One-Shot Averaging for Distributed TD(λ) Under Markov Sampling

Haoxing Tian[1], Ioannis Ch. Paschalidis[2] and Alex Olshevsky[2]

***Abstract*— We consider a distributed setup for reinforcement learning, where each agent has a copy of the same Markov Decision Process but transitions are sampled from the corresponding Markov chain independently by each agent. We show that in this setting, we can achieve a linear speedup for TD(λ), a family of popular methods for policy evaluation, in the sense that N agents can evaluate a policy N times faster provided the target accuracy is small enough. Notably, this speedup is achieved by "one shot averaging," a procedure where the agents run TD(λ) with Markov sampling independently and only average their results after the final step. This significantly reduces the amount of communication required to achieve a linear speedup relative to previous work.**

I. INTRODUCTION

Actor-critic method achieves state-of-the-art performance in many domains including robotics, game playing, and control systems ([1], [2], [3]). Temporal Difference (TD) Learning may be thought of as a component of actor critic, and better bounds for TD Learning are usually ingredients of actor-critic analyses. We consider the problem of policy evaluation in reinforcement learning: given a Markov Decision Process (MDP) and a policy, we need to estimate the value of each state (expected discounted sum of all future rewards) under this policy. Policy evaluation is important because it is effectively a subroutine of many other algorithms such as policy iteration and actor-critic. The main challenges for policy evaluation are that we usually do not know the underlying MDP directly and can only interact with it, and that the number of states is typically too large forcing us to maintain a low-dimensional approximation of the true vector of state values.

We focus here on the simplest class of methods overcoming this set of challenges, namely TD methods with linear function approximation. These methods try to maintain a low dimensional parameter which is continuously updated based on observed rewards and transitions to maintain consistency of estimates across states. The proof of convergence for these methods was first given in [4].

In this paper, we focus on the multi-agent version of policy evaluation: we consider N agents that have a copy of the same MDP and the same policy, but transitions in the MDP by different agents are independent. The question we wish to ask is whether the agents can cooperate to evaluate the underlying policy N times faster, since now N transitions are generated per unit time.

Although there is some previous work on distributed temporal difference methods (e.g., [5], [6], [7]), this question has only been considered in the recent papers [8], [9], [10], [11]. The answer was positive in both [8], [9] in a "federated learning" setting, provided the nodes have N rounds of communication with a central server before time T, with environment heterogeneity additionally considered in [9]. In [10], the answer was also positive (i.e., linear speedup was obtained) under a distributed erasure model where each node communicated with neighbors in a graph a constant fraction of time, leading to $O(T)$ communications in T steps. Our previous work [11] established that, in fact, only one communication round with a central server is sufficient in the case of i.i.d. observations and TD(0), the most basic method within the temporal difference family. This was accomplished via the "one-shot averaging" methods where the N agents just ignore each other for T steps, and then simply average their results. Further, the final averaging step could be replaced with $O(\log T)$ rounds of an average consensus method.

The i.i.d. observation assumption is a limiting feature of our previous work in [11]: it is assumed that at each time, we can generate a random state from the underlying MDP. This is convenient for analysis but rarely satisfied in practice.

In this paper, our contribution is to show that one-shot averaging suffices to give a linear speedup *without* the i.i.d. assumption and for the more general class of temporal difference methods TD(λ) (precise definitions are given later). Our method of proof is new and does not overlap with the arguments given in our previous work.

II. BACKGROUND

A. Markov Decision Process (MDP)

A finite discounted-reward MDP can be described by a tuple $(S, A, P_{\text{env}}, r, \gamma)$, where S is the state-space whose elements are vectors, with s_0 being the starting state; A is the action space; $P_{\text{env}} = (P_{\text{env}}(s'|s,a))_{s,s' \in S, a \in A}$ is the transition probability matrix, where $P_{\text{env}}(s'|s,a)$ is the probability of transitioning from s to s' after taking action a; $r : S \times S \to \mathbb{R}$ is the reward function, where $r(s,s')$ associates a deterministic reward with each state transition; and $\gamma \in (0,1)$ is the discount factor.

A policy π is a mapping $\pi : S \times A \to [0,1]$ where $\pi(a|s)$ is the probability that action a is taken in state s. Given a

Research partially supported by the NSF under grants CCF-2200052 and IIS-1914792, by the ONR under grants N00014-19-1-2571 and N00014-21-1-2844, by the DOE under grant DE-AC02-05CH11231, by the NIH under grant UL54 TR004130, by ARPA-E under grant DE-AR0001282, and by the Boston University Kilachand Fund for Integrated Life Science and Engineering.

[1]Department of Electrical Engineering, Boston University, Boston, MA, USA tianhx@bu.edu

[2]Department of Electrical Engineering and Division of System Engineering, Boston University, Boston, MA, USA {yannisp, alexols}@bu.edu.

policy π, the state transition matrix $P_\pi = (P_\pi(s'|s))_{s,s'\in S}$ and the state reward function $r_\pi(s)$ is defined as

$$P_\pi(s'|s) = \sum_{a\in A} P_{\text{env}}(s'|s,a)\pi(a|s),\ r_\pi(s) = \sum_{s'\in S} P_\pi(s'|s)r(s,s').$$

Since the policy is fixed throughout the paper, we will omit the subscript π and thus write $P(s'|s)$ and $r(s)$ instead of $P_\pi(s'|s)$ and $r_\pi(s)$.

The stationary distribution μ is a nonnegative vector with coordinates summing to one and satisfying $\mu^T = \mu^T P$. The Perron-Frobenius theorem guarantees that such a stationary distribution exists and is unique subject to some conditions on P, e.g., aperiodicity and irreducibility [12]. The entries of μ are denoted by $\mu(s)$. We also define $D = \text{diag}(\mu(s))$ as the diagonal matrix whose elements in the main diagonal are given by the entries of the stationary distribution μ.

The value function $V_\pi^*(s)$ is defined as $V_\pi^*(s) = \mathbb{E}_\pi\left[\sum_{t=0}^{+\infty}\gamma^t r(s_t)\right]$, where $\mathbb{E}_\pi$ stands for the expectation when actions are chosen according to policy π. Since the MDPs is finite, it is without loss of generality to assume a bound on rewards.

Assumption 2.1: For any $s, s' \in S\times S$, $|r(s,s')| \le r_{\max}$.

B. Value Function Approximation

Given a fixed policy, the problem is to efficiently estimate V_π^*. We consider a linear function approximation architecture $V_{\pi,\theta}(s) = \phi(s)^T\theta$, where $\phi(s) \in \mathbb{R}^d$ is a feature vector for state s and $\theta \in \mathbb{R}^d$. Without loss of generality, we assume $||\phi(s)|| \le 1$ for all states s. For simplicity, we define $V_{\pi,\theta} = (V_{\pi,\theta}(s))_{s\in S}$ to be a column vector, $\Phi = \left(\phi(s)^T\right)_{s\in S}$ to be a $|S| \times d$ matrix and $R = (r(s))_{s\in S}$ to be a column vector. We are thus trying to approximate $V_{\pi,\theta} \approx \Phi\theta$. We make the following assumptions:

Assumption 2.2 (Input assumption): It is standard to assume the following statements hold:

- The matrix Φ has linearly independent columns.
- The stationary distribution $\mu(s) > 0, \forall s \in S$.

With this assumption, if ω is the smallest eigenvalue of $\Sigma := \Phi^T D\Phi$, then $\omega > 0$ and $\omega\|x\|^2 \le x^T\Sigma x$.

C. Distributed Model and Algorithm

We assume that there are N agents and each agent shares the same tuple $(S, A, P_{\text{env}}, r, \gamma)$ as well as the same fixed policy π. However, each agent independently samples its trajectories and updates its own version of a parameter θ_t.

We will study an algorithm which mixes TD learning and one-shot averaging: after all agents finish T steps, they share their information and compute the average parameter as the final result. These agents do not communicate before the final step. The averaging can take place using average consensus (using any average consensus algorithm) or, in a federated manner, using a single communication with a coordinator.

We next spell out the details of our algorithm. Every agent runs TD(0) with Markov sampling as follows. Agent i generates an initial state $s_i(0)$ from some initial distribution. It also maintains an iterate $\theta_i(t)$, initialized arbitrarily. At time t, agent i generates a transition according to P. It then computes the so-called TD-error

$$\delta_{s,s'}(\theta_i(t)) = r(s,s') + \gamma\phi(s')^T\theta_i(t) - \phi(s)^T\theta_i(t),$$

with $s = s_i(t), s' = s_i(t+1)$ coming from the transition it just generated; and then updates

$$\theta_i(t+1) = \theta_i(t) + \alpha_t g_{s_i(t),s_i(t+1)}(\theta_i(t)) \tag{1}$$

where the update direction is $g_{s,s'}(\theta) = \delta_{s,s'}(\theta)\phi(s)$. At the end, the agents average their results.

We will further use $\bar{g}$ to denote the expectation of $g_{s,s'}$ assuming s are sampled from the stationary distribution μ and s' is generated by taking a step from P. We use $\mathbb{E}_{\text{I}}$ to denote this expectation. Therefore, $\bar{g}(\theta) = \mathbb{E}_{\text{I}}\left[g_{s,s'}(\theta)\right]$. We can also rewrite $\bar{g}$ in matrix notation: $\bar{g}(\theta) = \Phi^T D(R + (\gamma P - I)\Phi\theta)$.

In order to perform our analysis, we need to define the stationary point. We adopt the classic way of defining such point as shown in [4]. We call θ^* the stationary point if $\bar{g}(\theta^*) = 0$. According to [13], in matrix notation, it is equivalent to say that θ^* satisfies the following:

$$\Phi^T D\left(R + (\gamma P - I)\Phi\theta^*\right) = 0. \tag{2}$$

Naturally, each agent can easily compute θ^* by running TD(0) for infinite times and simply ignore all the other agents. However, this ignores the possibility that agents can benefit from communicating with each other.

We next focus on TD(λ), which is a popular generalization of the conceptually simpler TD(0) and attains better performance with an appropriate choice of λ [14]. For any fixed $\lambda \in [0,1]$, TD(λ) executes the update

$$\theta_i(t+1) = \theta_i(t) + \alpha_t x_{s_i(t),s_i(t+1)}(\theta_i(t), z_{0:t}). \tag{3}$$

Here, $z_{0:t} = \sum_{k=0}^{t}(\gamma\lambda)^k\phi(s_i(t-k))$ is called eligibility trace and $x_{s_i(t),s_i(t+1)}(\theta_i(t), z_{0:t})$ is defined as

$$x_{s_i(t),s_i(t+1)}(\theta_i(t), z_{0:t}) = \delta_{s_i(t),s_i(t+1)}(\theta_i(t))z_{0:t}.$$

For convenience of the analysis, we define the eligibility trace going back to minus infinity:

$$z_{-\infty:t} = \lim_{t\to\infty} z_{0:t} = \sum_{k=0}^{\infty}(\gamma\lambda)^k\phi(s_i(t-k)).$$

We also introduce the operator $T_\pi^{(\lambda)}$ which is defined as

$$\left(T_\pi^{(\lambda)}V\right)(s) = (1-\lambda)\sum_{k=0}^{\infty}\lambda^k$$
$$\cdot\,\mathbb{E}\left[\sum_{t=0}^{k}\gamma^t r(s_i(t), s_i(t+1)) + \gamma^{k+1}V(s_i(k+1)) \mid s_i(0) = s\right].$$

We also rewrite the above equation in matrix notation:

$$T_\pi^{(\lambda)}V = (1-\lambda)\sum_{k=0}^{\infty}\lambda^k\left(\sum_{t=0}^{k}\gamma^t P^t R + \gamma^{k+1}P^{k+1}V\right). \tag{4}$$

We denote $\bar{x}(\theta_i(t)) = \mathbb{E}_{\text{I}}\left[x_{s,s'}(\theta_i(t), z_{-\infty:t})\right]$. This expectation assumes that $s_i(k) \sim \mu, \forall k$, and that the history of

the process then extends to $-\infty$ according to a distribution consistent with each forward step being taken by P; for more details, see [4]. We also note that Lemma 8 in [4] implies $\bar{x}(\theta_t) = \Phi^T D\left(T_\pi^{(\lambda)}(\Phi\theta) - \Phi\theta\right)$.

We call θ^* the stationary point if $\bar{x}(\theta^*) = 0$. In matrix notation, it is equivalent to say that θ^* satisfies the following:

$$\Phi^T D\left(T_\pi^{(\lambda)}(\Phi\theta^*) - \Phi\theta^*\right) = 0. \tag{5}$$

Finally, we define $\kappa = \gamma\frac{1-\lambda}{1-\gamma\lambda}$ which will be useful later. An obvious result is that $\kappa \leq \gamma$.

D. Markov Sampling and Mixing

As mentioned in the previous section, the ideal way of generating $s_i(t)$ is to draw from stationary distribution μ. However, the typical way is generating a trajectory is $s_i(1), \ldots, s_i(T)$. Every state in this trajectory is sampled by taking a transition $s_i(t) \sim P(\cdot|s_i(t-1))$ (and recall our policy is always fixed). This way of sampling is called Markov sampling, and we denote $\mathbb{E}_{\mathrm{M}}$ as the expectation under Markov sampling. Analyzing algorithms under Markov sampling can be challenging since one cannot ignore the dependency on previous samples. The following "uniform mixing" assumption is standard [15]. We also note that this assumption always holds for irreducible and aperiodic Markov chains [16].

Assumption 2.3: There are constants m and ρ such that

$$||P(s(t) \in \cdot|s(0)) - \mu||_1 \leq m\rho^t, \quad \forall t.$$

A key definition from the uniform mixing assumption is called the mixing time. We define the mixing time $\tau_{\min}(\epsilon)$:

$$\tau_{\mathrm{mix}}(\epsilon) = \min\{t \mid m\rho^t \leq \epsilon\}.$$

In this paper, we always set $\epsilon = \alpha_t$ which is the step-size at time t, typically $\alpha_t = \beta/(c+t)$, and simplify $\tau_{\mathrm{mix}}(\alpha_t)$ as τ_{mix}. An obvious result is that

$$m\rho^t \leq \alpha_t, \quad \forall t \geq \tau_{\mathrm{mix}}.$$

E. Convergence times for centralized TD(0) and TD(λ)

We now state the state-of-the-art results for the centralized case which are based on using ideas from gradient descent to analyze TD(0) and TD(λ). These results considered Projected TD Learning, where $\theta(t)$ is projected onto a ball of fixed radius after the update is performed. We will use these results as a basis for comparison for our distributed results.

Lemma 2.1 ([15]): In Projected TD(0) with Markov sampling, suppose Assumptions 2.2, 2.3 hold. For a decaying stepsize sequence $\alpha_t = 2/(\omega(t+1)(1-\gamma))$,

$$\mathbb{E}\left[||\theta_i(T) - \theta^*||^2\right] \leq \nu_{\mathrm{central}} \sim O\left(\frac{(\log T)^2}{T}\right).$$

We next discuss convergence times for TD(λ). It is usually assumed that the algorithm extends back to negative infinity, and that every $s_i(t)$ has distribution μ (but the samples are, of course, correlated since each successive state is obtained by taking a step in the Markov chain P from the previous one). Similarly as before, we define $\tau_{\mathrm{mix}}^{(\lambda)}(\epsilon)$ as

$$\tau_{\mathrm{mix}}^{(\lambda)}(\epsilon) = \max\{\tau_{\mathrm{mix}}(\epsilon), \tau'_{\mathrm{mix}}(\epsilon)\},$$

where $\tau_{\mathrm{mix}}(\epsilon) = \min\{t \mid m\rho^t \leq \epsilon\}$ and $\tau'_{\mathrm{mix}}(\epsilon) = \min\{t \mid (\gamma\lambda)^t \leq \epsilon\}$. As before, we choose $\epsilon = \alpha_t$ and simplify $\tau_{\mathrm{mix}}^{(\lambda)}(\alpha_t)$ as $\tau_{\mathrm{mix}}^{(\lambda)}$. The same result applies to $\tau_{\mathrm{mix}}^{(\lambda)}$:

$$\max\{m\rho^t, (\gamma\lambda)^t\} \leq \alpha_t, \quad \forall t \geq \tau_{\mathrm{mix}}^{(\lambda)}(\alpha_t).$$

With these notations in place, we can now state the following result from the previous literature.

Lemma 2.2 ([15]): In Projected TD(λ) with the Markov sampling, suppose Assumptions 2.2, 2.3 hold. For a decaying stepsize sequence $\alpha_t = 2/(\omega(t+1)(1-\kappa))$,

$$\mathbb{E}\left[||\theta_i(T) - \theta^*||^2\right] \leq \nu_{\mathrm{central}}^{(\lambda)} \sim O\left(\frac{(\log T)^2}{T}\right).$$

III. Main Result

We now state our main result which claims a linear speed-up for both distributed TD(0) and distributed TD(λ). Recall that we use $\theta_i(t)$ to denote the parameters of agent i at time t, $\bar{\theta}(t) = (\sum_i \theta_i(t))/N$ to denote the averaged parameters among all N agents, and $\bar{\theta}_i(t) = \mathbb{E}[\theta_i(t)]$ to denote the expectation of $\theta_i(t)$. We now have the following two theorems for TD(0) and TD(λ) respectively. Notice that $\tilde{O}$ omits logarithm factors.

Theorem 3.1: Suppose Assumptions 2.2 and 2.3 hold. Denote $t_0 = \max\{\tau_{\mathrm{mix}}, \frac{8}{\omega\omega_{\mathrm{I}}(1-\gamma)} - 1\}$. With ν_{central} in Lemma 2.1. TD(0) with $\alpha_t = 2/(\omega(t+1)(1-\gamma))$ satisfies

$$\mathbb{E}\left[||\bar{\theta}(T+t_0) - \theta^*||^2\right] \leq \frac{1}{N}\nu_{\mathrm{central}} + \tilde{O}\left(\frac{1}{T^2}\right), \forall T \geq 0.$$

Theorem 3.2: Suppose Assumptions 2.2 and 2.3 hold. Denote $t_0^{(\lambda)} = \max\{2\tau_{\mathrm{mix}}^{(\lambda)}, \frac{8}{\omega\omega_{\mathrm{I}}^{(\lambda)}(1-\kappa)} - 1\}$. With $\nu_{\mathrm{central}}^{(\lambda)}$ in Lemma 2.2, TD(λ) with $\alpha_t = 2/(\omega(t+1)(1-\kappa))$ satisfies

$$\mathbb{E}\left[||\bar{\theta}(T+t_0) - \theta^*||^2\right] \leq \frac{1}{N}\nu_{\mathrm{central}}^{(\lambda)} + \tilde{O}\left(\frac{1}{T^2}\right), \forall T \geq 0.$$

In brief, the distributed version with N nodes is N times faster than the comparable centralized version for large enough T (note that v_{central} and $v_{\mathrm{central}}^{(\lambda)}$ are $\tilde{O}(1/T)$ whereas the term that does not get divided by N is $\tilde{O}(1/T^2)$ in both theorems). This significantly improves previous results from [11], which only showed this for TD(0).

We note that the proofs given in this paper have no overlaps with the proof from [11], which could not be extended to either TD(λ) or Markov sampling (and here both extensions are done simultaneously). Instead, our analysis here is based on the following simple observation, at each step of which we just use independence plus elementary

algebra:

$$\begin{aligned}
&\mathbb{E}\left[||\bar{\theta}(T)-\theta^*||^2\right]\\
\leq&\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\|\theta_i(T)-\theta^*\|^2\right]\\
&+\frac{2}{N^2}\sum_{1\leq i<j\leq N}\mathbb{E}\left[(\theta_i(T)-\theta^*)^T(\theta_j(T)-\theta^*)\right] \qquad (6)\\
=&\frac{1}{N^2}\sum_{i=1}^{N}\mathbb{E}\left[\|\theta_i(T)-\theta^*\|^2\right]\\
&+\frac{2}{N^2}\sum_{1\leq i<j\leq N}\left(\bar{\theta}_i(T)-\theta^*\right)^T\left(\bar{\theta}_j(T)-\theta^*\right).
\end{aligned}$$

Here, the last equality uses the fact that $\theta_i(T)$ are independent of each other since there is no communication during learning. This immediately implies a linear speed-up if only we could prove that the first term dominates. This is quite plausible, since the second term involves the convergence speed of the *expected* updates. In other words, all that is really needed is to prove that the expected update converges faster than the unexpected update.

IV. Proof of Theorem 3.1

Before we go into the proof, we first introduce some notations. We rewrite $\bar{g}(\theta)$ in matrix notation as

$$\bar{g}(\theta)=\Phi^T D(I-\gamma P)\Phi\left(\theta^*-\theta\right):=\Sigma_{\mathrm{I}}\left(\theta^*-\theta\right). \qquad (7)$$

The matrix Σ_{I} has some nice properties, which is pointed out in [17], [18]. Indeed, we have the following lemmas whose proof we postpone:

Lemma 4.1: There exists $\omega_{\mathrm{I}}\geq(1-\gamma)\omega>0$ such that $\inf_{\|x\|=1}x^T\Sigma_{\mathrm{I}}x\geq\omega_{\mathrm{I}}$.

Lemma 4.2: For any x, $\|\Sigma_{\mathrm{I}}x\|^2\leq 4\|x\|^2$.

We define $\bar{g}'(\theta_i(t))=\mathbb{E}_{\mathrm{M}}\left[g_{s,s'}(\theta_i(t))\right]$ where $\mathbb{E}_{\mathrm{M}}$ is defined in Section II-D. We call $\bar{g}'(\theta_i(t))-\bar{g}(\theta_i(t))$ Markov noise and write it as

$$\begin{aligned}
&\bar{g}'(\theta_i(t))-\bar{g}(\theta_i(t))\\
=&\sum_{s_i(t),s_i'(t)}\left(P_t(s_i(t)|s_i(0))-\mu(s_i(t))\right)\phi(s_i(t))\\
&\cdot\left(r(t)-\left(\phi(s_i(t))-\gamma P(s_i'(t)|s_i(t))\phi(s_i'(t))\right)\theta_i(t)\right),
\end{aligned}$$

where $P_t(s'|s)$ stands for the t step transition probability. To further address both the Markov noise and the recursion relations we will derive, we need the following lemmas, whose proofs we also postpone.

Lemma 4.3: In TD(0), $||\theta^*||\leq R:=r_{\max}/\omega_{\mathrm{I}}$.

Lemma 4.4: For a sequence of numbers $\{x_t\}$ and three constants a,b,c such that $a>1$, we have the following recursive inequality:

$$x_{t+1}\leq\left(1-\frac{a}{c+t}\right)x_t+\frac{b^2}{(c+t)^2},\quad\forall t\geq\tau.$$

Then we have the following result:

$$x_t\leq\frac{\nu}{c+t},\text{ where }\nu=\max\left\{\frac{b^2}{a-1},(c+\tau)x_\tau\right\}.$$

Given all preliminaries, we can begin our proof.

Proof: [Proof of Theorem 3.1]

We will omit the subscript i since the analysis holds for all $i\in\{1,2,\ldots,n\}$. For simplicity, denote $\Delta_t=\bar{\theta}(t)-\theta^*$. Letting $t=T+t_0$, we take expectation on both sides in (1),

$$\Delta_{t+1}=\Delta_t+\alpha_t\mathbb{E}\left[\bar{g}(\theta(t))\right]+\alpha_t\mathbb{E}\left[\bar{g}'(\theta(t))-\bar{g}(\theta(t))\right].$$

By (7), we know that $\mathbb{E}\left[\bar{g}(\theta(t))\right]=-\Sigma_{\mathrm{I}}\Delta_t$. Notice that, Assumption 2.2 immediately implies $\|\Sigma_{\mathrm{I}}\|\leq 1$. Therefore,

$$\begin{aligned}
&\|(I-\alpha_t\Sigma_{\mathrm{I}})\Delta_t\|^2\\
=&\|\Delta_t\|^2-\alpha_t\Delta_t^T(\Sigma_{\mathrm{I}}^T+\Sigma_{\mathrm{I}})\Delta_t+\alpha_t^2\Delta_t^T\Sigma_{\mathrm{I}}^T\Sigma_{\mathrm{I}}\Delta_t\\
\leq&(1-2\omega_{\mathrm{I}}\alpha_t+4\alpha_t^2)\|\Delta_t\|^2\\
\leq&(1-\omega_{\mathrm{I}}\alpha_t)\|\Delta_t\|^2.
\end{aligned}$$

Here we use Lemma 4.1, 4.2 and the fact that $4\alpha_t\leq\omega_{\mathrm{I}}$ (recall we assume $t\geq t_0$). This immediately implies

$$\|(I-\alpha_t\Sigma_{\mathrm{I}})\Delta_t\|\leq(1-\omega_{\mathrm{I}}\alpha_t/2)\|\Delta_t\|,$$

where we use the fact $\sqrt{1-x}\leq 1-x/2$. Therefore,

$$\|\Delta_{t+1}\|\leq(1-\omega_{\mathrm{I}}\alpha_t/2)\|\Delta_t\|+\alpha_t\mathbb{E}\left[\|\bar{g}'(\theta(t))-\bar{g}(\theta(t))\|\right].$$

To address the second term on the right-hand side, by Assumption 2.3, for all $t\geq t_0$,

$$\begin{aligned}
&\mathbb{E}\left[\|\bar{g}'(\theta(t))-\bar{g}(\theta(t))\|\right]\\
\leq&\mathbb{E}\Big[\left(r_{\max}+2\|\theta(t)-\theta^*\|+2\|\theta^*\|\right)\cdot\|P_t(\cdot|s(0))-\mu\|_1\Big]\\
\leq&\alpha_t\left(r_{\max}+2u+2R\right).
\end{aligned}$$

where u is defined as $u=\max_{i,t}\mathbb{E}\|\theta_i(t)-\theta^*\|$. Notice that Lemma 2.2 guarantees u is finite. This immediately indicates

$$||\Delta_{t+1}||\leq(1-\omega_{\mathrm{I}}\alpha_t/2)\,||\Delta_t||+\alpha_t^2\left(r_{\max}+2u+2R\right).$$

We set $x_t=\|\Delta_t\|, a=\omega_{\mathrm{I}}/(\omega(1-\gamma)), b^2=r_{\max}+2u+2R, c=1$ and $\tau=t_0$. By Lemma 4.4,

$$||\Delta_t||\leq\frac{\nu}{1+t},\quad\nu=\max\{\alpha,\beta\}$$

where

$$\alpha=\frac{r_{\max}+2u+2R}{\frac{\omega_{\mathrm{I}}}{\omega(1-\gamma)}-1},\quad\beta=(1+t_0)\|\Delta_{t_0}\|.$$

Since all the above facts holds for every agent i, the result directly follows after plugging the above fact as well as Lemma 2.1 into (6). ■

A. Proof of Lemma 4.1

Proof: Based on [17], [18], one can show that

$$x^T\Sigma_I x=(1-\gamma)\sum_{s\in S}\mu(s)y(s)^2+\gamma\sum_{s,s'\in S}\mu(s)P(s'|s)(y(s')-y(s))^2.$$

Here, $x\in\mathbb{R}^d$ is an arbitrary vector and $y=\Phi x\in\mathbb{R}^{|S|}$, whereas $x(s),y(s)$ is the entry of x,y corresponding to the state s. Then, it is obvious that $\omega_{\mathrm{I}}\geq(1-\gamma)\omega$. ■

B. Proof of Lemma 4.2

Proof: According to the definition of $\mathbb{E}_{\mathrm{I}}$ before (2), $\Sigma_{\mathrm{I}} x = \mathbb{E}_{\mathrm{I}}\left[(\phi(s)^T x - \gamma\phi(s')^T x)\phi(s)\right]$. Therefore,

$$\|\Sigma_{\mathrm{I}} x\|^2 \leq \mathbb{E}_{\mathrm{I}}\left[\left\|(\phi(s)^T x - \gamma\phi(s')^T x)\phi(s)\right\|^2\right] \leq 4\|x\|^2$$

where we use $\|\phi(s)\| \leq 1$ and $\gamma \leq 1$. ∎

C. Proof of Lemma 4.3

Proof: By the Gershgorin circle theorem, $D(\gamma P - I)$ is invertible, and thus so is $\Phi^T D(\gamma P - I)\Phi$. By (2),

$$\theta^* = \left[\Phi^T D(I - \gamma P)\Phi\right]^{-1}\Phi^T DR.$$

By Lemma 5.9 in [19] and Lemma 4.1, $\|\Sigma_{\mathrm{I}}\|^{-1} \leq \omega_{\mathrm{I}}^{-1}$. Furthermore, since $\|\Phi^T\sqrt{D}\|^2 \leq 1$ which is because all features vectors have norm at most one by assumption, and $\|\sqrt{D}R\|^2 \leq r_{\max}^2$, we obtain $||\theta^*|| \leq r_{\max}/\omega_{\mathrm{I}}$. ∎

D. Proof of Lemma 4.4

Proof: We prove it by induction. First, it is easy to see $x_\tau \leq \frac{\nu}{c+\tau}$. Now suppose $x_t \leq \frac{\nu}{c+t}$,

$$\begin{aligned} x_{t+1} \leq& (1 - a\cdot\alpha_t)\, x_t + \frac{b^2}{(c+t)^2} \\ \leq& \left(1 - \frac{a}{c+t}\right)\frac{\nu}{c+t} + \frac{b^2}{(c+t)^2} \\ =& \frac{c+t-1}{(c+t)^2}\nu + \frac{(1-a)\nu + b^2}{(c+t)^2} \\ \leq& \frac{1}{c+t+1}\nu, \end{aligned}$$

where the last inequality uses the facts that $x^2 \geq (x-1)(x+1), \forall x$ and $(1-a)\nu + b^2 \leq 0$ (This is because of the definition of ν as defined in Lemma 4.4). ∎

V. Proof of Theorem 3.2

Our proof here follows the same strategy as for TD(0). For simplicity, we define

$$\Sigma_I^{(\lambda)} := \Phi^T D\Phi - (1-\lambda)\sum_{k=0}^{\infty}\lambda^k\gamma^{k+1}\Phi^T DP^{k+1}\Phi.$$

One could use (4) and (5) to show that

$$\bar{x}(\theta) = \Phi^T D\left(T_\pi^{(\lambda)}(\Phi\theta) - \Phi\theta\right) = \Sigma_I^{(\lambda)}(\theta - \theta^*). \quad (8)$$

Inspired by [17], we claim the following lemmas whose proof we postpone:

Lemma 5.1: These exists $\omega_{\mathrm{I}}^{(\lambda)} \geq (1-\kappa)\omega > 0$ such that $\inf_{\|x\|=1} x^T\Sigma_{\mathrm{I}}^{(\lambda)}x \geq \omega_{\mathrm{I}}^{(\lambda)}$.

Lemma 5.2: For any x, $\|\Sigma_{\mathrm{I}}^{(\lambda)}x\|^2 \leq 4\|x\|^2$.

As before, we denote $\bar{x}'(\theta_i(t)) = \mathbb{E}_{\mathrm{M}}\left[x_{s,s'}(\theta_i(t), z_{0:t})\right]$. This expectation assumes that $s_i(0) \sim \mu$ while the subsequent states are sampled according to the transition probability of the policy, i.e., $s_i(k) \sim P(\cdot|s_i(k-1))$. We call the quantity $\bar{x}'(\theta_i(t)) - \bar{x}(\theta_i(t))$ Markov noise.

Finally, for the stationary point defined in (5), we have the following lemma whose proof is also postponed:

Lemma 5.3: In TD(λ), $\|\theta^*\| \leq R^{(\lambda)} := \frac{r_{\max}}{\omega_{\mathrm{I}}^{(\lambda)}(1-\gamma)}$.

Now we are ready to begin the proof.

Proof: [Proof of Theorem 3.2]

We will omit the subscript i since the analysis holds for all $i \in \{1, 2, \ldots, n\}$. For simplicity, denote $\Delta_t = \bar{\theta}(t) - \theta^*$. Letting $t = T + t_0$, we take expectation on both sides in (3),

$$\Delta_{t+1} = \Delta_t + \alpha_t\mathbb{E}\left[\bar{x}(\theta(t))\right] + \alpha_t\left(\mathbb{E}\left[\bar{x}'(\theta(t)) - \bar{x}(\theta(t))\right]\right).$$

By (8), we have $\mathbb{E}\left[\bar{x}(\theta(t))\right] = \Sigma_{\mathrm{I}}^{(\lambda)}\Delta_t$. Notice that

$$\begin{aligned} &\|(I - \alpha_t\Sigma_{\mathrm{I}}^{(\lambda)})\Delta_t\|^2 \\ =&\|\Delta_t\|^2 - \alpha_t\Delta_t^T({\Sigma_{\mathrm{I}}^{(\lambda)}}^T + \Sigma_{\mathrm{I}}^{(\lambda)})\Delta_t + \alpha_t^2\Delta_t^T{\Sigma_{\mathrm{I}}^{(\lambda)}}^T\Sigma_{\mathrm{I}}^{(\lambda)}\Delta_t \\ \leq&(1 - \omega_{\mathrm{I}}^{(\lambda)}\alpha_t)\|\Delta_t\|^2. \end{aligned}$$

Here we use both Lemma 5.1, 5.2 and the fact that $4\alpha_t \leq \omega_{\mathrm{I}}^{(\lambda)}$ (recall we assume $t \geq t_0^{(\lambda)}$). This immediately implies

$$\|(I - \alpha_t\Sigma_{\mathrm{I}}^{(\lambda)})\Delta_t\| \leq (1 - \omega_{\mathrm{I}}^{(\lambda)}\alpha_t/2)\|\Delta_t\|,$$

where we use the fact $\sqrt{1-x} \leq 1 - x/2$. Therefore,

$$\|\Delta_{t+1}\| \leq (1-\omega_{\mathrm{I}}^{(\lambda)}\alpha_t/2)\|\Delta_t\| + \alpha_t\mathbb{E}\left[\|\bar{x}'(\theta(t)) - \bar{x}(\theta(t))\|\right].$$

To deal with the Markov noise, for simplicity, we write

$$\begin{aligned} \bar{x}(\theta) =& \sum_{k=0}^{+\infty}(\gamma\lambda)^k\sum_{s_i(t-k)}\mu(s_i(t-k))\phi(s_i(t-k))l_{t-k}(\theta) \\ \bar{x}'(\theta) =& \sum_{k=0}^{t}(\gamma\lambda)^k\sum_{s_i(t-k)}P_{t-k}(s_i(t-k)|s_i(0))\phi(s_i(t-k))l_{t-k}(\theta), \end{aligned}$$

where $l_{t-k}(\theta) = \mathbb{E}_{s\sim P_k(\cdot|s_i(t-k)), s'\sim P(\cdot|s)}\left[\delta_{s,s'}(\theta)\right]$. Let $l_0 = |l_{t-k}(\theta)|/(1-\gamma\lambda)$. A simple bound for l_0 is

$$l_0 \leq \frac{r_{\max} + 2u^{(\lambda)} + 2R^{(\lambda)}}{1-\gamma\lambda},$$

where we both use Lemma 5.3 and denote $u^{(\lambda)} = \max_{i,t}\mathbb{E}\|\theta_i(t) - \theta^*\|$. Notice that Lemma 2.2 guarantees that $u^{(\lambda)}$ is finite. With these notations, we have

$$\begin{aligned} &\bar{x}(\theta) - \bar{x}'(\theta) \\ =&\sum_{k=0}^{t}(\gamma\lambda)^k\sum_{s(t-k)}\left[\mu - P_{t-k}(\cdot|s(0))\right]_{s(t-k)}\phi(s(t-k))l_{t-k}(\theta) \\ &+\sum_{k=t+1}^{+\infty}(\gamma\lambda)^k\sum_{s(t-k)}\mu(s(t-k))\phi(s(t-k))l_{t-k}(\theta). \end{aligned}$$

We denote the first term as $I_{0:t}$ and divide it into two terms, $I_{0:\tau_{\mathrm{mix}}^{(\lambda)}}$ and $I_{\tau_{\mathrm{mix}}^{(\lambda)}+1:t}$. By Assumption 2.3,

$$\|P_{t-k}(\cdot|s(0)) - \mu\|_1 \leq \alpha_t, \quad \forall t \leq \tau_{\mathrm{mix}}^{(\lambda)},$$

where we use the fact $t \geq 2\tau_{\mathrm{mix}}^{(\lambda)}$. Therefore,

$$\begin{aligned} \mathbb{E}\left[\|I_{0:\tau_{\mathrm{mix}}^{(\lambda)}}\|\right] \leq& \mathbb{E}\left[|l_{t-k}(\theta)|\sum_{k=0}^{\tau_{\mathrm{mix}}^{(\lambda)}}(\gamma\lambda)^k\cdot\|P_{t-k}(\cdot|s(0)) - \mu\|_1\right] \\ \leq& \mathbb{E}\left[l_0\right]\alpha_t. \end{aligned}$$

By the definition of $\tau_{\text{mix}}^{(\lambda)}$, $(\gamma\lambda)^t \leq \alpha_t$. Therefore,

$$\begin{aligned}\mathbb{E}\left[\|I_{\tau_{\text{mix}}^{(\lambda)}+1:t}\|\right] \leq & \mathbb{E}\left[|l_{t-k}(\theta)| \sum_{k=\tau_{\text{mix}}^{(\lambda)}+1}^{t} (\gamma\lambda)^k \left\|P_{t-k}(\cdot|s(0)) - \mu\right\|_1\right] \\ \leq & 2\mathbb{E}\left[l_0\right]\alpha_t.\end{aligned}$$

The second term (under expectation) also has upper-bound $\mathbb{E}\left[l_0\right]\alpha_t$ since $(\gamma\lambda)^t \leq \alpha_t$ and the remaining terms are bounded by $\mathbb{E}\left[l_0\right]$. So far, we have

$$\begin{aligned}\|\Delta_{t+1}\| = & (1-\alpha_t\omega_{\text{I}}^{(\lambda)}/2)\|\Delta_t\| \\ & + \alpha_t^2 \frac{4}{1-\gamma\lambda} \cdot \left(r_{\max} + 2u^{(\lambda)} + 2R^{(\lambda)}\right),\end{aligned}$$

We set $x_t = \|\Delta_t\|, a = \omega_{\text{I}}^{(\lambda)}/(\omega(1-\kappa)), b^2 = 4\left(r_{\max} + 2u^{(\lambda)} + 2R^{(\lambda)}\right)/(1-\gamma\lambda), c = 1$ and $\tau = t_0^{(\lambda)}$. By Lemma 4.4,

$$||\Delta_t|| \leq \frac{\nu^{(\lambda)}}{1+t}, \quad \nu^{(\lambda)} = \max\left\{\alpha^{(\lambda)}, \beta^{(\lambda)}\right\}$$

where

$$\alpha^{(\lambda)} = \frac{4r_{\max} + 8u^{(\lambda)} + 8R^{(\lambda)}}{(\frac{\omega_{\text{I}}^{(\lambda)}}{\omega(1-\kappa)} - 1)(1-\gamma\lambda)}, \quad \beta^{(\lambda)} = (1+t_0^{(\lambda)})\|\Delta_{t_0}\|.$$

Since all the above facts hold for every agent i, the result directly follows after plugging the above fact as well as Lemma 2.2 into (6).

■

A. Proof of Lemma 5.1

Proof: As pointed out in Theorem 2 in [17], $x^T\Sigma_{\text{I}}^{(\lambda)}x$ equals to a convex combination of D-norm and Dirichlet semi-norm. Since Dirichlet semi-norm is always no less than zero, we have

$$x^T\Sigma_{\text{I}}^{(\lambda)}x \geq (1-\kappa)x^T\Phi^T D\Phi x \geq (1-\kappa)\omega\|x\|^2, \quad \forall x \in \mathbb{R}^d.$$

■

B. Proof of Lemma 5.2

Proof: According to the definition of $\mathbb{E}_{\text{I}}$ after (4),

$$\Sigma_{\text{I}}^{(\lambda)}x = \mathbb{E}_{\text{I}}\left[\left(\phi(s_0)^T x - (1-\lambda)\sum_{k=0}^{\infty}\lambda^k\gamma^{k+1}\phi(s_{k+1})^T x\right)\phi(s_0)\right].$$

Since $\|\phi(s)\| \leq 1$ and $\kappa \leq 1$,

$$\|\Sigma_{\text{I}}^{(\lambda)}x\| \leq (1+(1-\lambda)\sum_{k=0}^{\infty}\lambda^k\gamma^{k+1})\|x\| = (1+\kappa)\|x\| \leq 2\|x\|.$$

■

C. Proof of Lemma 5.3

Proof: Solving for θ^* using (4) and (5), we arrive at

$$\theta^* = \left(\Sigma_{\text{I}}^{(\lambda)}\right)^{-1}\Phi^T D(1-\lambda)\sum_{k=0}^{\infty}\lambda^k\sum_{t=0}^{k}\gamma^t P^t R.$$

By Lemma 5.9 in [19] and Lemma 5.1, $\|(\Sigma_{\text{I}}^{(\lambda)})^{-1}\| \leq {\omega_{\text{I}}^{(\lambda)}}^{-1}$. Furthermore, since $\|\Phi^T\sqrt{D}\|^2 \leq 1$ which is because all features vectors have norm at most one by assumption, and $\|\sqrt{D}(1-\lambda)\sum_{k=0}^{\infty}\lambda^k\sum_{t=0}^{k}\gamma^t P^t R\|^2 \leq r_{\max}^2/(1-\gamma)^2$, we obtain $\|\theta^*\| \leq \frac{r_{\max}}{\omega_{\text{I}}^{(\lambda)}(1-\gamma)}$.

■

CONCLUSION

We have shown that one-shot averaging suffices to give a linear speedup for distributed TD(λ) under Markov sampling. This is an improvement over previous works, which had alternatively either $O(T)$ communication rounds per T steps or $O(N)$ averaging rounds per T steps to achieve the same. An open question is whether a similar result can be proven for tabular Q-learning.

REFERENCES

[1] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[2] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. Asynchronous methods for deep reinforcement learning. In *International conference on machine learning*, pages 1928–1937. PMLR, 2016.

[3] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

[4] John Tsitsiklis and Benjamin Van Roy. Analysis of temporal-diffference learning with function approximation. *Advances in Neural Information Processing Systems*, 9, 1996.

[5] Thinh Doan, Siva Maguluri, and Justin Romberg. 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. PMLR, 2019.

[6] Jun Sun, Gang Wang, Georgios B Giannakis, Qinmin Yang, and Zaiyue Yang. Finite-time analysis of decentralized temporal-difference learning with linear function approximation. In *International Conference on Artificial Intelligence and Statistics*, pages 4485–4495. PMLR, 2020.

[7] Gang Wang, Songtao Lu, Georgios Giannakis, Gerald Tesauro, and Jian Sun. Decentralized TD tracking with linear function approximation and its finite-time analysis. *Advances in Neural Information Processing Systems*, 33:13762–13772, 2020.

[8] Sajad Khodadadian, Pranay Sharma, Gauri Joshi, and Siva Theja Maguluri. Federated reinforcement learning: Linear speedup under Markovian sampling. In *International Conference on Machine Learning*, pages 10997–11057. PMLR, 2022.

[9] Han Wang, Aritra Mitra, Hamed Hassani, George J Pappas, and James Anderson. Federated temporal difference learning with linear function approximation under environmental heterogeneity. *arXiv preprint arXiv:2302.02212*, 2023.

[10] Chenyu Zhang, Han Wang, Aritra Mitra, and James Anderson. Finite-time analysis of on-policy heterogeneous federated reinforcement learning. *arXiv preprint arXiv:2401.15273*, 2024.

[11] Rui Liu and Alex Olshevsky. Distributed TD(0) with almost no communication. *IEEE Control Systems Letters*, 2023.

[12] FR Gantmacher. The theory of matrices. *New York*, 1964.

[13] Dimitri Bertsekas and John N Tsitsiklis. *Neuro-dynamic programming*. Athena Scientific, 1996.

[14] Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

[15] Jalaj Bhandari, Daniel Russo, and Raghav Singal. A finite time analysis of temporal difference learning with linear function approximation. In *Conference on learning theory*, pages 1691–1692. PMLR, 2018.

[16] David A Levin and Yuval Peres. *Markov chains and mixing times*, volume 107. American Mathematical Soc., 2017.

[17] Rui Liu and Alex Olshevsky. Temporal difference learning as gradient splitting. In *International Conference on Machine Learning*, pages 6905–6913. PMLR, 2021.

[18] Haoxing Tian, Ioannis C. Paschalidis, and Alex Olshevsky. On the performance of temporal difference learning with neural networks. In *The Eleventh International Conference on Learning Representations*, 2022.

[19] Alex Olshevsky and Bahman Gharesifard. A small gain analysis of single timescale actor critic. *SIAM Journal on Control and Optimization*, 61(2):980–1007, 2023.