DISTRIBUTED TD(0) WITH ALMOST NO COMMUNICATION

Rui Liu
Division of Systems Engineering
Boston University
Boston, MA, 02215
rliu@bu.edu

Alex Olshevsky
Department of ECE and Division of Systems Engineering
Boston University
Boston, MA, 02215
alexols@bu.edu

January 31, 2022

ABSTRACT

We provide a new non-asymptotic analysis of distributed TD(0) with linear function approximation. Our approach relies on "one-shot averaging," where N agents run local copies of TD(0) and average the outcomes only once at the very end. We consider two models: one in which the agents interact with an environment they can observe and whose transitions depends on all of their actions (which we call the global state model), and one in which each agent can run a local copy of an identical Markov Decision Process, which we call the local state model.

In the global state model, we show that the convergence rate of our distributed one-shot averaging method matches the known convergence rate of TD(0). By contrast, the best convergence rate in the previous literature showed a rate which, according to the worst-case bounds given, could underperform the non-distributed version by $O(N^3)$ in terms of the number of agents N. In the local state model, we demonstrate a version of the linear time speedup phenomenon, where the convergence time of the distributed process is a factor of N faster than the convergence time of TD(0). As far as we are aware, this is the first result rigorously showing benefits from parallelism for temporal difference methods.

1 Introduction

Recent years have seen reinforcement learning used in a variety of multi-agent systems, for example, cooperative control (Wang et al., 2020b), traffic control (Kuyer et al., 2008; Bazzan, 2009), networked robotics (Yang and Gu, 2004; Duan et al., 2016), and bidding and advertising (Jin et al., 2018). However, a rigorous understanding of how standard methods in reinforcement learning perform in a multi-agent setting with limited communication is only partially available.

One of the most fundamental problems in reinforcement learning is policy evaluation, and one of the most basic policy evaluation algorithms is temporal difference (TD) learning, originally proposed in Sutton (1988). TD learning works by updating a value function from differences in predictions over a succession of steps in the underlying Markov Decision Process (MDP).

Developments in the field of multi-agent reinforcement learning (MARL) have led to an increased interest in decentralizing TD methods, which is the subject of this paper. We will consider two different MARL settings. These are described formally below, but, in brief, the "global state" setting considers a collection of agents interacting with an environment which takes actions depending on the actions of all the agents, which may have different rewards; and the "local state" setting, involves each agent having its own copy of the same MDP. In both settings, the goal is to find a policy maximizing the average of the discounted reward streams.

1.1 Related Literature

Analysis of centralized TD algorithm: A natural benchmark to compare the performance of distributed TD methods to is the performance of centralized TD methods. In the context of linear function approximation, these date to Jaakkola et al. (1994); Tsitsiklis and Van Roy (1997). Precise conditions for the asymptotic convergence result was first given in

Tsitsiklis and Van Roy (1997) by viewing TD as a stochastic approximation for solving a Bellman equation. Recently, there has been an increased interest in non-asymptotic convergence results, e.g., Dalal et al. (2018a); Lakshminarayanan and Szepesvari (2018); Bhandari et al. (2018). The state of the art results show that, under i.i.d samples, TD algorithm with linear function approximation converge as fast as $O(1/\sqrt{T})$ for value function with step-size $1/\sqrt{T}$ and converge as fast as $O(1/t)$ with step-size $O(1/t)$ (Bhandari et al., 2018).

Analysis of distributed TD methods: There has been several recent non-asymptotic analyses of distributed TD with linear function approximation. All of these were in the global state model informally described above. The first paper on the subject Doan et al. (2019a) proposed a distributed variant of the TD algorithm by combining consensus step and local TD updates for the update of each agent. For distributed TD algorithm with projection step and i.i.d samples, Doan et al. (2019a) provided a $O((\log T)/\sqrt{T})$ convergence rate for value function with step-size that scales as $1/\sqrt{T}$ and $O((\log t)/t)$ convergence rate for optimal value with step-size $O(1/t)$.

However, the constants in these bounds scaled with the spectral gap of the matrix underlying inter-agent communications. It is mentioned in Doan et al. (2019b) that these could be as bad as $O(N^3)$ in terms of the number of agents N. This is of the main issues we will try to address in this paper: we would like to derive bounds that either benefit from, or are not hurt by, the multi-agent nature of the system. That is, we would like bounds that either do not get worse with N or are actually improved as N becomes large.

In Sun et al. (2020), the case of both i.i.d and Markov samples was studied using a Lyapunov approach. It was shown that all local estimates converge linearly to a small neighborhood of the optimum with constant step-size. The size of the neighborhood scaled with the inverse of the underlying spectral gap of a matrix based on the pattern of inter-agent communications (and thus, indirectly, with the number of agents for many communication patterns); this was removed in Wang et al. (2020a) using a more sophisticated "gradient-tracking" approach, which involves communicating twice as much information per each step.

Finally, we also mention Shen et al. (2020) although that work deals with actor-critic rather than temporal difference methods. It is shown there, up to a certain approximation error, it is possible to obtain a linear speedup for a distributed model of actor-critic with independent samples across agents; this is in the same spirit as what we are attempting to do in part of this work.

1.2 Our Contribution

This paper provides a simple scheme for both the global state model studied in the previous literature and the local state model we introduce here under i.i.d. sampling. In contrast to previous papers which required communication between neighbors at every step of the underlying methods, our schemes requires only one global average computation at the very end.

We show that we are able to replicate the standard bounds for TD(0) in the global state model, including convergence to the same limit θ^* without any dependence on the number of agents. In the local state model, we show that we can improve state-of-the-art convergence times for TD(0): in other words, there is a benefit from parallelism in the local state setting. In particular, to the extent that the variance of the temporal difference error enters the convergence bounds for the local state model, it can be divided by the number of agents N.

2 Preliminaries

We begin by standardizing notation and providing standard background information on Markov Decision Processes and temporal difference methods.

2.1 Markov Decision Processes

A discounted reward MDP is described by a 5-tuple $(\mathcal{S},\mathcal{A},\mathcal{P},r,\gamma)$, where $\mathcal{S}=[n]=\{1,2,\cdots,n\}$ is a finite state space, $\mathcal{A}$ is a finite action space, $\mathcal{P}(s'|s,a):\mathcal{S}\times\mathcal{A}\times\mathcal{S}\to[0,1]$ is transition probability from s to s' determined by a, $r(s,a,s'):\mathcal{S}\times\mathcal{A}\times\mathcal{S}\to\mathbb{R}$ are deterministic rewards and $\gamma\in(0,1)$ is the discount factor.

Let μ denote a fixed policy that maps a state $s\in\mathcal{S}$ to a probability distribution $\mu(\cdot|s)$ over the action space $\mathcal{A}$, so that $\sum_{a\in\mathcal{A}}\mu(a|s)=1$. For such a fixed policy μ, define the instantaneous reward vector $R^\mu:\mathcal{S}\to\mathbb{R}$ as

$$R^\mu(s)=\sum_{s'\in\mathcal{S}}\sum_{a\in\mathcal{A}}\mu(a|s)\mathcal{P}(s'|s,a)r(s,a,s').$$

Fixing the policy μ induces a probability transition matrix between states:

$$P^{\mu}(s,s') = \sum_{a\in\mathcal{A}} \mu(a|s)\mathcal{P}(s'|s,a).$$

We will use $r_t = r(s_t, a_t, s_{t+1})$ to denote the instantaneous reward at time t, where s_t, a_t are the state and action taken at step t. The value function of μ, denoted by $V^{\mu} : \mathcal{S} \to \mathbb{R}$ is defined as

$$V^{\mu}(s) = E_{\mu,s}\left[\sum_{t=0}^{\infty} \gamma^t r_t\right], \tag{1}$$

where $E_{\mu,s}[\cdot]$ indicates that s is the initial state and the actions are chosen according to the policy μ. In the following, we will treat V^{μ} and R^{μ} as vectors in $\mathbb{R}^n$ and treat P^{μ} as a matrix in $\mathbb{R}^{n\times n}$.

Next, we state a standard assumptions on the underlying Markov chain.

Assumption 1. *The Markov chain with transition matrix P^{μ} is irreducible and aperiodic.*

A consequence of Assumption 1 is that there exists a unique stationary distribution $\pi = (\pi_1, \pi_2, \cdots, \pi_n)$, a row vector whose entries are positive and sum to 1. This stationary distribution satisfies $\pi^T P^{\mu} = \pi^T$ and $\pi_{s'} = \lim_{t\to\infty}(P^{\mu})^t(s,s')$ for any two states $s, s' \in \mathcal{S}$. *Note that we use π to denote the stationary distribution and μ to denote the policy.*

We next provide definitions of two norms that we will have occasion to use later. For a positive definite matrix $A \in \mathbb{R}^{n\times n}$, we define the inner product $\langle x, y\rangle_A = x^T A y$ and the associated norm $\|x\|_A = \sqrt{x^T A x}$ respectively. Since the numbers π_s are positive for all $s \in \mathcal{S}$, then the diagonal matrix $D = \mathrm{diag}(\pi_1, \cdots, \pi_n) \in \mathbb{R}^{n\times n}$ is positive definite. Therefore, for any two vectors $V, V' \in \mathbb{R}^n$, we can also define an inner product as

$$\left\langle V, V'\right\rangle_D = V^T D V' = \sum_{s\in\mathcal{S}} \pi_s V(s) V'(s),$$

and the associated norm as

$$\|V\|_D^2 = V^T D V = \sum_{s\in\mathcal{S}} \pi_s V(s)^2. \tag{2}$$

Finally, we introduce the definition of Dirichlet seminorm, following the notation of Ollivier (2018):

$$\|V\|_{\mathrm{Dir}}^2 = \frac{1}{2}\sum_{s,s'\in\mathcal{S}} \pi_s P^{\mu}(s,s')(V(s') - V(s))^2. \tag{3}$$

Note that Dirichlet seminorm depends both on the transition matrix P^{μ} and the stationary distribution π. Similarly, we introduce the k-step Dirichlet seminorm:

$$\|V\|_{\mathrm{Dir},k}^2 = \frac{1}{2}\sum_{s,s'\in\mathcal{S}} \pi_s (P^{\mu})^k(s,s')(V(s') - V(s))^2.$$

2.2 Temporal Difference Learning

Evaluating the value function V^{μ} of a policy can be computationally expensive when the number of states is very large. The classical TD algorithm uses low dimensional approximation V_{θ}^{μ}. For brevity, we will omit the superscript μ throughout from now on.

We next introduce the update rule of the classical temporal difference method with linear function approximation V_{θ}, a linear function of θ:

$$V_{\theta}(s) = \sum_{l=1}^{K} \theta_l \phi_l(s) \quad \forall s \in \mathcal{S}, \tag{4}$$

where $\phi_l = (\phi_l(1), \cdots, \phi_l(n))^T \in \mathbb{R}^n$ for $l \in [K]$ are K given feature vectors. Together, all K feature vectors form a $n \times K$ matrix $\Phi = (\phi_1, \cdots, \phi_K)$. For $s \in \mathcal{S}$, let $\phi(s) = (\phi_1(s), \cdots, \phi_K(s))^T \in \mathbb{R}^K$ denote the s-th row of matrix Φ, a vector that collects the features of state s. Then, Eq. (4) can be written in a compact form $V_{\theta}(s) = \theta^T \phi(s)$.

The TD(0) method maintains a parameter $\theta(t)$ which is updated at every step to improve the approximation. Supposing that we observe a sequence of states $\{s(t)\}_{t\in\mathbb{N}_0}$, then the classical TD(0) algorithm updates as:

$$\theta(t+1) = \theta(t) + \alpha_t \delta(t) \phi(s(t)), \tag{5}$$

where $\{\alpha_t\}_{t\in\mathbb{N}_0}$ is the sequence of step-sizes, and letting $s'(t)$ denote the next state after $s(t)$, the quantity $\delta(t)$ is the temporal difference error

$$\delta(t) = r(t) + \gamma\theta^T(t)\phi(s'(t)) - \theta^T(t)\phi(s(t)). \tag{6}$$

A common assumption on feature vectors in the literature (Tsitsiklis and Van Roy, 1997; Bhandari et al., 2018) is that features are linearly independent and uniformly bounded, which is formally given next.

Assumption 2. *The matrix Φ has full column rank, i.e., the feature vectors $\{\phi_1,\ldots,\phi_K\}$ are linearly independent. Additionally, we have that $\|\phi(s)\|_2^2 \leq 1$ for $s \in \mathcal{S}$.*

Under Assumption 1 and 2, we introduce the steady-state feature covariance matrix $\Phi^T D\Phi$. That this is a positive definite matrix as an immediate consequence of Assumptions 1 and 2, and we let $\omega > 0$ be a lower bound on its smallest eigenvalue.

We will use the fact, shown in Tsitsiklis and Van Roy (1997), that under Assumptions 1-2 as well as an additional assumption on the decay of the step-sizes α_t, the sequence of iterates $\{\theta_t\}$ generated by TD(0) learning converges almost surely a vector satisfying a certain projected Bellman equation; we will use θ^* to refer to this vector.

2.3 Two Models of Distributed Temporal Difference Methods

We now introduce two distributed models, which we will refer to as the *global state* and *local state* models. The global state model was previously introduced in Doan et al. (2019a) and studied in Wang et al. (2020a); Sun et al. (2020). We are not aware of the local state model being considered in the previous literature.

2.3.1 The Global State Model

The problem is characterized by the 6-tuple $(\mathcal{S},\mathcal{V},\mathcal{A}_\mathcal{V},\mathcal{P},r_\mathcal{V},\gamma)$. Here, $\mathcal{S}$, $\mathcal{P}$, and γ have the same definition as before; $\mathcal{V} = [N] = \{1,\ldots,N\}$ is the set of agents; $\mathcal{A}_\mathcal{V} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_N$ is the set of joint actions, where $\mathcal{A}_v$ is a set of actions only available to agent v; and $r_\mathcal{V} = \{r_v\}_{v\in\mathcal{V}}$ is a set of reward functions, where $r_v(s,a,s') : \mathcal{S}\times\mathcal{A}\times\mathcal{S} \to \mathbb{R}$ is the reward function of agent v.

The policy μ now takes the form $\mu(a|s) = \prod_{v=1}^N \mu_v(a_v|s)$, where $\mu_v(a_v|s)$ is the probability to select action $a_v \in \mathcal{A}_v$ when in state s. After the action $a(t) = \{a_1(t),\cdots,a_N(t)\}$, the system moves to a new state $s'(t)$ with probability $\mathcal{P}(s'(t)|s(t),a(t))$; then all the agents observe the new state and obtain local rewards $r_v(t) = r_v(s_v(t),a(t),s'_v(t))$.

The value of a policy will depend on the average of the rewards obtained by the individual agents:

$$V(s) = E_{\mu,s}\left[\sum_{t=0}^{\infty}\frac{\gamma^t}{N}\sum_{v\in\mathcal{V}} r_v(s_v(t),a(t),s'_v(t)\right], \quad i \in \mathcal{S}. \tag{7}$$

In distributed policy evaluation, the agents wish to cooperate to estimate the reward V. We could do this by applying TD(0) to the reward average $(1/N)\sum_v r_v(t)$. However, this is not naturally distributed since the reward average depends on what happens at every node in the network. Nevertheless, let θ^*_{gl} be the limit point of this method; our goal is to converge to this θ^*_{gl} with a fully distributed method.

A natural way to do this is to distribute the TD(0) method as we do in Algorithm 1. Informally, Algorithm 1 starts from an arbitrary parameter vector $\theta_v(0)$. At each iteration $t \in \mathbb{N}_0$, the agents take actions, the system moves to a new random state $s'(t)$ based on these actions, and agent v observes the a tuple $(s(t),s'(t),r_v(t))$. It then executes the TD(0) algorithm on this tuple. This is done for T steps, and then the system "outputs" the average across the network of the running averages of the iterates maintained by each individual node, and the average across the network of the latest estimates. Communication among workers is required **only** in the final step to average their parameters.

Message complexity: Under the assumption that the nodes are connected to a server, computing the average in step 10 takes a single round of communication with a server. In the nearest-neighbor model where the nodes are connected over an undirected graph and nodes know the total number of nodes N, it is possible to find an ε-approximation of the average in $O(N\log(1/\varepsilon))$ time using the algorithm from Olshevsky (2017). If such knowledge is not available, and the communication graph is further time-varying, it is possible to do the same in $O(N^2\log(1/\varepsilon)$ using the algorithm from Nedic et al. (2009). As we will later discuss, it suffices to choose ε proportional to a power of $1/T$, so that the message complexity of step 10 in the fully distributed setting is at most $O(\log T)$. For large enough T, this is an exponential improvement over the previous papers Doan et al. (2019a, 2020); Sun et al. (2020); Wang et al. (2020a) which required communication at every step and thus needed T communications.

Algorithm 1 TD(0) with Global State

For $v \in \mathcal{V}$, initialize $\theta_v(0)$, $s(0)$
for $t = 0$ to $T-1$ **do**
  for $v \in \mathcal{V}$ **do**
    Observe a tuple $(s(t), s'(t), r_v(t))$.
    Compute temporal difference:

$$\delta_v(t) = r_v(t) - \left(\phi(s(t)) - \gamma\phi(s'(t))\right)^T \theta_v(t).$$

    Execute local TD update:

$$\theta_v(t+1) = \theta_v(t) + \alpha_t \delta_v(t)\phi(s(t)). \tag{8}$$

    Update running average:

$$\hat{\theta}_v(t+1) = \left(1 - \frac{1}{t+2}\right)\hat{\theta}_v(t) + \frac{1}{t+2}\theta_v(t+1).$$

  end for
end for
Return $\hat{\theta}(T)$ and $\bar{\theta}(T)$:

$$\hat{\theta}(T) = \frac{1}{N}\sum_{v\in\mathcal{V}} \hat{\theta}_v(T), \quad \bar{\theta}(T) = \frac{1}{N}\sum_{v\in\mathcal{V}} \theta_v(T).$$

2.3.2 The Local State Model

We next consider a model where each agent has its own independently evolving copy of the same MDP. Although this seems quite different form the global state model, the two models can be treated with a very similar analysis.

More formally, each agent has the same 6-tuple $(\mathcal{S}, \mathcal{V}, \mathcal{A}, \mathcal{P}, r, \gamma)$; at time t, agent v will be in a state $s_v(t)$; it will apply action $a_v(t) \in \mathcal{A}$ with probability $\mu(a_v(t)|s_v(t))$; then agent v moves to state $s'_v(t)$ with probability $\mathcal{P}(s'_v(t)|s_v(t), a_v(t))$, with the transitions of all agents being independent of each other; finally agent v gets a reward $r_v(t) = r(s_v(t), a_v(t), s'_v(t))$. Note that, although the rewards obtained by different agents can be different, the reward function $r(s,a,s')$ is identical across agents.

We define θ^*_{lc} to be the fixed point of TD(0) on the MDP $(\mathcal{S}, \mathcal{V}, \mathcal{A}, \mathcal{P}, r, \gamma)$. Naturally, each agent can easily compute θ^*_{lc} by simply ignoring all the other agents. However, this ignores the possibility that agents can benefit from communication with each other.

We propose a distributed TD method in this setting as Algorithm 2. It is very similar to to Algorithm 1: each agent runs TD(0) locally at each agent, and, at the end, the agents just average the results. The message complexity of this method is the same as the message complexity of Algorithm 1: one communication with a server if a server is assumed to be available and $O(\log T)$ communications in the fully distributed setting.

2.4 Why Two Models?

The global state model was introduced in the previous literature Doan et al. (2019a); Wang et al. (2020a); Sun et al. (2020). It is a natural starting point for multi-agent RL where the state of the system depends on what all the agents do.

However, a shortcoming of the global state model is that it cannot be used to parallelize temporal difference learning. Indeed, consider the global state model with the proviso that all rewards except the rewards of the first agent are zero, and only the action of the first agent affects the transition of the global state. Then the global model reduces to the regular policy evaluation problem for one agent. As a consequence, the best we can hope to achieve in the global state model is to recover the guarantees for classical TD learning.

It may be objected that it is somewhat artificial to only have the first agent make decisions that matter, but it is easy to come up with more involved examples with the same property (e.g., when all rewards are the same and the system evolves according to the action chosen by the most agents).

By contrast, in the local state model, there is the possibility of doing better than the regular TD(0) because the agents are "collectively" observing n tuples $(s_v(t), s'_v(t), r_v(t))$ per step, whereas classical TD(0) only gets to observe a single tuple (though it must be stressed that each agent in the local state model only observes its own tuple). Since these tuples

Algorithm 2 TD(0) with Local State

For $v \in \mathcal{V}$, initialize $\theta_v(0)$, $s_v(0)$
for $t = 0$ to $T-1$ **do**
  for $v \in \mathcal{V}$ **do**
    Observe a tuple $(s_v(t), s'_v(t), r_v(t))$.
    Compute temporal difference:

$$\delta_v(t) = r_v(t) - \left(\phi(s_v(t)) - \gamma\phi(s'_v(t))\right)^T \theta_v(t). \tag{9}$$

    Execute local TD update:

$$\theta_v(t+1) = \theta_v(t) + \alpha_t \delta_v(t)\phi_v(s_v(t)). \tag{10}$$

    Update running average:

$$\hat{\theta}_v(t+1) = \left(1 - \frac{1}{t+2}\right)\hat{\theta}_v(t) + \frac{1}{t+2}\theta_v(t+1).$$

  end for
end for
Return $\hat{\theta}(T)$ and $\bar{\theta}(T)$:

$$\hat{\theta}(T) = \frac{1}{N}\sum_{v\in\mathcal{V}} \hat{\theta}_v(T), \quad \bar{\theta}(T) = \frac{1}{N}\sum_{v\in\mathcal{V}} \theta_v(T).$$

are generated independently, to the extent that the variance of the temporal difference error affects the performance of TD(0), there is the possibility of achieving performance that is a factor of N times better.

3 Convergence Analyses of Our Methods

We next describe the main results of this paper, which are convergence analyses of Algorithms 1 and 2 under the assumption that the tuples are i.i.d. In the literature, the i.i.d model is sometimes referred to as having a "generator" for the MDP and is a more restrictive assumption compared to assuming that the state evolves as a Markov process with a fixed starting state. Nevertheless, this is a standard assumption under which many TD and Q-learning methods are analyzed (e.g., (Sutton et al., 2008; Dalal et al., 2018a,b; Lakshminarayanan and Szepesvari, 2018; Doan et al., 2019a; Chen et al., 2019; Kumar et al., 2019)).

3.1 Distributed TD(0) with Global State Model

Before stating our result, we need to introduce notation for the variance of the temporal difference error. Let $\bar{r}(t) = (1/|N|)\sum_v r_v(t)$ be the average of the instantaneous rewards received by the agents at time t. Then we define

$$\bar{\sigma}^2 = E\left[\left(\bar{r}(t) - \left(\phi(s(t)) - \gamma\phi(s'(t))\right)^T \theta^*_{\rm gl}\right)^2\right].$$

Recall that the expectation is taken with respect to the distribution that generates the state s with probability π_s, then actions $(a_1, \ldots, a_n)$ from the policy, and the next state $s'(t)$ from the transition of the MDP.

Our first main result bounds the performance of distributed TD(0) with global state in terms of $\bar{\sigma}^2$ as well as the initial distance to the optimal solution.

Theorem 1. *Suppose Assumptions 1-2 hold. Suppose further that $\{\theta_v(t)\}_{v\in\mathcal{V}}$ and $\{\hat{\theta}_v(t)\}_{v\in\mathcal{V}}$ are generated by Algorithm 1 in the global state model where the state $s(t)$ is sampled i.i.d according to the stationary distribution π. Then,*

(a) For any constant step-size sequence $\alpha_0 = \cdots = \alpha_T = \alpha \le (1-\gamma)/8$,

$$E\left[\left\|\bar{\theta}(T) - \theta^*_{\rm gl}\right\|_2^2\right] \le e^{-\alpha(1-\gamma)\omega T} E\left[\left\|\bar{\theta}(0) - \theta^*_{\rm gl}\right\|_2^2\right] + \frac{2\alpha\bar{\sigma}^2}{(1-\gamma)\omega}.$$

(b) For any $T \geq \frac{64}{(1-\gamma)^2}$ *and constant step-size sequence* $\alpha_0 = \cdots = \alpha_T = \frac{1}{\sqrt{T}}$,

$$(1-\gamma)E\left[\left\|V_{\theta^*_{\rm gl}} - V_{\hat{\theta}(T)}\right\|_D^2\right] + 2\gamma E\left[\left\|V_{\theta^*_{\rm gl}} - V_{\hat{\theta}(T)}\right\|_{\rm Dir}^2\right] \leq \frac{E\left[\left\|\bar{\theta}(0) - \theta^*_{\rm gl}\right\|_2^2\right] + \bar{\sigma}^2}{\sqrt{T}}.$$

(c) For a decaying step-size sequence $\alpha_t = \frac{\alpha}{t+\tau}$ *with* $\alpha = \frac{2}{(1-\gamma)\omega}$ *and* $\tau = \frac{16}{(1-\gamma)^2\omega}$, *we have,*

$$E\left[\left\|\bar{\theta}(T) - \theta^*_{\rm gl}\right\|_2^2\right] \leq \frac{\zeta}{T+\tau},$$

where $\zeta = \max\left\{2\alpha^2\bar{\sigma}^2, \tau\left\|\bar{\theta}(0) - \theta^*_{\rm gl}\right\|_2^2\right\}$.

A formal proof can be found in the supplementary material.

The first takeaway here is that the convergence rates to $\theta^*_{\rm gl}$ in parts (b) and (c) are *exactly the same as the existing convergence times for regular TD(0).* Indeed, the state-of-the-art finite-time convergence analysis for TD(0) was given in the paper Bhandari et al. (2018), and the bounds given there are exactly the same as the ones in the above theorem, up to constant factors[1]. As discussed earlier, the best one can hope for is to replicate the bounds for regular TD(0) as Algorithm 1 contains the usual TD(0) as a special case.

Similarly part (a) shows convergence of the distributed method to an $O(\alpha\bar{\sigma}^2/(\omega(1-\gamma))$ neighborhood of the optimal solution (measured in terms of the performance measure on the left-hand side). This is also equivalent to the asymptotic performance in their state-of-the-art analysis of regular TD(0) with fixed step-size from Bhandari et al. (2018).

Message complexity of the final consensus step: For simplicity, we have given Theorem 1 under the assumption that the final averages $\hat{\theta}(T), \bar{\theta}(T)$ are computed exactly. We now come back to the question of how many inter-neighbor communication steps are needed to implement step 10 (and preserve our theoretical guarantees) when only nearest-neighbor communications in a graph are allowed.

It is immediate that all the quantities we bound in Theorem 1 (i.e., the left-hand sides of all the equations) are Lipschitz in a neighborhood of θ^*_{gl}. Consequently, to preserve a constant error in part (a), or $1/\sqrt{T}$ and $1/T$ errors in parts (b) and (c), it suffices to average with an error that is a small enough constant in part (a), and a small enough multiple of $1/\sqrt{T}$ and $1/T$ respectively in parts (b) and (c).

As discussed earlier, to obtain an ε-approximate average using state of the art distributed "average consensus" methods takes $O(\log(1/\varepsilon))$ steps, where the constant in the $O(\cdot)$ notation will depend on N or the spectral gap as well as assumptions we make about the graph. Thus we need to take $\varepsilon = O(1/\sqrt{T}), \varepsilon = O(1/T)$ in cases (b),(c). This means we will need to run an average consensus method for $O(\log T)$ communications to approximately implement step 10 of Algorithm 1 after running the previous loop for T steps. Thus the final message complexity is $O(\log T)$.

Comparison to earlier work: A similar algorithm was analyzed in Doan et al. (2019a). The difference is that the agents were assumed to be connected by a (possibly time-varying) sequence of graphs; each agent communicated with neighbors at each step. The bound derived in Doan et al. (2019a) for step-size that scales as $1/\sqrt{k}$ with iteration k was of the form

$$(1-\gamma)E\left[\left\|V_{\theta^*_{\rm gl}} - V_{\hat{\theta}(T)}\right\|_D^2\right] \leq O\left(\frac{||\bar{\theta}(0) - \theta^*_{\rm gl}||_D^2 + \left(\frac{1}{1-\lambda}\right)\ln T}{\sqrt{T}}\right),$$

where λ was related to the spectral gap of the communication graphs; crucially, it was remarked in Doan et al. (2019b) that in the worst case, $1/(1-\lambda)$ was as large as N^3 on a fixed network of N nodes.

In other words, the bounds of Doan et al. (2019a) allowed for the possibility that distributed TD(0) performs worse than regular TD(0) in this setting. It might have been natural to guess that something like this is inevitable due to the multi-agent nature of the system. Our results show this is not the case. Not only can we match the performance of regular TD(0), we do not even need to communicate with neighbors except to average the estimates at the very end.

[1] Actually, the results here are slightly stronger compared to the results in Bhandari et al. (2018). The difference is that we give bounds on the quantity $(1-\gamma)E\left[\left\|V_{\theta^*} - V_{\hat{\theta}(T)}\right\|_D^2\right] + 2\gamma E\left[\left\|V_{\theta^*} - V_{\hat{\theta}(T)}\right\|_{\rm Dir}^2\right]$, whereas the bounds of Bhandari et al. (2018), after some rearrangement, give the same upper bound on just $(1-\gamma)E\left[\left\|V_{\theta^*_{\rm gl}} - V_{\hat{\theta}(T)}\right\|_D^2\right]$.

We remark that Doan et al. (2019a) also considered the fixed step-size, and the follow-up paper Doan et al. (2020) also considered step-sizes that scale with $1/t$. The comparison of parts (a) and (c) with these results is similar, as is the comparison of our results with the analysis of Sun et al. (2020): relative to all these papers, we both remove the scaling with the number of agents or with the eigengap of the underlying graph, while requiring no communication except for average computation at very end.

We next compare this theorem with the results of Wang et al. (2020a). That paper considered a more sophisticated "gradient tracking" scheme, where multiple quantities are shared among neighbors in an underlying graph at every step. Only the fixed step-size case was analyzed, and it was shown that the system converges to a neighborhood of the optimal solution whose size does not depend on the number of agents or the spectral gap of a communication matrix.

The above theorem improves on Wang et al. (2020a) in several ways. Besides significantly saving on communication by not requiring communication any communication until the very end, we give a clean expression for the final error: the right-hand side of Theorem 1(a) is $\alpha\bar{\sigma}^2/(\omega(1-\gamma))$ in the limit as $T \to \infty$. Most importantly, in Theorem 1(b) and Theorem 1(c) we show convergence to the optimal solution itself without any dependence n or a spectral gap (rather than only a neighborhood of it).

3.2 Distributed TD(0) with Local State Model

We now turn to the analysis of the local state model, beginning with some notation. Recall that, for regular TD(0) in the global state model, convergence analysis will scale both with the distance to the initial solution, and with the variance $\bar{\sigma}^2$ of the temporal difference error with average reward. For the local model, the variance is identical to the variance defined in the centralized model:

$$\sigma^2 = \mathrm{E}\left[\left(r(s,a,s') - \left(\phi(s) - \gamma\phi(s')\right)^T \theta^*_{\mathrm{lc}}\right)^2\right].$$

As before, the expectation is taken with respect to the distribution that generates the state s with probability π_s, then actions $(a_1,\ldots,a_n)$ from the policy, and the next state $s'(t)$ from the transition of the MDP.

In the multi-agent case, we need some notion of the initial distance to the optimal solution; we simply take the maximum over all the agents to define:

$$\hat{R}_0 = \max_{v\in\mathcal{V}} E\left[\|\theta_v(0) - \theta^*_{\mathrm{lc}}\|_2^2\right].$$

In the case where all agents start with the same initial condition, this reduces to the same quantities as we had before, i.e., $R_0 = \|\bar{\theta}(0) - \theta^*_{\mathrm{lc}}\|_2^2$.

The following theorem is our second main result.

Theorem 2. *Suppose Assumptions 1-2 hold. Suppose further that $\{\theta_v(t)\}_{v\in\mathcal{V}}$ and $\{\hat{\theta}_v(t)\}_{v\in\mathcal{V}}$ are generated by Algorithm 2 in the local state model under i.i.d sampling. Then,*

(a) For any constant step-size sequence $\alpha_0 = \cdots = \alpha_T = \alpha \le (1-\gamma)/8$, we have

$$E\left[(1-\gamma)\left\|V_{\theta^*_{\mathrm{lc}}} - V_{\hat{\theta}(T)}\right\|_D^2 + \gamma\left\|V_{\theta^*_{\mathrm{lc}}} - V_{\hat{\theta}(T)}\right\|_{\mathrm{Dir}}^2\right] \le \frac{1}{T}\left(\frac{1}{2\alpha}E\left[\|\bar{\theta}(0) - \theta^*_{\mathrm{lc}}\|_2^2\right] + \frac{4\hat{R}_0}{1-\gamma}\right) + \frac{\alpha\sigma^2}{N} + \frac{8\alpha^2\sigma^2}{1-\gamma}.$$

(b) For any $T \ge \frac{64}{(1-\gamma)^2}$ and constant step-size sequence $\alpha_0 = \cdots = \alpha_T = \frac{1}{\sqrt{T}}$, we have

$$E\left[(1-\gamma)\left\|V_{\theta^*_{\mathrm{lc}}} - V_{\hat{\theta}(T)}\right\|_D^2 + \gamma\left\|V_{\theta^*_{\mathrm{lc}}} - V_{\hat{\theta}(T)}\right\|_{\mathrm{Dir}}^2\right] \le \frac{1}{2\sqrt{T}}\left(E\left[\left\|\bar{\theta}(0) - \theta^*_{\mathrm{lc}}\right\|_2^2\right] + \frac{2\sigma^2}{N}\right) + \frac{1}{T}\left(\frac{4\hat{R}_0 + 8\sigma^2}{1-\gamma}\right).$$

(c) For the decaying step-size sequence $\alpha_t = \frac{\alpha}{t+\tau}$ with $\alpha = \frac{2}{(1-\gamma)\omega}$ and $\tau = \frac{16}{(1-\gamma)^2\omega}$. Then,

$$E\left[\left\|\bar{\theta}(t+1) - \theta^*_{\mathrm{lc}}\right\|_2^2\right] \le \frac{2\alpha^2\sigma^2/N}{t+\tau} + \frac{8\alpha^2\hat{\zeta}}{(t+\tau)^2} + \frac{(\tau-1)^4 E\left[\left\|\bar{\theta}(0) - \theta^*_{\mathrm{lc}}\right\|_2^2\right]}{(t+\tau)^4},$$

where $\hat{\zeta} = \max\left\{2\alpha^2\sigma^2, \tau\hat{R}_0\right\}$.

The proof of Theorem 2 is given in the supplementary material.

To parse Theorem 2, note that all the terms in brown are "negligible" in a limiting sense. Indeed, in part (a), the first term scales as $O(1/T)$ and consequently goes to zero as $T \to \infty$ (whereas the remaining terms do not). In parts (b) and

(c), the terms in brown go to zero at an asymptotically faster rate compared to the dominant term (i.e., as $1/T$ vs the dominant $1/\sqrt{T}$ term in part(b) and as $1/t^2, 1/t^4$ compared to the dominant $1/t$ in part (c)). Finally, the last term in part (a) scales as $O(\alpha^2)$ and will be negligible compared to the term preceding it, which scales as $O(\alpha)$, when α is small.

Moreover, among the non-negligible terms, whenever σ^2 appears, it is divided by N; this is highlighted in blue.

To summarize, parts (b) and (c) show that, when the number of iterations is large enough, we can divide the variance term by N as a consequence of the parallelism among N agents. Part (a) shows that, when the number of iterations is large enough and the step-size is small enough, the size of the final error will be divided by N.

Note that, in part (c), the result of this is a factor of N speed up of the entire convergence time (when T is large enough). In part (a), this results in a factor of N shrinking of the asymptotic error (when the step-size α is small enough). In part (b), however, this only shrinks the "variance term" by a factor of N; the term depending on the initial condition is unaffected. The explanation for this is that in parts (a) and (c), the variance of the temporal difference error dominates the convergence rate, while in part (b) this is not the case.

As far as we are aware, these results constitute the first example where parallelism was shown to help for distributed temporal difference learning. They also justify the introduction of the local state model in this paper: indeed, even if there is nothing multi-agent about the underlying problem, one might still choose to distribute the MDP among agents (which could be nodes in a computer cluster) in order to speed-up computation as guaranteed by this theorem.

4 Numerical Experiments

In this section, we perform some experiments to verify the conclusions of our theorems and compare Algorithm 2 with earlier work from Doan et al. (2019a) and Wang et al. (2020a). Our experiments are performed on classic control problems from OpenAI gym and Gridworld; details are given in the supplementary materials.

We focus on the case of constant step-size, since this both matches what is usually done in practice (where a fixed but small step-size is typically picked) and results in faster convergence fitting within our limited computation budget. Normally, a choice of step-size of $\alpha_t = \alpha$ results in an error of $O(\alpha)$ around the optimal solution. But according to Theorem 2(a), choosing $N = \frac{1-\gamma}{8\alpha}$ will result in a final error that is a much smaller $O(\alpha^2)$.

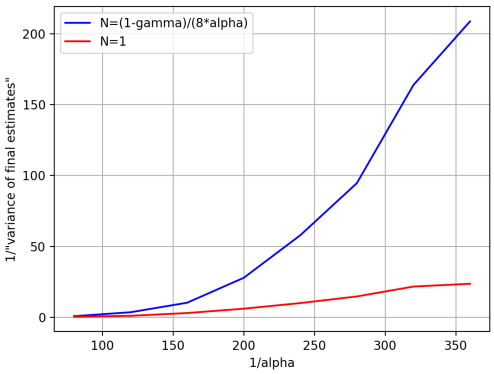
In other words, if we were to plot the inverse of the variance of the final answer, we should see it grows linearly in one agent, and quadratically with N agents chosen as above. This is exactly what Figure 1 below shows, plotting as a function of α^{-1} to make the quadratic vs linear distinction happen when $\alpha \to +\infty$, thus making it more visible.

Note that the step-size α can be thought of trading off between the quality of the final solution, which is $O(\alpha)$, and the convergence time (which scales with α^{-1}). These graphs show that we can use parallelism to get a much more accurate solution ($O(\alpha^2)$ error instead of $O(\alpha)$).

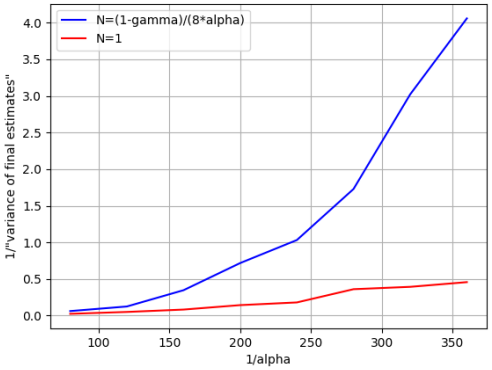
Our second set of simulations compare Algorithm 2 with earlier distributed TD methods stated in Doan et al. (2019a), Sun et al. (2020) and Wang et al. (2020a) in terms of TD error. The distributed TD methods of Doan et al. (2019a) and Sun et al. (2020) are the same expect that Doan et al. (2019a) has an additional projection step (these two methods can be viewed as the same if one chooses a large enough set for the projection step). Again, we consider constant step size. The number of agents $N = 100$. The communication graph among agents is generated by the Erdos–Renyi model, which is connected. **Recall that our method only uses one run of average consensus at the end, whereas the other methods require a communication at every step**. The graphs for our method show the TD error at each iteration if we stopped the method and run the average consensus to average the estimates across the network. Figure 2 shows that the TD errors of Algorithm 2 perform essentially identically to the other methods in spite of the reduced communication.

5 Conclusion

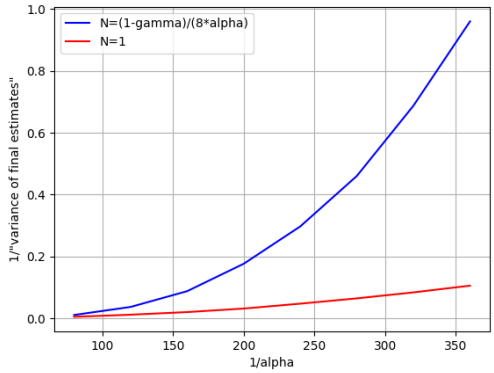
We have presented convergence results for distributed TD(0) with linear function approximation. Our results improve on the previous literature both in terms of utilizing almost no communication: only one run of average consensus is needed. The convergence bounds we derive match state-of-the-art analysis of TD(0) or reduce the variance by a factor of N when the nodes generate their samples independently.



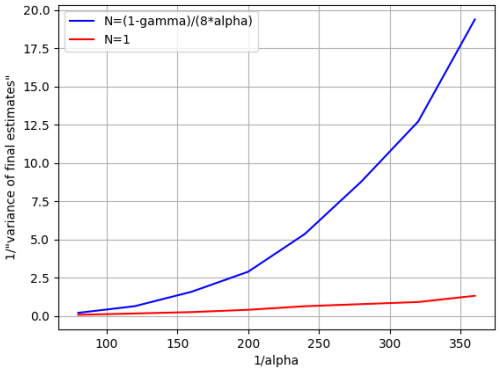
(a) Grid World



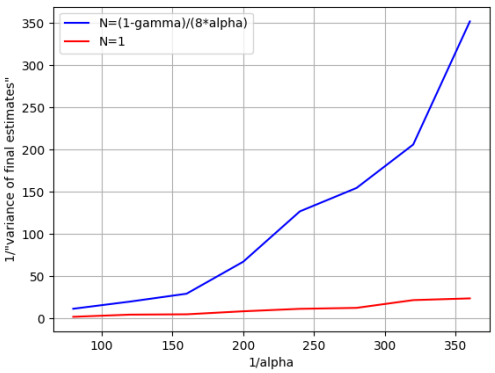
(b) Acrobot-v1



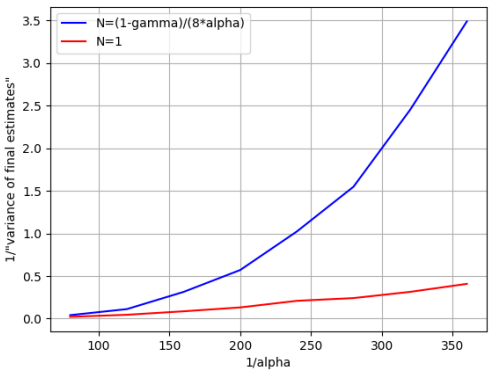
(c) CartPole-v1



(d) MountainCar-v1

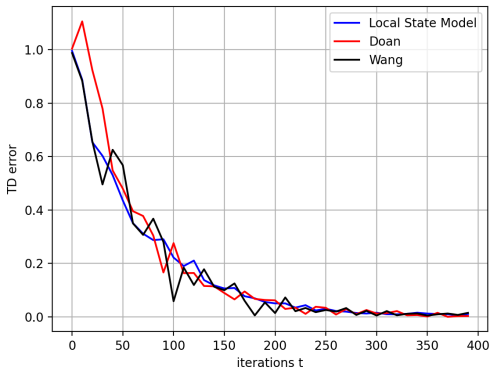


(e) MountainCarContinous-v0

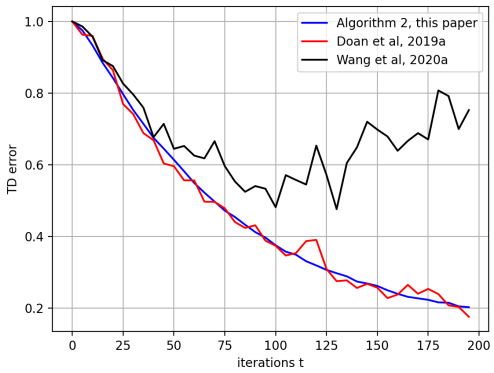


(f) Pendulum-v1

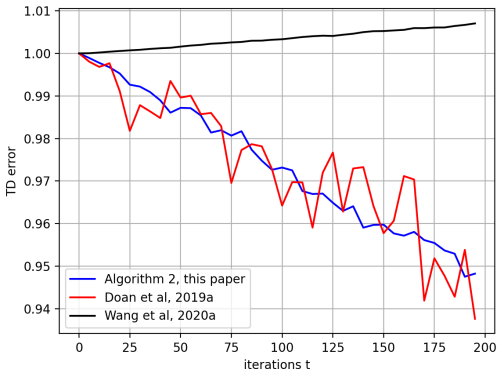
Figure 1: We plot the inverse of the empirical variance of the final solution vs α^{-1} on the x-axis. The policy takes uniformly random actions, and the function approximation uses tile coding. Details are available in the supplementary information.



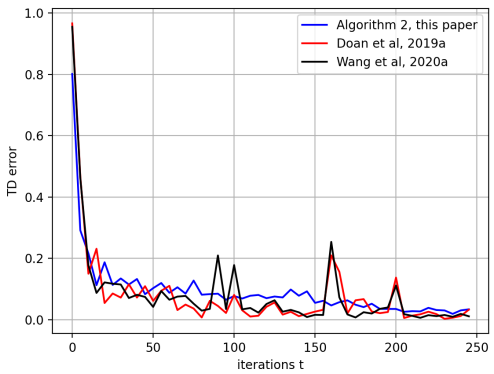
(a) Grid World



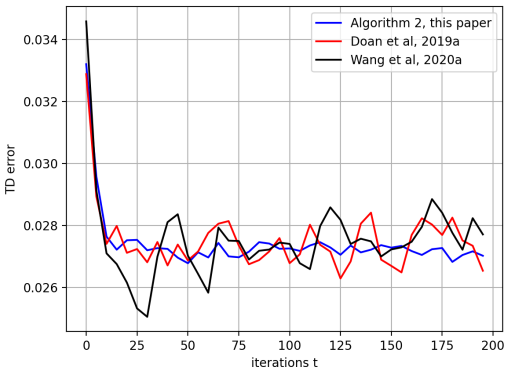
(b) Acrobot-v1



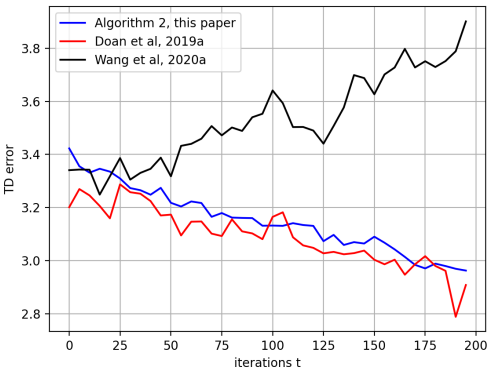
(c) CartPole-v1



(d) MountainCar-v1



(e) MountainCarCont-v0



(f) Pendulum-v1

Figure 2: Comparison of our method to the previous literature in the same setting as Figure 1. Graphs show the TD error vs iteration count.

References

Bazzan, A. L. (2009). Opportunities for multiagent systems and multiagent reinforcement learning in traffic control. *Autonomous Agents and Multi-Agent Systems*, 18(3):342.

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.

Chen, Z., Zhang, S., Doan, T. T., Clarke, J.-P., and Theja Maguluri, S. (2019). Finite-sample analysis of nonlinear stochastic approximation with applications in reinforcement learning. *arXiv e-prints*, pages arXiv–1905.

Dalal, G., Szörényi, B., Thoppe, G., and Mannor, S. (2018a). Finite sample analyses for td (0) with function approximation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32.

Dalal, G., Thoppe, G., Szörényi, B., and Mannor, S. (2018b). Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning. In *Conference On Learning Theory*, pages 1199–1233. PMLR.

Doan, T., Maguluri, S., and Romberg, J. (2019a). 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.

Doan, T. T., Maguluri, S. T., and Romberg, J. (2019b). Finite-time analysis of distributed td (0) with linear function approximation for multi-agent reinforcement learning. *arXiv preprint arXiv:1902.07393*.

Doan, T. T., Nguyen, L. M., Pham, N. H., and Romberg, J. (2020). Finite-time analysis of stochastic gradient descent under markov randomness. *arXiv preprint arXiv:2003.10973*.

Duan, Y., Chen, X., Houthooft, R., Schulman, J., and Abbeel, P. (2016). Benchmarking deep reinforcement learning for continuous control. In *International Conference on Machine Learning*, pages 1329–1338.

Jaakkola, T., Jordan, M. I., and Singh, S. P. (1994). Convergence of stochastic iterative dynamic programming algorithms. In *Advances in Neural Information Processing Systems*, pages 703–710.

Jin, J., Song, C., Li, H., Gai, K., Wang, J., and Zhang, W. (2018). Real-time bidding with multi-agent reinforcement learning in display advertising. In *Proceedings of the 27th ACM International Conference on Information and Knowledge Management*, pages 2193–2201.

Kumar, H., Koppel, A., and Ribeiro, A. (2019). On the sample complexity of actor-critic method for reinforcement learning with function approximation. *arXiv preprint arXiv:1910.08412*.

Kuyer, L., Whiteson, S., Bakker, B., and Vlassis, N. (2008). Multiagent reinforcement learning for urban traffic control using coordination graphs. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 656–671. Springer.

Lakshminarayanan, C. and Szepesvari, C. (2018). Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *International Conference on Artificial Intelligence and Statistics*, pages 1347–1355.

Liu, R. and Olshevsky, A. (2020). Temporal difference learning as gradient splitting. *arXiv preprint arXiv:2010.14657*.

Nedic, A., Olshevsky, A., Ozdaglar, A., and Tsitsiklis, J. N. (2009). On distributed averaging algorithms and quantization effects. *IEEE Transactions on automatic control*, 54(11):2506–2517.

Ollivier, Y. (2018). Approximate temporal difference learning is a gradient descent for reversible policies. *arXiv preprint arXiv:1805.00869*.

Olshevsky, A. (2017). Linear time average consensus and distributed optimization on fixed graphs. *SIAM Journal on Control and Optimization*, 55(6):3990–4014.

Shen, H., Zhang, K., Hong, M., and Chen, T. (2020). Asynchronous advantage actor critic: Non-asymptotic analysis and linear speedup. *arXiv preprint arXiv:2012.15511*.

Sun, J., Wang, G., Giannakis, G. B., Yang, Q., and Yang, Z. (2020). Finite-sample analysis of decentralized temporal-difference learning with linear function approximation. *In International Conference on Artificial Intelligence and Statistics*, pages 1–8.

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

Sutton, R. S. and Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT Press.

Sutton, R. S., Maei, H., and Szepesvári, C. (2008). A convergent $o(n)$ temporal-difference algorithm for off-policy learning with linear function approximation. *Advances in Neural Information Processing Systems*, 21:1609–1616.

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, G., Lu, S., Giannakis, G., Tesauro, G., and Sun, J. (2020a). Decentralized td tracking with linear function approximation and its finite-time analysis. *Advances in Neural Information Processing Systems*, 33.

Wang, Y., Dong, L., and Sun, C. (2020b). Cooperative control for multi-player pursuit-evasion games with reinforcement learning. *Neurocomputing*, 412:101–114.

Yang, E. and Gu, D. (2004). Multiagent reinforcement learning for multi-robot systems: A survey. Technical report, Tech. rep.

Supplementary Information

We now provide proofs of the theorems in the main text of the paper. We begin with a sequence of definitions, notation, and simple observations we will use later.

A Notations and Preliminary Results

A.1 Linear Equations Satisfied by the Fixed Point

Let $X(t) = (s(t), s'(t), r(t))$ be a triple that generates $s(t)$ with steady-state distribution π_s and generates $s'(t)$, $r(t)$ from the MDP. It is well-known Tsitsiklis and Van Roy (1997) that the limit point θ^* of centralized TD(0) is the unique solution of the linear system

$$A\theta = b, \tag{11}$$

where

$$A = E[A(X(t))] = E\left[\phi(s(t))\left(\gamma\phi(s'(t)) - \phi(s(t))\right)^T\right] \tag{12}$$

and

$$b = E[b(X(t))] = E\left[r(t)\phi(s(t))\right].$$

We give a quick generalization of this to the global state in the following lemma.

Lemma 1. *Suppose that Assumption 1-2 hold and suppose that the iterates $\{\theta_v(t)\}_{v\in\mathcal{V}}$ are generated by Algorithm 1 where the step-size sequence α_t is positive, nonincreasing, and satisfies $\sum_{t=0}^{\infty}\alpha_t = \infty$ and $\sum_{t=0}^{\infty}\alpha_t^2 < \infty$. Then $\bar{\theta}(t)$ converges to θ^*_{gl} with probability 1, where θ^*_{gl} is the unique solution of equation*

$$A\theta = \frac{1}{N}\sum_{v\in\mathcal{V}} b_v, \tag{13}$$

where

$$b_v = E\left[r_v(t)\phi(s(t))\right]. \tag{14}$$

Proof. Let $\bar{\theta}(t) = (1/N)\sum_v \theta_v(t)$ be the average of the iterates for all the agents in the network. Following the update Eq. (8), we have

$$\begin{aligned}\bar{\theta}(t+1) =& \bar{\theta}(t) + \alpha_t \frac{1}{N}\sum_{v=1}^{N}\delta_v(t)\phi(s(t)) \\ =& \bar{\theta}(t) + \alpha_t\left(\frac{1}{N}\sum_{v=1}^{N} r_v(t)\phi(s(t)) - \phi(s(t))\left(\phi(s(t)) - \gamma\phi(s'(t))\right)^T\bar{\theta}(t)\right)\end{aligned}$$

In other words, $\bar{\theta}(t)$ follows the TD(0) recursion with $(1/N)\sum_v E[r_v(t)]$ as the reward. We thus apply Theorem 2 of Tsitsiklis and Van Roy (1997) to obtain this lemma.

■

A.2 The Expectation of the TD Direction

For agent v, as shown in Eq. (8), the direction of the distributed TD(0) with global state update at iteration t is $\delta_v(t)\phi(s(t))$. We split it into two terms

$$\delta_v(t)\phi(s(t)) = h_v(t) + m_v(t),$$

where

$$h_v(t) = b_v - A\theta_v(t), \tag{15}$$

$$m_v(t) = \delta_v(t)\phi(s(t)) - h_v(t). \tag{16}$$

It is worth mentioning that, for a given $\theta_v(t)$, the quantity $h_v(t)$ can be interpreted as the conditional expectation of the direction $\delta_v(t)\phi(s(t))$:

$$h_v(t) = E\left[\delta_v(t)\phi(s(t))|\theta_v(t)\right].$$

Indeed,

$$E\left[\delta_v(t)\phi(s(t))|\theta_v(t)\right] = E\left[\left(r_v(t) - \left(\phi(s(t)) - \gamma\phi(s'(t))\right)^T\theta_v(t)\right)\phi(s(t))|\theta_v(t)\right]$$

$$
\begin{aligned}
&= E\left[r_v(t)\phi(s(t))\right] - E\left[\phi(s(t))\left(\phi(s(t)) - \gamma\phi(s'(t))\right)^T\right]\theta_v(t) \\
&= b_v - A\theta_v(t),
\end{aligned} \tag{17}
$$

where the second equality follows because $r_v(t)$, $\phi(s(t)), \phi(s'(t))$ have the same distribution regardless of $\theta_v(t)$.

It is immediate then that $m_v(t)$ represents the error, i.e., the difference between the direction $\delta_v(t)\phi(s(t))$ and its conditional expectation. Hence, the update of Eq. (8) for TD(0) can be written as:

$$
\theta_v(t+1) = \theta_v(t) + \alpha_t\left(h_v(t) + m_v(t)\right). \tag{18}
$$

Let $\Theta(t)$ denote the matrix whose v-th row is the estimates of agent v at time t, i.e., $\theta_v^T(t)$. Thus,

$$
\Theta(t) = \begin{bmatrix} — & \theta_1^T(t) & — \\ \cdots & \cdots & \cdots \\ — & \theta_N^T(t) & — \end{bmatrix} \in \mathbb{R}^{N\times K}.
$$

The matrix form of Algorithm 1 then can be written as

$$
\Theta(t+1) = \Theta(t) + \alpha_t\left[H(t) + M(t)\right], \tag{19}
$$

where $H(t)$, $M(t)$ are the matrices, whose v-th rows are $h_v^T(t)$, $m_v^T(t)$ respectively.

Let us adopt the convention that given a collection of vectors, one for each agent in the network, putting a bar will denote their average. Then

$$
\bar{\theta}(t+1) = \bar{\theta}(t) + \alpha_t\left[\bar{h}(t) + \bar{m}(t)\right]. \tag{20}
$$

Our next proposition introduces some useful properties of $\bar{h}(t)$ and $\bar{m}(t)$.

Proposition 1. *Suppose Assumptions 1-2 hold, and suppose that $\{\theta_v(t)\}_{v\in\mathcal{V}}$ are generated by Algorithm 1. Then,*

(a) $\bar{h}(t)$ is a linear function of $\bar{\theta}(t)$:

$$
\bar{h}(t) = \bar{b} - A\bar{\theta}(t),
$$

where A, b_v are defined in Eq. (12) and Eq. (14), and $\bar{b} = \frac{1}{N}\sum_{v\in\mathcal{V}} b_v$;

(b) The conditional expectation of $\bar{m}(t)$ given $\Theta(t)$ is equal to zero:

$$
E[\bar{m}(t)|\Theta(t)] = 0. \tag{21}
$$

Proof of Proposition 1. (a) Following the definition of $\bar{h}(t)$, it can be observed that:

$$
\bar{h}(t) = \frac{1}{N}\sum_{v\in\mathcal{V}} h_v(t) = \frac{1}{N}\sum_{v\in\mathcal{V}}\left(b_v - A\theta_v(t)\right) = \bar{b} - A\bar{\theta}(t),
$$

where $\bar{b} = \frac{1}{N}\sum_{v\in\mathcal{V}} b_v$.

(b) Recall that $h_v(t)$ is exactly the conditional expectation of $\delta_v(t)\phi(s(t))$ given $\theta_v(t)$. Actually, the more general statement

$$
E[\delta_v(t)\phi(s(t))|\Theta(t)] = h_v(t),
$$

is true; for a proof, one can simply repeat all the steps of Eq. (17) replacing $\theta_v(t)$ by $\Theta(t)$.

Furthermore, $m_v(t)$ is defined as

$$
m_v(t) = \delta_v(t)\phi(s(t)) - h_v(t),
$$

by Eq. (16). Therefore: $\theta_v(t)$:

$$
\begin{aligned}
E\left[m_v(t)|\Theta(t)\right] &= E\left[\delta_v(t)\phi(s(t)) - h_v(t)|\Theta(t)\right] \\
&= E\left[\delta_v(t)\phi(s(t))|\Theta(t)\right] - h_v(t) \\
&= 0
\end{aligned}
$$

By the definition of $\bar{m}(t)$, we then have

$$
E[\bar{m}(t)|\Theta(t)] = 0.
$$

■

A.3 Averages of TD Updates

In this subsection, we introduce an equality satisfied by the inner product between the averages of TD updates throughout the agents in the network and the direction from the averages of the iterates throughout the network to the fixed point of distributed TD algorithm.

Lemma 2. *Suppose Assumptions 1-2 hold. Further suppose that $\{\theta_v(t)\}_{v\in\mathcal{V}}$ are generated by Algorithm 1. For any integer $t \geq 0$, we have that*

$$E\left[\left[\bar{h}(t)+\bar{m}(t)\right]^T\left(\theta^*_{\text{gl}}-\bar{\theta}(t)\right)\right] = E\left[(1-\gamma)\left\|V_{\theta^*_{\text{gl}}}-V_{\bar{\theta}(t)}\right\|_D^2+\gamma\left\|V_{\theta^*_{\text{gl}}}-V_{\bar{\theta}(t)}\right\|_{\text{Dir}}^2\right] \tag{22}$$

Proof.

$$\begin{aligned}E\left[\left[\bar{h}(t)+\bar{m}(t)\right]^T\left(\theta^*_{\text{gl}}-\bar{\theta}(t)\right)\right] =& E\left[E\left[\left(\bar{h}(t)+\bar{m}(t)\right)^T\left(\theta^*_{\text{gl}}-\bar{\theta}(t)\right)|\Theta(t)\right]\right]\\ =& E\left[\bar{h}^T(t)(\theta^*_{\text{gl}}-\bar{\theta}(t))\right],\end{aligned}$$

where in the second equation, we use Eq. (21). Here, by Proposition 1 part (a), we have that $\bar{h}(t)$ is a linear function of $\bar{\theta}(t)$, i.e., $\bar{h}(t)=\bar{b}-A\bar{\theta}(t)$. Furthermore, if we let $\bar{h}(\theta)$ denote the linear function $\bar{b}-A\theta$, we can obtain that $\bar{h}(\theta^*_{\text{gl}})=0$.

Corollary 1 in Liu and Olshevsky (2020) states that for any $\theta\in\mathbb{R}^K$,

$$(\theta^*-\theta)^T\bar{g}(\theta) = (1-\gamma)\|V_{\theta^*}-V_\theta\|_D^2+\gamma\|V_{\theta^*}-V_\theta\|_{\text{Dir}}^2,$$

where $\bar{g}(\theta)$ in that paper denote the steady-state expectation of $r(s,s')\phi(s(t))-\phi(s(t))\left(\phi(s)-\gamma\phi(s')\right)^T\theta$, which is indeed $b-A\theta$ and θ^* is the limit point of centralized TD(0) method such that $\bar{g}(\theta^*)=0$.

Now applying Corollary 1 in Liu and Olshevsky (2020), we can obtain Eq.(22).

■

B Proof of Theorem 1

With the above preliminaries in place, we can now begin the proof of Theorem 1. Our first step is to analyze the recurrence relation satisfied by the averages of the iterates throughout the network.

Lemma 3. *Suppose Assumptions 1-2 hold. Further suppose that $\{\theta_v(t)\}_{v\in\mathcal{V}}$ are generated by Algorithm 1. For any integer $t \geq 0$, we have that*

$$\begin{aligned}E\left[\left\|\bar{\theta}(t+1)-\theta^*_{\text{gl}}\right\|_2^2\right] \leq& E\left[\left\|\bar{\theta}(t)-\theta^*_{\text{gl}}\right\|_2^2\right]+\alpha_t^2\left[2\bar{\sigma}^2+8\left\|V_{\theta^*_{\text{gl}}}-V_{\bar{\theta}(t)}\right\|_D^2\right]\\ &-2\alpha_t E\left[(1-\gamma)\left\|V_{\theta^*_{\text{gl}}}-V_{\bar{\theta}(t)}\right\|_D^2+\gamma\left\|V_{\theta^*_{\text{gl}}}-V_{\bar{\theta}(t)}\right\|_{\text{Dir}}^2\right],\end{aligned}$$

Proof of Lemma 3. By Eq. (20), we have

$$\begin{aligned}\left\|\bar{\theta}(t+1)-\theta^*_{\text{gl}}\right\|_2^2 =& \left\|\bar{\theta}(t)+\alpha_t\left[\bar{h}(t)+\bar{m}(t)\right]-\theta^*_{\text{gl}}\right\|_2^2\\ =& \left\|\bar{\theta}(t)-\theta^*_{\text{gl}}\right\|_2^2+2\alpha_t\left[\bar{h}(t)+\bar{m}(t)\right]^T\left(\bar{\theta}(t)-\theta^*_{\text{gl}}\right)+\alpha_t^2\left\|\bar{h}(t)+\bar{m}(t)\right\|_2^2.\end{aligned}$$

Taking expectations we obtain that

$$E\left[\left\|\bar{\theta}(t+1)-\theta^*_{\text{gl}}\right\|_2^2\right] = E\left[\left\|\bar{\theta}(t)-\theta^*_{\text{gl}}\right\|_2^2\right]+\alpha_t^2E\left[\left\|\bar{h}(t)+\bar{m}(t)\right\|_2^2\right]-2\alpha_tE\left[\left(\bar{h}(t)+\bar{m}(t)\right)^T\left(\theta^*_{\text{gl}}-\bar{\theta}(t)\right)\right]. \tag{23}$$

Consider the second term on the right hand side of Eq. (23). Following the definition of $\bar{h}(t)$ and $\bar{m}(t)$, we have that

$$E\left[\left\|\bar{h}(t)+\bar{m}(t)\right\|_2^2\right] = E\left[\left\|\frac{1}{N}\sum_{v\in\mathcal{V}}\delta_v(t)\phi(s(t))\right\|_2^2\right].$$

Plugging in the expression for TD error $\delta_v(t)$ with Eq. (6), we obtain

$$
\begin{aligned}
&E\left[\left\|\bar{h}(t)+\bar{m}(t)\right\|_2^2\right] \\
=&E\left[\left\|\frac{1}{N}\sum_{v\in\mathcal{V}} r_v(t)\phi(s(t)) \;\; -\frac{1}{N}\sum_{v\in\mathcal{V}}\phi(s(t))\left(\phi(s(t))-\gamma\phi(s'(t))\right)^T\theta_v(t)\right\|_2^2\right] \\
=&E\left[\left\|\frac{1}{N}\sum_{v\in\mathcal{V}}\left[r_v(t)-\left(\phi(s(t))-\gamma\phi(s'(t))\right)^T\theta_{\text{gl}}^*\right]\phi(s(t))-\frac{1}{N}\sum_{v\in\mathcal{V}}\phi(s(t))\left(\phi(s(t))-\gamma\phi(s'(t))\right)^T(\theta_v(t)-\theta_{\text{gl}}^*)\right\|_2^2\right]
\end{aligned}
$$

Denote

$$
\begin{aligned}
\boldsymbol{a}^* &= \frac{1}{N}\sum_{v\in\mathcal{V}}\left[r_v(t)-\left(\phi(s(t))-\gamma\phi(s'(t))\right)^T\theta_{\text{gl}}^*\right]\phi(s(t)), \\
\boldsymbol{b}^* &= \frac{1}{N}\sum_{v\in\mathcal{V}}\phi(s(t))\left(\phi(s(t))-\gamma\phi(s'(t))\right)^T(\theta_v(t)-\theta_{\text{gl}}^*).
\end{aligned}
$$

Using inequality $\|\boldsymbol{a}^*-\boldsymbol{b}^*\|^2 \le 2\|\boldsymbol{a}^*\|^2+2\|\boldsymbol{b}^*\|^2$, we obtain

$$
E\left[\left\|\bar{h}(t)+\bar{m}(t)\right\|_2^2\right] \le 2E\left[\|\boldsymbol{a}^*\|^2\right]+2E\left[\|\boldsymbol{b}^*\|^2\right]. \tag{24}
$$

We first bound $E\left[\|\boldsymbol{a}^*\|^2\right]$:

$$
\begin{aligned}
E\left[\|\boldsymbol{a}^*\|^2\right] =&E\left[\left\|\frac{1}{N}\sum_{v\in\mathcal{V}} r_v(t)\phi(s(t))-\left(\left(\phi(s(t))-\gamma\phi(s'(t))\right)^T\theta_{\text{gl}}^*\right)\phi(s(t))\right\|^2\right] \\
\le&E\left[\left(\frac{1}{N}\sum_{v\in\mathcal{V}} r_v(t)-\left(\phi(s(t))-\gamma\phi(s'(t))\right)^T\theta_{\text{gl}}^*\right)^2\right]=\bar{\sigma}^2,
\end{aligned} \tag{25}
$$

where the inequality follows by Assumption 2 and the equality is just the definition of $\bar{\sigma}^2$.

The next step is to find the bound for $E\left[\|\boldsymbol{b}^*\|^2\right]$:

$$
\begin{aligned}
E\left[\|\boldsymbol{b}^*\|^2\right] =&E\left[\left\|\frac{1}{N}\sum_{v\in\mathcal{V}}\phi(s(t))\left(\phi(s(t))-\gamma\phi(s'(t))\right)^T(\theta_v(t)-\theta_{\text{gl}}^*)\right\|^2\right] \\
=&E\left[\left\|\phi(s(t))\left(\phi(s(t))-\gamma\phi(s'(t))\right)^T(\bar{\theta}(t)-\theta_{\text{gl}}^*)\right\|^2\right] \\
\le&4\left\|V_{\theta_{\text{gl}}^*}-V_{\bar{\theta}(t)}\right\|_D^2,
\end{aligned} \tag{26}
$$

where the last line follows from the proof of Lemma 5 in Bhandari et al. (2018).

Plugging Eq. (25) and Eq. (26) into Eq. (24), we get

$$
E\left[\left\|\bar{h}(t)+\bar{m}(t)\right\|_2^2\right] \le 2\bar{\sigma}^2+8\left\|V_{\theta_{\text{gl}}^*}-V_{\bar{\theta}(t)}\right\|_D^2. \tag{27}
$$

The argument we just made bounds one of the terms in Eq. (23). We next consider a different term in the same equation, namely we consider the third term on the right hand side of Eq. (23) which satisfies the equality in Lemma 2:

$$
E\left[\left[\bar{h}(t)+\bar{m}(t)\right]^T(\theta_{\text{gl}}^*-\bar{\theta}(t))\right] = E\left[(1-\gamma)\left\|V_{\theta_{\text{gl}}^*}-V_{\bar{\theta}(t)}\right\|_D^2+\gamma\left\|V_{\theta_{\text{gl}}^*}-V_{\bar{\theta}(t)}\right\|_{\text{Dir}}^2\right] \tag{28}
$$

Finally, combining equations Eq. (23), Eq. (28), and Eq. (27), we obtain

$$E\left[\left\|\bar{\theta}(t+1)-\theta^*_{\text{gl}}\right\|_2^2\right] \leq E\left[\left\|\bar{\theta}(t)-\theta^*_{\text{gl}}\right\|_2^2\right]+\alpha_t^2\left[2\bar{\sigma}^2+8\left\|V_{\theta^*_{\text{gl}}}-V_{\bar{\theta}(t)}\right\|_D^2\right] - 2\alpha_t E\left[(1-\gamma)\left\|V_{\theta^*_{\text{gl}}}-V_{\bar{\theta}(t)}\right\|_D^2+\gamma\left\|V_{\theta^*_{\text{gl}}}-V_{\bar{\theta}(t)}\right\|_{\text{Dir}}^2\right],$$

which is what we needed to show. ■

With this lemma in place, we next prove Theorem 1.

Proof of Theorem 1. Starting with the statement of Lemma 3, which we reproduce here for convenience,

$$E\left[\left\|\bar{\theta}(t+1)-\theta^*_{\text{gl}}\right\|_2^2\right] \leq E\left[\left\|\bar{\theta}(t)-\theta^*_{\text{gl}}\right\|_2^2\right]+\alpha_t^2\left[2\bar{\sigma}^2+8\left\|V_{\theta^*_{\text{gl}}}-V_{\bar{\theta}(t)}\right\|_D^2\right] - 2\alpha_t E\left[(1-\gamma)\left\|V_{\theta^*_{\text{gl}}}-V_{\bar{\theta}(t)}\right\|_D^2+\gamma\left\|V_{\theta^*_{\text{gl}}}-V_{\bar{\theta}(t)}\right\|_{\text{Dir}}^2\right]. \tag{29}$$

we will plug in different choices of step-size.

Proof of part (a): We consider a constant step-size sequence $\alpha_0 = \cdots = \alpha_T \leq (1-\gamma)/8$. Let α denote this constant step-size. Since $\alpha \leq (1-\gamma)/8$, it follows that

$$8\alpha^2 - 2\alpha(1-\gamma) \leq -\alpha(1-\gamma).$$

Plugging this into Eq. (29) and rearranging,

$$\begin{aligned}
&E\left[\left\|\bar{\theta}(t+1)-\theta^*_{\text{gl}}\right\|_2^2\right]\\
\leq& E\left[\left\|\bar{\theta}(t)-\theta^*_{\text{gl}}\right\|_2^2\right]-\alpha(1-\gamma)E\left[\left\|V_{\theta^*_{\text{gl}}}-V_{\bar{\theta}(t)}\right\|_D^2\right]+2\alpha^2\bar{\sigma}^2-2\alpha\gamma E\left[\left\|V_{\theta^*_{\text{gl}}}-V_{\bar{\theta}(t)}\right\|_{\text{Dir}}^2\right]\\
\leq& E\left[\left\|\bar{\theta}(t)-\theta^*_{\text{gl}}\right\|_2^2\right]-\alpha(1-\gamma)E\left[\left\|V_{\theta^*_{\text{gl}}}-V_{\bar{\theta}(t)}\right\|_D^2\right]+2\alpha^2\bar{\sigma}^2\\
\leq& (1-\alpha(1-\gamma)\omega)\,E\left[\left\|\bar{\theta}(t)-\theta^*_{\text{gl}}\right\|_2^2\right]+2\alpha^2\bar{\sigma}^2,
\end{aligned}$$

where the second inequality follows that $\left\|V_{\theta^*_{\text{gl}}}-V_{\bar{\theta}(t)}\right\|_{\text{Dir}}^2$ is non-negative and the third inequality uses Lemma 1 in Bhandari et al. (2018) which states that

$$\sqrt{\omega}\|\theta\|_2 \leq \|V_\theta\|_D \leq \|\theta\|_2.$$

Iterating this inequality establishes that after T iterations

$$\begin{aligned}
E\left[\left\|\bar{\theta}(T)-\theta^*_{\text{gl}}\right\|_2^2\right] &\leq (1-\alpha(1-\gamma)\omega)^T\,E\left[\left\|\bar{\theta}(0)-\theta^*_{\text{gl}}\right\|_2^2\right]+2\alpha^2\bar{\sigma}^2\sum_{t=0}^{T-1}(1-\alpha(1-\gamma)\omega)^t\\
&\leq e^{-\alpha(1-\gamma)\omega T}E\left[\left\|\bar{\theta}(0)-\theta^*_{\text{gl}}\right\|_2^2\right]+\frac{2\alpha\bar{\sigma}^2}{(1-\gamma)\omega},
\end{aligned}$$

where the last inequality follows because $1-\alpha(1-\gamma)\omega \leq e^{-\alpha(1-\gamma)\omega}$ and the standard formula for the sum of a geometric series.

Proof of part (b): We now take the step-size $\alpha_0 = \cdots = \alpha_T = \frac{1}{\sqrt{T}}$. Since the step-size is constant once we fix T, we can denote it by α. Since $T \geq \frac{64}{(1-\gamma)^2}$, it can be observed that $\alpha = \frac{1}{\sqrt{T}} \leq \frac{1-\gamma}{8}$. Plugging this into Eq. (29) and rearranging it, we obtain

$$\alpha E\left[(1-\gamma)\left\|V_{\theta^*_{\text{gl}}}-V_{\bar{\theta}(t)}\right\|_D^2+2\gamma\left\|V_{\theta^*_{\text{gl}}}-V_{\bar{\theta}(t)}\right\|_{\text{Dir}}^2\right] \leq E\left[\left\|\bar{\theta}(t)-\theta^*_{\text{gl}}\right\|_2^2\right]-E\left[\left\|\bar{\theta}(t+1)-\theta^*_{\text{gl}}\right\|_2^2\right]+2\alpha^2\bar{\sigma}^2.$$

Summing over t gives

$$\sum_{t=0}^{T-1} \alpha E\left[(1-\gamma)\left\|V_{\theta^*_{\text{gl}}}-V_{\bar{\theta}(t)}\right\|_D^2+2\gamma\left\|V_{\theta^*_{\text{gl}}}-V_{\bar{\theta}(t)}\right\|_{\text{Dir}}^2\right] \leq E\left[\left\|\bar{\theta}(0)-\theta^*_{\text{gl}}\right\|_2^2\right]-E\left[\left\|\bar{\theta}(T)-\theta^*_{\text{gl}}\right\|_2^2\right]+2T\alpha^2\bar{\sigma}^2$$
$$\leq E\left[\left\|\bar{\theta}(0)-\theta^*_{\text{gl}}\right\|_2^2\right]+2T\alpha^2\bar{\sigma}^2,$$

Dividing by α on both sides, we obtain:

$$\sum_{t=0}^{T-1} E\left[(1-\gamma)\left\|V_{\theta^*_{\text{gl}}}-V_{\bar{\theta}(t)}\right\|_D^2+2\gamma\left\|V_{\theta^*_{\text{gl}}}-V_{\bar{\theta}(t)}\right\|_{\text{Dir}}^2\right] \leq \frac{1}{\alpha}E\left[\left\|\bar{\theta}(0)-\theta^*_{\text{gl}}\right\|_2^2\right]+2T\alpha\bar{\sigma}^2.$$

Finally, recall our notation $\hat{\theta}(T)=\frac{1}{T}\sum_{t=1}^{T}\bar{\theta}(t)$. Then, by convexity

$$E\left[(1-\gamma)\left\|V_{\theta^*_{\text{gl}}}-V_{\hat{\theta}(T)}\right\|_D^2+2\gamma\left\|V_{\theta^*_{\text{gl}}}-V_{\hat{\theta}(T)}\right\|_{\text{Dir}}^2\right] \leq \frac{1}{T}\sum_{t=1}^{T} E\left[(1-\gamma)\left\|V_{\theta^*_{\text{gl}}}-V_{\bar{\theta}(t)}\right\|_D^2+2\gamma\left\|V_{\theta^*_{\text{gl}}}-V_{\bar{\theta}(t)}\right\|_{\text{Dir}}^2\right]$$
$$\leq \frac{1}{\alpha T}E\left[\left\|\bar{\theta}(0)-\theta^*_{\text{gl}}\right\|_2^2\right]+2\alpha\bar{\sigma}^2. \tag{30}$$

Plugging in that $\alpha=\frac{1}{\sqrt{T}}$, we have that

$$E\left[(1-\gamma)\left\|V_{\theta^*_{\text{gl}}}-V_{\hat{\theta}(T)}\right\|_D^2+2\gamma\left\|V_{\theta^*_{\text{gl}}}-V_{\hat{\theta}(T)}\right\|_{\text{Dir}}^2\right] \leq \frac{1}{\sqrt{T}}E\left[\left\|\bar{\theta}(0)-\theta^*_{\text{gl}}\right\|_2^2\right]+\frac{2\bar{\sigma}^2}{\sqrt{T}}$$
$$=\frac{E\left[\left\|\bar{\theta}(0)-\theta^*_{\text{gl}}\right\|_2^2\right]+2\bar{\sigma}^2}{\sqrt{T}}.$$

Proof of part (c): From $\alpha_t \leq \alpha_0=\frac{\alpha}{\tau}=\frac{1-\gamma}{8}$, we have that for $t\geq 0$, it can once again be observed that

$$8\alpha_t^2-2\alpha_t(1-\gamma)\leq -\alpha_t(1-\gamma).$$

Applying this to Eq. (29), we have

$$E\left[\left\|\bar{\theta}(t+1)-\theta^*_{\text{gl}}\right\|_2^2\right] \leq E\left[\left\|\bar{\theta}(t)-\theta^*_{\text{gl}}\right\|_2^2\right]+2\alpha_t^2\bar{\sigma}^2-\alpha_t(1-\gamma)E\left[\left\|V_{\theta^*_{\text{gl}}}-V_{\bar{\theta}(t)}\right\|_D^2\right]-2\alpha_t\gamma E\left[\left\|V_{\theta^*_{\text{gl}}}-V_{\bar{\theta}(t)}\right\|_{\text{Dir}}^2\right]$$
$$\leq E\left[\left\|\bar{\theta}(t)-\theta^*_{\text{gl}}\right\|_2^2\right]+2\alpha_t^2\bar{\sigma}^2-\alpha_t(1-\gamma)E\left[\left\|V_{\theta^*_{\text{gl}}}-V_{\bar{\theta}(t)}\right\|_D^2\right],$$

Applying Lemma 1 in Bhandari et al. (2018) stating that $\sqrt{\omega}\|\theta\|_2\leq\|V_\theta\|_D$, we obtain

$$E\left[\left\|\bar{\theta}(t+1)-\theta^*_{\text{gl}}\right\|_2^2\right] \leq (1-\alpha_t(1-\gamma)\omega)E\left[\left\|\bar{\theta}(t)-\theta^*_{\text{gl}}\right\|_2^2\right]+2\alpha_t^2\bar{\sigma}^2.$$

Wwe will next prove by induction that this last inequality implies that

$$E\left[\left\|\bar{\theta}(t)-\theta^*_{\text{gl}}\right\|_2^2\right]\leq\frac{\zeta}{t+\tau},$$

where $\zeta=\max\left\{2\alpha^2\bar{\sigma}^2,\tau\left\|\bar{\theta}(0)-\theta^*_{\text{gl}}\right\|_2^2\right\}$.

Indeed, the assertion clearly holds at $t=0$. Suppose that the assertion holds at time t, i.e., suppose that $E\left[\left\|\bar{\theta}(t)-\theta^*_{\text{gl}}\right\|_2^2\right]\leq\frac{\zeta}{t+\tau}$. Then,

$$E\left[\left\|\bar{\theta}(t+1)-\theta^*_{\text{gl}}\right\|_2^2\right] \leq (1-\alpha_t(1-\gamma)\omega)\cdot\frac{\zeta}{t+\tau}+2\alpha_t^2\bar{\sigma}^2$$
$$=\frac{(t+\tau)\zeta-\alpha(1-\gamma)\omega\zeta+2\alpha^2\bar{\sigma}^2}{(t+\tau)^2}$$

$$
\begin{aligned}
&= \frac{(t+\tau-1)\zeta + \zeta - \alpha(1-\gamma)\omega\zeta + 2\alpha^2\bar{\sigma}^2}{(t+\tau)^2} \\
&= \frac{(t+\tau-1)\zeta - \zeta + 2\alpha^2\bar{\sigma}^2}{(t+\tau)^2} \\
&\leq \frac{(t+\tau-1)\zeta}{(t+\tau)^2} \leq \frac{(t+\tau-1)\zeta}{(t+\tau)^2-1} \\
&= \frac{\zeta}{t+1+\tau},
\end{aligned}
$$

where note that, at the first equal sign, we plug in the expression for α_t; and the second inequality, we use that $\zeta \geq 2\alpha^2\bar{\sigma}^2$. ■

Next, we state Eq. (30) in the proof of part(b) as the following Corollary.

Corollary 1. *Suppose Assumptions 1-2 hold. Suppose further that $\{\theta_v(t)\}_{v\in\mathcal{V}}$ are generated by Algorithm 1 in the global state model. Then, For any constant step-size sequence $\alpha_0 = \cdots = \alpha_T = \alpha \leq (1-\gamma)/8$,*

$$
E\left[(1-\gamma)\left\|V_{\theta^*_{\mathrm{gl}}} - V_{\hat{\theta}(T)}\right\|_D^2 + 2\gamma\left\|V_{\theta^*_{\mathrm{gl}}} - V_{\hat{\theta}(T)}\right\|_{\mathrm{Dir}}^2\right] \leq \frac{1}{\alpha T} E\left[\left\|\bar{\theta}(0) - \theta^*_{\mathrm{gl}}\right\|_2^2\right] + 2\alpha\bar{\sigma}^2.
$$

We believe this corollary is of independent interest, as it gives a very clean expression for steady-state error of $O(\alpha\bar{\sigma}^2)$ when evaluated in terms of the metric on the left-hand side. Note that there is no scaling here withe either the number of agents or the spectral gap, either in the final steady-state error or in the convergence time that comes from the first term on the right-hand side.

We state this as a corollary, rather than a theorem, since our theorem uses the metrics in the previous work, i.e., $||\theta^*_{\mathrm{gl}} - \hat{\theta}(T)||_2^2$ in the case of fixed step-size, to make the comparison between this paper and earlier papers clear. However, the relatively clean expression for the final steady-state error in this corollary suggests that, rather than using the distance to the optimal solution as the metric of performance, it is better to use the distances between the corresponding value vectors.

C Proof of Theorem 2

We now turn to the proof of Theorem 2. We now assume, for the remainder of this section, that we are analyzing Algorithm 2 in the local state model, subject to Assumptions 1 and 2. It turns out that, in many respects, the local state model can be treated analogously to the global state model.

Our starting point is similar. For a particular agent v, as shown in Eq. (10), the direction of the distributed TD(0) update at iteration t is $\delta_v(t)\phi(s_v(t))$. As before, we split this into two terms

$$
\delta_v(t)\phi(s_v(t)) = h_v(t) + m_v(t),
$$

where

$$
\begin{aligned}
h_v(t) &= b - A\theta_v(t), \\
m_v(t) &= \delta_v(t)\phi(s_v(t)) - h_v(t),
\end{aligned}
$$

where

$$
A = E_{\mu,s}\left[\phi(s_v(t))\left(\gamma\phi(s'_v(t)) - \phi(s_v(t))\right)^T\right],
$$

where, note that the right-hand side does not actually depend on v since $s_v(t), s'_v(t)$ have the same joint distribution regardless of v, and

$$
b = E_{\mu,s}\left[r_v(t)\phi(s_v(t))\right]
$$

where, again, the right-hand side actually does not depend on v.

It is worth mentioning that, although here we use the same notation h_v and m_v as in our earlier treatment of the global state model, the definitions are now slightly different since each agent maintains its own state $s_v(t)$.

As before we let $\Theta(t)$ be

$$
\Theta(t) = \begin{bmatrix} — & \theta_1^T(t) & — \\ \cdots & \cdots & \cdots \\ — & \theta_N^T(t) & — \end{bmatrix} \in \mathbb{R}^{N\times K}.
$$

Our first observation is that the quantity $h_v(t)$ can be interpreted as the conditional expectation of the update direction:

$$h_v(t) = E\left[\delta_v(t)\phi(s_v(t))|\Theta(t)\right].$$

An identical argument as in Eq. (17) can be carried out since $r_v(t)$, $\phi(s_v(t)), \phi(s'_v(t))$ have the same distribution for all agents v.

Similarly to before, we have the following proposition for the local state model. Recall here our notation of putting a bar to denote the network-wide average.

Proposition 2. *Suppose Assumptions 1-2 hold, and suppose that $\{\theta_v(t)\}_{v\in\mathcal{V}}$ are generated by Algorithm 2. Then,*

(a) $\bar{h}(t)$ is a linear function of $\bar{\theta}(t)$:

$$\bar{h}(t) = b - A\bar{\theta}(t).$$

(b) The conditional expectation of $\bar{m}(t)$ given $\Theta(t)$ is equal to zero:

$$E[\bar{m}(t)|\Theta(t)] = 0. \tag{31}$$

The proof of this proposition is essentially identical to the proof of Proposition 1 and we omit it.

Our next step is to prove a recurrence relation satisfied by the average of the iterates, stated as the following lemma. The key differences between this lemma and the previously-proved version in the global state model is that the quantity σ^2 that appears in this recursion will now be divided by N, at the cost of the addition of an extra term we will have to deal with. Recall that θ^*_{lc} is the fixed point of TD(0) on the MDP $(\mathcal{S}, \mathcal{V}, \mathcal{A}, \mathcal{P}, r, \gamma)$.

Lemma 4. *Suppose Assumptions 1-2 hold. Further suppose that $\{\theta_v\}_{v\in\mathcal{V}}$ are generated by Algorithm 2. For $t \in \mathbb{N}_0$, we have that*

$$\begin{aligned} E\left[\left\|\bar{\theta}(t+1) - \theta^*_{\text{lc}}\right\|_2^2\right] \leq & E\left[\left\|\bar{\theta}(t) - \theta^*_{\text{lc}}\right\|_2^2\right] + \alpha_t^2\left(\frac{2\sigma^2}{N} + \frac{8}{N}\sum_{v\in\mathcal{V}} E\left[\|V_{\theta_v(t)} - V_{\theta^*_{\text{lc}}}\|_D^2\right]\right) \\ & - 2\alpha_t E\left[(1-\gamma)\left\|V_{\theta^*_{\text{lc}}} - V_{\bar{\theta}(t)}\right\|_D^2 + \gamma\left\|V_{\theta^*_{\text{lc}}} - V_{\bar{\theta}(t)}\right\|_{\text{Dir}}^2\right], \end{aligned}$$

Proof of Lemma 4. Similarly to our argument for the global state model,

$$\bar{\theta}(t+1) = \bar{\theta}(t) + \alpha_t\left[\bar{h}(t) + \bar{m}(t)\right].$$

Therefore,

$$\left\|\bar{\theta}(t+1) - \theta^*_{\text{lc}}\right\|_2^2 = \left\|\bar{\theta}(t) - \theta^*_{\text{lc}}\right\|_2^2 + 2\alpha_t\left[\bar{h}(t) + \bar{m}(t)\right]^T\left(\bar{\theta}(t) - \theta^*_{\text{lc}}\right) + \alpha_t^2\left\|\bar{h}(t) + \bar{m}(t)\right\|_2^2.$$

Taking expectations:

$$E\left[\left\|\bar{\theta}(t+1) - \theta^*_{\text{lc}}\right\|_2^2\right] = E\left[\left\|\bar{\theta}(t) - \theta^*_{\text{lc}}\right\|_2^2\right] + \alpha_t^2 E\left[\left\|\bar{h}(t) + \bar{m}(t)\right\|_2^2\right] - 2\alpha_t E\left[\left(\bar{h}(t) + \bar{m}(t)\right)^T\left(\theta^*_{\text{lc}} - \bar{\theta}(t)\right)\right]. \tag{32}$$

We consider the second term on the right hand side of Eq. (32). Following the definition of $\bar{h}(t)$ and $\bar{m}(t)$, we have that

$$E\left[\left\|\bar{h}(t) + \bar{m}(t)\right\|_2^2\right] = E\left[\left\|\frac{1}{N}\sum_{v\in\mathcal{V}}\delta_v(t)\phi(s_v(t))\right\|_2^2\right].$$

Plugging in the expression for TD error $\delta_v(t)$ with Eq. (9), we obtain

$$\begin{aligned} & E\left[\left\|\bar{h}(t) + \bar{m}(t)\right\|_2^2\right] \\ =& E\left[\left\|\frac{1}{N}\sum_{v\in\mathcal{V}} r_v(t)\phi(s_v(t)) - \frac{1}{N}\sum_{v\in\mathcal{V}}\phi(s_v(t))\left(\phi(s_v(t)) - \gamma\phi(s'_v(t))\right)^T\theta_v(t)\right\|_2^2\right] \\ =& E\left[\left\|\frac{1}{N}\sum_{v\in\mathcal{V}}\left[r_v(t) - \left(\phi(s_v(t)) - \gamma\phi(s'_v(t))\right)^T\theta^*_{\text{lc}}\right]\phi(s_v(t)) - \frac{1}{N}\sum_{v\in\mathcal{V}}\phi(s_v(t))\left(\phi(s_v(t)) - \gamma\phi(s'_v(t))\right)^T\left(\theta_v(t) - \theta^*_{\text{lc}}\right)\right\|_2^2\right] \end{aligned}$$

Denote

$$\boldsymbol{a}^* = \frac{1}{N}\sum_{v\in\mathcal{V}}\left[r_v(t) - \left(\phi(s_v(t)) - \gamma\phi(s_v'(t))\right)^T\theta_{\text{lc}}^*\right]\phi(s_v(t)), \tag{33}$$

$$\boldsymbol{b}^* = \frac{1}{N}\sum_{v\in\mathcal{V}}\phi(s_v(t))\left(\phi(s_v(t)) - \gamma\phi(s_v'(t))\right)^T(\theta_v(t) - \theta_{\text{lc}}^*). \tag{34}$$

Using inequality $\|\boldsymbol{a}^* - \boldsymbol{b}^*\|^2 \le 2\|\boldsymbol{a}^*\|^2 + 2\|\boldsymbol{b}^*\|^2$, we obtain

$$E\left[\left\|\bar{h}(t) + \bar{m}(t)\right\|_2^2\right] \le 2E\left[\|\boldsymbol{a}^*\|^2\right] + 2E\left[\|\boldsymbol{b}^*\|^2\right]. \tag{35}$$

We first bound $E\left[\|\boldsymbol{a}^*\|^2\right]$. Let $\boldsymbol{a}^* = \frac{1}{N}\sum_{v\in\mathcal{V}}\rho_v$, where

$$\rho_v = \left[r_v(t) - \left(\phi(s_v(t)) - \gamma\phi(s_v'(t))\right)^T\theta_{\text{lc}}^*\right]\phi(s_v(t)).$$

Recall that, in the local state model, we are just running TD(0) on the identical MDP with the identical rewards across nodes. Hence the quantity θ_{lc}^* satisfies Equation (11), i.e., for all agents v,

$$E\left[\left(r_v(t) - \left(\phi(s_v(t)) - \gamma\phi(s_v'(t))\right)^T\theta_{\text{lc}}^*\right)\phi(s_v(t))\right] = 0.$$

We thus have that $E[\rho_v] = 0$ for $v \in \mathcal{V}$. Then

$$\begin{aligned}
E\left[\|\boldsymbol{a}^*\|^2\right] =& E\left[(\boldsymbol{a}^*)^T\boldsymbol{a}^*\right] \\
=& E\left[\left(\frac{1}{N}\sum_{v\in\mathcal{V}}\rho_v\right)^T\left(\frac{1}{N}\sum_{v\in\mathcal{V}}\rho_v\right)\right] \\
=& \frac{1}{N^2}E\left[\sum_{v\in\mathcal{V}}\rho_v^T\rho_v + \sum_{v\neq v'}\rho_v^T\rho_{v'}\right] \\
=& \frac{1}{N}E\left[\rho_1^T\rho_1\right] + \frac{1}{N^2}\sum_{v\neq v'}E[\rho_v]^TE[\rho_{v'}] \\
=& \frac{1}{N}E\left[\rho_1^T\rho_1\right] = \frac{1}{N}E\left[\|\rho_1\|^2\right],
\end{aligned}$$

where the forth line follows because we are assuming the quantities $\{s_v(t)\}_{v\in\mathcal{V}}$ are generated i.i.d. across time steps t and the last line uses that $E[\rho_v] = 0$. Next,

$$\begin{aligned}
E\left[\|\rho_1\|^2\right] =& E\left[\left\|\left(r_1(t) - \left(\phi(s_1(t)) - \gamma\phi(s_1'(t))\right)^T\theta_{\text{lc}}^*\right)\phi(s_1(t))\right\|^2\right] \\
\le& E\left[\left(r_1(t) - \left(\phi(s_1(t)) - \gamma\phi(s_1'(t))\right)^T\theta_{\text{lc}}^*\right)^2\right] = \sigma^2,
\end{aligned}$$

where, the inequality follows Assumption 2 and recall that σ^2 is defined by,

$$\sigma^2 = E\left[\left(r_v(t) - \left(\phi(s_v(t)) - \gamma\phi(s_v'(t))\right)^T\theta_{\text{lc}}^*\right)^2\right].$$

We have thus shown:

$$E\left[\|\boldsymbol{a}^*\|^2\right] \le \frac{\sigma^2}{N}. \tag{36}$$

Our next step is to bound $E\left[\|\boldsymbol{b}^*\|^2\right]$, where $\boldsymbol{b}^*$ is defined in Eq.(34):

$$E\left[\|\boldsymbol{b}^*\|^2\right] = E\left[\left\|\frac{1}{N}\sum_{v\in\mathcal{V}}\left(\phi(s_v(t)) - \gamma\phi(s_v'(t))\right)^T(\theta_v(t) - \theta_{\text{lc}}^*)\phi(s_v(t))\right\|^2\right]$$

$$
\begin{aligned}
&= \frac{1}{N^2} E\left[\left\| \sum_{v\in\mathcal{V}} \left(\phi(s_v(t)) - \gamma\phi(s_v'(t))\right)^T (\theta_v(t) - \theta_{\text{lc}}^*)\phi(s_v(t))\right\|^2\right] \\
&\leq \frac{1}{N^2}\cdot N \cdot \sum_{v\in\mathcal{V}} E\left[\left(\left(\phi(s_v(t)) - \gamma\phi(s_v'(t))\right)^T (\theta_v(t) - \theta_{\text{lc}}^*)\right)^2 \|\phi(s_v(t))\|^2\right] \\
&\leq \frac{1}{N^2}\cdot N \cdot \sum_{v\in\mathcal{V}} E\left[\left(\left(\phi(s_v(t)) - \gamma\phi(s_v'(t))\right)^T (\theta_v(t) - \theta_{\text{lc}}^*)\right)^2\right] \\
&\leq \frac{4}{N} \sum_{v\in\mathcal{V}} E\left[\|V_{\theta_v(t)} - V_{\theta_{\text{lc}}^*}\|_D^2\right], \qquad (37)
\end{aligned}
$$

where the first inequality uses $\|\sum_{i=1}^N a_i x_i\|^2 \leq N\sum_{i=1}^N a_i^2\|x_i\|^2$; the second inequality follows Assumption 2; and the last line follows from the proof of Lemma 5 in Bhandari et al. (2018).

Plugging Eq. (36) and Eq. (37) into Eq. (35), we get

$$
E\left[\left\|\bar{h}(t) + \bar{m}(t)\right\|_2^2\right] \leq \frac{2\sigma^2}{N} + \frac{8}{N}\sum_{v\in\mathcal{V}} E\left[\|V_{\theta_v(t)} - V_{\theta_{\text{lc}}^*}\|_D^2\right]. \qquad (38)
$$

This equation bounds bounds one of the terms in Eq. (32). We next consider a different term in the same equation, namely we consider the third term on the right hand side of Eq. (32):

$$
\begin{aligned}
E\left[\left[\bar{h}(t) + \bar{m}(t)\right]^T (\theta_{\text{lc}}^* - \bar{\theta}(t))\right] &= E\left[E\left[\left(\bar{h}(t) + \bar{m}(t)\right)^T (\theta_{\text{lc}}^* - \bar{\theta}(t))|\Theta(t)\right]\right] \\
&= E\left[\bar{h}^T(t)(\theta_{\text{lc}}^* - \bar{\theta}(t))\right],
\end{aligned}
$$

where in the second equation, we use Eq. (31).

By Proposition 2 part (a), we have that $\bar{h}(t) = b - A\bar{\theta}(t)$. Furthermore, if we let $\bar{h}(\theta)$ denote the linear function $b - A\theta$, we have that $\bar{h}(\theta_{\text{lc}}^*) = 0$. Now applying Corollary 1 in Liu and Olshevsky (2020), we have that

$$
E\left[\left[\bar{h}(t) + \bar{m}(t)\right]^T (\theta_{\text{lc}}^* - \bar{\theta}(t))\right] = E\left[(1-\gamma)\left\|V_{\theta_{\text{lc}}^*} - V_{\bar{\theta}(t)}\right\|_D^2 + \gamma\left\|V_{\theta_{\text{lc}}^*} - V_{\bar{\theta}(t)}\right\|_{\text{Dir}}^2\right] \qquad (39)
$$

Combining equations (32), (39), and (38), we obtain

$$
\begin{aligned}
E\left[\left\|\bar{\theta}(t+1) - \theta_{\text{lc}}^*\right\|_2^2\right] \leq & E\left[\left\|\bar{\theta}(t) - \theta_{\text{lc}}^*\right\|_2^2\right] + \alpha_t^2\left(\frac{2\sigma^2}{N} + \frac{8}{N}\sum_{v\in\mathcal{V}} E\left[\|V_{\theta_v(t)} - V_{\theta_{\text{lc}}^*}\|_D^2\right]\right) \\
& - 2\alpha_t E\left[(1-\gamma)\left\|V_{\theta_{\text{lc}}^*} - V_{\bar{\theta}(t)}\right\|_D^2 + \gamma\left\|V_{\theta_{\text{lc}}^*} - V_{\bar{\theta}(t)}\right\|_{\text{Dir}}^2\right].
\end{aligned}
$$

■

With this lemma in place, we are now ready to provide a proof of Theorem 2. This will be similar, but not identical, to the proof of Theorem 1, as the recursion we have just proved as an extra term multiplying $O(\alpha_t^2)$ relative to Lemma 3.

Proof of Theorem 2. Starting from Lemma 4,

$$
\begin{aligned}
E\left[\left\|\bar{\theta}(t+1) - \theta_{\text{lc}}^*\right\|_2^2\right] \leq & E\left[\left\|\bar{\theta}(t) - \theta_{\text{lc}}^*\right\|_2^2\right] + \alpha_t^2\left(\frac{2\sigma^2}{N} + \frac{8}{N}\sum_{v\in\mathcal{V}} E\left[\|V_{\theta_v(t)} - V_{\theta_{\text{lc}}^*}\|_D^2\right]\right) \\
& - 2\alpha_t E\left[(1-\gamma)\left\|V_{\theta_{\text{lc}}^*} - V_{\bar{\theta}(t)}\right\|_D^2 + \gamma\left\|V_{\theta_{\text{lc}}^*} - V_{\bar{\theta}(t)}\right\|_{\text{Dir}}^2\right], \qquad (40)
\end{aligned}
$$

we first consider the bound for the term $\sum_{t=1}^T\sum_{v=1}^N E\left[\|V_{\theta_v(t)} - V_{\theta_{\text{lc}}^*}\|_D^2\right]$. We can plug in that $N = 1$ into Lemma 4 to obtain the next inequality:

$$
E\left[\|\theta_v(t+1) - \theta_{\text{lc}}^*\|_2^2\right] \leq E\left[\|\theta_v(t) - \theta_{\text{lc}}^*\|_2^2\right] + \alpha_t^2\left(2\sigma^2 + 8E\left[\|V_{\theta_v(t)} - V_{\theta_{\text{lc}}^*}\|_D^2\right]\right)
$$

$$-2\alpha_t E\left[(1-\gamma)\left\|V_{\theta^*_{\text{lc}}}-V_{\theta_v(t)}\right\|_D^2+\gamma\left\|V_{\theta^*_{\text{lc}}}-V_{\theta_v(t)}\right\|_{\text{Dir}}^2\right].$$

If the sequence of step-sizes are non-increasing and satisfies

$$8\alpha_t^2-2\alpha_t(1-\gamma)\leq-\alpha_t(1-\gamma),$$

then we obtain

$$\alpha_t E\left[(1-\gamma)\left\|V_{\theta^*_{\text{lc}}}-V_{\theta_v(t)}\right\|_D^2+2\gamma\left\|V_{\theta^*_{\text{lc}}}-V_{\theta_v(t)}\right\|_{\text{Dir}}^2\right]\leq E\left[\|\theta_v(t)-\theta^*_{\text{lc}}\|_2^2\right]-E\left[\|\theta_v(t+1)-\theta^*_{\text{lc}}\|_2^2\right]+2\alpha_t^2\sigma^2.$$

Since $E\left[2\gamma\left\|V_{\theta^*_{\text{lc}}}-V_{\theta_v(t)}\right\|_{\text{Dir}}^2\right]$ is non-negative, it now follows that

$$\alpha_t E\left[(1-\gamma)\left\|V_{\theta^*_{\text{lc}}}-V_{\theta_v(t)}\right\|_D^2\right]\leq E\left[\|\theta_v(t)-\theta^*_{\text{lc}}\|_2^2\right]-E\left[\|\theta_v(t+1)-\theta^*_{\text{lc}}\|_2^2\right]+2\alpha_t^2\sigma^2.$$

Multiplying α_t on both sides and summing over t, we have

$$\begin{aligned}
&\sum_{t=0}^{T-1}\alpha_t^2 E\left[(1-\gamma)\left\|V_{\theta^*_{\text{lc}}}-V_{\theta_v(t)}\right\|_D^2\right]\\
=&\alpha_0 E\left[\|\theta_v(0)-\theta^*_{\text{lc}}\|_2^2\right]+\sum_{t=1}^{T-1}(\alpha_{t-1}-\alpha_t)E\left[\|\theta_v(t)-\theta^*_{\text{lc}}\|_2^2\right]-\alpha_{T-1}E\left[\|\theta_v(T)-\theta^*_{\text{lc}}\|_2^2\right]+2\sum_{t=0}^{T-1}\alpha_t^3\sigma^2\\
\leq&\alpha_0 E\left[\|\theta_v(0)-\theta^*_{\text{lc}}\|_2^2\right]+2\sum_{t=0}^{T-1}\alpha_t^3\sigma^2,
\end{aligned}$$

where the last inequality is because that $\{\alpha_t\}_t$ are non-increasing step-sizes. Summing over agents v , we get

$$\begin{aligned}
\sum_{v=1}^{N}\sum_{t=0}^{T-1}\alpha_t^2 E\left[(1-\gamma)\left\|V_{\theta^*_{\text{lc}}}-V_{\theta_v(t)}\right\|_D^2\right]\leq&\sum_{v=1}^{N}\alpha_0 E\left[\|\theta_v(0)-\theta^*_{\text{lc}}\|_2^2\right]+2\sum_{v=1}^{N}\sum_{t=0}^{T-1}\alpha_t^3\sigma^2\\
\leq&N\alpha_0\hat{R}_0+2N\sum_{t=0}^{T-1}\alpha_t^3\sigma^2,
\end{aligned}\tag{41}$$

where $\hat{R}_0=\max_{v\in\mathcal{V}}E\left[\left\|\theta_v(0)-\theta^*_{\text{lc}}\right\|_2^2\right]$. With this equation in place, we now turn to the proof of all the parts of the theorem.

Proof of part (a): We consider the constant step-size sequence $\alpha_0=\cdots=\alpha_T\leq(1-\gamma)/8$. Then let α denote the constant step-size. Plugging into Eq. (40) and rearranging it, we get

$$\begin{aligned}
2\alpha E\left[(1-\gamma)\left\|V_{\theta^*_{\text{lc}}}-V_{\bar{\theta}(t)}\right\|_D^2+\gamma\left\|V_{\theta^*_{\text{lc}}}-V_{\bar{\theta}(t)}\right\|_{\text{Dir}}^2\right]\leq&E\left[\left\|\bar{\theta}(t)-\theta^*_{\text{lc}}\right\|_2^2\right]-E\left[\left\|\bar{\theta}(t+1)-\theta^*_{\text{lc}}\right\|_2^2\right]\\
&+\alpha^2\left(\frac{2\sigma^2}{N}+\frac{8}{N}\sum_{v\in\mathcal{V}}E\left[\|V_{\theta_v(t)}-V_{\theta^*_{\text{lc}}}\|_D^2\right]\right).
\end{aligned}$$

Summing over t gives

$$\begin{aligned}
&2\sum_{t=0}^{T-1}\alpha E\left[(1-\gamma)\left\|V_{\theta^*_{\text{lc}}}-V_{\bar{\theta}(t)}\right\|_D^2+\gamma\left\|V_{\theta^*_{\text{lc}}}-V_{\bar{\theta}(t)}\right\|_{\text{Dir}}^2\right]\\
\leq&E\left[\left\|\bar{\theta}(0)-\theta^*_{\text{lc}}\right\|_2^2\right]-E\left[\left\|\bar{\theta}(T)-\theta^*_{\text{lc}}\right\|_2^2\right]+\frac{2T\alpha^2\sigma^2}{N}+\frac{8}{N}\sum_{t=0}^{T-1}\sum_{v\in\mathcal{V}}\alpha^2E\left[\|V_{\theta_v(t)}-V_{\theta^*_{\text{lc}}}\|_D^2\right]\\
\leq&E\left[\left\|\bar{\theta}(0)-\theta^*_{\text{lc}}\right\|_2^2\right]+\frac{2T\alpha^2\sigma^2}{N}+\frac{8}{N}\sum_{t=0}^{T-1}\sum_{v\in\mathcal{V}}\alpha^2E\left[\|V_{\theta_v(t)}-V_{\theta^*_{\text{lc}}}\|_D^2\right]\\
\leq&E\left[\left\|\bar{\theta}(0)-\theta^*_{\text{lc}}\right\|_2^2\right]+\frac{2T\alpha^2\sigma^2}{N}+\frac{8}{N(1-\gamma)}\left(N\alpha\hat{R}_0+2N\sum_{t=0}^{T-1}\alpha^3\sigma^2\right)
\end{aligned}$$

$$\leq E\left[\left\|\bar{\theta}(0)-\theta_{\text{lc}}^*\right\|_2^2\right]+\frac{2T\alpha^2\sigma^2}{N}+\frac{8\alpha}{1-\gamma}\left(\hat{R}_0+2T\alpha^2\sigma^2\right)$$

where the second inequality follows that $E\left[\left\|\bar{\theta}(T)-\theta_{\text{lc}}^*\right\|_2^2\right]$ is non-negative; the third inequality uses Eq. (41).

Now dividing by 2α on both sides:

$$\sum_{t=0}^{T-1} E\left[(1-\gamma)\left\|V_{\theta_{\text{lc}}^*}-V_{\bar{\theta}(t)}\right\|_D^2+\gamma\left\|V_{\theta_{\text{lc}}^*}-V_{\bar{\theta}(t)}\right\|_{\text{Dir}}^2\right]\leq\frac{1}{2\alpha}E\left[\left\|\bar{\theta}(0)-\theta_{\text{lc}}^*\right\|_2^2\right]+\frac{T\alpha\sigma^2}{N}+\frac{4}{1-\gamma}\left(\hat{R}_0+2T\alpha^2\sigma^2\right).$$

Let $\hat{\theta}(T)=\frac{1}{T}\sum_{t=1}^{T}\bar{\theta}(t)$. Then, by convexity

$$\begin{aligned}E\left[(1-\gamma)\left\|V_{\theta_{\text{lc}}^*}-V_{\hat{\theta}(T)}\right\|_D^2+\gamma\left\|V_{\theta_{\text{lc}}^*}-V_{\hat{\theta}(T)}\right\|_{\text{Dir}}^2\right]&\leq\frac{1}{T}\sum_{t=1}^{T}E\left[(1-\gamma)\left\|V_{\theta_{\text{lc}}^*}-V_{\bar{\theta}(t)}\right\|_D^2+\gamma\left\|V_{\theta_{\text{lc}}^*}-V_{\bar{\theta}(t)}\right\|_{\text{Dir}}^2\right]\\&\leq\frac{1}{T}\left(\frac{1}{2\alpha}E\left[\left\|\bar{\theta}(0)-\theta_{\text{lc}}^*\right\|_2^2\right]+\frac{4\hat{R}_0}{1-\gamma}\right)+\frac{\alpha\sigma^2}{N}+\frac{8\alpha^2\sigma^2}{1-\gamma},\end{aligned}$$

which is what we wanted to show.

Proof of part (b): We now consider the step-size $\alpha_0=\cdots=\alpha_T=\frac{1}{\sqrt{T}}$. When $T\geq\frac{64}{(1-\gamma)^2}$, it can be observed that $\alpha=\frac{1}{\sqrt{T}}\leq\frac{1-\gamma}{8}$. As a consequence of part (a), it is immediate that,

$$E\left[(1-\gamma)\left\|V_{\theta_{\text{lc}}^*}-V_{\hat{\theta}(T)}\right\|_D^2+\gamma\left\|V_{\theta_{\text{lc}}^*}-V_{\hat{\theta}(T)}\right\|_{\text{Dir}}^2\right]\leq\frac{1}{2\sqrt{T}}\left(E\left[\left\|\bar{\theta}(0)-\theta_{\text{lc}}^*\right\|_2^2\right]+\frac{2\sigma^2}{N}\right)+\frac{1}{T}\left(\frac{4\hat{R}_0+8\sigma^2}{1-\gamma}\right),$$

which is what we wanted to show.

Proof of part (c): Using that $\gamma\left\|V_{\theta_{\text{lc}}^*}-V_{\bar{\theta}(t)}\right\|_{\text{Dir}}^2$ is non-negative and rearranging Eq. (40), we have

$$E\left[\left\|\bar{\theta}(t+1)-\theta_{\text{lc}}^*\right\|_2^2\right]\leq E\left[\left\|\bar{\theta}(t)-\theta_{\text{lc}}^*\right\|_2^2\right]+\alpha_t^2\left(\frac{2\sigma^2}{N}+\frac{8}{N}\sum_{v\in\mathcal{V}}E\left[\left\|V_{\theta_v(t)}-V_{\theta_{\text{lc}}^*}\right\|_D^2\right]\right)-2\alpha_t(1-\gamma)E\left\|V_{\theta_{\text{lc}}^*}-V_{\bar{\theta}(t)}\right\|_D^2.$$

Applying Lemma 1 in Bhandari et al. (2018), which states that

$$\sqrt{\omega}\|\theta\|_2\leq\|V_\theta\|_D\leq\|\theta\|_2,$$

we get

$$E\left[\left\|\bar{\theta}(t+1)-\theta_{\text{lc}}^*\right\|_2^2\right]\leq(1-2\alpha_t(1-\gamma)\omega)E\left[\left\|\bar{\theta}(t)-\theta_{\text{lc}}^*\right\|_2^2\right]+\alpha_t^2\left(\frac{2\sigma^2}{N}+\frac{8}{N}\sum_{v\in\mathcal{V}}E\left[\left\|V_{\theta_v(t)}-V_{\theta_{\text{lc}}^*}\right\|_D^2\right]\right).\tag{42}$$

We first consider the last term on the right hand side, i.e., $E\left[\left\|V_{\theta_v(t)}-V_{\theta_{\text{lc}}^*}\right\|_D^2\right]$. Since each agent in the system executes the classical TD(0) at time t for $t\in\mathbb{N}_0$, then by part (c) of Theorem 2 and Lemma 1 in (Bhandari et al., 2018), for $v\in\mathcal{V}$, we have that

$$E\left[\left\|V_{\theta_v(t)}-V_{\theta_{\text{lc}}^*}\right\|_D^2\right]\leq E\left[\left\|\theta_v(t)-\theta_{\text{lc}}^*\right\|_2^2\right]\leq\frac{\hat{\zeta}}{t+\tau},$$

where

$$\hat{\zeta}=\max\left\{2\alpha^2\sigma^2,\tau\hat{R}_0\right\},$$

recall that $\hat{R}_0=\max_{v\in\mathcal{V}}E\left[\left\|\theta_v(0)-\theta_{\text{lc}}^*\right\|_2^2\right]$ and $\tau=\frac{16}{(1-\gamma)^2\omega}$. Hence,

$$\frac{8}{N}\sum_{v\in\mathcal{V}}E\left[\left\|V_{\theta_v(t)}-V_{\theta_{\text{lc}}^*}\right\|_D^2\right]\leq\frac{8\hat{\zeta}}{t+\tau},$$

and plugging it into Eq. (42), we can obtain

$$E\left[\left\|\bar{\theta}(t+1)-\theta_{\text{lc}}^*\right\|_2^2\right] \leq (1-2\alpha_t(1-\gamma)\omega)E\left[\left\|\bar{\theta}(t)-\theta_{\text{lc}}^*\right\|_2^2\right]+\alpha_t^2\left(\frac{2\sigma^2}{N}+\frac{8\hat{\zeta}}{t+\tau}\right)$$
$$=\left(1-\frac{4}{t+\tau}\right)E\left[\left\|\bar{\theta}(t)-\theta_{\text{lc}}^*\right\|_2^2\right]+\frac{2\alpha^2\sigma^2/N}{(t+\tau)^2}+\frac{8\alpha^2\hat{\zeta}}{(t+\tau)^3},$$

where we use that $\alpha_t=\frac{\alpha}{t+\tau}$ with $\alpha=\frac{2}{(1-\gamma)\omega}$ and $\tau=\frac{16}{(1-\gamma)^2\omega}$ to get the last line. This recursion implies that

$$E\left[\left\|\bar{\theta}(t+1)-\theta_{\text{lc}}^*\right\|_2^2\right] \leq \prod_{i=0}^{t}\left(1-\frac{4}{t+\tau-i}\right)E\left[\left\|\bar{\theta}(0)-\theta_{\text{lc}}^*\right\|_2^2\right]+\frac{2\alpha^2\sigma^2}{N}\sum_{i=0}^{t}\left[\frac{1}{(t+\tau-i)^2}\prod_{l=0}^{i-1}\left(1-\frac{4}{t+\tau-l}\right)\right]$$
$$+8\alpha^2\hat{\zeta}\sum_{i=0}^{t}\left[\frac{1}{(t+\tau-i)^3}\prod_{l=0}^{i-1}\left(1-\frac{4}{t+\tau-l}\right)\right]. \tag{43}$$

Consider the product

$$\prod_{i=0}^{t}\left(1-\frac{4}{t+\tau-i}\right)=\frac{t+\tau-4}{t+\tau}\cdot\frac{t+\tau-5}{t+\tau-1}\cdot\cdots\cdot\frac{\tau-4}{\tau}$$
$$=\frac{(\tau-1)(\tau-2)(\tau-3)(\tau-4)}{(t+\tau)(t+\tau-1)(t+\tau-2)(t+\tau-3)}$$
$$=\frac{\tau-1}{t+\tau}\cdot\frac{\tau-2}{t+\tau-1}\cdot\frac{\tau-3}{t+\tau-2}\cdot\frac{\tau-4}{t+\tau-3}$$
$$<\left(\frac{\tau-1}{t+\tau}\right)^4. \tag{44}$$

The last inequality follows because that last three terms in equation is smaller than $(\tau-1)/(t+\tau)$. Indeed, for $i=1,2,3$, we have that

$$\frac{\tau-i-1}{t+\tau-i}=\frac{\tau-1}{t+\tau}+\left(\frac{\tau-i-1}{t+\tau-i}-\frac{\tau-1}{t+\tau}\right)$$
$$=\frac{\tau-1}{t+\tau}+\frac{(t+\tau)(\tau-i-1)-(\tau-1)(t+\tau-i)}{(t+\tau)(t+\tau-i)}$$
$$=\frac{\tau-1}{t+\tau}-\frac{(i+1)t}{(t+\tau)(t+\tau-i)}$$
$$<\frac{\tau-1}{t+\tau}.$$

For the other product $\prod_{l=0}^{i-1}\left(1-\frac{4}{t+\tau-l}\right)$ in Eq. (43), applying the same method of Eq. (44), we thus have

$$\prod_{l=0}^{i-1}\left(1-\frac{4}{t+\tau-l}\right)\leq\left(\frac{t+\tau-i}{t+\tau}\right)^4. \tag{45}$$

Using Eq. (44) and Eq. (45), Eq. (43) becomes

$$E\left[\left\|\bar{\theta}(t+1)-\theta_{\text{lc}}^*\right\|_2^2\right]$$
$$\leq\left(\frac{\tau-1}{t+\tau}\right)^4 E\left[\left\|\bar{\theta}(0)-\theta_{\text{lc}}^*\right\|_2^2\right]+\frac{2\alpha^2\sigma^2}{N}\sum_{i=0}^{t}\left[\frac{1}{(t+\tau-i)^2}\left(\frac{t+\tau-i}{t+\tau}\right)^4\right]+8\alpha^2\hat{\zeta}\sum_{i=0}^{t}\left[\frac{1}{(t+\tau-i)^3}\left(\frac{t+\tau-i}{t+\tau}\right)^4\right]$$
$$\leq\left(\frac{\tau-1}{t+\tau}\right)^4 E\left[\left\|\bar{\theta}(0)-\theta_{\text{lc}}^*\right\|_2^2\right]+\frac{2\alpha^2\sigma^2}{N}\sum_{i=0}^{t}\frac{(t+\tau-i)^2}{(t+\tau)^4}+8\alpha^2\hat{\zeta}\sum_{i=0}^{t}\frac{t+\tau-i}{(t+\tau)^4}. \tag{46}$$

Next, we bound summations of the second and the third on the right hand side $\sum_{i=0}^{t}\frac{(t+\tau-i)^2}{(t+\tau)^4}$ and $\sum_{i=0}^{t}\frac{t+\tau-i}{(t+\tau)^4}$ separately.

For the summation in the second term, i.e., $\sum_{i=0}^{t}\frac{(t+\tau-i)^2}{(t+\tau)^4}$, we have

$$\sum_{i=0}^{t}\frac{(t+\tau-i)^2}{(t+\tau)^4}=\sum_{i=0}^{t}\frac{(i+\tau)^2}{(t+\tau)^4}$$

$$
\begin{aligned}
&\leq \frac{1}{(t+\tau)^4}\sum_{i=1}^{t+\tau} i^2 \\
&= \frac{1}{(t+\tau)^4}\cdot\frac{1}{6}\cdot(t+\tau)(t+\tau+1)(2t+2\tau+1) \\
&\leq \frac{1}{6(t+\tau)^4}(t+\tau)\left(2(t+\tau)\right)\left(3(t+\tau)\right) \\
&= \frac{1}{t+\tau}.
\end{aligned} \tag{47}
$$

For the summation in the third term, i.e., $\sum_{i=0}^{t}\frac{t+\tau-i}{(t+\tau)^4}$, it is immediately that

$$
\begin{aligned}
\sum_{i=0}^{t}\frac{t+\tau-i}{(t+\tau)^4} &= \sum_{i=0}^{t}\frac{i+\tau}{(t+\tau)^4} \\
&\leq \frac{1}{(t+\tau)^4}\sum_{i=0}^{t}(i+\tau) \\
&= \frac{1}{(t+\tau)^4}\cdot\frac{(t+1)(t+2\tau)}{2} \\
&\leq \frac{1}{2(t+\tau)^4}\left((t+\tau)\right)\left(2(t+\tau)\right) \\
&= \frac{1}{(t+\tau)^2}.
\end{aligned} \tag{48}
$$

Therefore, combining Eq. (46), Eq. (47), and Eq. (48), we obtain

$$
E\left[\left\|\bar{\theta}(t+1)-\theta^*_{\text{lc}}\right\|_2^2\right] \leq \frac{2\alpha^2\sigma^2/N}{t+\tau} + \frac{8\alpha^2\hat{\zeta}}{(t+\tau)^2} + \frac{(\tau-1)^4 E\left[\left\|\bar{\theta}(0)-\theta^*_{\text{lc}}\right\|_2^2\right]}{(t+\tau)^4}.
$$

■

D Numerical Experiments

In this section, we provide details of the simulations done in the main body of the paper. These simulations were done on OpenAI control problems and GridWorld. We first give the details of the Gridworld setup, which is fairly standard.

D.1 Settings on the Gridworld MDP

In this subsection, we introduce the specific problem settings for the grid-world MDP. We consider a 4×4 grid, where the states are $\mathcal{S}=\{1,2,\ldots,16\}$. There are four possible actions for each state, $\mathcal{A}$ = {left, right, up, down}. If the action leads out of the grid, then the next state will remain to be the current state. Set the discount factor in the MDP to be 0.8, $\gamma = 0.8$. Let deterministic rewards $r(s,a)$ be randomly chosen from a normal distribution $\mathcal{N}(1,100)$.

1	2	3	4
5	6	7	8
9	10	11	12
13	14	15	16

Table 1: State space of grid world

In this experiment, we will consider a random policy, i.e., each agent chooses an action from the 4 possible actions uniformly at random. Feature vectors are generated as $\phi(s)=(1,0,0,0)^T$ for four upper-left states $s\in\{1,2,5,6\}$; $\phi(s)=(0,1,0,0)^T$ for four upper-right states $s\in\{3,4,7,8\}$; $\phi(s)=(0,0,1,0)^T$ for four lower-left states $s\in\{9,10,13,14\}$; and $\phi(s)=(0,0,0,1)^T$ for four lower-right states $s\in\{11,12,15,16\}$. In this case, for any parameter $\theta\in\mathbb{R}^4$, we have $V_\theta(s)=\theta^T\phi(s)$ as the approximation for the value function of state s. Furthermore, samples are generated i.i.d and are equally likely chosen from the state space.

Due to the relatively small state space and the fixed policy, it is simply for us to use both the transition matrix P, and further get stationary distribution π. We can also get θ_{lc}^* by solving Eq.(11). Therefore, the left-hand side of Theorem 2(a):

$$E\left[(1-\gamma)\left\|V_{\theta_{\text{lc}}^*}-V_{\hat{\theta}(T)}\right\|_D^2+\gamma\left\|V_{\theta_{\text{lc}}^*}-V_{\hat{\theta}(T)}\right\|_{\text{Dir}}^2\right],$$

where recall that norm $\|\cdot\|_D$ and semi-norm $\|\cdot\|_{\text{Dir}}$ is defined as Eq.(2) and Eq.(3) with stationary distribution π, can be obtained exactly for any constant step-size.

D.2 Settings on the Classic Control Problems

Unlike the grid world case, for a more involved RL problem we do not have an explicit solution for the final limit θ_{lc}^* that we can compare to. In other words, it is not possible to plot the left-hand side of Theorem 2 which contains the optimal parameter vector θ_{lc}^*. As a consequence, we use the empirical variances among several runs of the method, which is a plausible measure for accuracy of the method in place of the left-hand side of Theorem 2(a).

For classic control problems, we use the tile coding Sutton and Barto (2018) to deal with multi-dimensional continuous spaces. A tiling is a partition of the state space and a tile is an element of a partition. We set the parameter dimension K be the total number of tiles among all tilings. The feature vector of state s, $\phi(s)\in\mathbb{R}^K$, is a vector has one component for each tile in each tiling. For a state s, it falls in exactly one tile for each tiling. The element in the $\phi(s)$ corresponding to the tile that s falls within is one and all others are zeros. Hence, the number of ones in the feature vector is always equal to the number of tilings.

The numbers of tiling and grid are similar to those used in Lakshminarayanan and Szepesvari (2018). We use 5 tilings, and each tiling has 7×7 grids for two dimensional MountainCar-v1 and MountainCarContinous-v0; $3\times 3\times 3$ grids for three dimensional Pendulum-v1; 3^4 grids for four dimensional CartPole-v1; and 2^6 grids for six dimensional Acrobot-v1. We considered uniform random policy for all problems. The discount factor was 0.8. The initial condition $\theta_v(0)$ were sampled form standard normal distribution and for a fixed initialization. We applied Algorithm 2 several times and then computed the empirical variance in the final estimates. As in the previous subsection, the step-size were chosen to be constant. In Figure 1, each subplot shows the empirical variance for many different choices of α with $N=1$ and $N=\frac{1-\gamma}{8\alpha}$. As we expected, all the blue lines for $N=\frac{1-\gamma}{8\alpha}$ are approximately quadratic in shapes while all the red lines are generally linear, consistent with our theoretical results.

D.3 TD Errors of Distributed TD Methods

We now discuss the details of the simulations that generated Figure 2, the comparison of our Algorithm 2 with earlier distributed TD methods from Doan et al. (2019a) and Wang et al. (2020a). We plot the averaged TD error among the network, i.e., $\frac{1}{N}\sum_{v\in\mathcal{V}}\delta_v(t)$ vs iteration t on the x-axis.

The number of agents $N=100$. For the distributed TD algorithms proposed in Doan et al. (2019a) and Wang et al. (2020a), the communication graph among agents is generated by the Erdos–Renyi model, which is connected. In the grid world case, all the settings are the same as stated in subsection D.1 except that deterministic rewards $r(s,a)$ be randomly chosen from a normal distribution $\mathcal{N}(1,0.01)$. The step-size is constant as $\alpha=0.3$. The parameters selection is mainly based on the parameters used in the simulations of Wang et al. (2020a).

For two dimensional MountainCar-v1 and MountainCarContinous-v0, we used 5 tilings, each tiling has 7×7 grids, and step-size $\alpha=0.3$. For three dimensional Pendulum-v1, we used 5 tilings, each tiling has $5\times 5\times 5$ grids, and step-size $\alpha=0.05$. For four dimensional CartPole-v1, we used 5 tilings, each tiling has 8^4 grids, and step-size $\alpha=0.3$. For six dimensional Acrobot-v1, we used 5 tilings, each tiling has 2^6 grids, and step-size $\alpha=0.05$. The constant step sizes for open AI gym problems are chosen from the the set $\Lambda=\{0.3,0.25,0.2,\cdots,0.05\}$. For each problem, we choose the largest step size from the set Λ such that all methods converge or the smallest of these step sizes 0.05 even if there exists one method does not converge for all step sizes in the set Λ. Note that the experiments of CartPole and Pendulum, the method in Wang et al. (2020a) does not converge with any step sizes in the set Λ; but Algorithm 2 in this paper and method in Doan et al. (2019a) do converge with all step sizes in the set Λ. We only show experimental result with $\alpha=0.05$ in the main text.