Finite-Time Analysis of Decentralized Temporal-Difference Learning with Linear Function Approximation

Jun Sun[1] **Gang Wang**[2] **Georgios B. Giannakis**[2] **Qinmin Yang**[1] **Zaiyue Yang**[3]
[1]Zhejiang University [2]University of Minnesota [3]Southern University of Science and Technology

Abstract

Motivated by the emerging use of multi-agent reinforcement learning (MARL) in various engineering applications, we investigate the policy evaluation problem in a fully decentralized setting, using temporal-difference (TD) learning with linear function approximation to handle large state spaces in practice. The goal of a group of agents is to collaboratively learn the value function of a given policy from locally private rewards observed in a shared environment, through exchanging local estimates with neighbors. Despite their simplicity and widespread use, our theoretical understanding of such decentralized TD learning algorithms remains limited. Existing results were obtained based on i.i.d. data samples, or by imposing an 'additional' projection step to control the 'gradient' bias incurred by the Markovian observations. In this paper, we provide a finite-sample analysis of the fully decentralized TD(0) learning under both i.i.d. as well as Markovian samples, and prove that all local estimates converge linearly to a neighborhood of the optimum. The resultant error bounds are the first of its type—in the sense that they hold under the most practical assumptions — which is made possible by means of a novel multi-step Lyapunov analysis.

1 INTRODUCTION

Reinforcement learning (RL) is concerned with how artificial agents ought to take actions in an unknown environment so as to maximize some notion of a cumulative reward. In recent years, combining with deep learning, RL has demonstrated its great potential in addressing challenging practical control and optimization problems [Mnih et al., 2015, Yang et al., 2019]. Among all possible algorithms, the temporal difference (TD) learning has arguably become one of the most popular RL algorithms so far, which is further dominated by the celebrated TD(0) algorithm [Sutton, 1988]. TD learning provides an iterative process to update an estimate of the value function $v^{\mu}(s)$ with respect to a given policy μ based on temporally successive samples. Dealing with a finite state space, the classical version of the TD(0) algorithm adopts a tabular representation for $v^{\mu}(s)$, which stores entry-wise value estimates on a per state basis.

Although it is conceptually simple as well as easy-to-implement, the tabular TD(0) learning algorithm can become intractable when the number of states grows large or even infinite, which emerges in many contemporary control and artificial intelligence problems of practical interest. This is also known as the "curse of dimensionality" [Bertsekas and Tsitsiklis, 1996]. The common practice to bypass this hurdle, is to approximate the exact tabular value function with a class of function approximators, including for example, linear functions or nonlinear ones using even deep neural networks [Sutton and Barto, 2018].

Albeit nonlinear approximators using e.g., deep neural networks [Mnih et al., 2015, Wang et al., 2019a], can be more powerful, linear approximation allows for an efficient implementation of TD learning even on large or infinite state spaces, which has been demonstrated to perform well in several applications [Sutton and Barto, 2018]. Specifically, TD learning with linear approximation parameterizes the value function with a linear combination of a set of preselected basis functions (a.k.a., feature vectors) induced by the states, and estimates the coefficients in the spirit of vanilla TD learning. Indeed, recent theoretical RL efforts have mostly been devoted to linear approximation; see e.g., [Baird, 1995, Bhandari et al., 2018, Hu and Syed, 2019, Xu et al., 2020].

Proceedings of the 23[rd]International Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108. The first two authors contributed equally.

Early theoretical convergence results of TD learning were mostly asymptotic [Sutton, 1988, Baird, 1995]; that is, results that hold only asymptotically when the number of updates (data samples) tends to infinity. By exploring the asymptotic behavior, TD(0) learning with linear function approximation can be viewed as a discretized version of an ordinary differential equation (ODE) [Tsitsiklis and Van Roy, 1997], or a linear dynamical system [Borkar, 2008]; therefore, TD(0) updates can be seen as tracking the trajectory of the ODE provided the learning rate is infinitely small [Tsitsiklis and Van Roy, 1997]. Motivated by the need for dealing with massive data in modern signal processing, control, and artificial intelligence tasks (e.g., [Chi et al., 2019, Mnih et al., 2015]), recent interests have centered around developing non-asymptotic performance guarantees that hold with even finite data samples and help us understand the efficiency of the algorithm or agent in using data.

Non-asymptotic analysis of RL algorithms, and TD learning in particular, is generally more challenging than their asymptotic counterpart, due mainly to two reasons that: i) TD updates do not correspond to minimizing any static objective function as standard optimization algorithms do; and, ii) data samples garnered along the trajectory of a single Markov chain are correlated across time, resulting in considerably large instantaneous 'gradient' bias in the updates. Addressing these challenges, a novel suite of tools has lately been put forward. Adopting the dynamical system viewpoint, the iterates of TD(0) updates after a projection step were shown converging to the equilibrium point of the associated ODE at a sublinear rate in [Dalal et al., 2018]. With additional projection steps, finite-time error bounds of a two-timescale TD learning algorithm developed by [Sutton et al., 2009] were established in [Gupta et al., 2019]. The authors in [Bhandari et al., 2018] unified finite-time results of TD(0) with linear function approximation, under both identically, and independently distributed (i.i.d.) noise and Markovian noise.

In summary, these aforementioned works in this direction either assume i.i.d. data samples [Dalal et al., 2018], or have to incorporate a projection step [Bhandari et al., 2018]. As pointed out in [Dalal et al., 2018] however, although widely adopted, i.i.d. samples are difficult to acquire in practice. On the other hand, the projection step is imposed only for analysis purposes, yet the projection can be difficult to implement. Moreover, most existing theoretical RL studies have considered the centralized setting, except for [Doan et al., 2019] dealing with finite-time analysis of decentralized TD(0) under the i.i.d. assumption and with a projection step. In a fully decentralized setting, multi-agents share a common environment but observe private rewards. With the goal of jointly maximizing the total accumulative reward, each agent can communicate with its neighbors, and updates the parameter locally. Such decentralized schemes appear naturally in several applications including robotics and mobile sensor networks [Krishnamurthy et al., 2008].

As a complementary to existing theoretical RL efforts, this paper offers a novel finite-sample analysis for a fully decentralized TD(0) algorithm with linear function approximation. For completeness of our analytical results, we investigate both the i.i.d. case as, well as, the practical yet challenging Markovian setting, where data samples are gathered along the trajectory of a single Markov chain. To render the finite-time analysis under the Markovian noise possible, we invoke a novel multi-step Lyapunov approach [Wang et al., 2019b], which successfully eliminates the need for a projection step as required by [Doan et al., 2019]. Our theoretical results show that a fully decentralized implementation of the original TD(0) learning, converges linearly to a neighborhood of the optimum under both i.i.d. and Markovian observations. Furthermore, the size of this neighborhood can be made arbitrarily small by choosing a small enough stepsize. In a nutshell, the main contributions of this paper are summarized as follows.

c1) We investigate the fully decentralized TD(0) learning with linear function approximation, and establish the multi-agent consensus, as well as their asymptotic convergence; and,

c2) We provide finite-time error bounds for all agents' local parameter estimates in a fully decentralized setting, under both i.i.d. and Markovian observations, through a multi-step Lyapunov analysis.

2 DECENTRALIZED REINFORCEMENT LEARNING

A discounted Markov decision process (MDP) is a discrete-time stochastic control process, which can be characterized by a 5-tuple $(\mathcal{S}, \mathcal{A}, P^a, R^a, \gamma)$. Here, $\mathcal{S}$ is a finite set of environment and agent states, $\mathcal{A}$ is a finite set of actions of the agent, $P^a(s, s') = \Pr(s'|s, a)$ is the probability of transition from state $s \in \mathcal{S}$ to state s' upon taking action $a \in \mathcal{A}$, $R^a(s, s') : \mathcal{S} \times \mathcal{S} \to \mathbb{R}$ represents the immediate reward received after transitioning from state s to state s' with action a, and γ is the discounting factor.

The core problem of MDPs is to find a policy for the agent, namely a mapping $\mu : \mathcal{S} \times \mathcal{A} \to [0, 1]$ that specifies the probability of choosing action $a \in \mathcal{A}$

when in state s. Once an MDP is combined with a policy, this fixes the action for each state and their combination determines the stochastic dynamics of a Markov chain [Bhatnagar et al., 2009]. Indeed, this is because the action a chosen in state s is completely determined by $\mu(s,a)$, then $\Pr(s'|s,a)$ reduces to $P^\mu(s,s') = \sum_{a\in\mathcal{A}} \mu(s,a)P^a(s'|s)$, a Markov transition matrix $\boldsymbol{P}^\mu$. Likewise, immediate reward $R^a(s,s')$ also simplifies to the expected reward $R^\mu(s,s') = \sum_{a\in\mathcal{A}} \mu(s,a)P^a(s'|s)R^a(s'|s)$.

The quality of policy μ is evaluated in terms of the expected sum of discounted rewards over all states in a finite-sample path while following policy μ to take actions, which is also known as the value function $v^\mu : \mathcal{S} \to \mathbb{R}$. In this paper, we focus on evaluating a given policy μ, so we neglect the dependence on μ hereafter for notational brevity. Formally, $v(s)$ is defined as

$$v(s) = \mathbb{E}\Big[\sum_{k=0}^{\infty} \gamma^k R(s(k), s(k+1))\Big| s(0) = s\Big], \quad \forall s \in \mathcal{S} \tag{1}$$

where the expectation is taken over all transitions from $k = 0$ to $k = +\infty$.

Assuming a canonical ordering on the elements of $\mathcal{S}$, say a renumbering $\{1, 2, \ldots, |\mathcal{S}|\}$, we can treat v as a $|\mathcal{S}|$-dimensional vector $\boldsymbol{v} := [v(1)\ \ v(2)\ \cdots\ v(|\mathcal{S}|)]^\top \in \mathbb{R}^{|\mathcal{S}|}$. It is well known that the value function $v(s)$ in (1) satisfies the so-called Bellman equation [Bertsekas and Tsitsiklis, 1996]

$$v(s) = \sum_{s'\in\mathcal{S}} P_{ss'}\big[R(s,s') + \gamma v(s')\big], \quad \forall s \in \mathcal{S}. \tag{2}$$

If the transition probabilities $\{P_{ss'}\}$ and the expected rewards $\{R(s,s')\}$ were known, finding $v \in \mathbb{R}^{|\mathcal{S}|}$ is tantamount to solving a system of linear equations given by (2). It is clear that when the number of states $|\mathcal{S}|$ is large or even infinite, exact computation of v can become intractable, which is also known as the "curse of dimensionality" [Bertsekas and Tsitsiklis, 1996]. This thus motivates well a low-dimensional (linear) function approximation of $v(s)$, parameterized by an unknown vector $\boldsymbol{\theta} \in \mathbb{R}^p$ as follows

$$v(s) \approx \tilde{v}(s, \boldsymbol{\theta}) = \boldsymbol{\phi}^\top(s)\boldsymbol{\theta}, \quad \forall s \in \mathcal{S} \tag{3}$$

where we oftentimes have the number of unknown parameters $p \ll |\mathcal{S}|$; and $\boldsymbol{\phi}(s) \in \mathbb{R}^p$ is a preselected feature or basis vector characterizing state $s \in \mathcal{S}$.

For future reference, let vector $\tilde{\boldsymbol{v}}(\boldsymbol{\theta}) := [\tilde{v}(1,\boldsymbol{\theta})\ \ \tilde{v}(2,\boldsymbol{\theta})\ \cdots\ \tilde{v}(|\mathcal{S}|,\boldsymbol{\theta})]^\top$ collect the value function approximations at all states, and define the feature matrix

$$\boldsymbol{\Phi} := \begin{bmatrix} \boldsymbol{\phi}^\top(1) \\ \boldsymbol{\phi}^\top(2) \\ \vdots \\ \boldsymbol{\phi}^\top(|\mathcal{S}|) \end{bmatrix} \in \mathbb{R}^{|\mathcal{S}|\times p}$$

then it follows that

$$\tilde{\boldsymbol{v}}(\boldsymbol{\theta}) = \boldsymbol{\Phi}\boldsymbol{\theta}. \tag{4}$$

Regarding the basis vectors $\{\boldsymbol{\phi}(s)\}$ (or equivalently, the feature matrix $\boldsymbol{\Phi}$), we make the next two standard assumptions [Tsitsiklis and Van Roy, 1997]: i) $\|\boldsymbol{\phi}(s)\| \le 1$, $\forall s \in \mathcal{S}$, that is, all feature vectors are normalized; and, ii) $\boldsymbol{\Phi}$ is of full column rank, namely, all feature vectors are linearly independent.

With the above linear approximation, the task of seeking $\boldsymbol{v}$ boils down to find the parameter vector $\boldsymbol{\theta}^*$ that minimizes the gap between the true value function $\boldsymbol{v}$ and the approximated one $\tilde{\boldsymbol{v}}(\boldsymbol{\theta})$. Among many possibilities in addressing this task, the original temporal difference learning algorithm, also known as TD(0), is arguably the most popular solution [Sutton, 1988]. The goal of this paper is to develop decentralized TD(0) learning algorithms and further investigate their finite-time performance guarantees in estimating $\boldsymbol{\theta}^*$. To pave the way for decentralized TD(0) learning, we first introduce standard centralized version below.

2.1 Centralized Temporal Difference Learning

The classical TD(0) algorithm with function approximation [Sutton, 1988] starts with some initial guess $\boldsymbol{\theta}(0) \in \mathbb{R}^p$. Upon observing the k^{th} transition from state $s(k)$ to state $s(k+1)$ with reward $r(k) = R(s(k, s(k+1)))$, it first computes the so-called temporal-difference error, given by

$$d(k) = r(k) + \gamma\tilde{v}(s(k+1), \boldsymbol{\theta}(k)) - \tilde{v}(s(k), \boldsymbol{\theta}(k)) \tag{5}$$

which is subsequently used to update the parameter vector $\boldsymbol{\theta}_k$ as follows

$$\boldsymbol{\theta}(k+1) = \boldsymbol{\theta}(k) + \alpha d(k)\nabla\tilde{v}(s(k), \boldsymbol{\theta}(k)). \tag{6}$$

Here, $\alpha > 0$ is a preselected constant stepsize, and the symbol $\nabla\tilde{v}(s(k), \boldsymbol{\theta}(k)) = \boldsymbol{\phi}(s(k))$ denotes the gradient of $\tilde{v}(s(k), \boldsymbol{\theta})$ with respect to $\boldsymbol{\theta}$ evaluated at the current estimate $\boldsymbol{\theta}(k)$. For ease of exposition, we define the 'gradient' estimate $\boldsymbol{g}(k)$ as follows

$$\begin{aligned} \boldsymbol{g}(\boldsymbol{\theta}(k), \xi_k) &:= d(k)\nabla\tilde{v}(s(k), \boldsymbol{\theta}(k)) \\ &= \boldsymbol{\phi}(s(k))\big[\gamma\boldsymbol{\phi}^\top(s(k+1)) - \boldsymbol{\phi}^\top(s(k))\big]\,\boldsymbol{\theta}(k) \\ &\quad + r(k)\boldsymbol{\phi}(s(k)). \end{aligned} \tag{7}$$

where ξ_k captures all the randomness corresponding to the k-th transition $(s(k), s(k+1), \{r_m(k)\}_{m\in\mathcal{M}})$. Thus, the TD(0) update (6) can be rewritten as

$$\boldsymbol{\theta}(k+1) = \boldsymbol{\theta}(k) + \alpha\boldsymbol{g}(\boldsymbol{\theta}(k), \xi_k). \tag{8}$$

Albeit viewing $\boldsymbol{g}(\boldsymbol{\theta}(k), \xi_k)$ as some negative 'gradient' estimate, the TD(0) update in (8) based on online rewards resembles that of the stochastic gradient descent (SGD). It is well known, however, that even the TD(0) learning update does not correspond to minimizing any fixed objective function [Sutton and Barto, 2018]. Indeed, this renders convergence analysis of TD algorithms rather challenging, letting alone the non-asymptotic (i.e., finite-time) analysis. To address this challenge, TD learning algorithms have been investigated in light of the stability of a dynamical system described by an ordinary differential equation (ODE) [Borkar, 2008, Tsitsiklis and Van Roy, 1997, Wang et al., 2019b].

Before introducing the ODE system for (8), let us first simplify the expression of $\boldsymbol{g}(\boldsymbol{\theta}(k))$. Upon defining

$$\boldsymbol{H}(\xi_k) := \boldsymbol{\phi}(s(k))\big[\gamma\boldsymbol{\phi}^\top(s(k+1)) - \boldsymbol{\phi}^\top(s(k))\big] \quad (9)$$

and

$$\boldsymbol{b}(\xi_k) := r(k)\boldsymbol{\phi}(s(k)) \quad (10)$$

the gradient estimate $\boldsymbol{g}(\boldsymbol{\theta}(k))$ can be re-expressed as

$$\boldsymbol{g}(\boldsymbol{\theta}(k), \xi_k) = \boldsymbol{H}(\xi_k)\boldsymbol{\theta}(k) + \boldsymbol{b}(\xi_k). \quad (11)$$

Assuming that the Markov chain is finite, irreducible, and aperiodic, there exists a unique stationary distribution $\boldsymbol{\pi} \in \mathbb{R}^{1\times|\mathcal{S}|}$ [Levin and Peres, 2017], adhering to $\boldsymbol{\pi}\boldsymbol{P} = \boldsymbol{\pi}$. Moreover, let $\boldsymbol{D}$ be a diagonal matrix holding entries of $\boldsymbol{\pi}$ on its main diagonal. We also introduce $r'(s) := \sum_{s'\in\mathcal{S}} P(s,s')R(s,s')$ for all $s \in \mathcal{S}$ and collect them into vector $\boldsymbol{r'} = \big[r'(1)\ r'(2)\ \cdots\ r'(|\mathcal{S}|)\big]^\top$.

It can be verified that after the Markov chain reaches the stationary distribution, the following limits hold

$$\bar{\boldsymbol{H}} := \lim_{k\to\infty} \mathbb{E}[\boldsymbol{H}(\xi_k)] = \boldsymbol{\Phi}\boldsymbol{D}(\gamma\boldsymbol{P}\boldsymbol{\Phi}^\top - \boldsymbol{\Phi}^\top) \quad (12)$$

$$\bar{\boldsymbol{b}} := \lim_{k\to\infty} \mathbb{E}[\boldsymbol{b}(\xi_k)] = \boldsymbol{\Phi}\boldsymbol{D}\boldsymbol{r'} \quad (13)$$

yielding

$$\bar{\boldsymbol{g}}(\boldsymbol{\theta}) := \bar{\boldsymbol{H}}\boldsymbol{\theta} + \bar{\boldsymbol{b}}. \quad (14)$$

It has been shown that, under mild conditions on the stepsize α, the TD(0) update (6) or (8) can be understood as tracking the following ODE [Tsitsiklis and Van Roy, 1997]

$$\dot{\boldsymbol{\theta}} = \bar{\boldsymbol{g}}(\boldsymbol{\theta}). \quad (15)$$

For any $\gamma \in [0,1)$, it can be further shown that albeit not symmetric, matrix $\bar{\boldsymbol{H}}$ is negative definite, in the sense that $\boldsymbol{\theta}^\top\bar{\boldsymbol{H}}\boldsymbol{\theta} < 0$ for any $\boldsymbol{\theta} \neq \boldsymbol{0}$. Appealing to standard linear systems theory (see e.g., [Bof et al., 2018]), we have that the ODE (15) admits a globally, asymptotically stable equilibrium point $\boldsymbol{\theta}^*$, dictated by

$$\bar{\boldsymbol{g}}(\boldsymbol{\theta}^*) = \bar{\boldsymbol{H}}\boldsymbol{\theta}^* + \bar{\boldsymbol{b}} = \boldsymbol{0}. \quad (16)$$

2.2 Decentralized TD(0) Learning

The goal of this paper is to investigate the policy evaluation problem in the context of multi-agent reinforcement learning (MARL), where a group of agents cooperate to evaluate the value function in an environment. Suppose there is a set $\mathcal{M}$ of agents with $|\mathcal{M}| = M$, distributed across a network denoted by $\mathcal{G} = (\mathcal{M}, \mathcal{E})$, where $\mathcal{E} \subseteq \mathcal{M}\times\mathcal{M}$ represents the edge set. Let $\mathcal{N}_m \subseteq \mathcal{M}$ collect the neighbor(s) of agent $m \in \mathcal{M}$, for all $m \in \mathcal{M}$. We assume that each agent locally implements a stationary policy μ_m. As explained in the centralized setting, when combined with fixed policies $\{\mu_m\}_{m\in\mathcal{M}}$, the multi-agent MDP can be described by the following 6-tuple

$$\big(\mathcal{S}, \{\mathcal{A}_m\}_{m=1}^M, P, \{R_m\}_{m=1}^M, \gamma, \mathcal{G}\big) \quad (17)$$

where $\mathcal{S}$ is a finite set of states shared by all agents, $\mathcal{A}_m$ is a finite set of actions available to agent m, and R_m is the immediate reward observed by agent m. It is worth pointing out that, here, we assume there is no centralized controller that can observe all information; instead, every agent can observe the joint state vector $s \in \mathcal{S}$, yet its action $a_m \in \mathcal{A}_m$ as well as reward $R_m(s,s')$ is kept private from other agents.

Specifically, at time instant k, each agent m observes the current state $s(k) \in \mathcal{S}$ and chooses action $a \in \mathcal{A}_m$ according to a stationary policy μ_m. Based on the joint actions of all agents, the system transits to a new state $s(k+1)$, for which an expected local reward $r_m(k) = R_m(s(k), s(k+1))$ is revealed to agent m. The objective of multi-agent policy evaluation is to cooperatively compute the average of the expected sums of discounted rewards from a network of agents, given by

$$v_{\mathcal{G}}(s) = \mathbb{E}\bigg[\frac{1}{M}\sum_{m\in\mathcal{M}}\sum_{k=0}^{\infty}\gamma^k R_m(s(k), s(k+1))\Big|s(0) = s\bigg]. \quad (18)$$

Similar to the centralized case, one can show that $v_{\mathcal{G}}(s)$ obeys the following multi-agent Bellman equation

$$v_{\mathcal{G}}(s) = \sum_{s'\in\mathcal{S}} P_{ss'}\Big[\frac{1}{M}\sum_{m\in\mathcal{M}} R_m(s,s') + \gamma v_{\mathcal{G}}(s')\Big], \quad \forall s \in \mathcal{S}. \quad (19)$$

Again, to address the "curse of dimensionality" in exact computation of $v_{\mathcal{G}}$ when the space $\mathcal{S}$ grows large, we are particularly interested in low-dimensional (linear) function approximation $\tilde{v}_{\mathcal{G}}(s)$ of $v_{\mathcal{G}}(s)$ as given in (3), or (4) in a matrix-vector representation.

Define $\boldsymbol{b}_m(k) := r_m(k)\boldsymbol{\phi}(s(k))$, $\bar{\boldsymbol{b}}_m = \mathbb{E}_\pi[\boldsymbol{b}_m(k)]$, $\boldsymbol{b}_{\mathcal{G}} := \frac{1}{M}\sum_{m\in\mathcal{M}}\boldsymbol{b}_m(k)$ and $\bar{\boldsymbol{b}}_{\mathcal{G}} := \frac{1}{M}\sum_{m\in\mathcal{M}}\bar{\boldsymbol{b}}_m$. As all agents share the same environment by observing a common state vector $s(k)$, and differ only in their

Algorithm 1 Decentralized TD(0) learning

Input: stepsize $\alpha > 0$, feature matrix $\boldsymbol{\Phi}$, and weight matrix $\boldsymbol{W}$.
Initialize: $\{\boldsymbol{\theta}_m(0)\}_{m\in\mathcal{M}}$.
for $k = 0, 1, \cdots, K$ **do**
  for $m = 1, 2, \cdots, M$ **do**
    Agent m receives $\boldsymbol{\theta}_{m'}(k)$ from its neighbors $m' \in \mathcal{N}_m$;
    Agent m observes $(s(k), s(k+1), r_m(k))$, and computes $\boldsymbol{g}_m(\boldsymbol{\theta}_m(k))$ according to (21);
    Agent m updates $\boldsymbol{\theta}_m(k)$ via (22), and broadcasts $\boldsymbol{\theta}_m(k+1)$ to its neighbors $m' \in \mathcal{N}_m$.
  end for
end for

rewards, the parameter vector $\boldsymbol{\theta}^*$ such that the linear function approximator $\tilde{\boldsymbol{v}}_{\mathcal{G}} = \boldsymbol{\Phi}\boldsymbol{\theta}^*$ satisfies the multi-agent Bellman equation (19); that is,

$$\bar{\boldsymbol{H}}\boldsymbol{\theta}^* + \bar{\boldsymbol{b}}_{\mathcal{G}} = \boldsymbol{0} \tag{20}$$

We are ready to study a standard consensus-based distributed variant of the centralized TD(0) algorithm, which is tabulated in Algorithm 1 for reference. Specifically, at the beginning of time instant k, each agent m first observes $(s(k), s(k+1), R_m(s(k), s(k+1)))$ and calculates the local gradient

$$\boldsymbol{g}_m(\boldsymbol{\theta}_m(k), \xi_k) := \boldsymbol{\phi}(s(k))\big[\gamma\boldsymbol{\phi}^\top(s(k+1)) - \boldsymbol{\phi}^\top(s(k))\big]\boldsymbol{\theta}_m(k) + r_m(k)\boldsymbol{\phi}(s(k)) \tag{21}$$

Upon receiving estimates $\{\boldsymbol{\theta}_{m'}(k)\}$ from its neighbors $m' \in \mathcal{N}_m$, agent m ($m \in \mathcal{M}$) updates its local estimate $\boldsymbol{\theta}_m(k)$ according to the following recursion

$$\boldsymbol{\theta}_m(k+1) = \sum_{m'\in\mathcal{M}} W_{mm'}\boldsymbol{\theta}_{m'}(k) + \alpha\boldsymbol{g}_m(\boldsymbol{\theta}_m(k), \xi_k), \tag{22}$$

where $W_{mm'}$ is a weight attached to the edge (m, m'); and $W_{mm'} > 0$ if $m' \in \mathcal{N}_m$, and $W_{mm'} = 0$, otherwise. Throughout this paper, we have following assumption on the network.

Assumption 1. *The communication network is connected and undirected, and the associated weight matrix $\boldsymbol{W}$ is a doubly stochastic matrix.*

For ease of exposition, we stack up all local parameter estimates $\{\boldsymbol{\theta}_m\}_{m\in\mathcal{M}}$ into matrix

$$\boldsymbol{\Theta} := \begin{bmatrix} \boldsymbol{\theta}_1^\top \\ \boldsymbol{\theta}_2^\top \\ \vdots \\ \boldsymbol{\theta}_M^\top \end{bmatrix} \in \mathbb{R}^{M\times p}. \tag{23}$$

and similarly for all local gradient estimates $\{\boldsymbol{g}_m(\boldsymbol{\theta}_m, \xi_k)\}_{m\in\mathcal{M}}$

$$\boldsymbol{G}(\boldsymbol{\Theta}, \xi_k) := \begin{bmatrix} \boldsymbol{g}_1^\top(\boldsymbol{\theta}_1, \xi_k) \\ \boldsymbol{g}_2^\top(\boldsymbol{\theta}_2, \xi_k) \\ \vdots \\ \boldsymbol{g}_M^\top(\boldsymbol{\theta}_M, \xi_k) \end{bmatrix} \in \mathbb{R}^{M\times p} \tag{24}$$

which admits the following compact representation

$$\boldsymbol{G}(\boldsymbol{\Theta}, \xi_k) = \boldsymbol{\Theta}\boldsymbol{H}^\top(\xi_k) + \boldsymbol{r}(k)\boldsymbol{\phi}^\top(s(k)) \tag{25}$$

where $\boldsymbol{r}(k) = [r_1(k)\ r_2(k)\ \cdots\ r_M(k)]^\top$ concatenates all local rewards. With the above definitions, the decentralized TD(0) updates in (22) over all agents can be collectively re-written as follows

$$\boldsymbol{\Theta}(k+1) = \boldsymbol{W}\boldsymbol{\Theta}(k) + \alpha\boldsymbol{G}(\boldsymbol{\Theta}(k), \xi_k). \tag{26}$$

In the sequel, we will investigate finite-sample analysis of the decentralized TD(0) learning algorithm in (26) in two steps. First, we will show that all local parameters reach a consensus, namely, converge to their average. Subsequently, we will prove that the average converges to the Bellman optimum $\boldsymbol{\theta}^*$.

To this end, let us define the average $\bar{\boldsymbol{\theta}} := (1/M)\boldsymbol{\Theta}^T\boldsymbol{1}$ of the parameter estimates by all agents, which can be easily shown using (26) to exhibit the following average system (AS) dynamics

$$\text{AS}: \quad \bar{\boldsymbol{\theta}}(k+1) = \bar{\boldsymbol{\theta}}(k) + \frac{\alpha}{M}\boldsymbol{G}^\top(\boldsymbol{\Theta}(k), \xi_k)\boldsymbol{1}. \tag{27}$$

Subtracting from each row of (26) (namely, each local parameter estimate) the average estimate in (27), yields

$$\boldsymbol{\Theta}(k+1) - \boldsymbol{1}\bar{\boldsymbol{\theta}}^\top(k+1) = \boldsymbol{W}\boldsymbol{\Theta}(k) - \boldsymbol{1}\bar{\boldsymbol{\theta}}^\top(k) + \alpha\Big(\boldsymbol{I} - \frac{\boldsymbol{1}\boldsymbol{1}^\top}{M}\Big)\boldsymbol{G}(\boldsymbol{\Theta}(k), \xi_k).$$

For notational convenience, we define the network difference operator $\boldsymbol{\Delta} := \boldsymbol{I} - (1/M)\boldsymbol{1}\boldsymbol{1}^\top$. Since $\boldsymbol{W}$ is a doubly stochastic matrix, it can be readily shown that $\boldsymbol{\Delta}\boldsymbol{\Theta} = \boldsymbol{\Theta} - \boldsymbol{1}\bar{\boldsymbol{\theta}}^\top$ capturing the difference between local estimates and the global average. After simple algebraic manipulations, we deduce that the parameter difference system (DS) evolves as follows

$$\text{DS}: \quad \boldsymbol{\Delta}\boldsymbol{\Theta}(k+1) = \boldsymbol{W}\boldsymbol{\Delta}\boldsymbol{\Theta}(k) + \alpha\boldsymbol{\Delta}\boldsymbol{G}(\boldsymbol{\Theta}(k)). \tag{28}$$

3 NON-ASYMPTOTIC PERFORMANCE GUARANTEES

The goal of this paper is to gain deeper understanding of statistical efficiency of decentralized TD(0) learning algorithms, and investigate their finite-time performance. We will start off by establishing convergence of the DS in (28), that is addressing consensus

among all agents. Formally, we have the following result. Proofs of the theorems, lemmas, and propositions can be found in the supplementary document.

Theorem 1. *Assume that all local rewards are uniformly bounded as $r_m(k) \in [0, r_{\max}]$, $\forall m \in \mathcal{M}$, and the feature vectors $\boldsymbol{\phi}(s)$ have been properly scaled such that $\|\boldsymbol{\phi}(s)\| \leq 1$, $\forall s \in \mathcal{S}$. For any deterministic initial guess $\boldsymbol{\Theta}(0)$ and any constant stepsize $0 < \alpha \leq (1 - \lambda_2^{\boldsymbol{W}})/4$, the parameter estimate difference over the network at any time instant $k \in \mathbb{N}_+$, satisfies*

$$\|\boldsymbol{\Delta\Theta}(k)\|_F \leq \left(\lambda_2^{\boldsymbol{W}} + 2\alpha\right)^k \|\boldsymbol{\Delta\Theta}(0)\|_F + \frac{2\alpha\sqrt{M} r_{\max}}{1 - \lambda_2^{\boldsymbol{W}}} \tag{29}$$

where $0 < \lambda_2^{\boldsymbol{W}} < 1$ denotes the second largest eigenvalue of $\boldsymbol{W}$.

Regarding Theorem 1, some remarks come in order.

To start, it is clear that the smaller $\lambda_2^{\boldsymbol{W}}$ is, the faster the convergence is. In practice, it is possible that the operator of the multi-agent system has the freedom to choose the weight matrix $\boldsymbol{W}$, so we can optimize the convergence rate by carefully designing $\boldsymbol{W}$. Furthermore, as the number k of updates grows large, the first term on the right-hand-side of (29) becomes negligible, implying that the parameter estimates of all agents converge to a small neighborhood of the global average $\bar{\boldsymbol{\theta}}(k)$, whose size is proportional to the constant stepsize $\alpha > 0$ (multiplied by a certain constant depending solely on the communication network).

So far, we have established the convergence of the DS. What remains is to show that the global average $\bar{\boldsymbol{\theta}}(k)$ converges to the optimal parameter value $\boldsymbol{\theta}^*$ [cf. (20)], which is equivalent to showing convergence of the AS in (27). In this paper, we investigate finite-time performance of decentralized TD(0) learning from data samples observed in two different settings, that is the i.i.d. setting as well as the Markovian setting, which occupy the ensuing two subsections.

3.1 The I.I.D. Setting

In the i.i.d. setting, we assume that data observations $\{(s(k), s(k+1), \{r_m(k)\}_{m\in\mathcal{M}})\}_{k\in\mathbb{N}^+}$ sampled along the trajectory of the underlying Markov chain are i.i.d.. Nevertheless, $s(k)$ and $s(k+1)$ are dependent within each data tuple. Indeed, the i.i.d. setting can be regarded as a special case of the Markovian setting detailed in the next subsection, after the Markov chain has reached a stationary distribution. To see this, consider the probability of the tuple $(s(k), s(k+1), r_m(k))$ taking any value $(s, s', r_m) \subseteq \mathcal{S} \times \mathcal{S} \times \mathbb{R}$

$$\Pr\{(s(k), s(k+1)) = (s, s')\} = \pi(s) P(s, s'). \tag{30}$$

An alternative way to obtain i.i.d. samples is to generate independently a number of trajectories and using first-visit methods; see details in [Bertsekas and Tsitsiklis, 1996].

With i.i.d. data samples, we can establish the following result which characterizes the relationship between $(1/M)\boldsymbol{G}^\top(\boldsymbol{\Theta}, \xi_j)\mathbf{1}$ and $\bar{\boldsymbol{g}}$.

Lemma 1. *Let $\{\mathcal{F}(k)\}_{k\in\mathbb{N}^+}$ be an increasing family of σ-fields, with $\boldsymbol{\Theta}(0)$ being $\mathcal{F}(0)$-measurable, and $\boldsymbol{G}(\boldsymbol{\Theta}(k), \xi_k)$ being $\mathcal{F}(k)$-measurable. The average $(1/M)\boldsymbol{G}^\top(\boldsymbol{\Theta}(k), \xi_k)\mathbf{1}$ of the gradient estimates at all agents is an unbiased estimate of $\bar{\boldsymbol{g}}(\bar{\boldsymbol{\theta}}(k))$; that is,*

$$\mathbb{E}_\pi\left[\frac{1}{M}\boldsymbol{G}^\top(\boldsymbol{\Theta}(k), \xi_k)\mathbf{1} - \bar{\boldsymbol{g}}(\bar{\boldsymbol{\theta}}(k))\Big|\mathcal{F}(k)\right] = \mathbf{0}, \forall \xi_k \tag{31}$$

and the variance satisfies

$$\mathbb{E}_\pi\left[\left\|\frac{1}{M}\boldsymbol{G}^\top(\boldsymbol{\Theta}(k), \xi_k)\mathbf{1} - \bar{\boldsymbol{g}}(\bar{\boldsymbol{\theta}}(k))\right\|^2 \Big|\mathcal{F}(k)\right] \leq 4\beta^2\|\bar{\boldsymbol{\theta}}(k) - \boldsymbol{\theta}^*\|^2 + 4\beta^2\|\boldsymbol{\theta}^*\|^2 + 8r_{\max}^2, \forall \xi_j \tag{32}$$

where β is the maximum spectral radius of matrices $\boldsymbol{H}(\xi_k) - \bar{\boldsymbol{H}}$ for all k.

This lemma suggests that $(1/M)\boldsymbol{G}^\top(\boldsymbol{\Theta}(k), \xi_j)\mathbf{1}$ is a noisy estimate of $\bar{\boldsymbol{g}}(\bar{\boldsymbol{\theta}}(k))$, and the noise is zero-mean and its variance depends only on $\bar{\boldsymbol{\theta}}(k)$. Evidently, the maximum spectral radius of $\boldsymbol{H}(\xi_k) - \bar{\boldsymbol{H}}$ can be upper bounded by $2(1+\gamma)$ using the definitions of $\boldsymbol{H}(\xi_k)$ in (9) and $\bar{\boldsymbol{H}}$ in (12).

We are now ready to state our main convergence result in the i.i.d. setting.

Theorem 2. *Letting $\lambda_{\max}^{\bar{\boldsymbol{H}}} < 0$ denote the largest eigenvalue of $\bar{\boldsymbol{H}}$ given in (12). For any constant stepsize $0 < \alpha \leq -\frac{\lambda_{\max}^{\bar{\boldsymbol{H}}}}{2[4\beta^2 + (\lambda_{\min}^{\bar{\boldsymbol{H}}})^2]}$, the average parameter estimate over all agents converges linearly to a small neighborhood of the equilibrium point $\boldsymbol{\theta}^*$; i.e.,*

$$\mathbb{E}\left[\left\|\bar{\boldsymbol{\theta}}(k) - \boldsymbol{\theta}^*\right\|^2\right] \leq c_1^k \left\|\bar{\boldsymbol{\theta}}(0) - \boldsymbol{\theta}^*\right\|^2 + c_2\alpha \tag{33}$$

where the constants $0 < c_1 := 1 + 2\alpha\lambda_{\max}^{\bar{\boldsymbol{H}}} + 8\alpha^2\beta^2 + 2\alpha^2(\lambda_{\min}^{\bar{\boldsymbol{H}}})^2 < 1$ and $c_2 := \frac{8\beta^2\|\boldsymbol{\theta}^*\|^2 + 16r_{\max}^2}{-\lambda_{\max}^{\bar{\boldsymbol{H}}}}$.

Particularly for the i.i.d. setting, the AS drives $\bar{\boldsymbol{\theta}}(k)$ to the optimal solution $\boldsymbol{\theta}^*$ as SGD does, which is indeed due to the fact that $(1/M)\boldsymbol{G}^\top(\boldsymbol{\Theta}(k), \xi_j)\mathbf{1}$ is an unbiased estimate of $\bar{\boldsymbol{g}}(\bar{\boldsymbol{\theta}}(k))$.

Putting together the convergence result of the global parameter estimate average in Theorem 2 as, well as, the established consensus among the multi-agents' parameter estimates in Theorem 1, it follows readily convergence of the local parameter estimates $\{\boldsymbol{\theta}_m\}_{m\in\mathcal{M}}$, summarized in the next proposition.

Proposition 1. *Choosing any constant stepsize* $0 < \alpha < \alpha_{\max} \triangleq \min\Big\{\frac{1-\lambda_2^{\boldsymbol{W}}}{4}, -\frac{\lambda_{\max}^{\bar{\boldsymbol{H}}}}{2[4\beta^2+(\lambda_{\min}^{\bar{\boldsymbol{H}}})^2]}\Big\}$*, then the decentralized TD(0) update in* (22) *guarantees that each local parameter estimate* $\boldsymbol{\theta}_m$ *converges linearly to a neighborhood of the optimum* $\boldsymbol{\theta}^*$*; that is,*

$$\mathbb{E}\Big[\big\|\boldsymbol{\theta}_m(k)-\boldsymbol{\theta}^*\big\|^2\Big] \le c_3^k V_0 + c_4\alpha, \quad \forall m\in\mathcal{M} \tag{34}$$

where the constants $c_3 := \max\{(\lambda_2^{\boldsymbol{W}}+2\alpha_{\max})^2, c_1\}$, $V_0 := 2\max\{4\|\boldsymbol{\Delta\Theta}(0)\|_F^2, 2\|\bar{\boldsymbol{\theta}}(0)-\boldsymbol{\theta}^*\|^2\}$, *and* $c_4 := \alpha_{\max}\frac{8M^2r_{\max}^2}{(1-\lambda_2^{\boldsymbol{W}})^2} + \frac{16\beta^2\|\boldsymbol{\theta}^*\|^2+32r_{\max}^2}{-\lambda_{\max}^{\bar{\boldsymbol{H}}}}$.

3.2 The Markovian Setting

Although the i.i.d. assumption on the data samples $\{(s(k), s(k+1), r_m(t))\}_k$ helps simplify the analysis of TD(0) learning, it represents only an ideal setting, and undermines the practical merits. In this subsection, we will consider a more realistic scenario, where data samples are collected along the trajectory of a single Markov chain starting from any initial distribution. For the resultant Markovian observations, we introduce an important result bounding the bias between the time-averaged 'gradient estimate' $\boldsymbol{G}(\boldsymbol{\Theta},\xi_k)$ and the limit $\bar{\boldsymbol{g}}(\bar{\boldsymbol{\theta}})$, where ξ_k captures all the randomness corresponding to the k-th transition $(s(k), s(k+1), \{r_m(k)\}_{m\in\mathcal{M}})$.

Lemma 2. *Let* $\{\mathcal{F}(k)\}_{k\in\mathbb{N}^+}$ *be an increasing family of* σ*-fields, with* $\boldsymbol{\Theta}(0)$ *being* $\mathcal{F}(0)$*-measurable, and* $\boldsymbol{G}(\boldsymbol{\Theta},\xi_k)$ *being* $\mathcal{F}(k)$*-measurable. Then, for any given* $\boldsymbol{\Theta}\in\mathbb{R}^p$ *and any integer* $j\in\mathbb{N}^+$*, the following holds*

$$\Big\|\frac{1}{KM}\sum_{j=k}^{k+K-1}\mathbb{E}\big[\boldsymbol{G}^\top(\boldsymbol{\Theta},\xi_j)\mathbf{1}\big|\mathcal{F}(k)\big]-\bar{\boldsymbol{g}}(\bar{\boldsymbol{\theta}})\Big\| \le \sigma_k(K)(\|\bar{\boldsymbol{\theta}}-\boldsymbol{\theta}^*\|+1). \tag{35}$$

where $\sigma_k(K) := \frac{(1+\gamma)\nu_0\rho^k}{(1-\rho)K}\times\max\{2\|\boldsymbol{\theta}^*\|+r_{\max},1\}$*, with constants* $\nu_0>0$ *and* $0<\rho<1$ *determined by the Markov chain. In particular for any* $k\in\mathbb{N}^+$*, it holds that* $\sigma_k(K)\le\frac{(1+\gamma)\nu_0}{(1-\rho)K}\times\max\big\{2\|\boldsymbol{\theta}^*\|+r_{\max},1\big\}\triangleq\sigma(K)$.

Comparing Lemma 2 with Lemma 1, the consequence on the update (26) due to the Markovian observations is elaborated in the following two remarks.

Remark 1. *In the Markovian setting, per time instant* $k\in\mathbb{N}$*, the term* $(1/M)\boldsymbol{G}^\top(\boldsymbol{\Theta}(k),\xi_k)\mathbf{1}$ *is a biased estimate of* $\bar{\boldsymbol{g}}(\bar{\boldsymbol{\theta}}(k))$*, but its time-averaged bias over a number of future consecutive observations can be upper bounded in terms of the estimation error* $\|\bar{\boldsymbol{\theta}}(k)-\boldsymbol{\theta}^*\|$.

Remark 2. *The results in Lemma 1 for i.i.d. samples correspond to requiring* $\sigma(K)=0$ *for all* $K\in\mathbb{N}^+$ *in Lemma 2. That is, the i.i.d. setting is indeed a special case of the Markovian one.*

In fact, due to the unbiased 'gradient' estimates under i.i.d. samples, we were able to directly investigate the convergence of $\bar{\boldsymbol{\theta}}(k)-\boldsymbol{\theta}^*$. In the Markovian setting however, since we have no control over the instantaneous gradient bias, it becomes challenging, if not impossible, to directly establish convergence of $\bar{\boldsymbol{\theta}}(k)-\boldsymbol{\theta}^*$ as dealt with in the i.i.d. setting. In light of the result on the bounded time-averaged gradient bias in Lemma 2, we introduce the following multi-step Lyapunov function that involves K future consecutive estimates $\{\bar{\boldsymbol{\theta}}(k)\}_{k=k_0}^{k_0+K-1}$:

$$\mathbb{V}(k) := \sum_{j=k}^{k+K-1}\big\|\bar{\boldsymbol{\theta}}(j)-\boldsymbol{\theta}^*\big\|^2, \quad k\in\mathbb{N}^+. \tag{36}$$

Concerning the multi-step Lyapunov function, we establish the following result.

Lemma 3. *Define the following functions*

$$\begin{aligned}\Gamma_1(\alpha,K) &= 32\alpha^3K^4(1+2\alpha)^{2K-4}+32K\alpha\\ &\quad+8\alpha K^2(1+2\alpha)^{K-2}+4K\sigma(K)\\ \Gamma_2(\alpha,K) &= \big[32\alpha^3K^4(1+2\alpha)^{2K-4}+32K\alpha\\ &\quad+\alpha K^2(1+2\alpha)^{K-2}\big]\|\boldsymbol{\theta}^*\|^2+\big[4\alpha^3K^4(1+2\alpha)^{2K-4}\\ &\quad+\frac{1}{2}\alpha K^2(1+2\alpha)^{K-2}+4\alpha K\big]r_{\max}^2+\frac{1}{2}K\sigma(K)\end{aligned}$$

There exists a pair of constants $(\alpha_{\max}, K_{\mathcal{G}})$ *such that* $0<1+2\alpha K_{\mathcal{G}}\lambda_{\max}^{\bar{\boldsymbol{H}}}+\alpha\Gamma_1(\alpha_{\max},K_{\mathcal{G}})<1$ *holds for any fixed* $\alpha\in(0,\alpha_{\max})$ *and* $K=K_{\mathcal{G}}$*. Moreover, the multi-step Lyapunov function satisfies*

$$\begin{aligned}&\mathbb{E}\big[\mathbb{V}(k+1)-\mathbb{V}(k)\big|\mathcal{F}(k)\big]\\ &\le\alpha\big[2K_{\mathcal{G}}\lambda_{\max}^{\bar{\boldsymbol{H}}}+\Gamma_1(\alpha_{\max},K_{\mathcal{G}})\big]\big\|\bar{\boldsymbol{\theta}}(k)-\boldsymbol{\theta}^*\big\|^2\\ &\quad+\alpha\Gamma_2(\alpha_{\max},K_{\mathcal{G}}).\end{aligned} \tag{37}$$

Here, we show by construction the existence of a pair $(\alpha_{\max}, K_{\mathcal{G}})$ meeting the conditions on the stepsize. Considering the monotonicity of function $\sigma(K)$, a simple choice for $K_{\mathcal{G}}$ is

$$K_{\mathcal{G}} = \min_K\Big\{K\big|\sigma(K)<-\frac{1}{4}\lambda_{\max}^{\bar{\boldsymbol{H}}}\Big\}. \tag{38}$$

Fixing $K=K_{\mathcal{G}}\ge1$, it follows that

$$2K\lambda_{\max}^{\bar{\boldsymbol{H}}}+\Gamma_1(\alpha,K)=\Gamma_0(\alpha,K_{\mathcal{G}}) \tag{39}$$

where $\Gamma_0(\alpha,K_{\mathcal{G}}) = 32\alpha^3K_{\mathcal{G}}^6(1+2\alpha)^{2K_{\mathcal{G}}-4}+32\alpha+8\alpha K_{\mathcal{G}}^3(1+2\alpha)^{K_{\mathcal{G}}-2}+K_{\mathcal{G}}\lambda_{\max}^{\bar{\boldsymbol{H}}}$ can be shown to be monotonically increasing in α. Considering further that $\Gamma_0(0,K_{\mathcal{G}})=K_{\mathcal{G}}\lambda_{\max}^{\bar{\boldsymbol{H}}}<0$, then there exist a stepsize α_0 such that $\Gamma_0(\alpha_0,K_{\mathcal{G}})=\frac{1}{2}K_{\mathcal{G}}\lambda_{\max}^{\bar{\boldsymbol{H}}}<0$ holds.

Setting now

$$\alpha_{\max} := \min\Big\{-\frac{1}{2K_{\mathcal{G}}\lambda_{\max}^{\bar{\boldsymbol{H}}}},\ \alpha_0\Big\} \tag{40}$$

then one can easily check that $0 < 1 + 2\alpha K\lambda_{\max}^{\bar{\boldsymbol{H}}} + \Gamma_1(\alpha, K) \leq 1 + \frac{1}{2}\alpha K_{\mathcal{G}}\lambda_{\max}^{\bar{\boldsymbol{H}}} < 1$ holds true for any constant stepsize $0 < \alpha < \alpha_{\max}$. In the remainder of this paper, we will work with $K = K_{\mathcal{G}}$ and $0 < \alpha < \alpha_{\max}$, yielding

$$\begin{aligned}\Gamma_0(0, K_{\mathcal{G}}) = K_{\mathcal{G}}\lambda_{\max}^{\bar{\boldsymbol{H}}} &\leq 2K_{\mathcal{G}}\lambda_{\max}^{\bar{\boldsymbol{H}}} + \Gamma_1(\alpha, K_{\mathcal{G}}) \\ &\leq \frac{1}{2}K_{\mathcal{G}}\lambda_{\max}^{\bar{\boldsymbol{H}}} \end{aligned} \tag{41}$$

where the first inequality uses the fact that $\Gamma_0(\alpha, K_{\mathcal{G}})$ is an increasing function of $\alpha > 0$, while the second inequality follows from the definition of α_0.

Before presenting the main convergence results in the Markovian setting, we provide a lemma that bounds the multi-step Lyapunov function along the trajectory of a Markov chain. This constitutes a building block for establishing convergence of the averaged parameter estimate.

Lemma 4. *The multi-step Lyapunov function is upper bounded as follows*

$$\mathbb{V}(k) \leq c_5\big\|\bar{\boldsymbol{\theta}}(k) - \boldsymbol{\theta}^*\big\|^2 + c_6\alpha^2, \quad \forall k \in \mathbb{N}^+ \tag{42}$$

where the constants c_5 and c_6 are given by

$$c_5 := \frac{(3 + 12\alpha_{\max}^2)^{K_{\mathcal{G}}} - 1}{2 + 3\alpha_{\max}^2}$$

$$c_6 := \frac{6(3 + 12\alpha_{\max}^2)\big[(3 + 12\alpha_{\max}^2)^{K_{\mathcal{G}}-1} - 1\big] - 6K_{\mathcal{G}} + 6}{2 + 12\alpha_{\max}^2} (4\|\boldsymbol{\theta}^*\|^2 + r_{\max}^2).$$

With the above two lemmas, we are now on track to state our convergence result for the averaged parameter estimate, in a Markovian setting.

Theorem 3. *Define constants* $c_7 := 1 + (1/2c_5)\alpha_{\max}K_{\mathcal{G}}\lambda_{\max}^{\bar{\boldsymbol{H}}} \in (0, 1)$, *and* $c_8' := \big[16\alpha_{\max}^2 K_{\mathcal{G}}^6(1 + 2\alpha_{\max})^{2K_{\mathcal{G}}-4} + 32K_{\mathcal{G}} + 2K_{\mathcal{G}}^3(1 + 2\alpha_{\max})^{K_{\mathcal{G}}-2}\big]\|\boldsymbol{\theta}^*\|^2 + 4K_{\mathcal{G}}r_{\max}^2 - \frac{1}{8}K_{\mathcal{G}}\lambda_{\max}^{\bar{\boldsymbol{H}}} - \frac{\alpha_{\max}c_6}{c_5}K_{\mathcal{G}}\lambda_{\max}^{\bar{\boldsymbol{H}}}$. *Then, fixing any constant stepsize $0 < \alpha < \alpha_{\max}$ and $K = K_{\mathcal{G}}$ defined in* (38), *the averaged parameter estimate $\bar{\boldsymbol{\theta}}(k)$ converges at a linear rate to a small neighborhood of the equilibrium point $\boldsymbol{\theta}^*$; that is,*

$$\begin{aligned}\mathbb{E}\Big[\big\|\bar{\boldsymbol{\theta}}(k) - \boldsymbol{\theta}^*\big\|^2\Big] \leq c_5 c_7^k\big\|\bar{\boldsymbol{\theta}}(0) - \boldsymbol{\theta}^*\big\|^2 - \frac{2c_5c_8'}{K_{\mathcal{G}}\lambda_{\max}^{\bar{\boldsymbol{H}}}}\alpha \\ + \min\big\{1, c_7^{k-k_\alpha}\big\}\Big(\alpha^2 c_6 - \frac{2c_5c_8'}{K_{\mathcal{G}}\lambda_{\max}^{\bar{\boldsymbol{H}}}}\Big)\end{aligned} \tag{43}$$

where $k_\alpha := \max\{k \in \mathbb{N}^+ | \rho^k \geq \alpha\}$.

As a direct result of Theorems 1 and 3, the convergence of all local parameter estimates comes ready.

Proposition 2. *Choosing a constant stepsize* $0 < \alpha < \min\big\{\alpha_{\max}, (1 - \lambda_2^{\boldsymbol{W}})/4\big\}$, *and any integer* $K \geq K_{\mathcal{G}}$, *each local parameter $\boldsymbol{\theta}_m(k)$ converges linearly to a neighborhood of the equilibrium point $\boldsymbol{\theta}^*$; that is, the following holds true for each $m \in \mathcal{M}$*

$$\begin{aligned}\mathbb{E}\Big[\big\|\boldsymbol{\theta}_m(k) - \boldsymbol{\theta}^*\big\|^2\Big] \leq c_9^k V_0' + \frac{8\alpha^2 M r_{\max}^2}{\big(1 - \lambda_2^{\boldsymbol{W}}\big)^2} - \frac{2c_5c_8'}{K_{\mathcal{G}}\lambda_{\max}^{\bar{\boldsymbol{H}}}}\alpha \\ + \min\big\{1, c_7^{k-k_\alpha}\big\}\Big(\alpha^2 c_6 - \frac{2c_5c_8'}{K_{\mathcal{G}}\lambda_{\max}^{\bar{\boldsymbol{H}}}}\Big)\end{aligned}$$

where the constants $c_9 := \max\{(\lambda_2^{\boldsymbol{W}} + 2\alpha_{\max})^2, c_7\}$, *and* $V_0' := 2\max\{4\|\boldsymbol{\Delta\Theta}(0)\|_F^2, 2c_5\|\bar{\boldsymbol{\theta}}(0) - \boldsymbol{\theta}^*\|^2\}$.

The proof is similar to that of Proposition 1, and hence is omitted. Proposition 2 establishes that even in a Markovian setting, the local estimates produced by decentralized TD(0) learning converge linearly to a neighborhood of the optimum. Interestingly, different than the i.i.d. case, the size of the neighborhood is characterized in two phases, which correspond to Phase I ($k \leq k_\alpha$), and Phase II ($k > k_\alpha$). In Phase I, the Markov is far from its stationary distribution π, giving rise to sizable gradient bias in Lemma 2, and eventually contributing to a constant-size neighborhood $-2c_5c_8'/(K_{\mathcal{G}}\lambda_{\max}^{\bar{\boldsymbol{H}}})$; while, after the Markov chain gets close to π in Phase II, confirmed by the geometric mixing property, we are able to have gradient estimates of size-$\mathcal{O}(\alpha)$ bias in Lemma 2, and the constant-size neighborhood vanishes with $c_7^{k-k_\alpha}$.

4 CONCLUSIONS

In this paper, we studied the dynamics of a decentralized linear function approximation variant of the vanilla TD(0) learning, for estimating the value function of a given policy. We proved that such decentralized TD(0) algorithms converge linearly to a small neighborhood of the optimum, under both i.i.d. data samples as, well as, the realistic Markovian observations collected along the trajectory of a single Markov chain. To address the 'gradient bias' in a Markovian setting, our novel approach has been leveraging a carefully designed multi-step Lyapunov function to enable a unique two-phase non-asymptotic convergence analysis. Comparing with previous contributions, this paper provides the first finite-sample error bound for fully decentralized TD(0) learning under challenging Markovian observations.

5 AKNOWLEDGEMENT

The work by J. Sun and Z. Yang was supported in part by NSFC Grants 61873118, 61673347, and the Dept. of Science and Technology of Guangdong Province under Grant 2018A050506003. The work by J. Sun was also supported by the China Scholarship Council. The work by G. Wang and G. B. Giannakis was supported in part by NSF grants 1711471, and 1901134. The work of Q. Yang was supported in part by NSFC grants U1609214, 61751205, and the Key R&D Program of Zhejiang Province under Grant 2019C01050.

References

[Baird, 1995] Baird, L. (1995). Residual algorithms: Reinforcement learning with function approximation. In *International Conference on Machine Learning*, pages 30–37.

[Bertsekas and Tsitsiklis, 1996] Bertsekas, D. P. and Tsitsiklis, J. N. (1996). *Neuro-Dynamic Programming*, volume 5. Athena Scientific Belmont, MA.

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

[Bhatnagar et al., 2009] Bhatnagar, S., Precup, D., Silver, D., Sutton, R. S., Maei, H. R., and Szepesvári, C. (2009). Convergent temporal-difference learning with arbitrary smooth function approximation. In *Advances in Neural Information Processing Systems*, pages 1204–1212.

[Bof et al., 2018] Bof, N., Carli, R., and Schenato, L. (2018). Lyapunov theory for discrete time systems. *arXiv:1809.05289*.

[Borkar, 2008] Borkar, V. S. (2008). *Stochastic Approximation: A Dynamical Systems Viewpoint*, volume 48. Cambridge, New York, NY.

[Chi et al., 2019] Chi, Y., Lu, Y. M., and Chen, Y. (2019). Nonconvex optimization meets low-rank matrix factorization: An overview. *IEEE Transactions on Signal Processing*, 67(20):5239–5269.

[Dalal et al., 2018] Dalal, G., Szörényi, B., Thoppe, G., and Mannor, S. (2018). Finite sample analyses for TD(0) with function approximation. In *AAAI Conference on Artificial Intelligence*, pages 6144–6152.

[Doan et al., 2019] Doan, T., Maguluri, S., and Romberg, J. (2019). Finite-time analysis of distributed TD(0) with linear function approximation on multi-agent reinforcement learning. In *International Conference on Machine Learning*, pages 1626–1635.

[Gupta et al., 2019] Gupta, H., Srikant, R., and Ying, L. (2019). Finite-time performance bounds and adaptive learning rate selection for two time-scale reinforcement learning. *Advances in Neural Information Processing Systems*.

[Hu and Syed, 2019] Hu, B. and Syed, U. A. (2019). Characterizing the exact behaviors of temporal difference learning algorithms using Markov jump linear system theory. *Advances in Neural Information Processing Systems*.

[Krishnamurthy et al., 2008] Krishnamurthy, V., Maskery, M., and Yin, G. (2008). Decentralized adaptive filtering algorithms for sensor activation in an unattended ground sensor network. *IEEE Transactions on Signal Processing*, 56(12):6086–6101.

[Levin and Peres, 2017] Levin, D. A. and Peres, Y. (2017). *Markov Chains and Mixing Times*, volume 107. American Mathematical Society.

[Ma et al., 2019] Ma, M., Li, B., and Giannakis, G. B. (2019). Tight linear convergence rate of ADMM for decentralized optimization. *arXiv:1905.10456*.

[Mnih et al., 2015] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540):529.

[Nedić et al., 2018] Nedić, A., Olshevsky, A., and Rabbat, M. G. (2018). Network topology and communication-computation tradeoffs in decentralized optimization. *IEEE Trans. Automat. Control.*, 106(5):953–976.

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

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

[Sutton et al., 2009] Sutton, R. S., Maei, H. R., and Szepesvári, C. (2009). A convergent $o(n)$ temporal-difference algorithm for off-policy learning with linear function approximation. In *Advances in neural information processing systems*, pages 1609–1616.

[Tsitsiklis and Van Roy, 1997] Tsitsiklis, J. N. and Van Roy, B. (1997). An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, 42(5):674–690.

[Wang et al., 2019a] Wang, G., Giannakis, G. B., and Chen, J. (2019a). Learning ReLU networks on linearly separable data: Algorithm, optimality, and generalization. *IEEE Transactions on Signal Processing*, 67(9):2357–2370.

[Wang et al., 2019b] Wang, G., Li, B., and Giannakis, G. B. (2019b). A multistep Lyapunov approach for finite-time analysis of biased stochastic approximation. *arXiv:1909.04299*.

[Xu et al., 2020] Xu, T., Wang, Z., Zhou, Y., and Liang, Y. (2020). Reanalysis of variance reduced temporal difference learning. *arXiv:2001.01898*.

[Yang et al., 2019] Yang, Q., Wang, G., Sadeghi, A., Giannakis, G. B., and Sun, J. (2019). Two-timescale voltage control in distribution grids using deep reinforcement learning. *IEEE Transactions on Smart Grid*, pages 1–11.