Multi-Agent Reinforcement Learning for Constrained Markov Decision Processes by Consensus-Based Primal-Dual Method

Gaochen Cui, Qing-Shan Jia, *Member, IEEE*, Xiaohong Guan, *Fellow, IEEE*

***Abstract*—In this work, we consider multi-agent reinforcement learning for constrained Markov decision processes and develop a consensus-based primal-dual method to solve the problem, which is model-free and with provable convergence. Compared with existing methods, our algorithm does not require the dynamic model of the system, nor ask the agents to share their local policies. The constraint is incorporated in the objective function to form the Lagrangian with the dual variables updated through the primal-dual method. The consensus-based method is applied to update the parameters of the approximate action-value functions and the dual variables in a distributed manner. The developed algorithm is shown to achieve consensus among the agents and converge to a locally optimal policy. For a certain type of constrained Markov decision processes, the method to ensure the feasibility of the final solution is developed. Numerical results show that the developed algorithm outperforms the multi-agent actor-critic algorithm [1] which incorporates the constraint in the objective directly.**

***Index Terms*—Reinforcement learning, Multi-agent, Primal-dual, Constrained Markov decision process.**

I. Introduction

Markov decision processes (MDPs) are widely applied to model various practical systems [2]. Reinforcement learning (RL) is a class of methods for solving MDPs by interacting with the simulator or the real environment to optimize the expected cumulative reward or cost [3]. With deep learning showing strong feature extraction ability in computer vision and natural language processing [4], reinforcement learning is empowered to deal with complicated tasks from games [5] to robotics [6].

Multi-agent Markov decision processes (MAMDPs) are used to model large systems such as multi-robot system [7], which could be solved using the multi-agent reinforcement learning (MARL) method [8]. However, in the real world, we usually aim at finding a policy to minimize the cost subject to some constraints. For instance, in multi-robot systems, a problem of interest is to minimize the time spent by each robot on pursuing its target subject to constraint on collision times [9]. In the power system, the generation cost is minimized while the chance constraint of voltage should be satisfied [10]. These problems could be modeled as constrained multi-agent Markov decision processes (CMAMDPs).

In practical systems, when the closed-form CMAMDP models are hard to acquire, the model-based methods [9], [11] are hard to apply. The model-free RL is a candidate to solve MDPs [3] and constrained MDPs (CMDPs) [12], [13]. For solving cooperative CMAMDPs, the safe decentralized policy gradient (Safe Dec-PG) method is developed in [14]. However, the Safe Dec-PG method requires each agent to maintain and share a copy of the global policy, which may not be preferred by the users who demand privacy preservation.

To solve CMAMDP problems without the closed-form model, the following contributions are made in this article. First, an algorithm is developed where agents do not share their local costs or policies. The consensus-based dual update is developed to address the difficulty that the agents need to satisfy the system-level constraint while only local costs are observed. Second, theoretical results are presented to show the convergence of the developed algorithm. To ensure the feasibility of the final solution, we provide a method for a certain type of CMDP problems. Third, numerical experiments are carried out to validate the theory. The results show that the developed algorithm is more effective at handling the system-level constraint than the existing methods [1], [8].

The authors are with the CFINS, Department of Automations, BNRist, Tsinghua University, Beijing, China. Xiaohong Guan is also with the MOEKLINNS Laboratory, Xi'an Jiaotong University, Xi'an, China. Emails: cgc19@mails.tsinghua.edu.cn, jiaqs@tsinghua.edu.cn, xhguan@xjtu.edu.cn.
This work is supported by the National Natural Science Foundation of China (No. 62125304 and 62073182) and the 111 International Collaboration Project (No. BP2018006).

II. Literature Review

RL is a class of methods that improve the policy through interacting with the environment repeatedly. In practice, the closed-form model is usually hard to acquire, and thus model-free methods such as Q-learning [3] are developed. The parametrized policy can handle large action space and is trained with the actor-critic framework [15]. To deal with the large action space, the deep deterministic policy gradient (DDPG) method [16] is developed and achieves high performance in robotics.

For solving MAMDPs, single-agent RL is extended to MARL, such as multi-agent deterministic policy gradient which is extended from DDPG by updating distributed actor and critic [8]. For cooperative tasks, QMIX [17] is developed with the global critic function estimated by summing all the local critic estimates. Further, a fully decentralized MARL algorithm with convergence guarantees is developed in [18] where all agents do not directly share their local costs. For agents only observe the local states, scalable actor-critic frameworks are proposed to optimize local policies under some specific assumptions [19], [20].

For solving constrained dynamic control problems, control barrier functions are developed to define an invariant set of the system [21]. In [9], distributed control barrier functions are designed to keep the multi-robot system free from collisions. However, these methods require the dynamic model in the closed form. CMDPs are another approach to modeling the constrained dynamic control problems, which have an important impact in many areas of applications [22]. A direct way [1] to solve the CMDPs is to incorporate the constraint in the objective function with a fixed factor. However, the value of this factor is hard to choose. the The primal-dual-based RL algorithms are developed to solve the CMDPs [12], [13] by automatically tuning this factor with the duality gap bounded in [23].

For multi-agent systems, distributed constrained optimization is a hot topic in the last decade [24]. In the area of cooperative CMAMDPs, the Safe Dec-PG method is developed in [14], which is claimed to be the first decentralized policy gradient-based MARL algorithm that accounts for the coupled safety constraints with a quantifiable convergence rate. However, in the framework of Safe Dec-PG, each agent maintains a local copy of the global policy parameters and shares it with the neighbors, which harms the users' privacy. In this article, we develop a model-free MARL algorithm (Algorithm 2 in Section IV) to solve CMAMDPs free from policy sharing.

III. Problem Formulation

A CMAMDP is characterized by a tuple $\langle \mathcal{N}, \mathcal{S}, \mathcal{A}, P, c, g \rangle$, where $\mathcal{N}$ is the agent set with $|\mathcal{N}| = N$ as its cardinality, i.e. the number of agents, $\mathcal{S}$ is the finite state space, $\mathcal{A} = \prod_{i=1}^{N} \mathcal{A}^i$ is the finite joint action space with $\mathcal{A}^i$ as agent i's action space, $P(s'|s,a) : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0,1]$ is the state transition probability from $s \in \mathcal{S}$ to $s' \in \mathcal{S}$ determined by action $a = \prod_{i=1}^{N} a^i \in \mathcal{A}$, and $c^i(s,a^i) : \mathcal{S} \times \mathcal{A}^i \to \mathbb{R}$ and $g^i(s,a^i) : \mathcal{S} \times \mathcal{A}^i \to \mathbb{R}$ are two local cost functions received by agent i. Each agent executes a local policy $\pi^i(s,a^i) : \mathcal{S} \times \mathcal{A} \to [0,1]$ with $\sum_{a^i \in \mathcal{A}^i} \pi^i(s,a^i) = 1$, where $\pi^i(s,a^i)$ is the probability of choosing action a^i at state s. At each time t, the system state is denoted as s_t and each agent instantly gains costs $c_t^i = c^i(s_t, a_t^i)$ and $g_t^i = g^i(s_t, a_t^i)$ after taking action a_t^i at s_t.

Since the cardinality of the state space, $|\mathcal{S}|$, usually exponentially increases as the agent number increases, we apply parametrized functions to generate the distribution of randomized policy $\pi_{\theta^i}^i$ with parameter $\theta^i \in \Theta^i$ for each agent i, where $\Theta^i \subset \mathbb{R}^{m_i}$ is a convex and compact set with m_i denoting its dimensionality. Let $\pi_\theta = \prod_{i=1}^{N} \pi_{\theta^i}^i$ denote the joint policy, where $\theta = ((\theta^1)^\top, (\theta^2)^\top, ..., (\theta^N)^\top) \in \Theta$ and $\Theta = \prod_{i=1}^{N} \Theta^i$. Let P^θ denote the transition probability matrix of the Markov chain $\{s_t\}_{t \geq 0}$ induced by the joint policy π_θ, i.e.,

$$P^\theta(s'|s) = \sum_{a \in \mathcal{A}} \pi_\theta(s,a) P(s'|s,a), \forall s, s' \in \mathcal{S}. \tag{1}$$

Assumption 1. *We make the following assumptions for the Markov chain.*
(1.1) $\forall i \in \mathcal{N}, s \in \mathcal{S}, a^i \in \mathcal{A}^i$, *and* $\forall \theta^i \in \Theta^i$, $\pi_{\theta^i}^i(s,a^i) > 0$.
(1.2) $\pi_{\theta^i}^i(s,a^i) > 0$ *is continuous differentiable with respect to (w.r.t.)* θ^i *over* Θ^i.
(1.3) The Markov chain $\{s_t\}_{t \geq 0}$ *is irreducible and aperiodic induced by any policy* π_θ *and the randomized actions chosen by all agents are statistically independent.*

Assumption 1 is standard in the literature. The assumption that $\pi_{\theta^i}^i$ is differentiable w.r.t. θ^i is required by most policy gradient methods. Moreover, an irreducible and aperiodic Markov process has a stationary distribution $d_\theta(s)$ which is induced by joint policy π_θ. We then formulate the CMAMDP problem as

$$\begin{aligned} \min_{\theta \in \Theta} J(\theta) &= \lim_{T \to \infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=0}^{T-1} \frac{1}{N} \sum_{i=1}^{N} c_{t+1}^i \right] \\ &= \sum_{s \in \mathcal{S}} d_\theta(s) \sum_{a \in \mathcal{A}} \pi_\theta(s,a) \bar{c}(s,a) \end{aligned} \tag{2}$$

$$\begin{aligned} \text{s.t. } G(\theta) &= \lim_{T \to \infty} \frac{1}{T} \mathbb{E}\left[\sum_{t=0}^{T-1} \frac{1}{N} \sum_{i=1}^{N} g_{t+1}^i \right] \\ &= \sum_{s \in \mathcal{S}} d_\theta(s) \sum_{a \in \mathcal{A}} \pi_\theta(s,a) \bar{g}(s,a) \leq \bar{G}, \end{aligned} \tag{3}$$

where

$$\bar{c}(s,a) = \frac{1}{N} \sum_{i=1}^{N} c^i(s,a^i), \tag{4a}$$

$$\bar{g}(s,a) = \frac{1}{N} \sum_{i=1}^{N} g^i(s,a^i) \tag{4b}$$

are the global average cost functions. Accordingly, we define the global action-value functions under policy π_θ as

$$Q_\theta^c(s,a) = \mathbb{E}\left[\sum_t^\infty (\bar{c}_{t+1} - J(\theta)) \middle| s_0 = s, a_0 = a, \pi_\theta \right], \tag{5a}$$

$$Q_\theta^g(s,a) = \mathbb{E}\left[\sum_t^\infty (\bar{g}_{t+1} - G(\theta)) \middle| s_0 = s, a_0 = a, \pi_\theta \right], \tag{5b}$$

where $\bar{c}_t = \bar{c}(s_t, a_t)$ and $\bar{g}_t = \bar{g}(s_t, a_t)$. The global state-value functions are defined as

$$V_\theta^c(s) = \sum_{a \in \mathcal{A}} \pi_\theta(s,a) Q_\theta^c(s,a), \tag{6a}$$

$$V_\theta^g(s) = \sum_{a \in \mathcal{A}} \pi_\theta(s,a) Q_\theta^g(s,a). \tag{6b}$$

Then, we have the Poisson equations [3]

$$J(\theta) + Q_\theta^c(s,a) = \bar{c}(s,a) + \sum_{s' \in \mathcal{S}} P(s'|s,a) V_\theta^c(s'), \tag{7a}$$

$$G(\theta) + Q_\theta^g(s,a) = \bar{g}(s,a) + \sum_{s' \in \mathcal{S}} P(s'|s,a) V_\theta^g(s'). \tag{7b}$$

Assumption 2. *There exists at least one feasible solution for problem (2)-(3).*

With Assumption 2 holds throughout this article, the dual problem is defined as

$$\begin{aligned} \max_{\lambda} \min_{\theta} & L(\theta, \lambda) = J(\theta) + \lambda(G(\theta) - \bar{G}) \\ \text{s.t. } & \lambda \geq 0, \end{aligned} \tag{8}$$

where $L(\theta, \lambda)$ is the Lagrangian function and λ is the dual variable associated with the constraint (3). In the following section, we develop a model-free MARL algorithm for solving the dual problem (8). In this algorithm, all agents do not share their local costs c^i and g^i, nor the estimates for their expectation, i.e. $\mathbb{E}_{s \sim d_\theta, a \sim \pi_\theta}\left[c^i(s,a)\right]$ and $\mathbb{E}_{s \sim d_\theta, a \sim \pi_\theta}\left[g^i(s,a)\right]$, which is the same as [14]. In addition, unlike the existing MARL algorithm [14], our methods do not require the agents to share their local policies, which would be preferred by the users who expect privacy preservation on their behavior.

IV. Main Results

In this section, we first extend the work in [13] and [18] to develop Algorithm 1 which needs a centralized dual updater. We then develop Algorithm 2 that is fully decentralized, which is our main contribution. The two algorithms are both based on the primal-dual method, which is to incorporate the inequality constraint in the objective to form the Lagrangian function and both need a fully connected communication network, which is the same as [18], to convey local parameters.

A. Preliminaries

We first characterize the gradient of $L^\lambda(\theta)$, i.e. the Lagrangian function $L(\theta, \lambda)$ with λ fixed.

Lemma 1. *(Extended from Theorem 3.1 in [18]) Under Assumption 1, the gradient of* $L^\lambda(\theta)$ *w.r.t.* θ^i *is given by*

$$\nabla_{\theta^i} L^\lambda(\theta) = \mathbb{E}_{s \sim d_\theta, a \sim \pi_\theta}\left[A_\theta^\lambda(s,a) \nabla_{\theta^i} \ln \pi_{\theta^i}^i(s,a^i) \right], \tag{9}$$

where $A_\theta^\lambda(s,a) = Q_\theta^c(s,a) + \lambda Q_\theta^g(s,a)$.

Proof. The proof follows the policy gradient theorem in the single-agent RL [3], which implies that

$$\begin{aligned} \nabla_\theta J(\theta) &= \mathbb{E}_{s \sim d_\theta, a \sim \pi_\theta}\left[Q_\theta^c(s,a) \nabla_\theta \ln \pi_\theta(s,a) \right], \\ \nabla_\theta G(\theta) &= \mathbb{E}_{s \sim d_\theta, a \sim \pi_\theta}\left[Q_\theta^g(s,a) \nabla_\theta \ln \pi_\theta(s,a) \right], \end{aligned} \tag{10}$$

then

$$\begin{aligned} &\nabla_\theta L^\lambda(\theta) \\ =&\mathbb{E}_{s \sim d_\theta, a \sim \pi_\theta}\left[Q_\theta^\lambda(s,a) \nabla_\theta \ln \pi_\theta(s,a) \right] \\ =&\sum_{s \in \mathcal{S}} d_\theta(s) \sum_{a \in \mathcal{A}} \pi_\theta(s,a) Q_\theta^\lambda(s,a) \nabla_\theta \sum_{i=1}^{N} \ln \pi_{\theta^i}^i(s,a^i). \end{aligned} \tag{11}$$

Hence, the policy gradient with respect to each θ^i becomes

$$\nabla_{\theta^i} L^{\lambda}(\theta) \sum_{s\in\mathcal{S}} d_\theta(s) \sum_{a\in\mathcal{A}} \pi_\theta(s,a) Q_\theta^{\lambda}(s,a) \nabla_{\theta^i} \ln \pi_{\theta^i}^i(s,a^i) \quad (12)$$

□

Similar to [18], we build a directed time-variant communication graph $\mathcal{G}_t = (\mathcal{N}, \mathcal{E}_t)$ to convey information, where each agent is a vertex and is connected with other agents through the edges in the edge set $\mathcal{E}_t \subseteq \{(i,j)|i,j \in \mathcal{N}, i \neq j\}$. Let W_t denote the communication weight matrix at time t, where $W_t(i,j) = 0$ if $(i,j) \notin \mathcal{E}_t$. Each agent broadcasts and receives information through the communication graph to help them estimate the policy gradient. Then, they update their parameters according to the weight matrix W_t. Let $\mathbb{1}$ denote an all-ones vector with appropriate dimension in the rest of this paper. We impose the following assumption.

Assumption 3. *The sequence of nonnegative random matrices $\{W_t \in \mathbb{R}^{N\times N}\}_{t\geq 0}$ satisfies*
(3.1) $W_t\mathbb{1} = \mathbb{1}$ and $\mathbb{1}^\top \mathbb{E}[W_t] = \mathbb{1}^\top$. There exists a constant ϵ such that for any entry $W_t(i,j) > 0$, $W_t(i,j) \geq \epsilon$.
(3.2) $W_t(i,j) = 0$ if $(i,j) \notin \mathcal{E}_t$.
(3.3) The spectral norm of $\mathbb{E}[W_t^\top(\boldsymbol{I} - \mathbb{1}\mathbb{1}^\top/N)W_t]$ is strictly smaller than 1, where $\boldsymbol{I}$ is an identity matrix of size N.
(3.4) Given the σ-algebra generated by the random variables before time t, W_t is conditionally independent of c_{t+1}^i and g_{t+1}^i for any $i \in \mathcal{N}$.

In general, W_t is random because of communication failures or encryption. Condition (3.1) is standard in consensus algorithms. Condition (3.2) implies that no information is delivered through the edge that is not in the graph. Condition (3.3) holds if and only if the communication graph is connected [25]. Condition (3.4) requires W_t is independent of the costs, which is also common in multi-agent systems.

B. Centralized dual updating

We first extend the work in [13] and [18] to develop Algorithm 1. Considering the large dimensionality of a multi-agent system, we apply parametrized action-value functions $Q_\omega^c(\cdot,\cdot)$ and $Q_\psi^g(\cdot,\cdot)$ to approximate $Q_\theta^c(\cdot,\cdot)$ and $Q_\theta^g(\cdot,\cdot)$ of the policy π_θ with parameters ω and ψ, respectively. In our distributed framework, each agent maintains its local parameters ω^i and ψ^i. Then, we define the local parametrized action-value functions as $Q_{\omega^i}^c(\cdot,\cdot)$ and $Q_{\psi^i}^g(\cdot,\cdot)$. We note that all the local functions $Q_{\omega^i}^c(\cdot,\cdot)$ and $Q_{\psi^i}^g(\cdot,\cdot)$ are used to estimate the expectations of the global average costs. Let $\{\alpha_t\}_{t\geq 0}$, $\{\beta_t\}_{t\geq 0}$, and $\{\gamma_t\}_{t\geq 0}$ be the step size sequences. The algorithm is shown in Algorithm 1.

There are four main steps in Algorithm 1. In the critic step, the local variables μ^i and ν^i are for estimating $\mathbb{E}_{s\sim d_\theta, a\sim\pi_\theta}\left[c^i(s,a)\right]$ and $\mathbb{E}_{s\sim d_\theta, a\sim\pi_\theta}\left[g^i(s,a)\right]$, which are updated in lines 9 and 10, respectively. The TD-errors δ_t^i and Δ_t^i are calculated in lines 11 and 12 to update the temporal local parameters $\tilde{\omega}^i$ and $\tilde{\psi}^i$ in linse 13 and 14. In the actor step, the local policy parameter is driven towards the direction to minimize $L^{\lambda_t^i}(\theta)$ in lines 16 and 17 where Γ_θ is a projection to the convex and compact set Θ. In the dual step, the centralized updater collects all local estimates ν^i and processes a dual ascent for λ in line 20 where Γ_λ is a projection onto $[0, \lambda_{max}]$, then broadcasts the updated λ_{t+1} to every agent in line 22. In the consensus step, each agent receives the parameters $\tilde{\omega}^j$ and $\tilde{\psi}^j$ from the agents in $\{j|(j,i) \in \mathcal{E}_t, \forall j \in \mathcal{N}\}$ and update its local parameters ω^i and ψ^i with the communication weight matrix W_t in lines 26 and 27.

Algorithm 1 RL algorithm based on centralized dual update

1: **repeat**
2: **for all** $i \in \mathcal{N}$ **do**
3: observe state s_{t+1}
4: take action $a_{t+1}^i \sim \pi_{\theta^i}^i(s_{t+1})$
5: **end for**
6: observe joint actions $a_{t+1} = (a_{t+1}^1, ..., a_{t+1}^N)$
7: **for all** $i \in \mathcal{N}$ **do**
8: **Critic Step**
9: $\mu_{t+1}^i \leftarrow (1-\alpha_t)\mu_t^i + \alpha_t c_{t+1}^i$
10: $\nu_{t+1}^i \leftarrow (1-\alpha_t)\nu_t^i + \alpha_t g_{t+1}^i$
11: $\delta_t^i \leftarrow c_{t+1}^i - \mu_t^i + Q_{\omega_t^i}^c(s_{t+1}, a_{t+1}) - Q_{\omega_t^i}^c(s_t, a_t)$
12: $\Delta_t^i \leftarrow g_{t+1}^i - \nu_t^i + Q_{\psi_t^i}^g(s_{t+1}, a_{t+1}) - Q_{\psi_t^i}^g(s_t, a_t)$
13: $\tilde{\omega}_t^i \leftarrow \omega_t^i + \alpha_t \delta_t^i \nabla_{\omega^i} Q_{\omega^i}^c(s_t, a_t)$
14: $\tilde{\psi}_t^i \leftarrow \psi_t^i + \alpha_t \Delta_t^i \nabla_{\psi^i} Q_{\psi^i}^g(s_t, a_t)$
15: **Actor Step**
16: $A_t^i \leftarrow Q_{\omega_t^i}^c(s_t, a_t) + \lambda_t^i Q_{\omega_t^i}^g(s_t, a_t)$
17: $\theta_{t+1}^i \leftarrow \Gamma_\theta\left[\theta_t^i - \beta_t A_t^i \nabla_{\theta^i} \ln \pi_{\theta^i}^i(s_t, a_t)\right]$
18: **end for**
19: **Dual Step**
20: $\lambda_{t+1} \leftarrow \Gamma_\lambda\left[\lambda_t + \gamma_t(N^{-1}\sum_{i=1}^N \nu_t^i - \bar{G})\right]$
21: **for all** $i \in \mathcal{N}$ **do**
22: $\lambda_{t+1}^i = \lambda_{t+1}$
23: **end for**
24: **Consensus Step**
25: **for all** $i \in \mathcal{N}$ **do**
26: $\omega_{t+1}^i \leftarrow \sum_{j=1}^N W_t(i,j)\tilde{\omega}_t^j$
27: $\psi_{t+1}^i \leftarrow \sum_{j=1}^N W_t(i,j)\tilde{\psi}_t^j$
28: **end for**
29: **until** Max loop number

In Algorithm 1, each agent shares its $\tilde{\omega}^i$,$\tilde{\psi}^i$, and ν^i. Although the local cost c^i and the local estimator μ^i are preserved, the agents still need to share their local estimator ν^i. Next, we develop the following fully decentralized algorithm for further privacy preservation.

C. Decentralized dual updating

In this subsection, we develop the fully decentralized algorithm, shown in Algorithm 2.

In Algorithm 2, the centralized updater is removed and the dual variable λ^i is updated locally in line 20. To achieve this, each agent maintains the local variable υ^i to estimate the expectation of the additional global average cost, i.e. $\mathbb{E}_{s\sim d_\theta, a\sim\pi_\theta}\left[\bar{g}(s,a)\right]$, which is updated in lines 11 and 24. Every agent also knows $\bar{G}$. In addition, the local dual variables λ^i should achieve consensus to guarantee that all the agents are optimizing the same objective function. This is achieved by the consensus step in line 27. In Algorithm 2, each agent shares its $\tilde{\omega}^i$,$\tilde{\psi}^i$, $\tilde{\upsilon}^i$, and $\tilde{\lambda}^i$ with its neighbors.

V. Theoretical Analysis

In this section, we show the consensus and convergence of the developed algorithms with linear functions as the action-value functions, i.e. $Q_{\omega^i}^c(s,a) = {\omega^i}^\top \phi(s,a)$ and $Q_{\psi^i}^g(s,a) = {\psi^i}^\top \phi(s,a)$, where $\phi(s,a) = [\phi_1(s,a), ..., \phi_K(s,a)]^\top \in \mathbb{R}^K$ is the feature vector of state-action pair (s,a). We also denote $\Phi \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|\times K}$ as the feature matrix of the entire state-action space $\mathcal{S}\times\mathcal{A}$. The following assumptions are required for both algorithms.

Assumption 4. *The feature vector $\phi(s,a)$ is bounded, $\forall s \in \mathcal{S}, \forall a \in \mathcal{A}$. Let Φ have full column rank. Moreover, $\Phi u \neq \mathbb{1}$ for any $u \in \mathbb{R}^K$, where $\mathbb{1}$ here is an all-ones vector with dimension $|\mathcal{S}||\mathcal{A}|$.*

Algorithm 2 RL algorithm based on decentralized dual update

repeat
  for all $i \in \mathcal{N}$ **do**
    observe state s_{t+1}
    take action $a^i_{t+1} \sim \pi^i_{\theta^i}(s_{t+1})$
  end for
  observe joint actions $a_{t+1} = (a^1_{t+1}, ..., a^N_{t+1})$
  for all $i \in \mathcal{N}$ **do**
    Critic Step
    $\mu^i_{t+1} \leftarrow (1-\alpha_t)\mu^i_t + \alpha_t c^i_{t+1}$
    $\nu^i_{t+1} \leftarrow (1-\alpha_t)\nu^i_t + \alpha_t g^i_{t+1}$
    $\tilde{v}^i_t \leftarrow (1-\alpha_t)v^i_t + \alpha_t g^i_{t+1}$
    $\delta^i_t \leftarrow c^i_{t+1} - \mu^i_t + Q^c_{\omega^i_t}(s_{t+1}, a_{t+1}) - Q^c_{\omega^i_t}(s_t, a_t)$
    $\Delta^i_t \leftarrow g^i_{t+1} - \nu^i_t + Q^g_{\psi^i_t}(s_{t+1}, a_{t+1}) - Q^g_{\psi^i_t}(s_t, a_t)$
    $\tilde{\omega}^i_t \leftarrow \omega^i_t + \alpha_t \delta^i_t \nabla_{\omega^i} Q^c_{\omega^i}(s_t, a_t)$
    $\tilde{\psi}^i_t \leftarrow \psi^i_t + \alpha_t \Delta^i_t \nabla_{\psi^i} Q^g_{\psi^i}(s_t, a_t)$
    Actor Step
    $A^i_t \leftarrow Q^c_{\omega^i_t}(s_t, a_t) + \lambda^i_t Q^g_{\omega^i_t}(s_t, a_t)$
    $\theta^i_{t+1} \leftarrow \Gamma_\theta \left[\theta^i_t - \beta_t A^i_t \nabla_{\theta^i} \ln \pi^i_{\theta^i}(s_t, a_t)\right]$
    Dual Step
    $\tilde{\lambda}^i_{t+1} \leftarrow \Gamma_\lambda \left[\lambda^i_t + \gamma_t(v^i_t - \bar{G})\right]$
  end for
  Consensus Step
  for all $i \in \mathcal{N}$ **do**
    $v^i_{t+1} \leftarrow \sum_{j=1}^N W_t(i,j)\tilde{v}^j_t$
    $\omega^i_{t+1} \leftarrow \sum_{j=1}^N W_t(i,j)\tilde{\omega}^j_t$
    $\psi^i_{t+1} \leftarrow \sum_{j=1}^N W_t(i,j)\tilde{\psi}^j_t$
    $\lambda^i_{t+1} \leftarrow \sum_{j=1}^N W_t(i,j)\tilde{\lambda}^j_t$
  end for
until Max loop number

Assumption 5. *The step-size sequences* $\{\alpha_t\}_{t\geq 0}$, $\{\beta_t\}_{t\geq 0}$, *and* $\{\gamma_t\}_{t\geq 0}$ *satisfy*

$$\sum_{t=0}^{\infty} \alpha_t = \sum_{t=0}^{\infty} \beta_t = \sum_{t=0}^{\infty} \gamma_t = \infty \tag{13}$$

$$\sum_{t=0}^{\infty} \alpha_t^2 + \beta_t^2 + \gamma_t^2 < \infty \tag{14}$$

$$\lim_{t\to\infty} \frac{\beta_t}{\alpha_t} = \lim_{t\to\infty} \frac{\gamma_t}{\beta_t} = 0 \tag{15}$$

$$\lim_{t\to\infty} \frac{\alpha_{t+1}}{\alpha_t} = \lim_{t\to\infty} \frac{\beta_{t+1}}{\beta_t} = \lim_{t\to\infty} \frac{\gamma_{t+1}}{\gamma_t} = 1 \tag{16}$$

A. Convergence of Algorithm 1

The convergence analysis of Algorithm 1 follows similar techniques in [13] and [18]. We now provide the following detailed analysis. For the critic step, we can regard the policy parameter θ and the dual variable λ as constants. Let $P_\theta(s', a'|s, a)$ denote $P(s'|s, a)\pi_\theta(s', a')$ and D_θ denote diag $[d_\theta(s)\pi_\theta(s, a), s \in \mathcal{S}, a \in \mathcal{A}]$ for simplicity, which are the state-action transition probability and stationary distribution matrix, respectively. We also define two operators $T^c_\theta : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$ and $T^g_\theta : \mathbb{R}^{|\mathcal{S}||\mathcal{A}|} \to \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, which are

$$\begin{aligned} T^c_\theta(Q) &= \bar{R} - J(\theta)\mathbb{1} + P_\theta Q \\ T^g_\theta(Q) &= \bar{G} - G(\theta)\mathbb{1} + P_\theta Q, \end{aligned} \tag{17}$$

where $\bar{R} = (\bar{c}(s, a), s \in \mathcal{S}, a \in \mathcal{A})^\top \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, $\bar{G} = (\bar{g}(s, a), s \in \mathcal{S}, a \in \mathcal{A})^\top \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$, and P_θ is the transition probability matrix. Based on the above definition, we have the following lemma,

Lemma 2. *(Theorem 4.6 in [18]) Under Assumptions 1, 3, 4, and 5, for any policy* π_θ, *with* $\{\mu^i_t\}_{t\geq 0}$, $\{\nu^i_t\}_{t\geq 0}$, $\{\omega^i_t\}_{t\geq 0}$, *and* $\{\psi^i_t\}_{t\geq 0}$ *generated by Algorithm 1, we have* $\forall i \in \mathcal{N}, \lim_{t\to\infty} \mu^i_t = \mathbb{E}_{s\sim d_\theta, a\sim\pi_\theta}\left[c^i(s, a)\right], \lim_{t\to\infty} \nu^i_t = \mathbb{E}_{s\sim d_\theta, a\sim\pi_\theta}\left[g^i(s, a)\right], \lim_{t\to\infty} \omega^i_t = \omega_\theta, \lim_{t\to\infty} \psi^i_t = \psi_\theta$, ω_θ *is the unique solution to*

$$\Phi^\top D_\theta \left[T^c_\theta(\Phi\omega_\theta) - \Phi\omega_\theta\right] = 0, \tag{18}$$

and ψ_θ *is the unique solution to*

$$\Phi^\top D_\theta \left[T^g_\theta(\Phi\psi_\theta) - \Phi\psi_\theta\right] = 0. \tag{19}$$

First, Lemma 2 implies that each agent maintains two estimates for $\mathbb{E}\left[c^i(s, a)\right]$ and $\mathbb{E}\left[g^i(s, a)\right]$ of the joint policy π_θ. Second, the local action-value function parameters ω^i and ψ^i converge to the consensus values, ω_θ and ψ_θ, which are the minimum of the mean square errors of equation (7). Thus, the policy parameters are updated to the direction with the gradient estimated using the global action-value function.

Similarly, for the actor step, we consider that λ is fixed. Then, for the actor step, with

$$\hat{\Gamma}(F(x)) = \lim_{\eta\to 0^+} \frac{\Gamma\left[x + \eta F(x)\right] - x}{\eta}, \tag{20}$$

where $x \in \mathbb{R}^X$ and $F : \mathbb{R}^X \to \mathbb{R}^X$, we have the following lemma, which is extended from [18].

Lemma 3. *(Extended from Theorem 4.7 in [18]) Under Assumptions 1, 4, and 5, suppose that there is a compact set* $\Theta^{i\,*}_\lambda$ *as the asymptotically stable equilibrium of ODE*

$$\dot{\theta}^i = \hat{\Gamma}_\theta \left[-\mathbb{E}_{s\sim d_\theta, a\sim\pi_\theta}\left[A^\lambda_\theta(s, a)\nabla_{\theta^i} \ln \pi_{\theta^i}(s, a)\right]\right], \tag{21}$$

where $A^\lambda_\theta(s, a) = Q^c_{\omega_\theta}(s, a) + \lambda Q^g_{\omega_\theta}(s, a)$ *for each agent* i. *Then the policy parameters* $\{\theta^i_t\}_{t\geq 0}$ *generated by Algorithm 1 converges almost surely to* $\Theta^{i\,*}_\lambda$, $\forall i \in \mathcal{N}$.

Proof. Recall the actor update formula

$$\theta^i_{t+1} = \Gamma_\theta \left[\theta^i_t - \beta_t A^i_t(s_t, a_t)\nabla_{\theta^i} \ln \pi^i_{\theta^i}(s_t, a_t)\right]. \tag{22}$$

We define $A_{t,\theta_t} = Q^c_{\omega_{\theta_t}}(s_t, a_t) + \lambda Q^g_{\psi_{\theta_t}}(s_t, a_t)$, where ω_{θ_t} and ψ_{θ_t} are the unique solutions to equations (18) and (19) with $\theta = \theta_t$. Let $\nabla^i_t = \nabla_{\theta^i} \ln \pi^i_{\theta^i}(s_t, a_t)$. We then have

$$\begin{aligned} \theta^i_{t+1} &= \Gamma_\theta \left[\theta^i_t - \beta_t A^i_t \nabla^i_t\right] \\ &= \Gamma_\theta \left[\theta^i_t - \beta_t (\mathbb{E}_{s_t\sim d_\theta, a_t\sim\pi_\theta}\left[A_{t,\theta_t}\nabla^i_t | \mathcal{F}^\theta{}_t\right] - \xi^i_{1,t} - \xi^i_{2,t})\right], \end{aligned} \tag{23}$$

where $\xi^i_{1,t} = \mathbb{E}_{s_t\sim d_\theta, a_t\sim\pi_\theta}\left[(A^i_t - A_{t,\theta_t})\nabla^i_t | \mathcal{F}^\theta_t\right]$, $\xi^i_{2,t} = A^i_t\nabla^i_t - \mathbb{E}_{s_t\sim d_\theta, a_t\sim\pi_\theta}\left[A^i_t\nabla^i_t | \mathcal{F}^\theta_t\right]$, and $\mathcal{F}^\theta_t = \sigma(\theta_\tau, \tau \leq t)$ is the σ − field generated by $\{\theta_\tau, \tau \leq t\}$.

We can show that $\{\xi_{1,t}\}_{t\geq 0}$ is a bounded sequence with $\xi_{1,t} \to 0$ almost surely as $t \to \infty$ by Lemma 2 at the faster time-scale and $d_{\theta_t}, \pi_t, A_{t,\theta_t}$ can be verified to be continuous in θ^i_t. Moreover, $\{\xi_{2,t}\}_{t\geq 0}$ is a martingale-difference sequence. Thus, Assumption (7.4) holds under Assumption 5. Therefore, the sequence $\{\theta^i_t\}_{t\geq 0}$ generated by (22) converges almost surely to $\Theta^{i\,*}$ by Lemma 6, which concludes the proof. □

We note that $F_\theta(\theta^i) = \mathbb{E}_{s\sim d_\theta, a\sim\pi_\theta}\left[A^\lambda_\theta(s, a)\nabla_{\theta^i} \ln \pi_{\theta^i}(s, a)\right]$, which is continuous w.r.t. θ^i [18], could be regarded as the gradient. Thus, the asymptotically stable equilibrium could be explained as the locally minimal points in Θ^i in this gradient field. This assumption that a compact set $\Theta^{i\,*}_\lambda$ as the asymptotically stable equilibrium exists

is common in the literature [12], [13], [18]. In the following, we denote θ_λ^i as the stationary point which corresponds to policy π_λ^i and $d_\lambda = d_{\theta_\lambda}$ as the stationary distribution for any λ. We then have the following results for the dual step, which is extended from [13]. Assumption 6 is needed to satisfy Assumption 7.

Assumption 6. *$\forall \lambda \in [0, \lambda_{max}]$, ${\Theta_\lambda^i}^*$ only contains one unique point ${\theta_\lambda^i}^*$, and $G({\theta_\lambda^i}^*)$ is continuous w.r.t. λ.*

Lemma 4. *(Extended from Theorem 3 in [13]) Under Assumptions 1, 4, 5, and 6, suppose that for any λ there is a compact set Λ^* as the asymptotically stable equilibrium of ODE*

$$\dot{\lambda} = \hat{\Gamma}_\lambda(\mathbb{E}_{s\sim d_\lambda, a\sim\pi_\lambda}[\bar{g}(s,a)] - \bar{G}), \tag{24}$$

Then the dual variable $\{\lambda_t\}_{t\geq 0}$ generated by Algorithm 1 converges almost surely to Λ^*.

Proof. Recall that dual update formula

$$\lambda_{t+1}^i = \Gamma_\lambda\left[\lambda_t^i + \gamma_t(\bar{\nu}_t - \bar{G})\right], \tag{25}$$

where $\bar{\nu}_t = N^{-1}\sum_{i=1}^N \nu_t^i$. We rewrite formula (25) as

$$\begin{aligned}\lambda_{t+1} &= \Gamma_\lambda\left[\lambda_t + \gamma_t(\bar{\nu}_t - \bar{G})\right] \\ &= \Gamma_\lambda\left[\lambda_t + \gamma_t(\mathbb{E}\left[\bar{c}(s_t,a_t)|\mathcal{F}_t^\lambda\right] - \bar{G} + \zeta_t)\right],\end{aligned} \tag{26}$$

where $\mathbb{E}$ is short for $\mathbb{E}_{s_t\sim d_\lambda, a_t\sim\pi_\lambda}$,

$$\zeta_t = \bar{\nu}_t - \mathbb{E}_{s_t\sim d_\lambda, a_t\sim\pi_\lambda}\left[\bar{c}(s_t,a_t)|\mathcal{F}_t^\lambda\right],$$

and $\mathcal{F}_t^\lambda = \sigma(\lambda_\tau, \tau \leq t)$ is the $\sigma-$field generated by $\{\lambda_\tau, \tau\leq t\}$. We can show that $\{\xi_{1,t}\}_{t\geq 0}$ is a bounded sequence with $\xi_{1,t}\to 0$ almost surely as $t\to\infty$ by Theorem 2 at the faster time-scale. Therefore, the sequence $\{\lambda_t\}_{t\geq 0}$ generated by formula (25) converges almost surely to Λ^* by Lemma 6, which concludes the proof. □

With the dual step shows the same asymptotic convergence with [13], the following propositions could be directly extended to conclude Algorithm 1.

Proposition 1. *(Proposition 1 in [13]) For any limiting point λ^* in $\hat{\Lambda}^* = \{\lambda\in[0,\lambda_{max})|\hat{\Gamma}_\lambda(\mathbb{E}_{s\sim d_\lambda,a\sim\pi_\lambda}[\bar{g}(s,a)] - \bar{G}) = 0\}$, the corresponding limiting point θ_{λ^*} satisfies the constraint $G(\theta_{\lambda^*}) \leq \bar{G}$.*

Proposition 2. *(Proposition 2 in [13]) For some limiting point $\lambda^* \in \Lambda^*$, if $G(\theta_{\lambda^*}) < \bar{G}$, we have $\lambda^* = 0$.*

Lemma 4 shows that λ_t converges to a local maximum of the "dual problem" where the primal problem reaches a "local minimum". Proposition 1 and 2 show that the theoretical result is similar to that of the convex optimization theory [26].

B. Convergence of Algorithm 2

The critic step and actor step of Algorithm 2 are mostly the same as Algorithm 1, so the theoretical results are omitted. The difference is that the local dual variables in Algorithm 2 are updated in a distributed manner, where each agent maintains a local estimate υ^i to track the expectation of the global average cost, $\mathbb{E}[\bar{g}(s,a)]$. For υ^i, we have the following lemma.

Lemma 5. *(Theorem 4.10 in [18]) Under Assumptions 1, 4, and 5, for any policy π_θ, with $\{\upsilon_t^i\}$ generated by Algorithm 2, we have $\forall i\in\mathcal{N}, \lim_{t\to\infty}\upsilon_t^i = \mathbb{E}_{s\sim d_\theta, a\sim\pi_\theta}[\bar{g}(s,a)]$.*

Remark 1. *Algorithm 2 is in the class of noise-based privacy-preserving methods [27], because in each local update, the agents could add zero-mean noises to their local costs c_{t+1}^i and g_{t+1}^i. This operation does not affect the convergence result [18]. In the critic step, considering fixed policy π_θ, the agents do not directly share ν_t^i. Under the condition that each agent only knows its own $W_t(i,)$, it is hard to infer the other agents' ν_t^j since the shared $\tilde{\upsilon}_t^j$ is the weighted sum of the agent j's neighbors [27]. If W_t is time-invariant and is known to agent i, additional random noise could be added before sharing local information. In this case, the local $\mathbb{E}[g_t^j]$ can not be precisely inferred by agent i if and only if the neighbors of agent i is not a subset of agent j's [28]. The updates of ω^i and ψ^i are similar.*

Next, we show that the local dual variables converge to consensus a.s. in the following theorem.

Theorem 1. *Under Assumption 5, $\forall i\in\mathcal{N}$, with $\{\lambda_t^i\}_{t\geq 0}$ generated by Algorithm 2, $\lim_{t\to\infty}\lambda_t^i = \bar{\lambda}_t$ a.s. with $\bar{\lambda}_t = \frac{1}{N}\mathbb{1}^\top\lambda_t$.*

Proof. The proof is extended from Lemma 5.3 in [18]. The difference is that there is a projection in the local dual update. We first define

$$\lambda_t^\perp = \lambda_t - \bar{\lambda}_t\mathbb{1}, \tag{27}$$

where $\lambda_t = [\lambda_t^1, ..., \lambda_t^N]^\top$. Then, with

$$h_t = [\upsilon_t^1 - \bar{G}, ..., \upsilon_t^N - \bar{G}]^\top \tag{28}$$

we have

$$\begin{aligned}\lambda_{t+1}^\perp &= \lambda_{t+1} - \bar{\lambda}_{t+1}\mathbb{1} \\ &= \Gamma_\lambda[W_t(\lambda_t + \gamma_t h_t)] - \frac{1}{N}\mathbb{1}\mathbb{1}^\top\Gamma_\lambda[W_t(\lambda_t+\gamma_t h_t)] \\ &= (\mathbf{I} - \frac{1}{N}\mathbb{1}\mathbb{1}^\top)\Gamma_\lambda[W_t(\lambda_t+\gamma_t h_t)],\end{aligned} \tag{29}$$

where $\mathbf{I}$ is the N-dimensional identity matrix, and

$$\begin{aligned}&\mathbb{E}[\|\gamma_{t+1}^{-1}\lambda_{t+1}^\perp\|^2|\mathcal{F}_t^\lambda] \\ &\overset{(a)}{=}\frac{\gamma_t^2}{\gamma_{t+1}^2}\cdot\frac{1}{\gamma_t^2}\mathbb{E}[\Gamma_\lambda[W_t(\lambda_t+\gamma_t h_t)]^\top(\mathbf{I}-\frac{1}{N}\mathbb{1}\mathbb{1}^\top) \\ &\quad\Gamma_\lambda[W_t(\lambda_t+\gamma_t g_t)]|\mathcal{F}_t^\lambda] \\ &\overset{(b)}{\leq}\frac{\gamma_t^2}{\gamma_{t+1}^2}\cdot\frac{1}{\gamma_t^2}\mathbb{E}[(W_t(\lambda_t+\gamma_t g_t))^\top(\mathbf{I}-\frac{1}{N}\mathbb{1}\mathbb{1}^\top) \\ &\quad W_t(\lambda_t+\gamma_t h_t)|\mathcal{F}_t^\lambda] \\ &\overset{(c)}{=}\frac{\gamma_t^2}{\gamma_{t+1}^2}\cdot\frac{1}{\gamma_t^2}\mathbb{E}[\|(\mathbf{I}-\frac{1}{N}\mathbb{1}\mathbb{1}^\top)W_t(\lambda_t^\perp+\gamma_t h_t)\|^2|\mathcal{F}_t^\lambda] \\ &\overset{(d)}{\leq}\frac{\gamma_t^2}{\gamma_{t+1}^2}\rho(\mathbb{E}[\|\gamma_t^{-1}\lambda_t^\perp\|^2|\mathcal{F}_t^\lambda] + 2K_1\mathbb{E}[\|\gamma_t^{-1}\lambda_t^\perp\||\mathcal{F}_t^\lambda] + K_1^2),\end{aligned} \tag{30}$$

where (a) is by plugging (29) into $\lambda_t^\perp$ and $(\mathbf{I}-N^{-1}\mathbb{1}\mathbb{1}^\top)^\top(\mathbf{I}-N^{-1}\mathbb{1}\mathbb{1}^\top) = (\mathbf{I}-N^{-1}\mathbb{1}\mathbb{1}^\top)$; (b) is because that $(\mathbf{I}-N^{-1}\mathbb{1}\mathbb{1}^\top)$ is positive definite matrix; (c) is because that $(\mathbf{I}-N^{-1}\mathbb{1}\mathbb{1}^\top)\bar{\lambda}_t\mathbb{1} = 0$; (d) is by Assumption (4.3) where $\rho < 1$ is the upper bound of the spectral norm of $(\mathbf{I}-\frac{1}{N}\mathbb{1}\mathbb{1}^\top)W_t$ and K_1 is the upper bound of $\|\gamma_t h_t\|$ since $\|h_t\|$ is upper bounded by the assumption of that the cost $g^i(s,a)$ is bounded. Since $\lim_{t\to\infty}\gamma_t^2\gamma_{t+1}^{-2} = 1$ and $\rho<1$, there exist some $\varepsilon\in(0,1)$ and large enough t_0 such that $\gamma_t^2\gamma_{t+1}^{-2}\rho \leq (1-\varepsilon)$, for $t\geq t_0$. Thus, we have

$$\begin{aligned}&\mathbb{E}[\|\gamma_{t+1}^{-1}\lambda_{t+1}^\perp\|^2|\mathcal{F}_t^\lambda] \\ \leq&(1-\varepsilon)(\mathbb{E}[\|\gamma_t^{-1}\lambda_t^\perp\|^2|\mathcal{F}_t^\lambda] + 2K_1\mathbb{E}[\|\gamma_t^{-1}\lambda_t^\perp\||\mathcal{F}_t^\lambda] + K_1^2) \\ \leq&(1-\frac{\varepsilon}{2})\mathbb{E}[\|\gamma_t^{-1}\lambda_t^\perp\|^2|\mathcal{F}_t^\lambda] + K_2.\end{aligned} \tag{31}$$

Then, by taking expectations on both sides and induction, we have

$$\mathbb{E}\|\gamma_t^{-1}\lambda_t^\perp\|^2 \leq (1-\frac{\varepsilon}{2})^{(t-t_0)}\mathbb{E}\|\gamma_{t_0}^{-1}\lambda_{t_0}^\perp\|^2 + \frac{2K_2}{\varepsilon} < \infty. \tag{32}$$

Therefore, $\sup_t\mathbb{E}\|\gamma_t^{-1}\lambda_t^\perp\|^2 < \infty$ and under Assumption 5,

$\sum_{t=0}^{\infty}\|\lambda_t^{\perp}\|^2$ is finite which yields $\lim_{t\to\infty}\lambda_t^{\perp}=0$, and thus concludes the proof. □

Theorem 1 shows that all the local dual variables will converge to consensus. With Lemma 5 showing that all the local υ_t^i converge to consensus, the sequence $\{\bar{\lambda}_t\}$ generated by Algorithm 2 converges to the same set as Algorithm 1, which is shown in the following theorem.

Theorem 2. *Under Assumptions 1, 4, 5, and 6, suppose that for any λ there is a compact set Λ^* as the asymptotically stable equilibrium of ODE*

$$\dot{\lambda}=\hat{\Gamma}_{\lambda}(\mathbb{E}_{s\sim d_{\lambda},a\sim\pi_{\lambda}}[\bar{g}(s,a)]-\bar{G}), \tag{33}$$

Then the dual variable $\{\bar{\lambda}_t\}_{t\geq 0}$ generated by Algorithm 2 converges almost surely to Λ^*.

Proof. The proof is based on Theorem 2.1, in Chapter 5, [29]. With

$$h_t=[\upsilon_t^1-\bar{G},...,\upsilon_t^N-\bar{G}]^{\top}, \tag{34}$$

recall the dual update

$$\begin{aligned}\lambda_{t+1}=&W_t\Gamma_{\lambda}[\lambda_t+\gamma_t h_t]\\ =&W_t(\lambda_t+\gamma_t h_t+\gamma_t Z_t),\end{aligned} \tag{35}$$

where $Z_t=(Z_t^1,...,Z_t^N)^{\top}$ and $\gamma_t Z_t^i$ is a vector of the shortest Euclidean length needed to take $\lambda_t+\gamma_t h_t$ back to the constraint set $[0,\lambda_{max}]^N$. We then have

$$\begin{aligned}\bar{\lambda}_{t+1}=&\frac{1}{N}\mathbb{1}^{\top}W_t\Gamma_{\lambda}[\lambda_t+\gamma_t h_t]\\ =&\frac{1}{N}\mathbb{1}^{\top}W_t(\lambda_t+\gamma_t h_t+\gamma_t Z_t)\\ =&\bar{\lambda}_t+\gamma_t\mathbb{E}[\bar{g}(s,a)-\bar{G}]+\gamma_t\bar{Z}_t+\gamma_t\zeta_t^1+\gamma_t\zeta_t^2,\end{aligned} \tag{36}$$

where $\gamma_t\bar{Z}_t$ is a scalar of the minimal absolute value needed to take $\bar{\lambda}_t+\gamma_t\bar{h}_t$ back to the constraint set $[0,\lambda_{max}]$

$$\zeta_t^1=\frac{1}{N}\mathbb{1}^{\top}(h_t-\mathbb{E}[\bar{g}(s,a)-\bar{G}]\mathbb{1}) \tag{37a}$$

$$\zeta_t^2=\frac{1}{N}\mathbb{1}^{\top}(Z_t-\bar{Z}_t). \tag{37b}$$

By Lemma 5, we have $\lim_{t\to\infty}\zeta_t^1=0$ a.s. Next, we show that $\lim_{t\to\infty}\zeta_t^1=0$. From Theorem 1, we have that $\sup_t\mathbb{E}\|\gamma_t^{-1}\lambda_t^{\perp}\|^2<\infty$, which means that $\lambda_t^{\perp}$ decreases at the same speed as γ_t. Also, $\lim_{t\to\infty}\upsilon_t^i=\mathbb{E}_{s\sim d_\theta,a\sim\pi_\theta}[\bar{g}(s,a)]$. We consider two circumstances. One is that if $h_t>>\gamma_t$ as $t\to\infty$, we have that $|\lambda_t^{\perp}|_{\infty}<<h_t$. So, $\zeta_t^2\to 0$ as $t\to\infty$. The other is that if $h_t\sim\gamma_t$ or $h_t=o(\gamma_t)$, we have $\bar{Z}_t\to 0$ directly since the limits of $\bar{Z}_t$ and Z_t are both 0. Then the proof follows that in [29]. □

Theorem 2 implies that the local dual variables of Algorithm 2 converge to the same asymptotically stable set as the centralized dual variable of Algorithm 1.

C. Discussions on the feasible solutions

In this subsection, we develop a method to ensure the feasibility of the final policy for a certain type of CMDPs. For an MDP that satisfies Assumption 1, each agent i applies a tabular policy that stores a vector $\boldsymbol{\theta}_s^i=[\theta_s^i(a_1),...,\theta_s^i(a_{|\mathcal{A}^i|})]$ for each state $s\in\mathcal{S}$, where $\theta_s^i(a_i)$ is the probability to choose action a_i at state s. Thus, the projection operation in line 17 (18) of Algorithm 1 (2) is onto the convex and compact set Θ^i, i.e.$\sum_{a\in\mathcal{A}^i}\theta_s^i(a)=1$ and $\theta_s^i(a)\geq\epsilon$, $\forall s\in\mathcal{S},i\in\mathcal{N}$, where ϵ is some small constant to satisfy Assumption (1.2). Under Assumption 1, we define the ergodic occupation measure, $f\in\mathcal{G}$, that $f(s,a)=d(s)\pi(s,a)$ and $\mathcal{G}$ is a simplex on a hyperplane [30]. Problem (2)-(3) is equivalent to

$$\begin{aligned}\min_{f\in\mathcal{G}}&\sum_{s,a}f(s,a)\bar{c}(s,a)\\ \text{s.t.}&\sum_{s,a}f(s,a)\bar{g}(s,a)\leq\bar{G},\end{aligned} \tag{38}$$

which is a linear programming problem. Let the feature dimension $K=|\mathcal{S}||\mathcal{A}|-1$ and Assumption 4 hold, then it could be verified that the solutions to Eqs. (18) and (19) are also the solutions to Poisson equations (7). This implies that for any fixed policy π_θ, the approximation error of the policy gradient a.s. converges to 0. Thus, for any fixed dual variable λ ($\bar{\lambda}$), Algorithm 1 and 2 could both find the globally optimal solution $f^*(\lambda)$ to the relaxed problem [31]

$$\min_{f\in\mathcal{G}}\sum_{s,a}f(s,a)(\bar{c}(s,a)+\lambda\bar{g}(s,a)). \tag{39}$$

However, $f^*(\lambda)$ is not continuous w.r.t. λ everywhere because the property of linear programming. If the optimal solution to the original problem (38) is not any extreme point of $\mathcal{G}$, then $f^*(\lambda)$ jumps from one extreme point to the other around the dual solution, λ^*. In this case, Assumption 6 does not hold. To address this, we propose to build a surrogate convex and compact set $\tilde{\Theta}^i\subset\Theta^i,\forall i\in\mathcal{N}$, of which all the points on its relative boundary are extreme points so that the induced ergodic occupation measure space $\tilde{\mathcal{G}}$ is with the same property [30]. In this way, Assumption 6 is satisfied. And the optimal point of the relaxed problem

$$\min_{f\in\tilde{\mathcal{G}}}\sum_{s,a}f(s,a)(\bar{c}(s,a)+\lambda^*\bar{g}(s,a))$$

is also the unique solution to problem

$$\begin{aligned}\min_{f\in\tilde{\mathcal{G}}}&\sum_{s,a}f(s,a)\bar{c}(s,a)\\ \text{s.t.}&\sum_{s,a}f(s,a)\bar{g}(s,a)\leq\bar{G},\end{aligned}$$

since the strong duality holds for linear programming problems.

Remark 2. *The developed algorithms and theoretical results could be extended to multiple constraints by replacing λ with $\boldsymbol{\lambda}\in\mathbb{R}^m$, where m is the constraint number. However, $m\geq 2$ constraints may cause the optimized policy infeasible. For example, in the above case, additional constraints may cause the solution to problem (38) is not on the relative boundary of $\mathcal{G}$. And it would be difficult to construct $\tilde{\mathcal{G}}$ to guarantee that $f^*(\boldsymbol{\lambda})$ is on its relative boundary.*

In general, when $K<|\mathcal{S}||\mathcal{A}|-1$ or using nonlinear functions such as neural networks to represent the policy, it would be much more complicated and the theoretical analysis is out of scope of this paper. In the following section, numerical experiments are carried out to evaluate the effectiveness.

VI. Numerical Experiments

A. Case 1

In this case, a 2-agent system is built. Each agent is with local state $s^i\in\{0,1\}$ and $s=[s^1,s^2]^{\top}$. The local action $a^i\in\{\text{left},\text{right},\text{hold}\}$, and each action is with different probability distributions to decrease, increase, or keep its local state. Local cost functions are $c^i(s,a)=b^i s^i$, where $b^1=2$ and $b^2=1$, and $g^i(s,a)=s^1*s^2+(1-s^1)*(1-s^2)$. $\bar{G}=0.5$. The parameters are as described in Subsection V-C, that $K=|\mathcal{S}||\mathcal{A}|-1$ and the state-action value functions are linear. Each agent maintains a vector $\boldsymbol{\theta}_s^i=[\theta_s^i(a_1),...,\theta_s^i(a_{|\mathcal{A}^i|})]$ for each state s. We first test Algorithm 2 with projection onto set Θ^i, $\sum_{a\in\mathcal{A}^i}\theta_s^i(a)=1$ and $\theta_s^i(a)\geq\epsilon$. This

induces a simplex set, $\mathcal{G}$. Then, we build set $\tilde{\Theta}^i \subset \Theta^i$ by making several balls and taking intersection, of which all the points on the relative boundary are extreme points, to take place Θ^i. This induces non-simplex convex and compact set $\tilde{\mathcal{G}}$.

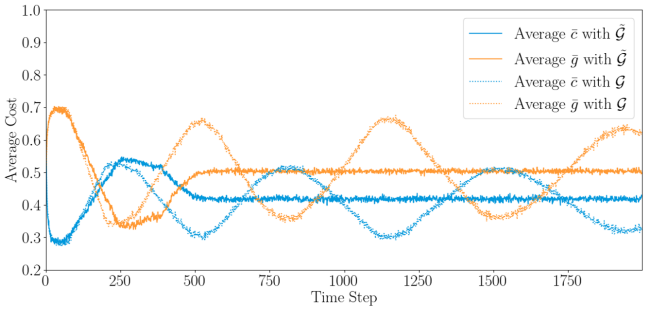

Fig. 1: The cost curves in case 1.

The results are shown in Fig. 1. It shows that with projection onto $\tilde{\Theta}^i$, Algorithm 2 converges to a feasible solution that $\bar{g}$ converges to 0.5. The policy with projection onto Θ^i heavily fluctuates between two deterministic policies. This comparison result validates the analysis in Subsection V-C and shows that Assumption 6 is critical to the convergence of the algorithms. Moreover, we randomly sample 1,000 policies and evaluate them through simulation. The result shows that among all the feasible policies, the average $\bar{c}$ with Algorithm 2 is the minimum.

B. Case 2

We test the developed algorithms on a 10-agent system on a two-dimensional map [8]. Each agent is aimed at moving to its own target landmark as shown in Fig. 2. The local action is a two-dimensional

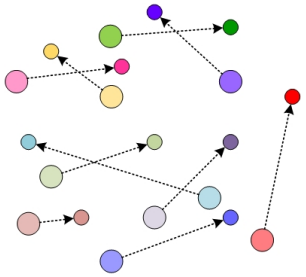

Fig. 2: The multi-agent system.

force put on the agent. Each agent observes the global state, including the location of all agents and landmarks as well as the velocity of each agent. At time t, the local cost c_t^i is the distance between the agent and its target landmark. The local cost g_t^i is the collision times with other agents. In this system, the agents are aimed at obtaining local policies to minimize the expectation of the global reward $\bar{c}_t$ such that the expectation of global collision times $\bar{g}_t$ is below a certain value $\bar{G}$. All the agents and landmarks are reset to some random locations every 25 time steps. The 3-layer neural networks are applied to represent the policies.

Algorithms 1 and 2 are tested with three different $\bar{G}$ values. For comparison, we apply the MARL method [1] that minimizes $J^{\hat{\lambda}}(\theta) = J(\theta) + \hat{\lambda}G(\theta)$, where $\hat{\lambda}$ is the fixed factor. The results are shown in Figs. 3-5[1]. And are compared against MADDPG [8] with surrogate local cost $\tilde{c}^i = c^i + 100 \times g^i$ as shown in TABLE I.

[1]Some results of Algorithm 1 are not presented since they are similar to Algorithm 2.

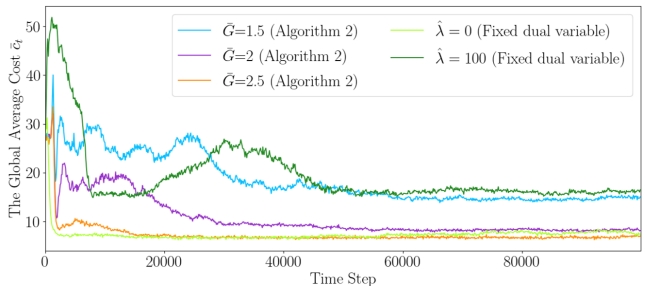

Fig. 3: The global average cost $\bar{c}_t$.

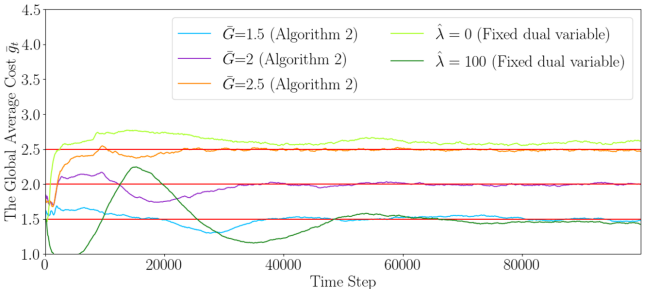

Fig. 4: The additional global average cost $\bar{g}_t$. The red lines are the $\bar{G}$ values, which are 1.5, 2, and 2.5, respectively.

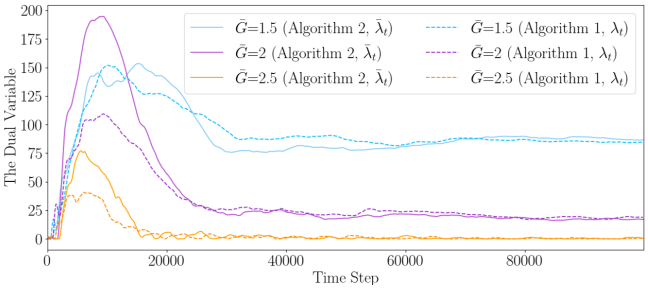

Fig. 5: The dual variables. For Algorithm 1, the plotted dual variable is the centralized λ_t, while for Algorithm 2 is the average value $\bar{\lambda}_t$.

TABLE I: Comparison results.

Method	Average $\bar{c}$	Average $\bar{g}$
Algorithm 2 ($\bar{G} = 1.5$)	14.95	1.51
Algorithm 2 ($\bar{G} = 2.0$)	8.30	1.94
Algorithm 2 ($\bar{G} = 2.5$)	7.01	2.46
Algorithm 2 (fixed $\hat{\lambda} = 100$)	16.47	1.15
MADDPG (surrogate cost $\tilde{c}^i$)	26.21	0.95

It shows that all the tested algorithms converge. In Fig. 3, $J(\theta)$ reaches a stable value in each test and decreases as $\bar{G}$ increases. This is because the agents need to take devious routes to avoid collisions. In Fig. 4, $G(\theta)$ generated with Algorithm 2 converges to $\bar{G}$. This implies that the limiting point of $\bar{\lambda}_t$ leads to the feasible policies that activate the system constraint. By comparison, $G(\theta)$ is not controlled directly with fixed $\hat{\lambda}$ [1] or MADDPG [8]. This results in violation of constraint or higher $J(\theta)$, while our algorithms adaptively tune λ to minimize $J(\theta)$ such that $G(\theta) \leq \bar{G}$. In addition, MADDPG achieves lower $G(\theta)$ than other methods but is with much higher $J(\theta)$. This is because each agent is aimed at its local costs with MADDPG, and thus it takes too much attention to avoiding local collisions.

We also verify that in all three tests for Algorithm 2, all the local dual variables λ_t^i converge to consensus although they have

different initialization. Fig. 5 shows that the consensus dual variables of Algorithm 2 converge to similar points as Algorithm 1. It also implies that the value of the dual variable increases as $\bar{G}$ decreases.

VII. Conclusion

In this article, we handle the problem of CMAMDPs which is applied in multiple areas. To preserve the users' local costs and policies, we develop a fully decentralized MARL algorithm, which is proven to converge. The comparison experiment validates the theoretical results and demonstrates the effectiveness of the developed algorithm.

Appendix

Consider the sequence $\{x_t \in \mathbb{R}^N_{t\geq 0}\}$ generated by recursion

$$x_{t+1} = \Gamma\big[x_t + a_t(h(x_t) + \xi_{1,t} + \xi_{2,t})\big],$$

where Γ projects the iterate x_t onto a compact and convex set. The ODE associated is given by

$$\dot{x} = \hat{\Gamma}(h(x)),$$

where

$$\hat{\Gamma}(h(x)) \lim_{c\to 0^+} \frac{\Gamma\big[x + ch(x)\big] - x}{c}$$

Assumption 7. *We make the following assumptions:*
(7.1) $h(\cdot)$ is a continuous $\mathbb{R}^N$-valued function.
(7.2) The sequence $\{\xi_{1,t}\}_{t\geq 0}$ is a bounded random sequence with $\xi_{1,t} \to 0$ almost surely as $t \to \infty$.
(7.3) The step-sizes $\{a_t\}_{t\geq 0}$ satisfy $a_t \to 0$ as $t \to \infty$ and $\sum_t a_t = \infty$.
(7.4) $\{\xi_{2,t}, t \leq 0\}$ is a sequence such that for any $\epsilon > 0$,

$$\lim_{n\to\infty} P(\sup_{m\leq n} \|\sum_{i=n}^{m} a_i \xi_{1,i}\| \leq \epsilon) = 0.$$

(7.5) The ODE has a compact subset χ of $\mathbb{R}^N$ as its set of asymptotically stable equilibrium points.

Lemma 6. *(Kushner-Clark Lemma [32]) Under Assumption 7, $\{x_t\}_{t\geq 0}$ converges almost surely to the set χ.*

References

[1] W. Liu, P. Zhuang, H. Liang, J. Peng, and Z. Huang, "Distributed economic dispatch in microgrids based on cooperative reinforcement learning," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 29, no. 6, pp. 2192–2203, 2018.

[2] C. Keerthisinghe, A. C. Chapman, and G. Verbič, "Pv and demand models for a markov decision process formulation of the home energy management problem," *IEEE Transactions on Industrial Electronics*, vol. 66, no. 2, pp. 1424–1433, 2019.

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

[4] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[5] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van Den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot *et al.*, "Mastering the game of go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016.

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

[7] L. Wang, A. D. Ames, and M. Egerstedt, "Safety barrier certificates for collisions-free multirobot systems," *IEEE Transactions on Robotics*, vol. 33, no. 3, pp. 661–674, 2017.

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

[9] L. Wang, A. D. Ames, and M. Egerstedt, "Safety barrier certificates for collisions-free multirobot systems," *IEEE Transactions on Robotics*, vol. 33, no. 3, pp. 661–674, 2017.

[10] Z. Wang, C. Shen, F. Liu, X. Wu, C.-C. Liu, and F. Gao, "Chance-constrained economic dispatch with non-gaussian correlated wind power uncertainty," *IEEE Transactions on Power Systems*, vol. 32, no. 6, pp. 4880–4893, 2017.

[11] X. Xu, J. W. Grizzle, P. Tabuada, and A. D. Ames, "Correctness guarantees for the composition of lane keeping and adaptive cruise control," *IEEE Transactions on Automation Science and Engineering*, vol. 15, no. 3, pp. 1216–1229, 2018.

[12] V. S. Borkar, "An actor-critic algorithm for constrained markov decision processes," *Systems & control letters*, vol. 54, no. 3, pp. 207–213, 2005.

[13] S. Bhatnagar, "An actor–critic algorithm with function approximation for discounted cost constrained markov decision processes," *Systems & Control Letters*, vol. 59, no. 12, pp. 760–766, 2010.

[14] S. Lu, K. Zhang, T. Chen, T. Basar, L. Horesh, R. Vinod, P. Y. Chen *et al.*, "Decentralized policy gradient descent ascent for safe multi-agent reinforcement learning," in *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 10, 2021, pp. 8767–8775.

[15] T. Degris, M. White, and R. Sutton, "Off-policy actor-critic," *Proceedings of the 29th International Conference on Machine Learning, ICML 2012*, vol. 1, 05 2012.

[16] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[17] T. Rashid, M. Samvelyan, C. Schroeder, G. Farquhar, J. Foerster, and S. Whiteson, "Qmix: Monotonic value function factorisation for deep multi-agent reinforcement learning," in *International Conference on Machine Learning*. PMLR, 2018, pp. 4295–4304.

[18] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Basar, "Fully decentralized multi-agent reinforcement learning with networked agents," in *International Conference on Machine Learning*. PMLR, 2018, pp. 5872–5881.

[19] G. Qu, A. Wierman, and N. Li, "Scalable reinforcement learning of localized policies for multi-agent networked systems," in *Learning for Dynamics and Control*. PMLR, 2020, pp. 256–266.

[20] G. Qu, Y. Lin, A. Wierman, and N. Li, "Scalable multi-agent reinforcement learning for networked systems with average reward," *Advances in Neural Information Processing Systems*, vol. 33, pp. 2074–2086, 2020.

[21] S. Prajna, A. Jadbabaie, and G. J. Pappas, "A framework for worst-case and stochastic safety verification using barrier certificates," *IEEE Transactions on Automatic Control*, vol. 52, no. 8, pp. 1415–1428, 2007.

[22] E. Altman, *Constrained Markov decision processes*. CRC Press, 1999, vol. 7.

[23] S. Paternain, L. Chamon, M. Calvo-Fullana, and A. Ribeiro, "Constrained reinforcement learning has zero duality gap," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[24] S. Lee, A. Nedić, and M. Raginsky, "Stochastic dual averaging for decentralized online optimization on time-varying communication graphs," *IEEE Transactions on Automatic Control*, vol. 62, no. 12, pp. 6407–6414, 2017.

[25] A. Nedić, A. Olshevsky, and W. Shi, "Achieving geometric convergence for distributed optimization over time-varying graphs," *SIAM Journal on Optimization*, vol. 27, no. 4, pp. 2597–2633, 2017.

[26] S. Boyd and L. Vandenberghe, *Convex Optimization*. Convex Optimization, 2004.

[27] J. He, L. Cai, C. Zhao, P. Cheng, and X. Guan, "Privacy-preserving average consensus: Privacy analysis and algorithm design," *IEEE Transactions on Signal and Information Processing over Networks*, vol. 5, no. 1, pp. 127–138, 2019.

[28] Y. Mo and R. M. Murray, "Privacy preserving average consensus," *IEEE Transactions on Automatic Control*, vol. 62, no. 2, pp. 753–765, 2017.

[29] H. J. Kushner and G. G. Yin, *Stochastic approximation algorithms and applications*. Stochastic approximation algorithms and applications, 1997.

[30] V. S. Borkar, "Convex analytic methods in markov decision processes," in *Handbook of Markov decision processes*. Springer, 2002, pp. 347–375.

[31] V. R. Konda and V. S. Borkar, "Actor-critic–type learning algorithms for markov decision processes," *Siam Journal on Control & Optimization*, vol. 38, no. 1, pp. 94–123, 1999.

[32] V. S. Borkar, *Stochastic approximation: a dynamical systems viewpoint*. Springer, 2009, vol. 48.