Multiagent Safe Reinforcement Learning by Decentralized Event-Triggered Min-Max Optimization

Shunsuke Otani, Naoki Hayashi, and Masahiro Inuiguchi

***Abstract*— This paper addresses decentralized event-triggered reinforcement learning with safety constraints. Each agent has an individual reward function and safety constraints that depend on the joint actions of agents. The objective is to maximize the team's long-term return while satisfying the safety constraints. We formulate the reinforcement learning problem as a nonconvex-concave min-max optimization problem and propose a decentralized policy gradient algorithm. Each agent has estimations for the optimal primal and dual solutions of the min-max optimization problem. Different from the existing decentralized reinforcement learning, the agents share these estimates only when the error exceeds a predefined threshold. We show that the estimates of agents converge to a neighborhood of a locally optimal solution while effectively reducing the communication overhead.**

I. Introduction

Large-scale AI systems often require handling vast amounts of data and computational resources. Recently, decentralized learning has attracted attention due to its efficiency and scalability. Within this context, cooperative optimization has played a crucial role in various engineering fields [1]–[4]. The cooperative optimization methods have been extended to reinforcement learning [5]–[7]. In multi-agent reinforcement learning (MARL), several agents interact with the environment in parallel and take actions based on local observation.

Despite its numerous advantages, developing MARL methods for real-world applications remains a significant challenge. Many real-world applications require agents to operate under strict safety conditions while ensuring optimal performance [8]. Without proper safety constraints, agents may take actions that deteriorate the stability of the system. To address these challenges, various approaches have been proposed for safe MARL. Ying et al. introduced a cooperative safe MARL framework based on a primal-dual actor-critic method [9]. Lu et al. proposed a decentralized descent-ascent algorithm for safe reinforcement learning with gradient tracking [10]. Hassan et al. developed a consensus-based primal-dual algorithm for safe MARL that accounts for both peak and average constraints [11].

However, these existing methods for safe MARL face significant practical limitations in real-world applications due to their requirement for constant communication at every iteration, which consumes substantial communication resources. To address this issue, an event-triggered scheme has been applied as an alternative communication approach [12]–[15]. In event-triggered algorithms, agents communicate selectively by transmitting information to their neighbors only when certain conditions are met. Unlike periodic communication schemes, event-triggered mechanisms dynamically adjust the frequency of communication based on real-time system behavior.

This paper considers a decentralized safe reinforcement learning problem and formulate it as a nonconvex-concave min-max optimization problem. We propose an event-triggered decentralized primal-dual policy gradient algorithm. In this approach, each agent has an estimate of the optimal policy parameters and updates these estimates through local policy gradient steps while considering safety constraints via Lagrange multipliers. To reduce communication overhead, an event-triggering mechanism is employed, where each agent transmits its current estimate only when the deviation from its last transmitted value exceeds a predefined threshold. The theoretical analysis of the algorithm shows that the agents' estimates reach consensus and converge within a neighborhood of a locally optimal solution.

The proposed event-triggered decentralized safe MARL algorithm provides a significant advantage over existing methods [8]–[11] by integrating safety constraints and an event-triggered communication scheme into the learning process. This integration ensures that agents optimize their decision-making while strictly adhering to operational limits. At the same time, the event-triggered mechanism reduces unnecessary communication by allowing agents to share information only when a critical change occurs in their local estimates. This selective communication strategy is particularly advantageous in bandwidth-constrained networks, where excessive data transmission can lead to congestion and increased latency.

Moreover, the analysis of decentralized min-max problems often relies on the assumption that the objective function is strongly concave with respect to the dual variable [16], [17]. However, in many practical applications, the strong concavity condition may not be satisfied. The proposed method addresses this gap by extending the theoretical framework of prior work to accommodate scenarios where the objective function is merely concave, rather than strongly concave. By relaxing this assumption, the proposed approach broadens the scope of decentralized MARL algorithms while still maintaining convergence and performance guarantees.

This paper is organized as follows. In Section II, we address the problem formulation of safe MARL. In Section

This work partially supported by JSPS KAKENHI Grant Number JP21K04121.

S. Otani, N. Hayashi, and M. Inuiguchi are with the Graduate School of Engineering Science, The University of Osaka, Toyonaka, Osaka 560-8531, Japan (e-mail: otani@inulab.sys.es.osaka-u.ac.jp, {n.hayashi, inuiguti}@sys.es.osaka-u.ac.jp). (Corresponding author: Shunsuke Otani)

III, we propose a decentralized learning algorithm with event-triggered communication. The analysis of the proposed algorithm is shown in Section IV. The numerical example is presented in Section V. Finally, concluding remarks are given in Section VI.

II. Safe Multiagent Reinforcement Learning

The set of nonnegative real numbers is represented by $\mathbb{R}_+$. For a set $\mathcal{S}$, $|\mathcal{S}|$ represents the cardinality of $\mathcal{S}$. For a set $\mathcal{S}$, $\Delta(\mathcal{S})$ represents the probability simplex. A sequence of vectors $(a_{i,1}, a_{i,2}, \dots, a_{i,t})$ is represented as $a_{[1,t]}$. The projection of $x \in \mathbb{R}^n$ onto a point in $\mathbb{R}^n_+$ is represented by $[x]_+$. The projection of $x \in \mathbb{R}^n$ to a closed convex set $\mathcal{S} \subseteq \mathbb{R}^n$ is given by $\Pi_{\mathcal{S}}(x) = \arg\min_{y\in\mathcal{S}} \|x-y\|$. For $\gamma > 0$ and $v \in \mathbb{R}^n$, the proximal operator of $f : \mathbb{R}^n \to \mathbb{R}$ is defined by $\operatorname{prox}_{f,\gamma}(v) = \arg\min_{x\in\mathbb{R}^n} \left\{f(x) + (1/2\gamma)\|x - v\|^2\right\}$.

We consider a multiagent system with N agents. Each agent makes independent decisions and exchanges information with others through a communication network. The network is modeled as an undirected graph $G = (\mathcal{V}, \mathcal{E})$, where $\mathcal{V} = \{1, 2, \dots, N\}$ is the set of nodes that correspond to agents and $\mathcal{E}$ is the set of edges that represents communication links between agents. The graph G satisfies the following assumption in this paper.

Assumption 1: G is connected.

The decentralized safe MARL problem is modeled as a Markov decision process $\langle \mathcal{S}, \{\mathcal{A}_i\}_{i=1}^N, P, \{R_i\}_{i=1}^N, \{C_i\}_{i=1}^N\rangle$, where $\mathcal{S}$ is the global state space, $\mathcal{A}_i$ represents the individual action space, $P : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the probability distribution function, $\mathcal{A} = \mathcal{A}_1 \times \mathcal{A}_2 \times \cdots \times \mathcal{A}_n$ is the joint action space, $R_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the individual reward function, and $C_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}^m$ is the constraint function.

A policy $\pi_i^\theta : \mathcal{S} \to \Delta(\mathcal{A}_i)$ defines a probabilistic decision-making rule for agent i by mapping each state $s_i \in \mathcal{S}$ to a probability distribution over the available actions in the action space $\mathcal{A}$. The policy π_i^θ is parameterized by $\theta \in \mathbb{R}^n$, which is a tunable parameter that influences how the agents select actions in different states. For simplicity of notation, we denote the parameterized policy π_i^θ as π_i throughout the remainder of this paper. The policy satisfies $0 \le \pi_i(a_i \mid s_i) \le 1$ for all $a_i \in \mathcal{A}$, $s_i \in \mathcal{S}$, and $i \in \mathcal{V}$, and $\sum_{a_i\in\mathcal{A}} \pi_i(a_i \mid s_i) = 1$ for all $s_i \in \mathcal{S}$ and $i \in \mathcal{V}$. The joint policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ is given by $\pi(a_t \mid s_t) = \prod_{i=1}^N \pi_i(a_{i,t} \mid s_{i,t})$, where $a_t = (a_{1,t}, a_{2,t}, \dots, a_{N,t})$ and $s_t = (s_{1,t}, s_{2,t}, \dots, s_{N,t})$.

At each iteration t, the agents select an action randomly according to the policy distribution, that is, $a_t \sim \pi(\cdot \mid s_t)$. Once an action is taken, the next state s_{t+1} is given by the transition probability function $P(\cdot \mid s_t, a_t)$. The probability distribution over trajectories, induced by a given policy π, shows the likelihood of observing a particular sequence of states and actions. This distribution is given by $p(\tau \mid \pi) = \rho(s_0) \prod_{t=0}^\infty \pi(a_t \mid s_t) P(s_{t+1} \mid s_t, a_t)$.

The goal of the agents is to estimate the parameters θ by solving the following safe reinforcement learning problem:

$$\text{minimize} \quad \frac{1}{N}\sum_{i=1}^N J_i^R(\theta) \tag{1a}$$

$$\text{subject to} \quad J_i^C(\theta) \le d_i, \quad \forall i \in \mathcal{V}, \tag{1b}$$

where $J_i^R(\theta) = \mathbb{E}_{\tau\sim p(\cdot|\pi)}\left[\sum_{t=0}^\infty \gamma_i^t R_i(s_t, a_t)\right]$ and $J_i^C(\theta) = \mathbb{E}_{\tau\sim p(\cdot|\pi)}\left[\sum_{t=0}^\infty \gamma_i^t C_i(s_t, a_t)\right]$ represent the expected cumulative reward and cost, respectively, and $d_i \in \mathbb{R}^{m_i}$ is a constant vector specifying the constraint bounds.

To address this constrained optimization problem, we introduce the following Lagrange function:

$$f(\theta, \lambda) = \frac{1}{N}\sum_{i=1}^N f_i(\theta, \lambda), \tag{2}$$

where $f_i(\theta, \lambda) = J_i^R(\theta) + (1/N)\lambda^\top J^C(\theta)$, $\lambda \in \mathbb{R}^m_+$ is the Lagrange variable, $m = \sum_{i=1}^N m_i$ is the total number of all constraints, and $J^C(\theta) = [(J_1^C(\theta))^\top, (J_2^C(\theta))^\top, \dots, (J_N^C(\theta))^\top]^\top$. By using the Lagrangian function, the original constrained reinforcement learning problem (1) can be reformulated into the following min-max optimization problem:

$$\underset{\theta\in\Theta}{\text{minimize}} \quad \frac{1}{N}\sum_{i=1}^N \Phi_i(\theta), \tag{3}$$

where

$$\Phi_i(\theta) = \max_{\lambda\in\mathbb{R}^m_+} f_i(\theta, \lambda). \tag{4}$$

The min-max optimization problem in (3) presents analytical challenges because it does not possess convexity with respect to the primal variable, nor does it exhibit strong concavity with respect to the dual variable. In particular, previous works [16], [17] rely on the assumption that the objective function is strongly concave in the dual variable. As a result, their analytical frameworks and convergence guarantees cannot be directly applied to this problem setting. To address these challenges, the following sections introduce a decentralized event-triggered primal-dual algorithm for the nonconvex-concave min-max optimization problem (3) through the analysis of the Moreau envelope [18].

III. Decentralized Event-Triggered Algorithm for Safe MARL

In this section, we propose a decentralized event-triggered algorithm for the safe multi-agent reinforcement learning problem (1). Let $\theta_{i,t}$ and $\lambda_{i,t}$ be agent i's optimal estimations for the parameter $\theta^* \in \Theta$ and the optimal solution to the Lagrange dual problem $\lambda^* \in \Lambda$ at iteration t, where $\Theta \subset \mathbb{R}^n$ and $\Lambda \subset \mathbb{R}^m_+$ are the feasible solution sets for the parameters and the Lagrange variables. The proposed algorithm for agent i is summarized in Algorithm 1. For convenience, we call the solution to the problem (1) and that of the problem (3) "primal solution" and "dual solution", respectively.

Each agent sets the initial primal solution $\hat{\theta}_{i,0} = \theta_{i,0} \in \Theta$ and the dual solution $\hat{\lambda}_{i,0} = \lambda_{i,0} \in \Lambda$. The initial step sizes $\eta_\theta > 0$ and $\eta_\lambda > 0$ should be set adequately.

Algorithm 1 Decentralized Event-Triggered Algorithm for Safe MARL

Input: Set initial values: $\hat{\theta}_{i,0} = \theta_{i,0} \in \Theta$ and $\hat{\lambda}_{i,0} = \lambda_{i,0} \in \Lambda$. Set stepsizes $\eta_\theta > 0$ and $\eta_\lambda > 0$.

for $t = 1, 2, \ldots, T$ **do**

Update the estimates by

$$z_{i,t} = \theta_{i,t-1} + \sum_{j=1}^{N} p_{ij}(\hat{\theta}_{j,t-1} - \hat{\theta}_{i,t-1}), \tag{5}$$

$$w_{i,t} = \lambda_{i,t-1} + \sum_{j=1}^{N} p_{ij}(\hat{\lambda}_{j,t-1} - \hat{\lambda}_{i,t-1}), \tag{6}$$

$$\theta_{i,t} = \Pi_\Theta \left[z_{i,t} - \eta_\theta \nabla_\theta f_i(\theta_{i,t-1}, \lambda_{i,t-1}) \right], \tag{7}$$

$$\lambda_{i,t} = \Pi_\Lambda \left[w_{i,t} + \eta_\lambda \nabla_\lambda f_i(\theta_{i,t-1}, \lambda_{i,t-1}) \right]. \tag{8}$$

if $\|\theta_{i,t} - \hat{\theta}_{i,t-1}\| \geq e^\theta_{i,t}$ or $\|\lambda_{i,t} - \hat{\lambda}_{i,t-1}\| \geq e^\lambda_{i,t}$ **then**

Set $\hat{\theta}_{i,t} = \theta_{i,t}$ and $\hat{\lambda}_{i,t} = \lambda_{i,t}$.

else

Set $\hat{\theta}_{i,t} = \hat{\theta}_{i,t-1}$ and $\hat{\lambda}_{i,t} = \hat{\lambda}_{i,t-1}$.

end if

end for

During each iteration t, every agent i first performs a consensus step by collecting information from its neighbors. The agent computes the values of intermediate variables $z_{i,t}$ and $w_{i,t}$ by combining its own previous values $\theta_{i,t-1}$ and $\lambda_{i,t-1}$ with weighted differences between the most recently received information from its neighboring agents, where the edge weight with agent j is predetermined as p_{ij}.

After the consensus step, each agent updates its primal variable $\theta_{i,t}$ by taking a projected gradient descent step on the Lagrange function f_i. At the same time, the dual variable $\lambda_{i,t}$ is also updated via a projected gradient ascent step.

The event-triggering mechanism determines whether an agent should communicate its updated values to neighbors. The agent checks if the difference between the newly computed values and the previously shared values exceeds the thresholds $e^\theta_{i,t}$ for the primal variable or $e^\lambda_{i,t}$ for the dual variable. If either threshold is exceeded, the agent updates its shared values $\hat{\theta}_{i,t}$ and $\hat{\lambda}_{i,t}$ to the newly computed values and communicates these to its neighbors. Otherwise, the agent keeps its previously shared values unchanged and skips communication for that iteration.

IV. Convergence Analysis

In this section, we present a theoretical analysis of the proposed decentralized event-triggered primal-dual algorithm. Due to space limitations, the proofs of the results in this section are not included in this paper.

Since Θ and Λ are closed and convex, there exist positive constants D_θ and D_λ such that $\|\theta_{i,t}\| \leq D_\theta$ and $\|\lambda_{i,t}\| \leq D_\lambda$ for any $i \in \mathcal{V}$ and $t \geq 1$. Moreover, there exists a positive constant L such that $\|\nabla_\theta f_i(\theta, \lambda)\| \leq L$ and $\|\nabla_\lambda f_i(\theta, \lambda)\| \leq L$ for all $i \in \mathcal{V}$, $\theta \in \Theta$, and $\lambda \in \Lambda$.

In this paper, we assume the smoothness of the Lagrange function in the primal variable.

Assumption 2: For any $i \in \mathcal{V}$ and any $\theta, \theta' \in \mathbb{R}^n$, there exists a positive constant ℓ such that $\|\nabla f_i(\theta, \cdot) - \nabla f_i(\theta', \cdot)\| \leq \ell \|\theta - \theta'\|$.

To facilitate the convergence analysis of the proposed algorithm, we introduce the concept of the Moreau envelope, which is a fundamental tool in nonsmooth optimization. The Moreau envelope provides a smooth approximation of potentially nonsmooth functions and allows us to analyze the convergence of the algorithm using differentiable surrogates. In this paper, we consider the following Moreau envelope [18]: $\Phi_{i,1/(2\ell)}(x) = \min_{w \in \mathbb{R}^n} \left\{ \Phi_i(w) + \ell \|w - x\|^2 \right\}$. According to the results in [18], it is known that under Assumption 2, the Moreau envelope $\Phi_{i,1/(2\ell)}$ is not only convex but also differentiable, even when the original function Φ_i may be nonsmooth. In fact, the gradient of the Moreau envelope is explicitly given by

$$\nabla \Phi_{i,1/(2\ell)}(x) = 2\ell(x - \mathrm{prox}_{\Phi_i, 1/(2\ell)}(x)). \tag{9}$$

To proceed with the analysis, we add the following assumption that the distance between a given point and its proximal point is uniformly bounded [18].

Assumption 3: For any $x \in \mathbb{R}^n$, $i \in \mathcal{V}$, and $t \geq 1$, there exists a positive constant D_ϕ such that $\|x - \mathrm{prox}_{\Phi_i, 1/(2\ell)}(x)\| \leq D_\phi$.

From (9), Assumption 3 ensures that the gradient of the Moreau envelope is bounded, which is essential for deriving convergence guarantees.

We further introduce the following assumption on the edge weight.

Assumption 4: For each agent $i \in \mathcal{V}$, there exists a positive constant $c_p \in (0, 1)$ such that

$$p_{ij} \begin{cases} \geq c_p, & j \in \mathcal{N}_i, \\ = 0, & j \notin \mathcal{N}_i, \end{cases}$$

and $p_{ii} = 1 - \sum_{j \in \mathcal{N}_i} p_{ij} > c_p$ for all $j \in \mathcal{V}$, where $\mathcal{N}_i = \{j \in \mathcal{V} \mid \{i, j\} \in \mathcal{E}\}$ is the neighbor set of agent i.

Finally, we introduce an assumption concerning the event-triggered communication mechanism. In the proposed algorithm, agents share their local information with neighbors only when a triggering condition is satisfied, that is, when the deviation of their local estimates exceeds a specified threshold. Let e_t be the upper bound of the allowable trigger error such that $e_t = \max\{\max_{i \in \mathcal{V}} e^\theta_{i,t}, \max_{i \in \mathcal{V}} e^\lambda_{i,t}\}$.

Assumption 5: There exists a positive constant E satisfying $\sum_{t=0}^{\infty} e_t < E$.

To analyze the effect of event-triggered communication on the algorithm's behavior, we introduce notation to represent the local trigger errors. Let $h^\theta_{i,t}$ and $h^\lambda_{i,t}$ be the trigger errors defined by $h^\theta_{i,t} = \theta_{i,t} - \hat{\theta}_{i,t}$ and $h^\lambda_{i,t} = \lambda_{i,t} - \hat{\lambda}_{i,t}$, respectively. From (5), we have $z_{i,t} = \sum_{j=1}^N p_{ij}\theta_{j,t-1} + q^\theta_{i,t-1}$ and $w_{i,t} = \sum_{j=1}^N p_{ij}\lambda_{j,t-1} + q^\lambda_{i,t-1}$, where $q^\theta_{i,t} = \sum_{j=1}^N p_{ij}(h^\theta_{i,t} - h^\theta_{j,t})$ and $q^\lambda_{i,t} = \sum_{j=1}^N p_{ij}(h^\lambda_{i,t} - h^\lambda_{j,t})$.

We provide several key lemmas that quantify the behavior of the primal and dual variables over time. The first result, Lemma 1, evaluates an upper bound on the deviation of

each agent's local variables from their corresponding average values.

Lemma 1: Under Assumptions 1–5, for any $i \in \mathcal{V}$ and $t \geq 1$, we have

$$\begin{aligned}&\|\theta_{i,t} - \bar{\theta}_t\| \\ &\leq CND_\theta \gamma^{t-1} + 2\left(\frac{CN}{1-\gamma} + 2\right)(2e_{t-1} + L\eta_\theta), \quad (10) \\ &\|\lambda_{i,t} - \bar{\lambda}_t\| \\ &\leq CND_\lambda \gamma^{t-1} + 2\left(\frac{CN}{1-\gamma} + 2\right)(2e_{t-1} + L\eta_\lambda), \quad (11)\end{aligned}$$

where $\bar{\theta}_t = (1/N)\sum_{i=1}^N \theta_{i,t}$, $\bar{\lambda}_t = (1/N)\sum_{i=1}^N \lambda_{i,t}$, $C > 0$, and $0 < \gamma < 1$.

From Assumption 5, the allowable trigger error converges to 0. Thus, this lemma shows that, for sufficiently small step sizes, agents can achieve consensus even in the presence of sparse and irregular communication by event-triggered scheme.

The next lemma shows a bound on the aggregate gradient of the Moreau envelope.

Lemma 2: Under Assumptions 1–5, for any $i \in \mathcal{V}$ and $t \geq 1$, we have

$$\begin{aligned}&\left\|\frac{1}{N}\sum_{i=1}^N \Phi_{i,1/(2\ell)}(\bar{\theta}_{t-1})\right\|^2 \\ &\leq \frac{4\ell}{N\eta_\theta}\sum_{i=1}^N(\|\tilde{\theta}_{i,t-1} - \bar{\theta}_{t-1}\|^2 - \|\tilde{\theta}_{i,t-1} - \bar{\theta}_t\|^2) \\ &\quad + \frac{8\ell}{N}\Delta_{t-1} + 24\ell L^2\eta_\theta + 24\ell LD \\ &\quad + 16\ell\left(L + \frac{D}{\eta_\theta}\right)e_{t-1} + \frac{32\ell}{\eta_\theta}e_{t-1}^2, \quad (12)\end{aligned}$$

where $\Delta_t = \sum_{i=1}^N \Delta_{i,t}$, $\Delta_{i,t} = \Phi_i(\bar{\theta}_t) - f_i(\bar{\theta}_t, \bar{\lambda}_t)$, and $D = \max\{D_\theta, D_\phi\}$.

The next lemma provides a bound on the time-averaged duality gap.

Lemma 3: Under Assumptions 1–4, for all $i \in \mathcal{V}$, we have

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^T \Delta_t \\ &\leq \left(\Delta_0 + NL\left(12E\left(\frac{CN}{1-\gamma} + 2\right) + \frac{CN(2D_\theta + D_\lambda)}{1-\gamma}\right)\right. \\ &\left. + \frac{4NE(2D_\lambda + L\eta_\lambda + 4E)}{\eta_\lambda} + \hat{\Delta}_0 + 2LND_\lambda\right)\frac{1}{T} \\ &+ 2NL^2\left(\frac{2CN}{1-\gamma} + 7\right)\eta_\theta \\ &+ NL^2\left(\frac{2CN}{1-\gamma} + \frac{13}{2}\right)\eta_\lambda + \frac{2ND_\lambda^2}{\eta_\lambda} + 2NL(3E + 2D_\lambda), \quad (13)\end{aligned}$$

where $\hat{\Delta}_0 = f(\bar{\theta}_0, \bar{\lambda}_0) - \min_x f(x, \bar{\lambda}_0)$.

The result shows that by properly tuning the step sizes and the trigger threshold, the duality gap is bounded.

By combining the results of the previous lemmas, we arrive at the main convergence theorem of the proposed algorithm. Theorem 1 characterizes the convergence behavior of the decentralized event-triggered primal-dual algorithm in terms of the average norm of the aggregated gradient of the Moreau envelopes.

Theorem 1: Under Assumptions 1–4, for all $i \in \mathcal{V}$, we have

$$\begin{aligned}&\frac{1}{T}\sum_{t=1}^T \left\|\frac{1}{N}\sum_{i=1}^N \Phi_{i,1/(2\ell)}(\bar{\theta}_t)\right\|^2 \\ &\leq \frac{8\ell}{N}\left(\Delta_0 + NL\left(12E\left(\frac{CN}{1-\gamma} + 2\right) + \frac{CN(2D_\theta + D_\lambda)}{1-\gamma}\right)\right. \\ &\quad + \frac{4NE(2D_\lambda + L\eta_\lambda + 4E)}{\eta_\lambda} + \frac{D^2 + 8NE^2}{2\eta_\theta} \\ &\quad \left. + \hat{\Delta}_0 + 2LND_\lambda\right)\frac{1}{T} \\ &\quad + \frac{8\ell L^2}{N}\left(\frac{4CN}{1-\gamma} + 14 + 3N\right)\eta_\theta + \frac{16\ell ED}{\eta_\theta} \\ &\quad + \frac{8\ell L^2}{N}\left(\frac{2CN}{1-\gamma} + \frac{13}{2}\right)\eta_\lambda + \frac{16\ell D_\lambda^2}{\eta_\lambda} \\ &\quad + 8\ell L(8E + 4D_\lambda + 3D). \quad (14)\end{aligned}$$

Theorem 1 confirms that, despite the reduced communication induced by the event-triggered scheme, the algorithm achieves near-optimal convergence.

V. Numerical Example

We consider a 4×4 Cliffwalk environment in Fig. 1 [19]. Each agent incurs a cost of 1 for every action, representing movement cost or energy consumption. Entering a cliff cell at $(3,1)$ or $(3,2)$ incurs a penalty of cost 5, after which the agent is returned to the starting point. Slippery cells at $(2,1)$ and $(2,2)$ have a 10% chance of causing an unintended transition into the cliff, modeling indirect risk.

Three agents, starting from $(3,0)$, $(2,2)$, and $(1,2)$, aim to collaboratively minimize the expected cumulative cost while avoiding risky areas. The communication topology among the agents is shown in Fig. 2.

We set the discount factor to $\gamma = 0.9999$ to emphasize long-term performance. Both the primal and dual step sizes are set to $\eta_\theta = \eta_\lambda = 0.001$. The event-triggering thresholds for both primal and dual updates are set as $e_{i,t}^\theta = e_{i,t}^\lambda = C_e/\sqrt{t}$ for all agents, where C_e is a positive constant. The agents aim to learn policies that satisfy the cost constraint in expectation, with a target threshold of 6 corresponding to the safety policy.

Fig. 3 compares the total loss of the agent starting at $(3,0)$ under event-triggered and time-triggered schemes. In the time-triggered method, where the communication threshold is set to $C_e = 0$, agents communicate their policy parameter updates at every iteration. In contrast, the event-triggered method with $C_e = 50$ allows agents to communicate only when the magnitude of their update exceeds the threshold. Despite this difference, Fig. 3 shows that both methods achieve a steady decline in loss over time, which confirms

Cost=1 (0,0)	Cost=1 (0,1)	Cost=1 (0,2)	Cost=1 (0,3)
Cost=1 (1,0)	Cost=1 (1,1)	Cost=1 (1,2)	Cost=1 (1,3)
Cost=1 (2,0)	Cost=1 (2,1)	Cost=1 (2,2)	Cost=1 (2,3)
Cost=1 (3,0) Start	Cost=5 (3,1) Cliff	Cost=5 (3,2) Cliff	Cost=0 (3,3) Goal

Fig. 1. Cliffwalk environment

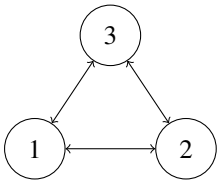

Fig. 2. Communication network

the effective learning of the proposed event-triggered algorithm. We note that the event-triggered scheme achieves this performance with reduced communication overhead as shown in Fig. 4, which illustrates that the total number of communications required by each agent is reduced by more than 60% compared to the time-triggered approach. These results demonstrate that the proposed event-triggered strategy offers an effective balance between learning performance and communication efficiency.

Figure 5 illustrates the evolution of the cumulative cost incurred by agent 1 under both constrained and unconstrained policy settings. In the constrained case, the agent adheres to safety requirements by selecting actions that avoid high-risk areas, where the total cost converges to 6. This is achieved by taking the longer but safer route: $(3,0) \rightarrow (2,0) \rightarrow (1,0) \rightarrow (1,1) \rightarrow (1,2) \rightarrow (1,3) \rightarrow (2,3) \rightarrow (3,3)$. In contrast, the unconstrained policy, which ignores the cumulative cost, chooses a shorter but riskier route: $(3,0) \rightarrow (2,0) \rightarrow (2,1) \rightarrow (2,2) \rightarrow (2,3) \rightarrow (3,3)$. This route results in a lower cumulative cost of 4, but increases the likelihood of encountering undesirable transitions. The comparison between these results demonstrates the trade-off between performance optimization and safety assurance. By incorporating safety constraints into the policy optimization process, agents are guided to avoid potentially hazardous scenarios even if it results in a higher cumulative cost. Therefore, the proposed event-triggered reinforcement learning framework is suitable for safety-critical and resource-constrained networks where constraint violations can lead to significant negative consequences.

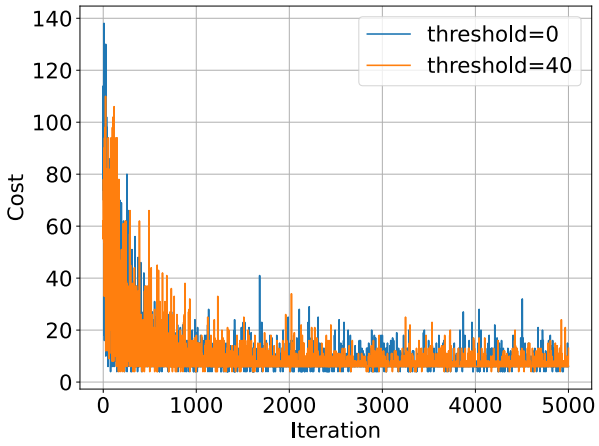

Fig. 3. Total loss comparison between event-triggered and time-triggered schemes.

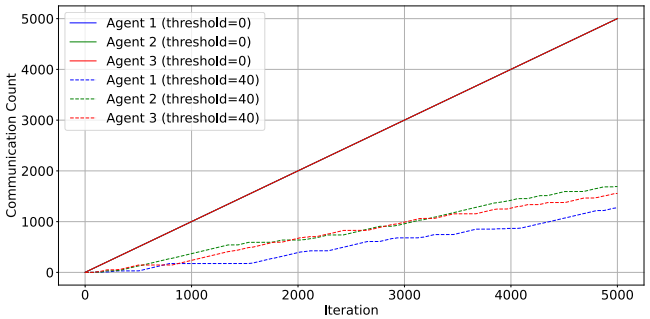

Fig. 4. Communication count comparison between event-triggered and time-triggered schemes.

VI. Conclusion

This paper investigated an event-triggered safe reinforcement learning algorithm for multi-agent systems under safety constraints. By formulating the learning problem as a nonconvex-concave min-max optimization, we introduced a decentralized primal-dual policy gradient method that enables agents to collaboratively optimize their policies while satisfying system constraints. Theoretical analysis showed that the proposed algorithm ensures consensus among agents and that their policy estimates converge within a bounded neighborhood of a locally optimal solution while effectively reducing the number of communications. Future work will consider further refinements to the event-triggering algorithm under more practical conditions such as communication delays and packet loss.

References

[1] A. Nedić and A. Ozdaglar, "Distributed subgradient methods for multi-agent optimization," *IEEE Transactions on Automatic Control*, vol. 54, no. 1, pp. 48–61, 2009.

[2] J. Li, G. Chen, Z. Wu, and X. He, "Distributed subgradient method for multi-agent optimization with quantized communication," *Mathematical Methods in the Applied Sciences*, vol. 40, no. 4, pp. 1201–1213, 2017.

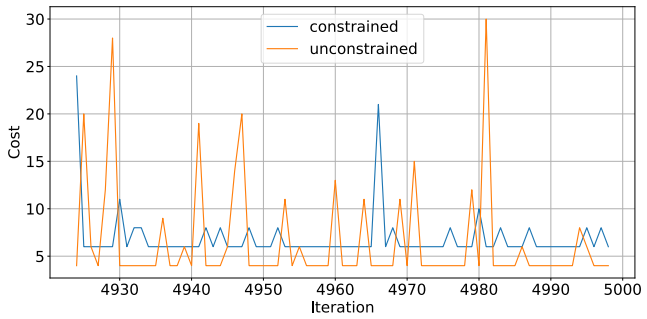

Fig. 5. Costs incurred by the agent under safety constrained and unconstrained schemes.

[3] G. Carnevale, F. Farina, I. Notarnicola, and G. Notarstefano, "GTAdam: Gradient tracking with adaptive momentum for distributed online optimization," *IEEE Transactions on Control of Network Systems*, vol. 10, no. 3, pp. 1436–1448, 2023.

[4] X. Zhou, Z. Ma, S. Zou, and K. Margellos, "Distributed momentum-based multiagent optimization with different constraint sets," *IEEE Transactions on Automatic Control*, vol. 70, no. 2, pp. 963–978, 2025.

[5] T. T. Nguyen, N. D. Nguyen, and S. Nahavandi, "Deep reinforcement learning for multiagent systems: A review of challenges, solutions, and applications," *IEEE Transactions on Cybernetics*, vol. 50, no. 9, pp. 3826–3839, 2020.

[6] Y. Zhang, M. Qiu, and H. Gao, "Communication-efficient stochastic gradient descent ascent with momentum algorithms," *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pp. 4602–4610, 2023.

[7] Z. Ning and L. Xie, "A survey on multi-agent reinforcement learning and its application," *Journal of Automation and Intelligence*, vol. 3, no. 2, pp. 73–91, 2024.

[8] Q. Zhang, K. Dehghanpour, Z. Wang, F. Qiu, and D. Zhao, "Multi-agent safe policy learning for power management of networked microgrids," *IEEE Transactions on Smart Grid*, vol. 12, no. 2, pp. 1048–1062, 2020.

[9] D. Ying, Y. Zhang, Y. Ding, A. Koppel, and J. Lavaei, "Scalable primal-dual actor-critic method for safe multi-agent rl with general utilities," *Advances in Neural Information Processing Systems*, vol. 36, pp. 36 524–36 539, 2023.

[10] S. Lu, K. Zhang, T. Chen, T. Başar, and L. Horesh, "Decentralized policy gradient descent ascent for safe multi-agent reinforcement learning," *Proceedings of the AAAI Conference on Artificial Intelligence*, vol. 35, no. 10, pp. 8767–8775, 2021.

[11] R. Hassan, K. S. Wadith, M. M. o. Rashid, and M. M. Khan, "Depaint: a decentralized safe multi-agent reinforcement learning algorithm considering peak and average constraints," *Applied Intelligence*, vol. 54, no. 8, pp. 6108–6124, 2024.

[12] Y. Kajiyama, N. Hayashi, and S. Takai, "Distributed subgradient method with edge-based event-triggered communication," *IEEE Transactions on Automatic Control*, vol. 63, no. 7, pp. 2248–2255, 2018.

[13] W. Wang, R. Postoyan, D. Nešić, and W. P. M. H. Heemels, "Periodic event-triggered control for nonlinear networked control systems," *IEEE Transactions on Automatic Control*, vol. 65, no. 2, pp. 620–635, 2020.

[14] X. Cao and T. Başar, "Decentralized online convex optimization with event-triggered communications," *IEEE Transactions on Signal Processing*, vol. 69, pp. 284–299, 2021.

[15] K. Okamoto, N. Hayashi, and S. Takai, "Distributed online adaptive gradient descent with event-triggered communication," *IEEE Transactions on Control of Network Systems*, vol. 11, no. 2, pp. 610–622, 2024.

[16] X. Zhang, Z. Liu, J. Liu, Z. Zhu, and S. Lu, "Taming communication and sample complexities in decentralized policy evaluation for cooperative multi-agent reinforcement learning," *Advances in Neural Information Processing Systems*, vol. 34, pp. 18 825–18 838, 2021.

[17] L. Chen, H. Ye, and L. Luo, "An efficient stochastic algorithm for decentralized nonconvex-strongly-concave minimax optimization," *Proceedings of The 27th International Conference on Artificial Intelligence and Statistics*, pp. 1990–1998, 2024.

[18] D. Davis and D. Drusvyatskiy, "Stochastic model-based minimization of weakly convex functions," *SIAM Journal on Optimization*, vol. 29, no. 1, pp. 207–239, 2019.

[19] X. Yu and L. Ying, "On the global convergence of risk-averse policy gradient methods with expected conditional risk measures," *Proceedings of the 40th International Conference on Machine Learning*, vol. 202, pp. 40 425–40 451, 2023.