Approximate Global Convergence of Independent Learning in Multi-Agent Systems

Ruiyang Jin
Peking University

Zaiwei Chen
Purdue University

Yiheng Lin
California Institute of Technology

Jie Song
Peking University

Adam Wierman
California Institute of Technology

Abstract

Independent learning (IL) is a popular approach for achieving scalability in large-scale multi-agent systems, yet it typically lacks global convergence guarantees. In this paper, we study two representative algorithms—independent Q-learning and independent natural actor-critic—within both value-based and policy-based frameworks, and provide the first finite-sample analysis for approximate global convergence. Our results show that IL can achieve global convergence up to a fixed error arising from agent interdependence, which characterizes the fundamental limit of IL in achieving true global convergence. To establish these results, we develop a novel approach by constructing a separable Markov decision process (MDP) for convergence analysis and then bounding the gap caused by the model discrepancy between this separable MDP and the original one. Finally, we present numerical experiments using a synthetic MDP and an electric vehicle charging example to demonstrate our findings and the practical applicability of IL.

1 INTRODUCTION

Reinforcement learning (RL) (Sutton and Barto, 2018) has emerged as a powerful framework for tackling sequential decision-making problems and has achieved remarkable success across diverse applications, including Atari games (Mnih et al., 2015), the game of Go (Silver et al., 2017), robotics (Gu et al., 2017), and nuclear fusion control (Degrave et al., 2022), among others. While early RL research primarily focused on the single-agent setting, the growing interest in applications involving multi-agent interactions—such as multiplayer games (Yang and Wang, 2020) and active voltage control (Wang et al., 2021)—has driven increasing attention toward multi-agent reinforcement learning (MARL).

Compared to single-agent RL, MARL presents unique challenges due to the interplay among agents, making it significantly harder to design provably efficient learning algorithms. One of the key challenges is scalability. Specifically, even when agents share a common objective, MARL algorithms that rely on centralized training become computationally intractable as the number of agents increases. To address this scalability issue, researchers have developed various decentralized MARL approaches, including centralized training with decentralized execution (CTDE) (Lowe et al., 2017), localized MARL in networked systems (Zhang et al., 2018b), and independent learning (IL) (Lanctot et al., 2017). Among these, IL imposes the least requirements for coordination and information sharing, as each agent makes decisions based solely on its local observations without gathering information from others. Consequently, IL maintains a constant communication and computational complexity regardless of the number of agents.

Although IL is simple, intuitive, and easy to implement, it remains unclear whether agents can jointly achieve any performance guarantees through IL. In IL, each agent disregards multi-agent interactions by treating all other agents as part of the environment, which is inherently non-stationary due to the changing actions of other agents. As a result, except in cases where the underlying model possesses a special structure—such as zero-sum stochastic games (Chen et al., 2024; Cai

Proceedings of the 28th International Conference on Artificial Intelligence and Statistics (AISTATS) 2025, Mai Khao, Thailand. PMLR: Volume 258. Copyright 2025 by the author(s).

et al., 2024) or Markov potential games (Leonardos et al., 2022; Jordan et al., 2024)—IL is generally not guaranteed to converge (Zhang et al., 2021a). This observation raises the following fundamental question:

Is it possible to establish (approximate) global convergence for IL in multi-agent systems?

In this paper, we provide a positive answer to this question in the cooperative setting by establishing the first finite-sample bounds for the approximate global convergence of IL. Furthermore, we introduce the concept of *dependence level* as a novel measure to characterize the fundamental limitation of IL. Our key contributions are summarized below.

- **Approximate Global Convergence for Independent Learning.** We consider two widely used and representative forms of IL: a value-based algorithm called independent Q-learning (IQL) and a policy-based algorithm within the actor-critic framework called independent natural actor-critic (INAC). We establish the first last-iterate finite-sample bounds measured by the global optimality gap. A key feature of our results is that the bound asymptotically converges within a fixed error term, which is proportional to the dependence level, denoted by $\mathcal{E}$, of the MDP model. The dependence level $\mathcal{E}$ quantifies how close the MARL model is to one where each agent's transitions are independent of others. Our finite-sample bound implies that, apart from this asymptotic error term, the sample complexity for the remaining terms to achieve ϵ-optimality is $\tilde{\mathcal{O}}(\epsilon^{-2})$, which is minimax-optimal (Gheshlaghi Azar et al., 2013).

- **A Novel Approach for Analyzing Independent Learning.** A key challenge in analyzing IL is that each agent disregards multi-agent interactions by treating other agents as part of the environment. To address this, we develop a novel approach (illustrated in Figure 1) that consists of three main steps: *(i)* constructing a separable MDP with local transition kernels to approximate the original MDP, *(ii)* analyzing IL as if it were executed on the separable MDP, and *(iii)* bounding the error introduced by the model discrepancy between the separable MDP and the original one to establish approximate global convergence.

Beyond IL, our proof technique has broader applicability. Specifically, in a general stochastic iterative algorithm, as long as the underlying random process driving the iterations can be approximated by a Markov chain (even if the original process is not Markovian), our approach can potentially be used to derive finite-sample guarantees. A more detailed discussion is provided in Appendix B.4.

- **Validation of Approximate Global Convergence of IL.** We conduct numerical simulations on a synthetic MDP to demonstrate that the asymptotic optimality gap of IL is determined by the dependence level. Additionally, we apply IL to an electric vehicle (EV) charging problem to evaluate its practical applicability, employing neural networks as function approximators. Notably, while our theoretical results are derived for the tabular setting, our empirical findings suggest that IL can achieve approximate global convergence even under function approximation.

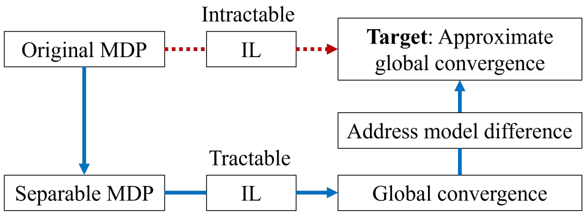

Figure 1: Our Roadmap for Analyzing IL.

The remainder of this paper is organized as follows. In Section 2, we introduce our model. Section 3 presents our main results, including two IL algorithms and their finite-sample convergence guarantees. We then provide numerical simulations in Section 4 and conclude the paper in Section 5. Due to space constraints, a detailed literature review and full proofs of all technical results are included in the appendix.

2 PROBLEM FORMULATION

Consider an MARL problem with n agents. Let $\mathcal{S}$ and $\mathcal{A}$ denote the global state space and global action space, respectively. For each agent $i \in \{1, 2, \dots, n\} := [n]$, let $\mathcal{S}^i$ and $\mathcal{A}^i$ represent its local state space and action space, respectively. Throughout this work, we assume that $|\mathcal{S}||\mathcal{A}| < \infty$. Let $\mathcal{P} = \{P_a \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|} \mid a \in \mathcal{A}\}$ be the set of transition probability matrices indexed by joint actions, where $P_a(s, s')$ denotes the probability of transitioning to the global state s' after taking the joint action a in global state s. For each agent $i \in [n]$, let $\mathcal{R}^i : \mathcal{S}^i \times \mathcal{A}^i \to [0, 1]$ be its reward function. Without loss of generality, we assume the reward function is bounded within $[0, 1]$, as we are considering a finite MARL problem. Given a global state-action pair (s, a), where $s = (s^1, s^2, \dots, s^n) \in \mathcal{S}$ and $a = (a^1, a^2, \dots, a^n) \in \mathcal{A}$, the total one-stage reward is defined as $\mathcal{R}(s, a) = \sum_{i\in[n]} \mathcal{R}^i(s^i, a^i)$. Let $\gamma \in (0, 1)$ be the discount factor, which determines the relative importance of future rewards. We denote the MDP associated with the MARL problem as $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$.

Given a joint policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$, where $\Delta(\mathcal{A})$ denotes the probability simplex over $\mathcal{A}$, the Q-function $Q_\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ is defined as

$$Q_\pi(s,a) = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty}\gamma^k \mathcal{R}(S_k, A_k)\,\middle|\, S_0 = s, A_0 = a\right]$$

for all $(s,a) \in \mathcal{S}\times\mathcal{A}$, where $\mathbb{E}_\pi[\,\cdot\,]$ indicates that actions are selected according to the joint policy π. The value function $V_\pi \in \mathbb{R}^{|\mathcal{S}|}$ is then defined as $V_\pi(s) = \mathbb{E}_{a\sim\pi(\cdot|s)}[Q_\pi(s,a)]$ for all $s \in \mathcal{S}$. Given an initial state distribution $\mu \in \Delta(\mathcal{S})$, we denote the expected value of policy π as $V_\pi^\mu = \mathbb{E}_{s\sim\mu}[V_\pi(s)]$. For an individual agent $i \in [n]$, we define its local Q-function $Q_\pi^i \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ under policy π as

$$Q^i_\pi(s,a) = \mathbb{E}_\pi\left[\sum_{k=0}^{\infty}\gamma^k \mathcal{R}^i(S^i_k, A^i_k)\,\middle|\, S_0 = s, A_0 = a\right]$$

for all $(s,a)\in\mathcal{S}\times\mathcal{A}$. Since the total reward is the sum of individual agent rewards, it follows that $Q_\pi(s,a) = \sum_{i\in[n]} Q^i_\pi(s,a)$ for all $(s,a)\in\mathcal{S}\times\mathcal{A}$.

As a remark, since we are considering the cooperative setting, global optimality and the optimality gap are well-defined. In more general settings where agents are competitive (such as zero-sum games), some notion of equilibrium is required to evaluate the performance of a joint policy.

2.1 The Dependence Level

We now introduce the metric of *dependence level* for a multi-agent MDP model to characterize globally coupled state transitions. This is a new notion for analyzing IL that we introduce in this paper. To define the dependence level, we first illustrate the concept of a *separable MDP*, which plays a crucial role in our analysis.

For any $i \in [n]$, let Δ^i be the set of transition probability matrices with state space $\mathcal{S}^i$. We then define the set of separable transition kernels as

$$\hat{\mathcal{Z}} = \left\{\{\hat{P}_a\}_{a\in\mathcal{A}}\,\middle|\, \hat{P}_a(s_1,s_2) = \prod_{i=1}^n \hat{P}_{a^i}(s_1^i, s_2^i),\right.$$
$$\left.\text{for any } s_1, s_2 \in \mathcal{S}, \text{where } \hat{P}_{a^i}\in\Delta^i \text{ for all } i\in[n]\right\}.$$

We call an MDP with transition matrices $\{\hat{P}_a\}_{a\in\mathcal{A}}$ a *separable MDP* if $\{\hat{P}_a\}_{a\in\mathcal{A}} \in \hat{\mathcal{Z}}$. Intuitively, while a separable MDP is defined on the joint state-action space, it can be decomposed into n independent MDPs, each evolving on its respective local state-action space. Now, we are ready to define the *dependence level* of a multi-agent MDP model.

Definition 2.1. *An MDP with n agents and transition probability matrices $\{P_a\}_{a\in\mathcal{A}}$ is said to be $\mathcal{E}$-dependent if and only if*

$$\min_{\{\hat{P}_a\}\in\hat{\mathcal{Z}}}\max_{s,a}\left\|P_a(s,\cdot) - \hat{P}_a(s,\cdot)\right\|_{\text{TV}} = \mathcal{E}, \qquad (1)$$

where $\mathcal{E}$ is called the dependence level and $\|\cdot\|_{TV}$ denotes the total variation distance.

Remark 2.2. The $\min(\cdot)$ in Eq. (1) is always well-defined since we are minimizing a continuous function over a compact set (Rudin et al., 1976).

To understand Definition 2.1, consider the case where the original MDP $\mathcal{M}$ is separable. In this case, the dependence level is zero. Since the problem reduces to finding the optimal policies for n decoupled single-agent RL problems, we would expect IL to achieve global convergence. More generally, when $\mathcal{E} > 0$, the original MDP is not separable, meaning exact global convergence may not be achievable. Thus, the dependence level $\mathcal{E}$ quantifies how close the original MDP is to the space of separable MDPs and characterizes the fundamental limit of IL in achieving global convergence.

For simplicity of notation, let $\hat{\mathcal{P}}$ be a solution to $\arg\min_{\{\hat{P}_a\}\in\hat{\mathcal{Z}}}\max_{s,a}\left\|P_a(s,\cdot)-\hat{P}_a(s,\cdot)\right\|_{\text{TV}}$, and denote $\hat{\mathcal{M}} = (\mathcal{S}, \mathcal{A}, \hat{\mathcal{P}}, \mathcal{R}, \gamma)$, which consists of n decoupled MDPs. We use the hat notation throughout the paper to denote the counterpart quantities for the separable MDP $\hat{\mathcal{M}}$. For example, we denote $\hat{Q}_\pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ as the Q-function of a policy π under $\hat{\mathcal{M}}$.

2.2 Illustrative Examples

To further illustrate the concepts of separable MDPs and the dependence level, we present several examples in this section to enhance understanding.

In multi-agent systems, local information sharing and coordinated decision-making can significantly enhance the performance of IL (Tan, 1993; Böhmer et al., 2019). However, selecting the relevant information to communicate and determining which agents should collaborate is challenging. The following example demonstrates how different choices of grouping and collaboration impact the dependence level, which in turn significantly influences the performance of IL.

Example 1. Consider an MDP consisting of three agents. The state space of each agent is $\{0,1\}$. The action space of each agent is also $\{0,1\}$. There are different dependencies among the states of the three agents with detailed transition probabilities shown in Appendix E.1. Suppose that we are allowed to group two of the three agents as one agent, in which case the two agents can share information and coordinate with each other. Then, we have three different grouping options shown in Figure 2. After solving the optimization

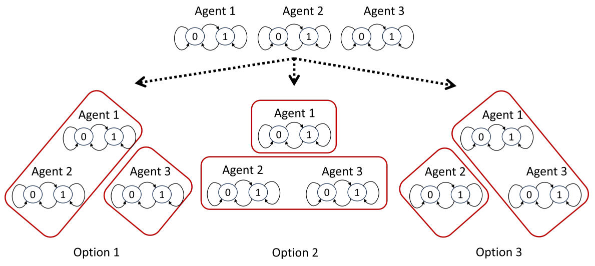

Figure 2: Illustration of Example 1.

problem in Eq. (1) based on the state transition model (cf. Appendix E.1), we can calculate the dependence levels: Option 1 leads to $\mathcal{E} = 0.5$; Option 2 leads to $\mathcal{E} = 0.75$, and Option 3 leads to $\mathcal{E} = 0.875$.

Example 2. Consider the operation of distributed energy storage (ES) in a distribution network comprising a set of buses $\mathcal{N}$. Each ES (i.e., an agent in the multi-agent system) performs IL to jointly maximize total revenue by controlling local charging/discharging while ensuring voltage safety in the distribution system (Xiang et al., 2023). For simplicity, consider an MDP where the state consists only of a voltage safety indicator, i.e., $s = (s^1, s^2, \ldots, s^B)$, where B is the number of ESs. The local state $s^i = 1$ if the local voltage is safe and $s^i = 0$ otherwise. The action $a = (a^1, a^2, \ldots, a^B)$ represents the charging/discharging decisions that influence voltage safety. Each ES's reward is determined by its local voltage safety, i.e., its local state s^i. See Appendix E.2 for the full problem formulation.

In this application, the placement of ESs significantly impacts the performance of IL. Suppose the locations are chosen such that each ES's local action has a limited effect on the voltage safety of others. Mathematically, this means $P_a(s_1, s_2) \approx \prod_{i=1}^{B} \mathbb{P}(s_2^i \mid s_1^i, a^i)$ for all $s_1, s_2 \in \mathcal{S}, a \in \mathcal{A}$. In this case, the dependence level is small, and global safety can be approximately ensured if each ES independently maintains its local safety. See Appendix E.2 for a more detailed discussion of this intuition.

3 MAIN RESULTS

This section presents our main results. First, we present the IQL and INAC algorithms in Sections 3.1 and 3.2, respectively. Then, we present the approximate finite-sample global convergence guarantees for both algorithms in Section 3.3. The proof sketch of our main theorems is given in Section 3.5.

3.1 Independent Q-Learning

In IQL, each agent treats all other agents as part of the environment and implements Q-learning on its local state-action space, as presented in Algorithm 1.

Algorithm 1 IQL (Agent i)

```
Input: Integer $K$, behavior policy $\pi_b^i$, and initialization $Q_0^i = 0$.
for $k = 0, 1, 2, \cdots, K-1$ do
    Implement $A_k^i \sim \pi_b^i(\cdot \mid S_k^i)$ (simultaneously with all other agents), and observes $S_{k+1}^i$.
    Update $Q_k^i$ according to Eq. (2).
end for
Output: $Q_K^i$
```

Algorithm Details. In IQL, each agent uses a local behavior policy $\pi_b^i : \mathcal{S}^i \to \Delta(\mathcal{A}^i)$ to interact with the environment to collect samples and updates its local Q-function estimate $Q_k^i \in \mathbb{R}^{|\mathcal{S}^i||\mathcal{A}^i|}$ according to the following formula for any $k \geq 0$:

$$Q_{k+1}^i(S_k^i, A_k^i) = (1 - \alpha_k) Q_k^i(S_k^i, A_k^i) + \alpha_k \big(\mathcal{R}^i(S_k^i, A_k^i) + \gamma \max_{\bar{a}^i} Q_k^i(S_{k+1}^i, \bar{a}^i) \big), \qquad (2)$$

where α_k is the learning rate (also known as the stepsize). Note that this is an asynchronous update because only one component of the vector-valued local Q-function is updated at each step. The update of Q-learning can be viewed as a stochastic approximation algorithm for solving the Bellman optimality equation. See Bertsekas and Tsitsiklis (1996); Watkins and Dayan (1992) for more details of Q-learning.

3.2 Independent Natural Actor-Critic

A detailed description of the INAC algorithm is provided in Algorithm 2. At a high level, INAC consists of an actor in the outer loop and a critic in the inner loop. The actor updates the policy using independent natural policy gradient (INPG), while the critic estimates the local Q-function using asynchronous independent TD-learning (ITD), which the actor relies on for updates. We next elaborate on the actor and critic in more detail.

INPG for the Actor. Policy gradient (Sutton and Barto, 2018) is a popular approach to solving the RL problem. The idea is to perform gradient ascent in the policy space. The NPG framework can be viewed as a variant of the policy gradient, where the Fisher information matrix is used as a preconditioner (Kakade, 2001). There are many equivalent formulations and interpretations of NPG (Kakade, 2001; Agarwal et al., 2021; Lan, 2023).

Algorithm 2 INAC (Agent i)

```
Input: Integers $K, T$, initializations $\theta_0^i = 0$ and $Q_{t,0}^i = 0$ for all $t \geq 0$.
for $t = 0, 1, 2, \cdots, T-1$ do
    for $k = 0, 1, 2, \cdots, K-1$ do
        Implement $A_k^i \sim \pi_{\theta_t^i}^i(\cdot \mid S_k^i)$ (simultaneously with all other agents), and observes $S_{k+1}^i$.
        Update $Q_{t,k}^i$ according to Eq. (6).
    end for
    $\theta_{t+1}^i = \theta_t^i + \eta_t Q_{t,K}^i$.
end for
Output: $\pi_{\theta_T^i}^i$
```

Next, we introduce the NPG algorithm. Consider using softmax policies of the form

$$\pi_\theta(a \mid s) = \frac{\exp(\theta_{s,a})}{\sum_{\bar{a} \in \mathcal{A}} \exp(\theta_{s,\bar{a}})}, \; \forall (s,a) \in \mathcal{S} \times \mathcal{A},$$

where $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ is the parameter. It was shown in Agarwal et al. (2021) that NPG takes the following form in the parameter space (which is also called Q-NPG):

$$\theta_{t+1} = \theta_t + \eta_t Q_{(t)}, \tag{3}$$

where η_t is the stepsize. Here we denote $\pi_{(t)} = \pi_{\theta_t}$ and $Q_{(t)} = Q_{\pi_{(t)}}$ for simplicity. In addition, the previous update equation in the parameter space is equivalent to the following update equation in the policy space (Agarwal et al., 2021):

$$\pi_{(t+1)}(a \,|\, s) = \frac{\pi_{(t)}(a|s) \exp\{\eta_t Q_{(t)}(s,a)\}}{\sum_{a' \in \mathcal{A}} \pi_{(t)}(a'|s) \exp\{\eta_t Q_{(t)}(s,a')\}} \tag{4}$$

for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. However, carrying out the update rule in Eq. (3) would require the agents to jointly estimate the global Q-function, which, in general, cannot be achieved with IL.

To enable the use of IL, we propose that agent i maintains its own parameter $\theta_t^i \in \mathbb{R}^{|\mathcal{S}^i||\mathcal{A}^i|}$ and updates it according to

$$\theta_{t+1}^i = \theta_t^i + \eta_t \hat{q}_{(t)}^i, \tag{5}$$

where $\hat{q}_{(t)}^i \in \mathbb{R}^{|\mathcal{S}^i||\mathcal{A}^i|}$ is the local Q-function[1] of agent i associated with the policy $\pi_{(t)}$ under the *separable MDP* model $\hat{\mathcal{M}}$. Specifically, given a separable joint policy $\pi = (\pi^1, \pi^2, \cdots, \pi^n)$ with $\pi^i : \mathcal{S}^i \to \Delta(\mathcal{A}^i)$ for any agent i, we define $\hat{q}_\pi^i(s^i, a^i) = \hat{\mathbb{E}}_\pi[\sum_{k=0}^\infty \gamma^k \mathcal{R}^i(S_k^i, A_k^i) \mid S_0^i = s^i, A_0^i = a^i]$ for all $(s^i, a^i) \in \mathcal{S}^i \times \mathcal{A}^i$ and $i \in [n]$, where $\hat{\mathbb{E}}_\pi[\,\cdot\,]$ denotes the expectation with respect to the separable transition kernel $\hat{\mathcal{P}}$ and policy π. To make sense of Eq. (5), note that Eq. (5) is equivalent to the following update in the policy space:

$$\pi_{(t+1)}^i(a^i \mid s^i) = \frac{\pi_{(t)}^i(a^i \mid s^i) \exp\{\eta_t \hat{q}_{(t)}^i(s^i, a^i)\}}{\sum_{\bar{a}^i \in \mathcal{A}^i} \pi_{(t)}^i(\bar{a}^i \mid s^i) \exp\{\eta_t \hat{q}_{(t)}^i(s^i, \bar{a}^i)\}}$$

for all $(s^i, a^i) \in \mathcal{S}^i \times \mathcal{A}^i$. Suppose that the original MDP $\mathcal{M}$ itself is separable, in which case we have $\mathcal{M} = \hat{\mathcal{M}}$. Then, for any $s = (s^1, s^2, \cdots, s^n) \in \mathcal{S}$ and $a = (a^1, a^2, \cdots, a^n) \in \mathcal{A}$, it is clear that $Q_\pi(s,a) = \sum_{i \in [n]} \hat{q}_\pi^i(s^i, a^i)$. As a result, when each agent updates its policy parameter θ_t^i according to Eq. (5), it is easy to see that the joint policy obeys the update rule in Eq. (4). In general, when the original n-agent MDP $\mathcal{M}$ is not separable, Eq. (5) can be viewed as an approximation of the Q-NPG update in Eq. (4). Explicitly characterizing such an approximation error is one of the major technical challenges in the analysis. As we shall see later, the approximation error will be captured by the dependence level $\mathcal{E}$.

In view of Eq. (5), each agent needs to estimate its local Q-function $\hat{q}_{(t)}^i$ to carry out the update. To achieve that, we use ITD, which is presented next.

ITD for the Critic. Within each iteration $t \in \{0, 1, \cdots, T-1\}$ of the outer loop, each agent performs policy evaluation independently according to the following TD-learning update rule for any $k \geq 0$:

$$\begin{aligned} &Q_{t,k+1}^i(S_k^i, A_k^i) = (1 - \alpha_k) Q_{t,k}^i(S_k^i, A_k^i) \\ &\quad + \alpha_k \big(\mathcal{R}^i(S_k^i, A_k^i) + \gamma \mathbb{E}_{a^i \sim \pi_{(t)}^i}\left[Q_{t,k}^i(S_{k+1}^i, a^i)\right]\big). \end{aligned} \tag{6}$$

Similar to Q-learning, TD-learning can also be viewed as a stochastic approximation algorithm for solving the Bellman equation for policy evaluation. See Sutton and Barto (2018); Bertsekas and Tsitsiklis (1996); Sutton (1988) for more details about TD-learning.

Although ITD is trying to estimate the local Q-function, the local sample trajectory $\{(S_k^i, A_k^i)\}_{k \geq 0}$ from the original MDP $\mathcal{M}$ does not form a Markov chain. Thus, the convergence of local Q-function $\{Q_{t,k}^i\}_{k \geq 0}$ is unclear and cannot be dealt with by existing standard RL theory, which presents a challenge in the analysis.

Finally, combining INPG in Eq. (5) and ITD in Eq. (6) leads to INAC in Algorithm 2.

3.3 Finite-Sample Analysis

To present our theoretical results, we first introduce our assumption regarding the MDP model.

Assumption 3.1. *For any joint policy π, the induced Markov chain $\{(S_k, A_k)\}_{k \geq 0}$ is irreducible and aperiodic with a unique stationary distribution, denoted*

[1] Here, we use the notation $\hat{q}_\pi^i$ to distinguish with the Q-function $\hat{Q}_\pi^i$ given policy π and agent i, which is defined in the global state-action space.

by $d_\pi \in \Delta(\mathcal{S} \times \mathcal{A})$. In addition, we assume that (1) $\sigma := \inf_\pi \min_{s,a} d_\pi(s,a) > 0$, and (2) there exist $M_1 \geq 0$ and $M_2 \geq 1$ such that

$$\left|\sum_{(s,a)\in\mathcal{N}} (d_\pi(s,a) - \mathbb{P}_k^\pi(\bar{s},\bar{a},s,a))\right| \leq M_1 \exp\left(-\frac{k}{M_2}\right)$$

for any $\mathcal{N} \subseteq \mathcal{S}\times\mathcal{A}$, $(\bar{s},\bar{a}) \in \mathcal{S}\times\mathcal{A}$, policy π, and $k \geq 0$, where $\mathbb{P}_k^\pi(\bar{s},\bar{a},s,a) := \mathbb{P}((S_k,A_k) = (s,a) \mid (S_0,A_0) = (\bar{s},\bar{a}))$ for any $(s,a),(\bar{s},\bar{a}) \in \mathcal{S}\times\mathcal{A}$.

Remark 3.2. Since we are working with a finite MDP, fixing a policy π, the induced Markov chain $\{(S_k,A_k)\}_{k\geq 0}$ being irreducible and aperiodic implies: (1) the existence and uniqueness of a stationary distribution d_π satisfying $d_\pi(s,a) > 0$ for all (s,a), and (2) the geometric mixing property (Levin and Peres, 2017). However, the quantity $\min_{s,a} d_\pi(s,a)$ and the mixing coefficients may depend on π. This is the reason we impose Assumption 3.1 (1) and (2), which can be viewed as a "uniform" mixing property.

In RL, successful policy learning requires a sufficient exploration component. Assumption 3.1 ensures this by stating that for any joint policy π, the induced Markov chain is uniformly ergodic, guaranteeing that agents can sufficiently explore the state-action space. This assumption is commonly used in the literature on RL with time-varying policies; see, for example, Lin et al. (2021); Wu et al. (2020); Zou et al. (2019); Zeng et al. (2022); Khodadadian et al. (2022a).

To proceed, we need the following notation. Let $d'_\pi(\bar{s}^i,\bar{a}^i) = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}, s^i=\bar{s}^i, a^i=\bar{a}^i} d_\pi(s,a)$ for all $(\bar{s}^i,\bar{a}^i) \in \mathcal{S}^i \times \mathcal{A}^i$. Denote $m = \max_{i\in[n]} |\mathcal{S}^i||\mathcal{A}^i|$ and $\sigma' = \inf_\pi \min_{i\in[n],(\bar{s}^i,\bar{a}^i)\in\mathcal{S}^i\times\mathcal{A}^i} d'_\pi(\bar{s}^i,\bar{a}^i)$, which is strictly positive under Assumption 3.1.

We first present the finite-sample analysis of IQL. Given a Q-function Q_k^i generated by Algorithm 1, let π_k^i be the policy greedily induced by Q_k^i, that is, $\pi_k^i(a^i \mid s^i) = 1$ if and only if $a^i = \arg\max_{\bar{a}^i} Q_k^i(s^i,\bar{a}^i)$, where we break the tie arbitrarily.

Theorem 3.3. *Consider $\{Q_k\}_{k\geq 0}$ generated by Algorithm 1. Suppose that Assumption 3.1 is satisfied and $\alpha_k = \frac{\alpha}{k+k_0}$ with $k_0 = \max(4\alpha, 2M_2\log K)$ and $\alpha \geq \frac{2}{\sigma'(1-\gamma)}$. Then, for any $\delta' \in (0,1)$, with probability at least $1-\delta'$, we have*

$$V_{\pi_*}^\mu - V_{\pi_K}^\mu \leq \underbrace{\frac{1}{1-\gamma}\left(\frac{2nC'_a}{\sqrt{K+k_0}} + \frac{2nC_b}{K+k_0}\right)}_{E_1\text{: Q-Learning Convergence Error}} + \underbrace{\frac{8n\gamma\mathcal{E}}{(1-\gamma)^3}}_{E_2\text{: Error due to } \mathcal{E}}, \tag{7}$$

where π_* is an optimal policy and $\pi_K = (\pi_k^1, \pi_k^2, \ldots, \pi_k^n)$. The constants C'_a and C_b are defined as $C'_a = \frac{40\alpha}{(1-\gamma)^2}\sqrt{2M_2\log K\log\left(\frac{4mnM_2K}{\delta'}\right)}$, and $C_b = 8\max\{\frac{144M_2\alpha\log K+4M_1\sigma'(1+2M_2+4\alpha)}{(1-\gamma)^2\sigma'}, \frac{2M_2\log K+k_0}{(1-\gamma)^2}\}$.

Remark 3.4. Since IQL uses a fixed behavior policy π_b to collect samples, Assumption 3.1 can be relaxed to the following weaker assumption: the Markov chain $\{(S_k,A_k)\}$ induced by π_b is irreducible and aperiodic.

The proof of Theorem 3.3 is provided in Appendix B. The convergence bound in Eq. (7) consists of two terms. The first term, E_1, converges to zero at a rate of $\tilde{\mathcal{O}}(1/\sqrt{K})$, matching the convergence rate of Q-learning in the single-agent setting (Qu and Wierman, 2020; Li et al., 2024). The second term, E_2, is asymptotically non-vanishing. Notably, E_2 is proportional to the dependence level $\mathcal{E}$, and captures the fundamental limit of IQL. In the special case where the original MDP $\mathcal{M}$ is separable, E_2 vanishes, and implies the global convergence of IQL. Based on Theorem 3.3, we have the following sample complexity of IQL.

Corollary 3.5. *Given $\epsilon > 0$, the sample complexity*[2] *is $\tilde{\mathcal{O}}(\epsilon^{-2})$ for Algorithm 1 to achieve $V_{\pi_*}^\mu - V_{\pi_k}^\mu \leq \epsilon + \frac{8n\gamma\mathcal{E}}{(1-\gamma)^3}$ with probability at least $1-\delta'$, where the $\tilde{O}$ notation hides the dependence on $\log(\delta'^{-1})$.*

As seen in Corollary 3.5, up to a model difference error proportional to the dependence level, Algorithm 1 achieves a sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-2})$ for finding an optimal policy. This sample complexity is known to be optimal, as it matches existing lower bounds for solving RL problems (Gheshlaghi Azar et al., 2013).

Next, we present the finite-sample guarantees for INAC.

Theorem 3.6. *Consider $\{\pi_{(t)} = (\pi_{(t)}^i)_{i\in[d]}\}_{t\geq 0}$ generated by Algorithm 2. Suppose that (i) Assumption 3.1 is satisfied, (ii) $\alpha_k = \alpha/(k+k_0)$ with $\alpha \geq \frac{2}{\sigma'(1-\gamma)}$ and $k_0 = \max(4\alpha, 2M_2\log K)$, and (iii) η_t satisfies $\eta_0 = \gamma\log|\mathcal{A}|$ and $\eta_t \geq (\frac{2n}{1-\gamma}\sum_{j=0}^{t-1}\eta_j + \log|\mathcal{A}|)/\gamma^{2t-1}$ for all $t \geq 0$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$, we have*

$$\|Q_{\pi_*} - Q_{(T)}\|_\infty \leq \underbrace{\frac{2}{(1-\gamma)^2}\left(\frac{nC_a}{\sqrt{K+k_0}} + \frac{nC_b}{K+k_0}\right)}_{G_1\text{: TD-Learning Convergence Error}}$$

[2] It was argued in Khodadadian et al. (2021) that, given a finite-sample bound with asymptotically non-vanishing terms on the right-hand side, the interpretation of the finite-sample bound in terms of sample complexity can be ambiguous, as it is possible to trade-off the vanishing terms and the non-vanishing terms to obtain "better" sample complexity guarantees. In Corollary 3.5, we present the sample complexity in this way to allow a fair comparison with the existing literature, as a finite-sample bound with non-vanishing terms can frequently occur in RL when (i) function approximation is used, (ii) off-policy sampling is used, and (iii) IL is used as in our paper.

$$+ \underbrace{\frac{4n\gamma^{T-1}}{(1-\gamma)^2}}_{G_2:\textit{Actor Convergence Error}} + \underbrace{\frac{8n\gamma\mathcal{E}}{(1-\gamma)^4}}_{G_3:\textit{Error due to }\mathcal{E}}, \qquad (8)$$

where π_* is an optimal policy, C_a is defined as $C_a = \frac{40\alpha}{(1-\gamma)^2}\sqrt{2M_2 \log K \log\left(\frac{4mnTM_2K}{\delta}\right)}$, and C_b is defined in Theorem 3.3.

The detailed proof of Theorem 3.6 is provided in Appendix B. Similar to Theorem 3.3, the convergence bound in Theorem 3.6 consists of terms that asymptotically converge to zero and a term proportional to the dependence level $\mathcal{E}$. Specifically, the terms G_1 and G_2 in Eq. (8) correspond to the convergence errors in ITD for the critic and INPG for the actor, respectively. The term G_3 is independent of the number of iterations and captures the model difference error between the original MDP and the separable one. Recall that our approach to analyzing IL involves introducing a separable MDP $\hat{\mathcal{M}}$, treating INAC as if it were implemented on $\hat{\mathcal{M}}$, and then separately bounding the model difference error, which leads to the term G_3.

At first glance, the increasing stepsize sequence η_t in INAC (see the recursive geometric form in Theorem 3.6) may seem counterintuitive. To illustrate this, consider the single-agent setting. From the equivalent form of Q-NPG in Eq. (4), it resembles classical policy iteration, which exhibits geometric convergence as η_t approaches infinity. This intuition has been theoretically justified in Xiao (2022); Chen and Maguluri (2022b); Khodadadian et al. (2022b). An alternative interpretation of Q-NPG based on mirror descent Lan (2023) also requires using increasing stepsizes to achieve geometric convergence. While theoretically justified, excessively large stepsizes in practice may hinder exploration. To address this issue, one can use an exploration-encouraging variant of INAC, such as selecting actions based on $\pi^i_{(t)}$ with probability $1-\epsilon$ and choosing actions uniformly at random with probability ϵ.

Next, we also derive the sample complexity of Algorithm 2 based on Theorem 3.6.

Corollary 3.7. *Given $\epsilon > 0$, the sample complexity is $\tilde{\mathcal{O}}\left(\epsilon^{-2}\right)$ for Algorithm 2 to achieve $\left\|Q_{\pi_*} - Q_{(t)}\right\|_\infty \leq \epsilon + \frac{8n\gamma\mathcal{E}}{(1-\gamma)^4}$ with probability at least $1-\delta$, where the $\tilde{O}$ notation hides the dependence on $\log(\delta^{-1})$.*

Similarly to IQL, we have an $\tilde{\mathcal{O}}(\epsilon^{-2})$ sample complexity for INAC to find a global optimal policy up to a model difference error.

3.4 Discussion

In this section, we further discuss our results in terms of their tightness, the estimation of the dependence level to characterize the asymptotic optimality gap, and closely related work in the literature.

Tightness of the Convergence Bounds. While the asymptotically vanishing terms (in both Theorems 3.3 and 3.6) have the optimal order-wise sample complexity, the non-vanishing terms scale polynomially with respect to the effective horizon $1/(1-\gamma)$. The fact that these non-vanishing terms increase with the effective horizon is intuitive, as any error incurred by employing IL at each step will contribute to an overall cumulative error. However, despite this dependence, while the derivation of our bound does not depend on $\mathcal{E}$, our convergence bound is theoretically meaningful only when $\mathcal{E}$ is sufficiently small. Therefore, it is an interesting future direction to reduce the polynomial dependence on the effective horizon. That being said, similar phenomena also appear in other related studies, such as RL with function approximation and networked MARL. Specifically, in the convergence bound of TD-learning with linear function approximation, the Bellman completeness error is scaled by $1/(1-\gamma)$ (Tsitsiklis and Van Roy, 1996). Similarly, in networked MARL, the induced error due to localized algorithms (which is not even IL) scales as $1/(1-\gamma)^5$ (Qu et al., 2022). Moreover, our finite-sample analysis only provides an upper bound, so having a small dependence level is merely a sufficient condition for approximate global convergence. This is reflected in our numerical simulations, where $\mathcal{E}$ can be relatively large, yet the empirical results are still successful.

Estimation/Calculation of the Dependence Level. The dependence level is an important notion introduced in this work to characterize the fundamental limit of IL in terms of global convergence. Although neither the implementation of our algorithms nor the derivation of the convergence guarantees requires the explicit calculation of the dependence level, it is of practical interest to get an estimate of it to validate the implementation of IL. For finite MDPs with relatively small state-action spaces, the dependence level can be relatively easily computed by solving the optimization problem in (1). However, it can be complicated or even in principle intractable when the state-action space is large or the transition probabilities are unknown, which are common in model-free RL settings. Developing an efficient approach to estimate the dependence level under Definition (1) can be a future direction.

Comparison with Results for Networked MARL. Among existing literature, the works closest to ours are those analyzing networked MARL problems. The main idea in networked MARL is that, from each agent's perspective, agents that are far away in graph distance should have a negligible impact on the agent. This is referred to as the "exponential decay property" in networked MARL (Qu et al., 2022; Lin et al., 2021). Therefore, by restricting the sharing of information among agents within their κ-hop neighborhood, and by choosing κ appropriately, localized RL algorithms can achieve scalability without compromising too much on optimality. Compared to Qu et al. (2022); Lin et al. (2021), our algorithm does not require any information exchange among agents. Furthermore, to rigorously establish the exponential decay property, certain assumptions must be imposed on the underlying MDP model (Qu et al., 2022; Lin et al., 2021). In this work, we do not impose structural assumptions on the underlying model except the one that guarantees exploration (cf. Assumption 3.1), which is commonly adopted in the existing literature. While our results are more general and applicable to networked MARL, their usefulness depends on how small the dependence level is. The local structure in a networked MDP may help reduce the dependence level.

3.5 Proof Sketch

Our proof follows the roadmap depicted in Figure 1. Here, we present a proof sketch of Theorem 3.6, which consists of the following three main steps. The proof of Theorem 3.3 follows a similar approach.

3.5.1 Step One: Convergence of ITD

The main challenge in analyzing ITD (and also IQL) is that, as a stochastic approximation algorithm, the randomness in the algorithm stems from the local sample trajectory $\{(S_k^i, A_k^i)\}$, which does not necessarily form a Markov chain. Therefore, the existing results on Markovian stochastic approximation (Srikant and Ying, 2019; Chen et al., 2022) do not apply here.

To overcome this challenge, and inspired by Lin et al. (2021), we model ITD (and also IQL) as a stochastic approximation algorithm with state aggregation. More specifically, for agent $i \in [n]$, while its algorithm is driven by the local sample trajectory $\{(S_k^i, A_k^i)\}$, it can be alternatively viewed as driven by the global trajectory $\{(S_k, A_k)\}$, but the values $\{(S_k^j, A_k^j)_{j\neq i}\}_{k\geq 0}$ do not affect the algorithm's behavior. Therefore, to analyze the convergence behavior of Agent i's algorithm, for each $\bar{s}^i \in \mathcal{S}^i$, we can aggregate the set of states $\{s \in \mathcal{S} \mid s^i = \bar{s}^i\}$ as one. Using this approach, we show that $Q_{t,K}^i$ approximates the solution to a variant of the projected Bellman equation, denoted by $\tilde{Q}_t^i$. In particular, the following convergence bound holds with high probability: $\|Q_{t,K}^i - \tilde{Q}_t^i\|_\infty \leq \tilde{\mathcal{O}}(1/\sqrt{K})$. To proceed, recall from Section 3.2 that ideally INPG would like to use $\hat{q}_{(t)}^i$ in its update equation. Therefore, we take a further step to show that $\|\tilde{Q}_t^i - \hat{q}_{(t)}^i\|_\infty \leq \mathcal{O}(\mathcal{E})$ using the definition of the dependence level. Altogether, we obtain the convergence bound for ITD:

$$\|Q_{t,K} - \hat{Q}_{(t)}\|_\infty = \left\|\sum_{i=1}^n (Q_{t,K}^i - \hat{q}_{(t)}^i)\right\|_\infty \leq \tilde{\mathcal{O}}(1/\sqrt{K}) + \mathcal{O}(\mathcal{E}),$$

where $Q_{t,k}(s,a) = \sum_{i\in[n]} Q_{t,k}^i(s^i, a^i)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$, and $\hat{Q}_{(t)}$ is the Q-function of the policy $\pi_{(t)}$ on $\hat{\mathcal{M}}$. It is clear that $\hat{Q}_{(t)} = \sum_{i=1}^n \hat{q}_{(t)}^i$ because the underlying MDP $\hat{\mathcal{M}}$ is separable.

3.5.2 Step Two: Global Convergence on the Separable MDP

Following our roadmap described in Figure 1, we analyze INAC as if it were implemented in the separable MDP $\hat{\mathcal{M}}$. Combining the results for ITD and INPG, the following inequality holds with high probability for INAC:

$$\|Q_{(T)} - \hat{Q}_{\hat{\pi}_*}\|_\infty \leq \tilde{\mathcal{O}}(1/\sqrt{K}) + \mathcal{O}(\gamma^T) + \mathcal{O}(\mathcal{E}), \quad (9)$$

where $\hat{\pi}_*$ is an optimal policy of $\hat{\mathcal{M}}$.

3.5.3 Step Three: Bounding the Model Difference Error

With the approximate convergence to the optimal Q-function of $\hat{\mathcal{M}}$, the last step is to bound the gap caused by the model difference in order to achieve approximate global convergence for the original MDP. Using the definition of $\mathcal{E}$, we can show that the difference between the optimal Q-functions of the original MDP $\mathcal{M}$ and the separable MDP $\hat{\mathcal{M}}$ is on the order of $\mathcal{E}$, i.e., $\|Q_{\pi_*} - \hat{Q}_{\hat{\pi}_*}\|_\infty \leq \mathcal{O}(\mathcal{E})$. Combining the above inequality with Eq. (9) completes the proof.

4 NUMERICAL SIMULATIONS

Our last technical section presents numerical experiments for IL. First, we present the results of INAC and IQL applied to the synthetic MDP introduced in Section 2.2 to illustrate the effects of the dependence level. Then, we apply IQL and INAC to an EV charging problem to demonstrate that our algorithms can also be extended to the function approximation setting with approximate global convergence. We simulate the EV charging in a 3-agent system and a 15-agent system and compare the performance of IQL and INAC under different levels of information contraction. The detailed problem formulation and experimental setting are provided in Appendices E.3 and E.4.

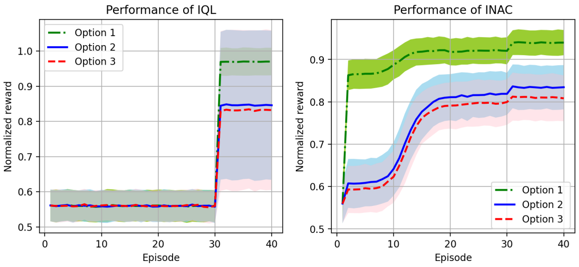

Figure 3: Performance of IQL and INAC for the synthetic MDP.

4.1 Results for Synthetic MDP

We run IQL and INAC on the synthetic MDP 100 times each. The rewards, normalized by the optimal average reward, are shown in Figure 3, where the shaded areas denote the standard deviation. From Figure 3, we see that both IQL and INAC with Option 1 achieve better average rewards and lower variance than Options 2 and 3. This observation corroborates the convergence results in Theorems 3.3 and 3.6, where the optimality gap of IL is controlled by the dependence levels.

Comparing IQL with INAC, we find that IQL achieves a smaller gap than INAC for any option, which is also consistent with our theoretical results. In particular, the asymptotically non-vanishing term (cf. E_2) in IQL is smaller than the corresponding term G_3 in INAC by a factor of $1/(1-\gamma)$.

4.2 Results for EV Charging

We repeated the experiments 20 times in both the 3-agent system and the 15-agent system, respectively, and the results are shown in Figure 4 and Figure 5. The rewards are also normalized by the average optimal reward calculated by the offline optimal policy (see Appendix E.4). The results in the 3-agent system show that IQL and INAC exhibit similar performances. Their performances with different information contractions are very close and much better than the baseline policy, and the performances with full information are slightly better. This phenomenon indicates that we can aggressively contract the information without compromising the optimality. In addition, the performance of IQL is significantly better than that of INAC.

In the 15-agent system, running IQL or INAC with no information contraction is computationally intractable due to the curse of dimensionality. The results show that the performance of INAC with average information and no information both hover around 80% of the optimal rewards, with the former slightly better. The performance of IQL with average information can significantly outperform the others.

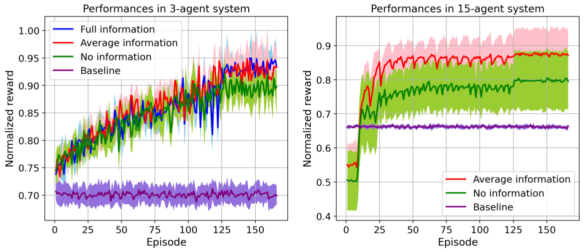

Figure 4: Performance of IQL with different contractions in the 3-agent (left) and 15-agent (right) systems.

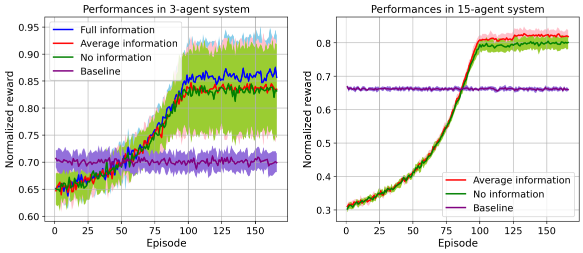

Figure 5: Performance of INAC with different contractions in the 3-agent (left) and 15-agent (right) systems.

5 CONCLUSION

In this paper, we consider IL for MARL in the cooperative setting and establish approximate global convergence for IQL and INAC. Both algorithms have an $\tilde{\mathcal{O}}(\epsilon^{-2})$ sample complexity to find a globally optimal policy, up to an error term that is proportional to the dependence level. Methodologically, we propose a new approach to analyze IL by constructing a separable MDP, where each agent has an independent local state transition model. The model difference between the original MDP and the separable one is captured by the dependence level. Our numerical experiments support these theoretical findings.

There are many interesting directions for future work. First, it is worth investigating whether our proof technique can be applied to other problems beyond IL, such as general non-Markovian stochastic iterative algorithms. Second, it is also promising to develop an efficient approach for estimating the dependence level, especially for MDPs with large-scale state-action spaces. Finally, our analysis suggests that carefully adding coordination and information sharing may reduce the dependence level of the model while preserving the algorithm's scalability. However, theoretically characterizing the trade-off between scalability and optimality remains an open question.

Acknowledgment

Ruiyang Jin and Jie Song were gratefully supported by the National Natural Science Foundation of China under grants T2121002 and 72131001. This work was also funded by NSF under grants CCF-2326609, CNS-2146814, CPS-2136197, CNS-2106403, and NGSDI-2105648. The research of Ruiyang Jin was partially supported by the scholarship "Peking University-BHP Carbon and Climate Wei-Ming PhD Scholars". The research of Yiheng Lin was additionally supported by Amazon AI4Science Fellowship and PIMCO Graduate Fellowship in Data Science.

References

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

Bertsekas, D. and Tsitsiklis, J. N. (1996). *Neuro-dynamic programming*. Athena Scientific.

Bhatnagar, S., Sutton, R. S., Ghavamzadeh, M., and Lee, M. (2009). Natural actor-critic algorithms. *Automatica*, 45(11):2471–2482.

Böhmer, W., Rashid, T., and Whiteson, S. (2019). Exploration with unreliable intrinsic reward in multi-agent reinforcement learning. In *ICML em Exploration in Reinforcement Learning workshop*.

Busoniu, L., Babuska, R., and De Schutter, B. (2008). A comprehensive survey of multiagent reinforcement learning. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, 38(2):156–172.

Cai, Y., Luo, H., Wei, C.-Y., and Zheng, W. (2024). Uncoupled and convergent learning in two-player zero-sum markov games with bandit feedback. *Advances in Neural Information Processing Systems*, 36.

Chen, Z. and Maguluri, S. T. (2022a). An approximate policy iteration viewpoint of actor-critic algorithms. *Preprint arXiv:2208.03247*.

Chen, Z. and Maguluri, S. T. (2022b). Sample complexity of policy-based methods under off-policy sampling and linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 11195–11214. PMLR.

Chen, Z., Zhang, K., Mazumdar, E., Ozdaglar, A., and Wierman, A. (2024). A finite-sample analysis of payoff-based independent learning in zero-sum stochastic games. *Advances in Neural Information Processing Systems*, 36.

Chen, Z., Zhang, S., Doan, T. T., Clarke, J.-P., and Maguluri, S. T. (2022). Finite-sample analysis of nonlinear stochastic approximation with applications in reinforcement learning. *Automatica*, 146:110623.

Chu, T., Chinchali, S., and Katti, S. (2020). Multi-agent reinforcement learning for networked system control. In *International Conference on Learning Representations*.

Daskalakis, C., Foster, D. J., and Golowich, N. (2020). Independent policy gradient methods for competitive reinforcement learning. *Advances in neural information processing systems*, 33:5527–5540.

Degrave, J., Felici, F., Buchli, J., Neunert, M., Tracey, B., Carpanese, F., Ewalds, T., Hafner, R., Abdolmaleki, A., de Las Casas, D., et al. (2022). Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897):414–419.

Ding, D., Wei, C.-Y., Zhang, K., and Jovanovic, M. (2022a). 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.

Ding, L., Lin, Z., Shi, X., and Yan, G. (2022b). Target-value-competition-based multi-agent deep reinforcement learning algorithm for distributed nonconvex economic dispatch. *IEEE Transactions on Power Systems*, 38(1):204–217.

Even-Dar, E., Mansour, Y., and Bartlett, P. (2003). Learning rates for Q-learning. *Journal of Machine Learning Research*, 5(1).

Foruzan, E., Soh, L.-K., and Asgarpoor, S. (2018). Reinforcement learning approach for optimal distributed energy management in a microgrid. *IEEE Transactions on Power Systems*, 33(5):5749–5758.

Gheshlaghi Azar, M., Munos, R., and Kappen, H. J. (2013). Minimax PAC bounds on the sample complexity of reinforcement learning with a generative model. *Machine learning*, 91:325–349.

Gu, S., Holly, E., Lillicrap, T., and Levine, S. (2017). Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates. In *2017 IEEE international conference on robotics and automation (ICRA)*, pages 3389–3396. IEEE.

Jin, R., Zhou, Y., Lu, C., and Song, J. (2022). Deep reinforcement learning-based strategy for charging station participating in demand response. *Applied Energy*, 328:120140.

Jordan, P., Barakat, A., and He, N. (2024). Independent learning in constrained Markov potential games. In *International Conference on Artificial Intelligence and Statistics*, pages 4024–4032. PMLR.

Kakade, S. and Langford, J. (2002). Approximately optimal approximate reinforcement learning. In *Proceedings of the Nineteenth International Conference on Machine Learning*, pages 267–274.

Kakade, S. M. (2001). A natural policy gradient. *Advances in neural information processing systems*, 14.

Khodadadian, S., Chen, Z., and Maguluri, S. T. (2021). Finite-sample analysis of off-policy natural actor-critic algorithm. In *International Conference on Machine Learning*, pages 5420–5431. PMLR.

Khodadadian, S., Doan, T. T., Romberg, J., and Maguluri, S. T. (2022a). Finite sample analysis of two-time-scale natural actor-critic algorithm. *IEEE Transactions on Automatic Control*.

Khodadadian, S., Jhunjhunwala, P. R., Varma, S. M., and Maguluri, S. T. (2022b). On linear and super-linear convergence of natural policy gradient algorithm. *Systems & Control Letters*, 164:105214.

Kiumarsi, B., Lewis, F. L., Modares, H., Karimpour, A., and Naghibi-Sistani, M.-B. (2014). Reinforcement Q-learning for optimal tracking control of linear discrete-time systems with unknown dynamics. *Automatica*, 50(4):1167–1175.

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

Lan, G. (2023). Policy mirror descent for reinforcement learning: Linear convergence, new sampling complexity, and generalized problem classes. *Mathematical programming*, 198(1):1059–1106.

Lanctot, M., Zambaldi, V., Gruslys, A., Lazaridou, A., Tuyls, K., Pérolat, J., Silver, D., and Graepel, T. (2017). A unified game-theoretic approach to multiagent reinforcement learning. *Advances in neural information processing systems*, 30.

Leonardos, S., Overman, W., Panageas, I., and Piliouras, G. (2022). Global convergence of multi-agent policy gradient in Markov potential games. In *International Conference on Learning Representations*.

Levin, D. A. and Peres, Y. (2017). *Markov chains and mixing times*, volume 107. American Mathematical Soc.

Li, G., Cai, C., Chen, Y., Wei, Y., and Chi, Y. (2024). Is Q-learning minimax optimal? A tight sample complexity analysis. *Operations Research*, 72(1):222–236.

Liang, L., Ye, H., and Li, G. Y. (2019). Spectrum sharing in vehicular networks based on multi-agent reinforcement learning. *IEEE Journal on Selected Areas in Communications*, 37(10):2282–2292.

Lin, Y., Qu, G., Huang, L., and Wierman, A. (2021). Multi-agent reinforcement learning in stochastic networked systems. *Advances in neural information processing systems*, 34:7825–7837.

Littman, M. L. (2009). A tutorial on partially observable Markov decision processes. *Journal of Mathematical Psychology*, 53(3):119–125.

Lowe, R., Wu, Y. I., Tamar, A., Harb, J., Pieter Abbeel, O., and Mordatch, I. (2017). Multi-agent actor-critic for mixed cooperative-competitive environments. *Advances in neural information processing systems*, 30.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *nature*, 518(7540):529–533.

Nguyen, K. K., Duong, T. Q., Vien, N. A., Le-Khac, N.-A., and Nguyen, L. D. (2019). Distributed deep deterministic policy gradient for power allocation control in d2d-based v2v communications. *IEEE Access*, 7:164533–164543.

Qu, G., Lin, Y., Wierman, A., and Li, N. (2020). Scalable multi-agent reinforcement learning for networked systems with average reward. *Advances in Neural Information Processing Systems*, 33:2074–2086.

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

Qu, G., Wierman, A., and Li, N. (2022). Scalable reinforcement learning for multiagent networked systems. *Operations Research*, 70(6):3601–3628.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The annals of mathematical statistics*, pages 400–407.

Rudin, W. et al. (1976). *Principles of mathematical analysis*, volume 3. McGraw-hill New York.

Sadeghianpourhamami, N., Deleu, J., and Develder, C. (2019). Definition and evaluation of model-free coordination of electrical vehicle charging with reinforcement learning. *IEEE Transactions on Smart Grid*, 11(1):203–214.

Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang, A., Guez, A., Hubert, T., Baker, L., Lai, M., Bolton, A., et al. (2017). Mastering the game of Go without human knowledge. *Nature*, 550(7676):354.

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

Sutton, R. S. (1988). Learning to predict by the methods of temporal differences. *Machine learning*, 3(1):9–44.

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction.* MIT press.

Tampuu, A., Matiisen, T., Kodelja, D., Kuzovkin, I., Korjus, K., Aru, J., Aru, J., and Vicente, R. (2017). Multiagent cooperation and competition with deep reinforcement learning. *PloS one*, 12(4):e0172395.

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

Tsitsiklis, J. and Van Roy, B. (1996). Analysis of temporal-difference learning with function approximation. *Advances in neural information processing systems*, 9.

Tsitsiklis, J. N. (1994). Asynchronous stochastic approximation and Q-learning. *Machine learning*, 16(3):185–202.

Wang, J., Xu, W., Gu, Y., Song, W., and Green, T. C. (2021). Multi-agent reinforcement learning for active voltage control on power distribution networks. *Advances in Neural Information Processing Systems*, 34:3271–3284.

Watkins, C. J. and Dayan, P. (1992). Q-learning. *Machine learning*, 8:279–292.

Wu, Y. F., Zhang, W., Xu, P., and Gu, Q. (2020). A finite-time analysis of two time-scale actor-critic methods. *Advances in Neural Information Processing Systems*, 33:17617–17628.

Xiang, Y., Lu, Y., and Liu, J. (2023). Deep reinforcement learning based topology-aware voltage regulation of distribution networks with distributed energy storage. *Applied Energy*, 332:120510.

Xiao, L. (2022). On the convergence rates of policy gradient methods. *The Journal of Machine Learning Research*, 23(1):12887–12922.

Yang, Y. and Wang, J. (2020). An overview of multi-agent reinforcement learning from game theoretical perspective. *Preprint arXiv:2011.00583.*

Yardim, B., Cayci, S., Geist, M., and He, N. (2023). Policy mirror ascent for efficient and independent learning in mean field games. In *International Conference on Machine Learning*, pages 39722–39754. PMLR.

Yongacoglu, B., Arslan, G., and Yüksel, S. (2023). Satisficing paths and independent multiagent reinforcement learning in stochastic games. *SIAM Journal on Mathematics of Data Science*, 5(3):745–773.

Yuan, R., Du, S. S., Gower, R. M., Lazaric, A., and Xiao, L. (2022). Linear convergence of natural policy gradient methods with log-linear policies. In *The Eleventh International Conference on Learning Representations.*

Zeng, S., Doan, T. T., and Romberg, J. (2022). Finite-time complexity of online primal-dual natural actor-critic algorithm for constrained Markov decision processes. In *2022 IEEE 61st Conference on Decision and Control (CDC)*, pages 4028–4033. IEEE.

Zhang, K., Yang, Z., and Basar, T. (2018a). Networked multi-agent reinforcement learning in continuous spaces. In *2018 IEEE conference on decision and control (CDC)*, pages 2771–2776. IEEE.

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

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

Zhang, K., Yang, Z., Liu, H., Zhang, T., and Başar, T. (2021b). Finite-sample analysis for decentralized batch multiagent reinforcement learning with networked agents. *IEEE Transactions on Automatic Control*, 66(12):5925–5940.

Zhang, Y., Qu, G., Xu, P., Lin, Y., Chen, Z., and Wierman, A. (2023). Global convergence of localized policy iteration in networked multi-agent reinforcement learning. *Proceedings of the ACM on Measurement and Analysis of Computing Systems*, 7(1):1–51.

Zhong, C., Gursoy, M. C., and Velipasalar, S. (2019). Deep multi-agent reinforcement learning based cooperative edge caching in wireless networks. In *ICC 2019-2019 IEEE International Conference on Communications (ICC)*, pages 1–6. IEEE.

Zhou, Z., Chen, Z., Lin, Y., and Wierman, A. (2023). Convergence rates for localized actor-critic in networked Markov potential games. In Evans, R. J. and Shpitser, I., editors, *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 2563–2573. PMLR.

Zou, S., Xu, T., and Liang, Y. (2019). Finite-sample analysis for SARSA with linear function approximation. *Advances in neural information processing systems*, 32.

Checklist

1. For all models and algorithms presented, check if you include:
   - (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model. [Yes]
   - (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm. [Yes]
   - (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries. [Yes]
2. For any theoretical claim, check if you include:
   - (a) Statements of the full set of assumptions of all theoretical results. [Yes]
   - (b) Complete proofs of all theoretical results. [Yes]
   - (c) Clear explanations of any assumptions. [Yes]
3. For all figures and tables that present empirical results, check if you include:
   - (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL). [Yes]
   - (b) All the training details (e.g., data splits, hyperparameters, how they were chosen). [Yes]
   - (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times). [Yes]
   - (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider). [No. The experiments in our paper mainly aim to verify the theoretical results, and are not computationally complex. Therefore, the needed computer resources to reproduce our results are not very restrictive.]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   - (a) Citations of the creator If your work uses existing assets. [No. We do not use any existing assets.]
   - (b) The license information of the assets, if applicable. [Not Applicable]
   - (c) New assets either in the supplemental material or as a URL, if applicable. [No. We do not release any new assets.]
   - (d) Information about consent from data providers/curators. [Not Applicable]
   - (e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content. [Not Applicable. We focus on the theoretical guarantees of independent learning which we think has no such content.]
5. If you used crowdsourcing or conducted research with human subjects, check if you include:
   - (a) The full text of instructions given to participants and screenshots. [Not Applicable]
   - (b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable. [Not Applicable]
   - (c) The estimated hourly wage paid to participants and the total amount spent on participant compensation. [Not Applicable]

Appendices

A Related Literature

In what follows, we discuss a non-exhaustive overview of related work on Q-learning, NAC, IL, and networked MARL.

Q-Learning. Q-learning was first proposed in Watkins and Dayan (1992). The main idea of Q-learning is to maintain a vector indexed by all state-action pairs and iteratively update the Q-function to improve the policy derived from it. Since its introduction, Q-learning has been widely studied in terms of both asymptotic convergence (Tsitsiklis, 1994) and finite-sample analysis (Even-Dar et al., 2003; Qu and Wierman, 2020; Li et al., 2024). Meanwhile, Q-learning has been successfully applied in various domains, such as games (Mnih et al., 2015) and control problems (Kiumarsi et al., 2014). Although Q-learning has also been applied to multi-agent systems (Tampuu et al., 2017), theoretical results for MARL with Q-learning remain relatively limited.

Natural Actor-Critic. In RL, a popular approach for finding an optimal policy is to implement gradient-based methods directly in the policy space (Sutton and Barto, 2018; Konda and Tsitsiklis, 2000), a typical example of which is NAC (Bhatnagar et al., 2009), where NPG (Kakade, 2001; Xiao, 2022; Yuan et al., 2022) is used for policy improvement and TD-learning (Sutton, 1988) is used for policy evaluation. Finite-sample analysis of NAC (with linear function approximation) has been conducted in Agarwal et al. (2021); Lan (2023); Khodadadian et al. (2022a); Chen and Maguluri (2022b) and the references therein, where the state-of-the-art sample complexity is $\tilde{\mathcal{O}}(\epsilon^{-2})$. However, these studies focus solely on the single-agent setting. In MARL, where each agent performs IL, existing results on single-agent NAC do not directly extend due to multi-agent interactions.

Independent learning in MARL. IL has been widely applied in various domains, such as power systems (Ding et al., 2022b; Jin et al., 2022; Foruzan et al., 2018) and communication networks (Liang et al., 2019; Nguyen et al., 2019; Zhong et al., 2019), among others. While IL is empirically popular, it has been demonstrated in Tan (1993) that IL may fail for tasks requiring coordination among agents. Therefore, understanding when IL can achieve global convergence is crucial (Busoniu et al., 2008; Zhang et al., 2021a). IL has been theoretically justified to some extent in certain multi-agent scenarios, such as zero-sum stochastic games (Chen et al., 2024; Daskalakis et al., 2020), Markov potential games (Ding et al., 2022a; Leonardos et al., 2022), and other structured settings (Yongacoglu et al., 2023; Yardim et al., 2023), where the specific game structures are leveraged to design customized learning dynamics with provable guarantees. However, results for the general setting remain limited. In this work, we establish finite-sample guarantees for IL in the cooperative setting and provide a characterization of the optimality gap based on the dependence level.

Networked MARL. In networked MARL, each agent engages in information exchange only with its neighbors through an interaction network. Such a localized interaction structure is commonly found in applications involving social networks, computer networks, traffic networks, and more (Lin et al., 2021; Zhou et al., 2023; Qu et al., 2022). This structure enables decentralized decision-making based on local observations and shared information from neighboring agents (Chu et al., 2020). Theoretically, Zhang et al. (2018b) incorporates a consensus algorithm into the design of their networked MARL algorithm and provides convergence analysis under linear function approximation. Various algorithms and analyses have been proposed for extended settings, such as continuous spaces (Zhang et al., 2018a) and stochastic networked MARL (Zhang et al., 2021b). A more detailed review of networked MARL can be found in Zhang et al. (2021a).

The works most relevant to IL within networked MARL are the scalable algorithms designed using the exponential decay property (Qu et al., 2022; Lin et al., 2021; Qu et al., 2020; Zhang et al., 2023; Zhou et al., 2023), where the value functions of each agent have exponentially decaying correlations with agents far away from it. Convergence results with respect to the range of neighbors from which each agent collects information have been established. However, there are two key differences between our setting and this line of work: First, we do not impose the assumption of local interaction structure which requires the next local state of an agent to be only affected by the current states of its direct neighbors; Second, our IL approach only requires local information to implement, while the scalable algorithms in Networked MARL require communications between agents that are within a certain range.

B Proofs of Theorem 3.3 and Theorem 3.6

B.1 Reformulation as Stochastic Approximation with State Aggregation

In this section, we model IQL (and ITD) as a special case of stochastic approximation with state aggregation and apply the results from Lin et al. (2021) (see Appendix D) to derive finite-sample bounds on its convergence to the fixed point of an appropriately defined operator.

B.1.1 Independent Q-Learning

Fixing $i \in [n]$, define a surjection $h_1 : \mathcal{S} \to \mathcal{S}^i$ and $h_2 : \mathcal{A} \to \mathcal{A}^i$ as $h_1(s) = s^i$ and $h_2(a) = a^i$ for all $s = (s^1, \dots, s^n) \in \mathcal{S}$ and $a = (a^1, \dots, a^n) \in \mathcal{A}$. Let $\Phi^i \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}^i||\mathcal{A}^i|}$ be binary matrix defined as $\Phi^i((s,a),(\bar{s}^i,\bar{a}^i)) = 1$ if and only if $h_1(s) = \bar{s}^i$ and $h_2(a) = \bar{a}^i$. The update of IQL for agent $i \in [n]$ can be equivalently written as:

$$Q_{k+1}^i(h_1(S_k), h_2(A_k)) = Q_k^i(h_1(S_k), h_2(A_k)) + \alpha_k \left(\left[F_*^i(\Phi^i Q_k^i)\right](S_k, A_k) - Q_k^i\left(h_1(S_k), h_2(A_k)\right) + w_k^i \right), \quad (10)$$

where $F_*^i : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ is an operator defined as

$$[F_*^i(Q)](s,a) = \mathcal{R}^i(s^i, a^i) + \gamma \mathbb{E}_{\bar{s} \sim P_a(s,\cdot)} \left[\max_{\bar{a} \in \mathcal{A}} Q(\bar{s}, \bar{a}) \right], \quad \forall\, (s,a) \in \mathcal{S} \times \mathcal{A}, \quad (11)$$

and the noise sequence w_k^i is defined as

$$w_k^i = \mathcal{R}^i(S_k^i, A_k^i) + \gamma \max_{\bar{a}} Q_k^i(h_1(S_{k+1}), h_2(\bar{a})) - [F_*^i(\Phi^i Q_k^i)](S_k, A_k), \quad \forall\, k \geq 0. \quad (12)$$

Next, we verify in the following lemma that all the assumptions (which are restated in Assumption D.1, Appendix D) needed to apply (Lin et al., 2021, Theorem 3.1) are satisfied in the context of IQL. Define the filtration $\mathcal{F}_k$ as the induced σ-algebra generated by $(S_0, A_0, \cdots, S_k, A_k)$.

Lemma B.1. *Suppose that Assumption 3.1 is satisfied. Then, we have the following results.*

(1) The operator $F_*^i(\cdot)$ is a γ-contraction with respect to $\|\cdot\|_\infty$. Moreover, we have $\|F_*^i(Q)\|_\infty \leq \gamma\|Q\|_\infty + 1$ for any $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$.

(2) It holds that $\|Q_k^i\|_\infty \leq 1/(1-\gamma)$ for all $k \geq 0$.

(3) The stochastic process $\{w_k^i\}$ satisfies (i) w_k^i is measurable with respect to $\mathcal{F}_{k+1}$, (ii) $\mathbb{E}[w_k^i \mid \mathcal{F}_k] = 0$, and (iii) $|w_k^i| \leq 2/(1-\gamma)$ almost surely for all $k \geq 0$.

Remark B.2. Since Assumption D.1 (2) clearly holds under Assumption 3.1, we only verify Assumption D.1 (1), (3), and (4) here.

Proof of Lemma B.1. (1) Using the definition of $F_*^i(\cdot)$, for any $Q_1, Q_2 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ and (s,a), we have

$$\begin{aligned} |[F_*^i(Q_1)](s,a) - [F_*^i(Q_2)](s,a)| &= \gamma \left| \mathbb{E}_{\bar{s} \sim P_a(s,\cdot)} \left[\max_{\bar{a} \in \mathcal{A}} Q_1(\bar{s},\bar{a}) - \max_{\bar{a} \in \mathcal{A}} Q_2(\bar{s},\bar{a}) \right] \right| \\ &\leq \gamma \mathbb{E}_{\bar{s} \sim P_a(s,\cdot)} \left[\left| \max_{\bar{a} \in \mathcal{A}} Q_1(\bar{s},\bar{a}) - \max_{\bar{a} \in \mathcal{A}} Q_2(\bar{s},\bar{a}) \right| \right] \\ &\leq \gamma \|Q_1 - Q_2\|_\infty. \end{aligned}$$

Since the right-hand side of the inequality does not depend on (s,a), we have

$$\|F_*^i(Q_1) - F_*^i(Q_2)\|_\infty = \max_{(s,a)} |[F_*^i(Q_1)](s,a) - [F_*^i(Q_2)](s,a)| \leq \gamma \|Q_1 - Q_2\|_\infty.$$

Moreover, for any $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, using the above inequality with $Q_1 = Q$ and $Q_2 = 0$, we have

$$\|F_*^i(Q)\|_\infty \leq \|F_*^i(Q) - F_*^i(0)\|_\infty + \|F_*^i(0)\|_\infty \leq \gamma\|Q\|_\infty + 1.$$

(2) In view of Eq. (2), if $(s^i, a^i) \neq (S_k^i, A_k^i)$, we have $|Q_{k+1}^i(s^i, a^i)| = |Q_k^i(s^i, a^i)|$. Otherwise if $(s^i, a^i) = (S_k^i, A_k^i)$, we have

$$\begin{aligned}|Q_{k+1}^i(S_k^i, A_k^i)| &\leq (1-\alpha_k)|Q_k^i(S_k^i, A_k^i)| + \alpha_k|\mathcal{R}^i(S_k^i, A_k^i)| + \alpha_k\gamma|\max_{\bar{a}^i} Q_k^i(S_{k+1}^i, \bar{a}^i)| \\ &\leq (1-\alpha_k)\|Q_k^i\|_\infty + \alpha_k + \alpha_k\gamma\|Q_k^i\|_\infty \\ &= (1-(1-\gamma)\alpha_k)\|Q_k^i\|_\infty + \alpha_k.\end{aligned}$$

In summary, the following inequality must hold:

$$\|Q_{k+1}^i\|_\infty \leq \max(\|Q_k^i\|_\infty, (1-(1-\gamma)\alpha_k)\|Q_k^i\|_\infty + \alpha_k).$$

Since $Q_0^i = 0$, it is straightforward to show by induction that $\|Q_k^i\|_\infty \leq 1/(1-\gamma)$ for all $k \geq 0$.

(3) In view of the definition of w_k^i in Eq. (12), since w_k^i is an explicit function of S_k, A_k, S_{k+1}, and Q_k^i (all of which are measurable with respect to $\mathcal{F}_{k+1}$), w_k^i is measurable with respect to $\mathcal{F}_{k+1}$. Moreover, we have for all $k \geq 0$ that

$$\begin{aligned}\mathbb{E}[w_k^i \mid \mathcal{F}_k] &= \mathbb{E}\left[\mathcal{R}^i(S_k^i, A_k^i) + \gamma \max_{\bar{a}} Q_k^i(h_1(S_{k+1}), h_2(\bar{a})) \,\middle|\, \mathcal{F}_k\right] - [F_*^i(\Phi^i Q_k^i)](S_k, A_k) \\ &= \mathbb{E}\left[\mathcal{R}^i(S_k^i, A_k^i) + \gamma \max_{\bar{a}} [\Phi^i Q_k^i](S_{k+1}, \bar{a}) \,\middle|\, \mathcal{F}_k\right] - [F_*^i(\Phi^i Q_k^i)](S_k, A_k)] \\ &= [F_*^i(\Phi^i Q_k^i)](S_k, A_k) - [F_*^i(\Phi^i Q_k^i)](S_k, A_k) \\ &= 0,\end{aligned}$$

and

$$|w_k^i| \leq 2\mathcal{R}^i(S_k^i, A_k^i) + 2\gamma\|Q_k^i\|_\infty \leq 2 + 2\gamma/(1-\gamma) = 2/(1-\gamma),$$

where $\|Q_k^i\|_\infty \leq 1/(1-\gamma)$ follows from Part (2) of this lemma.

□

Now, we are ready to use the results in Lin et al. (2021) to show that IQL converges to a fixed point of an appropriately defined operator. Specifically, let $D_1 := \text{diag}(d_{\pi_b}) \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}$ and let $\Pi_1 := ((\Phi^i)^\top D_1 \Phi^i)^{-1}(\Phi^i)^\top D_1$. It was show in (Lin et al., 2021, Proposition C.1) that the equation

$$\Pi_1 F_*^i(\Phi^i Q) = Q \tag{13}$$

admits a unique solution, which we denote by $\tilde{Q}_*^i$. The following theorem presents the finite-time convergence bounds of IQL to $\tilde{Q}_*^i$.

Theorem B.3. *Consider $\{Q_k\}_{k\geq 0}$ generated from IQL (cf. Algorithm 1). Under Assumption 3.1, suppose that the stepsizes satisfy $\alpha_k = \alpha/(k+k_0)$ with $k_0 = \max(4\alpha, 2M_2 \log K)$ and $\alpha \geq 2/(\sigma'(1-\gamma))$. Then, we have with probability at least $1 - \delta'/n$ that*

$$\left\|Q_K^i - \tilde{Q}_*^i\right\|_\infty \leq \frac{C_a'}{\sqrt{K+k_0}} + \frac{C_b}{K+k_0},$$

where C_a' and C_b are defined the same as in Theorem 3.3.

B.1.2 Independent TD-Learning

Similarly, ITD can also be modeled as a stochastic approximation with state aggregation. Specifically, fixing a policy π, define the operator $F_\pi^i : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \to \mathbb{R}$ as

$$[F_\pi^i(Q)](s,a) = \mathcal{R}^i(s^i, a^i) + \gamma\mathbb{E}_{\bar{s}\sim P_a(s,\cdot),\bar{a}\sim\pi(\cdot|\bar{s})}\left[Q(\bar{s}, \bar{a})\right], \quad \forall\, (s,a) \in \mathcal{S} \times \mathcal{A}. \tag{14}$$

Let w_k^i be the noise sequence defined as

$$w_k^i = \mathcal{R}^i(S_k^i, A_k^i) + \gamma \mathbb{E}_{\bar{a} \sim \pi(\cdot|S_{k+1})}[Q_k^i(h_1(S_{k+1}), h_2(\bar{a}))] - [F_\pi^i(\Phi^i Q_k^i)](S_k, A_k). \tag{15}$$

Then, the update equation for ITD in Eq. (6) can be equivalently written as

$$Q_{k+1}^i(h_1(S_k), h_2(A_k)) = Q_k^i(h_1(S_k), h_2(A_k)) + \alpha_k \left(\left[F_\pi^i(\Phi^i Q_k^i)\right](S_k, A_k) - Q_k^i\left(h_1(S_k), h_2(A_k)\right) + w_k^i \right)$$

with $\pi = \pi_{(t)}$.

The next lemma verifies the assumptions needed to apply (Lin et al., 2021, Theorem 3.1) to ITD. The proof is identical to that of Lemma B.1, and therefore is omitted.

Lemma B.4. *Suppose that Assumption 3.1 is satisfied. Then, we have the following results.*

(1) The operator F_π^i is a γ-contraction in $\|\cdot\|_\infty$. Moreover, we have $\|F_\pi^i(Q)\|_\infty \leq \gamma \|Q\|_\infty + 1$ for any $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$.

(2) It holds that $\|Q_k^i\|_\infty \leq 1/(1-\gamma)$ for all $k \geq 0$.

(3) The stochastic process $\{w_k^i\}$ satisfies (i) w_k^i is measurable with respect to $\mathcal{F}_{k+1}$, (ii) $\mathbb{E}[w_k^i \mid \mathcal{F}_k] = 0$, and (iii) $|w_k^i| \leq 2/(1-\gamma)$ almost surely for all $k \geq 0$.

Remark B.5. Since Assumption D.1 (2) clearly holds under Assumption 3.1, Lemma B.4, we only verify Assumption D.1 (1), (3), and (4).

Let $D_2 := \operatorname{diag}(d_\pi) \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|}$ and let $\Pi_2 := ((\Phi^i)^\top D_2 \Phi^i)^{-1} (\Phi^i)^\top D_2$. It was shown in (Lin et al., 2021, Proposition C.1) that the equation

$$\Pi_2 F_\pi^i(\Phi^i Q) = Q$$

admits a unique solution, which is denoted by $\tilde{Q}^i$. The convergence results ITD is stated below.

Theorem B.6. *Consider $\{Q_k\}_{k \geq 0}$ generated from ITD (cf. Algorithm 3). Under Assumption 3.1, suppose that the stepsizes satisfy $\alpha_k = \alpha/(k+k_0)$ with $k_0 = \max(4\alpha, 2M_2 \log K)$ and $\alpha \geq 2/(\sigma'(1-\gamma))$. Then, we have with probability at least $1 - \delta'/n$ that*

$$\left\| Q_K^i - \tilde{Q}^i \right\|_\infty \leq \frac{C_a'}{\sqrt{K+k_0}} + \frac{C_b}{K+k_0},$$

where C_a' and C_b are defined the same as in Theorem 3.3.

B.2 Proof of Theorem 3.3

Recall that we defined $Q_k \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ with $Q_k(s,a) = \sum_{i \in [n]} Q_k^i(s^i, a^i)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$ and $\hat{Q}_{\hat{\pi}_*}^i \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ as the optimal Q-function of agent i under model $\hat{\mathcal{M}}$. Applying Theorem B.3, we have with probability at least $1 - \delta'/n$ that

$$\left\| Q_K^i - \tilde{Q}_*^i \right\|_\infty \leq \frac{C_a'}{\sqrt{K+k_0}} + \frac{C_b}{K+k_0}, \tag{16}$$

where $\tilde{Q}_*^i$ is the unique fixed point of Eq. (13). To proceed, the following two lemmas are needed.

Lemma B.7 (Proof in Appendix C.1)**.** *It holds that $\|\Phi^i \tilde{Q}_*^i - \hat{Q}_{\hat{\pi}_*}^i\|_\infty \leq 2\gamma\mathcal{E}/(1-\gamma)^2$.*

Lemma B.8 (Proof in Appendix C.2)**.** *Consider the optimal Q-functions Q_{π_*} and $\hat{Q}_{\hat{\pi}_*}$ under the original MDP model $\mathcal{M}$ and the separable one $\hat{\mathcal{M}}$, respectively. Then, we have $\|Q_{\pi_*} - \hat{Q}_{\hat{\pi}_*}\|_\infty \leq 2n\gamma\mathcal{E}/(1-\gamma)^2$.*

Combining Lemma B.7 with Eq. (16), we have by the union bound that with probability at least $1 - \delta'$:

$$\|Q_K - \hat{Q}_{\hat{\pi}_*}\|_\infty = \max_{s,a} \left| \sum_{i \in [n]} Q_K^i(s^i, a^i) - \sum_{i \in [n]} \hat{Q}_{\hat{\pi}_*}^i(s,a) \right|$$

$$
\begin{aligned}
&\leq \max_{s,a} \sum_{i\in[n]} \left| Q_K^i(s^i,a^i) - \hat{Q}^i_{\hat{\pi}_*}(s,a) \right| \\
&\leq \max_{s,a} \sum_{i\in[n]} \left(\left| Q_K^i(s^i,a^i) - \tilde{Q}^i_*(s^i,a^i) \right| + \left| (\Phi^i \tilde{Q}^i_*)(s,a) - \hat{Q}^i_{\hat{\pi}_*}(s,a) \right| \right) && \text{(17a)} \\
&\leq \sum_{i\in[n]} \left(\|Q_K^i - \tilde{Q}^i_*\|_\infty + \|(\Phi^i \tilde{Q}^i_*) - \hat{Q}^i_{\hat{\pi}_*}\|_\infty \right) \\
&\leq \sum_{i\in[n]} \left(\frac{C'_a}{\sqrt{K+k_0}} + \frac{C_b}{K+k_0} + \frac{2\gamma\mathcal{E}}{(1-\gamma)^2} \right) && \text{(17b)} \\
&= \frac{nC'_a}{\sqrt{K+k_0}} + \frac{nC_b}{K+k_0} + \frac{2n\gamma\mathcal{E}}{(1-\gamma)^2},
\end{aligned}
$$

where Eq. (17a) follows from $(\Phi^i \tilde{Q}^i_*)(s,a) = \tilde{Q}^i_*(s^i,a^i)$ (see the proof of Lemma B.7 in Appendix C.1) and Eq. (17b) follows combining Lemma B.7 with Eq. (16). Combining the previous inequality with Lemma B.8, we have with probability at least $1-\delta'$ that

$$
\|Q_K - Q_{\pi_*}\|_\infty \leq \|Q_K - \hat{Q}_{\hat{\pi}_*}\|_\infty + \|Q_{\pi_*} - \hat{Q}_{\hat{\pi}_*}\|_\infty \leq \frac{nC'_a}{\sqrt{K+k_0}} + \frac{nC_b}{K+k_0} + \frac{4n\gamma\mathcal{E}}{(1-\gamma)^2}.
$$

To convert the Q-function gap into the policy gap, we will use the performance difference lemma. Specifically, define the advantage function $A_\pi : \mathcal{S}\times\mathcal{A} \to \mathbb{R}$ as $A_\pi(s,a) = Q_\pi(s,a) - V_\pi(s)$ for policy π and state-action pairs $(s,a) \in \mathcal{S}\times\mathcal{A}$. Then, we have for any $s \in \mathcal{S}$ that

$$
\begin{aligned}
|A_{\pi_*}(s,a_{K,s})| &= |Q_{\pi_*}(s,a_{K,s}) - V_{\pi_*}(s)| && (a_{K,s} \in \arg\max\nolimits_{a\in\mathcal{A}} Q_K(s,a)) \\
&= |Q_{\pi_*}(s,a_{K,s}) - Q_{\pi_*}(s,a_{*,s})| && (a_{*,s} \in \arg\max\nolimits_{a\in\mathcal{A}} Q_{\pi_*}(s,a)) \\
&\leq \frac{2nC'_a}{\sqrt{K+k_0}} + \frac{2nC_b}{K+k_0} + \frac{8n\gamma\mathcal{E}}{(1-\gamma)^2} \\
&\leq |Q_{\pi_*}(s,a_{K,s}) - Q_K(s,a_{K,s})| + |Q_K(s,a_{K,s}) - Q_{\pi_*}(s,a_{*,s})| \\
&= |Q_{\pi_*}(s,a_{K,s}) - Q_K(s,a_{K,s})| + |\max_{a\in\mathcal{A}} Q_K(s,a) - \max_{a\in\mathcal{A}} Q_{\pi_*}(s,a)| \\
&\leq |Q_{\pi_*}(s,a_{K,s}) - Q_K(s,a_{K,s})| + \max_{a\in\mathcal{A}} |Q_K(s,a) - Q_{\pi_*}(s,a)| \\
&\leq 2\|Q_K - Q_{\pi_*}\|_\infty \\
&\leq \frac{2nC'_a}{\sqrt{K+k_0}} + \frac{2nC_b}{K+k_0} + \frac{8n\gamma\mathcal{E}}{(1-\gamma)^2}.
\end{aligned}
$$

Finally, using the performance difference lemma (Kakade and Langford, 2002), with probability at least $1-\delta'$, we have for any initial state distribution μ that

$$
V^\mu_{\pi_*} - V^\mu_{\pi_K} = \frac{1}{1-\gamma}\mathbb{E}_{s\sim o^\mu_{\pi_K}} A_{\pi_*}(s,a_{K,s}) \leq \frac{1}{1-\gamma}\left(\frac{2nC'_a}{\sqrt{K+k_0}} + \frac{2nC_b}{K+k_0} \right) + \frac{8n\gamma\mathcal{E}}{(1-\gamma)^3},
$$

where we recall that π_K is the policy greedily induced by Q_K and $o^\mu_{\pi_K} \in \Delta(\mathcal{S})$ is the discounted state visitation distribution of policy π_K under initial distribution μ.

B.3 Proof of Theorem 3.6

We establish the convergence error of the critic in Appendix B.3.1 and the convergence rate of the actor in Appendix B.3.2, then combine them to complete the proof.

B.3.1 Analysis of the Critic

For simplicity of presentation, we first write down only the inner loop of Algorithm 2 in the following, where we may omit the subscript t. The results we derive for the inner loop can be easily combined with the analysis of the outer loop using the Markov property.

Next, we provide the convergence of the critic in the following theorem.

Algorithm 3 Inner Loop of Algorithm 2

```
Input: Integer $K$, policy $\pi^i := \pi^i_{\theta^i}$ from the outer loop, and initialization $Q^i_0 = 0$.
for $k = 0, 1, 2, \cdots, K-1$ do
    Implement $A^i_k \sim \pi^i(\cdot \mid S^i_k)$ (simultaneously with all other agents), and observes $S^i_{k+1}$.
    $Q^i_{k+1}(S^i_k, A^i_k) = (1-\alpha_k)Q^i_k(S^i_k, A^i_k) + \alpha_k(\mathcal{R}^i(S^i_k, A^i_k) + \gamma\mathbb{E}_{\bar{a}^i \sim \pi^i(\cdot|S^i_{k+1})} Q^i_k(S^i_{k+1}, \bar{a}^i))$
end for
Output: $Q^i_K$
```

Theorem B.9. *Consider $\{Q_k\}_{k\geq 0}$ generated by Algorithm 3. Suppose that Assumption 3.1 is satisfied. Let the stepsizes satisfy $\alpha_k = \alpha/(k+k_0)$ with $k_0 = \max(4\alpha, 2M_2 \log K)$ and $\alpha \geq 2/(\sigma'(1-\gamma))$. Then, with probability at least $1-\delta'$, we have*

$$\|Q_K - \hat{Q}_\pi\|_\infty \leq \frac{nC'_a}{\sqrt{K+k_0}} + \frac{nC_b}{K+k_0} + \frac{2n\gamma\mathcal{E}}{(1-\gamma)^2},$$

where the constants C'_a and C_b are defined in Theorem 3.3.

Proof of Theorem B.9. We have verified that ITD is a special case of stochastic approximation with state aggregation. Therefore, applying Theorem B.6, we have with probability $1-\delta'/n$ that

$$\|Q^i_K - \tilde{Q}^i\|_\infty \leq \frac{C'_a}{\sqrt{K+k_0}} + \frac{C_b}{K+k_0}. \tag{18}$$

To proceed, we need the following lemma.

Lemma B.10 (Proof in Appendix C.3). *It holds that $\|\Phi^i\tilde{Q}^i - \hat{Q}^i_\pi\|_\infty \leq 2\gamma\mathcal{E}/(1-\gamma)^2$.*

Combing Lemma B.10 with Eq. (18), we have by the union bound that with probability at least $1-\delta'$:

$$\begin{aligned}
\|Q_K - \hat{Q}_\pi\|_\infty &= \max_{s,a} \left| \sum_{i\in[n]} Q^i_K(s^i, a^i) - \sum_{i\in[n]} \hat{Q}^i_\pi(s,a) \right| \\
&\leq \max_{s,a} \sum_{i\in[n]} \left| Q^i_K(s^i, a^i) - \hat{Q}^i_\pi(s,a) \right| \\
&\leq \max_{s,a} \sum_{i\in[n]} \left(\left| Q^i_K(s^i, a^i) - \tilde{Q}^i(s^i, a^i) \right| + \left| (\Phi^i\tilde{Q}^i)(s,a) - \hat{Q}^i_\pi(s,a) \right| \right) \qquad (19a) \\
&\leq \sum_{i\in[n]} \left(\left\| Q^i_K(s^i, a^i) - \tilde{Q}^i(s^i, a^i) \right\|_\infty + \left\| (\Phi^i\tilde{Q}^i)(s,a) - \hat{Q}^i_\pi(s,a) \right\|_\infty \right) \\
&\leq \sum_{i\in[n]} \left(\frac{C'_a}{\sqrt{K+k_0}} + \frac{C_b}{K+k_0} + \frac{2\gamma\mathcal{E}}{(1-\gamma)^2} \right) \qquad (19b) \\
&= \frac{nC'_a}{\sqrt{K+k_0}} + \frac{nC_b}{K+k_0} + \frac{2n\gamma\mathcal{E}}{(1-\gamma)^2},
\end{aligned}$$

where Eq. (19a) follows from the fact that $(\Phi^i\tilde{Q}^i)(s,a) = \tilde{Q}^i(s^i, a^i)$; Eq. (19b) follows from combing Lemma B.10 with Eq. (18). □

B.3.2 Analysis of the Actor

For ease of presentation, we write down only the outer loop of Algorithm 2 in Algorithm 4.

The following theorem leverages the result in Chen and Maguluri (2022a) to provide the convergence rate of Algorithm 4. Recall that we denote $\pi_{(t)}$ as the policy with parameter π_{θ_t} and $Q_{(t)}$ as $Q_{\pi_{(t)}}$.

Algorithm 4 Outer Loop of Algorithm 2

```
Input: Integer $T$ and initialization $\theta_0^i = 0$.
for $t = 0, 1, 2 \cdots, T-1$ do
    $\theta_{t+1}^i = \theta_t^i + \eta_t Q_{t,K}^i$, where $Q_{t,K}^i$ is from the last iterate of Algorithm 3.
end for
```

Theorem B.11. *Consider $\{\pi_{(t)} = (\pi_{(t)}^1, \pi_{(t)}^2, \cdots, \pi_{(t)}^n)\}$ generated by Algorithm 4. Suppose that η_t satisfies the condition specified in Theorem 3.6. Then we have*

$$\left\|Q_{\pi_*} - Q_{(t)}\right\|_\infty \leq \frac{4n\gamma^{t-1}}{(1-\gamma)^2} + \frac{2}{1-\gamma}\sum_{j=0}^{t-1}\gamma^{t-j-1}\|Q_{j,K} - \hat{Q}_{(j)}\|_\infty + \frac{4n\gamma\mathcal{E}}{(1-\gamma)^2}.$$

Proof of Theorem B.11. We begin by decomposing the optimality gap according to

$$\|Q_{\pi_*} - Q_{(t)}\|_\infty \leq \underbrace{\|\hat{Q}_{\hat{\pi}_*} - \hat{Q}_{(t)}\|_\infty}_{v_1} + \underbrace{\|Q_{\pi_*} - \hat{Q}_{\hat{\pi}_*}\|_\infty}_{v_2} + \underbrace{\|\hat{Q}_{(t)} - Q_{(t)}\|_\infty}_{v_3}, \tag{20}$$

where we recall that $\hat{\pi}_*$ denotes the optimal policy of the separable MDP $\hat{\mathcal{M}}$.

To bound the term v_1, since $\hat{\mathcal{M}}$ is a separable MDP, when each agent implements INPG as in Algorithm 4, it is equivalent to implementing NPG globally. the update equation for the global policy parameter $\theta_t \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ can be written as

$$\theta_{t+1} = \theta_t + \eta_t Q_{t,K}, \tag{21}$$

where $Q_{t,K}(s,a) = \sum_{i=1}^n Q_{t,K}^i(s^i, a^i)$ for all $s = (s^1, \cdots, s^n) \in \mathcal{S}$ and $a = (a^1, \cdots, a^n) \in \mathcal{A}$. Therefore, we can use existing results on single-agent natural actor-critic to bound the term v_1. In particular, Theorem 2.1 in Chen and Maguluri (2022a) provides us the following result, the proof of which is omitted here.

Lemma B.12. *Consider $\{\pi_{(t)}\}_{t\geq 0}$ generated by each agent implementing Algorithm 4. When the stepsize η_t satisfies $\eta_t \geq \log(\frac{1}{\min_s \pi_{(t)}(a_{t,s}|s)})/\gamma^{2t-1}$, where $a_{t,s} = \arg\max_a Q_{t,K}(s,a)$, we have*

$$\left\|\hat{Q}_{\hat{\pi}_*} - \hat{Q}_{(t)}\right\|_\infty \leq \gamma^t\|\hat{Q}_{\pi_*} - \hat{Q}_{(0)}\|_\infty + \frac{2\gamma}{1-\gamma}\sum_{j=0}^{t-1}\gamma^{t-j-1}\|Q_{j,K} - \hat{Q}_{(j)}\|_\infty + \frac{2\gamma^{t-1}}{(1-\gamma)^2}.$$

To apply Lemma B.12, we need to verify that the stepsizes in Theorem 3.6 satisfy the condition in Lemma B.12. For the ease of readers, we restate our condition for η_t from Theorem 3.6 below:

$$\eta_0 = \gamma\log|\mathcal{A}|, \quad \eta_t \geq \left(\frac{2n}{1-\gamma}\sum_{j=0}^{t-1}\eta_j + \log|\mathcal{A}|\right)/\gamma^{2t-1}, \quad \forall\, t \geq 1. \tag{22}$$

Next, we show that Eq. (22) implies $\eta_t \geq \log(\frac{1}{\min_s \pi_{(t)}(a_{t,s}|s)})/\gamma^{2t-1}$. Using Eq. (21), we have for any $t \geq 0$ that

$$\|\theta_t\|_\infty \leq \sum_{j=0}^{t-1}\eta_j\|Q_{j,K}\|_\infty \leq \bar{Q}\sum_{j=0}^{t-1}\eta_j,$$

where $\bar{Q} := n/(1-\gamma)$. The previous inequality enables us to derive a lower bound for $\pi_{(t)}(a \mid s)$ as follows:

$$\begin{aligned}\pi_{(t)}(a \mid s) &= \frac{\exp\{(\theta_t)_{s,a}\}}{\sum_{a'\in\mathcal{A}}\exp\{(\theta_t)_{s,a'}\}}\\ &= \frac{1}{1 + \sum_{a'\neq a}\exp\{(\theta_t)_{s,a'} - (\theta_t)_{s,a}\}}\end{aligned}$$

$$\begin{aligned}
&\geq \frac{1}{1+\sum_{a'\neq a}\exp\{2\|\theta_t\|_\infty\}}\\
&\geq \frac{1}{1+\sum_{a'\neq a}\exp\{2\bar{Q}\sum_{j=0}^{t-1}\eta_j\}}\\
&\geq \frac{1}{|\mathcal{A}|\exp\{2\bar{Q}\sum_{j=0}^{t-1}\eta_j\}}.
\end{aligned}$$

Since the right-hand side of the bound does not depend on (s,a), we have

$$\min_{s,a}\pi_{(t)}(a\mid s)\geq \frac{1}{|\mathcal{A}|\exp\{2\bar{Q}\sum_{j=0}^{t-1}\eta_j\}}=\frac{1}{|\mathcal{A}|\exp\{\frac{2n}{1-\gamma}\sum_{j=0}^{t-1}\eta_j\}}$$

It follows that

$$\log\left(\frac{1}{\min_s\pi_{(t)}(a_{t,s}\mid s)}\right)\leq \frac{2n}{1-\gamma}\sum_{j=0}^{t-1}\eta_j+\log|\mathcal{A}|.$$

Therefore, as long as

$$\eta_t\geq\left(\frac{2n}{1-\gamma}\sum_{j=0}^{t-1}\eta_j+\log|\mathcal{A}|\right)/\gamma^{2t-1},$$

which is the condition stated in Eq. (22), we have $\eta_t\geq\log\left(\frac{1}{\min_s\pi_{(t)}(a_{t,s}|s)}\right)/\gamma^{2t-1}$.

Now that the condition on the stepsizes in Lemma B.12 is satisfied, using Lemma B.12 for the term v_1 on the right-hand side of Eq. (20), we have

$$\begin{aligned}
v_1&=\|\hat{Q}_{\hat{\pi}_*}-\hat{Q}_{(t)}\|_\infty\\
&\leq\gamma^t\|\hat{Q}_{\pi_*}-\hat{Q}_{(0)}\|_\infty+\frac{2\gamma}{1-\gamma}\sum_{j=0}^{t-1}\gamma^{t-j-1}\|Q_{j,K}-\hat{Q}_{(j)}\|_\infty+\frac{2\gamma^{t-1}}{(1-\gamma)^2}\\
&\leq\frac{4n\gamma^{t-1}}{(1-\gamma)^2}+\frac{2}{1-\gamma}\sum_{j=0}^{t-1}\gamma^{t-j-1}\|Q_{j,K}-\hat{Q}_{(j)}\|_\infty,
\end{aligned}$$

where the last inequality follows from:

$$\|\hat{Q}_{\pi_*}-\hat{Q}_{(0)}\|_\infty\leq\|\hat{Q}_{\pi_*}\|_\infty+\|\hat{Q}_{(0)}\|_\infty\leq\frac{2n}{1-\gamma}.$$

We next consider the terms v_2 and v_3 on the right-hand side of Eq. (20). Using Lemma B.8, we bound the term v_2 in Eq. (20) as

$$v_2=\|Q_{\pi_*}-\hat{Q}_{\hat{\pi}_*}\|_\infty\leq\frac{2n\gamma\mathcal{E}}{(1-\gamma)^2}.$$

To bound v_3, we also need the following lemma, which bounds the gap of Q-functions of the same policy applied to the original and separable MDPs.

Lemma B.13 (Proof in Appendix C.4)**.** *Given a policy π, let Q_π and $\hat{Q}_\pi$ be the Q-functions of the original MDP $\mathcal{M}$ and the separable MDP $\hat{\mathcal{M}}$. Then, we have*

$$\|Q_\pi-\hat{Q}_\pi\|_\infty\leq\frac{2n\gamma\mathcal{E}}{(1-\gamma)^2}.$$

The term v_3 can be bounded using Lemma B.13 as

$$v_3=\|Q_{(t)}-\hat{Q}_{(t)}\|_\infty\leq\frac{2n\gamma\mathcal{E}}{(1-\gamma)^2}.$$

Theorem B.11 then follows from using the upper bounds we derived for the terms v_1, v_2, and v_3 in Eq. (20). □

B.3.3 Combining the Actor and the Critic

To finish proving Theorem 3.6, we combine the bounds in Theorem B.9 and Theorem B.11. Specifically, for any $j \leq t-1$, we have with probability at least $1-\delta'$ that

$$\|Q_{j,K} - \hat{Q}_{(j)}\|_\infty \leq \frac{nC_a'}{\sqrt{K+k_0}} + \frac{nC_b}{K+k_0} + \frac{2n\gamma\mathcal{E}}{(1-\gamma)^2}.$$

Let $\delta' = \delta/T$ in the above result. Using union bound, we have with probability at least $1-\delta$ that for all $j \leq t-1$:

$$\|Q_{j,K} - \hat{Q}_{(j)}\|_\infty \leq \frac{nC_a}{\sqrt{K+k_0}} + \frac{nC_b}{K+k_0} + \frac{2n\gamma\mathcal{E}}{(1-\gamma)^2}.$$

Therefore, with probability at least $1-\delta$, we have

$$\begin{aligned}\sum_{j=0}^{T-1}\gamma^{T-j-1}\|Q_{j,K} - \hat{Q}_{\{j\}}\|_\infty &\leq \sum_{j=0}^{T-1}\gamma^{T-j-1}\left(\frac{nC_a}{\sqrt{K+k_0}} + \frac{nC_b}{K+k_0} + \frac{2n\gamma\mathcal{E}}{(1-\gamma)^2}\right)\\ &\leq \frac{1}{1-\gamma}\left(\frac{nC_a}{\sqrt{K+k_0}} + \frac{nC_b}{K+k_0} + \frac{2n\gamma\mathcal{E}}{(1-\gamma)^2}\right).\end{aligned}$$

Applying the bound in the above inequality to Theorem B.11 completes the proof of Theorem 3.6.

B.4 Extensions

In this work, our analysis of IL follows the following blueprint: (i) constructing a separable MDP to approximate the original MDP, (ii) analyzing the algorithm as if it were implemented on the separable MDP, and (iii) bounding the gap due to the model difference between the separable MDP and the original one. This framework can be applied to other learning scenarios beyond IL, as considered in this paper. Below, we briefly elaborate on some of them.

Partially Observable Markov Decision Processes (POMDPs). In sequential decision-making problems, the agent (or agents) may not have full state observability and must rely on partial observations (Littman, 2009). More specifically, instead of observing a trajectory of state-action pairs (S_k, A_k), the agent observes (O_k, A_k), where $O_k \in \mathcal{O}$ represents the partial observation at time k. Unlike (S_k, A_k), the trajectory (O_k, A_k) does *not* form a Markov chain, posing significant technical challenges in POMDP analysis.

Within our analysis framework, if we can approximate the non-Markovian trajectory (O_k, A_k) with a Markov chain defined on $\mathcal{O} \times \mathcal{A}$, this would enable a more tractable analysis of POMDPs. In this case, the approximation error—analogous to our dependence level—would appear in the analysis.

General Non-Markovian Stochastic Approximation Algorithms. RL algorithms broadly fall under the framework of stochastic iterative algorithms, also known as stochastic approximation algorithms (Robbins and Monro, 1951), which take the general form:

$$x_{k+1} = x_k + \alpha_k G(x_k, Y_k),$$

where Y_k represents noise, and $G(\cdot,\cdot)$ is a properly defined operator depending on the problem of interest. As seen in Appendix B.3.1, the update equation for the joint Q-function can be written in this stochastic approximation form. Moreover, the widely used stochastic gradient descent algorithm in large-scale continuous optimization is a special case, where $G(\cdot,\cdot)$ corresponds to a noisy version of the negative gradient operator of the objective function.

While stochastic approximation has been extensively studied, most existing results assume that the noise sequence $\{Y_k\}$ is either i.i.d. or forms a Markov chain. Leveraging our proof framework, we can analyze stochastic approximation algorithms driven by non-Markovian samples, provided that the trajectory $\{Y_k\}$ can be approximated by a Markov chain.

C Proof of All Supporting Lemmas

C.1 Proof of Lemma B.7

According to the definition of Π_1, $\Phi^i\Pi_1$ is the projection matrix that projects a vector in $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ to the set $\{\Phi^i r \mid r \in \mathbb{R}^{|\mathcal{S}^i||\mathcal{A}^i|}\}$. Mathematically, this means that for any $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, we have

$$\Phi^i\Pi_1 Q = \arg\min_{q\in\{\Phi^i r | r\in\mathbb{R}^{|\mathcal{S}^i||\mathcal{A}^i|}\}} \|Q - q\|_{D_1}, \tag{23}$$

where $\|\cdot\|_{D_1}$ denotes the weighted ℓ_2-norm with weights being the stationary distribution d_{π_b} of the Markov chain $\{(S_k, A_k)\}$ induced by π_b. The following lemma presents a non-expansive property of the projection operator $\Phi^i\Pi_1$.

Lemma C.1. *For any $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, we have $\|\Phi^i\Pi_1 Q\|_\infty \leq \|Q\|_\infty$.*

Proof of Lemma C.1. From Eq. (23), we have

$$\Pi_1 Q = \arg\min_{r\in\mathbb{R}^{|\mathcal{S}^i||\mathcal{A}^i|}} \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} d_{\pi_b}(s,a)\left(Q(s,a) - r(h_1(s), h_2(a))\right)^2.$$

In view of the above optimization problem, for any $(\bar{s}, \bar{a})$ in $\mathcal{S}\times\mathcal{A}$, we must have

$$\min_{s\in h_1^{-1}(\bar{s}^i), a\in h_2^{-1}(\bar{a}^i)} Q(s,a) \leq [\Pi_1 Q](\bar{s}^i, \bar{a}^i) \leq \max_{s\in h_1^{-1}(\bar{s}^i), a\in h_2^{-1}(\bar{a}^i)} Q(s,a),$$

which implies

$$\left|[\Phi^i\Pi_1 Q](\bar{s}, \bar{a})\right| = \left|[\Pi_1 Q](\bar{s}^i, \bar{a}^i)\right| \leq \max_{s\in h_1^{-1}(\bar{s}^i), a\in h_2^{-1}(\bar{a}^i)} |Q(s,a)| \leq \|Q\|_\infty.$$

Therefore, we have $\|\Phi^i\Pi_1 Q\|_\infty \leq \|Q\|_\infty$. □

Recall that $\hat{Q}^i_\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ is defined as

$$\hat{Q}^i_\pi(s,a) = \hat{\mathbb{E}}_\pi\left[\sum_{k=0}^\infty \gamma^k \mathcal{R}^i(S^i_k, A^i_k)\,\middle|\, S_0 = s, A_0 = a\right], \quad \forall\,(s,a)\in\mathcal{S}\times\mathcal{A},$$

where $\hat{\mathbb{E}}_\pi[\,\cdot\,]$ denotes the expectation with respect to the separable MDP transition kernel $\hat{\mathcal{P}}$. Due to $\hat{\mathcal{P}}$ being separable, we have

$$\hat{Q}^i_\pi(s,a) = \hat{\mathbb{E}}_\pi\left[\sum_{k=0}^\infty \gamma^k \mathcal{R}^i(S^i_k, A^i_k)\,\middle|\, S_0 = s, A_0 = a\right] = \hat{\mathbb{E}}_\pi\left[\sum_{k=0}^\infty \gamma^k \mathcal{R}^i(S^i_k, A^i_k)\,\middle|\, S^i_0 = s^i, A_0 = a^i\right],$$

which implies $\hat{Q}^i_\pi(s_1, a_1) = \hat{Q}^i_\pi(s_2, a_2)$ for any $(s_1, a_1), (s_2, a_2) \in \mathcal{S}\times\mathcal{A}$ satisfying $(s^i_1, a^i_1) = (s^i_2, a^i_2)$. As a result, we have $\hat{Q}^i_\pi \in \{\Phi^i r \mid r \in \mathbb{R}^{|\mathcal{S}^i||\mathcal{A}^i|}\}$, implying $\Phi^i\Pi_1\hat{Q}^i_\pi = \hat{Q}^i_\pi$ (cf. Lemma C.1).

Next, we bound the term $\|\Phi^i\tilde{Q}^i_* - \hat{Q}^i_{\hat{\pi}_*}\|_\infty$ as follows:

$$\begin{aligned}
\|\Phi^i\tilde{Q}^i_* - \hat{Q}^i_{\hat{\pi}_*}\|_\infty &= \|\Phi^i\Pi_1 F^i_*(\Phi^i\tilde{Q}^i_*) - \Phi^i\Pi_1\hat{Q}^i_{\hat{\pi}_*}\|_\infty && (\Pi_1 F^i_*(\Phi^i\tilde{Q}^i_*) = \tilde{Q}^i_*)\\
&\leq \|F^i_*(\Phi^i\tilde{Q}^i_*) - \hat{Q}^i_{\hat{\pi}_*}\|_\infty && (\Phi^i\Pi_1 \text{ is nonexpansive})\\
&\leq \|F^i_*(\Phi^i\tilde{Q}^i_*) - F^i_*(\hat{Q}^i_{\hat{\pi}_*})\|_\infty + \|F^i_*(\hat{Q}^i_{\hat{\pi}_*}) - \hat{Q}^i_{\hat{\pi}_*}\|_\infty && (\text{Triangle inequality})\\
&\leq \gamma\|\Phi^i\tilde{Q}^i_* - \hat{Q}^i_{\hat{\pi}_*}\|_\infty + \|F^i_*(\hat{Q}^i_{\hat{\pi}_*}) - \hat{Q}^i_{\hat{\pi}_*}\|_\infty. && (F^i_* \text{ is a } \gamma\text{-contraction})
\end{aligned}$$

Rearranging terms, we have

$$\|\Phi^i\tilde{Q}^i_* - \hat{Q}^i_{\hat{\pi}_*}\|_\infty \leq \frac{1}{1-\gamma}\|F^i_*(\hat{Q}^i_{\hat{\pi}_*}) - \hat{Q}^i_{\hat{\pi}_*}\|_\infty.$$

To proceed, observe that

$$\begin{aligned}\|F_*^i(\hat{Q}_{\hat{\pi}_*}^i) - \hat{Q}_{\hat{\pi}_*}^i\|_\infty &= \|F_*^i(\hat{Q}_{\hat{\pi}_*}^i) - \hat{F}_*^i(\hat{Q}_{\hat{\pi}_*}^i)\|_\infty \qquad (\hat{F}_*^i \text{ is the counterpart of } F_*^i \text{ for the model } \hat{\mathcal{M}})\\ &= \max_{(s,a)} \left|\gamma \mathbb{E}_{\bar{s}\sim P_a(s,\cdot)} \max_{\bar{a}} \hat{Q}_{\hat{\pi}_*}^i(\bar{s},\bar{a}) - \gamma \mathbb{E}_{\bar{s}\sim \hat{P}_a(s,\cdot)} \max_{\bar{a}} \hat{Q}_{\hat{\pi}_*}^i(\bar{s},\bar{a})\right|\\ &\leq \gamma \|Q\|_\infty \max_{(s,a)} \sum_{\bar{s}\in\mathcal{S}} \left|P_a(s,\bar{s}) - \hat{P}_a(s,\bar{s})\right|\\ &\leq \frac{2\gamma}{(1-\gamma)} \max_{(s,a)} \left\|P_a(s,\cdot) - \hat{P}_a(s,\cdot)\right\|_{\mathrm{TV}}\\ &\leq \frac{2\gamma\mathcal{E}}{1-\gamma},\end{aligned}$$

where the last two inequalities follows from $\|\hat{Q}_{\hat{\pi}_*}^i\|_\infty \leq 1/(1-\gamma)$ and the definition of the dependence level $\mathcal{E}$. Finally, we have

$$\|\Phi^i \tilde{Q}_*^i - \hat{Q}_{\hat{\pi}_*}^i\|_\infty \leq \frac{1}{1-\gamma}\|F_*^i(\hat{Q}_{\hat{\pi}_*}^i) - \hat{Q}_{\hat{\pi}_*}^i\|_\infty \leq \frac{2\gamma\mathcal{E}}{(1-\gamma)^2}.$$

C.2 Proof of Lemma B.8

Let $\mathcal{T}_* : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ be the Bellman optimality operator of the MDP model $\mathcal{M}$ defined as

$$[\mathcal{T}_* Q](s,a) = \mathcal{R}(s,a) + \gamma \mathbb{E}_{s'\sim P_a(s,\cdot)}\left[\max_{a'} Q(s',a')\right], \quad \forall\ (s,a).$$

The Bellman optimality operator $\hat{\mathcal{T}}_*$ for the separable MDP model $\hat{\mathcal{M}}$ is defined similarly.

Recall that π_* and $\hat{\pi}_*$ are the optimal policies for model $\mathcal{M}$ and $\hat{\mathcal{M}}$ respectively. Then, we have

$$\begin{aligned}\|Q_{\pi_*} - \hat{Q}_{\hat{\pi}_*}\|_\infty &\leq \|Q_{\pi_*} - \mathcal{T}_*\hat{Q}_{\hat{\pi}_*}\|_\infty + \|\mathcal{T}_*\hat{Q}_{\hat{\pi}_*} - \hat{Q}_{\hat{\pi}_*}\|_\infty\\ &\leq \gamma\|Q_{\pi_*} - \hat{Q}_{\hat{\pi}_*}\|_\infty + \|\mathcal{T}_*\hat{Q}_{\hat{\pi}_*} - \hat{Q}_{\hat{\pi}_*}\|_\infty,\end{aligned}$$

where last line follows from $\mathcal{T}_* Q_{\pi^*} = Q_{\pi^*}$ and the Bellman optimality operator $\mathcal{T}_*$ being a γ-contraction mapping with respect to $\|\cdot\|_\infty$. It follows that

$$\left\|Q_{\pi_*} - \hat{Q}_{\hat{\pi}_*}\right\|_\infty \leq \frac{1}{1-\gamma}\left\|\mathcal{T}_*\hat{Q}_{\hat{\pi}_*} - \hat{Q}_{\hat{\pi}_*}\right\|_\infty. \tag{24}$$

To proceed, for all $(s,a) \in \mathcal{S}\times\mathcal{A}$, we have

$$\begin{aligned}&\left|\hat{Q}_{\hat{\pi}_*}(s,a) - (\mathcal{T}_*\hat{Q}_{\hat{\pi}_*})(s,a)\right|\\ =&\left|\gamma\sum_{s'\in\mathcal{S}}\hat{P}_a(s,s')\max_{a'}\hat{Q}_{\hat{\pi}_*}(s',a') - \gamma\sum_{s'\in\mathcal{S}}P_a(s,s')\max_{a'}\hat{Q}_{\hat{\pi}_*}(s',a')\right|\\ \leq&\gamma\|\hat{Q}_{\hat{\pi}_*}\|_\infty\sum_{s'\in\mathcal{S}}\left|\hat{P}_a(s,s') - P_a(s,s')\right|\\ \leq&\frac{2\gamma n}{1-\gamma}\left\|\hat{P}_a(s,\cdot) - P_a(s,\cdot)\right\|_{\mathrm{TV}} &\text{(25a)}\\ \leq&\frac{2n\gamma\mathcal{E}}{1-\gamma}, &\text{(25b)}\end{aligned}$$

where Eq. (25a) follows from $\|\hat{Q}_{\hat{\pi}_*}\|_\infty \leq n/(1-\gamma)$ and Eq.(25b) follows from the definition of $\mathcal{E}$. Substituting the above result into Eq. (24), we have

$$\left\|Q_{\pi_*} - \hat{Q}_{\hat{\pi}_*}\right\|_\infty \leq \frac{2n\gamma\mathcal{E}}{(1-\gamma)^2}.$$

C.3 Proof of Lemma B.10

According to the definition of Π_2, $\Phi^i\Pi_2$ is the projection matrix that projects a vector in $\mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ to the set $\{\Phi^i r \mid r \in \mathbb{R}^{|\mathcal{S}^i||\mathcal{A}^i|}\}$ with respect to the weighted ℓ_2-norm $\|\cdot\|_{D_2}$, where $D_2 = \text{diag}(d_\pi)$. Mathematically, for any $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, we have

$$\Phi^i\Pi_2 Q = \arg\min_{\bar{Q}\in\{\Phi^i r | r\in\mathbb{R}^{|\mathcal{S}^i||\mathcal{A}^i|}\}} \|Q - \bar{Q}\|_{D_2}. \tag{26}$$

In the next lemma, we show that $\Phi^i\Pi_2$ is non-expansive with respect to $\|\cdot\|_\infty$.

Lemma C.2. *For any $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, we have $\|\Phi^i\Pi_2 Q\|_\infty \leq \|Q\|_\infty$.*

The proof of Lemma C.2 follows similarly to that of Lemma C.1. Thus, we omit it here.

For any policy π, since $\hat{Q}^i_\pi(s_1, a_1) = \hat{Q}^i_\pi(s_2, a_2)$ for any $(s_1, a_1), (s_2, a_2) \in \mathcal{S} \times \mathcal{A}$ satisfying $(s^i_1, a^i_1) = (s^i_2, a^i_2)$, we have $\Phi^i\Pi_2\hat{Q}^i_\pi = \hat{Q}^i_\pi$. Therefore, we have

$$\begin{aligned}
\|\Phi^i\tilde{Q}^i - \hat{Q}^i_\pi\|_\infty &= \|\Phi^i\Pi_2 F^i_\pi(\Phi^i\tilde{Q}^i) - \Phi^i\Pi_2\hat{Q}^i_\pi\|_\infty && \text{(27a)}\\
&\leq \|F^i_\pi(\Phi^i\tilde{Q}^i) - \hat{Q}^i_\pi\|_\infty && \text{(27b)}\\
&\leq \|F^i_\pi(\Phi^i\tilde{Q}^i) - F^i_\pi(\hat{Q}^i_\pi)\|_\infty + \|F^i_\pi(\hat{Q}^i_\pi) - \hat{Q}^i_\pi\|_\infty\\
&\leq \gamma\|\Phi^i\tilde{Q}^i - \hat{Q}^i_\pi\|_\infty + \|F^i_\pi(\hat{Q}^i_\pi) - \hat{Q}^i_\pi\|_\infty, && \text{(27c)}
\end{aligned}$$

where Eq. (27a) follows from $\Phi^i\Pi_2 F^i_\pi(\Phi^i\tilde{Q}^i) = \tilde{Q}^i$; Eq. (27b) follows from Lemma C.2; Eq. (27c) follows from $F^i_\pi(\cdot)$ being a γ-contraction mapping with respect to $\|\cdot\|_\infty$. Rearranging terms, we have

$$\|\Phi^i\tilde{Q}^i - \hat{Q}^i_\pi\|_\infty \leq \frac{1}{1-\gamma}\|F^i_\pi(\hat{Q}^i_\pi) - \hat{Q}^i_\pi\|_\infty.$$

To proceed, observe that

$$\begin{aligned}
\|F^i_\pi(\hat{Q}^i_\pi) - \hat{Q}^i_\pi\|_\infty &= \|F^i_\pi(\hat{Q}^i_\pi) - \hat{F}^i_\pi(\hat{Q}^i_\pi)\|_\infty \qquad (\hat{F}^i_\pi \text{ is the counterpart of } F^i_\pi \text{ for the model } \hat{\mathcal{M}})\\
&= \max_{(s,a)}\left|\gamma\mathbb{E}_{\bar{s}\sim P_a(s,\cdot),\bar{a}\sim\pi(\cdot|\bar{s})}[\hat{Q}^i_\pi(\bar{s},\bar{a})] - \gamma\mathbb{E}_{\bar{s}\sim\hat{P}_a(s,\cdot),\bar{a}\sim\pi(\cdot|\bar{s})}[\hat{Q}^i_\pi(\bar{s},\bar{a})]\right|\\
&\leq 2\gamma\|\hat{Q}^i_\pi\|_\infty\max_{(s,a)}\left\|P_a(s,\cdot) - \hat{P}_a(s,\cdot)\right\|_{\text{TV}}\\
&\leq \frac{2\gamma\mathcal{E}}{1-\gamma},
\end{aligned}$$

where the last inequality follows from $\|\hat{Q}^i_\pi\|_\infty \leq 1/(1-\gamma)$ and the definition of the dependence level $\mathcal{E}$. Finally, we obtain

$$\|\Phi^i\tilde{Q}^i - \hat{Q}^i_\pi\|_\infty \leq \frac{1}{1-\gamma}\|F^i_\pi(\hat{Q}^i_\pi) - \hat{Q}^i_\pi\|_\infty \leq \frac{2\gamma\mathcal{E}}{(1-\gamma)^2}.$$

C.4 Proof of Lemma B.13

Let $\mathcal{T}_\pi : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ be the Bellman operator associated with policy π, which is defined as

$$[\mathcal{T}_\pi Q](s,a) = \mathcal{R}(s,a) + \gamma\mathbb{E}_{s'\sim P_a(s,\cdot),a'\sim\pi(\cdot|s')}[Q(s',a')], \quad \forall\,(s,a) \in \mathcal{S}\times\mathcal{A}.$$

Similarly, we define $\hat{\mathcal{T}}_\pi(\cdot)$ as the Bellman operator associated with the policy π under the separable MDP model $\hat{\mathcal{M}}$. It is well known that both $\mathcal{T}_\pi(\cdot)$ and $\hat{\mathcal{T}}_\pi(\cdot)$ are contractive operators with respect to $\|\cdot\|_\infty$, with a common contraction factor γ (Sutton and Barto, 2018). In addition, Q_π (respectively, $\hat{Q}_\pi$) is the unique fixed point of $\mathcal{T}_\pi$ (respectively, $\hat{\mathcal{T}}_\pi$).

To proceed, note that the gap between the Q-functions Q_π and $\hat{Q}_\pi$ can be bounded as

$$\|Q_\pi - \hat{Q}_\pi\|_\infty = \|Q_\pi - \mathcal{T}_\pi\hat{Q}_\pi + \mathcal{T}_\pi\hat{Q}_\pi - \hat{Q}_\pi\|_\infty$$

$$
\begin{aligned}
&\leq \|Q_\pi - \mathcal{T}_\pi \hat{Q}_\pi\|_\infty + \|\mathcal{T}_\pi \hat{Q}_\pi - \hat{Q}_\pi\|_\infty && \text{(Triangle inequality)}\\
&= \|\mathcal{T}_\pi Q_\pi - \mathcal{T}_\pi \hat{Q}_\pi\|_\infty + \|\mathcal{T}_\pi \hat{Q}_\pi - \hat{Q}_\pi\|_\infty && \text{(This follows from } \mathcal{T}_\pi Q_\pi = Q_\pi.\text{)}\\
&\leq \gamma \|Q_\pi - \hat{Q}_\pi\|_\infty + \|\mathcal{T}_\pi \hat{Q}_\pi - \hat{Q}_\pi\|_\infty,
\end{aligned}
$$

where the last inequality follows from that the Bellman operator $\mathcal{T}_\pi$ is a γ-contraction mapping with respect to $\|\cdot\|_\infty$. It follows that

$$
\begin{aligned}
\|Q_\pi - \hat{Q}_\pi\|_\infty &\leq \frac{1}{1-\gamma}\|\mathcal{T}_\pi \hat{Q}_\pi - \hat{Q}_\pi\|_\infty\\
&= \frac{1}{1-\gamma}\|\mathcal{T}_\pi \hat{Q}_\pi - \hat{\mathcal{T}}_\pi \hat{Q}_\pi\|_\infty\\
&= \frac{\gamma}{1-\gamma}\max_{(s,a)}\left|\sum_{s'}(P_a(s,s') - \hat{P}_a(s,s'))\mathbb{E}_{a'\sim\pi(\cdot|s')}[\hat{Q}_\pi(s',a')]\right| && \text{(28a)}\\
&\leq \frac{\gamma\|\hat{Q}_\pi\|_\infty}{1-\gamma}\max_{(s,a)}\sum_{s'}\left|P_a(s,s') - \hat{P}_a(s,s')\right|\\
&\leq \frac{2n\gamma}{(1-\gamma)^2}\max_{(s,a)}\left\|\hat{P}_a(s,\cdot) - P_a(s,\cdot)\right\|_{\text{TV}} && \text{(28b)}\\
&\leq \frac{2n\gamma\mathcal{E}}{(1-\gamma)^2}, && \text{(28c)}
\end{aligned}
$$

where Eq. (28a) follows from the definitions of $\mathcal{T}_\pi$ and $\hat{\mathcal{T}}_\pi$, Eq. (28b) follows from $\|\hat{Q}_\pi\|_\infty \leq n/(1-\gamma)$ and the definition of the total variation distance, and Eq. (28c) follows from the definition of the dependence level $\mathcal{E}$.

D Stochastic Approximation with State Aggregation

In this section, we present the results of stochastic approximation with state aggregation in Lin et al. (2021).

Let $\{i_k\}_{k\geq 0}$ be a Markov chain with a finite state space $\mathcal{N} = \{1, 2, \cdots, n\}$. Let $\mathcal{L} = \{1, 2, \cdots, l\}$ (where $l \leq n$) be the abstract state space. The surjection $h : \mathcal{N} \to \mathcal{L}$ is used to convert every state in $\mathcal{N}$ to its abstraction in $\mathcal{L}$. Given parameter $x \in \mathbb{R}^l$ and function $F : \mathbb{R}^n \to \mathbb{R}^l$, we consider the stochastic approximation that updates $x(k) \in \mathbb{R}^l$ according to

$$
x_j(k+1) = \begin{cases} x_j(k) + \alpha_k \left(F_{i_k}(\Phi x(k)) - x_j(k) + w(k)\right), & j = h(i_k),\\ x_j(k), & j \neq h(i_k), \end{cases}
$$

where $x(0) = 0$, $F : \mathbb{R}^n \to \mathbb{R}^n$ is an operator and $\Phi \in \mathbb{R}^{n\times l}$ is a binary matrix such that $\Phi_{ij} = 1$ if and only if $h(i) = j$. The following assumption is imposed in Lin et al. (2021) to study the convergence behavior of the stochastic approximation algorithm presented above.

Assumption D.1. The following statements hold.

(1) The operator $F(\cdot)$ is a γ-contraction mapping with respect to $\|\cdot\|_\infty$.

(2) The Markov chain $\{i_k\}_{k\geq 0}$ is irreducible and aperiodic with stationary distribution $d \in \Delta(\mathcal{N})$. Moreover, letting $d'_j = \sum_{i\in h^{-1}(j)} d_i$ for all $j \in \mathcal{L}$ and $\sigma' = \inf_{j\in\mathcal{L}} d'_j$, there exist $M_1 \geq 0$ and $M_2 \geq 1$ such that

$$
\max_{\mathcal{K}\subseteq\mathcal{N}}\max_{j\in\mathcal{L}}\left|\sum_{i\in\mathcal{K}} d_i - \sum_{i\in\mathcal{K}}\mathbb{P}(i_k = i \mid i_0 = j)\right| \leq M_1 \exp\left(-\frac{k}{M_2}\right), \quad \forall\, k \geq 0.
$$

(3) The stochastic process $\{w(k)\}_{k\geq 0}$ is $\mathcal{F}_{k+1}$ measurable and satisfies $\mathbb{E}[w(k) \mid \mathcal{F}_k] = 0$ for all $k \geq 0$. Further, there exists $\bar{w} > 0$ such that $\sup_{k\geq 0}|w(k)| \leq \bar{w}$ almost surely.

(4) There exists some constant $\bar{x} > 0$ such that $\sup_{k\geq 0}\|x(k)\|_\infty \leq \bar{x}$ almost surely.

Remark D.2. As a consequence of Assumption D.1 (1), we have $\|F(x)\|_\infty - \|F(0)\|_\infty \leq \|F(x) - F(0)\|_\infty \leq \gamma\|x\|_\infty$ for all $x \in \mathbb{R}^n$. Therefore, there exists $C = F(0)$ such that $\|F(x)\|_\infty \leq \gamma\|x\|_\infty + C$ for all $x \in \mathbb{R}^n$. One can easily verify that under Assumption D.1 (1) and (3), Statement (4) is automatically satisfied with $\bar{x} = \bar{w}/(1-\gamma)$. However, we treat the upper bound $\bar{x}$ on $\sup_{k\geq 0}\|x(k)\|_\infty$ as a separate assumption, since one might obtain a tighter bound on $\bar{x}$ than what is implied by Assumption D.1 (1) and (3), based on the specific stochastic approximation algorithm of interest.

Next, we present a convergence result showing that the stochastic approximation converges to the unique solution (denoted by x^*) of the following equation:

$$\underbrace{(\Phi^\top D\Phi)^{-1}\Phi^\top DF(\Phi x^*)}_{:=F'(x^*)} = x^*.$$

where $D = \mathrm{diag}(d) \in \mathbb{R}^{n\times n}$. The fact that the above equation admits a unique solution follows from the composite operator $F'(\cdot)$ being a contraction mapping with respect to some suitable norm. See (Lin et al., 2021, Proposition C.1) for more details.

The following notation is needed to present the theorem. Let K be the final iteration index. Let the stepsize be $\alpha_k = H/(k+k_0)$ with $k_0 = \max(4H, 2M_2\log K)$ and $H \geq 2/(\sigma'(1-\gamma))$. Define constants $C_1 = 2\bar{x} + C + \bar{w}$, $C_2 = 4\bar{x} + 2C + \bar{w}$, and $C_3 = 2M_1(2\bar{x}+C)(1+2M_2+4H)$.

Theorem D.3. *Suppose that Assumption D.1 is satisfied. Then, with probability at least* $1-\delta$*, we have*

$$\|x(K) - x^*\|_\infty \leq \frac{\tilde{C}_a}{\sqrt{K+k_0}} + \frac{\tilde{C}_b}{K+k_0},$$

where

$$\begin{aligned}\tilde{C}_a =& \frac{4HC_2}{1-\gamma}\sqrt{2M_2\log K\log\left(\frac{4lM_2K}{\delta}\right)},\\ \tilde{C}_b =& 4\max\left\{\frac{48M_2C_1H\log K + \sigma'C_3}{(1-\gamma)\sigma'}, \frac{2\bar{x}(2M_2\log K + k_0)}{1-\gamma}\right\}.\end{aligned}$$

E Examples and Experiments

E.1 Details of the Synthetic MDP

The Original MDP Model The joint state of the three agents can be represented as a three-dimension vector $s = (s^1, s^2, s^3)$, where $s^i \in \{0,1\}$ for $i \in \{1,2,3\}$. Similarly, the joint action is denoted as $a = (a^1, a^2, a^3)$, where $a^i \in \{0,1\}$ for $i \in \{1,2,3\}$.

Next, we specify the transition probabilities. Agents 1 and 2 are strongly coupled, always sharing the same state value, with identical initial states. Their state remains unchanged when they take the same action; otherwise, it simultaneously transitions from 0 to 1 or from 1 to 0. Mathematically, we have $s_0^1 = s_0^2$, and for any $k \geq 0$:

$$\mathbb{P}(s_{k+1}^1 = s_k^1, s_{k+1}^2 = s_k^2 \mid a_k^1 = a_k^2) = 1, \quad \mathbb{P}(s_{k+1}^1 \neq s_k^1, s_{k+1}^2 \neq s_k^2 \mid a_k^i \neq a_k^2) = 1.$$

For agent 3, any action leads to a state transition to 0 or 1 with probability 0.5 when (i) its current state value is 1 or (ii) its state value is 0 and agent 2 takes the same action. However, if agent 3's state value is 0 and it takes a different action from agent 2, the next state will be 0 with probability 0 and 1 with probability 1. Mathematically, the transition model for Agent 3 is summarized as

$$\begin{aligned}\mathbb{P}(s_{k+1}^3 = 0 \mid s_k^3 = 0, a_k^2 \neq a_k^3) =& 0, & \mathbb{P}(s_{k+1}^3 = 1 \mid s_k^3 = 0, a_k^2 \neq a_k^3) =& 1,\\ \mathbb{P}(s_{k+1}^3 = 0 \mid s_k^3 = 0, a_k^2 = a_k^3) =& 0.5, & \mathbb{P}(s_{k+1}^3 = 1 \mid s_k^3 = 0, a_k^2 = a_k^3) =& 0.5,\\ \mathbb{P}(s_{k+1}^3 = 0 \mid s_k^3 = 1) =& 0.5, & \mathbb{P}(s_{k+1}^3 = 1 \mid s_k^3 = 1) =& 0.5.\end{aligned}$$

In terms of reward, each agent receives a reward of 1 if its state value remains unchanged; otherwise, it receives a reward of 0.

The Dependence level If Agents 1 and 2 coordinate, they can be viewed as a single agent controlling two action variables. It is straightforward to derive $\mathcal{E} = 0.5$ by solving the optimization problem in Eq. (1). If Agents 2 and 3 are grouped together, we can find $\hat{P}_a$ and determine the dependence level as $\mathcal{E} = 0.75$. Similarly, when agents 1 and 3 are grouped together, we obtain $\mathcal{E} = 0.875$. The detailed separable MDPs for different grouping options are shown in Figure 6.

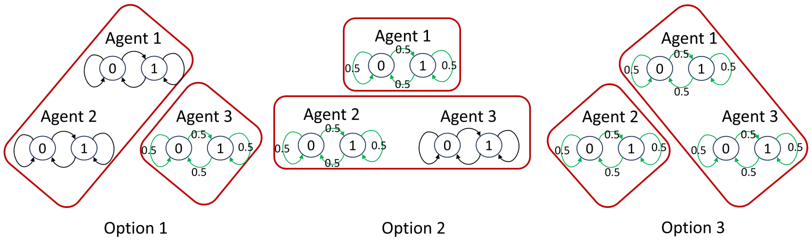

Figure 6: The separable MDPs for different grouping options. Here, black arrows mean the same transition probabilities as the original MDP. Green arrows mean random transition probabilities with 0.5 to any state under any state and action.

E.2 Details of Distributed Energy Storage Operation

Consider B distributed energy storage (ES) devices operating within a distribution network comprising a set of buses $\mathcal{N}$. Each ES, acting as an agent, performs IL by controlling its local charging/discharging power to optimize both global revenue and voltage safety of the network. For simplicity, we focus solely on the voltage safety of the distribution network. Let $V_j > 0$ represent the voltage at bus $j \in \mathcal{N}$. The voltage is safe when it satisfies $V_j \in [V_{\min}, V_{\max}]$, where $V_{\min}$ and $V_{\max}$ are the minimum and maximum voltage levels required to maintain safety. Note that the ESs are deployed at a subset of all buses $\mathcal{N}$, and each ES is responsible for keeping the voltage safe at the buses it can observe.

In the MDP model, the global state is defined by the voltage safety of B agents, i.e., $s = (s^1, s^2, \ldots, s^B)$. Each agent i can observe the local state s^i which indicates the voltage safety of the buses agent i can observe. The local state $s^i = 1$ means that the voltage of buses observed by agent i is safe, and $s^i = 0$ otherwise. Note that this is a simplified model for a practical problem, as the state in real-world applications may include additional information, such as power load and voltage levels, to form a complete MDP model. When an ES is located at bus $j \in \mathcal{N}$, it can observe the voltage safety of a set of neighboring buses, denoted by $\mathcal{N}^j$. We also assume $\mathcal{N}^i \cap \mathcal{N}^j = \emptyset, i \neq j$, and the collective observation of B agents can cover the voltage safety of all buses in the network. The global action is $a = (a^1, a^2, \cdots, a^B)$, where a^i is the binary action of turning on/off a local controller to actuate the charging/discharging power of the ith agent. The reward of each agent is related to its local voltage safety and local action (i.e., whether the local controller is activated).

Let P^j represent the power load at bus j. When the power P^j is adjusted, while the load of other buses remains unchanged, the voltages at all buses are affected due to the power flow constraints. Bus j can be considered relatively "independent" if the voltage change at buses in $\mathcal{N}^j_-$ is small, where $\mathcal{N}^j_- = \mathcal{N} \setminus \mathcal{N}^j$. To be specific, for example, we can define the voltage sensitivity at bus j as:

$$VS(j) = \max_{i \in \mathcal{N}^j_-} \left\{ \frac{\mathrm{d}V_i}{\mathrm{d}P_j} \right\}.$$

Note that the voltage sensitivity depends on the load distribution of the network. In practice, $VS(j)$ can be calculated using finite difference by introducing perturbations to the power at bus j under representative load conditions. When the ESs are located at buses with low voltage sensitivity, the local voltage safety is mainly affected by local charging/discharging actions. Consequently, we have

$$P_a(s_1, s_2) \approx \prod_i \mathbb{P}(s_2^i \mid s_1^i, a^i), \quad \forall\, s_1, s_2 \in \mathcal{S}, a \in \mathcal{A},$$

which leads to a small dependence level.

E.3 EV Charging Problem

Consider the problem where there are multiple charging stations in a distributed network and the total charging power must remain within a time-varying capacity to guarantee safety. A natural approach to address this problem is to formulate charging coordination as a hierarchical resource allocation problem, where the resource is the charging capacity. Each charging station receives an allocated charging capacity which denotes the maximum power that can be used to charge an EV. Consequently, safety is guaranteed because the total charging power consumed by all stations is upper bounded by the total charging capacity.

Consider allocating the charging capacity through a tree structure shown in Figure 7, where decisions are made layer by layer. Agents in the same layer perform IL and decide the charging capacity assigned to the children. Specifically, as illustrated in Figure 7, the maximum charging power is first allocated to the top agent and it needs to decide the capacities allocated to its left and right children. Each child receives a maximum charging capacity, which means that the total charging power induced by the charging stations on the child's side must be within the limit. Then, similar allocations are performed on all lower layers. Absolute safety is guaranteed since the charging capacity is allocated layer by layer to make sure the total charging power is within the upper limit. Finally, all charging stations on leaves charge the connected EVs no more than the maximum charging capacity they receive.

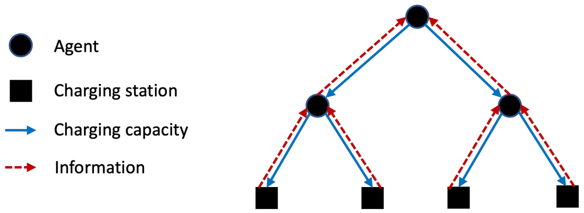

Figure 7: Illustration of a binary tree to allocate the charging capacity.

To make decisions, each agent observes the charging capacity assigned to it and the information of its children. The information from children is contracted from high dimensions to low dimensions for scalability before being transferred to the parents. In this formulation, the problem of EV charging is captured by an MARL with various dependence levels. To see this, consider agents in the same layer. Their parents' and children's policies can be seen as stationary by our design of the multi-scale learning rates (see Appendix E.4 for a detailed discussion). Information contraction controls the dependence level $\mathcal{E}$ among the agents in the same layer, which leads to different performance gaps in our analysis. To further elaborate, consider the following two extreme cases. In one extreme case, the contraction keeps the information unchanged, in which case the agents within the same layer are strongly correlated because their parent makes decisions based on their full information. In another extreme case, no information is communicated to the parents. As a result, all agents in the same layer are completely independent.

In our setting, all agents cooperate to maximize the long-term discounted total reward of all charging locations. Each agent can only communicate locally with its parent or children.

Each charging station j has its local information $I^j(t) = (p^j(t), d^j(t), f^j(t), l^j(t))$ of the connected EV at time t, where $p^i(t)$ is the rated charging power, $d^j(t)$ is the proportion of satisfied demand, $f^j(t)$ and $l^j(t)$ are the additional time needed to fully charge the EV and the remaining time before the EV leaves. All variables are 0 if no EV is connected or the connected EV is fully charged. The charging station j can induce a reward of $r^j(t) = \Delta d^j(t) = d^j(t+1) - d^j(t)$ after charging the connected EV with power $P_c^j(t) = \min\{P^j(t), p^j(t)\}$, where $P^j(t)$ denotes the maximum charging capacity it receives. The reward is $r^j(t) = 0$ if the connected EV is not charged or no EV is connected.

We define the states, actions, and rewards for the agents as follows:

States For each agent i, the state includes the information from its children and the charging capacity allocated to it, which can be written as

$$s^i(t) = (I^{C_l^i}(t), I^{C_r^i}(t), P^i(t)),$$

where C_l^i, C_r^i denote the left and right children of agent i. The information of agent i is constructed recursively from the bottom level to the top as a function of its children's information. Specifically, we define

$$I^i(t) = \mathbf{W}_1^i(t)(I^{C_l^i}(t))^T + \mathbf{W}_2^i(t)(I^{C_r^i}(t))^T,$$

where $\mathbf{W}_1^i(t), \mathbf{W}_2^i(t)$ are both matrices that can be designed to control the information communicated to the parents. We will discuss more about how to control the dependence level with different choices of $\mathbf{W}_1^i(t), \mathbf{W}_2^i(t)$ below.

Actions The action of agent i at time t is the proportion of charging capacity allocated to the left child $a^i(t) \in [0, 1]$. Consequently, the proportion of charging capacity allocated to the right child is $1 - a^i(t)$.

Rewards The reward of agent i is defined recursively as the sum of the rewards of its children from the bottom level to the top level. That is, $r^i(t) = r^{C_l^i}(t) + r^{C_r^i}(t)$. Thus, the reward of agent i is equal to the total reward of all charging stations that have i as an ancestor.

E.4 Detailed Experimental Setting

E.4.1 Experimental Setting for the Synthetic MDP

In this case, we consider the synthetic MDP in Example 2.2. Three agents are grouped into two with different options according to Figure 2. For both IQL and INAC, we will repeat the training and testing 100 times, respectively. Each run includes 3000 steps for training and 1000 steps for testing. The discount factor is $\gamma = 0.99$. Specifically, the inner loop of INAC has $K = 100$ steps. We set $\alpha = 0.05, k_0 = 4\alpha$ and $\alpha_k = \alpha/(k + k_0)$ for the critic. The stepsize for the actor is set as $\eta_0 = 0.2$ and $\eta_t = \sum_{i<t} \eta_i/\gamma^{2t-1}$. To encourage exploration during the training process, each agent takes action uniformly at random with probability $\epsilon_k = (1 - k/K)/10$. For IQL, we also set $\alpha = 0.05, k_0 = 4\alpha$ and $\alpha_k = \alpha/(k + k_0)$. The policy π_b is set as the uniform distribution over the action space. We calculate the normalized reward relative to the optimal expected reward of each episode consisting of 100 steps. The optimal reward can be easily computed with the model described in Appendix E.1.

E.4.2 Experimental Setting for EV Charging

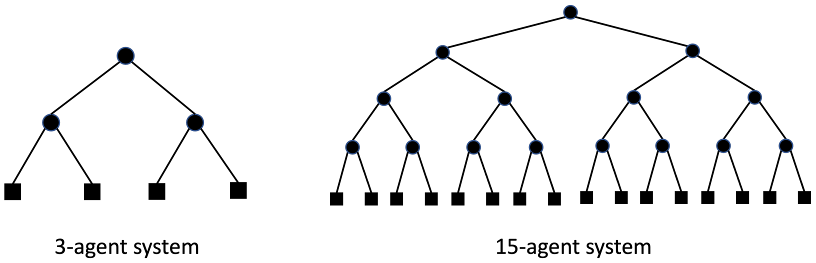

Figure 8: The structures of the 3-agent system (left) and the 15-agent system (right). The square nodes represent the charging stations and the circle nodes represent the agents. Note that our results do not require the interaction structure to be full binary trees.

Here we consider the 3-agent and 15-agent systems shown in Figure 8. Recall that $\mathbf{W}_1^i(t), \mathbf{W}_2^i(t)$ can be designed to derive different information contractions with different dependence levels among agents within the same layer. We design three types of information contractions to perform IL, including full information, average information, and no information. Full information means that each agent communicates the full information from its children to its parent, where the dimensions of agents' states increase exponentially from the bottom to the top. Average information means that the information communicated to the parent is taken as a weighted average of the information from its children. No information means that each agent communicates nothing to its parent. If $\mathbf{W}_1^i(t) = 0, \mathbf{W}_2^i(t) = 0$, the children do not send any information to their parents. The agent can also send average information by mixing the information of its children with $\mathbf{W}_1^i(t) = diag(1, e_l^i(t), e_l^i(t), e_l^i(t))$ and $\mathbf{W}_2^i(t) = diag(1, e_r^i(t), e_r^i(t), e_r^i(t))$, where $e_l^i(t) = \frac{p^{C_l^i}(t)}{p^{C_l^i}(t)+p^{C_r^i}(t)}$, $e_r^i(t) = \frac{p^{C_r^i}(t)}{p^{C_l^i}(t)+p^{C_r^i}(t)}$. We set $e_l^i(t) = e_r^i(t) = 0$ if

$p^{C_l^i}(t) = p^{C_r^i}(t) = 0$. In other words, the contracted $d^i(t), f^i(t)$, and $l^i(t)$ are the weighted average of information from its children according to the rated charging power. The agent sends the full information to its children when $\mathbf{W}_1^i(t) = (\mathbf{I}_{n^i}, \mathbf{0})^T \in \mathbb{R}^{2n^i \times n^i}, \mathbf{W}_2^i(t) = (\mathbf{0}, \mathbf{I}_{n^i})^T \in \mathbb{R}^{2n^i \times n^i}$, where $\mathbf{I}_{n^i}$ is the identity matrix, and n^i is the number of dimensions of the information from agent i's children.

EV arrivals are simulated with fixed arrival rates sampled from $(0, 1)$ for each charging station. Each EV is set with random charging demand, maximum charging power, and remaining time before leaving. We set the remaining time before leaving longer than the time needed to fully charge the EV. The time interval between two decisions is set to 1 hour.

Two other policies are considered for comparison to our algorithms. The first one is the offline optimal policy, which is non-causal and is computed via linear programming using all information collected throughout the time window. The second is a heuristic baseline policy similar to the business-as-usual policy that charges the EV immediately upon arrival (Sadeghianpourhamami et al., 2019). Due to the tree structure and the safety concern, the baseline policy here selects actions based on the proportion of rated power of the children, i.e., $a^i(t) = e_l^i(t)$.

Parameters of IQL Considering the continuity of the state space, we use two neural networks to learn the Q-functions. The current network is used to provide decisions and interact with the environment, and the target network is used to learn the optimal Q-fucntion. The parameters of the current network are set as the copy of the target network every 1000 steps. The target network is updated every 5 steps. The action space $[0, 1]$ is discretized as $\{0, 0.1, \cdots, 1\}$. We also apply multi-scale learning rates in different layers to make sure the agents in other layers are relatively stationary to the agents in each layer. Then, only the agents within the same layer are relatively non-stationary to each other. The learning rates are set to increase with depth by a multiplier of 10 for Q-networks and policy networks. To encourage exploration, the actions are given randomly with a linearly decreasing probability during the training process. The detailed parameters are listed in Table 1.

Table 1: Parameters of IQL for 3-agent and 15-agent systems

	3-agent system	15-agent system
Network	3 hidden layers, 64 neurons for each layer	
Learning rate (Top agent)	10^{-4}	10^{-5}
Max. exploration prob.	1	0.1
Min. exploration prob.	0.03	0.03
Batch size	32	64
Buffer size	5000	6000

Parameters of INAC Due to the continuity of the state and action space, we also use neural networks to act as the actor and critic. Similarly to IQL, we apply multi-scale learning rates in different layers for INAC. The learning rates are also set to increase with depth by a multiplier of 10 for Q-networks and policy networks. To encourage exploration, we add Gaussian noise to the action with zero mean and decreasing variance with steps. The detailed parameters are shown in Table 2.

Table 2: Parameters of INAC for 3-agent and 15-agent systems

	3-agent system	15-agent system
Network	3 hidden layers, 64 neurons for each layer	
Learning rate for critic (Top agent)	10^{-4}	10^{-6}
Learning rate for actor (Top agent)	5×10^{-5}	5×10^{-7}
Max. variance	10	10
Min. variance	0.4	0.4
Batch size	32	32
Buffer size	4000	4000

Both IQL and INAC are trained and tested 20 times. Each run includes 40000 steps, and each episode includes 10 days (240 steps). The first 125 episodes are for training, and the rest of the episodes are for testing. The normalized reward is calculated relative to the optimal average reward of the offline optimal policy.