Approximate Global Convergence of Independent Learning in Multi-Agent Systems

Ruiyang Jin[*,1], Zaiwei Chen[†,1], Yiheng Lin[†,2], Jie Song[*,2], and Adam Wierman[†,3]

[*]*Peking University,* [1]*jry@pku.edu.edu*, [2]*jie.song@pku.edu.cn*

[†]*California Institute of Technology,* [1]*zchen458@caltech.edu*, [2]*yihengl@caltech.edu*, [3]*adamw@caltech.edu*

Abstract

Independent learning (IL), despite being a popular approach in practice to achieve scalability in large-scale multi-agent systems, usually lacks global convergence guarantees. In this paper, we study two representative algorithms, independent Q-learning and independent natural actor-critic, within value-based and policy-based frameworks, and provide the first finite-sample analysis for approximate global convergence. The results imply a sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-2})$ up to an error term that captures the dependence among agents and characterizes the fundamental limit of IL in achieving global convergence. To establish the result, we develop a novel approach for analyzing IL by constructing a separable Markov decision process (MDP) for convergence analysis and then bounding the gap due to model difference between the separable MDP and the original one. Moreover, we conduct numerical experiments using a synthetic MDP and an electric vehicle charging example to verify our theoretical findings and to demonstrate the practical applicability of IL.

1 Introduction

Reinforcement learning (RL) [52] has become a popular learning framework to solve sequential decision-making problems in the real world and has achieved great success in applications across different domains, such as Atari games [40], the game of Go [49], robotics [21], nuclear fusion control [14], etc. While the early literature on RL predominantly focused on the single-agent setting, motivated by the recent development in applications involving multi-agent interactions, e.g., in areas such as multiplayer games [60] and active voltage control [56], multi-agent reinforcement learning (MARL) has received increasing attention.

Compared to single-agent RL, the interplay of agents in MARL presents significant challenges for developing provably efficient learning algorithms. One of those challenges is scalability. Specifically, even when all agents share an identical interest, any MARL algorithm that uses centralized training becomes computationally intractable as the number of agents increases. To address this scalability issue, various decentralized MARL algorithms were developed, such as centralized training with decentralized execution (CTDE) [39], localized MARL in networked systems [65], and independent learning (IL) [32]. Among them, IL has the minimum requirements on coordination and information sharing, as each agent makes decisions purely based on its local observations, without collecting any information from other agents. Therefore, increasing the number of agents does not introduce additional communication or computational complexity in IL.

Although IL is simple, intuitive, and easy to implement, it is unclear whether agents can jointly achieve any performance guarantees through IL. Note that, in IL, each individual agent overlooks the multi-agent interactions by treating all other agents as parts of the environment, which is non-stationary due to the actions taken by other agents. Therefore, except in cases where the underlying model has a special structure, such as in zero-sum stochastic games [6, 10] or Markov potential games [33], IL, in general, is not guaranteed to converge [64]. This leads to the following fundamental question:

Is it possible to achieve (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. In addition, we introduce the *dependence level* as a novel notion to capture the fundamental limit of IL. Detailed contributions of this work are presented below.

- **Approximate Global Convergence for Independent Learning.** We consider two popular and representative forms of IL, one of which is a value-based algorithm called independent Q-learning (IQL), and the other is 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 is that the bound asymptotically converges within a fixed error term, which is proportional to the dependence level, denoted by $\mathcal{E}$, of the Markov decision process (MDP) model. Here, the dependence level $\mathcal{E}$ characterizes how close the MARL model is to a model where each agent's transitions are independent of each other. Our finite-sample bound implies that, up to the aforementioned asymptotic error term, the sample complexity for the remaining terms to achieve an ϵ-optimality is $\tilde{\mathcal{O}}(\epsilon^{-2})$, which is not improvable [19].

- **A Novel Approach for Analyzing Independent Learning.** The main challenge in analyzing IL is that each agent overlooks the multi-agent interactions by treating other agents as parts of the environment. To overcome this challenge, we develop a novel approach (depicted in Figure 1) that involves (1) constructing a separable MDP consisting of local transition kernels to approximate the original MDP, (2) analyzing IL as if they were implemented on the separable MDP, and finally, (3) bounding the error due to model difference between the separable MDP and the original one to establish the approximate global convergence.

The proof technique discussed above can be used beyond the analysis of IL. Specifically, in a general stochastic iterative algorithm, as long as the random process that drives the iterative algorithm can be approximated by a Markov chain (even though the original random process is not Markovian), our proof technique can be potentially applied to get finite-sample guarantees. See Appendix A.6 for more details.
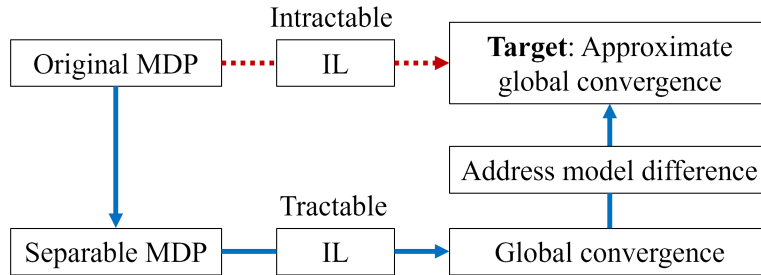

Figure 1: Our Roadmap for Analyzing IL.

- **Validation of Approximate Global Convergence of IL.** We first conduct numerical simulations on a synthetic MDP to verify the global convergence of IL, and to demonstrate that the asymptotic optimality gap of IL is determined by the dependence level. Then, we apply IL to an electric vehicle (EV) charging problem to showcase the practical applicability of our algorithms, where we employ a neural network as a form of function approximation. Notably, while our theoretical results are established for the tabular setting, our algorithms empirically achieve approximate global convergence under function approximation.

1.1 Related Literature

Q-learning and natural actor-critic are two popular and representative algorithms for value-based and policy-based methods, respectively. In the following, we discuss the related work in Q-learning, natural actor-critic, IL, and networked MARL.

Q-Learning. Q-learning was first proposed in [57] as a data-driven variant of the Q-value iteration for solving the Bellman optimality equation. Due to its popularity, the convergence behavior of Q-learning has been extensively studied. Specifically, the asymptotic convergence of Q-learning has been studied in [22, 55] and finite-sample analysis in [9, 17, 35, 44], among many others. Practically, Q-learning has been successfully employed in many applications in the form of the deep Q-network (DQN), such as games [40] and control problems [29]. Although Q-learning has also been applied to multi-agent systems [53], to our knowledge, the theoretical results in this domain are quite 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 [30, 52], a typical example of which is natural actor-critic [4], where natural policy gradient [25, 59, 61] is used for the policy improvement and TD-learning [3, 50, 51] is used for the policy evaluation. Finite-sample analysis of natural actor-critic (with linear function approximation) was performed in [1, 8, 27, 31] and the references therein, where the state-of-the-art sample complexity is $\tilde{\mathcal{O}}(\epsilon^{-2})$. However, they all focused on the single-agent setting. In MARL where each agent independently performs natural actor-critic, existing results 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 [16, 18, 23], communication networks [36, 41, 68], etc. While IL is empirically popular, it has been demonstrated in [54] that IL may fail for tasks where coordination among agents is needed. Therefore, it is vital to understand when it is possible for IL to achieve global convergence [5, 64]. Although IL has been theoretically justified to some extent in certain multi-agent scenarios, such as zero-sum stochastic games [10, 13] and Markov potential games [15, 33], the results are still lacking for the general setting. In this work, we establish finite-sample guarantees of 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 localized interaction structure is commonly found in applications that involve social networks, computer networks, traffic networks, etc. [37, 45, 69], and it enables decentralized decision-making based on local observations and shared information from neighboring agents [12]. Theoretically, the authors of [65] incorporate a consensus algorithm in the design of their networked MARL algorithm, and provide convergence analysis under linear function approximation. Many algorithms and analyses are proposed for the extended settings, such as continuous spaces [63], and stochastic networked MARL [66]. A more detailed review of networked MARL can be found in [64]. The works most relevant to IL within networked MARL are the scalable algorithms designed using the exponential decay property [37, 43, 45, 67, 69], 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.

2 Problem Formulation

Consider an MARL problem with n agents. Let $\mathcal{S}$ and $\mathcal{A}$ be the global state space and the global action space, respectively. Given $i \in \{1, 2, \cdots, n\} := [n]$, let $\mathcal{S}^i$ (respectively, $\mathcal{A}^i$) be the state space (respectively, the action space) of agent i. In 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 actions, i.e., $P_a(s, s')$ denotes the probability of transiting to the global state s' after taking the joint action a at the global state s. For each $i \in [n]$, let $\mathcal{R}^i : \mathcal{S}^i \times \mathcal{A}^i \mapsto [0, 1]$ be the reward function of agent i. Note that restricting the image space of the reward function to $[0, 1]$ is without loss of generality because we are working with a finite MARL problem. For any $s = (s^1, s^2, \cdots, s^n) \in \mathcal{S}$ and $a = (a^1, a^2, \cdots, a^n) \in \mathcal{A}$, the total one-stage reward is given by $\mathcal{R}(s, a) = \sum_{i \in [n]} \mathcal{R}^i(s^i, a^i)$. Let $\gamma \in (0, 1)$ be the discount factor, which captures the weight we place on future rewards. We denote the MDP for the MARL problem as $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$.

Given a joint policy $\pi : \mathcal{S} \mapsto \Delta(\mathcal{A})$, where $\Delta(\mathcal{A})$ stands for the probability simplex on $\mathcal{A}$, its associated Q-function $Q_\pi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ is defined as

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

where we use $\mathbb{E}_\pi[\,\cdot\,]$ to indicate that the actions are chosen according to the joint policy π. We further define the value function $V_\pi \in \mathbb{R}^{|\mathcal{S}|}$ as $V_\pi(s) = \mathbb{E}_{a \sim \pi(\cdot|s)}[Q_\pi(s, a)]$ for all $s \in \mathcal{S}$. Given an initial distribution μ on the states, denote $V_\pi^\mu := \mathbb{E}_{s \sim \mu}[V_\pi(s)]$ as the expected value of a policy π. The goal is to find a globally optimal joint policy π such that Q_π (or equivalently, V_π, V_π^μ) is uniformly maximized for all (s, a).

For an individual agent $i \in [n]$, we also define $Q_\pi^i \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ as the Q-function of agent i given policy π:

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

Note that we have $Q_\pi(s, a) = \sum_i Q_\pi^i(s, a)$ for all (s, a).

2.1 The Dependence Level

We now introduce the concept of *dependence level* for the multi-agent MDP model to characterize the globally coupled state transitions. This is a new notion for analyzing IL that we introduce in this paper. We first illustrate the concept of *separable MDPs*, which is crucial for defining the dependence level of the multi-agent MDP. For any $i \in [n]$, let $\mathcal{P}^i$ be the set of $|\mathcal{S}^i|$ by $|\mathcal{S}^i|$ right stochastic matrices, and let

$$\hat{\mathcal{Z}} = \left\{ \{\hat{P}_a\}_{a \in \mathcal{A}} \,\middle|\, \exists \hat{P}_{a^i} \in \mathcal{P}^i \text{ s.t. } \hat{P}_a(s_1, s_2) = \prod_{i=1}^{n} \hat{P}_{a^i}(s_1^i, s_2^i), \forall s_1, s_2 \in \mathcal{S} \right\}.$$

We call the MDP with transition matrices $\{\hat{P}_a\}_{a \in \mathcal{A}} \in \hat{\mathcal{Z}}$ a separable MDP. Intuitively, while a separable MDP is defined on the joint state-action space, it can be decomposed into n small MDPs that involve independently on their associated local state-action spaces. We next define the dependence level of a multi-agent MDP model.

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

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

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

Remark. Since the quantity $\max_{s,a}\|P_a(s,\cdot)-\hat{P}_a(s,\cdot)\|_{\mathrm{TV}}$ as a function of $\{\hat{P}_a\}_{a\in\mathcal{A}}$ is a continuous function and $\hat{\mathcal{Z}}$ is a compact set, the $\min(\cdot)$ in Eq. (1) is well-defined due to the extreme value theorem [47].

To understand Definition 2.1, suppose that the original MDP $\mathcal{M}$ is separable. Then the dependence level $\mathcal{E}=0$. In this case, we would expect IL to achieve global convergence because the problem is essentially to find the optimal policies of n decoupled MDPs. More generally, the dependence level $\mathcal{E}$ can be positive, in which case the original MDP is not separable, and the exact global convergence may not be achievable. Therefore, the dependence level $\mathcal{E}$ serves as a measure of how close the original MDP is to the space of separable MDPs and captures the fundamental limit of IL for global convergence. Let

$$\hat{\mathcal{P}} = \arg\min_{\{\hat{P}_a\}_{a\in\mathcal{A}}\in\hat{\mathcal{Z}}} \max_{s,a}\left\|P_a(s,\cdot)-\hat{P}_a(s,\cdot)\right\|_{\mathrm{TV}}$$

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 use $\hat{Q}_\pi\in\mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$ as the Q-function of a policy π under $\hat{\mathcal{M}}$.

Remark. In this work, although we assumed that the global reward is the summation of each agent's individual reward for simplicity of presentation, this assumption can be relaxed. Specifically, for each $i\in[n]$, let Λ^i be the set of real-valued functions with domain $\mathcal{S}^i\times\mathcal{A}^i$. Then, by defining the reward dependence level $\mathcal{E}_r$ as $\mathcal{E}_r = \min_{\hat{\mathcal{R}}^i\in\Lambda^i,\forall i\in[n]}\max_{s,a}|\mathcal{R}(s,a)-\sum_{i\in[n]}\hat{\mathcal{R}}^i(s^i,a^i)|$, we can replace the original reward function $\mathcal{R}(s,a)$ by $\sum_{i\in[n]}\hat{\mathcal{R}}^i(s^i,a^i)$ in our analysis. Note that in this case, the reward-dependence level $\mathcal{E}_r$ will also appear in the convergence bounds.

2.2 An Illustrative Example

In this subsection, we provide an illustrative example to further discuss the concept of dependence level.

In multi-agent systems, local information sharing and decision-making coordination can significantly improve the performance of IL [54]. However, it is challenging to select the pertinent information to communicate and the agents with whom to collaborate. The following example illustrates how different options of grouping and collaboration can lead to different dependence levels, which significantly affects the performance of IL.
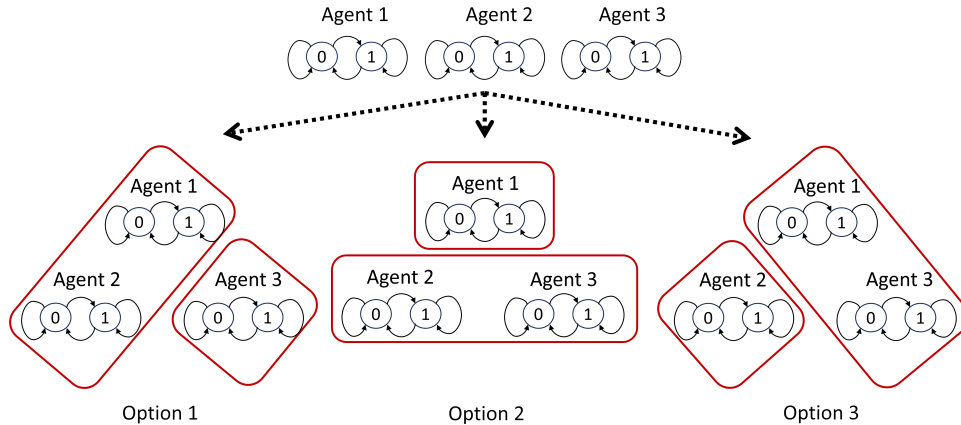

Figure 2: Illustration of the example.

Example 2.1. Consider an MDP consisting of 3 agents. The state space of each agent is $\{0,1\}$, and 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 B.1. Suppose that we are allowed to group 2 of the 3 agents as one agent, in which case the 2 agents can share information and coordinate with each other. Then, we have 3 different grouping options shown in Figure 2. After calculating the dependence level $\mathcal{E}$ for each grouping option (cf. Appendix B.1), we have: Option 1 leads to $\mathcal{E} = 0.5$; Option 2 leads to $\mathcal{E} = 0.75$, and Option 3 leads to $\mathcal{E} = 0.875$.

3 Main Results

This section presents our main results. We first present the IQL and the INAC as representative IL algorithms in Sections 3.1 and 3.2, respectively. Then, we present the finite-sample approximate global convergence guarantees for both algorithms in Section 3.3. The proof sketch of our main theorems is provided in Section 3.4, with detailed proof deferred to the appendix.

3.1 Independent Q-Learning

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

Algorithm 1 Independent Q-Learning (Agent i)

1: **Input:** An integer K, a behavior policy π_b^i, and an initialization $Q_0^i = 0$.
2: **for** $k = 0, 1, 2, \cdots, K - 1$ **do**
3:     Implement $A_k^i \sim \pi_b^i(\cdot \mid S_k^i)$ (simultaneously with all other agents), and observes S_{k+1}^i.
4:     Update Q_k^i according to

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

5: **end for**
6: **Output:** Q_K^i

Algorithm Details. In IQL, each agent uses a local behavior policy $\pi_b^i : \mathcal{S}^i \mapsto \Delta(\mathcal{A}^i)$ to interact with the environment to collect samples (cf. Algorithm 1 Line 3) and updates its local Q-function estimate $Q_k^i \in \mathbb{R}^{|\mathcal{S}^i||\mathcal{A}^i|}$ according to Algorithm 1 Line 4, where α_k is the learning rate. Note that this is an asynchronous update because only one component (i.e., the (S_k^i, A_k^i)-th 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 [2, 52, 57] for more details about Q-learning.

Remark. Unlike single-agent Q-learning, which is driven by a trajectory of Markovian samples, the local sample trajectory $\{(S_k^i, A_k^i)\}_{k \geq 0}$ used for IQL, in general, does *not* form a Markov chain. Therefore, the convergence of the local Q-function $\{Q_{t,k}^i\}_{k \geq 0}$ does not directly follow from the existing results studying single-agent Q-learning [9, 17, 35, 44], which presents a challenge in the analysis.

3.2 Independent Natural Actor-Critic

The detailed description of the INAC algorithm is presented 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 uses independent natural policy gradient (INPG) to update the policy and the critic uses independent TD-learning (ITD) to estimate the local Q-function, which is needed for the actor. We next elaborate on the actor and the critic in more detail.

Algorithm 2 Independent Natural Actor-Critic (Agent i)

1: **Input:** Integers K, T, initializations $\theta_0^i = 0$ and $Q_{t,0}^i = 0$ for all $t \geq 0$.
2: **for** $t = 0, 1, 2, \cdots, T - 1$ **do**
3:    **for** $k = 0, 1, 2, \cdots, K - 1$ **do**
4:       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.
5:       Update $Q_{t,k}^i$ according to

$$Q_{t,k+1}^i(S_k^i, A_k^i) = (1 - \alpha_k)Q_{t,k}^i(S_k^i, A_k^i) + \alpha_k \left(\mathcal{R}^i(S_k^i, A_k^i) + \gamma \mathbb{E}_{a^i \sim \pi_{(t)}^i(\cdot \mid S_{k+1}^i)} \left[Q_{t,k}^i(S_{k+1}^i, a^i) \right] \right).$$

6:    **end for**
7:    $\theta_{t+1}^i = \theta_t^i + \eta_t Q_{t,K}^i$.
8: **end for**
9: **Output:** $\pi_{\theta_T^i}^i$

Independent Natural Policy Gradient for the Actor.   Policy gradient [52] is a popular approach for solving the RL problem. The idea is to perform gradient ascent in the policy space. NPG can be viewed as a variant of the policy gradient, where the Fisher information matrix is used as a preconditioner [25]. See [1, 25, 31] for other equivalent formulations and interpretations of NPG.

To present the NPG algorithm, we consider using softmax policies with parameter $\theta \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$:

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

It was shown in [1] that when using softmax policies, NPG takes the following form in the parameter space (which is also called Q-NPG):

$$\theta_{t+1} = \theta_t + \eta_t Q_{(t)}, \tag{2}$$

where η_t is the stepsize and we denote $Q_{(t)} = Q_{\pi_{\theta_t}}$. In addition, the previous update equation in the parameter space is equivalent to the following update equation in the policy space [1]:

$$\pi_{(t+1)}(a \mid s) = \frac{\pi_{(t)}(a \mid s) \exp\{\eta_t Q_{(t)}(s, a)\}}{\sum_{a' \in \mathcal{A}} \pi_{(t)}(a' \mid s) \exp\{\eta_t Q_{(t)}(s, a')\}}, \ \forall (s, a) \in \mathcal{S} \times \mathcal{A}, \tag{3}$$

where we denote $\pi_{(t)} = \pi_{\theta_t}$ for simplicity of presentation. However, carrying out the update rule in Eq. (2) (or 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

$$\theta_{t+1}^i = \theta_t^i + \eta_t \hat{q}_{(t)}^i, \tag{4}$$

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 $\pi^i : \mathcal{S}^i \mapsto \Delta(\mathcal{A}^i)$ 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 $(s^i, a^i) \in \mathcal{S}^i \times \mathcal{A}^i$

---
[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.

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. (4), note that Eq. (4) is equivalent to the following update in the policy space:

$$\pi^i_{(t+1)}(a^i \mid s^i) = \frac{\pi^i_{(t)}(a^i \mid s^i) \exp\{\eta_t \hat{q}^i_{(t)}(s^i, a^i)\}}{\sum_{\bar{a}^i \in \mathcal{A}^i} \pi^i_{(t)}(\bar{a}^i \mid s^i) \exp\{\eta_t \hat{q}^i_{(t)}(s^i, \bar{a}^i)\}}, \ \forall (s^i, a^i) \in \mathcal{S}^i \times \mathcal{A}^i.$$

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 $Q_\pi(s, a) = \sum_{i \in [n]} \hat{q}^i_\pi(s^i, a^i)$. As a result, when each agent updates its policy parameter θ^i_t according to Eq. (4), the joint policy obeys the following update rule:

$$\begin{aligned}
\pi_{(t+1)}(a \mid s) &= \prod_{i=1}^n \pi^i_{(t+1)}(a^i \mid s^i) \\
&= \prod_{i=1}^n \left[\frac{\pi^i_{(t)}(a^i \mid s^i) \exp\{\eta_t \hat{q}^i_{(t)}(s^i, a^i)\}}{\sum_{\bar{a}^i \in \mathcal{A}^i} \pi^i_{(t)}(\bar{a}^i \mid s^i) \exp\{\eta^i_t \hat{q}^i_{(t)}(s^i, \bar{a}^i)\}} \right] \\
&= \frac{\pi_{(t)}(a \mid s) \exp\{\eta_t Q_{(t)}(s, a)\}}{\sum_{\bar{a} \in \mathcal{A}} \pi_{(t)}(\bar{a} \mid s) \exp\{\eta_t Q_{(t)}(s, \bar{a})\}}, \ \forall (s, a) \in \mathcal{S} \times \mathcal{A},
\end{aligned}$$

which is exactly the desired Q-NPG update presented in Eq. (3). In general, when the original n-agent MDP $\mathcal{M}$ is not separable, Eq. (4) can be viewed as an approximation of the Q-NPG update in Eq. (3). 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. (4), each agent needs to estimate its local Q-function $\hat{q}^i_{(t)}$ to carry out the update. To achieve that, we use ITD, which is presented next.

Independent TD-Learning 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 Algorithm 2 Line 5. Similarly to Q-learning, TD-learning can also be viewed as a stochastic approximation algorithm for solving the Bellman equation for policy evaluation. See [2, 51, 52] for more details about TD-learning.

Finally, combining the INPG in Eq. (4) with the ITD for estimating $\hat{q}^i_{(t)}$ leads to the INAC presented in Algorithm 2.

3.3 Finite-Sample Analysis

To present our theoretical results, we first introduce our assumption regarding the MDP model. For any policy π, state-action pairs $(s, a), (\bar{s}, \bar{a}) \in \mathcal{S} \times \mathcal{A}$, and $k \geq 0$, let $\mathbb{P}^\pi_k(\bar{s}, \bar{a}, s, a)$ be the probability of visiting $(\bar{s}, \bar{a})$ at time step k starting at (s, a) and following the policy π thereafter, that is, $\mathbb{P}^\pi_k(\bar{s}, \bar{a}, s, a) = \mathbb{P}((S_k, A_k) = (s, a) \mid (S_0, A_0) = (\bar{s}, \bar{a}))$, where $A_\ell \sim \pi(\cdot \mid S_\ell)$ for all $\ell \in \{1, 2, \cdots, k\}$.

Assumption 3.1. *For any joint policy π, the induced Markov chain $\{(S_k, A_k)\}_{k \geq 0}$ (from the original MDP $\mathcal{M}$) is irreducible and aperiodic with a unique stationary distribution $d_\pi \in \Delta(\mathcal{S} \times \mathcal{A})$. In addition, we assume that $\sigma := \inf_\pi \min_{s,a} d_\pi(s, a) > 0$, and there exist $M_1 \geq 0$ and $M_2 \geq 1$ such that*

$$\max_{\mathcal{N} \subseteq \mathcal{S} \times \mathcal{A}} \sup_\pi \max_{(\bar{s}, \bar{a}) \in \mathcal{S} \times \mathcal{A}} \left| \sum_{(s,a) \in \mathcal{N}} (d_\pi(s, a) - \mathbb{P}^\pi_k(\bar{s}, \bar{a}, s, a)) \right| \leq M_1 \exp\left(-\frac{k}{M_2}\right), \ \forall k \geq 0.$$

In RL, to successfully learn an optimal policy, it is well known that having a sufficient exploration component is necessary. Assumption 3.1 is imposed to ensure the exploration of RL agents, which states that for any joint policy π, the agents can sufficiently explore the state-action space. This type of assumption is commonly imposed in the existing literature studying RL algorithms, especially those with time-varying sampling policies. See for example [27, 37, 58, 62, 70].

To proceed and state our main results, we need to introduce more notation. For any $i \in [n]$ and $(\bar{s}^i, \bar{a}^i) \in \mathcal{S}^i \times \mathcal{A}^i$, let $d'_\pi(\bar{s}^i, \bar{a}^i) = \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}, s^i = \bar{s}^i, a^i = \bar{a}^i} d_\pi(s, a)$. 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. Next, we present the finite-sample analysis of IQL.

Theorem 3.1. *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*

$$V_{\pi_*}^\mu - V_{\pi_K}^\mu \leq \underbrace{\frac{1}{1-\gamma} \left(\frac{2nC'_a}{\sqrt{K+k_0}} + \frac{2nC_b}{K+k_0} \right)}_{E_1:\ Q\text{-}Learning\ Convergence\ Error} + \underbrace{\frac{8n\gamma\mathcal{E}}{(1-\gamma)^3}}_{E_2:\ Error\ due\ to\ \mathcal{E}}, \tag{5}$$

where π_* is an optimal policy, and $\pi_K = (\pi_k^1, \pi_k^2, \cdots, \pi_k^n)$ is the policy greedily induced by $(Q_k^1, Q_k^2, \cdots, Q_k^n)$, that is, $\pi_k^i(a^i \mid s^i) = 1$ if and only if $a^i = \arg\max_{\bar{a}^i} Q_k^i(s^i, \bar{a}^i)$ for all $i \in [n]$, where we break the tie arbitrarily, $C'_a = \frac{40\alpha}{(1-\gamma)^2} \sqrt{M_2 \log K \left(\log \left(\frac{4mnM_2 K}{\delta'} \right) + \log \log K \right)}$, 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. 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 3.1 is presented in Appendix A. Observe that the convergence bound presented in Eq. (5) consists of two terms. The term E_1 converges to zero at a rate of $\tilde{\mathcal{O}}(1/\sqrt{K})$, which matches the convergence rate of Q-learning in the single-agent setting [9, 35, 44]. The term E_2 is asymptotically non-vanishing. Note that 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 we have the global convergence of IQL.

Based on Theorem 3.1, we have the following sample complexity of IQL.

Corollary 3.2. *Given $\epsilon > 0$, 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 $1 - \delta'$, the sample complexity[2] is $\tilde{\mathcal{O}}\left(\epsilon^{-2} \right)$.*

As we see from Corollary 3.2, up to a model difference error that is proportional to the dependence level, Algorithm 1 achieves a sample complexity of $\tilde{\mathcal{O}}(\epsilon^{-2})$ to find a optimal policy. This sample complexity is known to be optimal due to the existing lower bounds for solving RL problems [19].

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

Theorem 3.3. *Consider $\{\pi_{(t)}\}_{t \geq 0}$ generated by Algorithm 2. Suppose that (1) Assumption 3.1 is satisfied, (2) $\alpha_k = \alpha/(k+k_0)$ with $\alpha \geq \frac{2}{\sigma'(1-\gamma)}$ and $k_0 = \max(4\alpha, 2M_2 \log K)$, and (3) η_t satisfies $\eta_0 = \gamma \log |\mathcal{A}|$*

---

[2]It was argued in [26] that, given a finite-sample bound with asymptotically non-vanishing terms on the right-hand side (RHS), 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 3.2, 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 (1) function approximation is used, (2) off-policy sampling is used, and (3) IL is used as in our paper.

and $\eta_t \geq 2n \log |\mathcal{A}| \sum_{i=0}^{t-1} \eta_i / [(1-\gamma)\gamma^{2t-1}]$ *for all* $t \geq 0$*. Then, for any* $\delta \in (0,1)$*, with probability at least* $1 - \delta$*, we have*

$$\left\| Q_{\pi_*} - Q_{(T)} \right\|_\infty \leq \underbrace{\frac{2}{(1-\gamma)^2} \left(\frac{nC_a}{\sqrt{K+k_0}} + \frac{nC_b}{K+k_0} \right)}_{G_1: \text{ TD-Learning Convergence Error}} + \underbrace{\frac{4n\gamma^{T-1}}{(1-\gamma)^2}}_{G_2: \text{Actor Convergence Error}} + \underbrace{\frac{8n\gamma\mathcal{E}}{(1-\gamma)^4}}_{G_3: \text{Error due to } \mathcal{E}} , \quad (6)$$

where π_* *is an optimal policy,* C_a *is defined as* $C_a = \frac{40\alpha}{(1-\gamma)^2} \sqrt{M_2 \log K \left(\log \left(\frac{4mnTM_2K}{\delta} \right) + \log \log K \right)}$, *and* C_b *is defined in Theorem 3.1.*

The detailed proof of Theorem 3.3 is deferred to Appendix A. Similarly to Theorem 3.1, the convergence bound in Theorem 3.3 is composed of terms that converge to 0 asymptotically and a term that is proportional to the dependence level $\mathcal{E}$. Specifically, the terms G_1 and G_2 in Eq. (6) represent the convergence error in ITD for the critic and INPG for the actor. The $\mathcal{O}(1/\sqrt{K})$ convergence rate of G_1 and the geometric convergence of G_2 agree with the existing results in the literature on TD-learning [3, 50] and NPG [7, 31]. The term G_3 is a constant independent of the number of iterations and captures the model difference error between the original MDP and the separable one, which is proportional to the dependence level $\mathcal{E}$.

At first glance, it may seem unnatural that our stepsize sequence $\{\eta_t\}$ for INAC is increasing (see the recursive geometric form in Theorem 3.3). To illustrate this, consider the single-agent setting. In view of the equivalent form of Q-NPG in Eq. (3), it resembles the classical policy iteration (which has geometric convergence) when η_t approaches infinity. This intuition was theoretically justified in [8, 28, 59], as well as from the perspective of mirror descent [31]. Although using increasing stepsizes is theoretically justified, in practice, excessively large stepsizes may discourage exploration. To overcome this practical issue, one can use a more exploration-encouraging variant of INAC, such as choosing actions based on $\pi_{(t)}^i$ with probability $1 - \epsilon$ and choosing actions uniformly at random with probability ϵ for some $\epsilon > 0$.

Next, we derive the sample complexity of Algorithm 2 based on Theorem 3.3.

Corollary 3.4. *Given* $\epsilon > 0$*, 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* $1 - \delta$*, the sample complexity is* $\tilde{\mathcal{O}}\left(\epsilon^{-2}\right)$*.*

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.

Comparison with Results for Networked MARL. In the existing literature, the work closest to ours are those analyzing networked MARL problems. The main idea in networked MARL is that, from each agent's perspective, the 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 [37, 45]. Therefore, by restricting the sharing of information between agents within their κ-hop neighborhood, with a properly chosen κ, decentralized RL algorithms can achieve scalability without compromising too much on optimality. Compared to [37, 45], 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 [37, 45]. 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 made in the existing literature. While our results are more general and applicable to networked MARL, the usefulness depends on how small the dependence level is. The local structure in networked MDP (which limits the impact of far-away agents) may help reduce the dependence level.

3.4 Proof Sketch

Our proof follows the roadmap in Figure 1. Next, we present the proof sketch of Theorem 3.3. The proof of Theorem 3.1 follows a similar approach. The proof consists of the following 3 main steps.

Step 1: Convergence of ITD. The main challenge in analyzing ITD (and also IQL) is that, as a stochastic approximation algorithm, the randomness in the algorithm comes from the stochastic process $\{(S_k^i, A_k^i)\}$, which does not necessarily form a Markov chain. Therefore, the existing results on the Markovian stochastic approximation [11, 50] do not apply directly here. To overcome this challenge, inspired by [37], we model ITD (and also IQL) as a stochastic approximation algorithm with state aggregation and show that $Q_{t,K}^i$ approximates a solution to a variant of the projected Bellman equation, which is denoted as $\tilde{Q}_t^i$. In the end, we obtain the following convergence result with high probability: $\|Q_{t,K}^i - \tilde{Q}_t^i\|_\infty \leq \tilde{\mathcal{O}}(1/\sqrt{K})$.

Step 2: Global Convergence on the Separable MDP. Following our blueprint described in Figure 1, we analyze INAC as if it were implemented in the separable MDP $\hat{\mathcal{M}}$. Then, we take a further step for ITD to prove $\|\tilde{Q}_t^i - \hat{q}_{(t)}^i\|_\infty \leq \mathcal{O}(\mathcal{E})$ using the definition of the dependence level. Altogether, we obtain the approximate convergence for ITD: $\|Q_{t,K} - \hat{Q}_{(t)}\|_\infty \leq \tilde{\mathcal{O}}(1/\sqrt{K}) + \mathcal{O}(\mathcal{E})$, 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 $\hat{\mathcal{M}}$. Combining the results for ITD and INPG, we have the following result for INAC:

$$\|Q_{(T)} - \hat{Q}_{\hat{\pi}_*}\|_\infty \leq \tilde{\mathcal{O}}(1/\sqrt{K}) + \mathcal{O}(\gamma^T) + \mathcal{O}(\mathcal{E}) \tag{7}$$

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

Step 3: 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 due to the model difference to get the approximate global convergence of the original MDP. In fact, we have $\|Q_{\pi_*} - \hat{Q}_{\hat{\pi}_*}\|_\infty \leq \mathcal{O}(\mathcal{E})$, where we recall that $\mathcal{E}$ is the dependence level. Combining the above inequality with Eq. (7) finishes 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 discussed in Section 2.2 to illustrate the effects of the dependence level. In Appendix B.3.3, we apply IQL and INAC to an EV charging problem to demonstrate that our algorithms can be extended to the function approximation setting with approximate global convergence as well.

We run IQL and INAC with the synthetic MDP 100 times, respectively. The rewards are normalized by the optimal average reward and 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 can achieve better performance than Options 2 and 3, which justifies the convergence results in Theorems 3.1 and 3.3, where the optimality gap of IL is controlled by the dependence levels. Comparing IQL with INAC, we find that IQL can achieve a smaller gap than INAC with any option, which is also consistent with our theoretical results where the asymptotically non-vanishing term (cf. E_2) in IQL is smaller than the corresponding term G_3 in INAC by a factor of $1/(1-\gamma)$.

5 Conclusion

In this paper, we investigate the theoretical merit of IL for MARL in the cooperative setting and establish approximate global convergence for 2 representative algorithms: IQL and INAC, both of which achieve
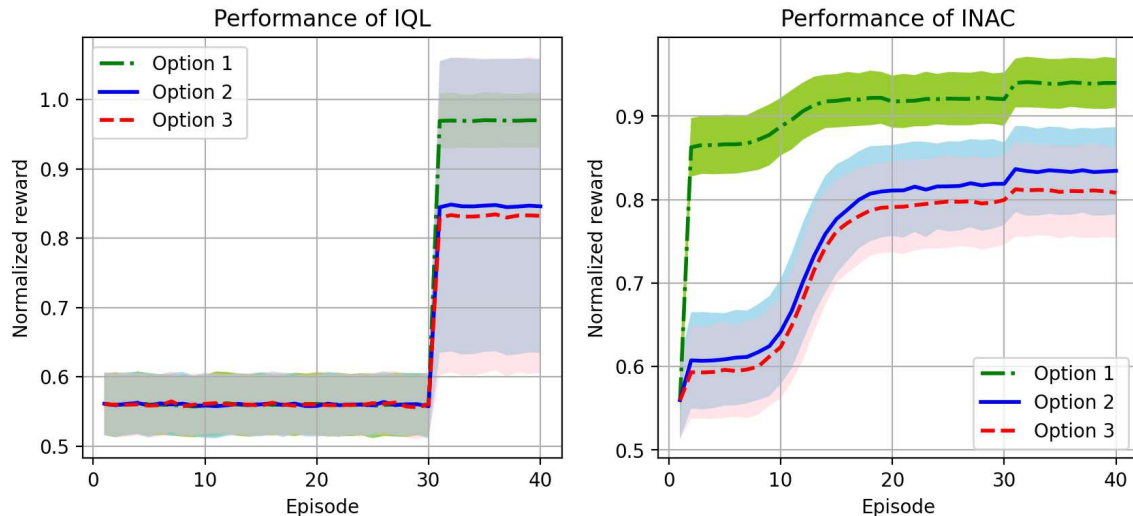
Figure 3: Performance of IQL and INAC

an $\tilde{\mathcal{O}}(\epsilon^{-2})$ sample complexity to find a global optimal policy up to an error term that is proportional to the dependence level. Methodologically, we propose a new method for analyzing 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 justify the theoretical findings.

There are many interesting directions for future work. First, it is worth investigating if our proof technique can be used to study other problems beyond IL, such as general non-Markovian stochastic iterative algorithms. Second, our analysis suggests that carefully adding coordination and information sharing may reduce the dependence level of the model while maintaining the scalability of the algorithm. However, theoretically characterizing the trade-off between scalability and optimality is still an open question.

References

[1] 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.

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

[3] Bhandari, J., Russo, D., and Singal, R. (2018). A finite-time analysis of temporal difference learning with linear function approximation. In *Conference On Learning Theory*, pages 1691–1692.

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

[5] 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.

[6] 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.

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

[8] 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.

[9] Chen, Z., Maguluri, S. T., Shakkottai, S., and Shanmugam, K. (2023). A Lyapunov theory for finite-sample guarantees of Markovian stochastic approximation. *Operations Research*.

[10] 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.

[11] 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.

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

[13] 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.

[14] 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.

[15] 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.

[16] 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.

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

[18] 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.

[19] 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.

[20] Gosavi, A. (2006). Boundedness of iterates in Q-learning. *Systems & control letters*, 55(4):347–349.

[21] 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.

[22] Jaakkola, T., Jordan, M. I., and Singh, S. P. (1994). Convergence of stochastic iterative dynamic programming algorithms. In *Advances in neural information processing systems*, pages 703–710.

[23] 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.

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

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

[26] 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.

[27] 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*.

[28] 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.

[29] 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.

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

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

[32] 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.

[33] 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*.

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

[35] Li, G., Cai, C., Chen, Y., Wei, Y., and Chi, Y. (2023). Is Q-learning minimax optimal? a tight sample complexity analysis. *Operations Research*.

[36] 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.

[37] 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.

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

[39] 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.

[40] 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.

[41] 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.

[42] Puterman, M. L. (2014). *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons.

[43] 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.

[44] 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.

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

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

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

[48] 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.

[49] 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.

[50] 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.

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

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

[53] 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.

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

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

[56] 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.

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

[58] 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.

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

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

[61] 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*.

[62] 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.

[63] 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.

[64] 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.

[65] 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.

[66] 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.

[67] 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.

[68] 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.

[69] 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.

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

Appendices

A   Proofs of Theorem 3.1 and Theorem 3.3

It is intractable to directly analyze IL using existing off-the-shelf RL theory since almost all of them require the samples to be Markovian or independent and identically distributed (i.i.d.), which is not the case for local state-actions that are essentially partial observations from the overall MDP model. Fortunately, note that there are some specific connections between IL and stochastic approximation with state aggregation, where an agent maintains an estimated vector with a smaller size than that of the global state space. Therefore, we will take a first step by showing that the update of IQL and ITD can be equivalently formulated as a stochastic approximation algorithm with state aggregation and derive that the corresponding algorithm converges to a fixed point. We further bound the error between the fixed point and the target Q-function on the separable MDP. Finally, we bound the optimality gap between the original and separable MDPs to finish the analysis.

A.1   Connections with Stochastic Approximation with State Aggregation

We follow the proof idea described in Section 3.4 by first deriving the convergence of IQL and the inner loop of INAC, which is ITD. To make this paper self-contained, we present the results of stochastic approximation with state aggregation [37] in Appendix A.5. Here, we show that IQL and ITD are special cases of such a stochastic approximation algorithm.

For stochastic approximation with state aggregation (see Appendix A.5), the agent only maintains a vector of function values the entries of which are abstract states with a much smaller size than that of the original state space. A surjection is defined as a mapping from the original state space to the abstract state space to decide which entry should be updated at each step.

In IQL (cf. Algorithm 1) and ITD (cf. Algorithm 3), each agent cannot estimate the Q-function for entries of global state-action pairs only with visibility of local states and actions. Agent i only maintains a local Q-function $Q_k^i \in \mathbb{R}^{|\mathcal{S}^i||\mathcal{A}^i|}$. Note that the notation of local Q-function is $Q_{t,k}^i$ for ITD in INAC, but the subscription t is omitted here for simplicity (due to the nested-loop structure of the algorithm). Furthermore, agent i uses $Q_k^i(\bar{s}^i, \bar{a}^i)$ to approximate the Q-function of global state-action (s, a) when $(\bar{s}^i, \bar{a}^i) \in \mathcal{S}^i \times \mathcal{A}^i$, $(s, a) \in \mathcal{S} \times \mathcal{A}$ and $(s^i, a^i) = (\bar{s}^i, \bar{a}^i)$. Therefore, IQL and ITD can be both seen as a special form of stochastic approximation with state aggregation in a very natural way. Specifically, define the surjection $h_1 : \mathcal{S} \mapsto \mathcal{S}^i$ and $h_2 : \mathcal{A} \mapsto \mathcal{A}^i$ as

$$h_1(s) = s^i, \qquad h_2(a) = a^i,$$

for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. That means when the global state-action is (s, a), agent i uses the surjections $h_1(s)$ and $h_2(a)$ to decide which local entry should be updated. Formally, define the matrix $\Phi^i \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}^i||\mathcal{A}^i|}$ as

$$\Phi^i(s, a, \bar{s}^i, \bar{a}^i) = \begin{cases} 1, & \text{if } h_1(s) = \bar{s}^i, \ h_2(a) = \bar{a}^i, \\ 0, & \text{otherwise,} \end{cases} \quad \forall (s, a) \in \mathcal{S} \times \mathcal{A}, (\bar{s}^i, \bar{a}^i) \in \mathcal{S}^i \times \mathcal{A}^i.$$

The update of IQL and ITD for any agent $i \in [n]$ can be written as a general stochastic approximation algorithm:

$$Q_{k+1}^i(h_1(S_k), h_2(A_k)) = Q_k^i(h_1(S_k), h_2(A_k))$$
$$+ \alpha_k \left(\left[F^i(\Phi^i Q_k^i) \right] (S_k, A_k) - Q_k^i (h_1(S_k), h_2(A_k)) + w_k^i \right), \qquad (8)$$

and $Q_{k+1}^i(s^i, a^i) = Q_k^i(s^i, a^i)$ for all $(s^i, a^i) \neq (h_1(S_k), h_2(A_k))$. Here in Eq. (8), $F^i : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ is an operator and can be defined differently for IQL and ITD. Note that $\{(S_k, A_k)\}_{k \geq 0}$ is sampled using

different policies, i.e., π_b for IQL and $\pi_{(t)}$ for ITD. For simplicity, here we do not distinguish them in notation with the same assumption (Assumption 3.1) on them.

For IQL, the operator $F^i(\cdot)$ is the Bellman optimality operator, which is defined for any $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ as

$$[F^i(Q)](s,a) = \mathcal{R}^i(s^i, a^i) + \gamma \mathbb{E}_{\bar{s} \sim P_a(s, \cdot)} \left[\max_{\bar{a} \in \mathcal{A}} Q(\bar{s}, \bar{a}) \right], \ \forall (s, a) \in \mathcal{S} \times \mathcal{A}. \tag{9}$$

Consequently, the noise sequence w_k^i for IQL is defined as

$$w_k^i = \mathcal{R}^i(S_k^i, A_k^i) + \gamma \max_{\bar{a}} Q_k^i(h_1(S_{k+1}), h_2(\bar{a})) - [F^i(\Phi^i Q_k^i)](S_k, A_k). \tag{10}$$

For ITD, we define the operator $F^i(\cdot)$ for any $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ as

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

where $\pi = \pi_{(t)}$ is the policy within the t-th outer loop in Algorithm 2. In ITD, the noise sequence w_k^i is defined as

$$w_k^i = \mathcal{R}^i(S_k^i, A_k^i) + \gamma \mathbb{E}_{\bar{a} \sim \pi(\cdot|S_{k+1})}[Q_k^i(h_1(S_{k+1}), h_2(\bar{a}))] - [F^i(\Phi^i Q_k^i)](S_k, A_k). \tag{12}$$

After reformulating IQL and ITD as stochastic approximation with state aggregation, we will present the convergence results of IQL and ITD using [37, Theorem 3.1]. We first introduce the following lemma that provides the required conditions for our results. Note that we only consider the ℓ_∞ norm here. Define $\mathcal{F}_k$ as the σ-algebra generated by $(S_0, A_0, \cdots, S_k, A_k)$.

Lemma A.1. *Suppose that Assumption 3.1 is satisfied. Let $F^i(\cdot)$ be defined as in Eq. (9) (or Eq. (11)), and let w_k^i be defined in Eq. (10) (or Eq. 12). Then, we have the following results.*

(1) The operator $F^i(\cdot)$ is a γ-contraction mapping with respect to $\|\cdot\|_\infty$. In addition, we have $\|F^i(Q)\|_\infty \le \gamma\|Q\|_\infty + 1$ for all $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$.

(2) The random process $\{w_k^i\}$ is measurable with respect to $\mathcal{F}_{k+1}$, and satisfies $\mathbb{E}[w_k^i \mid \mathcal{F}_k] = 0$. In addition, we have $|w_k^i| \le 2/(1-\gamma)$ almost surely for all $k \ge 0$.

Proof of Lemma A.1. (1) The contraction mapping follows from standard results in MDP theory [42]. Now, for any $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, we have

$$\|F^i(Q)\|_\infty \le \|F^i(Q) - F^i(0)\|_\infty + \|F^i(0)\|_\infty \le \gamma\|Q\|_\infty + 1,$$

where the last inequality follows from the fact that $F^i(\cdot)$ is a contraction mapping and $\|F^i(0)\|_\infty \le \max_{(s^i, a^i) \in \mathcal{S}^i \times \mathcal{A}^i} \mathcal{R}^i(s^i, a^i) \le 1$.

(2) The fact that w_k^i is measurable with respect to $\mathcal{F}_k$ follows from the definition of w_k^i. We next show that w_k^i is conditionally mean zero. For IQL, we have

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

$$= \mathcal{R}^i(S_k^i, A_k^i) + \gamma \mathbb{E}_{\bar{s} \sim P_{A_k}(S_k, \cdot)} \left[\max_{\bar{a}} Q_k^i(\bar{s}, \bar{a}) \right] - [F^i(\Phi^i Q_k^i)](S_k, A_k)$$

$$= 0,$$

where the second equality follows from the Markov property and the tower property of conditional expectations. The proof for $\mathbb{E}[w_k^i \mid \mathcal{F}_k] = 0$ in the case of ITD follows from an identical approach.

Finally, using the boundedness of Q-functions in finite MDPs [20], i.e., $\|Q_k^i\|_\infty \le 1/(1-\gamma)$, we have

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

$\square$

The results in Lemma A.1, together with Assumption 3.1, provide the conditions required for applying [37, Theorem 3.1]. Next, we present the convergence results of IQL and ITD. We will show that the stochastic approximation in Eq. (8) converges to the unique fixed point of $F^i(\cdot)$, denoted by $\tilde{Q}_*^i$ for IQL and $\tilde{Q}^i$ for ITD. Specifically, $\tilde{Q}_*^i$ is the unique solution to the following equation:

$$\Pi_1 F^i(\Phi^i x) = x,$$

where $\Pi_1 := ((\Phi^i)^\mathsf{T} D_1 \Phi^i)^{-1} (\Phi^i)^\mathsf{T} D_1$ with $D_1 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|} := \mathrm{diag}(d_{\pi_b})$ and $F^i(\cdot)$ is defined in Eq. (9). As for ITD, $\tilde{Q}^i$ is the unique solution to the equation

$$\Pi_2 F^i(\Phi^i x) = x,$$

where $\Pi_2 := ((\Phi^i)^\mathsf{T} D_2 \Phi^i)^{-1} (\Phi^i)^\mathsf{T} D_2$ with $D_2 \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}||\mathcal{A}|} := \mathrm{diag}(d_{\pi_{(t)}})$ and $F^i(\cdot)$ is defined in Eq. (11). Note that $\tilde{Q}^i$ and $\tilde{Q}_*^i$ are well defined because the operator $\Pi_1 F^i(\Phi^i \cdot)$ and $\Pi_2 F^i(\Phi^i \cdot)$ are also contraction mappings with respect to $\|\cdot\|_\infty$ [37, Proposition C.1]. The convergence results of IQL and ITD are stated below.

Theorem A.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, we have the following convergence bounds.*

(1) For $\{Q_k\}_{k\geq 0}$ generated by IQL (cf. Algorithm 1), we have with probability at least $1 - \delta'/n$ that

$$\left\| Q_K^i - \tilde{Q}_*^i \right\|_\infty \leq \frac{C_a'}{\sqrt{K + k_0}} + \frac{C_b}{K + k_0} = \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{K}}\right),$$

(2) For $\{Q_k\}_{k\geq 0}$ generated by ITD (cf. Algorithm 3), we have with probability at least $1 - \delta'/n$ that

$$\left\| Q_K^i - \tilde{Q}^i \right\|_\infty \leq \frac{C_a'}{\sqrt{K + k_0}} + \frac{C_b}{K + k_0} = \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{K}}\right),$$

where C_a' and C_b were defined in Theorem 3.1.

From Theorem A.1, we see that both IQL and ITD converge, which is not a straightforward result for IL. In the following sections, we will use the results in Theorem A.1 to continue the proof of Theorem 3.1 and Theorem 3.3.

A.2 Proof of Theorem 3.1

Let $Q_k(s,a) = \sum_{i \in [n]} Q_k^i(s^i, a^i)$ for all $(s,a) \in \mathcal{S} \times \mathcal{A}$. Denote the optimal Q-function of agent i under the separable MDP model $\hat{\mathcal{M}}$ as $\hat{Q}_{\hat{\pi}_*}^i$.

Theorem A.2. *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, with probability at least $1 - \delta'$, we have*

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

where the terms C_a', C_b were defined in Theorem 3.1.

Instead of directly considering the approximate convergence to the optimal Q-function of π_* on the original MDP $\mathcal{M}$, we first bound the difference between Q_K and the optimal Q-function on the separable MDP $\hat{\mathcal{M}}$ in Theorem A.2. Our approach follows the roadmap illustrated in Figure 1 to overcome the challenge of studying the convergence of IQL in the original MDP.

Proof of Theorem A.2. We first apply Theorem A.1 to obtain

$$\left\| Q_K^i - \tilde{Q}_*^i \right\|_\infty \leq \frac{C_a'}{\sqrt{K + k_0}} + \frac{C_b}{K + k_0} = \tilde{O}\left(\frac{1}{\sqrt{K}}\right)$$

with probability at least $1 - \delta'/n$. It remains to bound the difference between $\Phi^i \tilde{Q}_*^i$ and $\hat{Q}_{\hat{\pi}_*}^i$.

Lemma A.2 (Proof in Appendix A.4.1). *It holds that* $\|\Phi^i \tilde{Q}_*^i - \hat{Q}_{\hat{\pi}_*}^i\|_\infty \leq \frac{2\gamma\mathcal{E}}{(1-\gamma)^2}$.

Combining the above results, we have with probability at least $1 - \delta'$ that

$$
\begin{aligned}
\left\| Q_K - \hat{Q}_{\hat{\pi}_*} \right\|_\infty &= \max_{s,a} \left| \sum_{i \in [n]} Q_K^i(s^i, a^i) - \sum_{i \in [n]} \hat{Q}_{\hat{\pi}_*}^i(s, a) \right| \\
&\leq \max_{s,a} \sum_{i \in [n]} \left| Q_K^i(s^i, a^i) - \hat{Q}_{\hat{\pi}_*}^i(s, a) \right| \\
&\leq \max_{s,a} \sum_{i \in [n]} \left(\left| Q_K^i(s^i, a^i) - \tilde{Q}_*^i(s^i, a^i) \right| + \left| (\Phi^i \tilde{Q}_*^i)(s, a) - \hat{Q}_{\hat{\pi}_*}^i(s, a) \right| \right) \quad \text{(13a)} \\
&\leq \sum_{i \in [n]} \left(\left\| Q_K^i - \tilde{Q}_*^i \right\|_\infty + \left\| (\Phi^i \tilde{Q}_*^i) - \hat{Q}_{\hat{\pi}_*}^i \right\|_\infty \right) \\
&\leq \sum_{i \in [n]} \left(\frac{C_a'}{\sqrt{K + k_0}} + \frac{C_b}{K + k_0} + \frac{2\gamma\mathcal{E}}{(1-\gamma)^2} \right) \quad \text{(13b)} \\
&= \frac{nC_a'}{\sqrt{K + k_0}} + \frac{nC_b}{K + k_0} + \frac{2n\gamma\mathcal{E}}{(1-\gamma)^2},
\end{aligned}
$$

where Eq. (13a) follows from the fact that $(\Phi^i \tilde{Q}_*^i)(s, a) = \tilde{Q}_*^i(s^i, a^i)$ and Eq. (13b) follows from Lemma A.2 and Theorem A.1. The probability $1 - \delta'$ comes from the union bound. $\square$

Finally, to characterize the difference between the optimal Q-functions of the original MDP and the separable one, we need the following lemma.

Lemma A.3 (Proof in Appendix A.4.2). *Let Q_{π_*} and $\hat{Q}_{\hat{\pi}_*}$ be the optimal Q-functions under the original MDP model $\mathcal{M}$ and the separable one $\hat{\mathcal{M}}$. Then, we have*

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

Using Lemma A.3, 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 proceed and obtain the optimality gap in terms of policies, recall that we defined π_K as the policy greedily induced by Q_K. Denote $a_{K,s} = \max_a Q_K(s, a)$ and $a_{*,s} = \max_a Q_{\pi_*}(s, a)$. When $a_{K,s} \neq a_{*,s}$, using the previous inequality, we have

$$|Q_{\pi_*}(s, a_{K,s}) - Q_{\pi_*}(s, \pi_{*,s})| \leq \frac{2nC_a'}{\sqrt{K + k_0}} + \frac{2nC_b}{K + k_0} + \frac{8n\gamma\mathcal{E}}{(1-\gamma)^2}.$$

20

Define the advantage function as $A_\pi(s, a) = Q_\pi(s, a) - V_\pi(s)$ for any policy π and $(s, a) \in \mathcal{S} \times \mathcal{A}$. Then, we have

$$
\begin{aligned}
|A_{\pi_*}(s, a_{K,s})| &= |Q_{\pi_*}(s, a_{K,s}) - V_{\pi_*}(s)| \\
&= |Q_{\pi_*}(s, a_{K,s}) - Q_{\pi_*}(s, a_{*,s})| \\
&\leq \frac{2nC_a'}{\sqrt{K + k_0}} + \frac{2nC_b}{K + k_0} + \frac{8n\gamma\mathcal{E}}{(1 - \gamma)^2}.
\end{aligned}
$$

Using the performance difference lemma [24], with probability at least $1 - \delta'$, we have for any initial state distribution μ that

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

where $o_{\pi_K}^\mu$ is the occupancy measure defined as $o_{\pi_K}^\mu(s) = (1 - \gamma)\mathbb{E}_{\pi_K}[\sum_{k=0}^\infty \gamma^t \mathbb{P}(S_k = s \mid S_0 \sim \mu)]$ for all $s \in \mathcal{S}$. The proof is complete.

A.3 Proof of Theorem 3.3

To prove Theorem 3.3, we first by analyzing the convergence error of the critic in Appendix A.3.1. Then we analyze the convergence rate of the actor in Appendix A.3.2. Finally, we combine the analysis of the actor and the critic to finish the proof.

A.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.

Algorithm 3 Inner Loop of Algorithm 2

1: **Input:** Integer K, policy $\pi^i := \pi_{\theta^i}^i$ from the outer loop, and initialization $Q_0^i = 0$.
2: **for** $k = 0, 1, 2, \cdots, K - 1$ **do**
3:     Implement $A_k^i \sim \pi^i(\cdot \mid S_k^i)$ (simultaneously with all other agents), and observes S_{k+1}^i.
4:     Update Q-function:
5:     $Q_{k+1}^i(S_k^i, A_k^i) = (1 - \alpha_k)Q_k^i(S_k^i, A_k^i) + \alpha_k(\mathcal{R}^i(S_k^i, A_k^i) + \gamma \mathbb{E}_{\bar{a}^i \sim \pi^i(\cdot | S_{k+1}^i)} Q_k^i(S_{k+1}^i, \bar{a}^i))$
6: **end for**

Next, we provide the approximate convergence of the critic.

Theorem A.3. *Consider $\{Q_k\}_{k \geq 0}$ generated by Algorithm 3 with input policy $\pi = (\pi^1, \pi^2, \cdots, \pi^n)$. 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, with probability at least $1 - \delta'$, we have*

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

where the constants C_a' and C_b were defined in Theorem 3.1.

Proof of Theorem A.3. We have verified that ITD is a special case of the stochastic approximation algorithm with state aggregation. Therefore, apply Theorem A.1 and we get

$$\left\|Q_K^i - \tilde{Q}^i\right\|_\infty \leq \frac{C_a'}{\sqrt{K + k_0}} + \frac{C_b}{K + k_0} = \tilde{\mathcal{O}}\left(\frac{1}{\sqrt{K}}\right)$$

Our next step is to bound the gap between $\Phi^i \tilde{Q}^i$ and $\hat{Q}_\pi^i$.

Lemma A.4 (Proof in Appendix A.4.3). *It holds that $\|\Phi^i \tilde{Q}^i - \hat{Q}_\pi^i\|_\infty \leq \frac{2\gamma\mathcal{E}}{(1-\gamma)^2}$.*

It follows from Lemma A.4 and Theorem A.3 that with probability at least $1 - \delta'$, we have

$$
\begin{aligned}
\|Q_K - \hat{Q}_\pi\|_\infty &= \max_{s,a} \left| \sum_{i \in [n]} Q_K^i(s^i, a^i) - \sum_{i \in [n]} \hat{Q}_\pi^i(s, a) \right| \\
&\leq \max_{s,a} \sum_{i \in [n]} \left| Q_K^i(s^i, a^i) - \hat{Q}_\pi^i(s, a) \right| \\
&\leq \max_{s,a} \sum_{i \in [n]} \left(\left| Q_K^i(s^i, a^i) - \tilde{Q}^i(s^i, a^i) \right| + \left| (\Phi^i \tilde{Q}^i)(s, a) - \hat{Q}_\pi^i(s, a) \right| \right) \quad (14a) \\
&\leq \sum_{i \in [n]} \left(\left\| Q_K^i(s^i, a^i) - \tilde{Q}^i(s^i, a^i) \right\|_\infty + \left\| (\Phi^i \tilde{Q}^i)(s, a) - \hat{Q}_\pi^i(s, a) \right\|_\infty \right) \\
&\leq \sum_{i \in [n]} \left(\frac{C_a'}{\sqrt{K + k_0}} + \frac{C_b}{K + k_0} + \frac{2\gamma\mathcal{E}}{(1-\gamma)^2} \right) \quad (14b) \\
&= \frac{n C_a'}{\sqrt{K + k_0}} + \frac{n C_b}{K + k_0} + \frac{2n\gamma\mathcal{E}}{(1-\gamma)^2},
\end{aligned}
$$

where Eq. (14a) follows from the fact that $(\Phi^i \tilde{Q}^i)(s, a) = \tilde{Q}^i(s^i, a^i)$ and Eq. (14b) follows from Lemma A.4 and Theorem A.3. The probability $1 - \delta'$ comes from the union bound. □

A.3.2 Analysis of the Actor

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

Algorithm 4 Outer Loop of Algorithm 2

1: **Input:** Integer T and initialization $\theta_0^i = 0$.
2: **for** $t = 0, 1, 2 \cdots, T - 1$ **do**
3: $\quad \theta_{t+1}^i = \theta_t^i + \eta_t Q_{t,K}^i$
4: **end for**

The following theorem characterizes the convergence rates of Algorithm 4.

Theorem A.4. *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 3.3. Then we have*

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

Proof of Theorem A.4. We begin by decomposing the optimality gap in the following way:

$$\|Q_{\pi_*} - Q_{(t)}\|_\infty \leq \underbrace{\|\hat{Q}_{\hat{\pi}_*} - \hat{Q}_{(t)}\|_\infty}_{v_1} + \underbrace{\|Q_{\pi_*} - \hat{Q}_{\hat{\pi}_*}\|_\infty}_{v_2} + \underbrace{\|\hat{Q}_{(t)} - Q_{(t)}\|_\infty}_{v_3}, \tag{15}$$

where we recall that $\hat{\pi}_*$ denotes the optimal policy of the separable MDP $\hat{\mathcal{M}}$. On the RHS of Eq. (15), the term v_1 captures the optimality gap of the output of Algorithm 4 with respect to the separable MDP model $\hat{\mathcal{M}}$, and the terms v_2 and v_3 are both induced by the model difference between the original MDP $\mathcal{M}$ and the separable one $\hat{\mathcal{M}}$.

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

$$\theta_{t+1} = \theta_t + \eta_t Q_{t,K}, \tag{16}$$

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 [7] provides us the following result.

Lemma A.5. *Consider* $\{\pi_{(t)} = (\pi_{(t)}^1, \pi_{(t))}^2,, \cdots, \pi_{(t)}^n)\}_{t \geq 0}$ *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 A.5, we need to verify that the stepsizes in Theorem 3.3 satisfy the condition in Lemma A.5. Using Eq. (16), we have for any state-action pair (s,a) that

$$(\theta_t)_{s,a} = (\theta_{t-1})_{s,a} + \eta_{t-1} Q_{t-1,K}(s,a) = (\theta_0)_{s,a} + \sum_{j=0}^{t-1} \eta_j Q_{j,K}(s,a) \leq \sum_{j=0}^{t-1} \eta_j \frac{n}{1-\gamma} = \bar{Q} \sum_{j=0}^{t-1} \eta_j,$$

where $\bar{Q} := \frac{n}{1-\gamma}$ and the last inequality follows from the boundedness of the Q-functions in finite MDPs [20], that is,

$$Q_{t,K}(s,a) = \sum_{i=1}^n Q_{t,K}^i(s^i, a^i) \leq \sum_{i=1}^n \|Q_{t,K}^i\|_\infty \leq n/(1-\gamma).$$

The previous inequality allows us to derive a lower bound for $\pi_{(t)}(a \mid s)$:

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

23

It follows that

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

Therefore, the condition on the stepsizes in Lemma A.5 is satisfied. Now, using Lemma A.5 for the term v_1 on the RHS of Eq. (15), we have

$$v_1 = \|\hat{Q}_{\hat{\pi}_*} - \hat{Q}_{(t)}\|_\infty$$

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

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

where the last step 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 in Eq. (15). Using Lemma A.3, we bound the term v_2 in Eq. (15) as

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

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

Lemma A.6 (Proof in Appendix A.4.4). *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*

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

Using Lemma A.6, we have

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

Theorem A.4 follows from using the upper bounds we derived for the terms v_1, v_2, and v_3 in Eq. (15). □

In light of the analyses for the actor and the critic, to finish proving Theorem 3.3, we combine the bounds in Theorem A.3 and Theorem A.4. Specifically, for any $j \leq t-1$, with probability at least $1 - \delta'$, we have

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

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

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

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

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

$$\leq \frac{1}{1-\gamma}\left(\frac{nC_a}{\sqrt{K+k_0}} + \frac{nC_b}{K+k_0} + \frac{2n\gamma\mathcal{E}}{(1-\gamma)^2}\right).$$

Applying the bound in the above inequality to Theorem A.4 completes the proof of Theorem 3.3.

24

A.4 Proof of All Supporting Lemmas

A.4.1 Proof of Lemma A.2

Given a positive definite matrix D, define its corresponding weighted ℓ_2-norm as $\|x\|_D = (x^{\mathsf{T}} D x)^{1/2}$. 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 $\{\Phi^i r \mid r \in \mathbb{R}^{|\mathcal{S}^i||\mathcal{A}^i|}\}$ with respect to $\|\cdot\|_{D_1}$. Therefore, for any $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, we have:

$$\Phi^i \Pi_1 Q = \arg\min_{q \in \{\Phi^i r \mid r \in \mathbb{R}^{|\mathcal{S}^i||\mathcal{A}^i|}\}} \|Q - q\|_{D_1}. \tag{17}$$

We first show that $\Phi^i \Pi_1(\cdot)$ is nonexpansive with respect to $\|\cdot\|_\infty$.

Lemma A.7. *It holds that* $\|\Phi^i \Pi_1 Q\|_\infty \leq \|Q\|_\infty$ *for any* $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$.

Proof of Lemma A.7. From Eq. (17), we have

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

For any $(\bar{s}, \bar{a})$ in $\mathcal{S} \times \mathcal{A}$, the value of $r_*(\bar{s}^i, \bar{a}^i)$ must satisfy

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

which leads to

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

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

To proceed, recall the definition of $\hat{Q}_\pi^i$:

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

where $\hat{\mathbb{E}}_\pi[\cdot]$ denotes the expectation with respect to the separable MDP transition kernel $\hat{\mathcal{P}}$. Observe that $\hat{Q}_\pi^i(s_1, a_1) = \hat{Q}_\pi^i(s_2, a_2)$ when $(s_1, a_1), (s_2, a_2) \in \mathcal{S} \times \mathcal{A}$ and $(s_1^i, a_1^i) = (s_2^i, a_2^i)$. Therefore, it follows from Lemma A.7 that $\Phi^i \Pi_1 \hat{Q}_\pi^i = \hat{Q}_\pi^i$. In addition, since

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

we have

$$
\begin{aligned}
\|\Phi^i \tilde{Q}_*^i - \hat{Q}_{\hat{\pi}_*}^i\|_\infty &\leq \frac{1}{1-\gamma} \|F_*^i(\hat{Q}_{\hat{\pi}_*}^i) - \hat{Q}_{\hat{\pi}_*}^i\|_\infty \\
&= \frac{1}{1-\gamma} \max_{(s,a)} \left| \gamma \mathbb{E}_{\bar{s} \sim P_a(s,\cdot)} \max_{\bar{a}} \hat{Q}_{\hat{\pi}_*}^i(\bar{s}, \bar{a}) - \gamma \mathbb{E}_{\bar{s} \sim \hat{P}_a(s,\cdot)} \max_{\bar{a}} \hat{Q}_{\hat{\pi}_*}^i(\bar{s}, \bar{a}) \right|
\end{aligned}
$$

$$\leq \frac{\gamma}{1-\gamma} \max_{(s,a)} \sum_{\bar{s}\in\mathcal{S}} \left| P_a(s,\bar{s}) - \hat{P}_a(s,\bar{s}) \right| \max_{\bar{a}\in\mathcal{A}} \hat{Q}^i_{\hat{\pi}_*}(\bar{s},\bar{a})$$

$$\leq \frac{\gamma}{(1-\gamma)^2} \max_{(s,a)} \sum_{\bar{s}\in\mathcal{S}} \left| P_a(s,\bar{s}) - \hat{P}_a(s,\bar{s}) \right|$$

$$\leq \frac{2\gamma\mathcal{E}}{(1-\gamma)^2},$$

where the last two inequalities follow from $\max_{\bar{a}\in\mathcal{A}} \hat{Q}^i_{\hat{\pi}_*}(\bar{s},\bar{a}) \leq \|\hat{Q}^i_{\hat{\pi}_*}\|_\infty \leq \frac{1}{1-\gamma}$ and the definition of dependence level $\mathcal{E}$.

A.4.2 Proof of Lemma A.3

Let $\mathcal{T}_* : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \mapsto \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ 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

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

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

$$\left\|Q_{\pi_*} - \hat{Q}_{\hat{\pi}_*}\right\|_\infty \leq \frac{1}{1-\gamma} \left\|\mathcal{T}_* \hat{Q}_{\hat{\pi}_*} - \hat{Q}_{\hat{\pi}_*}\right\|_\infty. \tag{18}$$

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

$$\left|\hat{Q}_{\hat{\pi}_*}(s,a) - (\mathcal{T}_* \hat{Q}_{\hat{\pi}_*})(s,a)\right|$$

$$\leq \left|\gamma \sum_{s'\in\mathcal{S}} \hat{P}_a(s,s') \max_{a'} \hat{Q}_{\hat{\pi}_*}(s',a') - \gamma \sum_{s'\in\mathcal{S}} P_a(s,s') \max_{a'} \hat{Q}_{\hat{\pi}_*}(s',a')\right|$$

$$\leq 2\gamma \left\|\hat{P}_a(s,\cdot) - P_a(s,\cdot)\right\|_{\text{TV}} \frac{n}{1-\gamma} \tag{19a}$$

$$\leq \frac{2n\gamma\mathcal{E}}{1-\gamma}, \tag{19b}$$

where Eq.(19b) follows from the definition of $\mathcal{E}$, and Eq. (19a) follows from the following equivalent definition of the total variation distance [34]:

$$\|\nu_1 - \nu_2\|_{\text{TV}} = \frac{1}{2} \sup_{f:\|f\|_\infty\leq 1} \left|\int f\, d\nu_1 - \int f\, d\nu_2\right|.$$

Substituting the above result into Eq. (18), we have

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

A.4.3 Proof of Lemma A.4

Recall the definition of $\| \cdot \|_D$ in the proof of Lemma A.2 in Appendix A.4.1. 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 $\{\Phi^i r \mid r \in \mathbb{R}^{|\mathcal{S}^i||\mathcal{A}^i|}\}$. Then, for any $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, we have the following:

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

Next, we present a nonexpansive property of the operator $\Phi^i \Pi_2(\cdot)$.

Lemma A.8. *It holds that $\|\Phi^i \Pi_2 Q\|_\infty \leq \|Q\|_\infty$ for any $Q \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$.*

The proof of Lemma A.8 is similar to the proof of Lemma A.7, and is omitted here.

Observe that $\hat{Q}^i_\pi(s_1, a_1) = \hat{Q}^i_\pi(s_2, a_2)$ for any policy π when $(s_1, a_1), (s_2, a_2) \in \mathcal{S} \times \mathcal{A}$ and $(s_1^i, a_1^i) = (s_2^i, a_2^i)$. Then, it follows from Lemma A.8 that $\Phi^i \Pi_2 \hat{Q}^i_\pi = \hat{Q}^i_\pi$. In addition, note that we have

$$
\begin{aligned}
\|\Phi^i \tilde{Q}^i - \hat{Q}^i_\pi\|_\infty &= \|\Phi^i \tilde{Q}^i - \Phi^i \Pi_2 \hat{Q}^i_\pi\|_\infty \\
&= \|\Phi^i \Pi_2 F^i(\Phi^i \tilde{Q}^i) - \Phi^i \Pi_2 \hat{Q}^i_\pi\|_\infty &\text{(20a)} \\
&\leq \|F^i(\Phi^i \tilde{Q}^i) - \hat{Q}^i_\pi\|_\infty &\text{(20b)} \\
&\leq \|F^i(\Phi^i \tilde{Q}^i) - F^i(\hat{Q}^i_\pi)\|_\infty + \|F^i(\hat{Q}^i_\pi) - \hat{Q}^i_\pi\|_\infty \\
&\leq \gamma \|\Phi^i \tilde{Q}^i - \hat{Q}^i_\pi\|_\infty + \|F^i(\hat{Q}^i_\pi) - \hat{Q}^i_\pi\|_\infty, &\text{(20c)}
\end{aligned}
$$

where Eq. (20a) follows from the fact that $\tilde{Q}^i$ is the solution of equation $\Pi_2 F^i(\Phi^i r) = r$, Eq. (20b) follows from Lemma A.8, and the last inequality Eq. (20c) follows from the fact that $F^i(\cdot)$ is a γ-contraction in $\| \cdot \|_\infty$. Combined with the Bellman equation, the above inequality leads to

$$
\begin{aligned}
\|\Phi^i \tilde{Q}^i - \hat{Q}^i_\pi\|_\infty &\leq \frac{1}{1-\gamma} \|F^i(\hat{Q}^i_\pi) - \hat{Q}^i_\pi\|_\infty \\
&= \frac{1}{1-\gamma} \max_{(s,a)} \left| \gamma \mathbb{E}_{\bar{s} \sim P_a(s,\cdot), \bar{a} \sim \pi(\cdot|\bar{s})} [\hat{Q}^i_\pi(\bar{s}, \bar{a})] - \gamma \mathbb{E}_{\bar{s} \sim \hat{P}_a(s,\cdot), \bar{a} \sim \pi(\cdot|\bar{s})} [\hat{Q}^i_\pi(\bar{s}, \bar{a})] \right| \\
&\leq \frac{\gamma}{1-\gamma} \max_{(s,a)} \sum_{\bar{s} \in \mathcal{S}} \left| P_a(s, \bar{s}) - \hat{P}_a(s, \bar{s}) \right| \sum_{\bar{a} \in \mathcal{A}} \pi(\bar{a} \mid \bar{s}) \hat{Q}^i_\pi(\bar{s}, \bar{a}) \\
&\leq \frac{\gamma}{(1-\gamma)^2} \max_{(s,a)} \sum_{\bar{s} \in \mathcal{S}} \left| P_a(s, \bar{s}) - \hat{P}_a(s, \bar{s}) \right| \\
&\leq \frac{2\gamma \mathcal{E}}{(1-\gamma)^2},
\end{aligned}
$$

where the last two inequalities follow from $\sum_{\bar{a} \in \mathcal{A}} \pi(\bar{a} \mid \bar{s}) \hat{Q}^i_\pi(\bar{s}, \bar{a}) \leq \|\hat{Q}^i_\pi\|_\infty \leq \frac{1}{1-\gamma}$ and the definition of dependence level $\mathcal{E}$.

A.4.4 Proof of Lemma A.6

Let $\mathcal{T}_\pi : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \mapsto \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ 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 $\mathcal{T}_\pi$ and $\hat{\mathcal{T}}_\pi$ are contractive operators with respect to $\| \cdot \|_\infty$, with

27

a common contraction factor γ [52]. 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

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

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

$$
\begin{aligned}
&\left\|Q_\pi - \hat{Q}_\pi\right\|_\infty \\
&\leq \frac{1}{1-\gamma}\left\|\mathcal{T}_\pi\hat{Q}_\pi - \hat{Q}_\pi\right\|_\infty \\
&= \frac{1}{1-\gamma}\left\|\mathcal{T}_\pi\hat{Q}_\pi - \hat{\mathcal{T}}_\pi\hat{Q}_\pi\right\|_\infty \\
&= \frac{\gamma}{1-\gamma}\max_{(s,a)}\left|\sum_{s'} P_a(s,s')\mathbb{E}_{a'\sim\pi(\cdot|s')}[\hat{Q}_\pi(s',a')] - \sum_{s'}\hat{P}_a(s,s')\mathbb{E}_{a'\sim\pi(\cdot|s')}[\hat{Q}_\pi(s',a')]\right| && \text{(21a)} \\
&\leq \frac{2n\gamma}{(1-\gamma)^2}\left\|\hat{P}_a(s,\cdot) - P_a(s,\cdot)\right\|_{\text{TV}} && \text{(21b)} \\
&\leq \frac{2n\gamma\mathcal{E}}{(1-\gamma)^2}, && \text{(21c)}
\end{aligned}
$$

where Eq. (21a) follows from the definition of $\mathcal{T}_\pi$ and $\hat{\mathcal{T}}_\pi$, Eq. (21b) also follows from the equivalent definition of the total variation distance, and Eq. (21c) follows from the definition of $\mathcal{E}$.

A.5 Stochastic Approximation with State Aggregation

In this section, we present the results of stochastic approximation with state aggregation in [37].

Let $\mathcal{N} = \{1, 2, \cdots, n\}$ be the state space of $\{i_k\}$, where $\{i_k\}_{k=0}^\infty$ is the sequence of states visited by a Markov chain. Let $\mathcal{L} = \{1, 2, \cdots, l\}, (l \leq n)$ be the abstract state space. The surjection $h : \mathcal{N} \mapsto \mathcal{L}$ is used to convert every state in $\mathcal{N}$ to its abstraction set $\mathcal{L}$. Given parameter $x \in \mathbb{R}^l$ and function $F : \mathbb{R}^n \to \mathbb{R}^l$, we consider the generalized stochastic approximation that updates $x(k) \in \mathbb{R}^l$ starting from $x(0) = 0$:

$$
x_{h(i_k)}(k+1) = x_{h(i_k)}(k) + \alpha_k\left(F_{i_k}(\Phi x(k)) - x_{h(i_k)}(k) + w(k)\right), \tag{22}
$$
$$
x_j(k+1) = x_j(k), \text{for } j \neq h(i_k), j \in \mathcal{L}, \tag{23}
$$

where the matrix $\Phi \in \mathbb{R}^{n\times l}$ is given by

$$
\Phi_{ij} = \begin{cases} 1, & \text{if } h(i) = j, \\ 0, & \text{otherwise,} \end{cases} \forall i \in \mathcal{N}, j \in \mathcal{L}.
$$

Before presenting the convergence result, we first give the assumptions required for the theorem in [37].

Assumption A.1. *The random process $\{i_k\}_{k\geq 0}$ is an aperiodic and irreducible Markov chain on the space $\mathcal{S} \times \mathcal{A}$ with stationary distribution $d = \{d_1, d_2, \cdots, d_n\}$. In addition, 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 constants $M_1 \geq 0$ and $M_2 \geq 1$ such that*

$$\sup_{\mathcal{K} \subseteq \mathcal{N}} \left| \sum_{i \in \mathcal{K}} d_i - \sum_{i \in \mathcal{K}} \mathbb{P}(i_k = i \mid i_0 = j) \right| \leq M_1 \exp(-k/M_2) \tag{24}$$

for all $j \in \mathcal{L}$ and $k \geq 0$.

Assumption A.2. *The Operator $F(\cdot)$ is a γ-contraction with respect to $\|\cdot\|_\infty$. Further, there exists some constant $C > 0$ such that $\|F(x)\|_\infty \leq \gamma\|x\|_\infty + C$ for all $x \in \mathbb{R}^n$.*

Assumption A.3. *The random process $w(k)$ is $\mathcal{F}_{k+1}$ measurable and satisfies $\mathbb{E}[w(k) \mid \mathcal{F}_k] = 0$. Further, $|w(k)| \leq \bar{w}$ almost surely for some constant $\bar{w}$.*

The following theorem shows that the stochastic approximation presented in Eq. (22) converges to the unique x^* that solves

$$\Pi F(\Phi x^*) = x^*.$$

where $\Pi := (\Phi^\mathsf{T} D \Phi)^{-1} \Phi^\mathsf{T} D$. Here, D is the diagonal matrix with diagonal components being the stationary distribution of the Markov chain $\{i_k\}$, which we denote by d.

Theorem A.5. *Suppose that Assumptions A.1, A.2, A.3 hold. Further, assume that there exists some constant $\bar{x} \geq \|x^*\|_\infty$ such that $\|x(k)\|_\infty \leq \bar{x}$ for all k almost surely. Let the stepsize be $\alpha_k = \frac{H}{k+k_0}$ with $k_0 = \max(4H, 2M_2 \log K)$ and $H \geq \frac{2}{\sigma'(1-\gamma)}$. Define the constants $C_1 := 2\bar{x}+C+\bar{w}, C_2 := 4\bar{x}+2C+\bar{w}, C_3 := 2M_1(2\bar{x}+C)(1+2M_2+4H)$. Then, with probability at least $1 - \delta'$, we have*

$$\|x(K) - x^*\|_\infty \leq \frac{C'_a}{\sqrt{K+k_0}} + \frac{C_b}{K+k_0} = \tilde{O}\left(\frac{1}{\sqrt{K}}\right),$$

where

$$C'_a = \frac{4HC_2}{1-\gamma}\sqrt{K_2 \log K \left(\log\left(\frac{4lM_2 K}{\delta'}\right) + \log\log K\right)},$$

$$C_b = 4\max\left\{\frac{48M_2 C_1 H \log K + \sigma' C_3}{(1-\gamma)\sigma'}, \frac{2\bar{x}(2M_2 \log K + k_0)}{1-\gamma}\right\}.$$

A.6 Extensions

Our proof follows from the following blueprint: (1) constructing a separable MDP to approximate the original MDP, (2) analyzing the algorithm as if it were implemented on the separable MDP, and (3) bounding the dap due to model difference between the separable MDP and the original one. This framework can be used to study other learning scenarios beyond IL considered in this paper. We briefly elaborate on some of them in the following.

Partially Observable Markov Decision Processes (POMDPs)   In sequential decision-making problems, sometimes the agent (or agents) is not able to observe the complete state and must work with a partial observation of the state [38]. More specifically, instead of observing a trajectory of state-actions $\{(S_k, A_k)\}$, the agent actually observes $\{(O_k, A_k)\}$, where $O_k \in \mathcal{O}$ denotes the partial observation at time k. In this case, unlike $\{(S_k, A_k)\}$, the trajectory $\{(O_k, A_k)\}$ does *not* form a Markov chain, which imposes major technical difficulties in analyzing POMDPs. In light of our analysis framework, if we can approximate the non-Markovian trajectory $\{(O_k, A_k)\}$ by a Markov chain that is defined on $\mathcal{O} \times \mathcal{A}$, then, it would enable a tractable analysis of POMDPs. In this case, the approximation error (which is analogous to our dependence level) will appear in the analysis. Studying POMDPs is another future direction of this work.

General Non-Markovian Stochastic Iterative Algorithms   RL algorithms, in general, belong to the broad framework of stochastic iterative algorithms, also known as stochastic approximation algorithms [46], which can be represented in the following form

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

where Y_k is the noise and $G(\cdot, \cdot)$ is some properly defined operator, depending on the problem of interest. As we see in Section A.3.1, the update equation for the joint Q-function can be written as a stochastic approximation in the form above. Moreover, the popular stochastic gradient descent algorithm in large-scale continuous optimization is also a special case of stochastic approximation, where $G(\cdot, \cdot)$ is a noisy version of the negative gradient operator. While stochastic approximation has been extensively studied in the literature, most if not all existing results assume the noise sequence $\{Y_k\}$ is either an i.i.d. sequence or forms a Markov chain. In light of our proof framework, we are able to analyze stochastic approximation algorithms driven by non-Markovian samples, as long as the trajectory $\{Y_k\}$ can be approximated by a Markov chain. Rigorously developing asymptotic and finite-sample convergence results for non-Markovian stochastic approximation algorithms is also an interesting future direction.

B   Examples and Experiments

This section provides the details of the illustrative example presented in Section 2.2 and the numerical simulations.

B.1   Details of the Synthetic MDP

Original MDP.   The state of the three agents is a 3-dimension vector denoted as $s = (s^1, s^2, s^3)$, where $s^i \in \{0, 1\}$ for $i \in \{1, 2, 3\}$. The action is denoted as $a = (a^1, a^2, a^3)$, where $a^i \in \{0, 1\}$ for $i \in \{1, 2, 3\}$. For each agent, one unit of reward can be obtained when the value of the state variable stays unchanged and 0 is obtained otherwise. Agents 1 and 2 are strongly coupled and always have the same state value. Their state values remain unchanged when they take the same actions. Otherwise, their state values will transit from 0 to 1 or from 1 to 0 simultaneously. That means the following holds when $s_k^1 = s_k^2$,

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

For agent 3, any action leads to the state value 0 or 1 with probability 0.5 when (1) its state value is 1 or (2) its state value is 0 and it applies the same action to agent 2. However, when its state value is 0 and it applies a different action compared with agent 2, the next state will be 1 with probability 1. The transition model for agent 3 is summarized as

$$\mathbb{P}(s_{k+1}^3 = 0 \mid s_k^3 = 0, a_k^2 \neq a_k^3) = 0, \qquad \mathbb{P}(s_{k+1}^3 = 1 \mid s_k^3 = 0, a_k^2 \neq a_k^3) = 1,$$

$$\mathbb{P}(s_{k+1}^3 = 0 \mid s_k^3 = 0, a_k^2 = a_k^3) = 0.5, \qquad \mathbb{P}(s_{k+1}^3 = 1 \mid s_k^3 = 0, a_k^2 = a_k^3) = 0.5,$$
$$\mathbb{P}(s_{k+1}^3 = 0 \mid s_k^3 = 1) = 0.5, \qquad \mathbb{P}(s_{k+1}^3 = 1 \mid s_k^3 = 1) = 0.5.$$

The different dependencies among agents induce different dependence levels with different grouping options.

The Dependence level. Suppose that we group agents 1 and 2 together, that is, they can be seen as one agent controlling 2 action variables. Then, it is easy to derive $\mathcal{E} = 0.5$ by solving the optimization problem in Eq. (1). Suppose that agents 2 and 3 are grouped together. Then, we have $\mathcal{E} = 0.75$. Similarly, we have $\mathcal{E} = 0.875$ when agents 1 and 3 are grouped together. The detailed separable MDPs for different grouping options are shown in Figure 4.
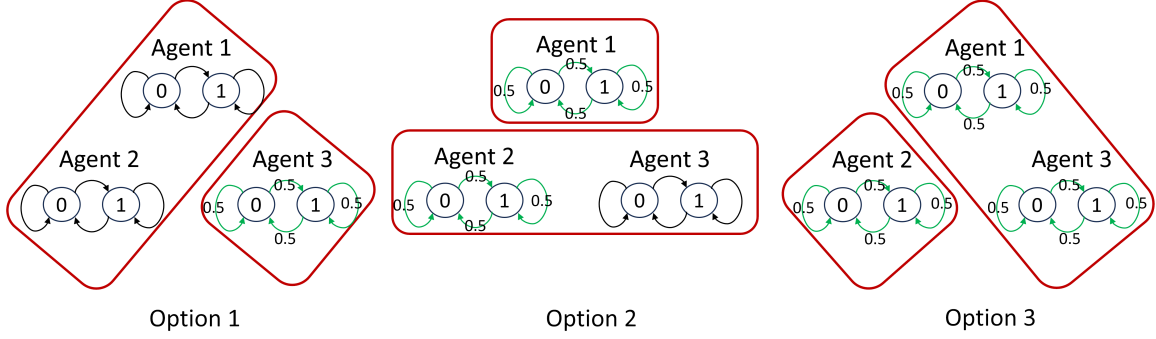

Figure 4: The separable MDPs for different grouping options

In Figure 4, black arrows mean the same transition probabilities as the original MDP, and green arrows mean random transition probabilities with 0.5 to any state under any state and action.

B.2   The 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 5, 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 5, 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.

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
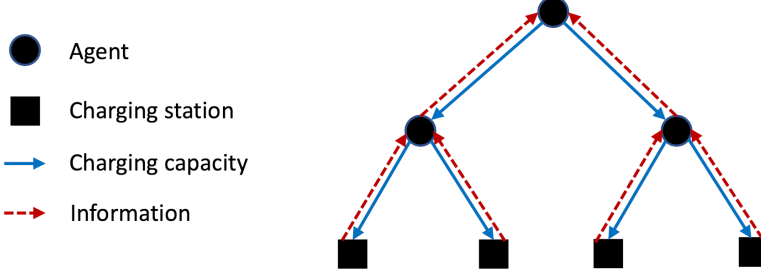
Figure 5: Illustration of a binary tree to allocate the charging capacity.

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 Section B.3 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.

B.3   Numerical Simulations

In this section, we present the details of our numerical experiments.

B.3.1 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 B.1.

B.3.2 Experimental Setting for EV Charging
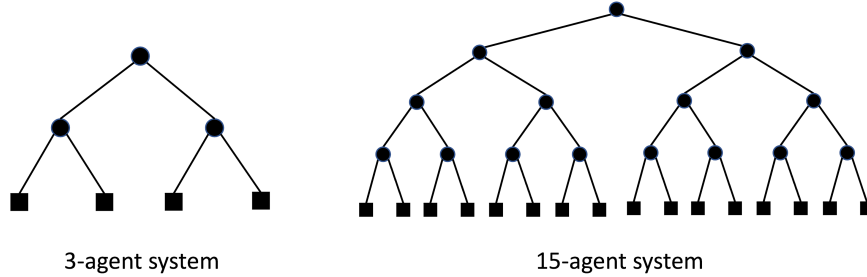

3-agent system   15-agent system

Figure 6: 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 6. 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 $\mathbf{W}_1^i(t) = \mathrm{diag}(1, e_l^i(t), e_l^i(t), e_l^i(t))$ and $\mathbf{W}_2^i(t) = \mathrm{diag}(1, e_r^i(t), e_r^i(t), e_r^i(t))$, where

$$e_l^i(t) = \frac{p^{C_l^i}(t)}{p^{C_l^i}(t) + p^{C_r^i}(t)}, \text{ and } e_r^i(t) = \frac{p^{C_r^i}(t)}{p^{C_l^i}(t) + p^{C_r^i}(t)}.$$

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

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

Table 2: Parameters of INAC for 3-agnet 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

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 [48], but 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., $a^i(t) = e^i_l(t)$.

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 multiscale 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.

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 multiscale 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 Guassian noise to the action with mean 0 and decreasing variance with steps. The detailed parameters are shown in Table 2.

In our simulations, both IQL and INAC are trained and tested for 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.

B.3.3 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 7 and Figure 8. The rewards are also normalized by the average optimal reward calculated by the offline optimal policy. 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.
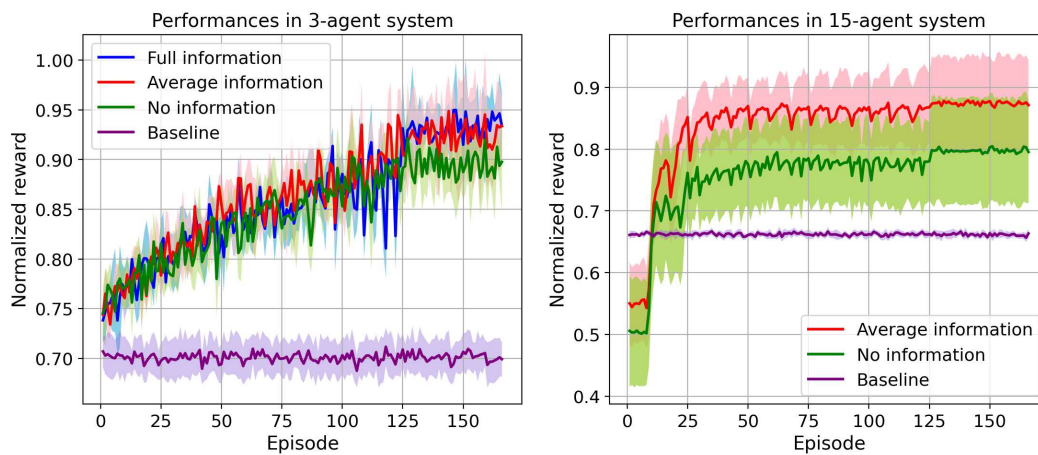

Figure 7: Performance of IQL with different contractions in the 3-agent (left) and 15-agent (right) systems.
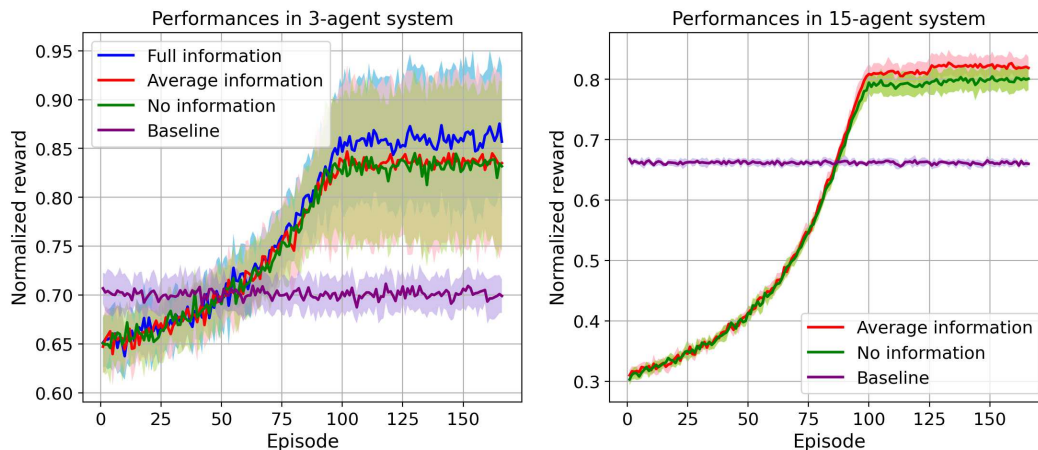

Figure 8: Performance of INAC with different contractions in the 3-agent (left) and 15-agent (right) systems.

In the 15-agent system, 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 are both around 80% of the optimal rewards, where the optimal rewards are calculated by an offline optimal algorithm (see Appendix B.3), and the former is slightly better. The performance of IQL with average information can significantly outperform others.