (b_4). We apply Lemma C.11 and get

$$
\begin{aligned}
(b_3) + (b_4) \leq &2L'_1 \mathbb{E}_\alpha \left[\|w(t - t'_\alpha) - w^*\| \, \|w(t) - w(t - t'_\alpha)\| \right] \\
\leq &8L'^2_1 \alpha t'_\alpha \mathbb{E}_\alpha \left[\|w(t - t'_\alpha) - w^*\| (\|w(t)\| + 1) \right] \quad \text{(Lemma C.12)} \\
\leq &8L'^2_1 \alpha t'_\alpha \mathbb{E}_\alpha \left[(\|w(t) - w(t - t'_\alpha)\| + \|w(t) - w^*\|)(\|w(t) - w^*\| + \|w^*\| + 1) \right] \\
\leq &8L'^2_1 \alpha t'_\alpha \mathbb{E}_\alpha \left[(\|w(t)\| + \|w(t) - w^*\| + 1)(\|w(t) - w^*\| + \|w^*\| + 1) \right] \quad \text{(Eq. (27))} \\
\leq &8L'^2_1 \alpha t'_\alpha \mathbb{E}_\alpha \left[(2 \|w(t) - w^*\| + \|w^*\| + 1)(\|w(t) - w^*\| + \|w^*\| + 1) \right] \\
\leq &16L'^2_1 \alpha t'_\alpha \mathbb{E}_\alpha \left[(\|w(t) - w^*\| + \|w^*\| + 1)^2 \right] \\
\leq &32L'^2_1 \alpha t'_\alpha \mathbb{E}_\alpha \left[\|w(t) - w^*\|^2 + (\|w^*\| + 1)^2 \right]. \quad (29)
\end{aligned}
$$

Now we combine the bounds for $(b_1), (b_2), (b_3), (b_4)$ (Eq. (26), (28), (29)) and get

$$
\begin{aligned}
(a_2) = &(b_1) + (b_2) + (b_3) + (b_4) \\
\leq &16L'^2_1 \alpha t'_\alpha \left(\mathbb{E}_\alpha \left[\|w(t) - w^*\|^2 \right] + (\|w^*\| + 1)^2 \right) \\
&+ 8L_1 \alpha \mathbb{E}_\alpha \left[\|w(t) - w^*\|^2 + (\|w^*\| + 1)^2 \right] \\
&+ 32L'^2_1 \alpha t'_\alpha \mathbb{E}_\alpha \left[\|w(t) - w^*\|^2 + (\|w^*\| + 1)^2 \right] \\
\leq &56L'^2_1 \alpha t'_\alpha \mathbb{E}_\alpha \left[\|w(t) - w^*\|^2 + (\|w^*\| + 1)^2 \right],
\end{aligned}
$$

where the last inequality is derived by $1 < L_1 < L'_1$ (cf. Assumption C.2) and $t'_\alpha \geq 1$.

C.5.6 Proof of Lemma C.14

By Assumption C.3, we have for any t

$$\left| \hat{C}(z_{N_{i_0}^{\kappa_c}}, w(t)) - \hat{C}(z, w^*) \right| \leq L_2 \left\| w(t) - w^* \right\|.$$

So we get for $K \geq t'_\alpha$ that

$$\mathbb{E}\left[\sup_z \left| \hat{C}(z_{N_{i_0}^{\kappa_c}}, w(K)) - \hat{C}(z_{N_{i_0}^{\kappa_c}}, w^*) \right|^2 \right]$$
$$\leq \mathbb{E}\left[L_2^2 \left\| w(K) - w^* \right\|^2 \right]$$
$$\leq L_2^2 \left(c_1 (1 - \alpha c_0)^{K - t'_\alpha} + c_2 \frac{\alpha t'_\alpha}{c_0} \right).$$

C.5.7 Proof of Lemma C.15

Denote by $\pi_{\kappa_r, t}$ the marginal distribution of $\mathcal{Z}_{N_{i_0}^{\kappa_r}}$ at time t in the original Markov chain (state space $\mathcal{Z}$). Let $\tilde{\pi}_{\kappa_r, t}$ denote the marginal state distribution at time t in the sub-chain (state space $\mathcal{Z}_{N_{i_0}^{\kappa_c}}$.)

Due to the local dependence of transition probability (cf. Eq. (11)), $\pi_{\kappa_r, t}$ is only dependent on the initial states of agents in $N_{i_0}^{\kappa_r + t}$, which is equal to $\tilde{\pi}_t^{\kappa_r}$ when $t \leq \kappa_c - \kappa_r$. Therefore we have for any $z \in \mathcal{Z}$

$$\left| \tilde{C}(z_{N_{i_0}^{\kappa_c}}) - C(z) \right|$$
$$\leq \sum_{t=0}^{\infty} \left| \mathbb{E}\left[\gamma^t \tilde{r}(z_{N_{i_0}^{\kappa_r}}(t)) | z_{N_{i_0}^{\kappa_c}}(0) = z_{N_{i_0}^{\kappa_c}} \right] - \mathbb{E}\left[\gamma^t \tilde{r}(z_{N_{i_0}^{\kappa_r}}(t)) | z(0) = z \right] \right|$$
$$= \sum_{t=0}^{\infty} \left| \gamma^t \mathop{\mathbb{E}}_{z_{N_{i_0}^{\kappa_r}} \sim \tilde{\pi}_{\kappa_r, t}} \tilde{r}(z_{N_{i_0}^{\kappa_r}}) - \gamma^t \mathop{\mathbb{E}}_{z_{N_{i_0}^{\kappa_r}} \sim \pi_{\kappa_r, t}} \tilde{r}(z_{N_{i_0}^{\kappa_r}}) \right|$$
$$= \sum_{t=\kappa_c - \kappa_r + 1}^{\infty} \left| \gamma^t \mathop{\mathbb{E}}_{z_{N_{i_0}^{\kappa_r}} \sim \tilde{\pi}_{\kappa_r, t}} \tilde{r}(z_{N_{i_0}^{\kappa_r}}) - \gamma^t \mathop{\mathbb{E}}_{z_{N_{i_0}^{\kappa_r}} \sim \pi_{\kappa_r, t}} \tilde{r}(z_{N_{i_0}^{\kappa_r}}) \right|$$
$$\leq \sum_{t=\kappa_c - \kappa_r + 1}^{\infty} \gamma^t \text{TV}(\tilde{\pi}_{\kappa_r, t}, \pi_{\kappa_r, t})$$
$$\leq \frac{1}{1 - \gamma} \gamma^{\kappa_c - \kappa_r + 1}.$$

Take $\sup_z$ on both sides above, we have

$$\sup_z \left| \tilde{C}(z_{N_{i_0}^{\kappa_c}}) - C(z) \right| \leq \frac{1}{1 - \gamma} \gamma^{\kappa_c - \kappa_r + 1}.$$

D PROOF OF THEOREM 5.8

In this section, we first derive the uniform properties of ϵ-exploration policies, which are prerequisites for the critic error bound. Then we show that assumptions of Appendix C can be satisfied by Algorithm 2 under certain conditions. After that we restate Theorem 5.8. Finally, we give the proof of the theorem, which is based on the results of localized stochastic approximation (Appendix C).

D.1 UNIFORMITY OF ϵ-EXPLORATION POLICIES

In order to derive a uniform critic error bound for all policies, we need uniform properties, , such as convergence rate, exploration, for the critic sampling policies, i.e., ϵ-exploration policy class Ξ^ϵ. This can be done by applying the results of Zhang et al. [2022c] when Assumption 5.3 holds.

Lemma D.1. *For any policy $\hat{\xi}$ from Ξ^ϵ for $\epsilon > 0$, the induced Markov chain $\{(s(t), a(t))\}$ is aperiodic and irreducible.*

Proof of Lemma D.1. The result is obvious by point 1. of [Zhang et al., 2022c, Proposition 3] and the fact that $\hat{\xi}(a|s) \geq \prod_{i \in \mathcal{N}} \frac{\epsilon}{|\mathcal{A}_i|} = \frac{\epsilon^n}{|\mathcal{A}|} > 0$ for any s, a. □

For any policy $\hat{\xi} \in \Xi^\epsilon$, we use the notation $\pi_t^{\hat{\xi}}(s, a)$ to denote the probability of (s, a) at time t and use $\overline{\pi}^{\hat{\xi}}(s, a)$ for the probability of (s, a) under stationary distribution. According to Lemma D.1, for any $\hat{\xi} \in \Xi^\epsilon$, there exists $c_{\hat{\xi}} > 1$ and $\rho_{\hat{\xi}} \in (0, 1)$ such that

$$\mathrm{TV}(\pi_t^{\hat{\xi}}, \overline{\pi}^{\hat{\xi}}) \leq c_{\hat{\xi}} \rho_{\hat{\xi}}^t. \tag{30}$$

Let $\overline{c} := \sup_{\hat{\xi} \in \Xi^\epsilon} c_{\hat{\xi}}$, $\overline{\rho} := \sup_{\hat{\xi} \in \Xi^\epsilon} \rho_{\hat{\xi}}$. Besides, recall the definitions of $\pi_{\min}$ and $\underline{\lambda}$ in Subsection 5.1. The uniformity of Ξ^ϵ can be shown by bounds of $\overline{\rho}$, $\pi_{\min}$ and $\underline{\lambda}$.

Lemma D.2 (Uniformity of ϵ-exploration policy class). *We have*

1. *$\overline{\rho} < 1$. So $\pi_t^{\hat{\xi}}$ has a uniform convergence rate for all $\hat{\xi} \in \Xi^\epsilon$.*
2. *$\pi_{\min} > 0$. So each $(s_{N_i^{\kappa_c}}, a_i)$ pair is visited with positive probability, which has a uniform positive lower bound for all ϵ-exploration polices.*
3. *$\underline{\lambda} > 0$.*

Proof of Lemma D.2. Fix any $\hat{\xi} \in \Xi^\epsilon$. For part 1, by [Zhang et al., 2022c, Proposition 3], for any $\hat{\xi} \in \Xi^\epsilon$, there exists $\hat{c} > 0$ and $\rho_0 < 1$, such that $\sup_{\hat{\xi} \in \Xi^\epsilon} \max_{s \in \mathcal{S}} \mathrm{TV}(\pi_{\mathcal{S},t}^{\hat{\xi}}(\cdot|s), \overline{\pi}_{\mathcal{S}}^{\hat{\xi}}) \leq \hat{c} \rho_0^t$ for any $t \geq 0$. Here $\pi_{\mathcal{S},t}^{\hat{\xi}}(s'|s) = \mathrm{Pr}_{\hat{\xi}}[s(t) = s'|s(0) = s]$ for any $s, s' \in \mathcal{S}$, and $\overline{\pi}_{\mathcal{S}}^{\hat{\xi}} \in \Delta(\mathcal{S})$ is the stationary distribution of $\mathcal{S}$ under policy $\hat{\xi}$. Then

$$\begin{aligned}
\mathrm{TV}(\pi_t^{\hat{\xi}}, \overline{\pi}^{\hat{\xi}}) =& \frac{1}{2} \sum_{s,a} \left| \overline{\pi}_t^{\hat{\xi}}(s, a) - \overline{\pi}^{\hat{\xi}}(s, a) \right| \\
=& \frac{1}{2} \sum_{s,a} \left| \mathbb{E}_{s^0 \sim \mu} \left[\pi_{\mathcal{S},t}^{\hat{\xi}}(s|s^0) \hat{\xi}(a|s) \right] - \overline{\pi}_{\mathcal{S}}^{\hat{\xi}}(s) \hat{\xi}(a|s) \right| \\
\leq& \frac{1}{2} \sum_{s,a} \mathbb{E}_{s^0 \sim \mu} \left[\left| \pi_{\mathcal{S},t}^{\hat{\xi}}(s|s^0) \hat{\xi}(a|s) - \overline{\pi}_{\mathcal{S}}^{\hat{\xi}}(s) \hat{\xi}(a|s) \right| \right] \\
=& \frac{1}{2} \mathbb{E}_{s^0 \sim \mu} \left[\sum_{s,a} \left| \pi_{\mathcal{S},t}^{\hat{\xi}}(s|s^0) - \overline{\pi}_{\mathcal{S}}^{\hat{\xi}}(s) \right| \hat{\xi}(a|s) \right] \\
=& \mathbb{E}_{s^0 \sim \mu} \left[\mathrm{TV}(\pi_{\mathcal{S},t}^{\hat{\xi}}(\cdot|s), \overline{\pi}_{\mathcal{S}}^{\hat{\xi}}) \right] \\
\leq& \hat{c} \rho_0^t.
\end{aligned}$$

This shows that $\overline{\rho} \leq \rho_0 < 1$.

As for part 2, by 3. of [Zhang et al., 2022c, Proposition 3], $\underline{\pi} := \inf_{\hat{\xi} \in \Xi^\epsilon} \min_{s \in \mathcal{S}} \overline{\pi}_{\mathcal{S}}^{\hat{\xi}}(s) > 0$. Thus for any $s_{N_i^{\kappa_c}} \in \mathcal{S}_{N_i^{\kappa_c}}, a_i \in \mathcal{A}_i$ and any $\hat{\xi} \in \Xi^\epsilon$,

$$\begin{aligned}
& \overline{\pi}^{\hat{\xi}}\left(s_{N_i^{\kappa_c}}, a_i\right) \\
=& \sum_{s_{-N_i^{\kappa_c}}, a_{-i}} \overline{\pi}^{\hat{\xi}}\left(s_{N_i^{\kappa_c}}, s_{-N_i^{\kappa_c}}, a_i, a_{-i}\right) \\
\geq& \underline{\pi}
\end{aligned}$$

Thus $\pi_{\min} \geq \underline{\pi} > 0$.

Finally we show part 3. For any square matrix X, denote by $\lambda_{\min}(X)$ the minimum eigenvalue of X. By definition of $\pi_{\min}$ (see Subsection 5.1), for any $\hat{\xi} \in \Xi^\epsilon$, $\lambda_{\min}(D^{\hat{\xi}}) \geq \pi_{\min}$.

Since Ω_i is full column rank, $\Omega_i^\top \Omega$ is positive definite, so $\lambda_{\min}(\Omega_i^\top \Omega_i) > 0$ for any $i \in \mathcal{N}$.

For any eigenvalue λ of $\Omega_i^\top D^{\hat{\xi}} \Omega_i$, let $x \in \mathbb{R}^{d_i}$ be the corresponding eigenvector. Then

$$
\begin{aligned}
\lambda \|x\|^2 &= \lambda x^\top x \\
&= x^\top \Omega_i^\top D^{\hat{\xi}} \Omega_i x \\
&\geq \pi_{\min} \|\Omega_i x\|^2 \\
&= \pi_{\min} x^T \Omega_i^\top \Omega_i x \\
&= \pi_{\min} \lambda_{\min}(\Omega_i^\top \Omega_i) \|x\|^2.
\end{aligned}
$$

Therefore $\lambda \geq \pi_{\min} \lambda_{\min}(\Omega_i^\top \Omega_i)$, and thus $\underline{\lambda}_i^{\hat{\xi}} \geq \pi_{\min} \lambda_{\min}(\Omega_i^\top \Omega_i)$. As a result, $\underline{\lambda} \geq \pi_{\min} \lambda_{\min}(\Omega_i^\top \Omega_i) > 0$. $\qquad \square$

Besides the uniform convergence rate of the distribution $(s(t), a(t))$, we also want to establish the uniform convergence of the marginal distribution $(s(t), a_i(t))$ for all $i \in \mathcal{N}$.

For any agent $i \in \mathcal{N}$ and policy $\hat{\xi} \in \Xi^\epsilon$, we use the notation $\pi_{i,t}^{\hat{\xi}}(s, a_i)$ to denote the marginal probability of (s, a_i) at time t and use $\overline{\pi}_i^{\hat{\xi}}(s, a_i)$ for the marginal probability of (s, a_i) under stationary distribution. The following lemma shows that the distribution $(s(t), a_i(t))$ converges as fast as the (s, a), with a uniform convergence rate for all ϵ-exploration policies.

Lemma D.3. *The distribution of $(s(t), a_i(t))$ has uniform convergence rate*

$$
\mathrm{TV}(\pi_{i,t}^{\hat{\xi}}, \overline{\pi}_i^{\hat{\xi}}) \leq \overline{c} \cdot \overline{\rho}^t.
$$

Here $\overline{c}$ and $\overline{\rho}$ are the uniform convergence rate of the distribution of $(s(t), a(t))$ (see Appendix D.1).

Proof of Lemma D.3. We have

$$
\begin{aligned}
&\mathrm{TV}(\pi_{i,t}^{\hat{\xi}}, \overline{\pi}_i^{\hat{\xi}}) \\
=& \frac{1}{2} \sum_{s, a_i} \left| \pi_{i,t}^{\hat{\xi}}(s, a_i) - \overline{\pi}_i^{\hat{\xi}} \right| \\
=& \frac{1}{2} \sum_{s, a_i} \left| \sum_{a_{-i}} \pi_t^{\hat{\xi}}(s, a_i, a_{-i}) - \overline{\pi}^{\hat{\xi}}(s, a_i, a_{-i}) \right| \\
\leq& \frac{1}{2} \sum_{s, a_i} \sum_{a_{-i}} \left| \pi_t^{\hat{\xi}}(s, a_i, a_{-i}) - \overline{\pi}^{\hat{\xi}}(s, a_i, a_{-i}) \right| \\
\leq& \overline{c} \cdot \overline{\rho}^t.
\end{aligned}
$$

Here the last inequality is due to Eq. (30), and $c_{\hat{\xi}} \leq \overline{c}$, $\rho_{\hat{\xi}} \leq \overline{\rho}$. $\qquad \square$

D.2 VERIFYING ASSUMPTIONS OF APPENDIX C

In order to apply the results of localized stochastic approximation (Appendix C), we verify the assumptions needed for Theorem C.7. First, we correspond the problem setting of NMPG to the Markov chain setting of localized stochastic approximation.

We fix any policy $\hat{\xi}$, and any agent $i \in \mathcal{N}$ for the rest of this section. Then $\{(s(t), a_i(t)\}$ forms a Markov chain $\mathcal{M}^{i,\hat{\xi}}$. Recall that $U_j^\kappa = N_j^\kappa/\{j\}$, which is the set of agents within j's κ-hop neighborhood excluding j it self. Construct Markov chain $\mathcal{M}^{i,\hat{\xi}} = (\mathcal{N}, \mathcal{E}, \mathcal{Z}, \mathbb{P}, \tilde{r}, \gamma, \mu')$, and choose $i_0 = i$. Here $\mathcal{N}$, $\mathcal{E}$ and γ have the same meaning as in the NMPG setting (Section

2), which are the set of agents, the edges of the graph and the discount factor, respectively. For the other elements, they are defined as

$$
\mathcal{Z}_j = \begin{cases} \mathcal{S}_j & j \neq i \\ \mathcal{S}_j \times \mathcal{A}_j & j = i \end{cases}
$$

$$
z_j = \begin{cases} s_j & j \neq i \\ (s_j, a_j) & j = i \end{cases}
$$

$$
\mathbb{P}_j(z_j' | z_{N_j}) = \hat{\xi}_j(a_j' | s_j') \mathcal{P}_j(s_j' | s_{N_j}, a_j), \forall j \in \mathcal{N}
$$

$$
\mu'(z) = \hat{\xi}(a|s)\mu(s)
$$

$$
\tilde{r}(z_{N_i^{\kappa r}}) = \bar{r}_i^{\hat{\xi}}(s_{N_i^{\kappa r}}, a_i) := \sum_{a_{U_i^{\kappa r}}} \hat{\xi}_{U_i^{\kappa r}}(a_{U_i^{\kappa r}} | s_{U_i^{\kappa r}}) r_i(s_{N_i^{\kappa r}}, a_i, a_{U_i^{\kappa r}}).
$$

Here $\bar{r}_i^{\hat{\xi}}(s_{N_i^{\kappa r}}, a_i)$ is the expected reward function with respect to $a_{U_i^{\kappa r}}$. Furthermore, cost function $C(z)$ of $\mathcal{M}^{i,\hat{\xi}}$ corresponds to

$$
\begin{aligned}
C(s, a_i) &= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\hat{\xi}} [\bar{r}_i^{\hat{\xi}}(s_{N_i^{\kappa r}}(t), a_i(t)) | s(0) = s, a_i(0) = a_i] \\
&= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\hat{\xi}} [r_i(s_{N_i^{\kappa r}}(t), a_{N_i^{\kappa r}}(t)) | s(0) = s, a_i(0) = a_i] \\
&= \overline{Q}_i^{\hat{\xi}}(s, a_i),
\end{aligned}
$$

which is the averaged Q-function.

Now we represent localized TD(λ) with linear function approximation in the form of generalized TD(λ) (Algorithm 4). Let

$$
\begin{aligned}
& d = d_i, \ t_0 = 0 \\
& X(t) = (z_{N_i^{\kappa c}}(t), z_{N_i^{\kappa c}}(t+1)) \\
& F(z_{N_i^{\kappa c}}(t), z_{N_i^{\kappa c}}(t+1), w) = -\delta_i(t).
\end{aligned} \tag{31}
$$

(Recall the definition of $\delta_i(t)$ in Algorithm 2.) Then Algorithm 4 analyzes the averaged Q-function of agent i.

Next we show that Assumptions C.1, C.2, C.3 are satisfied by localized TD(λ) under the conditions of Theorem 5.8.

Verify Assumption C.1. This can be shown by the following lemma.

Lemma D.4. *For any policy $\hat{\xi} \in \Xi^\epsilon$, $\epsilon > 0$, $i \in \mathcal{N}$, the induced Markov chain $\{(s(t), a_i(t))\}$ is aperiodic and irreducible.*

Proof of Lemma D.4. For any Markov chain with transition probability Γ on some state space $\mathcal{X}$, we write $\Gamma^l(x'|x) = \Pr[x(t+l) = x'|x(t) = x]$, for any $t \in \mathbb{N}$, $x, x' \in \mathcal{X}$.

We fix $i \in \mathcal{N}$. By Lemma $D.2$, the induced Markov chain $\{(s(t), a(t))\}$ is aperiodic and irreducible. We abuse the notation in this proof to use $\mathbb{P}$ for the transition probability of (s, a_i) and use $\tilde{\mathbb{P}}$ for that of (s, a).

For any $s, s' \in \mathcal{S}, a_i, a_i' \in \mathcal{A}_i$, we randomly pick $a_{-i}, a_{-i}' \in \mathcal{A}_{-i}$. Since $\tilde{\mathbb{P}}$ is irreducible, there exists $t > 0$, such that $\tilde{\mathbb{P}}^t(s', a_i', a_{-i}'|s, a_i, a_{-i}) > 0$. We choose the smallest $t_0 > 0$ such that $\tilde{\mathbb{P}}^{t_0}(s', a_i', a_{-i}'|s, a_i, a_{-i}) > 0$ for any $a_{-i}, a_{-i}' \in \mathcal{A}_{-i}$, so $\mathbb{P}_i^t(s', a_i'|s, a_i, a_{-i}) > 0$. $\mathbb{P}_i^t(s', a_i'|s, a_i, a_{-i})$ represents the marginal probability of s', a_i' given previous state-action pair (s, a_i, a_{-i}). Then we have

$$
\begin{aligned}
\mathbb{P}^{t_0}(s', a_i'|s, a_i) &= \sum_{\hat{a}_{-i} \in \mathcal{A}_{-i}} \hat{\xi}_{-i}(\hat{a}_{-i}|s_{-i}) \mathbb{P}_i^{t_0}(s', a_i'|s, a_i, \hat{a}_{-i}) \\
&\geq \min_{\hat{a}_{-i} \in \mathcal{A}_{-i}} \mathbb{P}_i^{t_0}(s', a_i'|s, a_i, \hat{a}_{-i}) \\
&> 0.
\end{aligned}
$$

Therefore, $\mathbb{P}$ is irreducible.

To show that $\mathbb{P}$ is aperiodic, we assume that $\mathbb{P}$ has period $T \geq 2$. Then for any t not divisible by T and any s, a_i, $\mathbb{P}^t(s, a_i | s, a_i) = 0$. For any s, a, since t not divisible by T, we have

$$0 = \mathbb{P}^t(s, a_i | s, a_i) = \sum_{\hat{a}_{-i} \in \mathcal{A}_{-i}} \hat{\xi}_{-i}(\hat{a}_{-i} | s_{-i}) \mathbb{P}_i^t(s, a_i | s, a_i, \hat{a}_{-i}).$$

Thus we have $\mathbb{P}_i^t(s, a_i | s, a_i, \hat{a}_{-i}) = 0$ for some $\hat{a}_{-i}$ and thus $\mathbb{P}_i^t(s, a_i, \hat{a}_{-i} | s, a_i, \hat{a}_{-i}) = 0$. This implies that the period of $(s, a_i, \hat{a}_{-i})$ is at least $T \geq 2$, which contradicts the assumption that $\mathbb{P}$ is aperiodic. (Notice that the period of any state is the same for an irreducible Markov chain.) Hence $\mathbb{P}$ is aperiodic.

In conclusion, $\mathbb{P}$ is irreducible and aperiodic. $\qquad\square$

Verify Assumption C.2. We show that for any agent $i \in \mathcal{N}$ and any policy $\hat{\xi} \in \Xi^\epsilon$, the assumption can be satisfied for $L_1 = 1 + \gamma$ and $c_0 = (1 - \gamma)\underline{\lambda}$.

We first verify 1. of Assumption C.2. For any w_1, w_2, x, we have

$$\begin{aligned}
&\|F(x, w_1) - F(x, w_2)\| \\
&= \left\| \langle \phi_i(s_{N_i^{\kappa c}}(t), a_i(t)), w_1 - w_2 \rangle - \gamma \langle \phi_i(s_{N_i^{\kappa c}}(t+1), a_i(t+1)), w_1 - w_2 \rangle \right\| \\
&\leq \left\| \langle \phi_i(s_{N_i^{\kappa c}}(t), a_i(t)), w_1 - w_2 \rangle \right\| + \gamma \left\| \langle \phi_i(s_{N_i^{\kappa c}}(t+1), a_i(t+1)), w_1 - w_2 \rangle \right\| \\
&\leq (1 + \gamma)\|w_1 - w_2\|.
\end{aligned}$$

For any x, notice that

$$\begin{aligned}
&\|F(x, 0)\| \\
&= \left\| \overline{r}_i^{\hat{\xi}}(s_{N_i^{\kappa r}}(t), a_i(t)) \right\| \\
&\leq 1 \\
&\leq 1 + \gamma.
\end{aligned}$$

Now we prove 2. in Assumption C.2. We point out that our localized TD(λ) can be easily reduced to the classical single-agent TD(λ) algorithm, with $\mathcal{Z}$ being the single-agent state space. Then by Tsitsiklis and Van Roy [1997], $\overline{G}(w)$ has a unique zero w^*. Furthermore, then according to the proof of Lemma 9 in Tsitsiklis and Van Roy [1997], we have

$$\begin{aligned}
\langle w - w^*, \overline{G}(w) \rangle &\leq -(1 - \gamma) \left\| \Omega_i w - \Omega_i w^* \right\|_{D^{\hat{\xi}}} \\
&= -(1 - \gamma)(w - w^*)^\top \Omega_i^\top D^{\hat{\xi}} \Omega_i (w - w^*) \\
&\leq -(1 - \gamma)\underline{\lambda}_k^{\hat{\xi}} \|w - w^*\|^2 \\
&\leq -(1 - \gamma)\underline{\lambda} \|w - w^*\|^2.
\end{aligned}$$

We will sometimes add a subscript i for w^* and write w_i^* to distinguish different agents.

Verify Assumption C.3. Approximate cost function $\hat{C}$ in the localized stochastic approximation problem corresponds to the approximate averaged Q-function of agent i, i.e., $\hat{Q}_i(s_{N_i^{\kappa c}}, a_i, w_i) = \langle \phi_i(s_{N_i^{\kappa c}}, a_i), w_i \rangle$. Since $\left\| \phi_i(s_{N_i^{\kappa c}}, a_i) \right\| \leq 1$, we have for any w_1, w_2

$$\begin{aligned}
&\left| \hat{C}(z_{N_i^{\kappa c}}, w_1) - \hat{C}(z_{N_i^{\kappa c}}, w_2) \right| \\
&= \left| \langle \phi_i(s_{N_i^{\kappa c}}, a_i), w_1 - w_2 \rangle \right| \\
&\leq \left\| \phi_i(s_{N_i^{\kappa c}}, a_i) \right\| \|w_1 - w_2\| \\
&\leq \|w_1 - w_2\|.
\end{aligned}$$

So Assumption C.3 can be satisfied with $L_2 = 1$.

D.3 RESTATEMENT OF THEOREM 5.8

Before giving the precise form of Theorem 5.8, We define two quantities needed for the theorem statement. The first quantity is the function approximation error. Denote by $\mathcal{M}_{N_i^{\kappa_c}}^{\hat{\xi}}$ the sub-chain of $\mathcal{M}^{i,\hat{\xi}}$ (defined in Appendix D.2) with respect to agents in $N_i^{\kappa_c}$, and let $\tilde{Q}_i^{\hat{\xi}}(s_{N_i^{\kappa_c}}, a_i)$ be the cost function of $\mathcal{M}^{i,\hat{\xi}}$ at state $(s_{N_i^{\kappa_c}}, a_i)$.

Definition D.5. *For any policy $\hat{\xi} \in \Xi^\epsilon$, and any agent $i \in \mathcal{N}$, define the function approximation error as*

$$\epsilon_{app} := \sup_{\hat{\xi} \in \Xi^\epsilon} \sup_{i \in \mathcal{N}} \inf_w \sup_{s,a_i} \left| \hat{Q}_i(s_{N_i^{\kappa_c}}, a_i, w) - \tilde{Q}_i^{\hat{\xi}}(s_{N_i^{\kappa_c}}, a_i) \right|.$$

Notice that we measure the representation power of the function approximation with respect to the sub-chain instead of the original MDP.

The second quantity we need is the uniform mixing time of the update term $-\delta_i(t)\zeta_i^{\kappa_c}(t)$ in Algorithm 2, corresponding to function G in Appendix C. Notice update term $-\delta_i(t)\zeta_i^{\kappa_c}(t)$ does not depend on policy $\hat{\xi}$, and there is a uniform decay rate of $(s(t), a_i(t))$ for any i and any policy $\hat{\xi} \in \Xi^\epsilon$ (Lemma D.3). Then we can apply Lemma C.16 for all policy $\hat{\xi} \in \Xi^\epsilon$ and get a uniform decay rate $c_{\text{TD}} := c_g(\bar{c}, \bar{\rho}, \lambda, 0)$, $\rho_{\text{TD}} := \rho_g(\bar{\rho}, \lambda)$. This indicates that there is a uniform "mixing time of function G with precision δ" (Definition C.4) for all policies $\hat{\xi} \in \Xi^\epsilon$ and all $\delta > 0$, which we denote by t_δ. In addition, $t_\delta = O(\log \frac{1}{\delta})$.

We now give a stronger version of Theorem 5.8. We can derive Theorem 5.8 by taking square root on both sides and apply Jenson's inequality on the right hand side.

Theorem D.6 (Restatement of Theorem 5.8). *Recall the definition of uniform mixing time t_α above. Choose step size α such that $\alpha t_\alpha \leq \min\left\{ \frac{1-\lambda}{4(1+\gamma)}, \frac{(1-\gamma)(1-\lambda)^2 \underline{\Delta}}{114(1+\gamma)^2} \right\}$. Then we have for any $K \geq t_\alpha$ and any $i \in \mathcal{N}$*

$$\epsilon_{critic}^2 = \max_{i \in \mathcal{N}} \sup_\theta \mathbb{E}\left[\max_{s,a_i} \left| \hat{Q}_i(s_{N_i^{\kappa_c}}, a_i, w_i(K)) - \overline{Q}_i^\theta(s, a_i) \right|^2 \right]$$

$$\leq 4 \left[c_1^*(1 - (1-\gamma)\underline{\lambda}\alpha)^{K-t_\alpha} + c_2^* \frac{\alpha t_\alpha}{(1-\gamma)\underline{\lambda}} + \left(\frac{(1-\lambda\gamma)\epsilon_{app}}{\pi_{\min}(1-\gamma)} \right)^2 \right.$$

$$\left. + \left(\frac{\gamma^{\kappa_c - \kappa_r + 1}}{1-\gamma} \right)^2 + \left(\frac{6n\epsilon}{(1-\gamma)^2} \right)^2 \right].$$

Here $c_1^* = (\max_{i \in \mathcal{N}} \|w_i^*\| + 1)^2$, $c_2^* = 114 \left(\frac{1+\gamma}{1-\lambda} \right)^2 (\max_{i \in \mathcal{N}} \|w_i^*\| + 1)^2$.

D.4 PROOF OF THEOREM D.6

Throughout the proof, we will fix agent $i \in \mathcal{N}$ and policy ξ^θ. For any policy parameter θ, let $\hat{\xi}$ be the ϵ-exploration policy of ξ^θ. We can decompose the critic error as

$$\mathbb{E}\left[\max_{s,a_i} \left| \hat{Q}_i(s_{N_i^{\kappa_c}}, a_i, w_i(K)) - \overline{Q}_i^\theta(s, a_i) \right|^2 \right]$$

$$\overset{(i)}{\leq} \mathbb{E}\left[\max_{s,a_i}(3+1) \left(\frac{1}{3} \left| \hat{Q}_i(s_{N_i^{\kappa_c}}, a_i, w_i(K)) - \overline{Q}_i^{\hat{\xi}}(s, a_i) \right|^2 + \left| \overline{Q}_i^{\hat{\xi}}(s, a_i) - \overline{Q}_i^\theta(s, a_i) \right|^2 \right) \right]$$

$$\overset{(ii)}{\leq} (3+1) \left\{ \mathbb{E}\left[\frac{1}{3} \max_{s,a_i} \left| \hat{Q}_i(s_{N_i^{\kappa_c}}, a_i, w_i(K)) - \overline{Q}_i^{\hat{\xi}}(s, a_i) \right|^2 \right] + \mathbb{E}\left[\max_{s,a_i} \left| \overline{Q}_i^{\hat{\xi}}(s, a_i) - \overline{Q}_i^\theta(s, a_i) \right|^2 \right] \right\}$$

$$\overset{(iii)}{=} 4 \left\{ \frac{1}{3} \underbrace{\mathbb{E}\left[\max_{s,a_i} \left| \hat{Q}_i(s_{N_i^{\kappa_c}}, a_i, w_i(K)) - \overline{Q}_i^{\hat{\xi}}(s, a_i) \right|^2 \right]}_{(a)} + \underbrace{\max_{s,a_i} \left| \overline{Q}_i^{\hat{\xi}}(s, a_i) - \overline{Q}_i^{\xi^\theta}(s, a_i) \right|^2}_{(b)} \right\}. \quad (32)$$

Here (i) uses Cauchy-Schwarz inequality, (ii) uses the fact that $\max_x \{f(x) + g(x)\} \le \max_x f(x) + \max_x g(x)$ for any two functions $f(x), g(x)$. For (iii), notice that $Q_i^\theta(s, a_i)$ is the short-hand notation of $Q_i^{\xi^\theta}(s, a_i)$, and term (b) is irrelevant of the sampled trajectory.

(a) is the policy evaluation error of estimating the value function of policy $\hat\xi$, which can be derived using the result in Appendix C. (b) is the difference of value function between policy $\hat\xi$ and ξ^θ, which can be bounded by a function of ϵ.

Bound (a). With all assumptions in Appendix C satisfied (Appendix D.2), the following result is a direct application of Theorem C.7 in localized TD(λ) with linear function approximation:

Corollary D.7. *Choose step size α such that $\alpha t_\alpha \le \min\left\{\frac{1-\lambda}{4(1+\gamma)}, \frac{(1-\gamma)(1-\lambda)^2\underline\lambda}{114(1+\gamma)^2}\right\}$. Then for $i_0 \ge t_\alpha$, we have*

$$\mathbb{E}\left[\max_{s,a_i}\left|\hat{Q}_i(s_{N_i^{\kappa_c}}, a_i, w_i(K)) - \overline{Q}_i^{\hat\xi}(s, a_i)\right|^2\right]$$

$$\le 3\left[c_1^*(1 - (1-\gamma)\underline\lambda\alpha)^{K-t_\alpha} + c_2^*\frac{\alpha t_\alpha}{(1-\gamma)\underline\lambda} + \epsilon_{i,red}^2 + \left(\frac{\gamma^{\kappa_c - \kappa_r + 1}}{1-\gamma}\right)^2\right].$$

Here $c_1^* = (\max_{i\in\mathcal{N}} \|w_i^*\| + 1)^2$, $c_2^* = 114\left(\frac{1+\gamma}{1-\lambda}\right)^2(\max_{i\in\mathcal{N}}\|w_i^*\| + 1)^2$, $\epsilon_{i,red} = \sup_{s_{N_i^{\kappa_c}}, a_i}\left|\hat{Q}_i(s_{N_i^{\kappa_c}}, a_i, w_i^*) - \tilde{Q}_i^{\hat\xi}(s_{N_i^{\kappa_c}}, a_i)\right|$.

For localized TD(λ) with linear function approximation, we can associate $\epsilon_{i,red}$ with function approximation error ϵ_{app} in a similar way as Tsitsiklis and Van Roy [1997].

Lemma D.8. *Agent i's reduction error can be bounded by the function approximation error*

$$\epsilon_{i,red} \le \frac{1}{\pi_{\min}} \cdot \frac{1-\lambda\gamma}{1-\gamma}\epsilon_{app}.$$

Proof of Lemma D.8. Denote by $\overline\pi_{\kappa_c}^{\hat\xi} \in \Delta(\mathcal{S}_{N_i^{\kappa_c}} \times \mathcal{A}_i)$ the marginal stationary distribution of $(s_{N_i^{\kappa_c}}, a_i)$. Define

$$\tilde{D} = \mathop{\mathrm{diag}}_{(s_{N_i^{\kappa_c}}, a_i)\in\mathcal{S}_{N_i^{\kappa_c}}\times\mathcal{A}_i}\{\overline\pi_{\kappa_c}^{\hat\xi}(s_{N_i^{\kappa_c}}, a_i)\}.$$

Then

$$\epsilon_{i,red} \le \frac{1}{\min_{s_{N_i^{\kappa_c}}, a_i}\{\overline\pi_{\kappa_c}^{\hat\xi}(s_{N_i^{\kappa_c}}, a_i)\}}\left\|\hat{Q}_i(\cdot,\cdot,w^*) - \tilde{Q}_i^{\hat\xi}(\cdot,\cdot)\right\|_{\tilde{D}}$$

$$\overset{(i)}{\le} \frac{1}{\min_{s_{N_i^{\kappa_c}}, a_i}\{\overline\pi_{\kappa_c}^{\hat\xi}(s_{N_i^{\kappa_c}}, a_i)\}} \cdot \frac{1-\lambda\gamma}{1-\gamma}\inf_w\left\|\hat{Q}_i(\cdot,\cdot,w) - \tilde{Q}_i^{\hat\xi}(\cdot,\cdot)\right\|_{\tilde{D}}$$

$$\le \frac{1}{\min_{s_{N_i^{\kappa_c}}, a_i}\{\overline\pi_{\kappa_c}^{\hat\xi}(s_{N_i^{\kappa_c}}, a_i)\}} \cdot \frac{1-\lambda\gamma}{1-\gamma}\inf_w\sup_{s,a_i}\left|\hat{Q}_i(s_{N_i^{\kappa_c}}, a_i, w) - \tilde{Q}_i^{\hat\xi}(s, a_i)\right|$$

$$\le \frac{1}{\min_{s_{N_i^{\kappa_c}}, a_i}\{\overline\pi_{\kappa_c}^{\hat\xi}(s_{N_i^{\kappa_c}}, a_i)\}} \cdot \frac{1-\lambda\gamma}{1-\gamma}\epsilon_{app}$$

$$\overset{(ii)}{\le} \frac{1}{\pi_{\min}} \cdot \frac{1-\lambda\gamma}{1-\gamma}\epsilon_{app}.$$

Here in (i), we applied [Tsitsiklis and Van Roy, 1997, Lemma 6] to the sub-chain $\mathcal{M}_{N_i^{\kappa_c}}^{\hat\xi}$, and in (ii), we used the fact that $\min_{s_{N_i^{\kappa_c}}, a_i}\{\overline\pi_{\kappa_c}^{\hat\xi}(s_{N_i^{\kappa_c}}, a_i)\} \ge \pi_{\min}$. $\square$

Combining Lemma D.8 and Corollary D.7, we have the bound of (a) immediately.

Corollary D.9. *Choose step size α such that $\alpha t_\alpha \leq \min \left\{ \frac{1-\lambda}{4(1+\gamma)}, \frac{(1-\gamma)(1-\lambda)^2 \underline{\lambda}}{114(1+\gamma)^2} \right\}$. Then for $K \geq t_\alpha$ and any $i \in \mathcal{N}$, we have*

$$
(a) = \mathbb{E} \left[\max_{s, a_i} \left| \hat{Q}_i(s_{N_i^{\kappa_c}}, a_i, w(K)) - \overline{Q}_i^{\hat{\xi}}(s, a_i) \right|^2 \right]
$$

$$
\leq 3 \left[c_1^*(1 - (1-\gamma)\underline{\lambda}\alpha)^{K-t_\alpha} + c_2^* \frac{\alpha t_\alpha}{(1-\gamma)\underline{\lambda}} + \left(\frac{(1-\lambda\gamma)\epsilon_{app}}{\pi_{\min}(1-\gamma)} \right)^2 + \left(\frac{\gamma^{\kappa_c - \kappa_r + 1}}{1-\gamma} \right) \right]^2.
$$

Here $c_1^* = (\max_{i \in \mathcal{N}} \|w_i^*\| + 1)^2$, $c_2^* = 114 \left(\frac{1+\gamma}{1-\lambda} \right)^2 (\max_{i \in \mathcal{N}} \|w_i^*\| + 1)^2$.

Bound (b). We first discuss the l_1-distance of policy ξ^θ and $\hat{\xi}$ with the two lemmas belows.

Lemma D.10. *For any agent i and policy ξ_i, let $\hat{\xi}_i(a_i, s_i) = (1 - \epsilon)\xi_i(a_i|s_i) + \epsilon \frac{1}{|\mathcal{A}_i|}$ for all s_i, a_i. Then*

$$
\left\| \xi_i(\cdot|s_i) - \hat{\xi}_i(\cdot|s_i) \right\|_1 \leq 2\epsilon, \ \forall s_i.
$$

Proof of Lemma D.10. We have by definition of the l_1-norm that

$$
\left\| \xi_i(\cdot|s_i) - \hat{\xi}_i(\cdot|s_i) \right\|_1 = \sum_{a_i} \left| \xi_i(a_i|s_i) - \hat{\xi}_i(a_i|s_i) \right|
$$

$$
= \epsilon \sum_{a_i} \left| \frac{1}{|\mathcal{A}_i|} - \xi_i(a_i|s_i) \right|
$$

$$
\leq \epsilon \sum_{a_i} \left(\frac{1}{|\mathcal{A}_i|} + \xi_i(a_i|s_i) \right)
$$

$$
= 2\epsilon.
$$

$\square$

Consider the difference of policy of a set of agents, we have the following result.

Lemma D.11. *For any set of agents $\mathcal{I} \subseteq \mathcal{N}$ and any policy $\xi_\mathcal{I}$. Define $\hat{\xi}_i$ in the same way as in Lemma D.10 for any $i \in \mathcal{I}$ and let $\hat{\xi}_\mathcal{I}$ be the product policy. Then*

$$
\left\| \xi_\mathcal{I}(\cdot|s_\mathcal{I}) - \hat{\xi}_\mathcal{I}(\cdot|s_\mathcal{I}) \right\|_1 \leq 2|\mathcal{I}| \epsilon, \ \forall s_\mathcal{I}.
$$

Proof of Lemma D.11. By Lemma F.15 and Lemma D.10, we immediately have

$$
\left\| \xi_\mathcal{I}(\cdot|s_\mathcal{I}) - \hat{\xi}_\mathcal{I}(\cdot|s_\mathcal{I}) \right\|_1 \leq \sum_{i \in \mathcal{I}} \left\| \xi_i(\cdot|s_i) - \hat{\xi}_i(\cdot|s_i) \right\|_1
$$

$$
\leq 2|\mathcal{I}| \epsilon.
$$

$\square$

Now we are ready to bound (b).

Lemma D.12. *For any policy ξ^θ, we have*

$$
(b) = \max_{s, a_i} \left| \overline{Q}_i^{\hat{\xi}}(s, a_i) - \overline{Q}_i^{\xi^\theta}(s, a_i) \right|^2 \leq \left(\frac{6n\epsilon}{(1-\gamma)^2} \right)^2.
$$

Here $\hat{\xi}$ is the ϵ-exploration policy of policy ξ^θ. Please refer to Line 2 of Algorithm 2 for the explicit definition of ϵ-exploration policy.

Proof of Lemma D.12. Notice that for any policy ξ, we have

$$Q_i^\xi(s, a_i)$$
$$= \sum_{a_{-k}} \xi_{-k}(a_{-k}|s_{-k})Q_i^\xi(s, a_i, a_{-k})$$
$$= \sum_{a_{-k}} \xi_{-k}(a_{-k}|s_{-k})\left(r_i(s_{N_i^{\kappa r}}, a_i, a_{U_{i_0}^{\kappa r}}) + \sum_{s'} \mathcal{P}(s'|s, a_i, a_{-k})V_i^\xi(s') \right).$$

Then we have for any two policies ξ and ξ' that

$$\left| \overline{Q}_i^\xi(s, a_i) - \overline{Q}_i^{\xi'}(s, a_i) \right|$$

$$= \left| \sum_{a_{-k}} \xi_{-k}(a_{-k}|s_{-k})\left(r_i(s_{N_i^{\kappa r}}, a_i, a_{U_{i_0}^{\kappa r}}) + \sum_{s'} \mathcal{P}(s'|s, a_i, a_{-k})V_i^\xi(s') \right) \right.$$

$$\left. - \sum_{a_{-k}} \xi'_{-k}(a_{-k}|s_{-k})\left(r_i(s_{N_i^{\kappa r}}, a_i, a_{U_{i_0}^{\kappa r}}) + \sum_{s'} \mathcal{P}(s'|s, a_i, a_{-k})V_i^{\xi'}(s') \right) \right|$$

$$\leq \left| \sum_{a_{-k}} (\xi_{-k}(a_{-k}|s_{-k}) - \xi'_{-k}(a_{-k}|s_{-k}))r_i(s_{N_i^{\kappa r}}, a_i, a_{U_{i_0}^{\kappa r}}) \right|$$

$$+ \left| \sum_{a_{-k}, s'} \mathcal{P}(s'|s, a_i, a_{-k})\left[\xi_{-k}(a_{-k}|s_{-k})V_i^\xi(s') - \xi'_{-k}(a_{-k}|s_{-k})V_i^{\xi'}(s') \right] \right|$$

$$\leq \left| \sum_{a_{-k}} (\xi_{-k}(a_{-k}|s_{-k}) - \xi'_{-k}(a_{-k}|s_{-k}))r_i(s_{N_i^{\kappa r}}, a_i, a_{U_{i_0}^{\kappa r}}) \right|$$

$$+ \left| \sum_{a_{-k}, s'} \mathcal{P}(s'|s, a_i, a_{-k})\left[\xi_{-k}(a_{-k}|s_{-k}) - \xi'_{-k}(a_{-k}|s_{-k}) \right] V_i^\xi(s') \right|$$

$$+ \left| \sum_{a_{-k}, s'} \mathcal{P}(s'|s, a_i, a_{-k})\xi'_{-k}(a_{-k}|s_{-k})\left[V_i^\xi(s') - V_i^{\xi'}(s') \right] \right|$$

$$\leq \sum_{a_{-k}} \left| \xi_{-k}(a_{-k}|s_{-k}) - \xi'_{-k}(a_{-k}|s_{-k}) \right|$$

$$+ \frac{1}{1-\gamma} \sum_{a_{-k}} \left| \xi_{-k}(a_{-k}|s_{-k}) - \xi'_{-k}(a_{-k}|s_{-k}) \right| \sum_{s'} \mathcal{P}(s'|s, a_i, a_{-k})$$

$$+ \sup_{s'} \left[V_i^\xi(s') - V_i^{\xi'}(s') \right].$$

To bound the last term, we apply a variant of performance difference lemma (Lemma F.13), and we have for any state s that

$$\left| V_i^\xi(s) - V_i^{\xi'}(s) \right| = \frac{1}{1-\gamma} \left| \sum_{s',a} d_s^{\xi'}(s')(\xi'(a|s') - \xi(a|s'))Q_i^\xi(s', a) \right|$$

$$\leq \frac{1}{(1-\gamma)^2} \left| \sum_{s',a} d_s^{\xi'}(s')(\xi'(a|s') - \xi(a|s')) \right|$$

$$\leq \frac{1}{(1-\gamma)^2} \sum_{s',a} d_s^{\xi'}(s') \left| \xi'(a|s') - \xi(a|s') \right|$$

$$\leq \frac{1}{(1-\gamma)^2} \sup_{s'} \left\| \xi'(\cdot|s') - \xi(\cdot|s') \right\|_1. \tag{33}$$

By Eq. (33), we can further bound $\left|\overline{Q}_i^\xi(s, a_i) - \overline{Q}_i^{\xi'}(s, a_i)\right|$ as

$$\left|\overline{Q}_i^\xi(s, a_i) - \overline{Q}_i^{\xi'}(s, a_i)\right|$$
$$\leq \left\|\xi_{-k}(\cdot|s_{-k}) - \xi'_{-k}(\cdot|s_{-k})\right\|_1 + \frac{1}{1-\gamma}\left\|\xi_{-k}(\cdot|s_{-k}) - \xi'_{-k}(\cdot|s_{-k})\right\|_1 + \frac{1}{(1-\gamma)^2}\sup_{s''}\left\|\xi'(\cdot|s'') - \xi(\cdot|s'')\right\|_1.$$

In particular, by choosing $\xi \leftarrow \hat{\xi}$, $\xi' \leftarrow \xi^\theta$ and applying Lemma D.11, we have

$$\left|\overline{Q}_i^{\hat{\xi}}(s, a_i) - \overline{Q}_i^{\xi^\theta}(s, a_i)\right| \leq 2(n-1)\epsilon + \frac{1}{1-\gamma}\cdot 2(n-1)\epsilon + \frac{1}{(1-\gamma)^2}\cdot 2n\epsilon$$
$$\leq \frac{6n\epsilon}{(1-\gamma)^2}.$$

Taking maximization on both sides w.r.t. s and a_i, we get

$$\max_{s,a_i}\left|\overline{Q}_i^{\hat{\xi}}(s, a_i) - \overline{Q}_i^{\xi^\theta}(s, a_i)\right|^2 \leq \left(\frac{6n\epsilon}{(1-\gamma)^2}\right)^2.$$

$\square$

Finally, apply Corollary D.9 and Lemma D.12 to Eq. (32), and we complete the proof.

E   PROOF OF THEOREM 5.7

For any $i \in \mathcal{N}$ and $m \geq 0$, we have

$$\Phi_i(\theta(m+1)) - \Phi_i(\theta(m)$$
$$= \left[\Phi_i(\theta(m+1)) - \Phi_i(\theta_{N_i^\kappa}(m+1), \theta_{-N_i^\kappa}(m))\right]$$
$$+ \left[\Phi_i(\theta_{N_i^\kappa}(m+1), \theta_{-N_i^\kappa}(m)) - \Phi_i(\theta(m))\right]. \tag{34}$$

Similar to the proof of Theorem 5.6, using Assumption 5.1 and the first term on the RHS of Eq. (34) can be lower-bounded as

$$\Phi_i(\theta(m+1)) - \Phi_i(\theta_{N_i^\kappa}(m+1), \theta_{-N_i^\kappa}(m)) \geq -\frac{\sqrt{2}\nu(\kappa)\beta}{(1-\gamma)^2}.$$

Now consider the second term. Denote $e_j(m) = \nabla_{\theta_j}J_j(\theta(m)) - \Delta_j^T(m)$. Using the smoothness property (Lemma B.2) of the local potential functions and we have

$$\Phi_i(\theta_{N_i^\kappa}(m+1), \theta_{-N_i^\kappa}(m)) - \Phi_i(\theta(m))$$
$$\geq \langle\nabla_{\theta_{N_i^\kappa}}\Phi_i(\theta(m)), \theta_{N_i^\kappa}(m+1) - \theta_{N_i^\kappa}(m)\rangle - \frac{L(\kappa)}{2}\left\|\theta_{N_i^\kappa}(m+1) - \theta_{N_i^\kappa}(m)\right\|^2$$
$$= \sum_{j\in N_i^\kappa}\left[\beta\langle\nabla_{\theta_j}J_j(\theta(m)), \Delta_j^T(m)\rangle - \frac{L(\kappa)\beta^2}{2}\left\|\Delta_j^T(m)\right\|^2\right]$$
$$= \sum_{j\in N_i^\kappa}\left[\beta\langle\nabla_{\theta_j}J_j(\theta(m)), \nabla_{\theta_j}J_j(\theta(m)) + e_j(m)\rangle - \frac{L(\kappa)\beta^2}{2}\left\|\nabla_{\theta_j}J_j(\theta(m) + e_j(m)\right\|^2\right]$$
$$\geq \sum_{j\in N_i^\kappa}\left[(\beta - L(\kappa)\beta^2)\|\nabla_{\theta_j}J_j(\theta(m))\|^2 + \beta\langle\nabla_{\theta_j}J_j(\theta(m)), e_j(m)\rangle - L(\kappa)\beta^2\left\|e_j(m)\right\|^2\right],$$

where the last line follows from $(a + b)^2 \leq 2(a^2 + b^2)$ for any $a, b \in \mathbb{R}$. Using the previous two inequalities in Eq. (34) and we have

$$
\mathbb{E}[\Phi_i(\theta(m + 1)) \mid \mathcal{F}_m] - \mathbb{E}[\Phi_i(\theta(m))]
$$

$$
\geq -\frac{\sqrt{2}\nu(\kappa)\beta}{(1 - \gamma)^2} + (\beta - L(\kappa)\beta^2) \sum_{j \in N_i^\kappa} \|\nabla_{\theta_j} J_j(\theta(m))\|^2
$$

$$
+ \beta \sum_{j \in N_i^\kappa} \langle \nabla_{\theta_j} J_j(\theta(m)), \mathbb{E}[e_j(m) \mid \mathcal{F}_m] \rangle - L(\kappa)\beta^2 \sum_{j \in N_i^\kappa} \mathbb{E}[\|e_j(m)\|^2 \mid \mathcal{F}_m]
$$

$$
\geq -\frac{\sqrt{2}\nu(\kappa)\beta}{(1 - \gamma)^2} + (\beta - L(\kappa)\beta^2) \sum_{j \in N_i^\kappa} \|\nabla_{\theta_j} J_j(\theta(m))\|^2
$$

$$
- \frac{\beta}{2} \sum_{j \in N_i^\kappa} (\|\nabla_{\theta_j} J_j(\theta(m))\|^2 + \|\mathbb{E}[e_j(m) \mid \mathcal{F}_m]\|^2) - L(\kappa)\beta^2 \sum_{j \in N_i^\kappa} \mathbb{E}[\|e_j(m)\|^2 \mid \mathcal{F}_m]
$$

$$
= -\frac{\sqrt{2}\nu(\kappa)\beta}{(1 - \gamma)^2} + \left(\frac{\beta}{2} - L(\kappa)\beta^2\right) \sum_{j \in N_i^\kappa} \|\nabla_{\theta_j} J_j(\theta(m))\|^2
$$

$$
- \frac{\beta}{2} \sum_{j \in N_i^\kappa} \|\mathbb{E}[e_j(m) \mid \mathcal{F}_m]\|^2 - L(\kappa)\beta^2 \sum_{j \in N_i^\kappa} \mathbb{E}[\|e_j(m)\|^2 \mid \mathcal{F}_m],
$$

where $\mathcal{F}_m$ represents the history up to the beginning of the m-th outer-loop iteration. Taking the total expectation on both sides of the previous inequality and we have

$$
\mathbb{E}[\Phi_i(\theta(m + 1))] - \mathbb{E}[\Phi_i(\theta(m))]
$$

$$
\geq -\frac{\sqrt{2}\nu(\kappa)\beta}{(1 - \gamma)^2} + \left(\frac{\beta}{2} - L(\kappa)\beta^2\right) \sum_{j \in N_i^\kappa} \mathbb{E}[\|\nabla_{\theta_j} J_j(\theta(m))\|^2]
$$

$$
- \frac{\beta}{2} \sum_{j \in N_i^\kappa} \mathbb{E}[\|\mathbb{E}[e_j(m) \mid \mathcal{F}_m]\|^2] - L(\kappa)\beta^2 \sum_{j \in N_i^\kappa} \mathbb{E}[\|e_j(m)\|^2],
$$

which implies

$$
\frac{1}{M}\mathbb{E}[\Phi_i(\theta(M))] - \mathbb{E}[\Phi_i(\theta(0))]
$$

$$
\geq -\frac{\sqrt{2}\nu(\kappa)\beta}{(1 - \gamma)^2} + \frac{(\beta/2 - L(\kappa)\beta^2)}{M} \sum_{m=0}^{M-1} \sum_{j \in N_i^\kappa} \mathbb{E}[\|\nabla_{\theta_j} J_j(\theta(m))\|^2]
$$

$$
- \frac{\beta}{2M} \sum_{m=0}^{M-1} \sum_{j \in N_i^\kappa} \mathbb{E}[\|\mathbb{E}[e_j(m) \mid \mathcal{F}_m]\|^2] - \frac{L(\kappa)\beta^2}{M} \sum_{m=0}^{M-1} \sum_{j \in N_i^\kappa} \mathbb{E}[\|e_j(m)\|^2]. \tag{35}
$$

Since $\Phi_{\min} \leq \Phi_i(\theta) \leq \Phi_{\max}$ for all θ and $\beta \leq \frac{1}{4L(\kappa)}$, after rearranging terms and we have

$$
\frac{1}{M} \sum_{m=0}^{M-1} \sum_{j \in N_i^\kappa} \mathbb{E}[\|\nabla_{\theta_j} J_j(\theta(m))\|^2]
$$

$$
\leq \frac{4(\Phi_{\max} - \Phi_{\min})}{\beta M} + \frac{4\sqrt{2}\nu(\kappa)}{(1 - \gamma)^2} + \frac{2}{M} \sum_{m=0}^{M-1} \sum_{j \in N_i^\kappa} \underbrace{\mathbb{E}[\|\mathbb{E}[e_j(m) \mid \mathcal{F}_m]\|^2]}_{\mathcal{T}_1}
$$

$$
+ \frac{4L(\kappa)\beta}{M} \sum_{m=0}^{M-1} \sum_{j \in N_i^\kappa} \underbrace{\mathbb{E}[\|e_j(m)\|^2]}_{\mathcal{T}_2}.
$$

We next bound the terms $\mathcal{T}_1$ and $\mathcal{T}_2$ from above. To begin with, we decompose $e_i(m)$ in the following way:

$$
e_i(m) = \nabla_{\theta_i} J_i(\theta(m)) - \Delta_i^T(m)
$$

$$
= \sum_{k=0}^{\infty} \gamma^k \mathbb{E}\left[\nabla_{\theta_i} \log \xi_i^{\theta_i(m)}(a_i(k)|s_i(k))\overline{Q}_i^{\theta(m)}(s(k), a_i(k))\right]
$$

$$
- \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=0}^{H-1}\gamma^k \nabla_{\theta_i}\log\xi_i^{\theta_i(m)}(a_i^t(k)|s_i^t(k))\phi_i(s_{\mathcal{N}_i^{\kappa_c}}^t(k), a_i^t(k))^\top w_i^m
$$

$$
= \sum_{k=H}^{\infty} \gamma^k \mathbb{E}\left[\nabla_{\theta_i} \log \xi_i^{\theta_i(m)}(a_i(k)|s_i(k))\overline{Q}_i^{\theta(m)}(s(k), a_i(k))\right] \tag{$\mathcal{X}_1$}
$$

$$
+ \sum_{k=0}^{H-1} \gamma^k \mathbb{E}\left[\nabla_{\theta_i} \log \xi_i^{\theta_i(m)}(a_i(k)|s_i(k))\overline{Q}_i^{\theta(m)}(s(k), a_i(k))\right] \tag{$\mathcal{X}_2$}
$$

$$
- \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=0}^{H-1}\gamma^k \nabla_{\theta_i}\log\xi_i^{\theta_i(m)}(a_i^t(k)|s_i^t(k))\overline{Q}_i^{\theta(m)}(s^t(k), a_i^t(k)) \tag{$\mathcal{X}_3$}
$$

$$
+ \frac{1}{T}\sum_{t=0}^{T-1}\sum_{k=0}^{H-1}\gamma^k \nabla_{\theta_i}\log\xi_i^{\theta_i(m)}(a_i^t(k)|s_i^t(k))
$$

$$
\times (\overline{Q}_i^{\theta(m)}(s^t(k), a_i^t(k)) - \phi_i(s_{\mathcal{N}_i^{\kappa_c}}^t(k), a_i^t(k))^\top w_i^m). \tag{$\mathcal{X}_4$}
$$

It follows that

$$
\mathbb{E}[e_i(m) \mid \mathcal{F}_m] = \mathcal{X}_1 + \mathcal{X}_2 + \mathbb{E}[\mathcal{X}_3 \mid \mathcal{F}_m] + \mathbb{E}[\mathcal{X}_4 \mid \mathcal{F}_m]
$$
$$
= \mathcal{X}_1 + \mathbb{E}[\mathcal{X}_4 \mid \mathcal{F}_m].
$$

Therefore, using Lemma F.9, the definition of ϵ_{critic}, and the fact that the averaged Q-function is bounded (in ℓ_∞-norm) by $\frac{1}{1-\gamma}$, and we have

$$
\|\mathbb{E}[e_i(m) \mid \mathcal{F}_m]\| \leq \|\mathcal{X}_1\| + \|\mathbb{E}[\mathcal{X}_4 \mid \mathcal{F}_m]\|
$$
$$
\leq \frac{\sqrt{2}\gamma^H}{(1-\gamma)^2} + \frac{\sqrt{2}\epsilon_{\text{critic}}}{1-\gamma},
$$

which implies that

$$
\mathcal{T}_1 = \mathbb{E}[\|\mathbb{E}[e_j(m) \mid \mathcal{F}_m]\|^2] \leq \frac{4\gamma^{2H}}{(1-\gamma)^4} + \frac{4\epsilon_{\text{critic}}^2}{(1-\gamma)^2}.
$$

As for the term $\mathcal{T}_2$, similarly we have

$$
\mathcal{T}_2 = \mathbb{E}[\|e_j(m)\|^2]
$$
$$
\leq \mathbb{E}[\|\mathcal{X}_1 + \mathcal{X}_2 + \mathcal{X}_3 + \mathcal{X}_4\|^2]
$$
$$
\leq 3\mathbb{E}[\|\mathcal{X}_1\|^2 + \|\mathcal{X}_2 + \mathcal{X}_3\|^2 + \|\mathcal{X}_4\|^2]
$$
$$
\leq \frac{6\gamma^{2H}}{(1-\gamma)^4} + \frac{24}{(1-\gamma)^2 T} + \frac{6\epsilon_{\text{critic}}^2}{(1-\gamma)^2}.
$$

Substituting the upper bounds we obtained for the terms $\mathcal{T}_1$ and $\mathcal{T}_2$ in Eq. (35) and we obtain

$$
\frac{1}{M}\sum_{m=0}^{M-1}\sum_{j\in N_i^\kappa}\mathbb{E}[\|\nabla_{\theta_j}J_j(\theta(m))\|^2]
$$

$$
\leq \frac{4(\Phi_{\max} - \Phi_{\min})}{\beta M} + \frac{4\sqrt{2}\nu(\kappa)}{(1-\gamma)^2} + \frac{8n(\kappa)\gamma^{2H}}{(1-\gamma)^4} + \frac{8n(\kappa)\epsilon_{\text{critic}}^2}{(1-\gamma)^2}
$$

$$
+ 24L(\kappa)n(\kappa)\beta\left(\frac{\gamma^{2H}}{(1-\gamma)^4} + \frac{\epsilon_{\text{critic}}^2}{(1-\gamma)^2} + \frac{4}{(1-\gamma)^2 T}\right).
$$

Recall that Lemma B.4 implies

$$\sum_{j \in N_i^\kappa} \left\| \nabla_{\theta_j} J_j(\theta(m)) \right\|^2 \geq \left\| \nabla_{\theta_i} J_i(\theta(m)) \right\|^2 \geq \frac{c^2}{\max_{j \in \mathcal{N}} |\mathcal{A}_j| D^2} \text{NE-Gap}_i(\theta(m))^2.$$

Therefore, by choosing $\kappa = \kappa_G$ and $\beta = \frac{1}{8L(\kappa)}$, we have

$$\frac{1}{M} \sum_{m=0}^{M-1} \mathbb{E}[\text{NE-Gap}_i(\theta(m))^2]$$

$$\leq \frac{\max_{j \in \mathcal{N}} |\mathcal{A}_j| D^2}{c^2} \left[\frac{4(\Phi_{\max} - \Phi_{\min})}{\beta M} + \frac{4\sqrt{2}\nu(\kappa)}{(1-\gamma)^2} + \frac{8n(\kappa)\gamma^{2H}}{(1-\gamma)^4} + \frac{8n(\kappa)\epsilon_{\text{critic}}^2}{(1-\gamma)^2} \right.$$

$$\left. + 24L(\kappa)n(\kappa)\beta \left(\frac{\gamma^{2H}}{(1-\gamma)^4} + \frac{\epsilon_{\text{critic}}^2}{(1-\gamma)^2} + \frac{4}{(1-\gamma)^2 T} \right) \right]$$

$$\leq \frac{12n(\kappa_G)\max_{j \in \mathcal{N}} |\mathcal{A}_j| D^2}{c^2(1-\gamma)^2} \left[\frac{16(\Phi_{\max} - \Phi_{\min})}{M(1-\gamma)} + \frac{\nu(\kappa_G)}{n(\kappa_G)} + \frac{\gamma^{2H}}{(1-\gamma)^2} + \epsilon_{\text{critic}}^2 + \frac{1}{T} \right].$$

Finally, using Jensen's inequality and we obtain

$$\mathbb{E}[\text{Avg-Nash-Regret}_i(M)]$$

$$= \frac{1}{M} \sum_{m=0}^{M-1} \mathbb{E}[\text{NE-Gap}_i(\theta(m))]$$

$$\leq \frac{4n(\kappa_G)^{1/2} \max_{j \in \mathcal{N}} \sqrt{|\mathcal{A}_j|}}{c(1-\gamma)} \left[\frac{4(\Phi_{\max} - \Phi_{\min})^{1/2}}{M^{1/2}(1-\gamma)^{1/2}} + \frac{\nu(\kappa_G)^{1/2}}{n(\kappa_G)^{1/2}} + \frac{\gamma^H}{1-\gamma} + \epsilon_{\text{critic}} + \frac{1}{T^{1/2}} \right].$$

Since the RHS of the bound is not a function of i, we have the desired result.

F   SUPPORTING RESULTS

F.1   NMPG IS A STRICT GENERALIZATION OF MPG

We present a simple example to show that NMPG is a strict generalization of the standard MPG.

Example: Consider a networked multi-agent Markov game, denoted by MG, with 4 agents $\mathcal{N} = \{1, 2, 3, 4\}$. Each agent has local state space $\mathcal{S}_i = \{s_b, s_g\}$ and local action space $\mathcal{A}_i = \{a_b, a_g\}$ for all agent $i \in \mathcal{N}$. The underlying undirected graph connects edges between every "neighboring" agents, i.e., the set of edges is $\mathcal{E} = \{(1, 2), (2, 3), (3, 4)\}$.

- Initial state: All agents starts at state s_b. That is, $s_i(0) = s_b$ for all $i \in \mathbb{N}$.
- Transition: MG has deterministic transitions. For agent 1, the next state only depends on its own action: For all $t \geq 0$, $s_1(t + 1) = s_g$ if $a_1(t) = a_g$ and $s_1(t + 1) = s_b$ if $a_1(t) = a_b$. For agent $i \in \{2, 3, 4\}$, the next state only depends on state of "previous" agent, i.e., $s_i(t + 1) = s_{i-1}(t)$, for all $t \geq 0$.
- Reward: Each agent's reward is only dependent on its own state and action. The local reward of agent 1,2,3 is always 0. That is, $r_i(s_i, a_i) = 0$ for all $i \in \{1, 2, 3\}$, $s_i \in \mathcal{S}_i$, $a_i \in \mathcal{A}_i$. For agent 4, it receives reward 1 if it is at state s_g and takes action a_g while it receives reward 0 in all other cases. In other words, $r_4(s_g, a_g) = 1$, $r_4(s_g, a_b) = 0$, $r_4(s_b, a_g) = 0$, $r_4(s_b, a_b) = 0$.

Obviously, the expected return of agent 1,2 and 3 is always 0. As for agent 4, its state is solely dependent on the state of agent 1 (three timesteps before), which is completely determined by agent 1's policy. As a result, agent 4's expected return, or objective function, depends on local policies of *both* agent 1 and 4. This observation is the key to showing that MG is not an MPG while being a 1-NMPG, which is summarized as the theorem below.

Theorem F.1. *MG is a 1-NMPG but not an MPG.*

Proof of Theorem F.1. We first compute the objective function $J_i(\xi)$ for all global policy ξ and agent $i \in \mathcal{N}$. Noticing that

agents 1,2,3 always receive reward 0, we have $J_i(\xi) = 0$ for all $i \in \{1, 2, 3\}$. As for agent 4, we have

$$
\begin{aligned}
J_4(\xi) &= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_\xi \left[r_4(s_4(t), a_4(t)) \right] \\
&= \sum_{t=0}^{\infty} \gamma^t \Pr[r_4(s_4(t), a_4(t) = 1] \\
&= \sum_{t=0}^{\infty} \gamma^t \Pr[s_4(t) = s_g, a_4(t) = a_g] \\
&\stackrel{(i)}{=} \sum_{t=4}^{\infty} \gamma^t \Pr[a_1(t-4) = a_g, a_4(t) = a_g] \\
&\triangleq f(\xi_1, \xi_4).
\end{aligned}
\tag{36}
$$

Here (i) uses the fact that $s_4(t) = s_b$ for $t = 0, 1, 2, 3$, and $s_4(t) = s_1(t-3) = s_g$ if and only if $a_1(t-4) = a_g$ for all $t \geq 4$.

Next we show that MG is a 1-NMPG. In fact, we can choose local potential functions $\Phi_1(\xi) = \Phi_2(\xi) = 0$, $\Phi_3(\xi) = \Phi_4(\xi) = f(\xi_1, \xi_4)$ for all global policy ξ, and Eq. (2) can be satisfied.

Finally, we prove that MG is not an MPG. Assume that there is a potential $\Phi(\xi)$ such that

$$
J_i(\xi_i', \xi_{-i}) - J_i(\xi_i, \xi_{-i}) = \Phi(\xi_i', \xi_{-i}) - \Phi(\xi_i, \xi_{-i})
\tag{37}
$$

for all $i \in \mathbb{N}$, $\xi_i, \xi_i' \in \Xi_i$, $\xi_{-i} \in \Xi_{-i}$. If we choose $i \in \{1, 2, 3\}$ in Eq. (37), we can see that $\Phi(\xi_i', \xi_{-i}) - \Phi(\xi_i, \xi_{-i}) = 0$ for all $\xi_i, \xi_i' \in \Xi_i$, $\xi_{-i} \in \Xi_{-i}$. Thus Φ is independent of ξ_i for all $i \in \{1, 2, 3\}$. As a result, the potential function can be represented as $\Phi(\xi_4)$. Then we let $i = 4$ in Eq. (37), and we have for all $\xi_1', \xi_1 \in \Xi_1$, $\xi_4' \in \Xi_4$ that

$$
\Phi(\xi_4') - \Phi(\xi_4) = J_4(\xi_4', \xi_{-4}) - J_4(\xi_4, \xi_{-4}) = f(\xi_1, \xi_4') - f(\xi_1, \xi_4).
\tag{38}
$$

To derive a contradiction, we take some special values of ξ_1, ξ_1', ξ_4. For any $i \in \mathbb{N}$, let ξ_i^g be the policy that agent i always takes a_g and ξ_i^b be the policy that agent i always takes a_b. Then we can derive from Eq. (36) that $f(\xi_1^g, \xi_4^g) = \frac{\gamma^4}{1-\gamma}$, $f(\xi_1^g, \xi_4^b) = f(\xi_1^b, \xi_4^g) = f(\xi_1^b, \xi_4^b) = 0$. In Eq. (38), let $\xi_4' = \xi_4^g$, $\xi_4 = \xi_4^b$, $\xi_1 \in \{\xi_1^b, \xi_1^g\}$, and we have

$$
\begin{cases}
\Phi(\xi_4^g) - \Phi(\xi_4^b) = f(\xi_1^b, \xi_4^g) - f(\xi_1^b, \xi_4^b) = 0 \\
\Phi(\xi_4^g) - \Phi(\xi_4^b) = f(\xi_1^g, \xi_4^g) - f(\xi_1^g, \xi_4^b) = \frac{\gamma^4}{1-\gamma},
\end{cases}
$$

which leads to contradiction. As a result, MG is not an MPG. $\qquad\square$

κ_G-NMPG in general cannot be reduced to a standard MPG. Please note that in a κ_G-NMPG, for any agent i, any agent $j \in N_i^{\kappa_G}$ do share a potential function Φ_i, but this potential is associated with the agent i. In the reviewer's example, the potential functions $\Phi_{i_1}, \Phi_{i_2}, \cdots, \Phi_{i_{\kappa_G+1}}$ are associated with agent i_1, while the potential functions $\Phi_{i_L}, \Phi_{i_{L-1}}, \cdots, \Phi_{i_{L-\kappa_G}}$ are associated with agent i_L. Therefore, unless $L \leq 2\kappa_G - 1$, i_1 and i_L are not guaranteed to share a same potential function.

In addition, we can give a simple example that is a for your reference. The transition is deterministic. The initial states of the 4 agents are all s_b. Then $J_1(\xi) = J_2(\xi) = J_3(\xi) = 0$.

$$
J_4(\xi) = \sum_{t=4}^{\infty} \gamma^t \Pr[a_1(t-4) = a_g] \Pr[a_4(t) = a_g] \triangleq f(\xi_1, \xi_4).
$$

The above example is a 1-NMPG, with $\Phi_1(\xi) = \Phi_2(\xi) = 0$, $\Phi_3(\xi) = \Phi_4(\xi) = f(\xi_1, \xi_4)$, but it is not an MPG.

F.2 POLICY GRADIENT THEOREM VARIANT

We prove Eq. (4), which is a variant of the policy gradient theorem [Sutton et al., 1999].

Lemma F.2 (Policy gradient theorem variant).

$$\nabla_{\theta_i} J_i(\theta) = \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\xi^{\theta}} \left[\nabla_{\theta_i} \log \xi_i^{\theta_i}(a_i(t)|s_i(t)) \overline{Q}_i^{\theta}(s(t), a_i(t)) \right].$$

Proof of Lemma F.2. By Lemma F.10, we have

$$\nabla_{\theta_i} J_i(\theta)$$
$$= \frac{1}{1-\gamma} \sum_{s,a_i} d^{\theta}(s) \nabla_{\theta_i} \xi_i^{\theta_i}(a_i|s_i) \overline{Q}_i^{\theta}(s, a_i)$$
$$= \sum_{s,a_i} \sum_{t=0}^{\infty} \gamma^t \mathrm{Pr}^{\xi^{\theta}}[s(t) = s|s(0) \sim \mu(\cdot)] \nabla_{\theta_i} \xi_i^{\theta_i}(a_i|s_i) \overline{Q}_i^{\theta}(s, a_i)$$
$$= \sum_{t=0}^{\infty} \sum_{s,a_i} \gamma^t \mathrm{Pr}^{\xi^{\theta}}[s(t) = s|s(0) \sim \mu(\cdot)] \xi_i^{\theta_i}(a_i|s_i) \nabla_{\theta_i} \log \xi_i^{\theta_i}(a_i|s_i) \overline{Q}_i^{\theta}(s, a_i)$$
$$= \sum_{t=0}^{\infty} \gamma^t \mathbb{E}_{\xi^{\theta}} \left[\nabla_{\theta_i} \log \xi_i^{\theta_i}(a_i(t)|s_i(t)) \overline{Q}_i^{\theta}(s(t), a_i(t)) \right].$$

Here $\mathrm{Pr}^{\xi^{\theta}}[s(t) = s|s(0) \sim \mu(\cdot)]$ represents the probability that $s(t) = s$ given that the policy is ξ^{θ} and initial state $s(0)$ is sampled from distribution μ. □

F.3 PROOF OF LEMMA 4.1

When $\kappa_c \le \kappa_r - 1$, the conclusion is obvious by the fact that both the truncated averaged Q-function and the averaged Q-function are in the range of $[0, \frac{1}{1-\gamma}]$. Below we only consider the case that $\kappa_c \ge \kappa_r$.

Consider any agent i, global policy parameter θ, and any truncated averaged Q-function $\overline{Q}_i^{\theta,\kappa_c} \in \mathcal{Q}_i^{\theta,\kappa_c}$. Then there exists $u_i \in \Delta(\mathcal{S}_{-N_i^{\kappa_c}})$, such that $\overline{Q}_i^{\theta,\kappa_c}(s_{N_i^{\kappa_c}}, a_i) = \sum_{s_{-N_i^{\kappa_c}}} u_i(s_{-N_i^{\kappa_c}}) \overline{Q}_i^{\theta}(s_{N_i^{\kappa_c}}, s_{-N_i^{\kappa_c}}, a_i)$ for any $(s_{N_i^{\kappa_c}}, a_i) \in \mathcal{S}_{N_i^{\kappa_c}} \times \mathcal{A}_i$. We have for any $(s, a_i) \in \mathcal{S} \times \mathcal{A}_i$ that

$$\left| \overline{Q}_i^{\theta,\kappa_c}(s_{N_i^{\kappa_c}}, a_i) - \overline{Q}_i^{\theta}(s, a_i) \right|$$
$$= \left| \sum_{s'_{-N_i^{\kappa_c}}} u_i(s'_{-N_i^{\kappa_c}}) \left(\overline{Q}_i^{\theta}(s_{N_i^{\kappa_c}}, s'_{-N_i^{\kappa_c}}, a_i) - \overline{Q}_i^{\theta}(s_{N_i^{\kappa_c}}, s_{-N_i^{\kappa_c}}, a_i) \right) \right|$$
$$\le \sum_{s'_{-N_i^{\kappa_c}}} u_i(s'_{-N_i^{\kappa_c}}) \left| \overline{Q}_i^{\theta}(s_{N_i^{\kappa_c}}, s'_{-N_i^{\kappa_c}}, a_i) - \overline{Q}_i^{\theta}(s_{N_i^{\kappa_c}}, s_{-N_i^{\kappa_c}}, a_i) \right|$$
$$\le \sum_{s'_{-N_i^{\kappa_c}}} \sum_{a_{-i}} u_i(s'_{-N_i^{\kappa_c}}) \xi_{-i}^{\theta_{-i}}(a_{-i}|s_{-i}) \left| Q_i^{\theta}(s_{N_i^{\kappa_c}}, s'_{-N_i^{\kappa_c}}, a_i, a_{-i}) - Q_i^{\theta}(s_{N_i^{\kappa_c}}, s_{-N_i^{\kappa_c}}, a_i, a_{-i}) \right|. \qquad (39)$$

We now try to give a perturbation bound for agent i's Q-function w.r.t. $s_{N_i^{\kappa_c}}$, the states of agents in κ-hop neighborhood.

Notice that for any $s \in \mathcal{S}, a \in \mathcal{A}$,

$$Q_i^\theta(s,a)$$

$$= \sum_{t=0}^\infty \gamma^t \sum_{s(t),a(t)} \mathrm{Pr}^{\xi^\theta}(s(t)|s(0)=s,a(0)=a)\xi^\theta(a(t)|s(t))r_i(s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t))$$

$$= \sum_{t=0}^\infty \gamma^t \sum_{s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t)} \mathrm{Pr}^{\xi^\theta}(s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t)|s(0)=s,a(0)=a)r_i(s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t))$$

$$= \sum_{t=0}^\infty \gamma^t \sum_{s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t)} \mathrm{Pr}^{\xi^\theta}(s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t)|s(0)=s,a(0)=a)r_i(s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t))$$

for simplicity of notation.

For any fixed $i \in \mathcal{N}$, $s_{N_i^{\kappa_c}} \in \mathcal{S}_{N_i^{\kappa_c}}$, $s_{-N_i^{\kappa_c}}, s'_{-N_i^{\kappa_c}} \in \mathcal{S}_{-N_i^{\kappa_c}}$, $a \in \mathcal{A}$ and policy ξ^θ, let

$$\pi_t^{\kappa_r}(s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t)) = \mathrm{Pr}^{\xi^\theta}(s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t)|s(0)=(s_{N_i^{\kappa_c}}, s_{-N_i^{\kappa_c}}), a(0)=a)$$

$$\tilde{\pi}_t^{\kappa_r}(s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t)) = \mathrm{Pr}^{\xi^\theta}(s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t)|s(0)=(s_{N_i^{\kappa_c}}, s'_{-N_i^{\kappa_c}}), a(0)=a).$$

Due to the local dependence of network and localized policy structure, $\pi_t^{\kappa_r}$ is only dependent on the initial states and actions of agents in $\mathcal{N}_i^{\kappa_r+t}$, which is equal to $\tilde{\pi}_t^{\kappa_r}$ when $t \leq \kappa_c - \kappa_r$. Therefore we have

$$\left| Q_i^\theta(s_{N_i^{\kappa_c}}, s'_{-N_i^{\kappa_c}}, a_i, a_{-i}) - Q_i^\theta(s_{N_i^{\kappa_c}}, s_{-N_i^{\kappa_c}}, a_i, a_{-i}) \right|$$

$$= \left| \sum_{t=0}^\infty \gamma^t \sum_{s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t)} \left(\xi_t^{\kappa_r}(s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t)) - \tilde{\pi}_t^{\kappa_r}(s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t)) \right) \right.$$

$$\left. \times r_i(s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t)) \right|$$

$$\leq \sum_{t=0}^\infty \gamma^t \sum_{s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t)} \left| \xi_t^{\kappa_r}(s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t)) - \tilde{\pi}_t^{\kappa_r}(s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t)) \right|$$

$$\times r_i(s_{\mathcal{N}_i^{\kappa_r}}(t), a_{\mathcal{N}_i^{\kappa_r}}(t))$$

$$\overset{(i)}{\leq} \sum_{t=0}^\infty \gamma^t \| \pi_t^{\kappa_r} - \tilde{\pi}_t^{\kappa_r} \|_1$$

$$= \sum_{t=\kappa_c-\kappa_r+1}^\infty \gamma^t \| \pi_t^{\kappa_r} - \tilde{\pi}_t^{\kappa_r} \|_1$$

$$\leq 2 \sum_{t=\kappa_c-\kappa_r+1}^\infty \gamma^t$$

$$\leq \frac{2}{1-\gamma} \gamma^{\kappa_c-\kappa_r+1}.$$

Here (i) is by $r_i(s_{N_i^{\kappa_r}}, a_{N_i^{\kappa_r}}) \leq 1$ for any $i \in \mathcal{N}$ and any $s_{N_i^{\kappa_r}}, a_{N_i^{\kappa_r}}$. Plug into Eq. (39), and we have

$$\left| \overline{Q}_i^{\theta,\kappa_c}(s_{N_i^{\kappa_c}}, a_i) - \overline{Q}_i^\theta(s, a_i) \right|$$

$$\leq \sum_{s'_{-N_i^{\kappa_c}}} \sum_{a_{-i}} u_i(s'_{-N_i^{\kappa_c}})\xi_{-i}^{\theta-i}(a_{-i}|s_{-i})\frac{2}{1-\gamma}\gamma^{\kappa_c-\kappa_r+1}$$

$$= \frac{2}{1-\gamma}\gamma^{\kappa_c-\kappa_r+1}.$$

Take $\max$ over s, a_i, sup over $\overline{Q}_i^{\theta,\kappa_c}$, and we complete the proof.

F.4 AVERAGED NASH REGRET

The relationship between the averaged Nash regret defined in this work and the Nash regret in Ding et al. [2022] is shown in the following lemma. Recall that we denote $n = |\mathcal{N}|$.

Lemma F.3. *Given any positive integer M, the following inequality holds for any sequence of policies $\{\xi(0), \xi(1), \cdots, \xi(M-1)\}$:*

$$\frac{1}{n} \textit{Nash-Regret}(M) \leq \textit{Avg-Nash-Regret}(M) \leq \textit{Nash-Regret}(M).$$

Proof of Lemma F.3. By definition of the averaged Nash Regret (cf. Definition 3.4) and the Nash Regret (cf. Eq. (3)), we have

$$
\begin{aligned}
\frac{1}{n}\text{Nash-Regret}(M) &= \frac{1}{n}\frac{1}{M}\sum_{m=0}^{M-1}\max_{i\in\mathcal{N}}\text{NE-Gap}_i(\xi(m)) \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\frac{1}{M}\sum_{m=0}^{M-1}\text{NE-Gap}_i(\xi(m)) \\
&\leq \max_{i\in\mathcal{N}}\frac{1}{M}\sum_{m=0}^{M-1}\text{NE-Gap}_i(\xi(m)) \\
&= \text{Avg-Nash-Regret}(M) \\
&\leq \frac{1}{M}\sum_{m=0}^{M-1}\max_{i\in\mathcal{N}}\text{NE-Gap}_i(\xi(m)) \qquad \text{(Jensen's inequality)} \\
&= \text{Nash-Regret}(M).
\end{aligned}
$$

$\square$

F.5 DECAY OF LOCAL POTENTIAL FUNCTIONS

We derive $\nu(\kappa) = O(\sum_{j \in -N_i^\kappa} |\mathcal{A}_j|)$ given a mild assumption, which guarantees existence of stage potential.

Assumption F.4. *For any agent $i \in \mathcal{N}$, there exists stage potential function $\varphi_i : \mathcal{S} \times \mathcal{A} \to [0, \overline{\varphi}]$, such that*

$$\Phi_i(\theta) = \mathbb{E}_\theta\left[\sum_{t=0}^{\infty}\gamma^t\varphi_i(s(t),a(t))\right]. \tag{40}$$

This assumption is common in recent MPG literature [Zhang et al., 2022a].

Lemma F.5. *With Assumption F.4 satisfied, we have*

$$\left|\Phi_i(\theta_{N_i^\kappa},\theta'_{-N_i^\kappa}) - \Phi_i(\theta_{N_i^\kappa},\theta_{-N_i^\kappa})\right| \leq \frac{\sqrt{2}\overline{\varphi}}{(1-\gamma)^2}\sum_{j\in-N_i^\kappa}|\mathcal{A}_j|\max_{j\in-N_i^\kappa}\left\|\theta'_j-\theta_j\right\|. \tag{41}$$

Proof of Lemma F.5. Similar to Q-function and averaged Q-function, we can define "Q-potential" function and averaged "Q-potential" function as

$$
\begin{aligned}
Q\Phi_i^\theta(s,a) &= \mathbb{E}_\theta\left[\sum_{t=0}^{\infty}\varphi_i(s(t),a(t))|s(0)=s,a(0)=a\right] \\
\overline{Q\Phi}_i^\theta(s,a_{-N_i^\kappa}) &= \sum_{a_{N_i^\kappa}}\xi_{N_i^\kappa}^{\theta_{N_i^\kappa}}(a_{N_i^\kappa}|s_{N_i^\kappa})Q\Phi_i(s,a_{N_i^\kappa},a_{-N_i^\kappa}).
\end{aligned}
$$

In Lemma F.14, replace the objective function J_i with potential function Φ_i, treat agents in $-N_i^\kappa$ as one agent, and we have

$$
\left| \Phi_i(\theta_{N_i^\kappa}, \theta'_{-N_i^\kappa}) - \Phi_i(\theta_{N_i^\kappa}, \theta_{-N_i^\kappa}) \right|
$$

$$
= \frac{1}{1-\gamma} \sum_{s, a_{-N_i^\kappa}} d^{\theta'}(s) \left| \xi_{-N_i^\kappa}^{\theta'_{-N_i^\kappa}}(a_{-N_i^\kappa}|s_{-N_i^\kappa}) - \xi_{-N_i^\kappa}^{\theta_{-N_i^\kappa}}(a_{-N_i^\kappa}|s_{-N_i^\kappa}) \right| \overline{Q\Phi}_i^{\theta}(s, a_{-N_i^\kappa})
$$

$$
\overset{(i)}{\leq} \frac{\overline{\varphi}}{(1-\gamma)^2} \sum_{s, a_{-N_i^\kappa}} d^{\theta'}(s) \left| \xi_{-N_i^\kappa}^{\theta'_{-N_i^\kappa}}(a_{-N_i^\kappa}|s_{-N_i^\kappa}) - \xi_{-N_i^\kappa}^{\theta_{-N_i^\kappa}}(a_{-N_i^\kappa}|s_{-N_i^\kappa}) \right|.
$$

Here (i) uses the fact that $\overline{Q\Phi}_i^{\theta}(s, a_{-N_i^\kappa}) \leq \frac{\overline{\varphi}}{1-\gamma}$. By Lagrange mean value theorem, for any j,

$$
\left| \xi_j^{\theta'_j}(a_j|s_j) - \xi_j^{\theta_j}(a_j|s_j) \right|
$$

$$
\leq \sup_{t \in [0,1], \hat{\theta}_j = t\theta'_j + (1-t)\theta_j} \left\| \nabla_{\theta_j} \xi_j^{\hat{\theta}_j}(a_j|s_j) \right\| \left\| \theta'_j - \theta_j \right\|
$$

$$
\leq \sqrt{2} \left\| \theta'_j - \theta_j \right\|.
$$

By Lemma F.15, we derive that

$$
\sum_{a_{-N_i^\kappa}} \left| \xi_{-N_i^\kappa}^{\theta'_{-N_i^\kappa}}(a_{-N_i^\kappa}|s_{-N_i^\kappa}) - \xi_{-N_i^\kappa}^{\theta_{-N_i^\kappa}}(a_{-N_i^\kappa}|s_{-N_i^\kappa}) \right|
$$

$$
\leq \sum_{j \in -N_i^\kappa} \left\| \xi_j^{\theta'_j}(\cdot|s_j) - \xi_j^{\theta_j}(\cdot|s_j) \right\|_1
$$

$$
= \sum_{j \in -N_i^\kappa} \sum_{a_i} \left| \xi_j^{\theta'_j}(a_j|s_j) - \xi_j^{\theta_j}(a_j|s_j) \right|
$$

$$
\leq \sqrt{2} \sum_{j \in -N_i^\kappa} |\mathcal{A}_j| \left\| \theta'_j - \theta_j \right\|
$$

$$
\leq \sqrt{2} \sum_{j \in -N_i^\kappa} |\mathcal{A}_j| \max_{j \in -N_i^\kappa} \left\| \theta'_j - \theta_j \right\|.
$$

Therefore, we arrive at the conclusion

$$
\left| \Phi_i(\theta_{N_i^\kappa}, \theta'_{-N_i^\kappa}) - \Phi_i(\theta_{N_i^\kappa}, \theta_{-N_i^\kappa}) \right|
$$

$$
\leq \frac{\overline{\varphi}}{(1-\gamma)^2} \sum_{s} d^{\theta'}(s) \sqrt{2} \sum_{j \in -N_i^\kappa} |\mathcal{A}_j| \max_{j \in -N_i^\kappa} \left\| \theta'_j - \theta_j \right\|
$$

$$
= \frac{\sqrt{2}\overline{\varphi}}{(1-\gamma)^2} \sum_{j \in -N_i^\kappa} |\mathcal{A}_j| \max_{j \in -N_i^\kappa} \left\| \theta'_j - \theta_j \right\|.
$$

$\square$

F.6 BOUNDEDNESS OF LOCAL POTENTIAL FUNCTIONS

Lemma F.6. *For any agent $i \in \mathcal{N}$, let $\xi_{N_i^{\kappa G}}, \xi'_{N_i^{\kappa G}} \in \Xi_{N_i^{\kappa G}}$ and $\xi_{-N_i^{\kappa G}} \in \Xi_{-N_i^{\kappa G}}$ be arbitrary. Then we have*

$$
\Phi_i(\xi'_{N_i^{\kappa G}}, \xi_{-N_i^{\kappa G}}) - \Phi_i(\xi_{N_i^{\kappa G}}, \xi_{-N_i^{\kappa G}}) \leq \frac{|N_i^{\kappa G}|}{1-\gamma}.
$$

Proof of Lemma F.6. Suppose that $N_i^{\kappa_G} = \{i_1, i_2, \cdots, i_k\}$, where $k = |N_i^{\kappa_G}|$. For any $r \in \{1, 2, \cdots, k+1\}$, denote $\tilde{\xi}_{N_i^{\kappa_G}}^r = (\xi_{i_1}, \cdots, \xi_{i_{r-1}}, \xi'_{i_r}, \cdots, \xi'_{i_k})$. Note that $\tilde{\xi}_{N_i^{\kappa_G}}^1 = \xi'_{N_i^{\kappa_G}}$ and $\tilde{\xi}_{N_i^{\kappa_G}}^{k+1} = \xi_{N_i^{\kappa_G}}$. Then, we have by Definition 3.1 that

$$
\begin{aligned}
\Phi_i(\xi'_{N_i^{\kappa_G}}, \xi_{-N_i^{\kappa_G}}) - \Phi_i(\xi_{N_i^{\kappa_G}}, \xi_{-N_i^{\kappa_G}}) &= \sum_{r=1}^{k} \left[\Phi_i(\tilde{\xi}_{N_i^{\kappa_G}}^r, \xi_{N_i^{\kappa_G}}) - \Phi_i(\tilde{\xi}_{N_i^{\kappa_G}}^{r+1}, \xi_{N_i^{\kappa_G}}) \right] \\
&= \sum_{r=1}^{k} \left[J_{i_r}(\tilde{\xi}_{N_i^{\kappa_G}}^r, \xi_{N_i^{\kappa_G}}) - J_{i_r}(\tilde{\xi}_{N_i^{\kappa_G}}^{r+1}, \xi_{N_i^{\kappa_G}}) \right] \\
&\leq \sum_{r=1}^{k} J_{i_r}(\tilde{\xi}_{N_i^{\kappa_G}}^r, \xi_{N_i^{\kappa_G}}) \\
&\leq \frac{k}{1-\gamma}.
\end{aligned}
$$

$\square$

Lemma F.7. *For an arbitrary NMPG, there exist a set of local potential functions $\{\hat{\Phi}_i\}_{i \in \mathcal{N}}$ and $\Phi_{\min}, \Phi_{\max} > 0$ satisfying $0 \leq \Phi_{\max} - \Phi_{\min} \leq \frac{2n(\kappa_G)}{1-\gamma}$ such that $\Phi_{\min} \leq \hat{\Phi}_i(\xi) \leq \Phi_{\max}$ for all $i \in \mathcal{N}$ and $\xi \in \Xi$.*

Proof of Lemma F.7. Let $\{\Phi_i\}_{i \in \mathcal{N}}$ be a set of local potential functions, and let $\bar{\xi} \in \Xi$ be an arbitrary policy. Define $\{\hat{\Phi}_i\}_{i \in \mathcal{N}}$ as

$$
\hat{\Phi}_i(\xi) = \Phi_i(\xi) - \Phi_i(\bar{\xi}) + \frac{n(\kappa_G)}{1-\gamma} + 1
$$

for all $i \in \mathcal{N}$ and $\xi \in \Xi$. It can be easily verified that $\{\hat{\Phi}_i\}_{i \in \mathcal{N}}$ is also a set of local potential functions. Now, for any $i \in \mathcal{N}$ and $\xi \in \Xi$, we have by Lemma F.6 that

$$
\begin{aligned}
\hat{\Phi}_i(\xi) &= \hat{\Phi}_i(\xi) - \hat{\Phi}_i(\bar{\xi}) + \frac{n(\kappa_G)}{1-\gamma} + 1 \leq \frac{2n(\kappa_G)}{1-\gamma} + 1, \\
\hat{\Phi}_i(\xi) &= - \left(\Phi_i(\bar{\xi}) - \Phi_i(\xi) - \frac{n(\kappa_G)}{1-\gamma} - 1 \right) \geq 1.
\end{aligned}
$$

Therefore, we have

$$
1 \leq \hat{\Phi}_i(\xi) \leq \frac{2n(\kappa_G)}{1-\gamma} + 1, \quad \forall i \in \mathcal{N}, \xi \in \Xi.
$$

The result follows by letting $\Phi_{\min} = 1$ and $\Phi_{\max} = \frac{2n(\kappa_G)}{1-\gamma} + 1$. $\square$

F.7 OTHER TECHNICAL LEMMAS

Lemma F.8. *Under softmax parameterization with weights $\{\theta_i\}_{i \in \mathcal{N}}$, the derivative of the policy ξ^θ is given by*

$$
\frac{\partial \xi_i^{\theta_i}(a_i|s_i)}{\partial \theta_{i,s'_i,a'_i}} = \xi_i^{\theta_i}(a_i|s_i) \mathbb{1}\{s'_i = s_i\} \left(\mathbb{1}\{a'_i = a_i\} - \xi_i^{\theta_i}(a'_i|s_i) \right)
$$

for all $i \in \mathcal{N}$, $s_i, s'_i \in \mathcal{S}_i$, $a_i, a'_i \in \mathcal{A}_i$, and $\theta_i \in \mathbb{R}^{|\mathcal{S}_i||\mathcal{A}_i|}$.

Proof of Lemma F.8. When $s'_i \neq s_i$, since $\xi_i^{\theta_i}(a_i|s_i)$ is not a function of θ_{i,s'_i,a'_i}, the result clearly holds. We next consider the case where $s'_i = s_i$.

Given an arbitrary positive integer d, let $f : \mathbb{R}^d \mapsto \mathbb{R}^d$ be the softmax operator defined as

$$
[f(x)](\ell) = \frac{\exp(x_\ell)}{\sum_{\ell'=1}^{d} \exp(x_{\ell'})}, \quad \forall \ell \in \{1, 2, \cdots, d\}.
$$

It was shown in [Gao and Pavel, 2017, Proposition 2] that $\nabla f(x) = \mathrm{diag}(f(x)) - f(x)f(x)^\top$. Therefore, we have for any $a_i' \in \mathcal{A}_i$ that

$$\frac{\partial \xi_i^{\theta_i}(a_i|s_i)}{\partial \theta_{i,s_i,a_i'}} = \xi_i^{\theta_i}(a_i|s_i)\mathbb{1}\{a_i' = a_i\} - \xi_i^{\theta_i}(a_i|s_i)\xi_i^{\theta_i}(a_i'|s_i).$$

The proof is complete. $\qquad\square$

Lemma F.9. *For any softmax policy ξ^θ, we have*

$$\|\nabla_{\theta_i}\xi_i^{\theta_i}(a_i|s_i)\| \leq \sqrt{2}\xi_i^{\theta_i}(a_i|s_i) \leq \sqrt{2}, \quad \forall\, i, s_i, a_i, \theta_i.$$

Proof of Lemma F.9. By Lemma F.8, we have

$$\|\nabla_{\theta_i}\xi_i^{\theta_i}(a_i|s_i)\| = \xi_i^{\theta_i}(a_i|s_i)\left[(1 - \xi_i^{\theta_i}(a_i|s_i))^2 + \sum_{a_i'\neq a_i}\xi_i^{\theta_i}(a_i'|s_i)^2\right]^{1/2}$$

$$\leq \xi_i^{\theta_i}(a_i|s_i)\left[(1 - \xi_i^{\theta_i}(a_i|s_i)) + \sum_{a_i'\neq a_i}\xi_i^{\theta_i}(a_i'|s_i)\right]^{1/2}$$

$$\leq \sqrt{2}\xi_i^{\theta_i}(a_i|s_i).$$

$\qquad\square$

Lemma F.10 (Multi-Agent Policy Gradient Theorem)**.** *It holds for all $i \in \mathcal{N}$ and $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ that*

$$\nabla_{\theta_i}J_i(\theta) = \frac{1}{1-\gamma}\sum_{s,a_i}d^\theta(s)\nabla_{\theta_i}\xi_i^{\theta_i}(a_i|s_i)\overline{Q}_i^\theta(s,a_i).$$

Proof of Lemma F.10. Using the policy gradient theorem [Sutton et al., 1999, Theorem 1] and we have for any $i \in \mathcal{N}$ that

$$\nabla_{\theta_i}J_i(\theta) = \frac{1}{1-\gamma}\sum_{s,a}d^\theta(s)\nabla_{\theta_i}\xi^\theta(a|s)Q_i^\theta(s,a)$$

$$= \frac{1}{1-\gamma}\sum_s\sum_{a_i}\sum_{a_{-i}}d^\theta(s)\nabla_{\theta_i}\xi^\theta(a_i,a_{-i}|s)Q_i^\theta(s,a_i,a_{-i})$$

$$= \frac{1}{1-\gamma}\sum_s\sum_{a_i}\sum_{a_{-i}}d^\theta(s)\nabla_{\theta_i}[\xi^{\theta_i}(a_i|s_i)\xi^{\theta_{-i}}(a_{-i}|s_{-i})]Q_i^\theta(s,a_i,a_{-i})$$

$$= \frac{1}{1-\gamma}\sum_s\sum_{a_i}d^\theta(s)\nabla_{\theta_i}\xi^{\theta_i}(a_i|s_i)\sum_{a_{-i}}\xi^{\theta_{-i}}(a_{-i}|s_{-i})Q_i^\theta(s,a_i,a_{-i})$$

$$= \frac{1}{1-\gamma}\sum_{s,a_i}d^\theta(s)\nabla_{\theta_i}\xi^{\theta_i}(a_i|s_i)\overline{Q}_i^\theta(s,a_i),$$

where the last line follows from the definition of the averaged Q-function. $\qquad\square$

Lemma F.11. *It holds for all $i \in \mathcal{N}$ and $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ that*

$$\|\nabla_{\theta_i}J_i(\theta)\| \leq \frac{\sqrt{2}}{(1-\gamma)^2}.$$

Proof of Lemma F.11. Using Lemma F.9 and Lemma F.10, and we have

$$
\begin{aligned}
\|\nabla_{\theta_i} J_i(\theta)\| &= \frac{1}{1-\gamma} \left\| \sum_{s,a_i} d^\theta(s) \nabla_{\theta_i} \xi_i^{\theta_i}(a_i|s_i) \overline{Q}_i^\theta(s,a_i) \right\| \\
&\leq \frac{1}{1-\gamma} \sum_{s,a_i} d^\theta(s) \left| \overline{Q}_i^\theta(s,a_i) \right| \left\| \nabla_{\theta_i} \xi_i^{\theta_i}(a_i|s_i) \right\| \\
&\leq \frac{\sqrt{2}}{(1-\gamma)^2} \sum_{s,a_i} d^\theta(s) \xi_i^{\theta_i}(a_i|s_i) \\
&= \frac{\sqrt{2}}{(1-\gamma)^2}.
\end{aligned}
$$

$\square$

Lemma F.12. *The following inequality holds for all i and θ:*

$$
\frac{\partial J_i(\theta)}{\partial \theta_{i,s_i',a_i'}} = \frac{1}{1-\gamma} \sum_{s_{-i}} d^\theta(s_i', s_{-i}) \xi_i^{\theta_i}(a_i'|s_i') \overline{A}_i^\theta(s_i', s_{-i}, a_i').
$$

Proof of Lemma F.12. Using Lemma F.8 and Lemma F.10, and we have

$$
\begin{aligned}
\frac{\partial J_i(\theta)}{\partial \theta_{i,s_i',a_i'}} &= \frac{1}{1-\gamma} \sum_{s,a_i} d^\theta(s) \frac{\partial \xi_i^{\theta_i}(a_i|s_i)}{\partial \theta_{i,s_i',a_i'}} \overline{Q}_i^\theta(s,a_i) \\
&= \frac{1}{1-\gamma} \sum_{s,a_i} d^\theta(s) \xi_i^{\theta_i}(a_i|s_i) \mathbb{1}\{s_i' = s_i\} \left(\mathbb{1}\{a_i' = a_i\} - \xi_i^{\theta_i}(a_i'|s_i) \right) \overline{Q}_i^\theta(s,a_i) \\
&= \frac{1}{1-\gamma} \sum_{s_{-i},a_i} d^\theta(s_i', s_{-i}) \xi_i^{\theta_i}(a_i|s_i') \left(\mathbb{1}\{a_i' = a_i\} - \xi_i^{\theta_i}(a_i'|s_i') \right) \overline{Q}_i^\theta(s_i', s_{-i}, a_i) \\
&= \frac{1}{1-\gamma} \sum_{s_{-i}} d^\theta(s_i', s_{-i}) \xi_i^{\theta_i}(a_i'|s_i') \left(\overline{Q}_i^\theta(s_i', s_{-i}, a_i') - \sum_{a_i} \xi_i^{\theta_i}(a_i|s_i') \overline{Q}_i^\theta(s_i', s_{-i}, a_i) \right) \\
&= \frac{1}{1-\gamma} \sum_{s_{-i}} d^\theta(s_i', s_{-i}) \xi_i^{\theta_i}(a_i'|s_i') \left(\overline{Q}_i^\theta(s_i', s_{-i}, a_i') - V_i^\theta(s_i', s_{-i}) \right) \\
&= \frac{1}{1-\gamma} \sum_{s_{-i}} d^\theta(s_i', s_{-i}) \xi_i^{\theta_i}(a_i'|s_i') \overline{A}_i^\theta(s_i', s_{-i}, a_i'),
\end{aligned}
$$

where the last line follows from the definition of the averaged advantage function. $\square$

Lemma F.13. *The following inequality holds for all $\xi, \xi' \in \Xi$, $i \in \mathcal{N}$, and $s \in \mathcal{S}$:*

$$
V_i^{\xi'}(s) - V_i^\xi(s) = \frac{1}{1-\gamma} \sum_{s',a} d_s^{\xi'}(s')(\xi'(a|s') - \xi(a|s')) Q_i^\xi(s',a).
$$

Proof of Lemma F.13. Using the performance difference lemma in the single agent setting [Agarwal et al., 2021, Lemma 2],

and we have

$$V_i^{\xi'}(s) - V_i^{\xi}(s) = \frac{1}{1-\gamma} \sum_{s',a} d_s^{\xi'}(s') \xi'(a|s') A_i^{\xi}(s',a)$$

$$= \frac{1}{1-\gamma} \sum_{s',a} d_s^{\xi'}(s') \xi'(a|s') (Q_i^{\xi}(s',a) - V_i^{\xi}(s'))$$

$$= \frac{1}{1-\gamma} \left(\sum_{s',a} d_s^{\xi'}(s') \xi'(a|s') Q_i^{\xi}(s',a) - \sum_{s'} d_s^{\xi'}(s') V_i^{\xi}(s') \right)$$

$$= \frac{1}{1-\gamma} \left(\sum_{s',a} d_s^{\xi'}(s') \xi'(a|s') Q_i^{\xi}(s',a) - \sum_{s',a} d_s^{\xi'}(s) \xi(a|s') Q_i^{\xi}(s',a) \right)$$

$$= \frac{1}{1-\gamma} \sum_{s',a} d_s^{\xi'}(s') (\xi'(a|s') - \xi(a|s')) Q_i^{\xi}(s',a).$$

$\square$

Lemma F.14. *It holds for any $i \in \mathcal{N}$, $\theta = (\theta_i, \theta_{-i})$, and $\theta' = (\theta_i', \theta_{-i})$ that*

$$J_i(\theta') - J_i(\theta) = \frac{1}{1-\gamma} \sum_{s,a_i} d^{\theta'}(s) (\xi_i^{\theta_i'}(a_i|s_i) - \xi_i^{\theta_i}(a_i|s_i)) \overline{Q}_i^{\theta}(s,a_i).$$

Proof of Lemma F.14. Using the performance difference lemma in the single agent setting [Agarwal et al., 2021, Lemma 2], and we have

$$J_i(\theta') - J_i(\theta) = \frac{1}{1-\gamma} \sum_{s,a} d^{\theta'}(s) \xi^{\theta'}(a|s) A_i^{\theta}(s,a)$$

$$= \frac{1}{1-\gamma} \sum_{s,a} d^{\theta'}(s) \xi^{\theta'}(a|s) (Q_i^{\theta}(s,a) - V_i^{\theta}(s))$$

$$= \frac{1}{1-\gamma} \left(\sum_{s,a_i} \sum_{a_{-i}} d^{\theta'}(s) \xi_i^{\theta_i'}(a_i|s_i) \xi_{-i}^{\theta_{-i}}(a_{-i}|s_{-i}) Q_i^{\theta}(s,a) - \sum_s d^{\theta'}(s) V_i^{\theta}(s) \right)$$

$$= \frac{1}{1-\gamma} \left(\sum_{s,a_i} d^{\theta'}(s) \xi_i^{\theta_i'}(a_i|s_i) \overline{Q}_i^{\theta}(s,a_i) - \sum_s d^{\theta'}(s) V_i^{\theta}(s) \right)$$

$$= \frac{1}{1-\gamma} \left(\sum_{s,a_i} d^{\theta'}(s) \xi_i^{\theta_i'}(a_i|s_i) \overline{Q}_i^{\theta}(s,a_i) - \sum_{s,a_i} d^{\theta'}(s) \xi_i^{\theta_i}(a_i|s_i) \overline{Q}_i^{\theta}(s,a_i) \right)$$

$$= \frac{1}{1-\gamma} \sum_{s,a_i} d^{\theta'}(s) (\xi_i^{\theta_i'}(a_i|s_i) - \xi_i^{\theta_i}(a_i|s_i)) \overline{Q}_i^{\theta}(s,a_i).$$

$\square$

Lemma F.15. *The following inequality holds for any $\mathcal{I} \subseteq \mathcal{N}$ and any $\xi_{\mathcal{I}}, \xi_{\mathcal{I}}' \in \Xi_{\mathcal{I}}$:*

$$\|\xi_{\mathcal{I}}(\cdot|s_{\mathcal{I}}) - \xi_{\mathcal{I}}'(\cdot|s_{\mathcal{I}})\|_1 \leq \sum_{i \in \mathcal{I}} \|\xi_i(\cdot|s_i) - \xi_i'(\cdot|s_i)\|_1, \quad \forall s_{\mathcal{I}} \in \mathcal{S}_{\mathcal{I}}.$$

Proof of Lemma F.15. The result follows by applying [Durrett, 2019, Lemma 3.4.3]. $\square$

Lemma F.16 (Property of NMPG). *In an NMPG, consider any $i \in \mathcal{N}$ and any policy parameter θ. Then the following equality holds for any $j \in N_i^{\kappa_G}$:*

$$\nabla_{\theta_j} J_j(\theta) = \nabla_{\theta_j} \Phi_i(\theta).$$

Proof of Lemma F.16. The proof essentially follows from Leonardos et al. [2022]. Using the definition of NMPG (cf. Definition 3.1), we have for any θ_j, θ'_j, and θ_{-j} that

$$J_j(\theta'_j, \theta_{-j}) - \Phi_i(\theta'_j, \theta_{-j}) = J_j(\theta_j, \theta_{-j}) - \Phi_i(\theta_j, \theta_{-j}).$$

Thus $J_j(\theta_j, \theta_{-j}) - \Phi_i(\theta_j, \theta_{-j})$ is independent of θ_j. Let $J_j(\theta_j, \theta_{-j}) - \Phi_i(\theta_j, \theta_{-j}) = U_j(\theta_{-j})$. Taking gradient with respect to θ_j on both sides, and we have

$$\nabla_{\theta_j} J_j(\theta_j, \theta_{-j}) = \nabla_{\theta_j} \Phi_i(\theta_j, \theta_{-j}).$$

$\square$