2024 American Control Conference (ACC)
July 8-12, 2024. Toronto, Canada

Finite-Time Analysis of Asynchronous Multi-Agent TD Learning

Nicolò Dal Fabbro, Arman Adibi, Aritra Mitra and George J. Pappas

Abstract— **Recent research endeavours have theoretically shown the beneficial effect of cooperation in multi-agent reinforcement learning (MARL). In a setting involving N agents, this beneficial effect usually comes in the form of an N-fold linear convergence speedup, i.e., a reduction - proportional to N - in the number of iterations required to reach a certain convergence precision. In this paper, we show for the first time that this speedup property also holds for a MARL framework subject to asynchronous delays in the local agents' updates. In particular, we consider a policy evaluation problem in which multiple agents cooperate to evaluate a common policy by communicating with a central aggregator. In this setting, we study the finite-time convergence of `AsyncMATD`, an asynchronous multi-agent temporal difference (TD) learning algorithm in which agents' local TD update directions are subject to asynchronous bounded delays. Our main contribution is providing a finite-time analysis of `AsyncMATD`, for which we establish a linear convergence speedup while highlighting the effect of time-varying asynchronous delays on the resulting convergence rate.**

I. Introduction

Training reinforcement learning (RL) algorithms is well-known to be a critically time consuming task that can require several data samples to achieve a desired level of accuracy. In an effort to reduce the sample-complexity of RL, recent research findings in multi-agent RL (MARL) have theoretically shown that, despite temporal correlations in each agent's observations, multi-agent cooperation provides collaborative performance gains [1], [2]. In a MARL framework where N agents communicate via a central entity - a common setup in the emerging federated RL (FRL) paradigm [3] - these gains manifest themselves in the form of an N-fold reduction in the sample-complexity relative to when each agent acts alone. To achieve such gains, the agents must communicate. However, in real-world multi-agent systems, inter-node communication imposes strong constraints, such as limited link capacity and transmission delays [4]–[6]. While the effects of these constraints have been widely analyzed for distributed optimization and federated learning (FL), analogous studies in the context of MARL are few and far between. Indeed, collaborative gains under communication constraints have only very recently been established for FRL under finite-rate links and wireless noisy channels [7], [8].

N. Dal Fabbro is with the Department of Information Engineering, University of Padova. Email: dalfabbron@dei.unipd.it. A. Adibi is with the Department of Electrical and Computer Engineering, Princeton University. Email: aadibi@princeton.edu. A. Mitra is with the Electrical and Computer Engineering Department, North Carolina State University. Email: amitra2@ncsu.edu. G. J. Pappas is with the Electrical and Systems Engineering Department, University of Pennsylvania. Email: pappasg@seas.upenn.edu.
This work was supported by NSF Award 1837253, ARL grant DCIST CRA W911NF-17-2-0181 and the Italian Ministry of Education, University and Research through the PRIN project no. 2017NS9FEY.

Given the above premise, our main goal in this paper is to contribute to the literature on *finite-time analyses of MARL under realistic communication models*. To that end, we consider an *asynchronous* MARL framework in which multiple agents cooperate to evaluate a common policy via temporal difference (TD) learning. To collaboratively evaluate the common policy, agents transmit their local TD update directions to a central server via up-link communication channels subject to asynchronous bounded delays. Asynchronous settings of this kind have been theoretically and empirically studied for FL and distributed optimization [5], [6], [9]. On the other hand, although asynchronous MARL implementations have shown promising empirical performance, like in the case of parallel actor-learner frameworks [10], [11], little to nothing is known regarding their non-asymptotic convergence guarantees and multi-agent collaborative gains. Indeed, the only existing study providing finite-sample convergence guarantees for asynchronous MARL [12] establishes collaborative performance gains only under a simplifying i.i.d. sampling assumption on the agents' observations, i.e., considering observations that are not temporally correlated. However, even in the non-delayed single-agent case, the major technical hurdle in the finite-time analysis of RL algorithms (like TD learning) relative to optimization/supervised learning, comes precisely from the fact that the agent's observation sequence is generated by a Markov chain, and, as such, exhibits temporal correlations. For such settings, finite-time convergence bounds have only recently been provided in [13], [14] via some fairly involved analysis. Thus, for the MARL setting we consider with Markovian sampling and asynchronous delays, establishing collaborative performance benefits turns out to be highly non-trivial. Nonetheless, providing such an analysis is the main contribution of this paper.

Contributions. We propose and analyze `AsyncMATD`, an asynchronous multi-agent TD learning algorithm in which transmission of agents' local TD update directions is subject to asynchronous bounded delays. At each iteration of `AsyncMATD`, the server updates the value function associated with the policy to be evaluated using linear function approximation. To do so, it uses potentially stale agents' TD update directions, where the staleness is a consequence of the delays. Our main contribution is to provide a finite-time analysis of `AsyncMATD` that clearly reveals the effect of the delay sequence on the resulting convergence rate; see Theorem 1. Remarkably, we establish an N-fold linear convergence speedup for `AsyncMATD` that shows the beneficial effect of collaboration *even in the presence of asynchronous delays in the agents' updates, and Markovian sampling*.

979-8-3503-8265-5/$31.00 ©2024 AACC

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 14:27:40 UTC from IEEE Xplore. Restrictions apply.

II. System Model and Problem Formulation

We consider a setting in which N agents independently interact with replicas of the *same* Markov Decision Process (MDP), which we denote by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$. For the MDP, we consider a finite set $\mathcal{S}$ of n states, a finite action space $\mathcal{A}$, a set of action-dependent Markov transition kernels $\mathcal{P}$, a reward function $\mathcal{R}$, and we denote by $\gamma \in (0,1)$ the discount factor. The goal is to collectively evaluate the value function associated with a policy $\mu : \mathcal{S} \to \mathcal{A}$. To do so, the agents communicate with a central aggregator (server) via up-link transmissions subject to asynchronous delays. We now review the key concepts regarding policy evaluation with value function approximation. Then, we formally describe the *asynchronous* communication model, outline our key objectives, and highlight the major technical challenges.

Policy Evaluation with Linear Function Approximation. The purpose of the considered *policy evaluation* setting is to evaluate, for each state $s \in \mathcal{S}$, the value function $\boldsymbol{V}_\mu(s)$, which is the discounted expected cumulative reward obtained by playing policy μ starting from initial state s. The policy μ interacts with the MDP $\mathcal{M}$ to generate a Markov Reward Process (MRP) characterized by a reward function $R_\mu : \mathcal{S} \to \mathbb{R}$, and a Markov chain with transition probability matrix $\mathbb{P}_\mu$. Formally, we have

$$\boldsymbol{V}_\mu(s) = \mathbb{E}\left[\sum_{k=0}^{\infty} \gamma^k R_\mu(s_k) | s_0 = s\right], \tag{1}$$

where s_k represents the state of the Markov chain associated with the kernel $\mathbb{P}_\mu$, when initialized from $s_0 = s$. The policy evaluation setting that we consider is *model-free*, i.e., we assume that the Markov transition kernel $\mathbb{P}_\mu$ and the reward function R_μ are *unknown*.

As is common, we consider a linear value function approximation setting in which the n-dimensional value function $\boldsymbol{V}_\mu$ is approximated by vectors in a linear subspace of $\mathbb{R}^n$ spanned by a set of m basis vectors $\{\boldsymbol{\phi}_\ell\}_{\ell=1}^m$. Resorting to value function approximation is particularly needed when the state space $\mathcal{S}$ is very large, and indeed usually $m \ll n$. Let $\boldsymbol{\Phi} \triangleq [\boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_m] \in \mathbb{R}^{n \times m}$ be the feature matrix. As is standard [13], we assume that the columns of $\boldsymbol{\Phi}$ are independent and that the rows are normalized. The parametric approximation $\hat{\boldsymbol{V}}_{\boldsymbol{\theta}}$ of $\boldsymbol{V}_\mu$ is given by $\boldsymbol{V}(\boldsymbol{\theta}) := \hat{\boldsymbol{V}}_{\boldsymbol{\theta}} = \boldsymbol{\Phi}\boldsymbol{\theta}$, where $\boldsymbol{\theta} \in \mathbb{R}^m$ is the approximation parameter. Denoting the s-th row of $\boldsymbol{\Phi}$ by $\boldsymbol{\phi}'_s$, the approximation of $\boldsymbol{V}_\mu(s)$ is given by $\hat{\boldsymbol{V}}_{\boldsymbol{\theta}}(s) = \langle \boldsymbol{\theta}, \boldsymbol{\phi}'_s \rangle$.

Asynchronous multi-agent TD Learning. In the setting described above, the collective aim of the agents is to estimate the best linear approximation parameter for $\boldsymbol{V}_\mu$ in the span of $\boldsymbol{\Phi}$; we denote this optimal parameter by $\boldsymbol{\theta}^*$. To achieve this goal, agents execute a multi-agent TD (MATD) variant of the classical TD(0) algorithm [15], as we describe next. We assume that agents initialize the MATD algorithm starting from the same state $s_0 \in \mathcal{S}$ and parameter $\boldsymbol{\theta}_0 \in \mathbb{R}^m$. In a *synchronous* MATD setting, at each time-step $k \in \mathbb{N}$, a global parameter $\boldsymbol{\theta}_k$ is broadcasted in down-link by the server to all agents. Each agent $i \in [N]$, in turn, based on the received parameter $\boldsymbol{\theta}_k$ and its k-th observation tuple $o_{i,k} = (s_{i,k}, r_{i,k}, s_{i,k+1})$, computes a local TD update direction $\mathbf{g}(\boldsymbol{\theta}_k, o_{i,k})$. Note that the observation tuple $o_{i,k}$ is obtained by playing policy μ at iteration k, which implies taking an action $a_{i,k} = \mu(s_{i,k})$, observing the next state $s_{i,k+1} \sim \mathbb{P}_\mu(\cdot | s_{i,k})$, and collecting an instantaneous reward $r_{i,k} = R_\mu(s_{i,k})$. Agent i's TD update direction is as follows:

$$\mathbf{g}(\boldsymbol{\theta}_k, o_{i,k}) = (r_{i,k} + \gamma \langle \boldsymbol{\phi}'_{s_{i,k+1}}, \boldsymbol{\theta}_k \rangle - \langle \boldsymbol{\phi}'_{s_{i,k}}, \boldsymbol{\theta}_k \rangle) \boldsymbol{\phi}'_{s_{i,k}}.$$

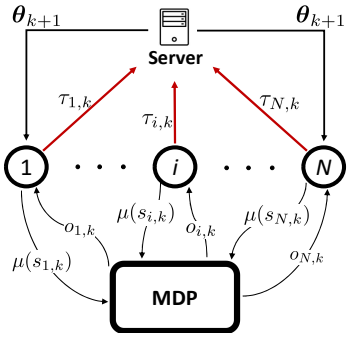


Fig. 1. System Model. Agents $1, \ldots, N$ cooperatively learn a common policy interacting with replicas of the same MDP. At each iteration k, the server uses the available delayed update directions with delays $\tau_{1,k}, \ldots, \tau_{N,k}$.

Although all agents play the same policy μ, and interact with replicas of the same MDP, they all experience different realizations of the local observation sequences $\{o_{i,k}\}$. We assume that the local observation sequences $\{o_{i,k}\}$ are *statistically independent* across the agents. For each agent i, however, the observations over time are correlated since they are all part of a single Markov chain.

Convergence speedup. Based on the independence property of the observation sequences, one would expect that exchanging agents' local TD update directions should help in reducing the variance in the estimate of $\boldsymbol{\theta}^*$. Achieving this variance reduction in the form of an N-fold linear convergence speedup is precisely the aim of multi-agent TD (MATD) learning. The model we described above is a synchronous MARL setting where at each time-step k, the server updates the model vector $\boldsymbol{\theta}_k$ using the average of the agents' local TD update directions from time-step k, transmitted over the up-link channels, i.e., *there are no delays*. Departing from this setting, we consider, for the first time, the crucial case in which agents' local TD update directions are computed in an asynchronous fashion. In what follows, we describe our model of asynchrony.

Asynchrony model of `AsyncMATD`. We now describe the model for asynchronous MATD learning that we consider in this paper, which is analogous to the models studied, for example, in FL [6], [9] and asynchronous MARL [12]. At each time-step k, the server updates the model vector $\boldsymbol{\theta}_k$ using the average of asynchronously delayed agents' local TD update directions. Specifically, for each agent i, at iteration k, the corresponding available TD update direction is subject to a *bounded* delay $\tau_{i,k}$. Define $t_{i,k} \triangleq (k - \tau_{i,k})_+$, where, for $x \in \mathbb{R}$, $(x)_+ = \max\{0, x\}$. The server updates

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 14:27:40 UTC from IEEE Xplore. Restrictions apply.

the model vector $\boldsymbol{\theta}_k$ according to the following rule:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \alpha \mathbf{v}_k, \tag{2}$$

where α is a constant step-size/learning rate, and

$$\mathbf{v}_k = \frac{1}{N}\sum_{i=1}^{N} \mathbf{g}(\boldsymbol{\theta}_{t_{i,k}}, o_{i,t_{i,k}}). \tag{3}$$

In this work, we assume that the down-link communication from the server to the agents is not subject to delays. Such an assumption is practically motivated by the fact that in most client-server architectures (e.g., wireless networks [4]), the main communication bottleneck comes from up-link transmissions, instead of down-link broadcasting. In the rest of the paper, we refer to the update rule in (2) as `AsyncMATD`. Note that the update direction $\mathbf{v}_k$ used by the server features iterates $\boldsymbol{\theta}_{t_{i,k}}$ and observations $o_{i,t_{i,k}}$ from potentially stale time-steps. Furthermore, the delays $\tau_{1,k}, \ldots, \tau_{N,k}$ can differ across agents. We make the following assumption on the delay sequence, which is common in the study of asynchronous distributed optimization and FL [5], [9].

Assumption 1. *There exists a positive integer $\tau_{max} > 0$ such that $0 \leq \tau_{i,k} \leq \tau_{max}$, for all i and for all k.*

Objective and Challenges. In the rest of the paper, we provide a *finite-time analysis* of `AsyncMATD`. This poses several challenges. In fact, even in the single-agent setting, providing a non-asymptotic analysis of `TD`(0) without performing intermediate projection steps is known to be challenging due to the temporal correlation between the Markov samples in the iterative learning process. Crucially, this challenge is absent in asynchronous stochastic optimization where one assumes i.i.d. data, precluding the use of techniques used in this line of work. For the analysis of `AsyncMATD`, we encounter further obstacles: the update rule involves the use of multiple *correlated iterates* $\boldsymbol{\theta}_{t_{i,k}}$, $i = 1, ..., N$, at which the local TD update directions are asynchronously computed. Indeed, note that, although the observation sequences $o_{i,k}$ are statistically independent across agents, the iterates used to compute the local TD update directions are all correlated. This aspect introduces the need for a much finer analysis when we want to provide finite-time convergence guarantees. Furthermore, unlike a single-agent setting, we aim to establish an N-fold linear convergence speedup while jointly dealing with the challenges outlined above. This necessitates a very careful study, which we illustrate below.

III. Main Result

In this section, we state and discuss our main result, which provides a non-asymptotic convergence bound for `AsyncMATD`. We start by providing the necessary assumptions and technical machinery. We make the standard assumption that the rewards are uniformly bounded, i.e., that $R_\mu(s) \leq \bar{r}, \forall s \in \mathcal{S}$ and for some $\bar{r} > 0$, which ensures that the value function in (1) is well-defined. Next, we make a key assumption which is crucial for the non-asymptotic analysis of TD learning algorithms [1], [7], [13], [14], [16].

Assumption 2. *The Markov chain induced by the policy μ is aperiodic and irreducible.*

Assumption 2 implies the existence of a unique stationary distribution π for the Markov transition matrix $\mathbb{P}_\mu$ [17]. Let $\mathbf{D} \in \mathbb{R}^{n\times n}$ be a diagonal matrix with entries given by the n elements of π and define $\boldsymbol{\Sigma} \triangleq \boldsymbol{\Phi}^\top \mathbf{D} \boldsymbol{\Phi}$. Since $\boldsymbol{\Phi}$ is assumed to be full column rank, $\boldsymbol{\Sigma}$ is full rank with a strictly positive smallest eigenvalue $\omega < 1$ that shows up in the convergence analysis. Next, we define the steady-state local TD direction:

$$\bar{\mathbf{g}}(\boldsymbol{\theta}) \triangleq \mathbb{E}_{s_{i,k}\sim\pi, s_{i,k+1}\sim\mathbb{P}_\mu(\cdot|s_{i,k})}[\mathbf{g}(\boldsymbol{\theta}, o_{i,k})], \forall \boldsymbol{\theta} \in \mathbb{R}^m. \tag{4}$$

The *deterministic* recursion $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \alpha \bar{\mathbf{g}}(\boldsymbol{\theta}_k)$ captures the limiting behavior of the `TD`(0) update rule. In [13], it was shown that the iterates generated by this recursion converge exponentially fast to $\boldsymbol{\theta}^*$, where $\boldsymbol{\theta}^*$ is the unique solution of the projected Bellman equation $\Pi_{\mathbf{D}} \mathcal{T}_\mu(\boldsymbol{\Phi}\boldsymbol{\theta}^*) = \boldsymbol{\Phi}\boldsymbol{\theta}^*$. Here, $\Pi_{\mathbf{D}}(\cdot)$ is the projection operator onto the subspace spanned by $\{\boldsymbol{\phi}_\ell\}_{\ell\in[m]}$ with respect to the inner product $\langle\cdot,\cdot\rangle_{\mathbf{D}}$, and $\mathcal{T}_\mu : \mathbb{R}^n \to \mathbb{R}^n$ is the policy-specific Bellman operator [16]. Next, we define the notion of mixing time τ_ϵ that will play a crucial role in our non-asymptotic analysis.

Definition 1. *Let τ_ϵ be the minimum time step such that $\|\mathbb{E}[\mathbf{g}(\boldsymbol{\theta}, o_{i,k})|o_{i,0}] - \bar{\mathbf{g}}(\boldsymbol{\theta})\| \leq \epsilon(\|\boldsymbol{\theta}\| + 1), \forall k \geq \tau_\epsilon, \forall \boldsymbol{\theta} \in \mathbb{R}^m, \forall i \in [N], \forall o_{i,0}$.*[1]

Assumption 2 implies that the Markov chain induced by μ mixes at a geometric rate [17], i.e., the total variation distance between $\mathbb{P}(s_{i,k} = \cdot|s_{i,0} = s)$ and the stationary distribution π decays exponentially fast $\forall k \geq 0, \forall i \in [N], \forall s \in \mathcal{S}$. This, in turn, implies the existence of some $K \geq 1$ such that τ_ϵ in Definition 1 satisfies $\tau_\epsilon \leq K\log(\frac{1}{\epsilon})$ [18]. We infer that the mixing time τ_ϵ scales logarithmically in the precision dictated by ϵ. For our purpose, we will set $\epsilon = \alpha^q$, where q is an integer satisfying $q \geq 2$. Unlike the centralized setting where $q = 1$ suffices [13], [14], to establish the linear speedup property, we will require $q \geq 2$. Henceforth, we will drop the subscript of $\epsilon = \alpha^q$ in τ_ϵ and simply refer to τ as the mixing time. Let $\delta_k^2 \triangleq \|\boldsymbol{\theta}^* - \boldsymbol{\theta}_k\|^2$ and define by $\sigma \triangleq \max\{1, \bar{r}, \|\boldsymbol{\theta}^*\|, \delta_0\}$ the "variance" of the observation model for our problem. We can now state our main result.

Theorem 1. *Consider the update rule of `AsyncMATD` in (2). There exist universal constants $C_0, C_1, C_2, C_3 \geq 1$, such that, for $\alpha \leq \frac{\omega(1-\gamma)}{C_0(\tau+\tau_{max})}$ and $T \geq \tau + 2\tau_{max}$,*

$$\begin{aligned}\mathbb{E}\left[\delta_T^2\right] \leq \exp\left(-\frac{\alpha(1-\gamma)\omega T}{2(\tau+\tau_{max})}\right) C_1\sigma^2 \\ + \frac{(\tau+\tau_{max})\sigma^2}{\omega(1-\gamma)}\left(\frac{C_2\alpha}{N} + C_3\alpha^3\right).\end{aligned} \tag{5}$$

Discussion: We now remark on the main takeaways from Theorem 1. From the bound in (5), we note that `AsyncMATD` guarantees linear convergence (in mean-square sense) to a ball around $\boldsymbol{\theta}^*$ whose radius depends on the variance σ^2 of the noise model. We now comment on the effect of the

[1] Unless otherwise specified, we use $\|\cdot\|$ to denote the Euclidean norm.

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 14:27:40 UTC from IEEE Xplore. Restrictions apply.

asynchronous delays on the convergence bound, and on the linear convergence speedup established by the theorem.

Effect of asynchronous delays. From (5), note how both the exponent of the linear convergence term and the radius of the noise ball are impacted by the delay sequence via the maximum delay τ_{max}. Indeed, compared to the centralized TD case [14, Theorem 7], and to the synchronous federated TD case [7, Theorem 1], we see that for AsyncMATD, the noise ball gets multiplied by the sum of mixing time and maximum delay, i.e., $\tau + \tau_{max}$. In essence, our analysis reveals that $\tau + \tau_{max}$ plays the role of an *effective* delay. Interestingly, an immediate implication is that if the underlying Markov chain mixes slowly, i.e., has a larger mixing time τ, then the effect of the delay is less pronounced. This appears to be a novel observation.

Linear convergence speedup. Compared to the centralized setting [14, Theorem 7], the noise variance term in (5) gets scaled down by a factor of N up to a higher-order $O(\alpha^3)$ term that, for small enough α, is dominated by (α/N). To better illustrate the linear speedup effect, consider the following choice of α and T (define $\bar{\tau} \triangleq \tau + \tau_{max}$):

$$\alpha = \frac{\bar{\tau}\log NT}{\omega(1-\gamma)T}, \quad \text{and} \quad T \geq \frac{2C_0 N\bar{\tau}^2 \log NT}{\omega^2(1-\gamma)^2}. \tag{6}$$

With the above choices, and simple manipulations of the bound in (5), it can be explicitly shown that

$$\mathbb{E}[\delta_T^2] \leq O\left(\frac{\sigma^2\bar{\tau}^2\log(NT)}{\omega^2(1-\gamma)^2 NT}\right). \tag{7}$$

The above bound tells us that AsyncMATD yields a convergence rate of $O(1/(NT))$, which is a factor of N better than the $O(1/T)$ rate in the centralized case [13]. *We remark that this is the first analysis for asynchronous multi-agent and federated RL that provides finite-time convergence guarantees, while jointly establishing an N-fold linear convergence speedup under Markovian sampling.*

IV. Proof of the Main Result

In this section, we prove Theorem 1. We start by introducing the following definitions to lighten the notation:

$$\begin{aligned} \eta_{k,\tau}^{(i)}(\boldsymbol{\theta}) &\triangleq \|\mathbb{E}\left[\mathbf{g}_{i,k}(\boldsymbol{\theta}, o_{i,k})|o_{i,k-\tau}\right] - \bar{\mathbf{g}}(\boldsymbol{\theta})\|, \ k \geq \tau, \\ \delta_{k,h} &\triangleq \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-h}\|, \ k \geq h \geq 0, \\ d_k &\triangleq \max_{k-2\tau_{max}-\tau \leq j \leq k} \mathbb{E}\left[\delta_j^2\right], \ k \geq \tau + 2\tau_{max}. \end{aligned} \tag{8}$$

For our analysis, we will need the following result from [13].

Lemma 1. *The following holds $\forall \boldsymbol{\theta} \in \mathbb{R}^m$:*

$$\langle \boldsymbol{\theta}^* - \boldsymbol{\theta}, \bar{\mathbf{g}}(\boldsymbol{\theta})\rangle \geq \omega(1-\gamma)\|\boldsymbol{\theta}^* - \boldsymbol{\theta}\|^2.$$

We will also use the fact that the random TD update directions and their steady-state versions are 2-Lipschitz [13], i.e., $\forall i \in [N], \forall k \in \mathbb{N}$, and $\forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^m$, we have:

$$\max\{\|\mathbf{g}(\boldsymbol{\theta}) - \mathbf{g}(\boldsymbol{\theta}')\|, \|\bar{\mathbf{g}}(\boldsymbol{\theta}) - \bar{\mathbf{g}}(\boldsymbol{\theta}')\|\} \leq 2\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|. \tag{9}$$

From [14], we also have that $\forall i \in [N], \forall k \in \mathbb{N}, \forall \boldsymbol{\theta} \in \mathbb{R}^m$:

$$\|\mathbf{g}(\boldsymbol{\theta}, o_{i,k})\| \leq 2\|\boldsymbol{\theta}\| + 2\bar{r}, \tag{10}$$

which, squared, using $\bar{r} \leq \sigma$, yields

$$\|\mathbf{g}(\boldsymbol{\theta}, o_{i,k})\|^2 \leq 8(\|\boldsymbol{\theta}\|^2 + \sigma^2). \tag{11}$$

We will often use the fact that, from the definition of the mixing time in Definition 1, we have, for a given iteration $k \geq \tau$, defining $\Theta_{i,k} \triangleq \{\boldsymbol{\theta}_{k-\tau}, o_{i,k-\tau}\}$,

$$\|\mathbb{E}\left[\mathbf{g}(\boldsymbol{\theta}_{k-\tau}, o_{i,k})|\Theta_{i,k}\right] - \bar{\mathbf{g}}(\boldsymbol{\theta}_{k-\tau})\| \leq \alpha^q\left(\|\boldsymbol{\theta}_{k-\tau}\| + 1\right) \tag{12}$$

The proof also relies on the following result from [19]:

Lemma 2. *Let V_k be non-negative real numbers that satisfy*

$$V_{k+1} \leq pV_k + q \max_{(k-d(k))_+ \leq \ell \leq k} V_\ell + \beta,$$

for $\beta, p, q > 0$. Here, $k \geq 0$ and $0 \leq d(k) \leq d_{\max}$ for some $d_{\max} \geq 0$. If $p + q < 1$, then we have

$$V_k \leq \rho^k V_0 + \epsilon,$$

where $\rho = (p+q)^{1/(1+d_{\max})}$ and $\epsilon = \frac{\beta}{1-p-q}$.

We will also use the fact that, for any $a, b \in \mathbb{R}, c \geq 0$,

$$ab = a\sqrt{c}\frac{b}{\sqrt{c}} \leq \frac{1}{2}\left(ca^2 + \frac{b^2}{c}\right), \tag{13}$$

and also the fact that, for $a_i \in \mathbb{R}, i = 1, ..., N$,

$$\left(\sum_{i=1}^N a_i\right)^2 \leq N\sum_{i=1}^N a_i^2. \tag{14}$$

Equipped with the above basic results, we now provide an outline of our proof before illustrating the technical details.

Outline of the proof. Recall $t_{i,k} \triangleq (k - \tau_{i,k})_+$. We write the update rule (2) as

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \alpha\mathbf{v}_k = \boldsymbol{\theta}_k + \alpha\bar{\mathbf{g}}(\boldsymbol{\theta}_k) - \alpha\mathbf{e}_k, \tag{15}$$

with $\mathbf{e}_k \triangleq \bar{\mathbf{g}}(\boldsymbol{\theta}_k) - \mathbf{v}_k$. Thus,

$$\mathbf{e}_k = \frac{1}{N}\sum_{i=1}^N \left(\bar{\mathbf{g}}(\boldsymbol{\theta}_k) - \mathbf{g}(\boldsymbol{\theta}_{t_{i,k}}, o_{i,t_{i,k}})\right). \tag{16}$$

We analyze the following recursion:

$$\begin{aligned} \delta_{k+1}^2 &= T_1 + \alpha^2 T_2 - 2\alpha T_3, \ \text{ with} \\ T_1 &= \|\boldsymbol{\theta}_k - \boldsymbol{\theta}^* + \alpha\bar{\mathbf{g}}(\boldsymbol{\theta}_k)\|^2 \\ T_2 &= \|\mathbf{e}_k\|^2 \\ T_3 &= \langle \boldsymbol{\theta}_k - \boldsymbol{\theta}^* + \alpha\bar{\mathbf{g}}(\boldsymbol{\theta}_k), \mathbf{e}_k\rangle. \end{aligned} \tag{17}$$

The most important part of the proof consists of obtaining a bound of the following form:

$$\mathbb{E}\left[\delta_{k+1}^2\right] \leq p\mathbb{E}\left[\delta_k^2\right] + O(\alpha^2(\tau + \tau_{max}))d_k + B_{\alpha,N}, \tag{18}$$

where d_k was defined in (8), $p < 1$ is a contraction factor and

$$B_{\alpha,N} = O(\alpha^2(\tau + \tau_{max}))\frac{\sigma^2}{N} + O(\alpha^4)\sigma^2, \tag{19}$$

which guarantees the linear speedup effect with N. The bound in (18) allows us to obtain the desired result, by picking a step size small enough and applying Lemma 2.

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 14:27:40 UTC from IEEE Xplore. Restrictions apply.

Given this outline, in the following we provide bounds for $\mathbb{E}[T_1]$, $\mathbb{E}[T_2]$ and $\mathbb{E}[T_3]$.

Bounding $\mathbb{E}[T_1]$. Note that

$$\begin{aligned} T_1 &= \|\boldsymbol{\theta}_k - \boldsymbol{\theta}^* + \alpha\bar{\mathbf{g}}(\boldsymbol{\theta}_k)\|^2 \\ &= \delta_k^2 + 2\alpha\langle\boldsymbol{\theta}_k - \boldsymbol{\theta}^*, \bar{\mathbf{g}}(\boldsymbol{\theta}_k)\rangle + \alpha^2\|\bar{\mathbf{g}}(\boldsymbol{\theta}_k)\|^2. \end{aligned} \tag{20}$$

Note that, using Lemma 1, we get

$$\langle\boldsymbol{\theta}_k - \boldsymbol{\theta}^*, \bar{\mathbf{g}}(\boldsymbol{\theta}_k)\rangle \leq -(1-\gamma)\omega\delta_k^2, \tag{21}$$

and using (9) we get

$$\|\bar{\mathbf{g}}(\boldsymbol{\theta}_k)\|^2 = \|\bar{\mathbf{g}}(\boldsymbol{\theta}_k) - \bar{\mathbf{g}}(\boldsymbol{\theta}^*)\|^2 \leq 4\delta_k^2. \tag{22}$$

Combining the two bounds above and taking the expectation,

$$\mathbb{E}[T_1] \leq (1 - 2\alpha(1-\gamma)\omega)\mathbb{E}\left[\delta_k^2\right] + 4\alpha^2\mathbb{E}\left[\delta_k^2\right]. \tag{23}$$

Bounding $\mathbb{E}[T_2]$. We need the following result.

Lemma 3. *For $k \geq \tau + \tau_{max}$, we have*

$$\mathbb{E}\left[\|\mathbf{v}_k\|^2\right] \leq 8 \max_{k-\tau_{max}\leq j\leq k} \mathbb{E}\left[\delta_j^2\right] + 32\frac{\sigma^2}{N} + 8\sigma^2\alpha^{2q} \tag{24}$$

Proof. We write

$$\begin{aligned} \|\mathbf{v}_k\|^2 &\leq \frac{2}{N^2}(V_1 + V_2), \quad \text{with} \\ V_1 &= \|\sum_{i=1}^N \mathbf{g}(\boldsymbol{\theta}_{t_{i,k}}, o_{i,t_{i,k}}) - \mathbf{g}(\boldsymbol{\theta}^*, o_{i,t_{i,k}})\|^2, \\ V_2 &= \|\sum_{i=1}^N \mathbf{g}(\boldsymbol{\theta}^*, o_{i,t_{i,k}})\|^2. \end{aligned} \tag{25}$$

We now bound V_1.

$$\begin{aligned} V_1 &\leq N\sum_{i=1}^N \|\mathbf{g}(\boldsymbol{\theta}_{t_{i,k}}, o_{i,t_{i,k}}) - \mathbf{g}(\boldsymbol{\theta}^*, o_{i,t_{i,k}})\|^2 \\ &\overset{(9)}{\leq} 4N\sum_{i=1}^N \delta_{t_{i,k}}^2. \quad \text{Thus,} \\ \mathbb{E}[V_1] &\leq 4N\sum_{i=1}^N \mathbb{E}\left[\delta_{t_{i,k}}^2\right] \leq 4N^2 \max_{k-\tau_{max}\leq j\leq k} \mathbb{E}\left[\delta_j^2\right] \end{aligned} \tag{26}$$

We now proceed to bound V_2.

$$\begin{aligned} V_2 &= V_{21} + V_{22}, \quad \text{with} \\ V_{21} &= \sum_{i=1}^N \|\mathbf{g}(\boldsymbol{\theta}^*, o_{i,t_{i,k}})\|^2 \\ V_{22} &= \sum_{\substack{i,j=1 \\ i\neq j}}^N \langle\mathbf{g}(\boldsymbol{\theta}^*, o_{i,t_{i,k}}), \mathbf{g}(\boldsymbol{\theta}^*, o_{j,t_{j,k}})\rangle. \end{aligned} \tag{27}$$

We see that, using (11), we get

$$V_{21} \leq 8\sum_{i=1}^N (\|\boldsymbol{\theta}^*\|^2 + \sigma^2) \quad \leq 16N\sigma^2. \tag{28}$$

Now, using the fact that the observations $o_{i,k}$ and $o_{j,k'}$ are independent for $i \neq j$ and for any $k, k' \geq 0$,

$$\begin{aligned} \mathbb{E}[V_{22}] = \sum_{\substack{i,j=1 \\ i\neq j}}^N \langle &\mathbb{E}\left[\mathbb{E}\left[\mathbf{g}(\boldsymbol{\theta}^*, o_{i,t_{i,k}})|o_{i,t_{i,k}-\tau}\right]\right], \\ &\mathbb{E}\left[\mathbb{E}\left[\mathbf{g}(\boldsymbol{\theta}^*, o_{j,t_{j,k}})|o_{j,t_{j,k}-\tau}\right]\right]\rangle. \end{aligned} \tag{29}$$

Using the fact that $\bar{\mathbf{g}}(\boldsymbol{\theta}^*) = 0$, and Cauchy-Schwarz inequality followed by Jensen's inequality, we can write

$$\begin{aligned} \mathbb{E}[V_{22}] &\leq \sum_{\substack{i,j=1 \\ i\neq j}}^N \mathbb{E}\left[\eta_{t_{i,k},\tau}^{(i)}(\boldsymbol{\theta}^*)\right] \times \mathbb{E}\left[\eta_{t_{j,k},\tau}^{(j)}(\boldsymbol{\theta}^*)\right] \\ &\leq N^2\alpha^{2q}(\|\boldsymbol{\theta}^*\| + \sigma)^2 \leq 4N^2\alpha^{2q}\sigma^2. \end{aligned} \tag{30}$$

Plugging the above bounds on $\mathbb{E}[V_1]$ and $\mathbb{E}[V_2]$ in (25), we can conclude the proof of the lemma. □

We are now in the position to proceed bounding $\mathbb{E}[T_2]$.

$$\begin{aligned} \mathbb{E}[T_2] = \mathbb{E}\left[\|\mathbf{e}_k\|^2\right] &= \mathbb{E}\left[\|\bar{\mathbf{g}}(\boldsymbol{\theta}_k) - \mathbf{v}_k\|^2\right] \\ &\leq 2\mathbb{E}\left[\|\bar{\mathbf{g}}(\boldsymbol{\theta}_k)\|^2 + \|\mathbf{v}_k\|^2\right]. \end{aligned} \tag{31}$$

Note that $\|\bar{\mathbf{g}}(\boldsymbol{\theta}_k)\|^2 = \|\bar{\mathbf{g}}(\boldsymbol{\theta}_k) - \bar{\mathbf{g}}(\boldsymbol{\theta}^*)\|^2 \leq 4\delta_k^2$, and so using Lemma 3 we get

$$\mathbb{E}[T_2] \leq 24 \max_{k-\tau_{max}\leq j\leq k} \mathbb{E}\left[\delta_j^2\right] + 64\frac{\sigma^2}{N} + 16\sigma^2\alpha^{2q}. \tag{32}$$

Bounding $\mathbb{E}[T_3]$. We now bound $\mathbb{E}[T_3]$, which represents the major technical burden of the proof. We need the following result.

Lemma 4. *Let $k \geq \tau_{max} + h$, for some $h \geq 0$. Then,*

$$\mathbb{E}\left[\delta_{k,h}^2\right] \leq 8\alpha^2h^2\left(d_k + 4\frac{\sigma^2}{N} + \sigma^2\alpha^{2q}\right) \tag{33}$$

Proof. Note that, using Lemma 3,

$$\begin{aligned} \mathbb{E}\left[\delta_{k,h}^2\right] &= \mathbb{E}\left[\|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-h}\|^2\right] \\ &\leq h\sum_{l=k-h}^{k-1} \mathbb{E}\left[\|\boldsymbol{\theta}_{l+1} - \boldsymbol{\theta}_l\|^2\right] \\ &\leq \alpha^2h\sum_{l=k-h}^{k-1} \mathbb{E}\left[\|\mathbf{v}_l\|^2\right] \\ &\leq \alpha^2h\sum_{l=k-h}^{k-1} (8\max_{l-\tau_{max}\leq j\leq l} \mathbb{E}\left[\delta_j^2\right] \\ &\quad + 32\frac{\sigma^2}{N} + 8\sigma^2\alpha^{2q}) \\ &\leq 8\alpha^2h^2\left(d_k + 4\frac{\sigma^2}{N} + \sigma^2\alpha^{2q}\right). \end{aligned} \tag{34}$$

□

Now, we can write

$$\begin{aligned} T_3 &= K + T_{32}, \text{ with} \\ K &= \langle\boldsymbol{\theta}_k - \boldsymbol{\theta}^*, \mathbf{e}_k\rangle, \\ T_{32} &= \alpha\langle\bar{\mathbf{g}}(\boldsymbol{\theta}_k), \mathbf{e}_k\rangle. \end{aligned} \tag{35}$$

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 14:27:40 UTC from IEEE Xplore. Restrictions apply.

Note that, using Cauchy-Schwarz and (13),

$$T_{32} \leq \frac{\alpha}{2}\left(\|\bar{\mathbf{g}}(\boldsymbol{\theta}_k)\|^2 + \|\mathbf{e}_k\|^2\right) \leq 2\alpha\delta_k^2 + \frac{\alpha}{2}\|\mathbf{e}_k\|^2. \tag{36}$$

Taking the expectation and using the bound on $\mathbb{E}[T_2]$,

$$\mathbb{E}[T_{32}] \leq \alpha(14d_k + 32\frac{\sigma^2}{N} + 8\sigma^2\alpha^{2q}). \tag{37}$$

Now we bound K. Define $\bar{\mathbf{g}}_{N,k} \triangleq \frac{1}{N}\sum_{i=1}^N \bar{\mathbf{g}}(\boldsymbol{\theta}_{k-\tau_{i,k}})$. Adding and subtracting $\bar{\mathbf{g}}_{N,k}$, we write

$$\begin{aligned} K &= K_1 + K_2, \quad \text{with} \\ K_1 &= \langle \boldsymbol{\theta}_k - \boldsymbol{\theta}^*, \bar{\mathbf{g}}(\boldsymbol{\theta}_k) - \bar{\mathbf{g}}_{N,k}\rangle, \\ K_2 &= \langle \boldsymbol{\theta}_k - \boldsymbol{\theta}^*, \bar{\mathbf{g}}_{N,k} - \mathbf{v}_k\rangle. \end{aligned} \tag{38}$$

Now note that, taking the sum of the second term outside of the inner product in K_1,

$$\begin{aligned} K_1 &= \frac{1}{N}\sum_{i=1}^N K_{1,i}, \quad \text{with} \\ K_{1,i} &= \langle \boldsymbol{\theta}_k - \boldsymbol{\theta}^*, \bar{\mathbf{g}}(\boldsymbol{\theta}_k) - \bar{\mathbf{g}}(\boldsymbol{\theta}_{t_{i,k}})\rangle. \end{aligned} \tag{39}$$

We now bound $\mathbb{E}[K_{1,i}]$. Using (13) and (9),

$$\begin{aligned} K_{1,i} &\leq \alpha\tau_{max}\delta_k^2 + \frac{1}{4\alpha\tau_{max}}\|\bar{\mathbf{g}}(\boldsymbol{\theta}_k) - \bar{\mathbf{g}}(\boldsymbol{\theta}_{t_{i,k}})\|^2 \\ &\leq \alpha\tau_{max}\delta_k^2 + \frac{1}{\alpha\tau_{max}}\delta_{k,\tau_{i,k}}^2. \end{aligned} \tag{40}$$

Note that, using Lemma 4, which requires $k \geq \tau_{max} + \tau_{i,k}$, which holds for $k \geq 2\tau_{max}$,

$$\mathbb{E}\left[\delta_{k,\tau_{i,k}}^2\right] \leq 8\alpha^2\tau_{max}^2\left(d_k + 4\frac{\sigma^2}{N} + \sigma^2\alpha^{2q}\right). \tag{41}$$

Taking the expectation and applying (41), we get

$$\mathbb{E}[K_{1,i}] \leq \alpha\tau_{max}\left(9d_k + 32\frac{\sigma^2}{N} + 8\sigma^2\alpha^{2q}\right), \tag{42}$$

and note that $\mathbb{E}[K_1]$ is bounded by the same quantity. We now proceed to bound $\mathbb{E}[K_2]$.

$$\begin{aligned} K_2 &= \frac{1}{N}\sum_{i=1}^N K_{2,i}, \quad \text{with} \\ K_{2,i} &= \langle \boldsymbol{\theta}_k - \boldsymbol{\theta}^*, \bar{\mathbf{g}}(\boldsymbol{\theta}_{t_{i,k}}) - \mathbf{g}(\boldsymbol{\theta}_{t_{i,k}}, o_{i,t_{i,k}})\rangle \\ &= \Delta_{1,i} + \Delta_{2,i} + \Delta_{3,i}, \quad \text{where} \\ \Delta_{1,i} &= \langle \boldsymbol{\theta}_k - \boldsymbol{\theta}^*, \bar{\mathbf{g}}(\boldsymbol{\theta}_{t_{i,k}}) - \bar{\mathbf{g}}(\boldsymbol{\theta}_{t_{i,k}-\tau})\rangle, \\ \Delta_{2,i} &= \langle \boldsymbol{\theta}_k - \boldsymbol{\theta}^*, \bar{\mathbf{g}}(\boldsymbol{\theta}_{t_{i,k}-\tau}) - \mathbf{g}(\boldsymbol{\theta}_{t_{i,k}-\tau}, o_{i,t_{i,k}})\rangle \\ \Delta_{3,i} &= \langle \boldsymbol{\theta}_k - \boldsymbol{\theta}^*, \mathbf{g}(\boldsymbol{\theta}_{t_{i,k}-\tau}, o_{i,t_{i,k}}) - \mathbf{g}(\boldsymbol{\theta}_{t_{i,k}}, o_{i,t_{i,k}})\rangle. \end{aligned} \tag{43}$$

Note that, using Cauchy-Schwarz inequality and (9),

$$\begin{aligned} \Delta_{1,i} &\leq \delta_k\|\bar{\mathbf{g}}(\boldsymbol{\theta}_{t_{i,k}}) - \bar{\mathbf{g}}(\boldsymbol{\theta}_{t_{i,k}-\tau})\| \\ &\leq 2\delta_k\delta_{t_{i,k},\tau} \\ &\overset{(13)}{\leq} \left(\alpha\tau\delta_k^2 + \frac{\delta_{t_{i,k},\tau}^2}{\alpha\tau}\right). \end{aligned} \tag{44}$$

For $k \geq \tau + \tau_{max}$, we can apply Lemma 4 and get

$$\mathbb{E}\left[\delta_{t_{i,k},\tau}^2\right] \leq \alpha^2\tau^2\left(8d_k + 32\frac{\sigma^2}{N} + 8\sigma^2\alpha^{2q}\right). \tag{45}$$

Thus, taking the expectation, we can get

$$\mathbb{E}[\Delta_{1,i}] \leq 9\alpha\tau d_k + 32\alpha\tau\frac{\sigma^2}{N} + 8\alpha\tau\sigma^2\alpha^{2q}. \tag{46}$$

Note that, using the Lipschitz property (see (9)), we get the exact same bound for $\Delta_{3,i}$. Now note that

$$\begin{aligned} K_2 &= \frac{1}{N}\sum_{i=1}^N (\Delta_{1,i} + \Delta_{3,i}) + \bar{\Delta}, \quad \text{with} \\ \bar{\Delta} &= \frac{1}{N}\sum_{i=1}^N \Delta_{2,i}. \end{aligned} \tag{47}$$

To bound $\bar{\Delta}$, we want to use the geometric mixing property of the Markov chain, that follows from Assumption 2. However, due to the correlations existing between the iterates $\boldsymbol{\theta}_{t_{i,k}}$, this needs to be done with special care. Defining $k' \triangleq k - \tau_{max} - \tau$, we start by adding and subtracting $\boldsymbol{\theta}_{k'}$ from the first term in the inner product of $\Delta_{2,i}$, getting

$$\begin{aligned} \bar{\Delta} &= \bar{\Delta}_1 + \bar{\Delta}_2, \quad \text{with} \\ \bar{\Delta}_1 &= \langle \boldsymbol{\theta}_k - \boldsymbol{\theta}_{k'}, \frac{1}{N}\sum_{i=1}^N \bar{\mathbf{g}}(\boldsymbol{\theta}_{t_{i,k}-\tau}) - \mathbf{g}(\boldsymbol{\theta}_{t_{i,k}-\tau}, o_{i,t_{i,k}})\rangle, \\ \bar{\Delta}_2 &= \langle \boldsymbol{\theta}_{k'} - \boldsymbol{\theta}^*, \frac{1}{N}\sum_{i=1}^N \bar{\mathbf{g}}(\boldsymbol{\theta}_{t_{i,k}-\tau}) - \mathbf{g}(\boldsymbol{\theta}_{t_{i,k}-\tau}, o_{i,t_{i,k}})\rangle. \end{aligned} \tag{48}$$

Note that we can write, using (13) and (14),

$$\begin{aligned} \bar{\Delta}_1 \leq& \frac{1}{2\alpha(\tau_{max}+\tau)}\delta_{k,\tau+\tau_{max}}^2 \\ &+ \frac{\alpha(\tau_{max}+\tau)}{2N^2}\underbrace{\|\sum_{i=1}^N \bar{\mathbf{g}}(\boldsymbol{\theta}_{t_{i,k}-\tau}) - \mathbf{g}(\boldsymbol{\theta}_{t_{i,k}-\tau}, o_{i,t_{i,k}})\|^2}_{G} \end{aligned} \tag{49}$$

For $k \geq \tau + 2\tau_{max}$, we can apply Lemma 4 and get

$$\mathbb{E}\left[\delta_{k,\tau+\tau_{max}}^2\right] \leq 8\alpha^2(\tau+\tau_{max})^2(d_k + 4\frac{\sigma^2}{N} + \sigma^2\alpha^{2q}). \tag{50}$$

Now note that

$$G \leq 2\underbrace{\|\sum_{i=1}^N \bar{\mathbf{g}}(\boldsymbol{\theta}_{t_{i,k}-\tau})\|^2}_{G_1} + 2\underbrace{\|\sum_{i=1}^N \mathbf{g}(\boldsymbol{\theta}_{t_{i,k}-\tau}, o_{i,t_{i,k}})\|^2}_{G_2}, \tag{51}$$

with $\mathbb{E}[G_1] \leq 4N^2\mathbb{E}\left[\delta_{t_{i,k}-\tau}^2\right] \leq 4N^2 d_k$. Also note that

$$\begin{aligned} G_2 &\leq 2G_{21} + 2G_{22}, \quad \text{with} \\ G_{21} &= \|\sum_{i=1}^N \mathbf{g}(\boldsymbol{\theta}_{t_{i,k}-\tau}, o_{i,t_{i,k}}) - \mathbf{g}(\boldsymbol{\theta}^*, o_{i,k-\tau_{i,k}})\|^2, \\ G_{22} &= \|\sum_{i=1}^N \mathbf{g}(\boldsymbol{\theta}^*, o_{i,t_{i,k}})\|^2. \end{aligned} \tag{52}$$

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 14:27:40 UTC from IEEE Xplore. Restrictions apply.

Note that $\mathbb{E}[G_{21}]$ and $\mathbb{E}[G_{22}]$ can be bounded in the same way as $\mathbb{E}[V_1]$ and $\mathbb{E}[V_2]$ in the proof of Lemma 3. We get

$$\mathbb{E}[G_2] \leq 8\left(N^2 d_k + 4N\sigma^2 + N^2\sigma^2\alpha^{2q}\right). \tag{53}$$

Hence, we get

$$\mathbb{E}\left[\bar{\Delta}_1\right] \leq 16\alpha\left(\tau_{max}+\tau\right)\left(d_k + 3\frac{\sigma^2}{N} + \sigma^2\alpha^{2q}\right). \tag{54}$$

Defining $\bar{\Delta}_{2,i} \triangleq \langle \boldsymbol{\theta}_{k'} - \boldsymbol{\theta}^*, \bar{\mathbf{g}}(\boldsymbol{\theta}_{t_{i,k}-\tau}) - \mathbf{g}(\boldsymbol{\theta}_{t_{i,k}-\tau}, o_{i,t_{i,k}})\rangle$,

$$\mathbb{E}\left[\bar{\Delta}_2\right] = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}\left[\bar{\Delta}_{2,i}\right]. \tag{55}$$

Now, defining $\bar{\Theta}_{i,k} \triangleq \{\boldsymbol{\theta}_{k'}, \boldsymbol{\theta}_{t_{i,k}-\tau}, o_{t_{i,k}-\tau}\}$,

$$\begin{aligned}
\mathbb{E}\left[\bar{\Delta}_{2,i}\right] =& \mathbb{E}[\langle \boldsymbol{\theta}_{k'} - \boldsymbol{\theta}^*, \bar{\mathbf{g}}(\boldsymbol{\theta}_{t_{i,k}-\tau}) - \mathbf{g}(\boldsymbol{\theta}_{t_{i,k}-\tau}, o_{i,t_{i,k}})\rangle] \\
=& \mathbb{E}[\langle \boldsymbol{\theta}_{k'} - \boldsymbol{\theta}^*, \\
& \bar{\mathbf{g}}(\boldsymbol{\theta}_{t_{i,k}-\tau}) - \mathbb{E}\left[\mathbf{g}(\boldsymbol{\theta}_{t_{i,k}-\tau}, o_{i,t_{i,k}})|\bar{\Theta}_{i,k}\right]\rangle] \\
\leq& \mathbb{E}[\delta_{k'} \\
& \cdot \underbrace{\|\bar{\mathbf{g}}(\boldsymbol{\theta}_{t_{i,k}-\tau}) - \mathbb{E}\left[\mathbf{g}(\boldsymbol{\theta}_{t_{i,k}-\tau}, o_{i,t_{i,k}})|\bar{\Theta}_{i,k}\right]\|}_{\bar{\eta}_{k,i}}].
\end{aligned} \tag{56}$$

Now, recall $\Theta_{i,k} = \{\boldsymbol{\theta}_{t_{i,k}-\tau}, o_{i,t_{i,k}-\tau}\}$. Note that, for the memoryless property of the Markov chain $\{o_{i,k}\}$, we have

$$\mathbb{E}\left[\mathbf{g}(\boldsymbol{\theta}_{t_{i,k}-\tau}, o_{i,t_{i,k}})|\bar{\Theta}_{i,k}\right] = \mathbb{E}\left[\mathbf{g}(\boldsymbol{\theta}_{t_{i,k}-\tau}, o_{i,t_{i,k}})|\Theta_{i,k}\right]. \tag{57}$$

Indeed, by inspecting the update rule (2) we see that the parameter $\boldsymbol{\theta}_{k'}$ is a function of agent i observations only up to time step $k'-1$, so up to observation $o_{i,k'-1}$. Hence, all the statistical information contained in $\boldsymbol{\theta}_{k'}$ has no influence on the random variable $o_{i,t_{i,k}}$ once we condition on $o_{i,t_{i,k}-\tau}$, because $t_{i,k} - \tau = k - \tau_{i,k} - \tau > k' - 1 = k - \tau_{max} - \tau - 1$. Therefore,

$$\begin{aligned}
\bar{\eta}_{k,i} &= \|\bar{\mathbf{g}}(\boldsymbol{\theta}_{t_{i,k}-\tau}) - \mathbb{E}\left[\mathbf{g}(\boldsymbol{\theta}_{t_{i,k}-\tau}, o_{i,t_{i,k}})|\Theta_{i,k}\right]\| \\
&\overset{(12)}{\leq} \alpha^q\left(\|\boldsymbol{\theta}_{t_{i,k}-\tau}\| + \sigma\right).
\end{aligned} \tag{58}$$

Hence, we get

$$\begin{aligned}
\mathbb{E}\left[\bar{\Delta}_{2,i}\right] &\leq \mathbb{E}\left[\delta_{k'}\alpha^q\left(\|\boldsymbol{\theta}_{t_{i,k}-\tau}\| + \sigma\right)\right] \\
&\leq \mathbb{E}\left[\delta_{k'}\alpha^q\left(\delta_{t_{i,k}-\tau} + 2\sigma\right)\right] \\
&\leq \mathbb{E}\left[\frac{\alpha}{2}\delta_{k'}^2 + \alpha^{2q-1}\left(\delta_{t_{i,k}-\tau}^2 + 4\sigma^2\right)\right] \\
&\leq \alpha d_k + 4\alpha\sigma^2\alpha^{2(q-1)} \\
&\leq \alpha d_k + 4\alpha\sigma^2\alpha^q
\end{aligned} \tag{59}$$

where we used the fact that $\alpha \leq \frac{1}{2}$ and the fact that $2(q-1) \geq q$ for $q \geq 2$. Putting all the above bounds together, and also the fact that $\alpha \leq \frac{1}{16(\tau_{max}+\tau)}$ we get

$$\mathbb{E}[T_3] \leq \alpha\left(\tau_{max}+\tau\right)\left(44 d_k + 144\frac{\sigma^2}{N}\right) + 7\alpha\sigma^2\alpha^q. \tag{60}$$

Concluding the proof. Using the bounds on $\mathbb{E}[T_1], \mathbb{E}[T_2]$ and $\mathbb{E}[T_3]$, we obtain, for $k \geq \tau' \triangleq 2\tau_{max} + \tau$,

$$\begin{aligned}
\mathbb{E}\left[\delta_{k+1}^2\right] \leq& \left(1 - 2\alpha(1-\gamma)\omega\right)\mathbb{E}\left[\delta_k^2\right] + 15\alpha^2\sigma^2\alpha^q \\
&+ \alpha^2\left(\tau + \tau_{max}\right)\left(112 d_k + 352\frac{\sigma^2}{N}\right).
\end{aligned} \tag{61}$$

We can now write the above bound in the following way:

$$\begin{aligned}
V_{k+1} &\leq pV_k + q\max_{k-\tau' \leq j \leq k} V_j + \beta, \quad \text{with} \\
V_k &= \mathbb{E}\left[\delta_k^2\right], \\
p &= \left(1 - 2\alpha(1-\gamma)\omega\right), \\
q &= 112\alpha^2\left(\tau_{max}+\tau\right), \\
\beta &= 352\alpha^2\left(\tau_{max}+\tau\right)\left(\frac{\sigma^2}{N}\right) + 15\alpha^2\sigma^2\alpha^q.
\end{aligned} \tag{62}$$

Imposing $\alpha \leq \frac{(1-\gamma)\omega}{112(\tau+\tau_{max})}$, we get $p + q \leq 1 - \alpha(1-\gamma)\omega$, and we can apply Lemma 2 getting

$$\mathbb{E}\left[\delta_T^2\right] \leq \rho^{T-\tau'}\mathbb{E}\left[\delta_{\tau'}^2\right] + \epsilon, \tag{63}$$

with $\rho = \left(1 - \alpha(1-\gamma)\omega\right)^{\frac{1}{1+\tau'}}$ and $\epsilon = \frac{\beta}{1-p-q}$. Note that we can easily show that $\rho^{-\tau'} \leq 2$ for $\alpha \leq \frac{1}{16(\tau_{max}+\tau)}$. Furthermore, we can show that $\mathbb{E}\left[\delta_{\tau'}^2\right] \leq 3\sigma^2$, as we do next. Note that, using (11),

$$\begin{aligned}
\|\mathbf{v}_k\|^2 &\leq \frac{1}{N}\sum_{i=1}^{N} 8(2\delta_{t_{i,k}}^2 + 3\sigma^2) \\
&\leq 16\max_{i=1,\ldots,N}\delta_{t_{i,k}}^2 + 24\sigma^2.
\end{aligned} \tag{64}$$

Using this bound, note that, for any $k \geq 0$

$$\begin{aligned}
\delta_{k+1}^2 &= \delta_k^2 + 2\alpha\langle \boldsymbol{\theta}_t - \boldsymbol{\theta}^*, \mathbf{v}_k\rangle + \alpha^2\|\mathbf{v}_k\|^2 \\
&\overset{(13)}{\leq} \delta_k^2 + \alpha\delta_k^2 + \alpha\|\mathbf{v}_k\|^2 + \alpha^2\|\mathbf{v}_k\|^2 \\
&\leq (1+\alpha)\delta_k^2 + 2\alpha\|\mathbf{v}_k\|^2 \\
&\leq (1+\alpha)\delta_k^2 + 32\alpha\max_{i=1,\ldots,N}\delta_{t_{i,k}}^2 + 48\alpha\sigma^2,
\end{aligned} \tag{65}$$

where recall that $t_{i,k} \leq k$. Now define $\bar{p} \triangleq 1 + \alpha$, $\bar{q} \triangleq 32\alpha$ and $\nu \triangleq \bar{p} + \bar{q}$ and $\bar{\beta} \triangleq 48\alpha\sigma^2$. We now prove by induction that, for all $k \geq 0$,

$$\delta_k^2 \leq \nu^k\delta_0^2 + \epsilon_k, \tag{66}$$

where $\epsilon_k = \nu\epsilon_{k-1} + \bar{\beta}$ for $k \geq 1$ and $\epsilon_0 = 0$. The base case is trivially satisfied, because $\delta_0^2 \leq \delta_0^2$. As an inductive step, suppose that (66) is true for $0 \leq s \leq k$, for some $k \geq 0$, so

$$\delta_s^2 \leq \nu^s\delta_0^2 + \epsilon_s, \quad 0 \leq s \leq k. \tag{67}$$

Now, we check the property for $k+1$, using (65), and noting that ϵ_k is an increasing sequence:

$$\begin{aligned}
\delta_{k+1}^2 &\leq \bar{p}\delta_k^2 + \bar{q}\max_{i=1,\ldots,N}\delta_{t_{i,k}}^2 + \bar{\beta}, \\
&\leq \bar{p}(\nu^k\delta_0^2 + \epsilon_k) + \bar{q}\left(\max_{i=1,\ldots,N}\nu^{t_{i,k}}\delta_0^2 + \epsilon_{t_{i,k}}\right) + \bar{\beta} \\
&\leq \bar{p}(\nu^k\delta_0^2 + \epsilon_k) + \bar{q}(\nu^k\delta_0^2 + \epsilon_k) + \bar{\beta} \\
&\leq (\bar{p}+\bar{q})\nu^k\delta_0^2 + (\bar{p}+\bar{q})\epsilon_k + \bar{\beta} \\
&= \nu^{k+1}\delta_0^2 + \epsilon_{k+1}.
\end{aligned} \tag{68}$$

From which we can conclude the proof of (66). Now, note that $\epsilon_k = \bar{\beta}\sum_{j=0}^{k-1}\nu^j$, and that for $0 \leq k \leq \tau'$,

$$\nu^k \leq \nu^{\tau'} \leq (1+33\alpha)^{\tau'} \leq e^{33\alpha\tau'} \leq e^{0.25} \leq 2, \tag{69}$$

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 14:27:40 UTC from IEEE Xplore. Restrictions apply.

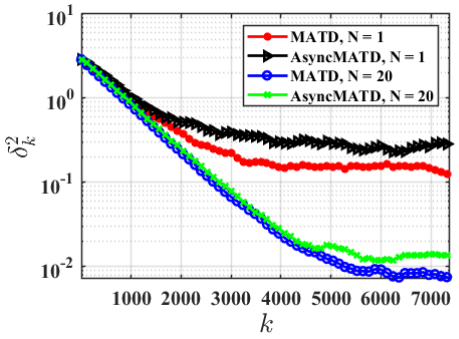


Fig. 2. Comparison between vanilla `MATD` and `AsyncMATD` in single-agent ($N = 1$) and multi-agent ($N = 20$) settings. For `AsyncMATD`, we set $\tau_{max} = 100$.

imposing $\alpha \leq \frac{1}{132\tau'}$. Hence, for $0 \leq k \leq \tau'$,

$$\begin{aligned}\delta_k^2 &\leq \nu^k \delta_0^2 + \epsilon_k \leq 2\delta_0^2 + \bar{\beta} \sum_{j=0}^{\tau'-1} \nu^j \leq 2\delta_0^2 + 2\bar{\beta}\tau' \\ &= 2\delta_0^2 + 2(48\alpha\sigma^2)\tau' \leq 2\delta_0^2 + \sigma^2 \leq 3\sigma^2,\end{aligned} \tag{70}$$

where we used the fact that $\alpha \leq \frac{1}{100\tau'}$ and that $\delta_0^2 \leq \sigma^2$. We can therefore conclude, writing the bound in (63) as

$$\begin{aligned}\mathbb{E}\left[\delta_T^2\right] \leq \exp\left(-\frac{\alpha(1-\gamma)\omega T}{2\left(\tau+\tau_{max}\right)}\right) 6\sigma^2 \\ + 352\frac{\alpha\left(\tau+\tau_{max}\right)\sigma^2}{(1-\gamma)\,\omega N} + 15\frac{\alpha^3\sigma^2}{(1-\gamma)\omega}.\end{aligned} \tag{71}$$

V. Simulations

In this section, we provide simulation results to validate our theory. We consider an MDP with $|\mathcal{S}| = 100$ states and a feature space spanned by $d = 10$ orthonormal basis vectors; we set the discount factor to $\gamma = 0.5$ and the step size to $\alpha = 0.05$. To simulate the asynchronous delays in the TD update directions, we generate random delays at each iteration k for each agent i, by generating a uniform random variable $\tau_{i,k}$ in the range $[1, \tau_{max}]$. We set $\tau_{max} = 100$. In the multi-agent seting, we set the number of agents to $N = 20$. For each configuration, we plot the average of 20 experiments. From the results, it is apparent how the linear speedup property also holds for `AsyncMATD`. Furthermore, we can see how the gap between the delayed and non-delayed settings decreases in the multi-agent case, compared to the single-agent configuration, by more than one order of magnitude.

VI. Conclusion and Future Work

We presented and analysed `AsyncMATD`, providing the first finite-time analysis for asynchronous multi-agent TD learning that jointly establishes a linear convergence speedup with the number of agents. Future work includes more complex delay models, for example considering random and potentially unbounded delays. Future studies could also include analysing non-linear function approximation and studying the trade-off between using only a portion of the agents' updates - e.g., the set of agents whose transmissions are subject to the smallest delays - instead of using all the available stale updates. In this regard, an interesting venue would be extending the single-agent delay-adaptive algorithm proposed in [20] to the multi-agent setting.

References

[1] S. Khodadadian, P. Sharma, G. Joshi, and S. T. Maguluri, "Federated reinforcement learning: Linear speedup under markovian sampling," in *ICML*. PMLR, 2022, pp. 10 997–11 057.

[2] H. Wang, A. Mitra, H. Hassani, G. J. Pappas, and J. Anderson, "Federated temporal difference learning with linear function approximation under environmental heterogeneity," *arXiv:2302.02212*, 2023.

[3] J. Qi, Q. Zhou, L. Lei, and K. Zheng, "Federated reinforcement learning: techniques, applications, and open challenges," *arXiv preprint arXiv:2108.11887*, 2021.

[4] M. Chen, Z. Yang, W. Saad, C. Yin, H. V. Poor, and S. Cui, "A joint learning and communications framework for federated learning over wireless networks," *IEEE Transactions on Wireless Communications*, vol. 20, no. 1, pp. 269–283, 2020.

[5] A. Koloskova, S. U. Stich, and M. Jaggi, "Sharper convergence guarantees for asynchronous sgd for distributed and federated learning," *Advances in Neural Information Processing Systems*, vol. 35, pp. 17 202–17 215, 2022.

[6] S. Dutta, G. Joshi, S. Ghosh, P. Dube, and P. Nagpurkar, "Slow and stale gradients can win the race: Error-runtime trade-offs in distributed sgd," in *International conference on artificial intelligence and statistics*. PMLR, 2018, pp. 803–812.

[7] N. Dal Fabbro, A. Mitra, and G. J. Pappas, "Federated TD learning over finite-rate erasure channels: Linear speedup under markovian sampling," *IEEE Control Systems Letters*, 2023.

[8] N. Dal Fabbro, A. Mitra, R. Heath, L. Schenato, and G. J. Pappas, "Over-the-air federated TD learning," in *MLSys23 Workshop on Resource-Constrained Learning in Wireless Networks*, 2023.

[9] J. Nguyen, K. Malik, H. Zhan, A. Yousefpour, M. Rabbat, M. Malek, and D. Huba, "Federated learning with buffered asynchronous aggregation," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 3581–3607.

[10] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," in *International conference on machine learning*. PMLR, 2016, pp. 1928–1937.

[11] A. Nair, P. Srinivasan, S. Blackwell, C. Alcicek, R. Fearon, A. De Maria, V. Panneershelvam, M. Suleyman, C. Beattie, S. Petersen *et al.*, "Massively parallel methods for deep reinforcement learning," *arXiv preprint arXiv:1507.04296*, 2015.

[12] H. Shen, K. Zhang, M. Hong, and T. Chen, "Towards understanding asynchronous advantage actor-critic: Convergence and linear speedup," *IEEE Transactions on Signal Processing*, vol. 71, pp. 2579–2594, 2023.

[13] J. Bhandari, D. Russo, and R. Singal, "A finite time analysis of temporal difference learning with linear function approximation," in *Conference on learning theory*. PMLR, 2018, pp. 1691–1692.

[14] R. Srikant and L. Ying, "Finite-time error bounds for linear stochastic approximation and TD learning," in *Conference on Learning Theory*. PMLR, 2019, pp. 2803–2830.

[15] R. S. Sutton, "Learning to predict by the methods of temporal differences," *Machine learning*, vol. 3, no. 1, pp. 9–44, 1988.

[16] J. N. Tsitsiklis and B. Van Roy, "An analysis of temporal-difference learning with function approximation," in *IEEE Transactions on Automatic Control*, 1997.

[17] D. A. Levin and Y. Peres, *Markov chains and mixing times*. American Mathematical Soc., 2017, vol. 107.

[18] Z. Chen, S. Zhang, T. T. Doan, S. T. Maguluri, and J.-P. Clarke, "Performance of q-learning with linear function approximation: Stability and finite-time analysis," *arXiv preprint arXiv:1905.11425*, p. 4, 2019.

[19] H. R. Feyzmahdavian, A. Aytekin, and M. Johansson, "A delayed proximal gradient method with linear convergence rate," in *2014 IEEE international workshop on machine learning for signal processing (MLSP)*. IEEE, 2014, pp. 1–6.

[20] A. Adibi, N. Dal Fabbro, L. Schenato, S. Kulkarni, H. V. Poor, G. J. Pappas, H. Hassani, and A. Mitra, "Stochastic approximation with delayed updates: Finite-time rates under markovian sampling," *arXiv preprint arXiv:2402.11800*, 2024.

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 14:27:40 UTC from IEEE Xplore. Restrictions apply.