Decentralized Constrained Policy Optimization for Safe Multi-agent Reinforcement Learning

Anonymous Authors[1]

Abstract

A challenging problem in seeking to bring multi-agent reinforcement learning (MARL) techniques into real-world applications, such as autonomous driving and drone swarms, is how to control multiple agents cooperatively and safely to accomplish tasks. Most existing safe MARL methods learn the centralized value function by introducing a global state to promote safety cooperation. However, the global coupling arising from agents' safety constraints and the exponential growth of the state-action space size limit the applicability of these methods in instant communication or computing resource-constrained systems and larger multi-agent systems. In this paper, we develop a novel decentralized and theoretically-justified policy optimization method to meet joint policy improvement and safety guarantee, which adopts a sequential update scheme to optimize κ-hop policies under a form of correlation decay property. Further, a practical algorithm for decentralized safe MARL, called decentralized MAPPO-Lagrangian (DEC-MAPPO-L), is proposed. The effectiveness of the proposed method is verified on a collection of benchmark tasks, and the results support our theory that decentralized training with local interactions can still meet performance improvement and safe constraints.

1. Introduction

With the advanced and rapid developments of reinforcement learning technology, a growing number of researchers have gradually shifted their focus from virtual simulation to real-world cyber-physical applications, which inevitably face safety challenges, especially in multi-agent systems (Brunke et al., 2022; Gu et al., 2022b). Such safety challenges generally exist in safety-critical scenarios such as autonomous vehicle navigation (Zhou et al., 2022), power grids (Cui et al., 2022), and drone swarms (Chen et al., 2020), in which agents perform complex cooperative tasks while meeting a variety of local and system-wide limitations or constraints. These constraints can be derived from domain-specific knowledge and are intended to prevent damage to people or other elements in the environment, e.g., equipment and infrastructure, or to prevent the inability to accomplish specific tasks or objectives. Take multi-robot control as an example. In order to cooperatively and safely accomplish tasks, each running robot must not take certain actions or not visit certain states, which can imply unsafe for itself, its collaborators, or the infrastructure of its environment (Hsu et al., 2023). These widely existing potential dangers, e.g., agents collide with each other, exacerbate the difficulty of safety decision-making in applying MARL. Consequently, it is necessary to research the safe decision-making problem in MARL to ensure that joint behaviors have performance improvement with a safety guarantee.

There are two main approaches concerning safe MARL techniques in the existing literature. The first type is shielded-based reactive methods (Zhang et al., 2019; Melcer et al., 2022), which combines environmental dynamics and safety specification constraints to predict whether agents' chosen actions will violate these constraints. Nevertheless, due to the reliance on precise modeling knowledge, these methods may lead to poor performance when the accurate state transition model is unavailable. The second type commonly employs a constrained Markov game to formulate the safe MARL problem, which requires agents to maximize total reward while avoiding violating cost constraints by solving a constrained optimization problem. To mention a few, several safe MARL variants, such as CMIX (Liu et al., 2021) and MAPPO-L (Gu et al., 2023), have been proposed, which learn the centralized value function by introducing the global state to overcome policy conflicts caused by the partially observable and non-stationarity nature of the environment faced by each agent. Unfortunately, the global coupling arising from agents' safety constraints and the exponential growth of the state-action space size make the usability of these algorithms in instant communication or computing resource-constrained systems and the scalability in larger

[1] Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

multi-agent systems become a bottleneck, limiting their applicability. To avoid these shortcomings mentioned above, Safe Dec-PG (Lu et al., 2021), a decentralized policy gradient descent-ascent method under a consensus network, is proposed, which adopts a primal-dual framework to find the saddle point between maximizing reward and minimizing cost. Yet, reaching a consensus is equivalent to imposing extra parameter-sharing constraints among neighboring agents, which may result in suboptimal solutions. Recent research (Ying et al., 2023), in a similar vein, introduces a primal-dual method and uses shadow reward and κ -hop neighbor truncation under a form of correlation decay property. However, since the truncated policy gradient for each agent depends on the states and actions of its κ-hop neighbors, this method cannot but adopt joint training in a local area, which is still cursed by non-stationary issues. Motivated by the urgent desire for decentralized learning in practical applications and the fact that most methods find it challenging to meet both safety guarantee and joint policy improvement in this context, we investigate a novel decentralized safe MARL with theoretical analysis, practical algorithm, and simulation verification.

In this study, we focus on decentralized learning without global observability and aim to provide rigorously theoretical analysis results and a feasible algorithm for decentralized safe MARL. Specifically, our main contributions are summarized as follows:

- We quantify the maximum information loss regarding the advantage function based on two assumptions for the spatial correlation of the transition dynamics and policies. Furthermore, we develop a novel decentralized policy optimization method, which follows a sequential update scheme to optimize κ-hop policies and meets the safety guarantee and the joint policy improvement.
- We propose a practical algorithm for decentralized safe MARL, called decentralized MAPPO-Lagrangian (DEC-MAPPO-L), which obtains an approximate theoretical solution.
- Experimentally, we provide the results on several safe MARL tasks to evaluate the effectiveness of our proposed method and the sensitivity for the hyperparameter κ.

2. Preliminaries

2.1. Constrained Markov game

A safe MARL problem can be formulated as a constrained Markov game $\langle \mathcal{N}, \mathcal{S}, \mathcal{A}, P, \gamma, \boldsymbol{\rho}_0, \boldsymbol{R}, \boldsymbol{C}, \boldsymbol{c} \rangle$. Here, $\mathcal{N} = \{1, \ldots, n\}$ is the set of agents; $\mathcal{S}$ and $\mathcal{A}$ are the global state and action spaces, which are the product of local spaces, i.e., $\mathcal{S} = \times_{i \in \mathcal{N}} \mathcal{S}^i$, $\mathcal{A} = \times_{i \in \mathcal{N}} \mathcal{A}^i$, meaning that for every $\mathbf{s} \in \mathcal{S}$ and $\mathbf{a} \in \mathcal{A}$, we can write $\mathbf{s} = (\mathrm{s}^1, \ldots, \mathrm{s}^n)$ and $\mathbf{a} = (\mathrm{a}^1, \ldots, \mathrm{a}^n)$; $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \longrightarrow \mathbb{R}$ is the probabilistic transition dynamics function, which satisfies the Dobrushin condition (Dembo & Montanari, 2009) as follows:

$$C^{ij} = \sup_{\mathrm{z}^j, \mathrm{z}'^j, \mathbf{z}^{-j}} \left\| P^i(\cdot | \mathrm{z}^j, \mathbf{z}^{-j}) - P^i(\cdot | \mathrm{z}'^j, \mathbf{z}^{-j}) \right\|_1, \quad (1)$$

where $\mathrm{z}^j = (\mathrm{s}^j, \mathrm{a}^j)$ and $\mathrm{z}'^j = (\mathrm{s}'^j, \mathrm{a}'^j)$ represent two different state-action pairs of the agent j respectively, and $\mathbf{z}^{-j}$ represents the state-action pair of the agent other than j. $\boldsymbol{\rho}_0$ is the initial state distribution, $\boldsymbol{R} : \mathcal{S} \times \mathcal{A} \longrightarrow \mathbb{R}$ is the joint reward function, $\boldsymbol{C} = \{C^i_j\}^{i \in \mathcal{N}}_{1 \leq j \leq m^i}$ is the sets of cost functions (every agent i has m^i cost functions) of the form $C^i_j : \mathcal{S}^i \times \mathcal{A}^i \longrightarrow \mathbb{R}$, and finally the set of corresponding cost-constraining values is given by $\boldsymbol{c} = \{c^i_j\}^{i \in \mathcal{N}}_{1 \leq j \leq m^i}$. At each timestep t, every agent i is in a state $s^i_t \in \mathcal{S}$, and takes an action a^i_t according to its policy $\pi^i = (\mathrm{a}^i | \mathrm{s}^i_t)$. Together with other agents' actions, it gives a joint action $\mathbf{a}_t = (\mathrm{a}^1_t, \ldots, \mathrm{a}^n_t)$ and the joint policy $\boldsymbol{\pi} = \prod_{i=1}^n \pi^i(\mathrm{a}^i | \mathrm{s}^i_t)$. The agents receive the reward $\boldsymbol{R}(\mathbf{s}_t, \mathbf{a}_t)$, meanwhile each agent i pays the costs $C^i_j\left(\mathrm{s}^i_t, \mathrm{a}^i_t\right), \forall\ j = 1, \ldots, m^i$, and their goal is to maximize the expected total reward of

$$J(\boldsymbol{\pi}) \triangleq \mathbb{E}_{\mathbf{s}_0 \sim \boldsymbol{\rho}_0, \mathbf{a}_{0:\infty} \sim \boldsymbol{\pi}} \left[\sum_{t=0}^{\infty} \gamma^t \boldsymbol{R}(\mathbf{s}_t, \mathbf{a}_t) \right], \quad (2)$$

meanwhile satisfying every agent i's safety constraints, written as

$$\begin{aligned} J^i_j(\boldsymbol{\pi}) \triangleq & \mathbb{E}_{\mathbf{s}_0 \sim \boldsymbol{\rho}_0, \mathbf{a}_{0:\infty} \sim \boldsymbol{\pi}} \left[\sum_{t=0}^{\infty} \gamma^t C^i_j(\mathbf{s}_t, \mathbf{a}^i_t) \right] \leq c^i_j, \\ & \forall\ j = 1, \ldots, m^i. \end{aligned} \quad (3)$$

2.2. Centralized joint constrained policy optimization

In an ideal centralized learning setting (Liu et al., 2021; Gu et al., 2023), the global state $\mathbf{s}$ and the action $\mathbf{a}^{-i}$ of all other agents can be explicitly or implicitly accessed for every agent i. Thus, the state-action value function and advantage function of agent i can defined as follows:

$$Q^i_{\boldsymbol{\pi}}(\mathbf{s}, \mathrm{a}^i) = \mathbb{E}_{\mathbf{a}^{-i} \sim \boldsymbol{\pi}^{-i}} Q^i_{\boldsymbol{\pi}}(\mathbf{s}, \mathbf{a}^{-i}, \mathrm{a}^i), \quad (4)$$

$$A^i_{\boldsymbol{\pi}}(\mathbf{s}, \mathrm{a}^j, \mathrm{a}^i) = Q^{j,i}_{\boldsymbol{\pi}}(\mathbf{s}, \mathrm{a}^j, \mathrm{a}^i) - Q^j_{\boldsymbol{\pi}}(\mathbf{s}, \mathrm{a}^j). \quad (5)$$

Further, updating agents' policies by following a sequential update scheme (Kuba et al., 2022), the multi-agent joint advantage function $\boldsymbol{A}_{\boldsymbol{\pi}}(\mathbf{s}, \mathbf{a})$ can be written as a sum of sequentially unfolding multi-agent advantages of individual agents, as stated by Lemma 2.1.

Lemma 2.1. *(Multi-agent advantage decomposition). For any agent $i \subseteq \mathcal{N}$, the action a^i, and the state $\mathbf{s} \in \mathcal{S}$, the following identity holds*

$$\boldsymbol{A}_{\boldsymbol{\pi}}(\mathbf{s}, \mathbf{a}) = \sum_{i=1}^{n} A^i_{\boldsymbol{\pi}}(\mathbf{s}, \mathbf{a}^{-i}, \mathrm{a}^i). \tag{6}$$

The proof of Lamma 2.1 can be seen in (Gu et al., 2023), and a rewritten version is reported in Appendix A.1.

Definition 2.2. Let $\boldsymbol{\pi}$ be a joint policy, $\bar{\boldsymbol{\pi}}^{1:i-1}$ be some other joint policy of agents $1 : i-1$, and $\hat{\pi}^i$ be a policy of agent i. Then, one define

$$\begin{aligned} & L^{1:i}_{\boldsymbol{\pi}}\left(\bar{\boldsymbol{\pi}}^{1:i-1}, \hat{\pi}^i\right) \\ & \triangleq \mathbb{E}_{\mathbf{s}\sim\boldsymbol{\rho}_{\boldsymbol{\pi}}, \mathbf{a}^{1:i-1}\sim\bar{\boldsymbol{\pi}}^{1:i-1}, \mathrm{a}^i\sim\hat{\pi}^i}\left[A^i_{\boldsymbol{\pi}}\left(\mathbf{s}, \mathbf{a}^{1:i-1}, \mathrm{a}^i\right)\right]. \end{aligned} \tag{7}$$

Building on Lemma 2.1 and Definition 2.2, one can obtain

$$\begin{aligned} & L^{1:i}_{\boldsymbol{\pi}}\left(\bar{\boldsymbol{\pi}}^{1:i-1}, \bar{\pi}^i\right) \\ & = \mathbb{E}_{\mathbf{s}\sim\boldsymbol{\rho}_{\boldsymbol{\pi}}, \mathbf{a}^{1:i}\sim\bar{\boldsymbol{\pi}}^{1:i}}\left[\sum_{h=1}^{i} A^j_{\boldsymbol{\pi}}\left(\mathbf{s}, \mathbf{a}^{1:h-1}, \mathrm{a}^h\right)\right]. \end{aligned} \tag{8}$$

Then, according to the results in (Schulman et al., 2015) and (Gu et al., 2023), the following proposition can be obtained.

Proposition 2.3. *If each agent i sequentially solves the following constrained optimization problem, then the joint policy $\boldsymbol{\pi}$ has the monotonic improvement property, $J(\bar{\boldsymbol{\pi}}) \geq J(\boldsymbol{\pi})$, as well as it satisfies the safety constraints, $J^i_j(\boldsymbol{\pi}) \leq c^i_j$, for all $i \in \mathcal{N}$, and $j \in \{1, \ldots, m^i\}$.*

$$\begin{aligned} & \bar{\pi}^i = \underset{\hat{\pi}^i\in\bar{\Pi}^i}{\arg\max}\, L^{1:i}_{\boldsymbol{\pi}}\left(\bar{\boldsymbol{\pi}}^{1:i-1}, \hat{\pi}^i\right) - \nu D^{\max}_{\mathrm{KL}}\left(\pi^i, \hat{\pi}^i\right), \\ & s.t. \left\{\hat{\pi}^i \in \bar{\Pi}^i \mid D^{\max}_{\mathrm{KL}}\left(\pi^i, \hat{\pi}^i\right) \leq \delta^i, \textit{ and} \right. \\ & J^i_j(\boldsymbol{\pi}) + L^i_{j,\boldsymbol{\pi}}(\hat{\pi}^i) + \nu^i_j D^{\max}_{\mathrm{KL}}\left(\pi^i, \bar{\pi}^i\right) \leq \\ & \left. c^i_j - \textstyle\sum_{h=1}^{i-1} \nu^h_j D^{\max}_{\mathrm{KL}}\left(\pi^h, \bar{\pi}^h\right), \forall\, j = 1, \ldots, m^i \right\}, \end{aligned} \tag{9}$$

where

$$\begin{aligned} & \nu = \frac{2\gamma \max_{\mathbf{s},\mathbf{a}} \left|A_{\boldsymbol{\pi}}\left(\mathbf{s}, \mathbf{a}\right)\right|}{(1-\gamma)^2}, \\ & \nu^i_j = \frac{2\gamma \max_{\mathbf{s},\mathrm{a}^i} \left|A^i_{j,\boldsymbol{\pi}}\left(\mathbf{s}, \mathrm{a}^i\right)\right|}{(1-\gamma)^2}, \\ & D^{\max}_{\mathrm{KL}}\left(\pi^i, \hat{\pi}^i\right) = \max_{\mathbf{s}} D_{\mathrm{KL}}\left(\pi^i(\cdot \mid \mathbf{s}), \hat{\pi}^i(\cdot \mid \mathbf{s})\right), \\ & \delta^i = \min\left\{\min_{h\leq i-1} \min_{1\leq j\leq m^h} \frac{\Xi^h_j - L^h_{j,\boldsymbol{\pi}}\left(\hat{\pi}^h\right)}{\nu^h_j},\right. \\ & \qquad \left. \min_{h\geq i+1} \min_{1\leq j\leq m^h} \frac{\Xi^h_j}{\nu^h_j}\right\}, \\ & \Xi^h_j = c^h_j - J^h_j\left(\pi^h\right) - \nu^h_j \sum_{l=1}^{i-1} D^{\max}_{\mathrm{KL}}\left(\pi^l, \hat{\pi}^l\right). \end{aligned} \tag{10}$$

When each agent adopts the policy update approach in Proposition 2.3, the joint policy can be guaranteed safety and have the monotonic improvement property. The proof of Proposition 2.3 is reported in Appendix A.2. However, it is worth noting that this approach imposes demanding requirements on both strong instant communication and computational capabilities during training. It is costly and even may be unachievable in some multi-agent environments. There are two main reasons as follows: (1) the safe guarantee and monotonic improvement property of the joint policy rely on the condition that the global information is available; (2) the joint advantage function is trained during each policy update round.

3. Decentralized constrained policy optimization

Although the centralized safe MARL method in Subsection 2.2 has good performance guarantees, we prioritize decentralized safe MARL as there are still many settings in which the global state cannot be available, and also for better robustness and scalability (Zhang et al., 2021; Munir et al., 2021). Moreover, the idea of decentralized learning is direct, comprehensible, and easy to realize in practice.

In this section, we quantify the maximum information loss regarding the advantage function based on two assumptions for the spatial correlation of the transition dynamics and policies. Further, we develop a novel decentralized and theoretically-justified policy optimization method, which follows a sequential update scheme to optimize κ-hop policies and meets the safety guarantee and the joint policy improvement. Finally, we parameterize each agent's policy and propose a practical algorithm for decentralized safe MARL, i.e., DEC-MAPPO-L, which obtains an approximate theoretical solution.

3.1. Optimization objective

According to Proposition 2.3 and its proof in Appendix A.2, we can obtain the following bounds:

$$J\left(\bar{\boldsymbol{\pi}}\right) - J\left(\boldsymbol{\pi}\right) \geq \sum_{i=1}^{n}\left(L^{1:i}_{\boldsymbol{\pi}}\left(\bar{\boldsymbol{\pi}}^{1:i}\right) - \nu D^{\max}_{\mathrm{KL}}\left(\pi^i, \bar{\pi}^i\right)\right), \tag{11}$$

$$J^i_j(\bar{\boldsymbol{\pi}}) \leq J^i_j(\boldsymbol{\pi}) + L^i_{j,\boldsymbol{\pi}}\left(\bar{\pi}^i\right) + \nu^i_j \sum_{h=1}^{i} D^{\max}_{\mathrm{KL}}\left(\pi^h, \bar{\pi}^h\right). \tag{12}$$

where ν, ν^i_j and $D^{\max}_{\mathrm{KL}}\left(\pi^i, \bar{\pi}^i\right)$ are defined in (10).

The right-hand side of (11) is a joint TRPO objective, i.e., a lower bound for the difference between the new joint policy $\bar{\boldsymbol{\pi}}$ and the old joint policy $\boldsymbol{\pi}$ in term of expected return when update policies by (9). The right-hand side of (12) is an upper bound for the new joint policy $\bar{\boldsymbol{\pi}}$, which can be used to restrict agents only to choose safe actions. Therefore, we use the objective, i.e., jointly optimizing the lower bound for the difference of joint policies and the upper bound for the safety constraints, as a surrogate. However, from the perspective of policy optimization, we cannot directly optimize this objective under a decentralized learning framework as this objective is involved in the global state and the new policies for other agents. Thus, we will focus on converting the objective into one that can be optimized independently in a decentralized learning framework in the next subsection.

3.2. Spatial correlation decay and κ-hop policy

Inspired by (Alfano & Rebeschini, 2021; Ying et al., 2023), we make the following two assumptions for the spatial correlation of the transition dynamics and policies (Dembo & Montanari, 2009; Gamarnik, 2013).

Assumption 3.1. (Spatial Decay of Correlation for the Dynamics) Assume that there exist $\omega > 0$ in (1), for any agents $i, j \subseteq \mathcal{N}$, such that

$$\max_{i \in \mathcal{N}} \sum_{j \in \mathcal{N}} e^{\omega d(i,j)} C_{ij} \leq \zeta, \tag{13}$$

where $d(i,j)$ represents the distance between agent i and agent j, $\gamma \in [0,1)$ is the discount factor and $\zeta \in [0, 2/\gamma)$ is a constant.

Assumption 3.2. (Spatial Decay of Correlation for the Policies) Assume that there exist $\xi, \beta \geq 0$, for any agents $i, j \subseteq \mathcal{N}$, such that

$$\begin{aligned} &\sup_{\mathrm{s}_{\mathcal{N}_\kappa^i}, \mathbf{s}_{\mathcal{N}_\kappa^{-i}}, \mathbf{s}'_{\mathcal{N}_\kappa^{-i}}} \left| \pi^i\left(\cdot | \mathrm{s}_{\mathcal{N}_\kappa^i}, \mathbf{s}_{\mathcal{N}_\kappa^{-i}}\right) - \pi^i\left(\cdot | \mathrm{s}_{\mathcal{N}_\kappa^i}, \mathbf{s}'_{\mathcal{N}_\kappa^{-i}}\right) \right| \\ &\leq \xi e^{-\beta\kappa}, \forall\, \mathrm{s}_{\mathcal{N}_\kappa^i} \in \mathcal{S}_{\mathcal{N}_\kappa^i}, \text{and } \mathbf{s}_{\mathcal{N}_\kappa^{-i}}, \mathbf{s}'_{\mathcal{N}_\kappa^{-i}} \in \mathcal{S}_{\mathcal{N}_\kappa^{-i}}, \end{aligned} \tag{14}$$

where $\mathrm{s}_{\mathcal{N}_\kappa^i}$ represents the state of agent i's κ-hop neighborhood $\mathcal{N}_\kappa^i$. For simplicity, we use the notation $\pi_\kappa^i = \pi^i(\cdot | \mathrm{s}_{\mathcal{N}_\kappa^i})$ for κ-hop policies when it is clear from context.

Assumption 3.1 portrays a common phenomenon: the transition dynamic of each agent is exponentially less sensible to perturbations of the states and actions of more distant agents, which is commonly seen in wireless communication, epidemics, traffic, and so on scenarios (Mei et al., 2017; Zocca, 2019). The value of C_{ij} reflects the extent to which the local transition probability of agent i is affected by the state and action of agent j. Assumption 3.2 imposes a design constraint for the policy class that encodes a weaker correlation decay property than the assumptions on the nature of Assumption 3.1. Moreover, Assumption 3.2 reveals how much information is lost compared with access to the global state and allows us to consider a policy class with the necessary properties for the optimal policy under Assumption 3.1, as stated in Appendix B.1. Below, we quantify the maximum information loss regarding the advantage function based on these two assumptions.

Proposition 3.3. *If Assumption 3.1 and Assumption 3.2 hold, then there exist the parameters $(c, \phi) = \left(\frac{\xi\gamma\zeta}{1-\gamma\zeta}, e^{-\beta}\right)$ such that the exponential decay property holds for the advantage function of any agent $i \subseteq \mathcal{N}$, i.e., we have that*

$$\begin{aligned} &\sup_{\mathrm{z}_{\mathcal{N}_\kappa^i}, \mathbf{z}_{\mathcal{N}_\kappa^{-i}}, \mathbf{z}'_{\mathcal{N}_\kappa^{-i}}} \left| A^i\left(\mathrm{z}_{\mathcal{N}_\kappa^i}, \mathbf{z}_{\mathcal{N}_\kappa^{-i}}\right) - A^i\left(\mathrm{z}_{\mathcal{N}_\kappa^i}, \mathbf{z}'_{\mathcal{N}_\kappa^{-i}}\right) \right| \\ &\leq c\phi^\kappa, \forall\, \mathrm{z}_{\mathcal{N}_\kappa^i} = \left(\mathrm{s}_{\mathcal{N}_\kappa^i}, \mathrm{a}_{\mathcal{N}_\kappa^i}\right) \in \mathcal{S}_{\mathcal{N}_\kappa^i} \times \mathcal{A}_{\mathcal{N}_\kappa^i}. \end{aligned} \tag{15}$$

Proposition 3.3 shows that when the transition dynamics and policies correlation satisfy the exponential correlation decay property, the advantage functions also have exponential decay dependence on the states and actions of the more distant agent. The proof of Proposition 3.3 is reported in Appendix B.2. In addition, based on this proposition, we can obtain the following two corollaries.

Corollary 3.4. *If Proposition 3.3 holds, for all agent $i \subseteq \mathcal{N}$, we have that*

$$\left| L_{\boldsymbol{\pi}}^{1:i}\left(\bar{\boldsymbol{\pi}}^{1:i-1}, \bar{\pi}^i\right) - L_{\pi_\kappa^i}^i\left(\bar{\pi}_\kappa^i\right) \right| \leq c^i \phi^\kappa, \tag{16}$$

where $(c^i, \phi) = \left(\frac{M^i\xi}{1-\gamma} + \frac{(2+\xi)\gamma\zeta}{1-\gamma\zeta}, e^{-\beta}\right)$, and M^i is a constant.

Corollary 3.5. *Let $\boldsymbol{\pi}$ and $\bar{\boldsymbol{\pi}}$ be joint policies. Let $i \in \mathcal{N}$ be an agent, and $j \in \{1, \ldots, m^i\}$ be an index of one of its costs. The following inequality holds*

$$J_j^i(\bar{\boldsymbol{\pi}}) \leq J_j^i(\boldsymbol{\pi}) + L_{j,\pi_\kappa^i}^i\left(\bar{\pi}_\kappa^i\right) + c_j\phi^\kappa + \nu_{j,\kappa}^i \sum_{h=1}^{i} D_{\mathrm{KL}}^{\max}\left(\pi_\kappa^h, \bar{\pi}_\kappa^h\right). \tag{17}$$

where $L_{j,\pi_\kappa^i}^i(\bar{\pi}_\kappa^i) = \mathbb{E}_{\mathrm{s}_{\mathcal{N}_\kappa^i} \sim \rho_{\pi_\kappa^i}, \mathrm{a}^i \sim \bar{\pi}_\kappa^i}\left[A_{j,\pi_\kappa^i}^i\left(\mathrm{s}_{\mathcal{N}_\kappa^i}, \mathrm{a}^i\right)\right]$, $\nu_{j,\kappa}^i = \frac{2\gamma \max_{\mathrm{s}_{\mathcal{N}_\kappa^i}, \mathrm{a}^i}\left|A_{j,\pi_\kappa^i}^i\left(\mathrm{s}_{\mathcal{N}_\kappa^i}, \mathrm{a}^i\right)\right|}{(1-\gamma)^2}$, $(c_j, \phi) = \left(\frac{M_j\xi}{1-\gamma} + \frac{(2+\xi)\gamma\zeta}{1-\gamma\zeta}, e^{-\beta}\right)$, and M_j is a constant.

The proofs of Corollary 3.4 and Corollary 3.5 are reported in Appendix B.3-B.4. Then, we can obtain an approximation TRPO surrogate objective, which can be learned independently for each agent to update their policies by following a sequential update scheme. The joint policy $\bar{\boldsymbol{\pi}}$ indeed improves the expected return, meanwhile satisfying the safety constraints, as stated by Theorem 3.6.

Theorem 3.6. *The joint policy $\boldsymbol{\pi}$ has the monotonic improvement property, $J\left(\bar{\boldsymbol{\pi}}\right) \geq J\left(\boldsymbol{\pi}\right)$, as well as it satisfies the safety constraints, $J_j^i\left(\boldsymbol{\pi}\right) \leq c_j^i$, for all $i \in \mathcal{N}$, and*

$j \in \{1, \ldots, m^i\}$, when the policy is updated by following a sequential update scheme, that is, each agent sequentially solves the following optimization problem:

$$\begin{aligned}&\bar{\pi}^i_\kappa = \underset{\hat{\pi}^i_\kappa \in \bar{\Pi}^i_\kappa}{\arg\max} \left(L^i_{\pi^i_\kappa}\left(\hat{\pi}^i_\kappa\right) - c^i \phi^\kappa - \nu^i_\kappa D^{\max}_{\mathrm{KL}}\left(\pi^i_\kappa | \hat{\pi}^i_\kappa\right)\right), \\ &s.t. \left\{\hat{\pi}^i_\kappa \in \bar{\Pi}^i_\kappa \mid D^{\max}_{\mathrm{KL}}\left(\pi^i_\kappa, \hat{\pi}^i_\kappa\right) \leq \delta^i_\kappa, and\right. \\ &\ J^i_j\left(\pi_\kappa\right) + L^i_{j,\pi^i_\kappa}\left(\hat{\pi}^i_\kappa\right) + c_j \phi^\kappa + \nu^i_{j,\kappa} D^{\max}_{\mathrm{KL}}\left(\pi^i_\kappa, \hat{\pi}^i_\kappa\right) \\ &\ \left.\leq c^i_j - \textstyle\sum_{h=1}^{i-1} \nu^h_{j,\kappa} D^{\max}_{\mathrm{KL}}\left(\pi^h_\kappa, \hat{\pi}^h_\kappa\right), \forall\ j = 1, \ldots, m^i\right\},\end{aligned} \tag{18}$$

where

$$\begin{aligned}\nu^i_\kappa &= \frac{2\gamma \max_{\mathrm{s}_{\mathcal{N}^i_\kappa}, \mathrm{a}^i} \left|A^i_{\pi^i_\kappa}\left(\mathrm{s}_{\mathcal{N}^i_\kappa}, \mathrm{a}^i\right)\right|}{(1-\gamma)^2}, \\ \nu^i_{j,\kappa} &= \frac{2\gamma \max_{\mathrm{s}_{\mathcal{N}^i_\kappa}, \mathrm{a}^i} \left|A^i_{j,\pi^i_\kappa}\left(\mathrm{s}_{\mathcal{N}^i_\kappa}, \mathrm{a}^i\right)\right|}{(1-\gamma)^2}, \\ (c^i, \phi) &= \left(\frac{M^i \xi}{1-\gamma} + \frac{(2+\xi)\gamma\zeta}{1-\gamma\zeta}, e^{-\beta}\right), \\ (c_j, \phi) &= \left(\frac{M_j \xi}{1-\gamma} + \frac{(2+\xi)\gamma\zeta}{1-\gamma\zeta}, e^{-\beta}\right), \\ \delta^i_\kappa &= \min\left\{\min_{h \leq i-1} \min_{1 \leq j \leq m^h} \frac{\Xi^h_j - L^h_{j,\pi^h_\kappa}\left(\hat{\pi}^h_\kappa\right)}{\nu^i_{j,\kappa}},\right. \\ &\qquad \left.\min_{h \geq i+1} \min_{1 \leq j \leq m^h} \frac{\Xi^h_j}{\nu^i_{j,\kappa}}\right\}, \\ \Xi^h_j &= c^h_j - J^h_j\left(\pi^h_\kappa\right) - \nu^h_j \sum_{l=1}^{i-1} D^{\max}_{\mathrm{KL}}\left(\pi^l_\kappa, \hat{\pi}^l_\kappa\right).\end{aligned}$$

Theorem 3.6 assures that if one follows (18) to update policies, agents will not only explore safe policies independently; meanwhile, every new policy will be guaranteed to result in performance improvement. It is worth mentioning that these two properties hold only under the condition that the only policy update restriction, i.e., $\bar{\pi}^i_\kappa \in \bar{\Pi}^i_\kappa$, is satisfied; this is due to the KL-penalty term in every agent's objective (i.e., $\nu^i_\kappa D^{\max}_{\mathrm{KL}}\left(\pi^i_\kappa, \bar{\pi}^i_\kappa\right)$, as well as the constraints on cost surrogates. The proof of Theorem 3.6 is reported in Appendix B.5.

3.3. Algorithm

In this section, we focus on how to practically implement policy update in Theorem 3.6 for each agent. Specifically, we parameterize each local policy $\pi^i_{\theta^i_\kappa}$ by a neural network with parameter θ^i_κ. At each policy update, every agent i maximizes its surrogate return and is subject to surrogate cost constraints and a form of expected KL-divergence constraint $\widetilde{D}_{\mathrm{KL}}\left(\pi^i_\kappa, \bar{\pi}^i_\kappa\right) \leq \delta^i_\kappa$, which avoids the computation of KL-divergence at every state. Then, we introduce a scalar variable λ and convert the constrained optimization problem from (18) into a min-max optimization problem with Lagrangian multipliers by subsuming the cost constraints. As such, the new optimization problem for all $i \in \mathcal{N}$ is as follows:

Algorithm 1 DEC-MAPPO-L

```
Input: Initialize policy parameters $\theta^1_\kappa, \ldots, \theta^n_\kappa$ and Lagrangian multipliers $\lambda^1_j, \ldots, \lambda^n_j$, $\forall\ j \in \{1, \ldots, m^i\}$.
for $t = 0, 1, \ldots$ do
    for $i = 1 : n$ do
        Select action $a^i \sim \pi^i_{\theta^i_\kappa}$.
        Execute action $a^i$ and receive reward $R$ and next state $\mathrm{s}_{\mathcal{N}^i_\kappa}$.
        Compute the sourragte advantage functions $L^i_{\pi^i_\kappa}\left(\hat{\pi}^i_\kappa\right)$ and $L^i_{j,\pi^i_\kappa}\left(\hat{\pi}^i_\kappa\right), \forall\ j \in \{1, \ldots, m^i\}$.
        Compute $\nu^i_\kappa$ and $\nu^i_{j,\kappa}, \forall\ j \in \{1, \ldots, m^i\}$.
        Compute the radius of the KL-constraint $\delta^i_\kappa$.
        Update decentralized critic according to (23).
        Update policy according to (22).
    end for
end for
```

$$\begin{aligned}&\max_{\theta^i_\kappa} \min_{\lambda^i_{1:m^i} \geq 0} \left[\mathbb{E}_{\mathrm{s}_{\mathcal{N}^i_\kappa} \sim \rho_{\pi^i_{\theta^i_\kappa}}, \mathrm{a}^i \sim \pi^i_{\theta^i_\kappa}}\left[A^i_{\pi^i_{\theta^i_\kappa}}\left(\mathrm{s}_{\mathcal{N}^i_\kappa}, \mathrm{a}^i\right)\right]\right. \\ &\left.- \sum_{u=1}^{m^i} \lambda^i_u \left(\mathbb{E}_{\mathrm{s}_{\mathcal{N}^i_\kappa} \sim \rho_{\pi^i_{\theta^i_\kappa}}, \mathrm{a}^i \sim \pi^i_{\theta^i_\kappa}}\left[A^i_{u,\pi^i_{\theta^i_\kappa}}\left(\mathrm{s}_{\mathcal{N}^i_\kappa}, \mathrm{a}^i\right)\right] + d^i_u\right)\right], \\ &\text{s.t. } \widetilde{D}_{\mathrm{KL}}\left(\pi^i_{\theta^i_\kappa}, \bar{\pi}^i_{\theta^i_\kappa}\right) \leq \delta^i_\kappa.\end{aligned} \tag{19}$$

where $\lambda^i_{1:m^i}$ is a scalar variable and θ^i_κ is a parameter of neural network for agent i.

Further, to alleviate the complications caused by computing the KL-divergence constraint, we simplify it by adopting the PPO-clip objective (Schulman et al., 2017), i.e., replacing the KL-divergence constraint with the clip operator and updating the policy parameter with first-order methods. We rewriting the (19) as

$$\begin{aligned}&\max_{\theta^i_\kappa} \min_{\lambda^i_{1:m^i} \geq 0} \left[\mathbb{E}_{\mathrm{s}_{\mathcal{N}^i_\kappa} \sim \rho_{\pi^i_{\theta^i_\kappa}}, \mathrm{a}^i \sim \pi^i_{\theta^i_\kappa}}\left[A^{i,(\lambda)}_{\pi^i_{\theta^i_\kappa}}\left(\mathrm{s}_{\mathcal{N}^i_\kappa}, \mathrm{a}^i\right)\right]\right], \\ &\text{s.t. } \widetilde{D}_{\mathrm{KL}}\left(\pi^i_{\theta^i_k}, \bar{\pi}^i_{\theta^i_\kappa}\right) \leq \delta^i_\kappa,\end{aligned} \tag{20}$$

where

$$\begin{aligned}A^{i,(\lambda)}_{\pi^i_{\theta^i_\kappa}}\left(\mathrm{s}_{\mathcal{N}^i_\kappa}, \mathrm{a}^i\right) &= A^i_{\pi^i_{\theta^i_\kappa}}\left(\mathrm{s}_{\mathcal{N}^i_\kappa}, \mathrm{a}^i\right) \\ &\quad - \sum_{u=1}^{m^i} \lambda^i_u \left(A^i_{u,\pi^i_{\theta^i_\kappa}}\left(\mathrm{s}_{\mathcal{N}^i_\kappa}, \mathrm{a}^i\right) + d^i_u\right).\end{aligned} \tag{21}$$

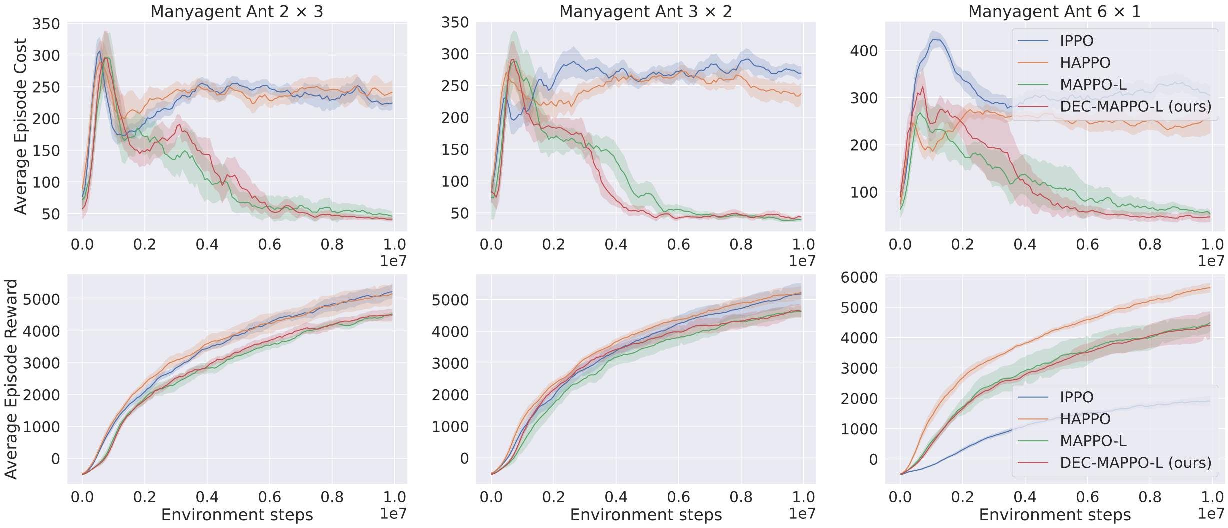

Figure 1. Performance comparisons in terms of cost and reward on three Safe ManyAgent Ant tasks. Each column subfigure represents a different task, and we plot the cost curves (the lower the better) in the upper row and the reward curves (the higher the better) in the bottom row for each task.

The objective in (20) takes the form of an expectation with a KL-divergence constraint on the policy. To approximate the distance between new and old policies, we adopt the PPO-clip operator (Schulman et al., 2017) to adjust it. Then, the final optimization problem takes the form

$$\max_{\theta^i_\kappa} \min_{\lambda^i_{1:m^i}\geq 0} \mathbb{E}_{\mathrm{s}_{\mathcal{N}^i_\kappa}\sim\rho_{\pi^i_{\theta^i_\kappa}}, \mathrm{a}^i\sim\pi^i_{\theta^i_\kappa}} \left[\min\left(\frac{\bar{\pi}^i_{\theta^i_\kappa}}{\pi^i_{\theta^i_\kappa}} A^{i,(\lambda)}_{\pi^i_{\theta^i_\kappa}}\left(\mathrm{s}_{\mathcal{N}^i_\kappa}, \mathrm{a}^i\right), \left(\frac{\bar{\pi}^i_{\theta^i_\kappa}}{\pi^i_{\theta^i_\kappa}}, 1\pm\epsilon\right) A^{i,(\lambda)}_{\pi^i_{\theta^i_\kappa}}\left(\mathrm{s}_{\mathcal{N}^i_\kappa}, \mathrm{a}^i\right)\right)\right]. \tag{22}$$

Let each agent i independently learn an individual value function $V^i\left(\mathrm{s}'_{\mathcal{N}^i_\kappa}\right)$, which follows the traditional idea in decentralized learning. Then, the loss for the decentralized critic is as follows:

$$\mathcal{L}^i_{\text{critic}} = \mathbb{E}\left[\left(V^i(\mathrm{s}_{\mathcal{N}^i_\kappa}) - y_i\right)^2\right], \text{where } y_i = r + \gamma V^i\left(\mathrm{s}'_{\mathcal{N}^i_\kappa}\right). \tag{23}$$

To summarize, we give a decentralized learning procedure for each agent i, name DEC-MAPPO-L, and provide its main pseudocode in Algorithm 1. The algorithm is a decentralized version of the MAPPO-L algorithm (Gu et al., 2023) with a simple idea that each agent independently optimizes the surrogate objective (22), and some approximations of the decentralized surrogate objective are employed in the actual algorithm execution. Most of these approximations are traditional practices in RL, yet they may make it impossible for the practical algorithm to rigorously maintain the theoretical guarantees in Theorem 3.6.

4. Experiments

We evaluate our method via several numerical experiments in this section. Our experiments aim to answer the following questions: First, how does the cost and reward performance of Algorithm 1 compare with existing methods on challenging multi-agent safe tasks? Second, how does the different κ affect the performance of Algorithm 1, and could the surrogate advantage truncation effectively alleviate computational load?

4.1. Experimental setup

Safe MAMuJoCo (Gu et al., 2023) is an extension of MAMuJoCo (Peng et al., 2021), which preserves the agents, physics simulator, background environment, and reward function and comes with obstacles, like walls or pitfalls. To answer the first question, we compare our method against the other PPO family algorithms, i.e., IPPO (Yu et al., 2022), HAPPO (Kuba et al., 2022), and MAPPO-L (Gu et al., 2023) and choose three games from Safe MAMuJoCo: Safe ManyAgent Ant task with 2 agents (2 × 3), 3 agents (3 × 2), 6 agents (6 × 1) to evaluate their performance. Concerning the second question, we choose three games with different agent numbers and tasks from Safe MAMuJoCo: Safe ManyAgent Ant task with 6 agents (6 × 1), Safe Ant task with 8 agents (8 × 1), and Safe Coupled HalfCheetah task with 12 agents (12 × 1). We train DEC-MAPPO-L with same network architecture and hyperparameters as the original MAPPO-L implementation. All reported results are averaged over three or more random seeds, and the curves are smooth over time.

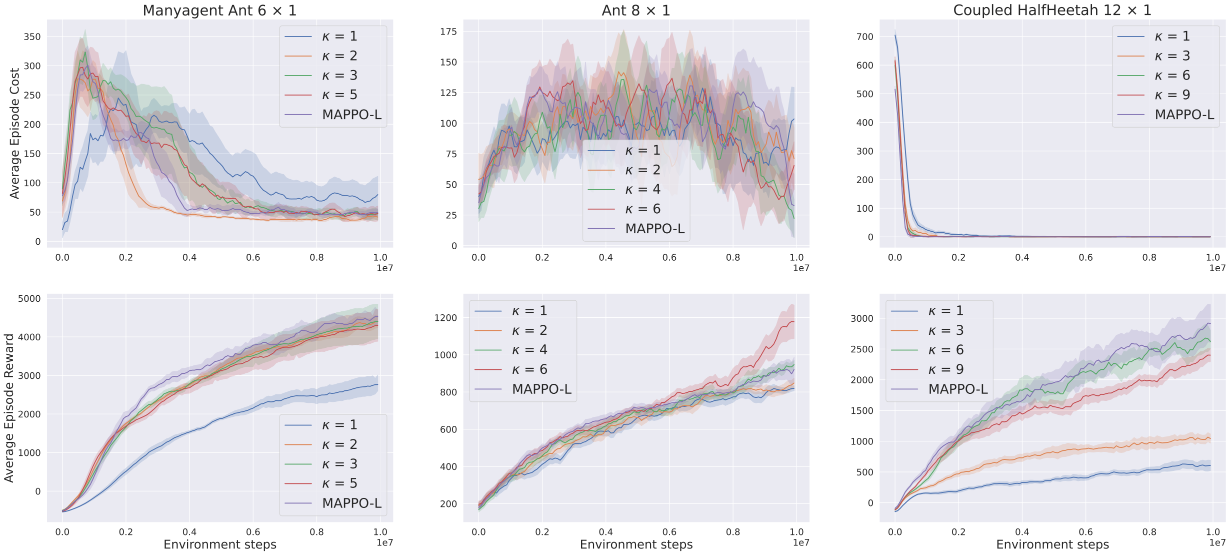

Figure 2. Performance comparisons in terms of cost and reward on Safe ManyAgent Ant task, Safe Ant task, and Safe Coupled HalfCheetah task. In each task, the performance of DEC-MAPPO-L with different κ and MAPPO-L are demonstrated.

4.2. Results

Comparisons with baselines: From Figure 1, we can see that compared to IPPO and HAPPO, on all three tasks, both DEC-MAPPO-L and MAPPO-L has fewer constraint violations and good performance (in terms of reward); that is to say, they keep their explorations within the feasible policy space and quickly learn to satisfy safety constraints. Moreover, it should be further pointed out that on all tasks, DEC-MAPPO-L performs similarly to MAPPO-L (which accesses the global state) even though it only accesses half of the state information. This means that the sensitivity of each agent to the states and actions perturbations of distant agents is minimal, and DEC-MAPPO-L is effective.

Performance with different κ: Figure 2 demonstrates the performance of Algorithm 1 in different environments with different κ. We observe that the algorithm performs poorly when the truncation with $\kappa = 1$, i.e., each agent can only access the state of itself and part of the information in the environment (not including the states and actions of its collaborators). However, when the truncation with $\kappa >= 3$, i.e., each agent has access to the states of at least two neighbors, we can observe that the performance of Algorithm 1 improves considerably and can approach or even outperform the MAPPO-L in some environments, such as $\kappa = 6$ in the Safe Ant task (8×1). This may be due to the difficulty in extracting useful information in many messages that causes them to have lower performance than Algorithm 1. These results underscore the efficiency of Algorithm 1 since it employs a smaller communication radius that can significantly reduce the computation.

5. Related Work

5.1. Safe RL

Safety is one of the bottlenecks preventing RL use in real-life applications, such as physical robotics (García & Shafie, 2020), medical applications (Datta et al., 2021) and autonomous driving (Gu et al., 2022a). Thus, safe RL has become a research hotspot in recent years and a growing number of safe RL approaches, such as trust region with safety constraints methods (Achiam et al., 2017), Lyapunov methods (Chow et al., 2018), primal-dual methods (Ding et al., 2020), and safety augmented methods (Sootla et al., 2022), have been developed. However, when it comes to multi-agent systems, this challenge is exacerbated by policy conflicts caused by multiple agents interacting within a shared environment and learning simultaneously. In order to address the above issue, CMIX (Liu et al., 2021) and MAPPO-L (Gu et al., 2023) have been proposed with the in-depth study of MARL. These algorithms follow the centralized training with decentralized execution (CTDE) framework (Sunehag et al., 2017; Rashid et al., 2018; Kuba et al., 2022), which learns the centralized value function by introducing the global state. Unfortunately, the global coupling arising from agents' safety constraints and the exponential growth of the state-action space size make the usability in communication or computing resource-constrained systems and the scalability of these algorithms in larger multi-agent systems become a bottleneck, limiting their applicability. Recent works have provided some theoretical results for safe MARL under decentralized learning to avoid these shortcomings. However, to the best of our knowledge, most methods fail to ensure both guarantee safety and joint policy improvement under a decentralized learning framework.

5.2. Spatial correlation decay

Exponential decay property (Qu et al., 2020), also known as spatial decay, is a powerful property associated with local interactions, which says that the impact of agents on each other decays exponentially in their graph distance. Over the past decades, many researchers have utilized spatial correlation property to design scalable, distributed algorithms for optimization and control problems in scenarios such as epidemics (Mei et al., 2017) and wireless communication (Zocca, 2019). Inspired by the studies mentioned above, a recent line of work (Qu & Li, 2019) has formally considered spatial decay of correlation assumptions and propose a method that finds nearly optimal local policies. An application (Gu et al., 2021) with the same principles adopts the setting of mean-field MARL (Yang et al., 2018), which proposes an actor-critic algorithm with global convergence. However, compared with the mean-field setting that requires the transition scheme of an agent to be only affected by the mean effect from its neighbors and effective only when agents are homogeneous, we allow each agent to have different transition probabilities and local policies.

5.3. Centralized training

In cooperative MARL settings, the training of agents can be broadly divided into two paradigms, namely centralized and decentralized (Gronauer & Diepold, 2022). The centralized training paradigm describes agent policies updated based on mutual information, which can be further differentiated into the centralized and decentralized execution framework. However, due to the obvious flaws that the state-action spaces grow exponentially by the number of agents and assume instantaneous and unconstrained information exchange between agents, the applicability of centralized training centralized execution (CTCE) is limited and has gradually faded out of the research field. Recently, centralized training decentralized execution (CTDE) has become the most popular framework (Rashid et al., 2018; Yu et al., 2022; Kuba et al., 2022; Gallici et al., 2023), since the fact that it addresses the non-stationarity issue with the centralized value function, remove the dependency on global state and actions during execution, and demonstrates state-of-the-art performance on challenging tasks, such as unit micromanagement in StarCraft II (Whiteson et al., 2019). However, although this framework does not require agents to access the global state during execution, the reliance on the global state only during training still poses a significant barrier to real-world applications, especially in scenarios where communication and computational resources are constrained (Du & Ding, 2021; Oroojlooy & Hajinezhad, 2023).

5.4. Decentralized training

In a decentralized learning paradigm, each agent learns independently and accesss local observations rather than the global state; the idea is direct, comprehensible, and easy to realize in practice (Zhang et al., 2021; Du & Ding, 2021). There are two mainline research approaches concerning decentralized learning in the existing literature. One line of research pursues fully decentralized learning and has carried out extensive experimental validation, such as independent Q-learning (IQL) (Tan, 1993; Tampuu et al., 2015) and independent actor-critic (IAC) (de Witt et al., 2020; Yu et al., 2022), which make agents directly execute the single-agent Q-learning or actor-critic algorithm individually. Another line of research establishes rational local communication networks, such as setting certain distance or neighbor graph (Jiang et al., 2019; Chu et al., 2019), to expand the perceptual capabilities of agents and mitigate the decision conflicts or errors caused by partial observability. It is worth noting that both lines of studies achieved good experimental results on a collection of benchmark tasks, even though they violate the stationary condition of MDP. Recently, motivated by good performance from experiments, some studies have tried to provide a theoretical basis for these phenomenons. To mention a few, DPO (Su & Lu, 2022) decomposes the joint TRPO goal into a cumulative sum of individual goals related only to the local information of each agent and proposes an algorithm with convergence guarantee in a fully decentralized learning setting. In addition, a series of scalable MARL theoretical results (Qu et al., 2020; Lin et al., 2021; Ying et al., 2023) under spatial correlation decay property are presented, which gives theoretical guarantees for decentralized algorithms in such cases. Unfortunately, the policy interactions or interference between neighboring agents has not yet been addressed in the latest results of the above studies, which is one of the motivations for the research in this paper.

6. Conclusion

Safety is a tremendous challenge for MARL when applied to real-world applications. In this paper, we quantize the approximation errors arising from both policy implementation and advantage estimation and then derive a novel decentralized lower bound for joint policy improvement and an upper bound for the safety constraints. Furthermore, we propose a novel decentralized and theoretically-justified multi-agent policy optimization method that follows a sequential update scheme to optimize κ-hop policies under a form of correlation decay property. Finally, we propose a practical decentralized constrained policy optimization algorithm called DEC-MAPPO-L and experimentally validate the effectiveness of the proposed algorithm on a collection of benchmark tasks.

References

Achiam, J., Held, D., Tamar, A., and Abbeel, P. Constrained policy optimization. In *ICML*, pp. 22–31, 2017.

Alfano, C. and Rebeschini, P. Dimension-free rates for natural policy gradient in multi-agent reinforcement learning. *arXiv preprint arXiv:2109.11692*, 2021.

Brunke, L., Greeff, M., Hall, A. W., Yuan, Z., Zhou, S., Panerati, J., and Schoellig, A. P. Safe learning in robotics: From learning-based control to safe reinforcement learning. *Annual Review of Control, Robotics, & Autonomous Systems*, 5:411–444, 2022.

Chen, Y.-J., Chang, D.-K., and Zhang, C. Autonomous tracking using a swarm of uavs: A constrained multi-agent reinforcement learning approach. *IEEE Transactions on Vehicular Technology*, 69(11):13702–13717, 2020.

Chow, Y., Nachum, O., Duenez-Guzman, E., and Ghavamzadeh, M. A lyapunov-based approach to safe reinforcement learning. In *NeurIPS*, pp. 8103–8112, 2018.

Chu, T., Chinchali, S., and Katti, S. Multi-agent reinforcement learning for networked system control. In *ICLR*, 2019.

Cui, W., Li, J., and Zhang, B. Decentralized safe reinforcement learning for inverter-based voltage control. *Electric Power Systems Research*, 211:108609, 2022.

Datta, S., Li, Y., Ruppert, M. M., Ren, Y., Shickel, B., Ozrazgat-Baslanti, T., Rashidi, P., and Bihorac, A. Reinforcement learning in surgery. *Surgery*, 170(1):329–332, 2021.

de Witt, C. S., Gupta, T., Makoviichuk, D., Makoviychuk, V., Torr, P. H., Sun, M., and Whiteson, S. Is independent learning all you need in the starcraft multi-agent challenge? *arXiv preprint arXiv:2011.09533*, 2020.

Dembo, A. and Montanari, A. Gibbs measures and phase transitions on sparse random graphs. *arXiv preprint arXiv:0910.5460*, 2009.

Ding, D., Zhang, K., Basar, T., and Jovanovic, M. R. Natural policy gradient primal-dual method for constrained markov decision processes. In *NeurIPS*, pp. 8378–8390, 2020.

Du, W. and Ding, S. A survey on multi-agent deep reinforcement learning: from the perspective of challenges and applications. *Artificial Intelligence Review*, 54:3215–3238, 2021.

Gallici, M., Martín Muñoz, M., and Masmitja Rusiñol, I. Transfqmix: transformers for leveraging the graph structure of multi-agent reinforcement learning problems. In *AAMAS*, pp. 1679–1687, 2023.

Gamarnik, D. Correlation decay method for decision, optimization, and inference in large-scale networks. In *Theory Driven by Influential Applications*, pp. 108–121. 2013.

García, J. and Shafie, D. Teaching a humanoid robot to walk faster through safe reinforcement learning. *Engineering Applications of Artificial Intelligence*, 88:103360, 2020.

Gronauer, S. and Diepold, K. Multi-agent deep reinforcement learning: a survey. *Artificial Intelligence Review*, pp. 1–49, 2022.

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

Gu, S., Chen, G., Zhang, L., Hou, J., Hu, Y., and Knoll, A. Constrained reinforcement learning for vehicle motion planning with topological reachability analysis. *Robotics*, 11(4):81, 2022a.

Gu, S., Yang, L., Du, Y., Chen, G., Walter, F., Wang, J., Yang, Y., and Knoll, A. A review of safe reinforcement learning: Methods, theory and applications. *arXiv preprint arXiv:2205.10330*, 2022b.

Gu, S., Kuba, J. G., Chen, Y., Du, Y., Yang, L., Knoll, A., and Yang, Y. Safe multi-agent reinforcement learning for multi-robot control. *Artificial Intelligence*, 319:103905, 2023.

Hsu, K.-C., Ren, A. Z., Nguyen, D. P., Majumdar, A., and Fisac, J. F. Sim-to-lab-to-real: Safe reinforcement learning with shielding and generalization guarantees. *Artificial Intelligence*, 314:103811, 2023.

Jiang, J., Dun, C., Huang, T., and Lu, Z. Graph convolutional reinforcement learning. In *ICLR*, 2019.

Kuba, J., Chen, R., Wen, M., Wen, Y., Sun, F., Wang, J., and Yang, Y. Trust region policy optimisation in multi-agent reinforcement learning. In *ICLR*, pp. 1046, 2022.

Lin, Y., Qu, G., Huang, L., and Wierman, A. Multi-agent reinforcement learning in stochastic networked systems. In *NeurIPS*, 2021.

Liu, C., Geng, N., Aggarwal, V., Lan, T., Yang, Y., and Xu, M. Cmix: Deep multi-agent reinforcement learning with peak and average constraints. In *ECML-PKDD*, pp. 157–173, 2021.

Lu, S., Zhang, K., Chen, T., Başar, T., and Horesh, L. Decentralized policy gradient descent ascent for safe multi-agent reinforcement learning. In *AAAI*, volume 35, pp. 8767–8775, 2021.

Mei, W., Mohagheghi, S., Zampieri, S., and Bullo, F. On the dynamics of deterministic epidemic propagation over networks. *Annual Reviews in Control*, 44:116–128, 2017.

Melcer, D., Amato, C., and Tripakis, S. Shield decentralization for safe multi-agent reinforcement learning. In *NeurIPS*, 2022.

Munir, M. S., Tran, N. H., Saad, W., and Hong, C. S. Multi-agent meta-reinforcement learning for self-powered and sustainable edge computing systems. *IEEE Transactions on Network and Service Management*, 18(3):3353–3374, 2021.

Oroojlooy, A. and Hajinezhad, D. A review of cooperative multi-agent deep reinforcement learning. *Applied Intelligence*, 53(11):13677–13722, 2023.

Peng, B., Rashid, T., de Witt, C. S., Kamienny, P.-A., Torr, P., Boehmer, W., and Whiteson, S. Facmac: Factored multi-agent centralised policy gradients. In *NeurIPS*, 2021.

Qu, G. and Li, N. Exploiting fast decaying and locality in multi-agent mdp with tree dependence structure. In *CDC*, pp. 6479–6486, 2019.

Qu, G., Lin, Y., Wierman, A., and Li, N. Scalable multi-agent reinforcement learning for networked systems with average reward. In *NeurIPS*, 2020.

Rashid, T., De Witt, C., Farquhar, G., Foerster, J., Whiteson, S., and Samvelyan, M. Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning. In *ICML*, pp. 6846–6859, 2018.

Schulman, J., Levine, S., Abbeel, P., Jordan, M., and Moritz, P. Trust region policy optimization. In *ICML*, pp. 1889–1897, 2015.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Sootla, A., Cowen-Rivers, A. I., Jafferjee, T., Wang, Z., Mguni, D. H., Wang, J., and Ammar, H. Sauté rl: Almost surely safe reinforcement learning using state augmentation. In *ICML*, pp. 20423–20443, 2022.

Su, K. and Lu, Z. Decentralized policy optimization. *arXiv preprint arXiv:2211.03032*, 2022.

Sunehag, P., Lever, G., Gruslys, A., Czarnecki, W. M., Zambaldi, V., Jaderberg, M., Lanctot, M., Sonnerat, N., Leibo, J. Z., Tuyls, K., et al. Value-decomposition networks for cooperative multi-agent learning. *arXiv preprint arXiv:1706.05296*, 2017.

Tampuu, A., Matiisen, T., Kodelja, D., Kuzovkin, I., Korjus, K., Aru, J., Aru, J., and Vicente, R. Multiagent cooperation and competition with deep reinforcement learning. *arXiv preprint arXiv:1511.08779*, 2015.

Tan, M. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *ICML*, pp. 330–337, 1993.

Whiteson, S., Samvelyan, M., Rashid, T., De Witt, C., Farquhar, G., Nardelli, N., Rudner, T., Hung, C., Torr, P., and Foerster, J. The starcraft multi-agent challenge. In *AAMAS*, pp. 2186–2188, 2019.

Yang, Y., Luo, R., Li, M., Zhou, M., Zhang, W., and Wang, J. Mean field multi-agent reinforcement learning. In *NeurIPS*, pp. 5571–5580, 2018.

Ying, D., Zhang, Y., Ding, Y., Koppel, A., and Lavaei, J. Scalable primal-dual actor-critic method for safe multi-agent rl with general utilities. *arXiv preprint arXiv:2305.17568*, 2023.

Yu, C., Velu, A., Vinitsky, E., Gao, J., Wang, Y., Bayen, A., and Wu, Y. The surprising effectiveness of ppo in cooperative multi-agent games. In *NeurIPS*, 2022.

Zhang, K., Yang, Z., and Başar, T. Multi-agent reinforcement learning: A selective overview of theories and algorithms. *Studies in Systems, Decision and Control*, pp. 321–384, 2021.

Zhang, W., Bastani, O., and Kumar, V. Mamps: Safe multi-agent reinforcement learning via model predictive shielding. *arXiv preprint arXiv:1910.12639*, 2019.

Zhou, W., Chen, D., Yan, J., Li, Z., Yin, H., and Ge, W. Multi-agent reinforcement learning for cooperative lane changing of connected and autonomous vehicles in mixed traffic. *Autonomous Intelligent Systems*, 2(1):5, 2022.

Zocca, A. Temporal starvation in multi-channel csma networks: an analytical framework. *ACM SIGMETRICS Performance Evaluation Review*, 46(3):52–53, 2019.

A. Supplementary materials for Section 2

A.1. The proof of Lamma 2.1

Proof. We write the multi-agent advantage function as in its definition, and then expand it in a telescoping sum.

$$\begin{aligned} A_{\boldsymbol{\pi}}\left(\mathbf{s},\mathbf{a}\right) &= Q_{\boldsymbol{\pi}}\left(\mathbf{s},\mathbf{a}\right) - V_{\boldsymbol{\pi}}(\mathbf{s}) \\ &= \sum_{i=1}^{n}\left[Q_{\boldsymbol{\pi}}^{1:i}\left(\mathbf{s},\mathbf{a}^{1:i}\right) - Q_{\boldsymbol{\pi}}^{1:i-1}\left(\mathbf{s},\mathbf{a}^{1:i-1}\right)\right] \\ &= \sum_{i=1}^{n} A_{\boldsymbol{\pi}}^{i}\left(\mathbf{s},\mathbf{a}^{1:i-1},a^{i}\right). \end{aligned} \tag{24}$$

□

A.2. The proof of Proposition 2.3

Firstly, by generalizing the result about the surrogate return in Equation (3), one can derive how the expected costs change when the agents update their policies. Inspired by (Gu et al., 2023), we provide the following lemma.

Lemma A.1. *Let $\boldsymbol{\pi}$ and $\bar{\boldsymbol{\pi}}$ be joint policies. Let $i \in \mathcal{N}$ be an agent, and $j \in \{1, \ldots, m^i\}$ be an index of one of its costs. The following inequality holds*

$$J_j^i(\bar{\boldsymbol{\pi}}) \leq J_j^i(\boldsymbol{\pi}) + L_{j,\boldsymbol{\pi}}^i\left(\bar{\pi}^i\right) + \nu_j^i \sum_{h=1}^{i} D_{\mathrm{KL}}^{\max}\left(\pi^h, \bar{\pi}^h\right). \tag{25}$$

where $L_{j,\boldsymbol{\pi}}^i(\bar{\pi}^i) = \mathbb{E}_{\mathbf{s}\sim\rho_{\boldsymbol{\pi}},\mathrm{a}^i\sim\bar{\pi}^i}\left[A_{j,\boldsymbol{\pi}}^i\left(\mathbf{s},\mathrm{a}^i\right)\right]$, $\nu_j^i = \frac{2\gamma\max_{\mathbf{s},\mathrm{a}^i}\left|A_{j,\pi}^i\left(\mathbf{s},\mathrm{a}^i\right)\right|}{(1-\gamma)^2}$.

Proof. From the upper bound version of Theorem 1 of (Schulman et al., 2015) applied to joint policies $\bar{\boldsymbol{\pi}}$ and $\boldsymbol{\pi}$, we conclude that

$$J_j^i(\bar{\boldsymbol{\pi}}) \leq J_j^i(\boldsymbol{\pi}) + \mathbb{E}_{\mathbf{s}\sim\rho_{\boldsymbol{\pi}},\mathbf{a}^{1:i}\sim\bar{\boldsymbol{\pi}}^{1:i}}\left[A_{j,\boldsymbol{\pi}}^i\left(\mathbf{s},\mathrm{a}^i\right)\right] + \frac{4\alpha^2\gamma\max_{\mathbf{s},\mathrm{a}^i}\left|A_{j,\boldsymbol{\pi}}^i\left(\mathbf{s},\mathrm{a}^i\right)\right|}{(1-\gamma)^2}, \tag{26}$$

where $\alpha = D_{\mathrm{TV}}^{\max}(\boldsymbol{\pi}^{1:i}, \bar{\boldsymbol{\pi}}^{1:i}) = \max_{\mathbf{s}} D_{\mathrm{TV}}(\boldsymbol{\pi}^{1:i}(\cdot \mid \mathbf{s}), \bar{\boldsymbol{\pi}}^{1:i}(\cdot \mid \mathbf{s}))$.

Then, using Pinsker's inequality $D_{\mathrm{TV}}(p,q)^2 \leq D_{\mathrm{KL}}(p,q)/2$, we obtain

$$J_j^i(\bar{\boldsymbol{\pi}}) \leq J_j^i(\boldsymbol{\pi}) + \mathbb{E}_{\mathbf{s}\sim\rho_{\boldsymbol{\pi}},\mathbf{a}^{1:i}\sim\bar{\boldsymbol{\pi}}^{1:i}}\left[A_{j,\boldsymbol{\pi}}^i\left(\mathbf{s},\mathrm{a}^i\right)\right] + \frac{2\gamma\max_{\mathbf{s},\mathrm{a}^i}\left|A_{j,\pi}^i\left(\mathbf{s},\mathrm{a}^i\right)\right|}{(1-\gamma)^2} D_{\mathrm{TV}}^{\max}(\boldsymbol{\pi}^{1:i}, \bar{\boldsymbol{\pi}}^{1:i}), \tag{27}$$

where $D_{\mathrm{KL}}^{\max}(\boldsymbol{\pi}^{1:i}, \bar{\boldsymbol{\pi}}^{1:i}) = \max_{\mathbf{s}} D_{\mathrm{KL}}(\boldsymbol{\pi}^{1:i}(\cdot \mid \mathbf{s}), \bar{\boldsymbol{\pi}}^{1:i}(\cdot \mid \mathbf{s}))$.

Notice that we have $\mathbb{E}_{\mathbf{s}\sim\rho_{\boldsymbol{\pi}},\mathbf{a}^{1:i}\sim\bar{\boldsymbol{\pi}}^{1:i}}\left[A_{j,\boldsymbol{\pi}}^i\left(\mathbf{s},\mathrm{a}^i\right)\right] = \mathbb{E}_{\mathbf{s}\sim\rho_{\boldsymbol{\pi}},\mathrm{a}^i\sim\bar{\pi}^i}\left[A_{j,\boldsymbol{\pi}}^i\left(\mathbf{s},\mathrm{a}^i\right)\right]$ as the action of agents other that i do not change the value of the variable inside of the expectation. Furthermore,

$$\begin{aligned} D_{\mathrm{KL}}^{\max}(\boldsymbol{\pi}^{1:i}, \bar{\boldsymbol{\pi}}^{1:i}) &= \max_{\mathbf{s}} D_{\mathrm{KL}}(\boldsymbol{\pi}^{1:i}(\cdot \mid \mathbf{s}), \bar{\boldsymbol{\pi}}^{1:i}(\cdot \mid \mathbf{s})) \\ &\leq \max_{\mathbf{s}}\left(\sum_{h=1}^{i} D_{\mathrm{KL}}\left(\pi^h(\cdot \mid \mathbf{s}), \bar{\pi}^h(\cdot \mid \mathbf{s})\right)\right) \\ &\leq \sum_{h=1}^{i}\max_{\mathbf{s}}\left(D_{\mathrm{KL}}\left(\pi^h(\cdot \mid \mathbf{s}), \bar{\pi}^h(\cdot \mid \mathbf{s})\right)\right) \\ &= \sum_{h=1}^{i} D_{\mathrm{KL}}^{\max}\left(\pi^h, \bar{\pi}^h\right). \end{aligned} \tag{28}$$

Setting $\nu_j^i = \frac{2\gamma \max_{\mathbf{s},\mathrm{a}^i} \left| A_{j,\pi}^i\left(\mathbf{s},\mathrm{a}^i\right)\right|}{(1-\gamma)^2}$, we finally obtain

$$J_j^i(\bar{\boldsymbol{\pi}}) \leq J_j^i(\boldsymbol{\pi}) + L_{j,\boldsymbol{\pi}}^i\left(\bar{\pi}^i\right) + \nu_j^i \sum_{h=1}^{i} D_{\mathrm{KL}}^{\max}\left(\pi^h, \bar{\pi}^h\right). \tag{29}$$

□

Then, we compute the size of KL constraint in Equation (9) as

$$\delta^i = \min_{h \leq i-1} \min_{1 \leq j \leq m^h} \frac{c_j^h - J_j^h(\boldsymbol{\pi}) - L_{j,\pi}^h\left(\bar{\pi}^h\right) - \nu_j^h \sum_{l=1}^{i-1} D_{\mathrm{KL}}^{\max}\left(\pi^l, \bar{\pi}^l\right)}{\nu_j^h}. \tag{30}$$

Note that δ^1 is guaranteed to be non-negative if $\boldsymbol{\pi}$ satisfies safety constraints; that is because then $c_j^h \geq J_j^h(\boldsymbol{\pi})$ for all $h \in \mathcal{N}$, and $j \in \left\{1, \ldots, m^i\right\}$, and the set $\{h \mid h < i\}$ is empty.

This formula for δ^i, combined with Lemma A.1, assures that the policies π^i within δ^i max-KL distance from π^i will not violate other agents' safety constraints, as long as the base joint policy $\boldsymbol{\pi}$ did not violate them (which assures $\delta^1 \geq 0$). To see this, notice that for every $h = 1, \ldots, i-1$, and $j = 1, \ldots, m^h$,

$$\begin{aligned}
&D_{\mathrm{KL}}^{\max}\left(\pi^i, \bar{\pi}^i\right) \leq \delta^i \leq \frac{c_j^h - J_j^h(\boldsymbol{\pi}) - L_{j,\boldsymbol{\pi}}^h\left(\bar{\pi}^h\right) - \nu_j^h \sum_{l=1}^{i-1} D_{\mathrm{KL}}^{\max}\left(\pi^l, \bar{\pi}^l\right)}{\nu_j^h}, \\
&\text{implies } J_j^h(\boldsymbol{\pi}) + L_{j,\boldsymbol{\pi}}^h\left(\bar{\pi}^h\right) + \nu_j^h \sum_{l=1}^{i-1} D_{\mathrm{KL}}^{\max}\left(\pi^l, \bar{\pi}^l\right) + \nu_j^h D_{\mathrm{KL}}^{\max}\left(\pi^i, \hat{\pi}^i\right) \leq c_j^h.
\end{aligned} \tag{31}$$

By Lemma A.1, the left-hand side of the above inequality is an upper bound of $J_j^h\left(\bar{\boldsymbol{\pi}}^{1:i-1}, \pi^i\right)$, which implies that the update of agent i does not violate the constraint of J_j^h.

Finally, we can obtain that the joint policy $\boldsymbol{\pi}$ has the monotonic improvement property when the agents' update size is bounded by Equation (9). By Theorem 1 from (Schulman et al., 2015), we have

$$\begin{aligned}
J(\bar{\boldsymbol{\pi}}) - J(\boldsymbol{\pi}) &\geq \mathbb{E}_{\mathbf{s}\sim\boldsymbol{\rho}_{\boldsymbol{\pi}}, \mathbf{a}\sim\bar{\boldsymbol{\pi}}}\left[A_{\boldsymbol{\pi}}(\mathbf{s}, \mathbf{a})\right] - \nu D_{\mathrm{KL}}^{\max}(\boldsymbol{\pi}, \bar{\boldsymbol{\pi}}) \\
&\geq \mathbb{E}_{\mathbf{s}\sim\boldsymbol{\rho}_{\boldsymbol{\pi}}, \mathbf{a}\sim\bar{\boldsymbol{\pi}}}\left[A_{\boldsymbol{\pi}}(\mathbf{s}, \mathbf{a})\right] - \sum_{i=1}^{n} \nu D_{\mathrm{KL}}^{\max}\left(\pi^i, \bar{\pi}^i\right) \\
&= \sum_{i=1}^{n} \mathbb{E}_{\mathbf{s}\sim\boldsymbol{\rho}_{\boldsymbol{\pi}}, \mathbf{a}^{1:i}\sim\boldsymbol{\pi}^{1:i}}\left[A_{\boldsymbol{\pi}}^i\left(\mathbf{s}, \mathbf{a}^{1:i-1}, a^i\right)\right] - \sum_{i=1}^{n} \nu D_{\mathrm{KL}}^{\max}\left(\pi^i, \bar{\pi}^i\right) \\
&= \sum_{i=1}^{n} \left(L_{\boldsymbol{\pi}}^{1:i}\left(\bar{\boldsymbol{\pi}}^{1:i-1}, \bar{\pi}^i\right) - \nu D_{\mathrm{KL}}^{\max}\left(\pi^i, \bar{\pi}^i\right)\right) \\
&\geq \sum_{i=1}^{n} \left(L_{\boldsymbol{\pi}}^{1:i}\left(\bar{\boldsymbol{\pi}}^{1:i-1}, \pi^i\right) - \nu D_{\mathrm{KL}}^{\max}\left(\pi^i, \bar{\pi}^i\right)\right) \\
&= \sum_{i=1}^{n} 0 \\
&= 0.
\end{aligned} \tag{32}$$

B. Supplementary materials for Section 3

B.1. Example of Assumption 3.2

Firstly, we start from Assumption 3.1, letting $\widetilde{\kappa} = \max_{i,j\in\mathcal{N}} d(i,j)$ be the maximum distance between agent i and agent j. Define a set of differentiable functions $\left\{f_\kappa : \mathcal{S}_{\mathcal{N}_\kappa^i} \times \mathcal{A}_\kappa \rightarrow \mathcal{K} \mid 0 \leq \kappa \leq \widetilde{\kappa}\right\}$, where $\mathcal{K} \subset [-K, K]$ and $K > 0$, a set of parameters $\{\alpha_\kappa \geq 0 \mid 0 \leq \kappa \leq \widetilde{\kappa}\}$ and let, for each agent i,

$$\begin{aligned} f^i\left(\mathbf{s}, \mathrm{a}^i\right) &= \sum_{\kappa=0}^{\widetilde{\kappa}} \alpha_\kappa f^i_\kappa\left(\mathrm{s}_{\mathcal{N}^i_\kappa}, \mathrm{a}^i\right), \\ \pi^i\left(\mathrm{a} \mid \mathbf{s}\right) &= \frac{\exp\left(f^i\left(\mathbf{s}, \mathrm{a}\right)\right)}{\sum_{\mathrm{a}' \in \mathcal{A}^i} \exp\left(f^i\left(\mathbf{s}, \mathrm{a}'\right)\right)}. \end{aligned} \tag{33}$$

Then, by tuning the parameters α_κ, we can make any policy belonging to this policy class respect Assumptions 3.2 , as we show in the following. Let $\kappa \in \{0, \ldots, \widetilde{\kappa}\}$, let $\mathbf{s}, \widetilde{\mathbf{s}} \in \mathcal{S}$ be such that $\mathrm{s}_{\mathcal{N}^i_\kappa} = \widetilde{\mathrm{s}}_{\mathcal{N}^i_\kappa}$, then

$$\begin{aligned} \left\|\pi^i(\cdot|\mathbf{s}) - \pi^i(\cdot|\widetilde{\mathbf{s}})\right\|_1 &= \sum_{a \in \mathcal{A}^i} \left|\pi^i(\mathrm{a} \mid \mathbf{s}) - \pi^i(\mathrm{a} \mid \widetilde{\mathbf{s}})\right| \\ &= \sum_{\mathrm{a} \in \mathcal{A}^i} \left| \frac{\exp\left(f^i\left(\mathbf{s}, \mathrm{a}\right)\right)}{\sum_{\mathrm{a}' \in \mathcal{A}^i} \exp\left(f^i\left(\mathbf{s}, \mathrm{a}'\right)\right)} - \frac{\exp\left(f^i\left(\widetilde{\mathbf{s}}, \mathrm{a}\right)\right)}{\sum_{\mathrm{a}' \in \mathcal{A}^i} \exp\left(f^i\left(\widetilde{\mathbf{s}}, \mathrm{a}'\right)\right)} \right| \\ &= \frac{\sum_{\mathrm{a} \in \mathcal{A}^i} \left| \sum_{\mathrm{a}' \in \mathcal{A}^i} \exp\left(f^i\left(\mathbf{s}, \mathrm{a}\right)\right) \exp\left(f^i\left(\widetilde{\mathbf{s}}, \mathrm{a}'\right)\right) - \sum_{\mathrm{a}' \in \mathcal{A}^i} \exp\left(f^i\left(\widetilde{\mathbf{s}}, \mathrm{a}\right)\right) \exp\left(f^i\left(\mathbf{s}, \mathrm{a}'\right)\right) \right|}{\sum_{\mathrm{a}' \in \mathcal{A}^i} \exp\left(f^i\left(\mathbf{s}, \mathrm{a}'\right)\right) \sum_{\mathrm{a}' \in \mathcal{A}^i} \exp\left(f^i\left(\widetilde{\mathbf{s}}, \mathrm{a}'\right)\right)} \\ &\leq \frac{\sum_{\mathrm{a} \in \mathcal{A}^i} \sum_{\mathrm{a}' \in \mathcal{A}^i} \left| \exp\left(f^i\left(\mathbf{s}, \mathrm{a}\right)\right) \exp\left(f^i\left(\widetilde{\mathbf{s}}, \mathrm{a}'\right)\right) - \exp\left(f^i\left(\widetilde{\mathbf{s}}, \mathrm{a}\right)\right) \exp\left(f^i\left(\mathbf{s}, \mathrm{a}'\right)\right) \right|}{\sum_{\mathrm{a}' \in \mathcal{A}^i} \exp\left(f^i\left(\mathbf{s}, \mathrm{a}'\right)\right) \sum_{\mathrm{a}' \in \mathcal{A}^i} \exp\left(f^i\left(\widetilde{\mathbf{s}}, \mathrm{a}'\right)\right)} \\ &\leq \frac{\sum_{\mathrm{a} \in \mathcal{A}^i} \left| \exp\left(f^i(\widetilde{\mathbf{s}}, \mathrm{a})\right) - \exp\left(f^i(\mathbf{s}, \mathrm{a})\right) \right|}{\sum_{\mathrm{a} \in \mathcal{A}^i} \exp\left(f^i(\widetilde{\mathbf{s}}, \mathrm{a})\right)} \\ &\leq \frac{\sum_{\mathrm{a} \in \mathcal{A}^i} \left| f^i(\widetilde{\mathbf{s}}, \mathrm{a}) - f^i(\mathbf{s}, \mathrm{a}) \right| \exp\left(\sup_{s' \in \{\mathbf{s}, \widetilde{\mathbf{s}}\}} f^i\left(\mathbf{s}', \mathrm{a}\right) \right)}{\sum_{\mathrm{a} \in \mathcal{A}^i} \exp\left(f^i(\widetilde{\mathbf{s}}, \mathrm{a})\right)} \\ &\leq e^{2K(\widetilde{\kappa} - \kappa)} \frac{\sum_{\mathrm{a} \in \mathcal{A}^i} \left| f^i(\widetilde{\mathbf{s}}, \mathrm{a}) - f^i(\mathbf{s}, \mathrm{a}) \right| \exp\left(f^i(\widetilde{\mathbf{s}}, a)\right)}{\sum_{\mathrm{a} \in \mathcal{A}^i} \exp\left(f^i(\widetilde{\mathbf{s}}, \mathrm{a})\right)} \\ &\leq e^{2K(\widetilde{\kappa} - \kappa)} \mathbb{E}_{\pi^i} \left| \sum_{\kappa' = \kappa + 1}^{\widetilde{\kappa}} \alpha_{\kappa'} \left(f^i_{\kappa'}\left(\widetilde{\mathrm{s}}_{\mathcal{N}^i_{\kappa'}}, \mathrm{a}\right) - f^i_{\kappa'}\left(\mathrm{s}_{\mathcal{N}^i_{\kappa'}}, \mathrm{a}\right) \right) \right| \\ &\leq e^{2K(\widetilde{\kappa} - \kappa)} \sum_{\kappa' = \kappa + 1}^{\widetilde{\kappa}} \alpha_{\kappa'} \mathbb{E}_{\pi^i} \left| \left(f^i_{\kappa'}\left(\widetilde{\mathrm{s}}_{\mathcal{N}^i_{\kappa'}}, \mathrm{a}\right) - f^i_{\kappa'}\left(\mathrm{s}_{\mathcal{N}^i_{\kappa'}}, \mathrm{a}\right) \right) \right| \\ &\leq 2K e^{2K(\widetilde{\kappa} - \kappa)} \sum_{\kappa' = \kappa + 1}^{\widetilde{\kappa}} \alpha_{\kappa'}. \end{aligned} \tag{34}$$

Denote that $(\xi, \beta) = \left(2K e^{2K\widetilde{\kappa}} \sum_{\kappa'=\kappa+1}^{\widetilde{\kappa}} \alpha_{\kappa'}, e^{-2K\kappa}\right)$, and setting the parameters $\{\alpha_{\kappa'}\}_{\kappa' \in \{\kappa+1, \ldots, \widetilde{\kappa}\}}$ small enough ensures that the policy respects Assumption 3.2.

B.2. The proofs of Proposition 3.3

Proof. The following holds for every $i \in \mathcal{N}$. Let $\mathbf{s}, \widetilde{\mathbf{s}} \in \mathcal{S}$, $\mathbf{a}, \widetilde{\mathbf{a}} \in \mathcal{A}$ be such that $\mathrm{s}_{\mathcal{N}^i_\kappa} = \widetilde{\mathrm{s}}_{\mathcal{N}^i_\kappa}$ and $\mathrm{a}_{\mathcal{N}^i_\kappa} = \widetilde{\mathrm{a}}_{\mathcal{N}^i_\kappa}$. Notice that

$$\begin{aligned} \left| A^i_{\boldsymbol{\pi}}(\mathbf{s}, \mathbf{a}) - A^i_{\boldsymbol{\pi}}(\widetilde{\mathbf{s}}, \widetilde{\mathbf{a}}) \right| &= \left| \left(Q^i_{\boldsymbol{\pi}}(\mathbf{s}, \mathbf{a}) - V^i_{\boldsymbol{\pi}}(\mathbf{s}) \right) - \left(Q^i_{\boldsymbol{\pi}}(\widetilde{\mathbf{s}}, \widetilde{\mathbf{a}}) - V^i_{\boldsymbol{\pi}}(\widetilde{\mathbf{s}}) \right) \right| \\ &= \left| \left(Q^i_{\boldsymbol{\pi}}(\mathbf{s}, \mathbf{a}) - Q^i_{\boldsymbol{\pi}}(\widetilde{\mathbf{s}}, \widetilde{\mathbf{a}}) \right) + \left(V^i_{\boldsymbol{\pi}}(\mathbf{s}) - V^i_{\boldsymbol{\pi}}(\widetilde{\mathbf{s}}) \right) \right| \\ &\leq \left| Q^i_{\boldsymbol{\pi}}(\mathbf{s}, \mathbf{a}) - Q^i_{\boldsymbol{\pi}}(\widetilde{\mathbf{s}}, \widetilde{\mathbf{a}}) \right| + \left| V^i_{\boldsymbol{\pi}}(\mathbf{s}) - V^i_{\boldsymbol{\pi}}(\widetilde{\mathbf{s}}) \right| \end{aligned} \tag{35}$$

Next, we analyze $\left| Q^i_{\boldsymbol{\pi}}(\mathbf{s}, \mathbf{a}) - Q^i_{\boldsymbol{\pi}}(\widetilde{\mathbf{s}}, \widetilde{\mathbf{a}}) \right|$ and $\left| V^i_{\boldsymbol{\pi}}(\mathbf{s}) - V^i_{\boldsymbol{\pi}}(\widetilde{\mathbf{s}}) \right|$ separately.

Firstly, we have

$$
\begin{aligned}
&\left|Q^i_{\boldsymbol{\pi}}(\mathbf{s},\mathbf{a}) - Q^i_{\boldsymbol{\pi}}(\widetilde{\mathbf{s}},\widetilde{\mathbf{a}})\right| \\
&= \left|\sum_{t=0}^{\infty}\gamma^t \mathbb{E}\left[r\left(\mathbf{s}_t,\mathbf{a}_t\right) \mid \boldsymbol{\pi}, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a}\right] - \sum_{t=0}^{\infty}\gamma^t \mathbb{E}\left[r\left(\mathbf{s}_t,\mathbf{a}_t\right) \mid \boldsymbol{\pi}, \mathbf{s}_0=\widetilde{\mathbf{s}}, \mathbf{a}_0=\widetilde{\mathbf{a}}\right]\right| \\
&\leq \sum_{t=0}^{\infty}\gamma^t \left|\mathbb{E}\left[r\left(\mathbf{s}_t,\mathbf{a}_t\right) \mid \boldsymbol{\pi}, \mathbf{s}_0=\mathbf{s}, \mathbf{a}_0=\mathbf{a}\right] - \mathbb{E}\left[r\left(\mathbf{s}_t,\mathbf{a}_t\right) \mid \boldsymbol{\pi}, \mathbf{s}_0=\widetilde{\mathbf{s}}, \mathbf{a}_0=\widetilde{\mathbf{a}}\right]\right| \\
&\leq \sum_{t=1}^{\infty}\gamma^t D_{\mathrm{TV}}\left(\rho^i_t, \widetilde{\rho}^i_t\right)
\end{aligned}
\tag{36}
$$

where ρ^i_t and $\widetilde{\rho}^i_t$ are the distributions at time t with starting point $(\mathbf{s},\mathbf{a})$ and $(\widetilde{\mathbf{s}},\widetilde{\mathbf{a}})$, respectively. We use the result in Lemma 13 of (Alfano & Rebeschini, 2021) to bound $D_{\mathrm{TV}}\left(\rho^i_t, \widetilde{\rho}^i_t\right)$. The structure of our MDP implies that:

$$
P(\mathbf{s}_{t+1},\mathbf{a}_{t+1} \mid \mathbf{s}_t,\mathbf{a}_t) = \prod_{i\in\mathcal{N}} \pi^i\left(\mathrm{a}^i_{t+1} \mid \mathrm{s}_{\mathcal{N}^i_\kappa, t+1}\right) P^i\left(\mathrm{s}^i_{t+1} \mid \mathrm{s}_{\mathcal{N}^i_\kappa, t}, \mathrm{a}^i_t\right). \tag{37}
$$

Then, if Assumption 3.1 holds, the requirements of Lemma 13 of (Alfano & Rebeschini, 2021) are satisfied. Therefore, $D_{\mathrm{TV}}\left(\rho^i_t, \widetilde{\rho}^i_t\right) \leq \zeta^t e^{-\beta\kappa}$ and

$$
\begin{aligned}
\left|Q^i_{\pi}(\mathbf{s},\mathbf{a}) - Q^i_{\pi}(\widetilde{\mathbf{s}},\widetilde{\mathbf{a}})\right| &\leq \sum_{t=1}^{\infty}\gamma^t D_{\mathrm{KL}}\left(\rho^i_t, \widetilde{\rho}^i_t\right) \\
&\leq e^{-\beta\kappa}\sum_{t=1}^{\infty}\gamma^t\zeta^t \\
&= \frac{\gamma\zeta}{1-\gamma\zeta}e^{-\beta\kappa}.
\end{aligned}
\tag{38}
$$

where ζ is defined in Assumption 3.1.

Moreover, according to Lemma 14 of in (Alfano & Rebeschini, 2021), we can obtain the following lemma.

Lemma B.1. *Consider the setting in (37). Let* $P^{i,t}\left(\mathrm{s}' \mid \mathrm{s}\right) = P\left(\mathrm{s}_t = \mathrm{s}' \mid \mathrm{s}_0 = \mathrm{s}\right)$ *and*

$$
\delta^j P^{i,t} = \sup_{\mathrm{s}^j, \mathrm{s}'^j, \mathbf{s}^{-j}} D_{\mathrm{TV}}\left(P^{i,t}\left(\cdot \mid \mathrm{s}^j, \mathbf{s}^{-j}\right), P^{i,t}\left(\cdot \mid \mathrm{s}'^{,j}, \mathbf{s}^{-j}\right)\right). \tag{39}
$$

If $\sum_{j\in\mathcal{N}} e^{\beta d(i,j)} C_{ij} \leq \zeta$, *we have*

$$
\sum_{j\in\mathcal{N}} e^{\beta d(i,j)} \delta^j P^{i,t} \leq \zeta^t, \forall\, i \in \mathcal{N} \tag{40}
$$

Let

$$
\delta^j Q^i_{\boldsymbol{\pi}}(\mathbf{s},\mathbf{a}) = \sup_{\mathrm{z}_j, \mathbf{z}_{-j}, \mathrm{z}'_j} \left|Q^i_{\boldsymbol{\pi}}\left(\mathrm{z}_j, \mathbf{z}_{-j}\right) - Q^i_{\boldsymbol{\pi}}\left(\mathrm{z}'_j, \mathbf{z}_{-j}\right)\right| \tag{41}
$$

Based this result of the exponential decay property for the Q-value function. We have already shown that the MDP satisfies the condition of Lemma B.1 . Using it, we can obtain

$$
\sum_{j\in\mathcal{N}} e^{\beta d(i,j)} \delta^j \left(Q^i_{\boldsymbol{\pi}}(\mathbf{s},\cdot)\right) \leq \sum_{t=1}^{\infty}\gamma^t \sum_{j\in\mathcal{N}} e^{\beta d(i,j)} \delta^j P^i \leq \sum_{t=1}^{\infty}\gamma^t\zeta^t = \frac{\gamma\zeta}{1-\gamma\zeta} \tag{42}
$$

Building on Assumption 3.2 and Lemma B.1, we have

$$
\begin{aligned}
\left|V_{\boldsymbol{\pi}}^{i}(\mathbf{s})-V_{\boldsymbol{\pi}}^{i}(\widetilde{\mathbf{s}})\right| &= \left|\mathbb{E}_{\mathbf{a}\sim\boldsymbol{\pi}(\cdot|\mathbf{s})}Q_{\boldsymbol{\pi}}^{i}(\mathbf{s},\mathbf{a})-\mathbb{E}_{\mathbf{a}\sim\boldsymbol{\pi}(\cdot|\widetilde{\mathbf{s}})}Q_{\boldsymbol{\pi}}^{i}(\widetilde{\mathbf{s}},\mathbf{a})\right| \\
&= \left|\mathbb{E}_{\mathbf{a}\sim\boldsymbol{\pi}(\cdot|\mathbf{s})}Q_{\boldsymbol{\pi}}^{i}(\mathbf{s},\mathbf{a})-\mathbb{E}_{\mathbf{a}\sim\boldsymbol{\pi}(\cdot|\widetilde{\mathbf{s}})}Q_{\boldsymbol{\pi}}^{i}(\mathbf{s},\mathbf{a})+\mathbb{E}_{\mathbf{a}\sim\boldsymbol{\pi}(\cdot|\widetilde{\mathbf{s}})}Q_{\boldsymbol{\pi}}^{i}(\mathbf{s},\mathbf{a})-\mathbb{E}_{\mathbf{a}\sim\boldsymbol{\pi}(\cdot|\widetilde{\mathbf{s}})}Q_{\boldsymbol{\pi}}^{i}(\widetilde{\mathbf{s}},\mathbf{a})\right| \\
&\leq \left|\mathbb{E}_{\mathbf{a}\sim\boldsymbol{\pi}(\cdot|\mathbf{s})}Q_{\boldsymbol{\pi}}^{i}(\mathbf{s},\mathbf{a})-\mathbb{E}_{\mathbf{a}\sim\boldsymbol{\pi}(\cdot|\widetilde{\mathbf{s}})}Q_{\boldsymbol{\pi}}^{i}(\mathbf{s},\mathbf{a})\right|+\left|\mathbb{E}_{\mathbf{a}\sim\boldsymbol{\pi}(\cdot|\widetilde{\mathbf{s}})}Q_{\boldsymbol{\pi}}^{i}(\mathbf{s},\mathbf{a})-\mathbb{E}_{\mathbf{a}\sim\boldsymbol{\pi}(\cdot|\widetilde{\mathbf{s}})}Q_{\boldsymbol{\pi}}^{i}(\widetilde{\mathbf{s}},\mathbf{a})\right| \\
&\leq \sum_{j\in\mathcal{N}} D_{\mathrm{KL}}\left(\pi_j(\cdot\mid\mathbf{s}),\pi_j(\cdot\mid\widetilde{\mathbf{s}})\right)\delta^{j}Q_{\boldsymbol{\pi}}^{i}(\mathbf{s},\mathbf{a})+\frac{\gamma\zeta}{1-\gamma\zeta}e^{-\beta\kappa} \\
&\leq \xi e^{-\beta\kappa}\sum_{j\in\mathcal{N}} e^{-\beta d(j,i)}\delta^{j}Q_{\boldsymbol{\pi}}^{i}(\mathbf{s},\mathbf{a})+\frac{\gamma\zeta}{1-\gamma\zeta}e^{-\beta\kappa} \\
&\leq \frac{\gamma\zeta}{1-\gamma\zeta}\xi e^{-\beta\kappa}+\frac{\gamma\zeta}{1-\gamma\zeta}e^{-\beta\kappa} \\
&\leq \frac{(1+\xi)\gamma\zeta}{1-\gamma\zeta}e^{-\beta\kappa}.
\end{aligned}
\tag{43}
$$

Then, bringing (38) and (43) into (35), we have

$$
\begin{aligned}
\left|A_{\boldsymbol{\pi}}^{i}(\mathbf{s},\mathbf{a})-A_{\boldsymbol{\pi}}^{i}(\widetilde{\mathbf{s}},\widetilde{\mathbf{a}})\right| &\leq \left|Q_{\boldsymbol{\pi}}^{i}(\mathbf{s},\mathbf{a})-Q_{\boldsymbol{\pi}}^{i}(\widetilde{\mathbf{s}},\widetilde{\mathbf{a}})\right|+\left|V_{\boldsymbol{\pi}}^{i}(\mathbf{s})-V_{\boldsymbol{\pi}}^{i}(\widetilde{\mathbf{s}})\right| \\
&\leq \left|\mathbb{E}_{\mathbf{a}\sim\boldsymbol{\pi}(\cdot|\mathbf{s})}Q_{\boldsymbol{\pi}}^{i}(\mathbf{s},\mathbf{a})-\mathbb{E}_{\mathbf{a}\sim\boldsymbol{\pi}(\cdot|\widetilde{\mathbf{s}})}Q_{\boldsymbol{\pi}}^{i}(\mathbf{s},\mathbf{a})\right|+2\left|\mathbb{E}_{\mathbf{a}\sim\boldsymbol{\pi}(\cdot|\widetilde{\mathbf{s}})}Q_{\boldsymbol{\pi}}^{i}(\mathbf{s},\mathbf{a})-\mathbb{E}_{\mathbf{a}\sim\boldsymbol{\pi}(\cdot|\widetilde{\mathbf{s}})}Q_{\boldsymbol{\pi}}^{i}(\widetilde{\mathbf{s}},\widetilde{\mathbf{a}})\right| \\
&\leq \frac{\gamma\zeta}{1-\gamma\zeta}\xi e^{-\beta\kappa}+\frac{2\gamma\zeta}{1-\gamma\zeta}e^{-\beta\kappa} \\
&\leq \frac{(2+\xi)\gamma\zeta}{1-\gamma\zeta}e^{-\beta\kappa}.
\end{aligned}
\tag{44}
$$

Finally, denoting $(c,\phi)=\left(\frac{(2+\xi)\gamma\zeta}{1-\gamma\zeta},e^{-\beta}\right)$, we can obtain the Proposition 3.3. □

B.3. The proof of Corollary 3.4

Proof. Firstly, according to Lamma 2.1, we have

$$
\begin{aligned}
&\left|L_{\boldsymbol{\pi}}^{1:i}\left(\bar{\boldsymbol{\pi}}^{1:i-1},\bar{\pi}^{i}\right)-L_{\pi_{\kappa}^{i}}^{i}\left(\bar{\pi}_{\kappa}^{i}\right)\right| \\
&=\left|\mathbb{E}_{\mathrm{s}\sim\boldsymbol{\rho}_{\boldsymbol{\pi}},\mathbf{a}^{1:i}\sim\bar{\boldsymbol{\pi}}^{1:i}}\left[A_{\boldsymbol{\pi}}^{i}\left(\mathbf{s},\mathbf{a}^{1:i-1},\mathrm{a}^{i}\right)\right]-\mathbb{E}_{\mathrm{s}_{\mathcal{N}_{\kappa}^{i}}\sim\rho_{\pi_{\kappa}^{i}},\mathrm{a}^{i}\sim\bar{\pi}_{\kappa}^{i}}\left[A_{\pi_{\kappa}^{i}}^{i}\left(\mathrm{s}_{\mathcal{N}_{\kappa}^{i}},\mathrm{a}^{i}\right)\right]\right| \\
&=\left|\mathbb{E}_{\mathbf{s}\sim\boldsymbol{\rho}_{\boldsymbol{\pi}},\mathbf{a}^{1:i}\sim\bar{\boldsymbol{\pi}}^{1:i}}\left[A_{\boldsymbol{\pi}}^{i}\left(\mathbf{s},\mathbf{a}^{1:i-1},\mathrm{a}^{i}\right)\right]-\mathbb{E}_{\mathbf{s}\sim\boldsymbol{\rho}_{\boldsymbol{\pi}},\mathrm{a}^{i}\sim\bar{\pi}^{i}}\left[A_{\boldsymbol{\pi}}^{i}\left(\mathbf{s},\mathrm{a}^{i}\right)\right]+\mathbb{E}_{\mathbf{s}\sim\boldsymbol{\rho}_{\boldsymbol{\pi}},\mathrm{a}^{i}\sim\bar{\pi}^{i}}\left[A_{\boldsymbol{\pi}}^{i}\left(\mathbf{s},\mathrm{a}^{i}\right)\right]-\mathbb{E}_{\widetilde{\mathbf{s}}\sim\widetilde{\boldsymbol{\rho}}_{\boldsymbol{\pi}},\mathrm{a}^{i}\sim\bar{\pi}_{\kappa}^{i}}\left[A_{\boldsymbol{\pi}}^{i}\left(\widetilde{\mathbf{s}},\mathrm{a}^{i}\right)\right]\right| \\
&\leq\left|\mathbb{E}_{\mathbf{s}\sim\boldsymbol{\rho}_{\boldsymbol{\pi}},\mathbf{a}^{1:i}\sim\bar{\boldsymbol{\pi}}^{1:i}}\left[A_{\boldsymbol{\pi}}^{i}\left(\mathbf{s},\mathbf{a}^{1:i-1},\mathrm{a}^{i}\right)\right]-\mathbb{E}_{\mathbf{s}\sim\boldsymbol{\rho}_{\boldsymbol{\pi}},\mathrm{a}^{i}\sim\bar{\pi}^{i}}\left[A_{\boldsymbol{\pi}}^{i}\left(\mathbf{s},\mathrm{a}^{i}\right)\right]\right|+\left|\mathbb{E}_{\mathbf{s}\sim\boldsymbol{\rho}_{\boldsymbol{\pi}},\mathrm{a}^{i}\sim\bar{\pi}^{i}}\left[A_{\boldsymbol{\pi}}^{i}\left(\mathbf{s},\mathrm{a}^{i}\right)\right]-\mathbb{E}_{\widetilde{\mathbf{s}}\sim\widetilde{\boldsymbol{\rho}}_{\boldsymbol{\pi}},\mathrm{a}^{i}\sim\bar{\pi}^{i}}\left[A_{\boldsymbol{\pi}}^{i}\left(\widetilde{\mathbf{s}},\mathrm{a}^{i}\right)\right]\right|
\end{aligned}
\tag{45}
$$

Then, using the results of Proposition 3.3, we have

$$
\begin{aligned}
&\left|\mathbb{E}_{\mathbf{s}\sim\boldsymbol{\rho}_{\boldsymbol{\pi}},\mathbf{a}^{1:i}\sim\bar{\boldsymbol{\pi}}^{1:i}}\left[A_{\boldsymbol{\pi}}^{i}\left(\mathbf{s},\mathbf{a}^{1:i-1},\mathrm{a}^{i}\right)\right]-\mathbb{E}_{\mathbf{s}\sim\boldsymbol{\rho}_{\boldsymbol{\pi}},\mathrm{a}^{i}\sim\bar{\pi}^{i}}\left[A_{\boldsymbol{\pi}}^{i}\left(\mathbf{s},\mathrm{a}^{i}\right)\right]\right| \\
&=\left|\mathbb{E}_{\mathbf{s}\sim\boldsymbol{\rho}_{\boldsymbol{\pi}},\mathrm{a}^{i}\sim\bar{\pi}^{i}}\left[\sum_{h=1}^{i-1}\left(\bar{\pi}^{h}-\pi^{h}\right)A_{\boldsymbol{\pi}}^{i}\left(\mathbf{s},\mathrm{a}^{i}\right)\right]\right| \\
&\leq\mathbb{E}_{\mathbf{s}\sim\boldsymbol{\rho}_{\boldsymbol{\pi}},\mathrm{a}^{i}\sim\bar{\pi}^{i}}\left[\sum_{h=1}^{i-1}\left|\bar{\pi}^{h}-\pi^{h}\right|\left|A_{\boldsymbol{\pi}}^{i}\left(\mathbf{s},\mathrm{a}^{i}\right)\right|\right] \\
&\leq\mathbb{E}_{\mathbf{s}\sim\boldsymbol{\rho}_{\boldsymbol{\pi}},\mathrm{a}^{i}\sim\bar{\pi}^{i}}\left[M\sum_{h=1}^{i-1}\left|\bar{\pi}^{h}-\pi^{h}\right|\right] \\
&\leq\frac{M^{i}}{1-\gamma}\sum_{h=1}^{i-1}\max_{\mathrm{s}}D_{\mathrm{TV}}(\bar{\pi}^{h},\pi^{h}) \\
&\leq\frac{M^{i}\xi}{1-\gamma}e^{-\beta\kappa},
\end{aligned}
\tag{46}
$$

where M^i is a constant. And, according to (44), we have

$$\begin{aligned}
&\left|\mathbb{E}_{\mathbf{s}\sim\boldsymbol{\rho}_{\boldsymbol{\pi}},\mathrm{a}^i\sim\bar{\pi}^i}\left[A^i_{\boldsymbol{\pi}}\left(\mathbf{s},\mathrm{a}^i\right)\right]-\mathbb{E}_{\widetilde{\mathbf{s}}\sim\widetilde{\boldsymbol{\rho}}_{\boldsymbol{\pi}},\mathrm{a}^i\sim\bar{\pi}^i}\left[A^i_{\boldsymbol{\pi}}\left(\widetilde{\mathbf{s}},\mathrm{a}^i\right)\right]\right|\\
&\leq\mathbb{E}\left|\left[A^i_{\boldsymbol{\pi}}(\mathbf{s},\mathbf{a})-A^i_{\boldsymbol{\pi}}(\widetilde{\mathbf{s}},\widetilde{\mathbf{a}})\right]\right|\\
&\leq\frac{(2+\xi)\gamma\zeta}{1-\gamma\zeta}e^{-\beta\kappa}.
\end{aligned}\tag{47}$$

Then, bringing (46) and (47) into (45), we have that

$$\begin{aligned}
&\left|L^{1:i}_{\boldsymbol{\pi}}\left(\bar{\boldsymbol{\pi}}^{1:i-1},\bar{\pi}^i\right)-L^i_{\boldsymbol{\pi}}\left(\bar{\pi}^i_\kappa\right)\right|\\
&\leq\left|\mathbb{E}_{\mathbf{s}\sim\boldsymbol{\rho}_{\boldsymbol{\pi}},\mathbf{a}^{1:i}\sim\bar{\boldsymbol{\pi}}^{1:i}}\left[A^i_{\boldsymbol{\pi}}\left(\mathbf{s},\mathbf{a}^{1:i-1},\mathrm{a}^i\right)\right]-\mathbb{E}_{\mathbf{s}\sim\boldsymbol{\rho}_{\boldsymbol{\pi}},\mathrm{a}^i\sim\bar{\pi}^i}\left[A^i_{\boldsymbol{\pi}}\left(\mathbf{s},\mathrm{a}^i\right)\right]\right|+\left|\mathbb{E}_{\mathbf{s}\sim\boldsymbol{\rho}_{\boldsymbol{\pi}},\mathrm{a}^i\sim\bar{\pi}^i}\left[A^i_{\boldsymbol{\pi}}\left(\mathbf{s},\mathrm{a}^i\right)\right]-\mathbb{E}_{\widetilde{\mathbf{s}}\sim\widetilde{\boldsymbol{\rho}}_{\boldsymbol{\pi}},\mathrm{a}^i\sim\bar{\pi}^i}\left[A^i_{\boldsymbol{\pi}}\left(\widetilde{\mathbf{s}},\mathrm{a}^i\right)\right]\right|\\
&\leq\frac{M^i}{1-\gamma}\xi e^{-\beta\kappa}+\frac{(2+\xi)\gamma\zeta}{1-\gamma\zeta}e^{-\beta\kappa}\\
&\leq\left(\frac{M^i\xi}{1-\gamma}+\frac{(2+\xi)\gamma\zeta}{1-\gamma\zeta}\right)e^{-\beta\kappa}.
\end{aligned}\tag{48}$$

Finally, denoting $(c^i,\phi)=\left(\frac{M^i\xi}{1-\gamma}+\frac{(2+\xi)\gamma\zeta}{1-\gamma\zeta},e^{-\beta}\right)$, we can obtain the Corollary 3.4.

□

B.4. The proof of Corollary 3.5

Proof. From (48), we can obtain $-c^i\phi^\kappa\leq L^{1:i}_{\boldsymbol{\pi}}\left(\bar{\boldsymbol{\pi}}^{1:i-1},\bar{\pi}^i\right)-L^i_{\pi^i_\kappa}\left(\bar{\pi}^i_\kappa\right)\leq c^i\phi^\kappa$. Similarly, we can further obtain the bounds about cost $-c_j\phi^\kappa\leq L^{1:i}_{j,\boldsymbol{\pi}}\left(\bar{\boldsymbol{\pi}}^{1:i-1},\bar{\pi}^i\right)-L^i_{j,\pi^i_\kappa}\left(\bar{\pi}^i_\kappa\right)\leq c_j\phi^\kappa$. By generalizing the result about the surrogate return, we can derive

$$\begin{aligned}
J^i_j(\bar{\boldsymbol{\pi}})&\leq J^i_j(\boldsymbol{\pi})+L^i_{j,\boldsymbol{\pi}}(\bar{\pi}^i)+\nu^i_j\sum_{h=1}^{n}D^{max}_{KL}(\pi^h,\widetilde{\pi}^h)\\
&\leq J^i_j(\boldsymbol{\pi})+L^i_{j,\pi^i_\kappa}\left(\bar{\pi}^i_\kappa\right)+c_j\phi^\kappa+\nu^i_{j,\kappa}\sum_{h=1}^{i}D^{\max}_{\mathrm{KL}}\left(\pi^h_\kappa,\bar{\pi}^h_\kappa\right).
\end{aligned}\tag{49}$$

where $L^i_{j,\pi^i_\kappa}(\bar{\pi}^i_\kappa)=\mathbb{E}_{\mathrm{s}_{\pi^i_\kappa}\sim\rho_{\pi^i_\kappa},\mathrm{a}^i\sim\bar{\pi}^i_\kappa}\left[A^i_{j,\pi^i_\kappa}\left(\mathrm{s}_{\mathcal{N}^i_\kappa},\mathrm{a}^i\right)\right]$, $(c_j,\phi)=\left(\frac{M_j\xi}{1-\gamma}+\frac{(2+\xi)\gamma\zeta}{1-\gamma\zeta},e^{-\beta}\right)$, $\nu^i_{j,\kappa}=\frac{2\gamma\max_{\mathrm{s}_{\mathcal{N}^i_\kappa},\mathrm{a}^i}\left|A^i_{j,\pi^i_\kappa}\left(\mathrm{s}_{\mathcal{N}^i_\kappa},\mathrm{a}^i\right)\right|}{(1-\gamma)^2}$, and M_j is a constant. □

B.5. The proof of Theorem 3.6

Proof. Theorem 3.6 is a scalable way of updating policies based on κ-hop dependencies under the condition that Assumption 3.1 and Assumption 3.2 hold. Bringing (48) into (11), we have

$$\begin{aligned}
&J\left(\bar{\boldsymbol{\pi}}\right)-J\left(\boldsymbol{\pi}\right)\\
&\geq\sum_{i=1}^{n}\left(L^{1:i}_{\boldsymbol{\pi}}\left(\bar{\boldsymbol{\pi}}^{1:i-1},\bar{\pi}^i\right)-\nu D^{\max}_{\mathrm{KL}}\left(\pi^i,\bar{\pi}^i\right)\right)\\
&\geq\sum_{i=1}^{n}\left(L^i_{\pi^i_\kappa}\left(\hat{\pi}^i_\kappa\right)-c^i\phi^\kappa-\nu D^{\max}_{\mathrm{KL}}\left(\pi^i|\hat{\pi}^i\right)\right)\\
&\geq\sum_{i=1}^{n}\left(L^i_{\pi^i_\kappa}\left(\hat{\pi}^i_\kappa\right)-c^i\phi^\kappa-\nu^i_\kappa D^{\max}_{\mathrm{KL}}\left(\pi^i_\kappa|\hat{\pi}^i_\kappa\right)\right)
\end{aligned}\tag{50}$$

where $(c^i,\phi)=\left(\frac{M^i\xi}{1-\gamma}+\frac{(2+\xi)\gamma\zeta}{1-\gamma\zeta},e^{-\beta}\right)$. □