Differentially Private Federated Temporal Difference Learning

Yiming Zeng, Yixuan Lin, Yuanyuan Yang, *Fellow, IEEE*, and Ji Liu

Abstract—This article considers a federated temporal difference (TD) learning algorithm and provides both asymptotic and finite-time analyses. To protect each worker agent's cost information from being acquired by possible attackers, we propose a privacy-preserving variant of the algorithm by adding perturbation to the exchanged information. We show the rigorous differential privacy guarantee by using moments accountant and derive an upper bound of the utility loss for the privacy-preserving algorithm. Evaluations are also provided to corroborate the efficiency of the algorithms.

Index Terms—Multi-agent reinforcement learning, TD learning, federated learning, differential privacy

1 Introduction

Reinforcement learning (RL) techniques have been widely applied in various domains, such as board games [1], MOBA games [2], autonomous driving [3], robotics [4], [5], and helicopter flights [6]. RL is a learning process to achieve the optimal behaviour in a given environment [7]. Multi-agent reinforcement learning (MARL) [8], [9], [10], [11] is a practical scenario of RL in which agents share the common state and cooperate to reach the same goal. Besides the popular domains mentioned above, MARL algorithms have shown their great potential in many areas related to distributed and decentralized settings such as edge network [12] and distributed computing [13].

An essential part of RL is about how to evaluate policies in the learning process. TD learning [14] is a classic efficient approach to address this challenge. Federated learning (FL) [15] is a promising distributed machine learning technique in which agents share the global model which is trained under the coordination of a central agent. This enables the agents' collaboration on an abstracted layer rather than the raw data itself. Motivated by the advantage of keeping multi-agents raw data private in FL, we propose a federated TD(λ) learning framework in which a trusted master agent gathers the agents' parameters and pushes back the aggregated model to agents. We further provide the error bound for the federated TD(λ) learning with a row stochastic weighted matrix under constant and time-varying step-size.

- *Yiming Zeng, Yuanyuan Yang, and Ji Liu are with the Department of Electrical and Computer Engineering, Stony Brook University, Stony Brook, NY 11794 USA. E-mail: {yiming.zeng, yuanyuan.yang, ji.liu}@stonybook.edu.*
- *Yixuan Lin is with the Department of Applied Mathematics and Statistics, Stony Brook University, Stony Brook, NY 11794 USA. E-mail: yixuan.lin.1@stonybook.edu.*

Manuscript received 1 Sept. 2021; revised 30 Oct. 2021; accepted 24 Nov. 2021. Date of publication 10 Dec. 2021; date of current version 23 May 2022.
This work was supported in part by U.S. National Science Foundation under Grants 1513719 and 1730291.
(Corresponding author: Yuanyuan Yang.)
Recommended for acceptance by A.J. Peña, M. Si and J. Zhai.
Digital Object Identifier no. 10.1109/TPDS.2021.3133898

Privacy is another important issue in MARL [16]. In MARL, agents carry out actions and receive rewards independently by observing the global states and the actions, the individual agent learned model or parameters are exchanged frequently between the agent and the master agent for the convergence of the RL model. Even though in the FL setting, agents do not share the raw data in the learning process, the privacy leakage risk still sharply increases because agents' learning models and parameters could be accessed by the adversary in the exchanging process. From the exchanged information, the adversary could deduce the individual private information. In MARL, with accessible environment settings such as states and actions, the reward information of the individual agent is sensitive and private which can reveal the total processing task. Meanwhile, the reward information is vulnerable. We list several examples applied in MARL. In MOBA game [2], players can directly observe their allies and components states information and actions, yet reward information described by value heads (e.g., hurt information) is hidden. In the autonomous driving [3], each car has complete related traffic information in the operation area but keeps its reward/utility information private. Note that in all these situations state information is globally observable, while rewards are purely local. [17] and [18] reveal that reward function is the most vulnerable and valuable component in MARL, which could be deduced by inverse RL techniques.

In this paper, we aim to preserve the privacy of reward information in TD(λ) learning.

Differential privacy (DP) is a widely implemented and effective method to preserve individual privacy being distinguished from a gathered data set which contains multiple individuals' information. The mathematical definition of DP is proposed by Dwork *et al*.[19]. Ever since its introduction, DP has been adopted in RL [20], [21], [22] to preserve privacy in the learning process. To the best of our knowledge, there is no literature considering TD(λ) learning with the privacy-preserving. To preserve the privacy in the TD(λ) learning process, we investigate the Gaussian mechanism proposed in [19] which adds Gaussian perturbation to the

privacy-preserving target. Similar studies have been done in other learning domains, such as deep learning [23], [24], [25], and federated learning [26].

In this paper, we propose a federated TD(λ) learning process to address the policy evaluation problem. Based on this model, we propose a DP algorithm to preserve the privacy of individual agents' reward information.

The contributions of this paper are as follows:

- We propose a federated TD(λ) algorithm that enables agents to keep their raw data private.
- Based on the federated TD(λ) learning, we provide the asymptotic performance and the finite-time analysis for both constant and time-varying step-size.
- We propose a differentially private algorithm to preserve agents' privacy in the federated TD(λ) learning process and provide a rigorous differential privacy guarantee of our algorithm. Notably, we make use of moment accountants to derive a tight privacy bound that can offer a specific choice of the Gaussian noise variance σ^2.
- We gain insight by deriving the upper bound of the federated TD(λ) learning utility loss.
- Evaluations are conducted to corroborate the efficiency of our proposed algorithms.

The organization of the rest of the paper is as follows: we first introduce the related work in Section 2. Section 3 presents the problem formulation. In Section 4, we propose a federated TD(λ) algorithm with the asymptotic performance and the finite-time analysis for both constant and time-varying step-sizes. Then, we enhance the federated TD (λ) algorithm with differential privacy to preserve agents' reward information and provide analyses about the privacy guarantee and utility loss in Section 5. Extensive evaluations are conducted in Section 6, followed by the conclusion in Section 7.

2 Related Works

An overview of recent achievements of MARL is summarized in [27]. In general, MARL problems are investigated in settings that are either collaborative, competitive or a mixture of the two. For collaborative MARL, the canonical multi-agent Markov decision process (MDP) [28], [29] and the team Markov game model [30], [31] are two rudimentary frameworks, where the agents share a common reward function. Then, these two frameworks were extended to the setting where agents are allowed to have heterogeneous reward functions [8], [9], [10], [11], [32], collaborating with the goal of maximizing the long-term return corresponding to the team averaged reward. In particular, these works focused on a decentralized setting, where there exists no central controller to coordinate the agents to achieve the overall team goal. Different from this setting, [33], [34] consider the distributed setting, where the agents can only communicate with a central controller. There is also an ever-growing number of works on MARL in competitive and mixed settings [34], [35], [36], [37]. The temporal difference (TD) learning as a key tool used for policy evaluation RL algorithms, is proposed in [14]. Then, [38], [39], [40] analyze the convergence of TD(λ) under sets of assumptions. In addition, [11], [41], [42] consider the policy evaluation problem with multiple agents under a common environment. In this paper, we will focus on the multi-agent TD algorithm under the distributed setting.

The finite-time analysis of the TD(λ) method with linear function approximation for the single agent is studied in [43], [43], [45], [46], where [46] provides a finite-time error bound for TD(λ) algorithm with linear stochastic approximation and Markov noise under the constant and time-varying step-size. Then, [11], [42] extended [46]'s result to multi-agent TD(λ) algorithm under the decentralized setting, and they proposed a general finite-time analysis under constant and time-varying step-size with a doubly stochastic weighted matrix, where they required to use the property of the second largest singular value of doubly stochastic matrix in their analysis. Different from the setting in [42], in this paper, we will propose the finite-time analysis for distributed TD(λ) with the fixed row stochastic weight matrix under constant and time-varying step-size.

There is a recent line of research that addresses privacy-preserving approaches on RL [21], [22]. [21] aims to protect the value function approximator in Q-learning by adding functional noise in the infinite continuous state space. [22] considers a distributed asynchronous actor-critic (A3C) based model and proposes a locally differentially private algorithm to protect individual agent models by adding the Laplace noise on gradients. Different from them, we add noise to shared parameters from agents to preserve the privacy of agent reward information instead of the reward function or the gradient. We make use of moment accountants to analyze the privacy loss which has been applied in many studies [23], [47], [48], [49]. [47] presents the provable lower and upper bound for r-fold approximate DP. [48] defines the central limit in DP by thoroughly analyzing the privacy loss in varying DP definitions. [49] provides a tight bound on the Renyi DP for the sub-sampling problem. The most related work to ours in the methodology aspect is [23]. [23] considers the differentially private stochastic gradient descent by adding the Gaussian noise to the gradient. Note that the privacy guarantee in [23] does not fit the condition without subsampling, and the variance of Gaussian noise is hard to be determined to preserve the privacy guarantee. However, in this paper, we drive the specific bound of Gaussian noise which is more practical to be applied in the real algorithm.

3 Problem Formulation

Consider a team of $N+1$ agents consisting of one master agent, denoted by 0, and N worker agents, denoted by $\mathcal{N} = \{1, 2, \ldots, N\}$, operating in a common environment. Each worker agent can exchange information only with the master agent.

At each step $t \geq 0$, each agent i can observe the state of the environment $s_t \in \mathcal{S}$, and take action $a_t^i = \mu^i(s_t) \in \mathcal{A}^i$. Given the action $a_t = [a_t^1 \times \cdots \times a_t^N]^\top$ and state s_t, the system will pick the state s_{t+1} for $t+1$ following the probability $P_s : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$, and agents receive a corresponding

cost $R^i(s_t, s_{t+1})$, where $\mathcal{S}$ is the global finite set of states including S states, $\mathcal{A}^i$ is the set of control actions for agent i, $\mu^i : \mathcal{S} \to \mathcal{A}^i$ is a function mapping the state to a control action is $\mathcal{A}^i$, $\mathcal{A} = \mathcal{A}^1 \times \cdots \times \mathcal{A}^N$ is the joint action space of all agents, and R^i is agent i's local cost function. Moreover, we assume that the states and the joint actions are globally observable whereas the cost functions are observed only locally.

The network can be characterized by a discount cost MDP: $(\mathcal{S}, \{\mathcal{A}^i\}, P, \{R^i\}, \gamma)$ for $i \in \mathcal{N}$. Here, $P = [p_{ij}]_{S\times S}$ is the transition probability matrix, i.e., p_{ij} is the probability from the state i to state j, and $\gamma \in (0, 1)$ is the discount factor.

The objective of the agents is to cooperatively estimate a fixed convex combination of the global discounted accumulative cost J, which is defined for all $s \in \mathcal{S}$ as

$$J(s) = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t \sum_{i\in\mathcal{V}} c^i R^i(s_t, s_{t+1}) | s_0 = s\right],$$

where $c^i > 0$ for all $i \in \mathcal{N}$ and $\sum_{i=1}^N c^i = 1$. When the number of states is very large, it is hard to track the computation of J. To mitigate this, as the setting in [42], we use a low-dimensional linear function $\tilde{J}$ to approximate J, where $\tilde{J}$ is parameterized by $\theta \in \mathbb{R}^K$

$$\tilde{J}(s, \theta) = \sum_{k=1}^{K} \theta_k \phi_k(s),$$

for a given set of K vectors $\phi_k \in \mathbb{R}^S, k = 1, \ldots, K$. Assume that $K \ll S$. Let $\Phi = [\phi_1 \quad \cdots \quad \phi_K]$ and $\phi(i)$ be the feature vector, where $(\phi(i))^\top$ is the ith row of Φ, for $i = 1, \ldots |\mathcal{S}|$. Thus, $\tilde{J} = \Phi\theta$. Then the goal is to find the optimal θ^* such that the distance between $\tilde{J}$ and J is minimized. Next, we will use the TD(λ) method to find the value of θ^*.

In this paper, our goal is two-fold. First, we want to keep the privacy of agents' raw data in the distributed TD(λ) learning setting with the asymptotic performance and finite-time analysis. Second, we aim to keep privacy in the entire learning process by protecting the agents' reward information with the privacy guarantee and the expected utility loss performance analysis.

4 Federated TD(λ)

FL keeps the privacy of agents' raw data by sharing learning parameters in the learning process. We also aim to track the convex combination of the global discounted accumulative cost in TD(λ) learning. Motivated by those, we present a federated TD(λ) algorithm using the row stochastic weighted matrix in this section. The detail is summarized in Algorithm 1.

The most related work to us is [42] in which a decentralized TD(λ) algorithm with a doubly stochastic matrix is proposed to track the average of the global discounted accumulative cost. [42] requires bidirectional communication between each pair of neighboring agents. However, this requirement restricts the applications of the algorithm in scenarios with possibly uni-directional communication. To address this, we propose a Federated TD (λ) algorithm with distributed setting using the row stochastic matrix to track a more general case considering a convex combination of the global discounted accumulative cost.

Next, we illustrate the algorithm step by step. At every iteration $t + 1 \geq 0$, each agent shares its own estimated θ_t^i to the master agent, meanwhile it will receives a cost r_t^i, which is a random variable with expected value $R^i(s_t, s_{t+1})$. After doing the convex combination, i.e., $\theta_{t+1}^0 = \sum_{j=1}^N c^j \theta_t^j$, where θ_{t+1}^0 is the value of master agent, master agent will share θ_{t+1}^0 back to all slave agents. Then each agent will update it's own estimate: $\theta_{t+1}^i = \theta_{t+1}^0 + \alpha_t z_t^i l_t^i$, where α_t is the step-size, $l_t^i = (\gamma\phi(s_{t+1}) - \phi(s_t))^\top \theta_t^i + r_t^i$, is the local temporal difference, and r_t^i is the cost for agent i at time t.

Algorithm 1. Federated TD(λ) Algorithm

Input: $N + 1$ agents, the initial environment state and the cost function $R^i(\cdot, \cdot), \forall i \in \mathcal{N}$

Parameters: discount factor γ, step-size α_t;
Initialize: the initial state s_0, each agent $i \in \mathcal{N}$ arbitrarily initializes θ_0^i, and $z_0^i = 0$;
repeat
  for *each iteration* $t + 1$ **do**
    // first, observe a tuple (s_t, s_{t+1}, r_t^i), and the master agent 0 gathers and updates θ_t^0

$$\theta_{t+1}^0 = \sum_{j=1}^{N} c^j \theta_t^j,$$

    // then, master agent broadcasts θ_t^0 to agent i for the local update

$$\theta_{t+1}^i = \theta_{t+1}^0 + \alpha_t z_t^i l_t^i,$$
$$l_t^i = (\gamma\phi(s_{t+1}) - \phi(s_t))^\top \theta_t^i + r_t^i,$$
$$z_{t+1}^i = \gamma\lambda z_t^i + \phi(s_{t+1}).$$

  end
until *convergence*

Note that all agents will get the same $z_t^i = z_t$ at iteration t, where $z_t = \sum_{k=0}^t (\gamma\lambda)^{t-k}\phi(s_k)$. Let $X_t = (s_t, s_{t+1}, z_t)$ be a Markov chain,

$$A(X_t) = z_t(\gamma\phi(s_{t+1}) - \phi(s_t))^\top, \quad b^i(X_t) = r_t^i z_t, \tag{1}$$

where r_t^i is the immediate cost of agent i at time t. Then we can rewrite the update of θ_t^i as $\theta_{t+1}^i = \theta_{t+1}^0 + \alpha_t(A(X_t)\theta_t^i + b^i(X_t))$.

Combined with the update of θ_t^0, we have

$$\theta_{t+1}^i = \sum_{j=1}^{N} c^j \theta_t^j + \alpha_t(A(X_t)\theta_t^i + b^i(X_t)) \tag{2}$$

Let $\pi = [\pi(1), \ldots, \pi(S)]$ be the unique stationary distribution associated with the transition matrix P, $D \in \mathbb{R}^{S\times S}$ be the diagonal matrix with $D = diag(\pi)$, and

$$A = \Phi^\top D(U-I)\Phi, \quad U = (1-\lambda)\sum_{t=0}^{\infty}\lambda^t(\gamma P)^{t+1}, \quad (3)$$
$$b = \sum_{i=1}^{N} d_i b^i, \qquad b^i = \Phi^\top D\sum_{t=0}^{\infty}(\gamma\lambda P)^t r_i,$$

where $r_i \in \mathbb{R}^S, \forall i = 1,\ldots,N$ and the kth entry is $r_i^k = \sum_{s=1}^{S} p_{ks}R^v(k,s), \forall 1 \le k \le S$.

Next, we impose the following standard assumptions, which are made in [42], [46].

Assumption 1. *All the costs are uniformly bounded, i.e., there exists a constant R, such that $|R^i(s,s')| \le R$ for all $i \in \mathcal{N}$ and $s, s' \in \mathcal{S}$.*

Assumption 2. *The basis vectors $\{\phi_k\}, \forall k = 1,\ldots,K$ are linearly independent, i.e Φ has full column rank. Meanwhile, we assume all feature vectors are uniformly bounded, i.e., $\|\phi(s)\|_2 \le \sqrt{\frac{1-\gamma\lambda}{1+\gamma}} \le 1$.*

Assumption 3. *The Markov chain associated with P is irreducible and aperiodic.*

Assumption 4. *The step-size α_t is positive, non-increasing, and satisfies $\sum_{t=0}^{\infty}\alpha_t = \infty$ and $\sum_{t=0}^{\infty}\alpha_t^2 < \infty$.*

Assumption 5. *There exist $T_1 > 0$ and $\Psi_6 > 0$, such that for any $t \ge T_1$, $t - \tau(\alpha_t) \ge 0$ and $\frac{\alpha_m}{\alpha_t} \le \Psi_6$, where $m = t - \tau(\alpha_t)$.*

4.1 Asymptotic Analysis

Under Assumptions 1, 2, and 3, we have $\lim_{t\to\infty}\mathbb{E}[A(X_t)] = A$, and $\lim_{t\to\infty}\mathbb{E}[B(X_t)] = [b^1,\ldots,b^N]^\top$. Moreover, from [38] we know A is a negative definite matrix/stable matrix, i.e., $x^\top Ax < 0$, for all $x \in \mathbb{R}^{K\times K}$. Then based on the Theorem 3 of [50], we can get the following theorem.

Theorem 1. *Under Assumptions 1, 2, 3, and 4, let $\{\theta_t^i\}$ be generated by Eq. (2) for all $i \in \mathcal{N}$, and θ^* be the equilibrium point of the ODE $\dot{\theta} = A\theta + \sum_{i=1}^{N} c^i b^i$, where A and b are defined in Eq. (3). Then, for any agent i, θ_t^i will converge to θ^* in mean square.*

Theorem 1 is an immediate consequence of item (2) of Theorem 2 and its proof will be discussed later.

4.2 Finite-Time Analysis

Given a positive constant α, we use $\tau(\alpha)$ to denote the mixing time of the Markov chain $\{X_t\}$ for which

$$\begin{cases} \|\mathbb{E}[A(X_t) - A|X_0 = X]\| \le \alpha, & \forall X,\ \forall t \ge \tau(\alpha), \\ \|\mathbb{E}[\langle B(X_t)\rangle - b|X_0 = X]\| \le \alpha, & \forall X,\ \forall t \ge \tau(\alpha), \end{cases}$$

where A and b are defined in Eq. (3). In addition, under Assumption 3 the Markov chain $\{X_t\}$ has a geometric mixing time [51], i.e., there exists a constant H such that given a small constant α we have

$$\tau(\alpha) = -H\log\alpha.$$

Then, we can get the following finite-time bound.

Theorem 2. *Let β and Ψ_1–Ψ_5 be the constants defined as*

$$e_1 = 1 + \left(\frac{1+\lambda}{1-\gamma\lambda}\right)^2 + \frac{2R\sqrt{NK}}{1-\gamma\lambda},$$
$$e_2 = \frac{1+\lambda}{(1-\gamma\lambda)^2}(1+\lambda+2R\sqrt{NK}),$$
$$\beta = 1 - c_{\min}^2/2 + \alpha^2 e_2 + \alpha e_1,$$
$$\Psi_1 = \left(\sqrt{e_1^2 + 2c_{\min}^2 e_2} - e_1\right)/(2e_2),$$
$$\Psi_2 = 4\alpha^2 KN\left(\frac{R}{1-\gamma\lambda}\right)^2 + 2\alpha\frac{R\sqrt{NK}}{1-\gamma\lambda}\left(1+\alpha\frac{1+\lambda}{1-\gamma\lambda}\right),$$
$$\Psi_3 = 62\sigma_{\max}\left(1+\frac{R}{1-\gamma\lambda}\right),$$
$$\Psi_4 = 55\sigma_{\max}\left(1+\frac{R}{1-\gamma\lambda}\right)^3,$$
$$\Psi_5 = 4\alpha_0 KN\left(\frac{R}{1-\gamma\lambda}\right)^2 + 2\frac{R\sqrt{NK}}{1-\gamma\lambda}\left(1+\alpha_0\frac{1+\lambda}{1-\gamma\lambda}\right).$$

(1) When the step-size is fixed, i.e., $\alpha_t = \alpha$ for all t. Suppose the Assumptions 1, 2, and 3 hold. When α satisfies

$$0 < \alpha < \min\left\{\frac{1}{4\tau(\alpha)}, \frac{0.05}{\Psi_3\tau(\alpha)+\sigma_{\min}}, \Psi_1\right\},$$

for all $t \ge \tau(\alpha)$ we have

$$\sum_{i=1}^{N} c^i\mathbb{E}[\|\theta_t^i - \theta^*\|_2^2] \le 2\beta^t\sum_{i=1}^{N} c^i\|\theta_0^i - \langle\Theta_0\rangle^\top\|_2^2 + \frac{2\Psi_2}{1-\beta} + \frac{2\Psi_4\sigma_{\max}}{0.9\sigma_{\min}}\alpha\tau(\alpha) + 2\frac{\sigma_{\max}}{\sigma_{\min}}\left(1-\frac{0.9\alpha}{\sigma_{\max}}\right)^{t-\tau(\alpha)} \cdot \left(1.5\|\langle\Theta_0\rangle^\top - \theta^*\|_2 + \frac{0.5R}{1-\gamma\lambda}\right)^2.$$

(2) When the step-size is time-varying. Suppose that Assumptions 1, 2, 3, 4, and 5 hold. Define T^* to be the smallest integer such that $T^* \ge T_1$, $T^*\alpha_0 \le \frac{1}{4}$, $\Psi_3\Psi_6\alpha_{T^*}\tau(\alpha_{T^*}) + \sigma_{\max}\alpha_{T^*} \le 0.05$ and $\alpha_{T^*} \le \Psi_1$. Then, for all $t \ge T^*$ we have

$$\sum_{i=1}^{N} c^i\mathbb{E}[\|\theta_t^i - \theta^*\|_2^2] \le 2\beta^{t-T^*}\sum_{i=1}^{N} c^i\|\theta_{T^*}^i - \langle\Theta_{T^*}\rangle^\top\|_2^2 + \frac{\Psi_5}{1-\beta}(\alpha_0\beta^{\frac{t-1}{2}} + \alpha_{\lceil\frac{t-1}{2}\rceil}) + 2\frac{\sigma_{\max}}{\sigma_{\min}}\left(1.5\|\langle\Theta_0\rangle^\top - \theta^*\|_2 + \frac{0.5R}{1-\gamma\lambda}\right)^2\left(\Pi_{k=T^*}^{t-1}a_k\right) + 4\left(\Psi_4\Psi_6 + \sigma_{\max}\left(\frac{R}{1-\gamma\lambda}\right)^2\right)\sum_{k=T^*}^{t-1} d_k\left(\Pi_{l=k+1}^{t-1}a_l\right),$$

where $a_t = 1 - \frac{0.9\alpha_t}{\sigma_{\max}}$ and $d_t = \alpha_t^2\tau(\alpha_t)$.

Remark 1. Since $0 < \beta < 1$ and $0 < 1 - \frac{0.9\alpha}{\sigma_{\max}} < 1$, then for the fixed step-size case, the summands in the above finite-time bound are exponentially decaying expect for the two constants, which implies

$$\limsup_{t\to\infty}\sum_{i=1}^{N} c^i\mathbb{E}[\|\theta_t^i-\theta^*\|_2^2]\le\frac{2\Psi_2}{1-\beta}+\frac{2\Psi_4\sigma_{\max}}{0.9\sigma_{\min}}\alpha\tau(\alpha).$$

It provides a constant limiting bound. Similarly, for the time-varying step-size case, since $\lim_{t\to\infty}\alpha_t=0$, we have $\limsup_{t\to\infty}\sum_{i=1}^{N} c^i\mathbb{E}[\|\theta_t^i-\theta^*\|_2^2]=0$.

Let

$$\Theta_t=\begin{bmatrix}(\theta_t^1)^\top\\ \cdots\\ (\theta_t^N)^\top\end{bmatrix},\quad W=\begin{bmatrix}c^1,\ldots,c^N\\ \cdots\\ c^1,\ldots,c^N\end{bmatrix},$$

$\langle\Theta_t\rangle=c^\top\Theta_t$, and $Y_t=\Theta_t-\mathbf{1}_N\langle\Theta_t\rangle=(I-\mathbf{1}_Nc^\top)\Theta_t$, where $\mathbf{1}_N$ denotes the N vectors with all entries equal to 1, and $c=[c^1,\ldots,c^N]^\top$. Then, we have the update for Θ_t and $\langle\Theta_t\rangle$ that

$$\begin{aligned}\Theta_{t+1}&=W\Theta_t+\alpha_t(\Theta_tA^\top(X_t)+B(X_t))\\ \langle\Theta_{t+1}\rangle&=\langle\Theta_t\rangle+\alpha_t(\langle\Theta_t\rangle A^\top(X_t)+\langle B(X_t)\rangle).\end{aligned}$$

Note that we have $W\Theta_t-\mathbf{1}_N\langle\Theta_t\rangle=W(\Theta_t-\mathbf{1}_N\langle\Theta_t\rangle)$, and we have the update for Y_t that

$$\begin{aligned}Y_{t+1}&=\Theta_{t+1}-\mathbf{1}_N\langle\Theta_{t+1}\rangle\\ &=WY_t+\alpha Y_tA^\top(X_t)+\alpha(I-\mathbf{1}_Nc^\top)B(X_t).\end{aligned}$$

Moreover, let $(Y_{t+1}^i)^\top$ be the ith row of matrix Y_{t+1}, and we have

$$Y_{t+1}^i=\sum_{j=1}^{N}c^jY_t^j+\alpha A(X_t)Y_t^i+\alpha(b^i(X_t)-B^\top(X_t)c).$$

To prove Theorem 2, we need the following lemmas.

Lemma 1. *Let $c_{\min}$ be the minimum entry of the vector c. For $t\ge 0$, we have*

$$\sum_{j=1}^{N}\sum_{l=1}^{N}c^jc^l\|Y_t^j-Y_t^l\|_2^2\ge c_{\min}^2\sum_{i=1}^{N}c^i\|Y_t^i\|_2^2.$$

Proof. Note that

$$\sum_{j=1}^{N}\sum_{l=1}^{N}c^jc^l\|Y_t^j-Y_t^l\|_2^2=\sum_{k=1}^{K}\sum_{j=1}^{N}\sum_{l=1}^{N}c^jc^l(Y_t^{jk}-Y_t^{lk})^2,$$

where Y_t^{jk} is the kth entry of vector Y_t^j. Let j_k^* and l_k^* be the agents that

$$\begin{aligned}&\max\nolimits_{1\le j,l\le N}|Y_t^{jk}-Y_t^{lk}|=|Y_t^{j_k^*k}-Y_t^{l_k^*k}|\\ &=\max\nolimits_{1\le j,l\le N}|\theta_t^{jk}-\theta_t^{lk}|=|\theta_t^{j_k^*k}-\theta_t^{l_k^*k}|.\end{aligned}$$

Then we have

$$\begin{aligned}&\sum_{j=1}^{N}\sum_{l=1}^{N}c^jc^l\|Y_t^j-Y_t^l\|_2^2\ge c_{\min}^2\sum_{k=1}^{K}(Y_t^{j_k^*k}-Y_t^{l_k^*k})^2\\ &=c_{\min}^2\sum_{k=1}^{K}(\theta_t^{j_k^*k}-\theta_t^{l_k^*k})^2=c_{\min}^2\sum_{k=1}^{K}\max\nolimits_{1\le j,l\le N}(\theta_t^{jk}-\theta_t^{lk})^2.\quad(4)\end{aligned}$$

where θ_t^{jk} is the kth entry of vector θ_t^j. Let $\Theta_t^{\cdot k}$ be the kth column of matrix Θ_t. Since $2x_1x_2\le x_1^2+x_2^2$. Then for any entry $k=1,\ldots,K$, we have

$$\begin{aligned}\sum_{i=1}^{N}c^i\|\theta_t^{ik}-c^\top\Theta_t^{\cdot k}\|_2^2&\le\max\nolimits_{1\le i\le N}\left[\theta_t^{ik}-c^\top\Theta_t^{\cdot k}\right]^2\\ &=\max\nolimits_{1\le i\le N}\left[\sum_{j=1}^{N}c^j(\theta_t^{ik}-\theta_t^{jk})\right]^2\\ &\le\max\nolimits_{1\le i\le N}\max\nolimits_{1\le j\le N}(\theta_t^{ik}-\theta_t^{jk})^2.\end{aligned}$$

Then, based on Eq. (4), we have

$$\begin{aligned}\sum_{j=1}^{N}\sum_{l=1}^{N}c^jc^l\|Y_t^j-Y_t^l\|_2^2&\ge c_{\min}^2\sum_{k=1}^{K}\max\nolimits_{1\le j,l\le N}(\theta_t^{jk}-\theta_t^{lk})^2\\ &\ge c_{\min}^2\sum_{i=1}^{N}\sum_{k=1}^{K}c^i(\theta_t^{ik}-c^\top\Theta_t^{\cdot k})^2\\ &=c_{\min}^2\sum_{i=1}^{N}c^i\|Y_t^i\|_2^2.\end{aligned}$$

This completes the proof. □

Lemma 2. *Under Assumption 2, for the fixed step-size, when $\alpha\in(0,\Psi_1)$, for any $t\ge\tau(\alpha)$ we have*

$$\sum_{i=1}^{N}c^i\|\theta_t^i-\langle\Theta_t\rangle\|_2^2\le\beta^t\sum_{i=1}^{N}c^i\|\theta_0^i-\langle\Theta_0\rangle\|_2^2+\frac{\Psi_2}{1-\beta},$$

where β,Ψ_1 and Ψ_2 are defined in Theorem 2.

Proof. Define the weighted norm $\|X\|_C^2=\sum_{i=1}^{N}c^i\|X_i\|_2^2$, where $C=diag(c)$, and X_i is the ith row of matrix $X\in\mathbb{R}^{N\times K}$. Then we have

$$\|Y_{t+1}\|_C^2=\sum_{i=1}^{N}c^i\sum_{j=1}^{N}\sum_{l=1}^{N}c^jc^l(Y_t^j)^\top Y_t^l\quad(5)$$

$$+\alpha^2\sum_{i=1}^{N}c^i\|A(X_t)Y_t^i\|_2^2$$
$$+\alpha^2\sum_{i=1}^{N}c^i\|b^i(X_t)-B^\top(X_t)c\|_2^2\quad(6)$$
$$+2\alpha\sum_{i=1}^{N}c^i\sum_{j=1}^{N}c^j(Y_t^j)^\top A(X_t)Y_t^i\quad(7)$$

$$+2\alpha\sum_{i=1}^{N}c^i\sum_{j=1}^{N}c^j(b^i(X_t)-B^\top(X_t)c)^\top Y_t^j$$
$$+2\sum_{i=1}^{N}c^i\alpha^2(b^i(X_t)-B^\top(X_t)c)^\top A(X_t)Y_t^i.\quad(8)$$

For Eq. (5), since $2(x_1)^\top x_2=\|x_1\|_2^2+\|x_2\|_2^2-\|x_1-x_2\|_2^2$ for any two vectors x_1 and x_2,

$$\begin{aligned}&\sum_{i=1}^{N}c^i\sum_{j=1}^{N}\sum_{l=1}^{N}c^jc^l(Y_t^j)^\top Y_t^l\\ &=\sum_{i=1}^{N}c^i\|Y_t^i\|_2^2-\frac{1}{2}c^i\sum_{j=1}^{N}\sum_{l=1}^{N}c^jc^l\|Y_t^j-Y_t^l\|_2^2.\end{aligned}$$

Applying the result from Lemma 1, we have

$$\sum_{j=1}^{N}\sum_{l=1}^{N} c^j c^l (Y_t^j)^\top Y_t^l \le \left(1-\frac{c_{\min}^2}{2}\right)\sum_{i=1}^{N} c^i \|Y_t^i\|_2^2. \tag{9}$$

As for Eq. (6), we have that

$$\alpha^2 \sum_{i=1}^{N} c^i \|A(X_t)Y_t^i\|_2^2 \le \alpha^2 \left(\frac{1+\gamma}{1-\gamma\lambda}\right)^2 \sum_{i=1}^{N} c^i \|Y_t^i\|_2^2,$$

and

$$\begin{aligned}
&\alpha^2 \sum_{i=1}^{N} c^i \|b^i(X_t) - B^\top(X_t)c\|_2^2 \\
&\le \alpha^2 \sum_{i=1}^{N} c^i \|(2\cdot\|B(X_t)\|_{\max})\mathbf{1}_K\|_2^2 = 4\alpha^2\|B(X_t)\|_{\max}^2 K \\
&\le 4\alpha^2\|B(X_t)\|_F^2 K = 4\alpha^2 K \sum_{i=1}^{N}\sum_{k=1}^{K}(b^{ik}(X_t))_2^2 \\
&\le 4\alpha^2 KN\left(\frac{R}{1-\gamma\lambda}\right)^2.
\end{aligned}$$

Then,

$$\begin{aligned}
&\alpha^2 \sum_{i=1}^{N} c^i \|A(X_t)Y_t^i\|_2^2 + \alpha^2 \sum_{i=1}^{N} c^i \|b^i(X_t) - B^\top(X_t)c\|_2^2 \\
&\le \alpha^2 \left(\frac{1+\gamma}{1-\gamma\lambda}\right)^2 \sum_{i=1}^{N} c^i \|Y_t^i\|_2^2 + 4\alpha^2 KN\left(\frac{R}{1-\gamma\lambda}\right)^2.
\end{aligned} \tag{10}$$

As for Eq. (7), we have

$$\begin{aligned}
&2\alpha \sum_{i=1}^{N} c^i \sum_{j=1}^{N} c^j (Y_t^j)^\top A(X_t) Y_t^i \\
&\le \alpha \sum_{i=1}^{N} c^i \sum_{j=1}^{N} c^j \left[\|Y_t^j\|_2^2 + \|A(X_t)Y_t^i\|_2^2\right] \\
&\le \alpha\left(1+\left(\frac{1+\gamma}{1-\gamma\lambda}\right)^2\right)\sum_{i=1}^{N} c^i \|Y_t^i\|_2^2.
\end{aligned} \tag{11}$$

As for Eq. (8), we get

$$\begin{aligned}
&2\alpha \sum_{i=1}^{N} c^i \sum_{j=1}^{N} c^j (b^i(X_t) - B^\top(X_t)c)^\top Y_t^j \\
&\quad + 2\sum_{i=1}^{N} c^i \alpha^2 (b^i(X_t) - B^\top(X_t)c)^\top A(X_t) Y_t^i \\
&\le 2\alpha \|B(X_t) - \mathbf{1}c^\top B(X_t)\|_{\max} \cdot \sum_{i=1}^{N}\left[c^i \mathbf{1}^\top + \alpha c^i \mathbf{1}^\top A(X_t)\right] Y_t^i.
\end{aligned}$$

Note that $\|B(X_t) - \mathbf{1}d^\top B(X_t)\|_{\max} \le 2\|B(X_t)\|_F \le 2\frac{R\sqrt{N}}{1-\gamma\lambda}$, and

$$\begin{aligned}
&\sum_{i=1}^{N}\left[c^i \mathbf{1}^\top + \alpha c^i \mathbf{1}^\top A(X_t)\right] Y_t^i \\
&\quad \le \sum_{i=1}^{N}\left[c^i\sqrt{K}\|Y_t^i\|_2 + \alpha c^i \sqrt{K}\|A(X_t)\|_2 \|Y_t^i\|_2\right].
\end{aligned}$$

Since for the vector x, we have inequality that $2\|x\|_2 \le 1 + \|x\|_2^2$. Then we have

$$\begin{aligned}
&2\alpha \sum_{i=1}^{N} c^i \sum_{j=1}^{N} c^j (b^i(X_t) - B^\top(X_t)c)^\top Y_t^j \\
&\quad + 2\sum_{i=1}^{N} c^i \alpha^2 (b^i(X_t) - B^\top(X_t)c)^\top A(X_t) Y_t^i \\
&\le 4\alpha \frac{R\sqrt{N}}{1-\gamma\lambda}\sum_{i=1}^{N}\left[c^i\sqrt{K}\|Y_t^i\|_2 + \alpha c^i\sqrt{K}\|A(X_t)\|_2\|Y_t^i\|_2\right] \\
&\le 2\alpha \frac{R\sqrt{NK}}{1-\gamma\lambda}\cdot \\
&\quad \left[\sum_{i=1}^{N} c^i(1+\|Y_t^i\|_2^2) + \alpha\sum_{i=1}^{N} c^i \|A(X_t)\|_2 (1+\|Y_t^i\|_2^2)\right] \\
&\quad = 2\alpha \frac{R\sqrt{NK}}{1-\gamma\lambda}\left(1+\alpha\frac{1+\gamma}{1-\gamma\lambda}\right)\sum_{i=1}^{N} c^i \|Y_t^i\|_2^2 \\
&\quad\quad + 2\alpha \frac{R\sqrt{NK}}{1-\gamma\lambda}\left(1+\alpha\frac{1+\gamma}{1-\gamma\lambda}\right).
\end{aligned} \tag{12}$$

Based on Eqs. (9), (10), (11), and (12), we know that

$$\begin{aligned}
\|Y_{t+1}\|_C^2 \le{}& \left(1-\frac{c_{\min}^2}{2}\right)\sum_{i=1}^{N} c^i \|Y_t^i\|_2^2 \\
&+ \alpha^2\left(\frac{1+\gamma}{1-\gamma\lambda}\right)^2 \sum_{i=1}^{N} c^i \|Y_t^i\|_2^2 \\
&+ 4\alpha^2 KN\left(\frac{R}{1-\gamma\lambda}\right)^2 \\
&+ \alpha\left(1+\left(\frac{1+\gamma}{1-\gamma\lambda}\right)^2\right)\sum_{i=1}^{N} c^i \|Y_t^i\|_2^2 \\
&+ 2\alpha\frac{R\sqrt{NK}}{1-\gamma\lambda}\left(1+\alpha\frac{1+\gamma}{1-\gamma\lambda}\right)\sum_{i=1}^{N} c^i \|Y_t^i\|_2^2 \\
&+ 2\alpha\frac{R\sqrt{NK}}{1-\gamma\lambda}\left(1+\alpha\frac{1+\gamma}{1-\gamma\lambda}\right).
\end{aligned}$$

Since $\alpha \in (0, \Psi_1)$, $0 < \beta < 1$. From the definition of Ψ_2,

$$\|Y_{t+1}\|_C^2 \le \beta\|Y_t\|_C^2 + \Psi_2 \le \beta^{t+1}\|Y_0\|_C^2 + \frac{\Psi_2}{1-\beta},$$

which implies that

$$\sum_{i=1}^{N} c^i \|\theta_t^i - \langle\Theta_t\rangle^\top\|_2^2 \le \beta^t \sum_{i=1}^{N} c^i \|\theta_0^i - \langle\Theta_0\rangle^\top\|_2^2 + \frac{\Psi_2}{1-\beta}.$$

This completes the proof. □

Lemma 3. *Under Assumptions 1, 2, and 3, when the fixed stepsize*

$$0 < \alpha < \min\left\{\frac{1}{4\tau(\alpha)}, \frac{0.05}{\Psi_3\tau(\alpha)+\sigma_{\min}}\right\},$$

for any $t \geq \tau(\alpha)$, we have

$$\begin{aligned}&\mathbb{E}[\|\langle\boldsymbol{\Theta}_t\rangle^\top - \theta^*\|_2^2]\\ &\leq \frac{\sigma_{\max}}{\sigma_{\min}}\left(1-\frac{0.9\alpha}{\sigma_{\max}}\right)^{t-\tau(\alpha)}\left(1.5\|\langle\boldsymbol{\Theta}_0\rangle^\top - \theta^*\|_2 + \frac{0.5R}{1-\gamma\lambda}\right)^2 + \frac{\Psi_4\sigma_{\max}}{0.9\sigma_{\min}}\alpha\tau(\alpha),\end{aligned}$$

where Ψ_3 and Ψ_4 are defined in Theorem 2.

Proof. Since $\lim_{t\to\infty}\mathbb{E}[\langle B(X_t)\rangle] = b$, the corresponding ODE is $\dot{\langle\boldsymbol{\Theta}\rangle} = A\langle\boldsymbol{\Theta}\rangle + b$, where A and b are defined in Eq. (3). From Theorem 1, θ^* is the equilibrium point of this ODE. Let $\hat{b}(X_t) = \langle B(X_t)\rangle^\top + A(X_t)\theta^*$, and $\hat{\theta}_t = \langle\boldsymbol{\Theta}_t\rangle^\top - \theta^*$, then $\hat{\theta}_{t+1} = \hat{\theta}_t + \alpha(A(X_t)\hat{\theta} + \hat{b}(X_t))$. By using Theorem 7 in [46], we have

$$\begin{aligned}&\mathbb{E}[\|\hat{\theta}\|_2^2] = \mathbb{E}[\|\langle\boldsymbol{\Theta}_t\rangle^\top - \theta^*\|_2^2]\\ &\leq \frac{\sigma_{\max}}{\sigma_{\min}}\left(1-\frac{0.9\alpha}{\sigma_{\max}}\right)^{t-\tau(\alpha)}\left(1.5\|\langle\boldsymbol{\Theta}_0\rangle^\top - \theta^*\|_2 + \frac{0.5R}{1-\gamma\lambda}\right)^2 + \frac{\Psi_4\sigma_{\max}}{0.9\sigma_{\min}}\alpha\tau(\alpha).\end{aligned}$$

This completes the proof. □

Lemma 4. *Under Assumptions 1, 2 and 4. For the time-varying step-size, if there exists a constant time T^*, for all $t \geq T^*$, $0 < \alpha_t < \Psi_1$. Then for any $t \geq T^*$ we have*

$$\begin{aligned}\sum_{i=1}^N c^i\|\theta_t^i - \langle\boldsymbol{\Theta}_t\rangle\|_2^2 \quad &\leq \beta^{t-T^*}\sum_{i=1}^N c^i\|\theta_{T^*}^i - \langle\boldsymbol{\Theta}_{T^*}\rangle\|_2^2\\ &+ \frac{\Psi_5}{1-\beta}\left(\alpha_0\beta^{\frac{t-1}{2}} + \alpha_{\lceil\frac{t-1}{2}\rceil}\right),\end{aligned}$$

where β, Ψ_1 and Ψ_2 are defined in Theorem 2.

Proof. Following the proof of Lemma 2, we have

$$\begin{aligned}\|Y_{t+1}\|_C^2 &\leq \left(1 - \frac{c_{\min}^2}{2} + \alpha_t^2\left(\frac{1+\lambda}{1-\gamma\lambda}\right)^2 + \alpha_t\left(1 + \left(\frac{1+\lambda}{1-\gamma\lambda}\right)^2\right) + 2\alpha_t\frac{R\sqrt{NK}}{1-\gamma\lambda}\left(1 + \alpha_t\frac{1+\lambda}{1-\gamma\lambda}\right)\right)\|Y_t\|_C^2\\ &+ 4\alpha_t^2 KN\left(\frac{R}{1-\gamma\lambda}\right)^2 + 2\alpha_t\frac{R\sqrt{NK}}{1-\gamma\lambda}\left(1 + \alpha_t\frac{1+\lambda}{1-\gamma\lambda}\right).\end{aligned}$$

Let

$$\begin{aligned}\beta_t &= 1 - \frac{c_{\min}^2}{2} + \alpha_t^2\left(\frac{1+\lambda}{1-\gamma\lambda}\right)^2 + \alpha_t\left(1 + \left(\frac{1+\lambda}{1-\gamma\lambda}\right)^2\right)\\ &+ 2\alpha_t\frac{R\sqrt{NK}}{1-\gamma\lambda}\left(1 + \alpha_t\frac{1+\lambda}{1-\gamma\lambda}\right).\end{aligned}$$

Since for $t \geq T^*$, $0 < \alpha_t < \Psi_1$, then $0 < \beta_t \leq \beta < 1$, where β is defined in Theorem 2. Besides, based on the definition of Ψ_2 in Theorem 2, for $t \geq T^*$ we have

$$\begin{aligned}\|Y_{t+1}\|_C^2 &\leq \beta\|Y_t\|_C^2 + \alpha_t\Psi_5\\ &\leq \beta^{t+1-T^*}\|Y_{T^*}\|_C^2 + \Psi_5\sum_{s=0}^t\beta^{t-s}\alpha_s\\ &\leq \beta^{t+1-T^*}\|Y_{T^*}\|_C^2 + \Psi_5\sum_{s=0}^{\lfloor t/2\rfloor}\beta^{t-s}\alpha_s + \Psi_5\sum_{s=\lceil t/2\rceil}^t\beta^{t-s}\alpha_s\\ &\leq \beta^{t+1-T^*}\|Y_{T^*}\|_C^2 + \frac{\Psi_5}{1-\beta}\left(\alpha_0\beta^{\frac{t}{2}} + \alpha_{\lceil\frac{t}{2}\rceil}\right),\end{aligned}$$

which implies that

$$\begin{aligned}&\sum_{i=1}^N c^i\|\theta_t^i - \langle\boldsymbol{\Theta}_t\rangle^\top\|_2^2 \leq\\ &\beta^{t-T^*}\sum_{i=1}^N c^i\|\theta_{T^*}^i - \langle\boldsymbol{\Theta}_{T^*}\rangle^\top\|_2^2 + \frac{\Psi_5}{1-\beta}\left(\alpha_0\beta^{\frac{t-1}{2}} + \alpha_{\lceil\frac{t-1}{2}\rceil}\right).\end{aligned}$$

This completes the proof. □

Lemma 5. *Under Assumptions 1, 2, 3, 4, and 5, for the time-varying step-size, for any $t \geq T^*$, where T^* is defined in Theorem 2, we have that*

$$\begin{aligned}&\mathbb{E}[\|\langle\boldsymbol{\Theta}_t\rangle^\top - \theta^*\|_2^2]\\ &\leq \frac{\sigma_{\max}}{\sigma_{\min}}\left(1.5\|\langle\boldsymbol{\Theta}_0\rangle^\top - \theta^*\|_2 + \frac{0.5R}{1-\gamma\lambda}\right)^2\left(\Pi_{k=T^*}^{t-1}a_k\right)\\ &+ 2\left(\Psi_4\Psi_6 + \sigma_{\max}\left(\frac{R}{1-\gamma\lambda}\right)^2\right)\sum_{k=T^*}^{t-1}d_k\left(\Pi_{l=k+1}^{t-1}a_l\right),\end{aligned}$$

where a_t, d_t and Ψ_4 are defined in Theorem 2.

Proof. By using Theorem 11 in [46], we have

$$\begin{aligned}&\mathbb{E}[\|\hat{\theta}\|_2^2] = \mathbb{E}[\|\langle\boldsymbol{\Theta}_t\rangle^\top - \theta^*\|_2^2]\\ &\leq \frac{\sigma_{\max}}{\sigma_{\min}}\left(1.5\|\langle\boldsymbol{\Theta}_0\rangle^\top - \theta^*\|_2 + \frac{0.5R}{1-\gamma\lambda}\right)^2\left(\Pi_{k=T^*}^{t-1}a_k\right)\\ &+ 2\left(\Psi_4\Psi_6 + \sigma_{\max}\left(\frac{R}{1-\gamma\lambda}\right)^2\right)\sum_{k=T^*}^{t-1}d_k\left(\Pi_{l=k+1}^{t-1}a_l\right).\end{aligned}$$

This completes the proof. □

We are now in a position to prove Theorem 2.

Proof of Theorem 2. (1) When the step-size is fixed,

$$\begin{aligned}&\sum_{i=1}^N c^i\mathbb{E}[\|\theta_t^i - \theta^*\|_2^2]\\ &\leq 2\sum_{i=1}^N c^i\mathbb{E}[\|\theta_t^i - \langle\boldsymbol{\Theta}_t\rangle^\top\|_2^2] + 2\mathbb{E}[\|\langle\boldsymbol{\Theta}_t\rangle^\top - \theta^*\|_2^2].\end{aligned}$$

By using Lemmas 2 and 3 , for any $t \geq \tau(\alpha)$, we have

$$\begin{aligned}&\sum_{i=1}^N c^i\mathbb{E}[\|\theta_t^i - \theta^*\|_2^2]\\ &\leq 2\beta^t\sum_{i=1}^N c^i\|\theta_0^i - \langle\boldsymbol{\Theta}_0\rangle^\top\|_2^2 + \frac{2\Psi_2}{1-\beta} + \frac{2\Psi_4\sigma_{\max}}{0.9\sigma_{\min}}\alpha\tau(\alpha)\\ &+ 2\frac{\sigma_{\max}}{\sigma_{\min}}\left(1-\frac{0.9\alpha}{\sigma_{\max}}\right)^{t-\tau(\alpha)}\left(1.5\|\langle\boldsymbol{\Theta}_0\rangle^\top - \theta^*\|_2 + \frac{0.5R}{1-\gamma\lambda}\right)^2.\end{aligned}$$

(2) When the step-size is time-varying,

$$\sum_{i=1}^{N} c^i \mathbb{E}[\|\theta_t^i - \theta^*\|_2^2]$$
$$\leq 2\sum_{i=1}^{N} c^i \mathbb{E}[\|\theta_t^i - \langle \Theta_t \rangle^\top \|_2^2] + 2\mathbb{E}[\|\langle \Theta_t \rangle^\top - \theta^*\|_2^2].$$

By using Lemmas 4 and 5 , for any $t \geq T^*$, we have

$$\sum_{i=1}^{N} c^i \mathbb{E}[\|\theta_t^i - \theta^*\|_2^2]$$
$$\leq 2\beta^{t-T^*} \sum_{i=1}^{N} c^i \|\theta_{T^*}^i - \langle \Theta_{T^*} \rangle^\top \|_2^2$$
$$+ \frac{2\Psi_5}{1-\beta}\left(\alpha_0 \beta^{\frac{t-1}{2}} + \alpha_{\lceil \frac{t-1}{2} \rceil}\right)$$
$$+ 2\frac{\sigma_{\max}}{\sigma_{\min}}\left(1.5\|\langle \Theta_0 \rangle^\top - \theta^*\|_2 + \frac{0.5R}{1-\gamma\lambda}\right)^2 \left(\Pi_{k=T^*}^{t-1} a_k\right)$$
$$+ 4\left(\Psi_4\Psi_6 + \sigma_{\max}\left(\frac{R}{1-\gamma\lambda}\right)^2\right) \sum_{k=T^*}^{t-1} d_k \left(\Pi_{l=k+1}^{t-1} a_l\right).$$

This completes the proof. □

5 Differentially Private Federated TD(λ)

In this section, we first illustrate the motivation to protect the reward information of agents in the federated TD(λ) learning and introduce the background of DP. Then, we propose a privacy-preserving algorithm to protect the privacy of individual agents and the entire team and analyze the privacy guarantee and the utility loss of our proposed DP algorithm.

5.1 Motivation and Background

In each iteration t of Algorithm 1, for each agent $i \in \mathcal{N}$, θ_t^i is gathered by the master agent 0. Then, the sum-average value θ_t^0 is shared with all agents for updating. Some other environment information such as the step-size α_t, state s_t, s_{t-1}, and feature vector $\phi(s_t), \phi(s_{t-1})$ are available to be accessed by all agents. From the environment setting above, $A(X_t)$ from Eq. (1) and $z_t = \sum_{k=0}^{t} (\gamma\lambda)^{t-k} \phi(s_k)$ can be deduced. Once θ_t^i is deduced or accessed by the potential attacker, the cost r_i^t for the agent i can be inferred easily from Eq. (2) since all the environment setting parameters are available. The cost function is sensitive and private in the TD-reinforcement learning process which can describe the task sufficiently. The attacker could be the agent in MARL which goal is to attack the total group or some certain agents, or an adversary out of the MARL group and tries to attack the group by accessing the group information.

Our goal is to preserve the privacy of the agent's cost function by protecting the individual θ_t^i from being distinguished by θ_{t+1}^0. We apply differential privacy (DP) [19], a nature and well-implemented privacy mechanism, to protect agents' cost functions. The definition of DP, mechanism sensitivity $\Delta\mathcal{M}$, and Gaussian Mechanism are listed in Definitions 1, 2 and 3 respectively.

Definition 1. *Assume $\mathcal{D}$ and $\hat{\mathcal{D}}$ are neighboring databases which differs by only one row. Let ϵ be a possible real number and δ be a non-negative real number. A randomized mechanism $\mathcal{M}$ is (ϵ, δ)-differential private if for any neighboring databases $\mathcal{D}$ and $\hat{\mathcal{D}}$, and for any $\mathcal{S} \in \mathcal{O}$, where $\mathcal{O}$ is the set of all possible output of mechanism $\mathcal{M}$ and $\mathcal{S}$ is the subset of $\mathcal{O}$.*

$$Pr[\mathcal{M}(\mathcal{D}) \in \mathcal{S}] \leq e^{\epsilon} Pr\big[\mathcal{M}(\hat{\mathcal{D}}) \in \mathcal{S}\big] + \delta.$$

When $\delta = 0$, $\mathcal{M}$ is called ϵ-differential private.

Definition 2. *For all pairs of $\mathcal{D}$, $\hat{\mathcal{D}}$ of neighboring inputs belong to the input set of the mechanism $\mathcal{M}$, the sensitivity of the mechanism $\mathcal{M}$ is defined as:$\Delta\mathcal{M} = \max_{\mathcal{D},\hat{\mathcal{D}}} \|\mathcal{M}(\mathcal{D}) - \mathcal{M}((\hat{\mathcal{D}})\|_2$, where $\|\cdot\|_2$ denotes the l_2 norm.*

Definition 3. *(Gaussian Mechanism) If $0 < \epsilon < 1$, and $\sigma \geq \sqrt{2\ln\frac{1.25}{\delta}}\frac{\Delta\mathcal{M}}{\epsilon}$, then $\mathcal{M}(\mathcal{D}) + g$ is (ϵ, δ)-differentially private, where g is drawn from $\mathcal{N}(0, \sigma^2)$.*

Lemma 6. *(Abadi, 2016[23]). Suppose that a mechanism $\mathcal{M}$ consists of a sequence of adaptive mechanism $\mathcal{M}_1, \ldots, \mathcal{M}_T$ where $\mathcal{M}_t : \Pi_{j=1}^{t-1}\mathbb{R}_j \to \mathbb{R}_t$. Then, for any $\nu > 0$.*

$$\Gamma_{\mathcal{M}}(\nu) \leq \sum_{t=1}^{T} \Gamma_{\mathcal{M}_t}(\nu) = T \cdot \Gamma_{\mathcal{M}_t}(\nu),$$

with

$$\Gamma_{\mathcal{M}_t}(\nu) = \ln \mathbb{E}_{\mathcal{S}\sim\mathcal{M}(\mathcal{D})} \exp\left(\nu \ln \frac{Pr[\mathcal{M}(\mathcal{D}) = \mathcal{S}]}{Pr[\mathcal{M}(\hat{\mathcal{D}}) = \mathcal{S}]}\right)$$
$$= \ln \mathbb{E}_{\mathcal{S}\sim\mathcal{M}(\mathcal{D})}\left[\left(\frac{\mathcal{M}(\mathcal{D}) = \mathcal{S}}{\mathcal{M}(\hat{\mathcal{D}}) = \mathcal{S}}\right)^{\nu}\right],$$

For any $\epsilon > 0$, the mechanism $\mathcal{M}$ is (ϵ, δ)-differential private for

$$\delta \leq exp(\Gamma_{\mathcal{M}}(\nu) - \nu\epsilon).$$

5.2 Differential Privacy Algorithm

We propose a privacy-preserving algorithm based on the setting of Algorithm 1. The details of the algorithm are illustrated in Algorithm 2. We achieve privacy by adding perturbation μ_t^0 to θ_t^0 at each iteration t.

Insight to the Algorithm Design. From Definition 3 proposed in [19], each iteration of Algorithm 2 is (ϵ, δ)-differentially private if we choose σ as $\sqrt{2\log\frac{1.25}{\delta}}\frac{\Delta\mathcal{M}}{\epsilon}$. [23] proposes a $(O(q\epsilon\sqrt{T}), \delta)$-deferentially private algorithm which saves a $\sqrt{\log(\frac{1}{\delta})}$ in ϵ part, and a Tq factor in δ than the standard composition theory (Theorem 3.14 in [19]). However, Theorem 1 in [23] has several limitations. First, it does not fit the condition without subsampling. Second, it does not present the exact guidance to determine the Gaussian noise to satisfy differential privacy (constants c_1 and c_2 in [23] are difficult to determine in the real algorithm operation, because of the loose bound for moment accountant $\Gamma_{\mathcal{M}}(\nu)$). We address these challenges in our paper.

5.3 Privacy and Utility Analysis

Privacy Analysis. There are three main components in the privacy analysis.

- First, we define the mechanism sensitive $\Delta\mathcal{M}$ (presented in Lemma 7). It is not straightforward to calculate the sensitivity directly. Instead, we seek to derive an upper bound of $\Delta\mathcal{M}$ that suffices to provide the differential privacy guarantee.
- Second, we prove the upper bound of the moment accountants $\Gamma_{\mathcal{M}}(\nu)$. This is a tight bound without subsampling.
- Third, we derive the privacy guarantee in Theorem 3 in which the noise variance can be determined in specific.

We estimate the sensitive denoted as $\Delta\mathcal{M}$ of data set Ω=$\{\theta_t^i|\forall i \in \mathcal{N}, t \in \mathcal{T}\}$. $\mathcal{M}$ is the privacy-preserving mechanism shown in Eq. (13). Let Ω' denote the neighboring data set of Ω that differs by only one item. The added perturbation part u_{t+1}^0 follows the Gaussian distribution $\mathbb{N}(0, \sigma^2)$.

Algorithm 2. Differentially Private Federated TD(λ) Algorithm

Input: $N+1$ agents, the initial environment state and the cost function $R^i(\cdot,\cdot), \forall i \in \mathcal{N}$

Parameters: target privacy (ϵ, δ), discount factor γ, step-size α_t, iteration time T;

Initialize: the initial state s_0, each agent $i \in \mathcal{N}$ arbitrarily initializes θ_0^i, and $z_0^i = 0$, $t = 0$;

for $t \in T$ **do**

// first, observe a tuple (s_t, s_{t+1}, r_t^i), and the master agent 0 gathers and updates θ_t^0

$$\theta_{t+1}^0 = \sum_{j=1}^{N} c^j \theta_t^j,$$

// add the perturbation part u_{t+1}^0 to θ_{t+1}^0

$$\hat{\theta}_{t+1}^0 = \theta_{t+1}^0 + u_{t+1}^0, \tag{13}$$

// then, master agent broadcasts $\hat{\theta}_t^0$ to agent i for the local update

$$\theta_{t+1}^i = \hat{\theta}_{t+1}^0 + \alpha_t z_t^i l_t^i,$$
$$l_t^i = (\gamma\boldsymbol{\phi}(s_{t+1}) - \boldsymbol{\phi}(s_t))^\top \theta_t^i + r_t^i,$$
$$z_{t+1}^i = \gamma\lambda z_t^i + \boldsymbol{\phi}(s_{t+1}).$$

end

Lemma 7. $\|\mathcal{M}(\Omega) - \mathcal{M}(\Omega')\|_2 \leq \frac{2R}{1-\lambda\gamma}$.

Proof. Assume r_t^i and $(r_t^i)'$ are two possible rewards for agent i at time t, let θ_{t+1}^i and $(\theta_{t+1}^i)'$ be the corresponding output, i.e.,

$$\theta_{t+1} = \theta_{t+1}^0 + \alpha_t(A(X_t)\theta_t^i + r_t^i z_t)$$
$$(\theta_{t+1}^i)' = \theta_{t+1}^0 + \alpha_t(A(X_t)\theta_t^i + (r_t^i)' z_t).$$

Then, we have

$$\|(\theta_{t+1}^i) - (\theta_{t+1}^i)'\|_2 = \|(r_t^i - (r_t^i)')z_t\|_2$$
$$\leq 2R\|z_t\|_2 = 2R\|\sum_{k=0}^{t}(\lambda\gamma)^{t-k}\boldsymbol{\phi}(s_k)\|_2$$
$$\leq 2R\sum_{k=0}^{t}(\lambda\gamma)^{t-k}\|\boldsymbol{\phi}(s_k)\|_2 \leq \frac{2R}{1-\lambda\gamma}.$$

Let $\theta_{t+1}^i \in \Omega$ and $(\theta_{t+1}^i)' \in \Omega'$ be the different elements in the neighboring dataset Ω and Ω'.

$$\|\mathcal{M}(\Omega) - \mathcal{M}(\Omega')\|_2 = \|c_i \cdot (\theta_{t+1}^i - (\theta_{t+1}^i)')\|_2$$
$$= c_i\|\theta_{t+1}^i - (\theta_{t+1}^i)'\|_2 \leq c_i\frac{2R}{1-\lambda\gamma} \leq \frac{2R}{1-\lambda\gamma}.$$

This completes the proof. □

Lemma 7 denotes the upper bound of the mechanism discrepancy with two neighboring inputs which is also the upper bound of the $\Delta\mathcal{M}$. We let the $\Delta\mathcal{M} = \frac{2R}{1-\lambda\gamma}$.

To keep track of the privacy loss, we apply the moments accountant technique as illustrated in Lemma 6 proposed in [23]. We first calculate the moment $\Gamma_{\mathcal{M}_i}(\nu)$. Then, we derive the upper bound of $\Gamma_{\mathcal{M}}(\nu)$ in Lemma 8 from the results in Lemma 6.

Lemma 8. *For all $\nu > 0, T > 0$ and $k \in [0, \nu+1]$,*

$$\Gamma_{\mathcal{M}}(\nu) \leq T\sum_{k=0}^{\nu+1}\nu + 1k\frac{(k^2 + k\Delta\mathcal{M})}{2\sigma^2}.$$

Proof. Suppose that $g : \mathcal{D} \to \mathcal{R}^p$ with $\|g(\cdot)\| \leq \Delta\mathcal{M}$. Let $\mathcal{M}(\theta) = \sum_{i\in\mathcal{N}} g_i(\theta^i)$. Given fixed θ' and $\theta = \theta' \cup \theta^i$. Without loss of generality, let $\sum_{j\in\mathcal{N}\setminus[i]} g(\theta^i) = \mathbf{0}$. Let μ_0 denote the pdf of $\mathcal{N}(0, \sigma^2)$ and let μ_1 denote the pdf of $\mathcal{N}(\Delta\mathcal{M}, \sigma^2)$. Thus: $\mathcal{M}(\theta) \sim \mathcal{N}(0, 1)$ and $\mathcal{M}(\theta') \sim \mathcal{N}(\Delta\mathcal{M}, 1)$. We want to calculate the value of the $\mathbb{E}_{z\sim\mu_1}\left[\left(\frac{\mu_0(z)}{\mu_1(z)}\right)\right]$.

$$\mathbb{E}_{z\sim\mu_1}\left[\left(\frac{\mu_0(z)}{\mu_1(z)}\right)^{\nu+1}\right]$$
$$= \mathbb{E}_{z\sim\mu_1}\left[\left(1 + \frac{\mu_0(z) - \mu_1(z)}{\mu_1(z)}\right)^{\nu+1}\right]$$
$$= \sum_{k=0}^{\nu+1}\nu + 1k\mathbb{E}_{z\sim\mu_1}\left[\left(\frac{\mu_0(z) - \mu_1(z)}{\mu_1(z)}\right)^{k}\right]$$
$$= \sum_{k=0}^{\nu+1}\nu + 1k\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi}\sigma}\exp{-\frac{(z-\Delta\mathcal{M})^2}{2\sigma^2}}\cdot$$
$$\exp\frac{k\cdot(2z - \Delta\mathcal{M})}{2\sigma^2}dz$$
$$= \sum_{k=0}^{\nu+1}\nu + 1k\int_{-\infty}^{+\infty}\frac{1}{\sqrt{2\pi\sigma}}\cdot$$
$$\exp\frac{-(z-(\Delta\mathcal{M}+k))^2}{2\sigma^2}dz\ \exp\frac{(k^2+k\Delta\mathcal{M})}{2\sigma^2},$$
$$= \sum_{k=0}^{\nu+1}\nu + 1k\ \exp\frac{(k^2+k\Delta\mathcal{M})}{2\sigma^2},$$

then from Lemma 1,

$$\Gamma_{\mathcal{M}}(\nu) \leq \sum_{t=1}^{n}\Gamma_{\mathcal{M}_T}(\nu) = T\cdot\Gamma_{\mathcal{M}_t}(\nu)$$
$$= T\sum_{k=0}^{\nu+1}\nu + 1k\frac{(k^2+k\Delta\mathcal{M})}{2\sigma^2}.$$

This completes the proof. □

Based on Lemmas 6 and 8, our privacy guarantee is introduced in Theorem 3.

Theorem 3. *Given the privacy budget $\epsilon > 0$, $\delta > 0$ and the iteration time T, when $\epsilon \leq \frac{2}{\nu}\ln(\frac{1}{\delta})$ Algorithm 2 is (ϵ, δ)-deferentially private for all integers $\nu > 0$ if we choose*

$$\sigma^2 \geq \frac{1}{\nu\epsilon} T \sum_{k=0}^{\nu+1} \nu + 1k(k^2 + k\Delta\mathcal{M}).$$

Remark 2. Theorem 3 provides a rigorous guarantee on the privacy of the reward function. The bound in Theorem 3 might appear complicated, this is partly because this bound is tighter and more precise than the non-asymptotic bound, i.e., including $O(\cdot)$. From Theorem 3, the added noise variance σ^2 has a specific boundary that is more feasible to be applied in the real algorithm process. Even though we can not compare the privacy bound directly with Theorem 1 in [23] because it is not feasible without subsampling. We illustrate a case from [23], with subsampling rate is 0.01, $\sigma = 4$, $\delta = 10^{-5}$, iteration time is 4000, they have $\epsilon = 2.55$. From the strong composition theorem, $\epsilon = 24.22$. As a comparison, without subsampling, our algorithm would get a much smaller one $\epsilon = 1.07$.

Utility Analysis. To the best of our knowledge, our study is the first paper to analyze the rigorous utility loss for MARL when preserving privacy. The utility loss is defined by the discrepancy between the value of cumulative function $J(s, \theta)$ output by Algorithms 1 and 2. Adding the noise to θ to preserve the privacy of individual agents has a disparate impact on the algorithm performance [52]. We give the theoretical analysis in Theorem 4 to show the upper bound of the utility loss.

Note that Algorithm 2 only executes T times, we have the following assumption for the iteration time T_2.

Assumption 6. *There exists iteration time $T_2 > 0$, and an arbitrarily small constant $\xi > 0$, that $\|\theta^* - \theta^i_{T_2}\|_1 \leq \xi$ for all $i \in \mathcal{N}$.*

Then, we formally state the utility loss as follows.

Theorem 4. *Under Assumptions 1, 2, 3 and 6, let $\tilde{J}(s, \theta^i_{T_2})$ and $\tilde{J}(s, \hat{\theta}^i_{T_2})$ denote the cumulative cost for the agent i at iteration T_2 generated by Algorithms 1 and 2 respectively. Given the constant $\zeta > 0$, the privacy target (ϵ, δ), for any agent $i \in \mathcal{N}$, the expected average increase of the of the $\tilde{J}(s, \theta)$ at iteration T_2 satisfies if $\|\theta^i_{T_2} - \hat{\theta}^i_{T_2}\|_1 \leq \zeta$,*

$$\begin{aligned}&\mathbb{E}[|\tilde{J}(s, \theta^i_{T_2}) - \tilde{J}(s, \hat{\theta}^i_{T_2})|] \\ &\leq \sqrt{K}\left(\left(\zeta - \frac{2R}{1-\lambda\gamma}\right)\frac{\sqrt{2}}{\sqrt{\pi}}\int_{-\infty}^{\frac{\zeta}{\sqrt{2}\sigma}} e^{-l^2} dl + \frac{2R}{1-\lambda\gamma}\right).\end{aligned}$$

where K is the size of vector θ^i_t, γ is the discount factor and R is defined in Assumption 1.

Proof. First, we calculate the probability of $\|\theta^i_{T_2} - \hat{\theta}^i_{T_2}\|_1 \leq \zeta$. From $u^0_t \sim \mathbb{N}(0, \sigma^2)$.

$$P(\|\theta^i_{T_2} - \hat{\theta}^i_{T_2}\|_1 \leq \zeta) = \int_{-\zeta}^{\zeta} \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{l^2}{2\sigma^2}} dl.$$

Let θ^{ik}_t denote value of θ of agent i, kth entry. The cumulative cost lost with $P(\|\theta^i_{T_2} - \hat{\theta}^i_{T_2}\| \leq \zeta)$ is,

$$\begin{aligned}&|\tilde{J}(s, \theta^i_{T_2}) - \tilde{J}(s, \hat{\theta}^i_{T_2})| = |\sum_{k=1}^{K} \theta^{ik}_{T_2}\phi_k(s) - \sum_{k=1}^{K} \hat{\theta}^{ik}_{T_2}\phi_k(s)| \\ &= |\sum_{k=1}^{K} (\theta^{ik}_{T_2} - \hat{\theta}^{ik}_{T_2})\phi_k(s)| \leq \sum_{k=1}^{K} |\theta^{ik}_{T_2} - \hat{\theta}^{ik}_{T_2}| \cdot \|\phi_k(s)\|_\infty \\ &\leq \sqrt{K}\|\theta^i_{T_2} - \hat{\theta}^i_{T_2}\|_1 \|\phi_k(s)\|_2 \leq \sqrt{K}\|\theta^i_{T_2} - \hat{\theta}^i_{T_2}\|_1 = \sqrt{K}\zeta.\end{aligned}$$

The probability that $\|\theta^i_{T_2} - \hat{\theta}^i_{T_2}\|_1 > \zeta$ is $1 - P(\|\theta_{T_2} - \hat{\theta}_{T_2}\|_1 \leq \zeta$.

From Lemma 7, the upper bound of $\|\theta^i_{T_2} - \hat{\theta}^i_{T_2}\|_1$ is $\frac{2R}{1-\lambda\gamma}$, then,

$$|\tilde{J}(s, \theta^i_{T_2}) - \tilde{J}(s, \hat{\theta}^i_{T_2})| \leq \sqrt{K}\frac{2R}{1-\lambda\gamma},$$

which is the worst case with the probability $1 - P(\|\theta^i_{T_2} - \hat{\theta}^i_{T_2}\|_1 \leq \zeta)$ could happen.

Then we have the following inequality,

$$\begin{aligned}&\mathbb{E}[|\tilde{J}(s, \theta_{T_2}) - \tilde{J}(s, \hat{\theta}_{T_2})|] \\ &\leq P(\|\theta^i_{T_2} - \hat{\theta}^i_{T_2}\|_1 \leq \zeta)\sqrt{K}\zeta \\ &\quad + (1 - P(\|\theta^i_{T_2} - \hat{\theta}^i_{T_2}\|_1 \leq \zeta))\frac{2R}{1-\lambda\gamma}\sqrt{K} \\ &= \sqrt{K}\left(\frac{\sqrt{2}}{\sqrt{\pi}}\int_{-\infty}^{\frac{\zeta}{\sqrt{2}\sigma}} e^{-l^2} dl \cdot \left(\zeta - \frac{2R}{1-\lambda\gamma}\right) + \frac{2R}{1-\lambda\gamma}\right).\end{aligned}$$

This completes the proof. □

Remark 3. Theorem 4 reveals the impact of adding noise in the algorithm by analyzing the relationship between the expected utility loss and the variance of Gaussian noise. Then, from Theorem 3 that presents the relationship between the privacy budget ϵ and the Gaussian noise variance σ, we have the following conclusion. The expected utility loss will decrease with the increase of the Gaussian noise variance σ. From Theorem 3, σ is decreasing with the increase of the privacy budget ϵ. Hence, the utility loss will increase with the increase of ϵ.

6 Performance Evaluation

We present the empirical results to demonstrate the efficiency of our proposed algorithms in this section.

The Environment Settings. We consider the system with $N = 20$ agents, a state space with $|\mathcal{S}| = 20$ states in total, and each agent has a binary action space, i.e., $\mathcal{A}^i = 0, 1$. The elements in the transition probability matrix $\mathcal{P}_s$ are uniformly sampled from [0,1] which are normalized to be stochastic. In order to satisfy the ergodicity of the MDP, we also add a small constant 10^{-5} onto each element in the matrix. Moreover, the selected feature matrices Φ is ensured to have full column rank as required in Assumption 2. For each state-value pair (s_1, s_2), the mean reward for each agent i, i.e., $R^i(s_1, s_2)$, is sampled uniformly from the

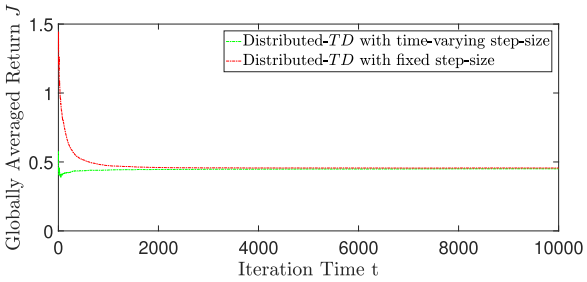

Fig. 1. Empirical results of Algorithm 1, J value versus iteration times.

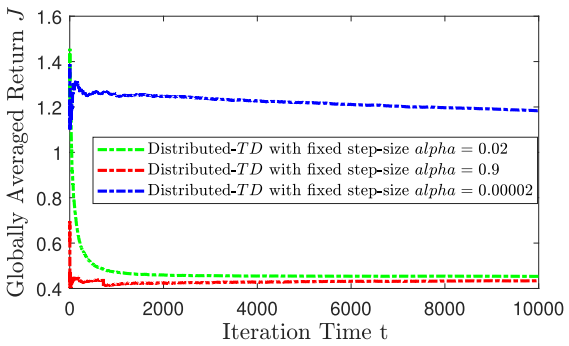

Fig. 2. Algorithm 1 with varying fixed step-sizes: J value versus iteration times.

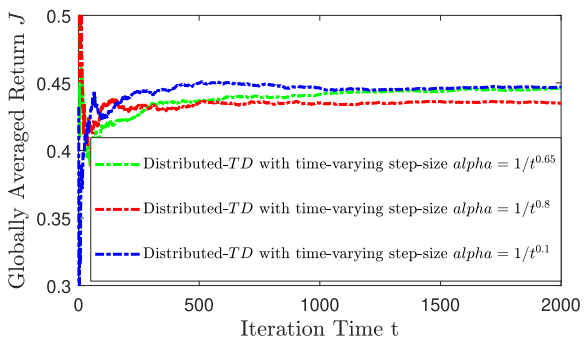

Fig. 3. Algorithm 1 with different time-varying step-sizes: J value versus iteration times.

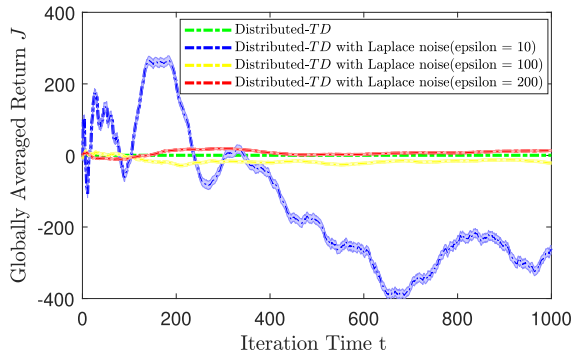

Fig. 4. Algorithm 2 on Laplace mechanism with time-varying step-size: J value with Laplace mechanism of DP budget versus iteration times.

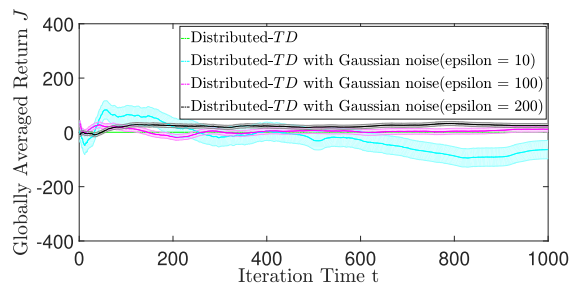

Fig. 5. Algorithm 2 on Gaussian mechanism with time-varying step-size: J value with Gaussian mechanism of DP budget versus iteration times.

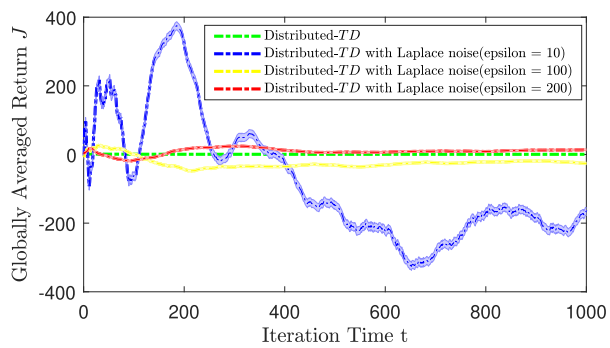

Fig. 6. Algorithm 2 on Laplace mechanism with fixed step size, J value versus iteration times.

interval [0.05,0.45], and it varies among agents. Besides, the instantaneous rewards r_t^i are uniformly sampled from $[R^i(s_1, s_2) - 0.05, R^i(s_1, s_2) + 0.05]$. The iteration time T is 1,000.

Parameters of Our Approach. We consider two groups of scenarios. The first group is the fixed step-size ($\alpha = 0.02$) algorithm and the time-varying step-size ($\alpha_t = t^{-0.65}$) algorithm. The second group contains the Laplace mechanism and the Gaussian mechanism. For the Laplace mechanism, from Theorem 3.6 in [19], the added Laplace perturbation follows the Laplace distribution $Lap(0, \frac{\Delta \mathcal{M}}{\epsilon})$ which can preserve (ϵ, δ) differential privacy. For the Gaussian mechanism, the added perturbation follows the Gaussian distribution $\mathcal{N}(0, \sigma^2)$ where σ^2 satisfies Theorem 3.

Convergence of Algorithm 1. Fig. 1 shows that Algorithm 1 converges to a same point of J value with both fixed step-size and time-varying step-size, which justifies the result from Theorem 1.

Effect of Fixed Step-Size. Fig. 2 presents the results with varying fixed step-sizes. Fig. 2 illustrates that when fixed step-size is small, the convergence time of Algorithm 1 will be large. For example, when $\alpha = 0.02$, Algorithm 1 dose not converge even the iteration time $t = 10000$, yet the conditions with $\alpha = 0.02$ and $\alpha = 0.9$ converged around $t = 4000$.

Effect of Time-Varying Step-Size. Fig. 3 tests Algorithm 1 with different time-varying step-sizes, i.e., $\alpha = \frac{1}{t^{0.1}}, \frac{1}{t^{0.65}}, \frac{1}{t^{0.8}}$. When α is small, the convergence speed of Algorithm 1 is relative slow. For example, when $t = 1000$, Algorithm 1 has converged with the conditions when $\alpha = \frac{1}{t^{0.1}}$ $\alpha = \frac{1}{t^{0.65}}$, yet $\alpha = \frac{1}{t^{0.8}}$ did not.

Effect of Noise Pattern. Figs. 4 and 5 show results with different privacy budgets of Laplace and Gaussian mechanisms in time-varying step-size, and Figs. 6 and 7 present the results of J value with Laplace mechanism and Gaussian mechanism in fixed step size. The global privacy budget is set as $\epsilon = \{10, 100, 200\}$, which is the same for the Laplace mechanism and Gauss mechanism. For the Laplace mechanism, it follows $(10, 0)$-differential private, $(100, 0)$-differential private, and

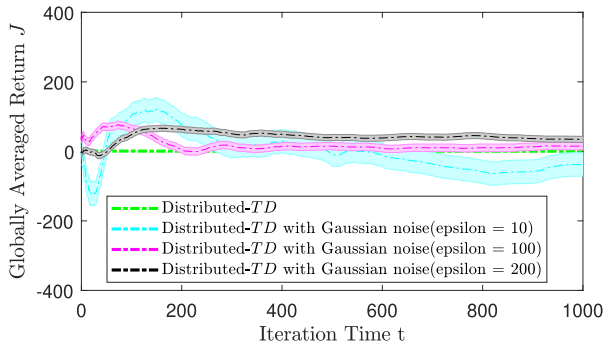

Fig. 7. Algorithm 2 on Gaussian mechanism fixed step size, J value versus iteration times., J value versus iteration times.

$(200, 0)$-differential private. For the Gaussian mechanism, it follows $(10, 10^{-7})$-differential private, $(100, 10^{-66})$-differential private, and $(200, 0)$-differential private. When $\epsilon = 10$, J value varies sharply because σ is mainly affected by the iteration time T. When ϵ is larger, J value keeps tight with the condition without noise, and in the Gaussian mechanism, δ can be kept at a very low level. We also gain insight that when the privacy budget is the same for Laplace and Gaussian mechanisms, the Gaussian mechanism outperforms the Laplace mechanism in the Federated TD(λ) learning model.

7 Conclusion

We develop a federated TD(λ) learning algorithm with both asymptotic and finite-time analyses considering both constant and time-varying step sizes. Furthermore, we propose a DP algorithm based on the federated TD(λ) learning model, show the rigorous differential privacy guarantee which offers a specific Gaussian variance selection guide and analyze the utility loss. For future work, we will focus on differential privacy problems in the decentralized TD(λ) learning and other RL algorithms, such as multi-agent actor-critic, by using different methods, including adding bounded Gaussian perturbation and via truncated concentrated DP [53].

Acknowledgments

Yiming Zeng and Yixuan Lin have contributed equally to this work.

References

[1] D. Silver, G. Lever, N. Heess, T. Degris, D. Wierstra, and M. Riedmiller, "Deterministic policy gradient algorithms," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 387–395.
[2] D. Ye *et al.*, "Towards playing full moba games with deep reinforcement learning," 2020, *arXiv:2011.12692*.
[3] C. Chen, A. Seff, A. Kornhauser, and J. Xiao, "Deepdriving: Learning affordance for direct perception in autonomous driving," in *Proc. IEEE Int. Conf. Comput. Vis.*, 2015, pp. 2722–2730.
[4] J. Kober and J. Peters, "Reinforcement learning in robotics: A survey," in *Proc. Reinforcement Learn.*, 2012, pp. 579–610.
[5] S. Gu, E. Holly, T. Lillicrap, and S. Levine, "Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates," in *Proc. IEEE Int. Conf. Robot. Autom.*, 2017, pp. 3389–3396.
[6] P. Abbeel, A. Coates, M. Quigley, and A. Y. Ng, "An application of reinforcement learning to aerobatic helicopter flight," in *Proc. Adv. Neural Inf. Process. Syst.*, 2007, pp. 1–8.
[7] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 1998.
[8] S. Kar, J. Moura, and H. Poor, "QD-learning: A collaborative distributed strategy for multi-agent reinforcement learning through consensus + innovations," *IEEE Trans. Signal Process.*, vol. 61, no. 7, pp. 1848–1862, Apr. 2013.
[9] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Başar, "Fully decentralized multi-agent reinforcement learning with networked agents," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 5872–5881.
[10] D. Lee, H. Yoon, and N. Hovakimyan, "Primal-dual algorithm for distributed reinforcement learning: Distributed GTD," in *Proc. IEEE Conf. Decis. Control*, 2018, pp. 1967–1972.
[11] T. Doan, S. Maguluri, and J. Romberg, "Finite-time analysis of distributed TD(0) with linear function approximation on multi-agent reinforcement learning," in *Proc. 36th Int. Conf. Mach. Learn.*, 2019, pp. 1626–1635.
[12] Z. Cao, P. Zhou, R. Li, S. Huang, and D. Wu, "Multiagent deep reinforcement learning for joint multichannel access and task offloading of mobile-edge computing in industry 4.0," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 6201–6213, Jul. 2020.
[13] L. Buşoniu, R. Babuška, and B. De Schutter, "Multi-agent reinforcement learning: An overview," *Innovations in Multi-Agent Systems and Applications-1*, Berlin, Germany: Springer, 2010, pp. 183–221.
[14] R. Sutton, "Learning to predict by the methods of temporal differences," *Mach. Learn.*, vol. 3, no. 1, pp. 9–44, 1988.
[15] J. Konečnỳ, H. B. McMahan, F. X. Yu, P. Richtárik, A. T. Suresh, and D. Bacon, "Federated learning: Strategies for improving communication efficiency," 2016, *arXiv:1610.05492*.
[16] X. Pan, W. Wang, X. Zhang, B. Li, J. Yi, and D. Song, "How you act tells a lot: Privacy-leaking attack on deep reinforcement learning," in *Proc. 18th Int. Conf. Autonomous Agents MultiAgent Syst.*, 2019, pp. 368–376.
[17] A. Y. Ng *et al.*, "Algorithms for inverse reinforcement learning," in *Proc. Int. Conf. Mach. Learn.*, 2000, pp. 663–670.
[18] P. Abbeel and A. Y. Ng, "Apprenticeship learning via inverse reinforcement learning," in *Proc. 21st Int. Conf. Mach. Learn.*, 2004.
[19] C. Dwork, A. Roth, "The algorithmic foundations of differential privacy," *Found. Trends Theor. Comput. Sci.*, vol. 9, no. 3–4, pp. 211–407, 2014.
[20] B. Balle, M. Gomrokchi, and D. Precup, "Differentially private policy evaluation," in *Proc. Int. Conf. Mach. Learn.*, 2016, pp. 2130–2138.
[21] B. Wang and N. Hegde, "Privacy-preserving q-learning with functional noise in continuous spaces," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 11323–11333.
[22] H. Ono and T. Takahashi, "Locally private distributed reinforcement learning," 2020, *arXiv:2001.11718*.
[23] M. Abadi *et al.*, "Deep learning with differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2016, pp. 308–318.
[24] P. C. M. Arachchige, P. Bertok, I. Khalil, D. Liu, S. Camtepe, and M. Atiquzzaman, "Local differential privacy for deep learning," *IEEE Internet Things J.*, vol. 7, no. 7, pp. 5827–5842, Jul. 2020.
[25] D. Wang, M. Ye, and J. Xu, "Differentially private empirical risk minimization revisited: Faster and more general," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 2722–2731.
[26] R. C. Geyer, T. Klein, and M. Nabi, "Differentially private federated learning: A client level perspective," 2017, *arXiv:1712.07557*.
[27] K. Zhang, Z. Yang, and T. Başar, "Multi-agent reinforcement learning: A selective overview of theories and algorithms," pp. 321–384, 2021.
[28] C. Boutilier, "Planning, learning and coordination in multi-agent decision processes," in *Proc. Conf. Theor. Aspects Rationality Knowl.*, 1996, pp. 195–210.
[29] M. Lauer and M. Riedmiller, "An algorithm for distributed reinforcement learning in cooperative multi-agent systems," in *Proc. Int. Conf. Mach. Learn.*, 2000, pp. 535–542.
[30] M. Littman, "Value-function reinforcement learning in Markov games," *Cogn. Syst. Res.*, vol. 2, no. 1, pp. 55–66, 2001.
[31] X. Wang and T. Sandholm, "Reinforcement learning to play an optimal Nash equilibrium in team Markov games," in *Proc. Adv. Neural Inf. Process. Syst.*, 2003, pp. 1603–1610.
[32] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Başar, "Finite-sample analysis for decentralized batch multi-agent reinforcement learning with networked agents," *IEEE Trans. Autom. Control*, 2021.
[33] T. Chen, K. Zhang, G. B. Giannakis, and T. Başar, "Communication-efficient policy gradient methods for distributed reinforcement learning," *IEEE Control Netw. Syst.*, 2021.
[34] R. Lowe, Y. Wu, A. Tamar, J. Harb, P. Abbeel, and I. Mordatch, "Multi-agent actor-critic for mixed cooperative-competitive environments," in *Proc. Adv. Neural Inf. Process. Syst.*, 2017, pp. 6379–6390.

[35] J. Hu and M. Wellman, "Nash Q-learning for general-sum stochastic games," *J. Mach. Learn. Res.*, vol. 4, no. 11, pp. 1039–1069, 2003.
[36] J. Foerster, Y. Assael, N. Freitas, and S. Whiteson, "Learning to communicate with deep multi-agent reinforcement learning," in *Proc. Adv. Neural Inf. Process. Syst.*, 2016, pp. 2137–2145.
[37] S. Omidshafiei, J. Pazis, C. Amato, J. P. How, and J. Vian, "Deep decentralized multi-task multi-agent reinforcement learning under partial observability," in *Proc. Int. Conf. Mach. Learn.*, 2017, pp. 2681–2690.
[38] J. N. Tsitsiklis and B. Van Roy, "An analysis of temporal-difference learning with function approximation," *IEEE Trans. Autom. Control*, vol. 42, no. 5, pp. 674–690, May 1997.
[39] P. Dayan, "The convergence of TD (λ) for general λ," *Mach. Learn.*, vol. 8, no. 3–4, pp. 341–362, 1992.
[40] F. Pineda, "Mean-field theory for batched TD (λ)," *Neural Comput.*, vol. 9, no. 7, pp. 1403–1419, 1997.
[41] M. Stanković and S. Stanković, "Multi-agent temporal-difference learning with linear function approximation: Weak convergence under time-varying network topologies," in *Proc. Amer. Control Conf.*, 2016, pp. 167–172.
[42] T. Doan, S. Maguluri, and J. Romberg, "Finite-time performance of distributed temporal-difference learning with linear function approximation," *SIAM J. Math. Data Sci.*, vol. 3, no. 1, pp. 298–320, 2021.
[43] G. Dalal, B. Szörényi, G. Thoppe, and S. Mannor, "Finite sample analyses for TD(0) with function approximation," in *Proc. 32nd AAAI Conf. Artif. Intell.*, 2018.
[44] J. Bhandari, D. Russo, and R. Singal, "A finite time analysis of temporal difference learning with linear function approximation," in *Proc. Conf. Learn. Theory*, 2018, pp. 1691–1692.
[45] C. Lakshminarayanan and C. Szepesvari, "Linear stochastic approximation: How far does constant step-size and iterate averaging go?," in *Proc. Int. Conf. Artif. Intell. Statist.*, 2018, pp. 1347–1355.
[46] R. Srikant and L. Ying, "Finite-time error bounds for linear stochastic approximation and td learning," in *Proc. Conf. Learn. Theory PMLR*, 2019, pp. 2803–2830.
[47] S. Meiser and E. Mohammadi, "Tight on budget? tight bounds for r-fold approximate differential privacy," in *Proc. ACM SIGSAC Conf. Comput. Commun. Secur.*, 2018, pp. 247–264.
[48] D. M. Sommer, S. Meiser, and E. Mohammadi, "Privacy loss classes: The central limit theorem in differential privacy," *Proc. Privacy Enhancing Technol.*, vol. 2019, no. 2, pp. 245–269, 2019.
[49] Y. -X. Wang, B. Balle, and S. P. Kasiviswanathan, "Subsampled rényi differential privacy and analytical moments accountant," in *Proc. 22nd Int. Conf. Artificial Intell. Statistics. PMLR*, 2019, pp. 1226–1235.
[50] H. Kushner and G. Yin, "Asymptotic properties of distributed and communicating stochastic approximation algorithms," *SIAM J. Control Optim.*, vol. 25, no. 5, pp. 1266–1290, 1987.
[51] P. Brémaud, *Markov Chains: Gibbs Fields, Monte Carlo Simulation, and Queues*. Berlin, Germany: Springer, 2013.
[52] E. Bagdasaryan, O. Poursaeed, and V. Shmatikov, "Differential privacy has disparate impact on model accuracy," in *Proc. Adv. Neural Inf. Process. Syst.*, 2019, pp. 15453–15 462.
[53] M. Bun, C. Dwork, G. N. Rothblum, and T. Steinke, "Composable and versatile privacy via truncated CDP," in *Proc. 50th Annu. ACM SIGACT Symp. Theory Comput.*, 2018, pp. 74–86.

Yiming Zeng received the BEng degree in information engineering from Shanghai Jiao Tong University, Shanghai, China. He is currently working toward the PhD degree in computer and electrical engineering with Stony Brook University, New York, NY. His research focuses on addressing computing, privacy, and caching issues in edge networks.

Yixuan Lin received the BS degree in mathematics from Fudan University, Shanghai, China, in 2017 and the MS degree in 2019 in mathematics and statistics from Stony Brook University, Stony Brook, NY, USA, where she is currently working toward the PhD degree in applied mathematics and statistics from Stony Brook University, Stony Brook, NY, USA. Her research interests include finite-time analysis, distributed reinforcement learning algorithm, and resilient problem.

Yuanyuan Yang (Fellow, IEEE) received the BEng and MS degrees in computer science and engineering from Tsinghua University, Beijing, China, and the MSE and PhD degrees in computer science from Johns Hopkins University, Baltimore, Maryland. She is currently a SUNY distinguished professor of computer engineering and computer science with Stony Brook University, New York and is currently on leave with National Science Foundation as the program director. She has authored or coauthored more than 460 papers in major journals and refereed conference proceedings and holds seven U.S. patents in the areas of her research interests, which include edge computing, data center networks, cloud computing, and wireless networks. She is currently the editor-in-chief of *IEEE Transactions on Cloud Computing* and an associate editor for *IEEE Transactions on Parallel and Distributed Systems* and *ACM Computing Surveys*. She was an associate editor-in-chief for *IEEE Transactions on Cloud Computing*, an associate editor-in-chief and an associated editor for *IEEE Transactions on Computers*, and an associate editor for *IEEE Transactions on Parallel and Distributed Systems*. She was the general chair, program chair, or vice chair for several major conferences and a program committee member for numerous conferences.

Ji Liu received the BS degree in information engineering from Shanghai Jiao Tong University, Shanghai, China, in 2006 and the PhD degree in electrical engineering from Yale University, New Haven, CT, USA, in 2013. He is currently an assistant professor with the Department of Electrical and Computer Engineering, Stony Brook University, Stony Brook, NY, USA. He is an associate editor for the *IEEE Transactions on Signal and Information Processing over Networks*. His research interests include distributed control and optimization, reinforcement learning, federated learning, epidemic networks, social networks, and cyber-physical systems.