Automatica 159 (2024) 111368

Contents lists available at ScienceDirect

Automatica

journal homepage: www.elsevier.com/locate/automatica

Finite-time error bounds for distributed linear stochastic approximation☆,☆☆

Yixuan Lin [a], Vijay Gupta [b], Ji Liu [c,*]

[a] Department of Applied Mathematics and Statistics, Stony Brook University, United States of America
[b] School of Electrical and Computer Engineering, Purdue University, United States of America
[c] Department of Electrical and Computer Engineering, Stony Brook University, United States of America

ARTICLE INFO

Article history:
Received 25 August 2022
Received in revised form 2 April 2023
Accepted 30 September 2023
Available online xxxx

Keywords:
Multi-agent systems
Distributed stochastic approximation
Finite-time analysis

ABSTRACT

This paper considers a novel multi-agent linear stochastic approximation algorithm driven by Markovian noise and general consensus-type interaction, in which each agent evolves according to its local stochastic approximation process which depends on the information from its neighbors. The interconnection structure among the agents is described by a time-varying directed graph. While the convergence of consensus-based stochastic approximation algorithms when the interconnection among the agents is described by doubly stochastic matrices (at least in expectation) has been studied, less is known about the case when the interconnection matrix is simply stochastic. For any uniformly strongly connected graph sequences whose associated interaction matrices are stochastic, the paper derives finite-time bounds on the mean-square error, defined as the deviation of the output of the algorithm from the unique equilibrium point of the associated ordinary differential equation. For the case of interconnection matrices being stochastic, the equilibrium point can be any unspecified convex combination of the local equilibria of all the agents in the absence of communication. Both the cases with constant and time-varying step-sizes are considered. In the case when the convex combination is required to be a straight average and interaction between any pair of neighboring agents may be uni-directional, so that doubly stochastic matrices cannot be implemented in a distributed manner, the paper proposes a push-sum-type distributed stochastic approximation algorithm and provides its finite-time bound for the time-varying step-size case by leveraging the analysis for the consensus-type algorithm with stochastic matrices and developing novel properties of the push-sum algorithm. Distributed temporal difference learning is discussed as an illustrative application.

1. Introduction

The use of reinforcement learning (RL) to obtain policies that describe solutions to a Markov decision process (MDP) in which an autonomous agent interacting with an unknown environment aims to optimize its long term reward is now standard (Sutton & Barto, 2018). Multi-agent RL is useful when a team of agents interacts with an unknown environment or system and aims to collaboratively accomplish tasks involving distributed decision-making.

Stochastic approximation is a family of model-free stochastic algorithms tailored for seeing the extrema of unknown functions via noisy observations only (Robbins & Monro, 1951). It is a key tool for designing and analyzing RL algorithms, including temporal difference (TD) learning as a special case (Sutton & Barto, 2018). Convergence study of stochastic approximation based on ordinary differential equation (ODE) methods has a long history (Borkar & Meyn, 2000). Notable examples are Dayan (1992), Tsitsiklis and Van Roy (1997) which prove asymptotic convergence of TD(λ). Recently, finite-time performance of single-agent stochastic approximation and TD algorithms has been studied in Bhandari, Russo, and Singal (2018), Chen, Maguluri, Shakkottai, and Shanmugam (2020), Dalal, Szörényi, Thoppe, and Mannor (2018), Gupta, Srikant, and Ying (2019), Lakshminarayanan and Szepesvari (2018), Ma, Zhou, and Zou (2020), Srikant and Ying (2019), Wang, Chen, Liu, Ma, and Liu (2017), Xu, Zou, and

☆ Proofs of all the lemmas in this paper are omitted due to space limitations and can be found in Lin et al. (2021). The material in this paper was not presented at any conference. This paper was recommended for publication in revised form by Associate Editor Giacomo Como under the direction of Editor Christos G. Cassandras.

☆☆ The work of Y. Lin and J. Liu was supported in part by the National Science Foundation under Grant No. 2230101 and by the Air Force Office of Scientific Research under award number FA9550-23-1-0175. The work of V. Gupta was supported in part by ARO W911NF2310111, W911NF2310266, and W911NF-23-1-0316, AFOSR F.10052139.02.005, and ONR F.10052139.02.009.

* Corresponding author.
E-mail addresses: yixuan.lin.1@stonybrook.edu (Y. Lin), gupta869@purdue.edu (V. Gupta), ji.liu@stonybrook.edu (J. Liu).

https://doi.org/10.1016/j.automatica.2023.111368
0005-1098/

Liang (2019); many other works have now appeared that perform finite-time analysis for other RL algorithms, see, e.g., Borkar and Pattathil (2018), Chen, Devraj, Bušić, and Meyn (2020), Dalal, Thoppe, Szörényi, and Mannor (2018), Ma, Chen, Zhou, and Zou (2021), Qu and Wierman (2020), Wang, Li, and Giannakis (2019), Wang and Zou (2020), Weng, Gupta, He, Ying, and Srikant (2020), Wu, Zhang, Xu, and Gu (2020), Xu and Gu (2020), Zou, Xu, and Liang (2019), just to name a few. Many distributed multi-agent RL algorithms have been proposed in the literature (Zhang, Yang, & Başar, 2021). In this setting, each agent can receive information only from its neighbors, and no single agent can solve the problem alone or by 'taking the lead'. Many works have analyzed asymptotic convergence of such RL algorithms using ODE methods (Lin et al., 2019; Suttle et al., 2020; Zhang, Yang, & Başar, 2018; Zhang, Yang, Liu, Zhang, & Başar, 2018; Zhang & Zavlanos, 2019). This can be viewed as an application of ideas from distributed stochastic approximation (Bianchi, Fort, & Hachem, 2013; Huang, 2012; Kushner & Yin, 1987; Stanković, Ilić, & Stanković, 2016; Stanković & Stanković, 2016; Stanković, Stanković, & Stipanović, 2010). Finite-time performance guarantees for distributed RL have also been provided in works, most notably in Doan, Maguluri, and Romberg (2019, 2021), Sun, Wang, Giannakis, Yang, and Yang (2020), Wang, Lu, Giannakis, Tesauro, and Sun (2020), Zeng, Doan, and Romberg (2020), Zhang, Yang, Liu, Zhang, and Başar (2021).

The assumption that is the central concern of this paper and is made in all the existing finite-time analyses for distributed RL algorithms is that the consensus interaction is characterized by doubly stochastic matrices (Doan et al., 2019, 2021; Sun et al., 2020; Wang et al., 2020; Zeng et al., 2020; Zhang, Yang, Liu, et al., 2021) at every time step, or at least in expectation (Bianchi et al., 2013). In a realistic network, especially with mobile agents such as autonomous vehicles, drones, or robots, uni-directional communication is inevitable due to various reasons such as asymmetric communication and privacy constraints, non-zero communication failure probability between any two agents at any given time, and application of resilient consensus in the presence of adversary attacks (LeBlanc, Zhang, Koutsoukos, & Sundaram, 2013; Vaidya, Tseng, & Liang, 2012), all leading to an interaction among the agents characterized by a stochastic matrix, which may further be time-varying. The problem of design of distributed RL algorithms with time-varying stochastic matrices and characterizing either their asymptotic convergence or finite time analysis remains open. Technical challenges in removing the assumption of doubly stochastic matrices are discussed in detail in Lin, Gupta, and Liu 2021, Section 1.

We propose a novel consensus-based distributed linear stochastic approximation algorithm driven by Markovian noise in which each agent evolves according to its local stochastic approximation process and the information from its neighbors. We assume only a (possibly time-varying) stochastic matrix being used during the consensus phase, which is a more practical assumption when only unidirectional communication is possible among agents. We establish both convergence guarantees and finite-time bounds on the mean-square error, defined as the deviation of the output of the algorithm from the unique equilibrium point of the associated ordinary differential equation. The equilibrium point can be an "uncontrollable" convex combination of the local equilibria of all the agents in the absence of communication. We consider both the cases of constant and time-varying step-sizes. Our results subsume the existing results on convergence and finite-time analysis of distributed RL algorithms that assume doubly stochastic matrices and bi-directional communication as special cases. In the case when the convex combination is required to be a straight average and interaction between any pair of neighboring agents may be uni-directional, we propose a push-type distributed stochastic approximation algorithm and establish its finite-time performance bound. It is worth emphasizing that it is straightforward to extend our algorithm from the straight average point to any pre-specified convex combination. Since it is well known that TD algorithms can be viewed as a special case of linear stochastic approximation (Tsitsiklis & Van Roy, 1997), our distributed linear stochastic approximation algorithms and their finite-time bounds can be applied to TD algorithms in a straightforward manner.

Notation We use X_t to represent that a variable X is time-dependent and $t \in \{0, 1, 2, \ldots\}$ is the discrete time index. The ith entry of a vector x will be denoted by x^i and, also, by $(x)^i$ when convenient. The ijth entry of a matrix A will be denoted by a^{ij} and, also, by $(A)^{ij}$ when convenient. We use $\mathbf{1}_n$ to denote the vectors in $\mathbb{R}^n$ whose entries all equal to 1's, and I to denote the identity matrix, whose dimension is to be understood from the context. Given a set $\mathcal{S}$ with finitely many elements, we use $|\mathcal{S}|$ to denote the cardinality of $\mathcal{S}$. We use $\lceil\cdot\rceil$ to denote the ceiling function.

A vector is called a stochastic vector if its entries are nonnegative and sum to one. A square nonnegative matrix is called a row stochastic matrix, or simply stochastic matrix, if its row sums all equal one. Similarly, a square nonnegative matrix is called a column stochastic matrix if its column sums all equal one. A square nonnegative matrix is called a doubly stochastic matrix if its row sums and column sums all equal one. The graph of an $n \times n$ matrix is a direct graph with n vertices and a directed edge from vertex i to vertex j whenever the ji-th entry of the matrix is nonzero. A directed graph is strongly connected if it has a directed path from any vertex to any other vertex. For a strongly connected graph $\mathbb{G}$, the distance from vertex i to another vertex j is the length of the shortest directed path from i to j; the longest distance among all ordered pairs of distinct vertices i and j in $\mathbb{G}$ is called the diameter of $\mathbb{G}$.

2. Distributed linear stochastic approximation

The stochastic approximation is a method for approximating the solution of an optimization problem when the objective function is not known, but where only noisy observations are available (Kushner & Yin, 1997). The linear stochastic approximation is a specific form of stochastic approximation that is used to solve linear regression problems with stochastic noise.

Consider a network consisting of N agents. For the purpose of presentation, we label the agents from 1 through N. The agents are not aware of such a global labeling, but can differentiate between their neighbors. The neighbor relations among the N agents are characterized by a time-dependent directed graph $\mathbb{G}_t = (\mathcal{V}, \mathcal{E}_t)$ whose vertices correspond to agents and whose directed edges (or arcs) depict neighbor relations, where $\mathcal{V} = \{1, \ldots, N\}$ is the vertex set and $\mathcal{E}_t = \mathcal{V} \times \mathcal{V}$ is the edge set at time t. Specifically, agent j is an in-neighbor of agent i at time t if $(j, i) \in \mathcal{E}_t$, and similarly, agent k is an out-neighbor of agent i at time t if $(i, k) \in \mathcal{E}_t$. Each agent can send information to its out-neighbors and receive information from its in-neighbors. Thus, the directions of edges represent the directions of information flow. For convenience, we assume that each agent is always an in- and out-neighbor of itself, which implies that $\mathbb{G}_t$ has self-arcs at all vertices for all time t. We use $\mathcal{N}_t^i$ and $\mathcal{N}_t^{i-}$ to denote the in- and out-neighbor set of agent i at time t, respectively, i.e.,

$$\mathcal{N}_t^i = \{j \in \mathcal{V} : (j, i) \in \mathcal{E}_t\}, \quad \mathcal{N}_t^{i-} = \{k \in \mathcal{V} : (i, k) \in \mathcal{E}_t\}.$$

It is clear that $\mathcal{N}_t^i$ and $\mathcal{N}_t^{i-}$ are nonempty as they both contain index i.

We propose the following distributed linear stochastic approximation over a time-varying neighbor graph sequence $\{\mathbb{G}_t\}$. Each

agent $i \in \mathcal{V}$ has control over a random vector $\theta_t^i \in \mathbb{R}^d$ for any $t \in \{0, 1, 2, \ldots\}$, which is updated by

$$\theta_{t+1}^i = \sum_{j \in \mathcal{N}_t^i} w_t^{ij} \theta_t^j + \alpha_t \left(A(X_t) \sum_{j \in \mathcal{N}_t^i} w_t^{ij} \theta_t^j + b^i(X_t) \right), \tag{1}$$

where w_t^{ij} are consensus weights, α_t is the step-size at time t, $A(X_t) \in \mathbb{R}^{d \times d}$ is a random matrix and $b^i(X_t) \in \mathbb{R}^d$ is a random vector, both generated based on the Markov chain $\{X_t\}$ with state spaces $\mathcal{X}$. It is worth noting that the update (1) of each agent only uses its own and in-neighbors' information and thus is distributed.

Remark 1. The work of Kushner and Yin (1987) considers a different consensus-based networked linear stochastic approximation for any $i \in \mathcal{V}, t \in \{0, 1, 2, \ldots\}$ as follows:

$$\theta_{t+1}^i = \sum_{j \in \mathcal{N}_t^i} w_t^{ij} \theta_t^j + \alpha_t \left(A(X_t) \theta_t^i + b^i(X_t) \right), \tag{2}$$

whose state form is $\Theta_{t+1} = W_t \Theta_t + \alpha_t \Theta_t A(X_t)^\top + \alpha_t B(X_t)$, and mainly focuses on asymptotically weakly convergence for the fixed step-size case (i.e., $\alpha_t = \alpha$ for all t). Under the similar set of conditions, with its condition (C3.4') being a stochastic analogy for Assumption 6, Theorem 3.1 in Kushner and Yin (1987) shows that (2) has a limit which can be verified to be the same as θ^*, the limit of (1). How to apply the finite-time analysis tools in this paper to (2) has so far eluded us. The two updates (1) and (2) are analogous to the "combine-then-adapt" and "adapt-then-combine" diffusion strategies in distributed optimization (Chen & Sayed, 2012). □

We impose the following assumption on the weights w_t^{ij} which has been widely adopted in consensus literature (Jadbabaie, Lin, & Morse, 2003; Nedić & Liu, 2017; Olfati-Saber, Fax, & Murray, 2007).

Assumption 1. There exists a constant $\beta > 0$ such that for all $i, j \in \mathcal{V}$ and t, $w_t^{ij} \geq \beta$ whenever $j \in \mathcal{N}_t^i$. For all $i \in \mathcal{V}$ and t, $\sum_{j \in \mathcal{N}_t^i} w_t^{ij} = 1$.

Let W_t be the $N \times N$ matrix whose ijth entry equals w_t^{ij} if $j \in \mathcal{N}_t^i$ and zero otherwise. From Assumption 1, each W_t is a stochastic matrix that is compliant with the neighbor graph $\mathbb{G}_t$. Since each agent i is always assumed to be an in-neighbor of itself, all diagonal entries of W_t are positive. Thus, if $\mathbb{G}_t$ is strongly connected, W_t is irreducible and aperiodic. To proceed, define

$$\Theta_t = \begin{bmatrix} (\theta_t^1)^\top \\ \vdots \\ (\theta_t^N)^\top \end{bmatrix}, \quad B(X_t) = \begin{bmatrix} (b^1(X_t))^\top \\ \vdots \\ (b^N(X_t))^\top \end{bmatrix}.$$

Then, the N linear stochastic recursions in (1) for any $t \in \{0, 1, 2, \ldots\}$ can be combined and written as

$$\Theta_{t+1} = W_t \Theta_t + \alpha_t W_t \Theta_t A(X_t)^\top + \alpha_t B(X_t). \tag{3}$$

The goal of this section is to characterize the finite-time performance of (1), or equivalently (3), with the following standard assumptions, which were adopted e.g. in Doan et al. (2019), Srikant and Ying (2019).

Assumption 2. There exists a matrix A and vectors b^i, $i \in \mathcal{V}$, such that

$$\lim_{t \to \infty} \mathbf{E}[A(X_t)] = A, \quad \lim_{t \to \infty} \mathbf{E}[b^i(X_t)] = b^i, \quad i \in \mathcal{V}.$$

Define $b_{\max} = \max_{i \in \mathcal{V}} \sup_{x \in \mathcal{X}} \|b^i(x)\|_2 < \infty$ and $A_{\max} = \sup_{x \in \mathcal{X}} \|A(x)\|_2 < \infty$. Then, $\|A\|_2 \leq A_{\max}$ and $\|b^i\|_2 \leq b_{\max}$, $i \in \mathcal{V}$.

Assumption 3. Given a positive constant α, we use $\tau(\alpha)$ to denote the mixing time of the Markov chain $\{X_t\}$ for which

$$\begin{cases} \|\mathbf{E}[A(X_t) - A | X_0 = X]\|_2 \leq \alpha, & \forall X, \quad \forall t \geq \tau(\alpha), \\ \|\mathbf{E}[b^i(X_t) - b^i | X_0 = X]\|_2 \leq \alpha, & \forall X, \quad \forall t \geq \tau(\alpha), \quad \forall i \in \mathcal{V}. \end{cases}$$

The Markov chain $\{X_t\}$ mixes at a geometric rate, i.e., there exists a constant C such that $\tau(\alpha) \leq -C \log \alpha$.

Assumption 4. All eigenvalues of A have strictly negative real parts, i.e., A is a Hurwitz matrix. Then, there exists a symmetric positive definite matrix P, such that $A^\top P + PA = -I$. Let $\gamma_{\max}$ and $\gamma_{\min}$ be the maximum and minimum eigenvalues of P, respectively.

Assumption 5. The step-size sequence $\{\alpha_t\}$ is positive, non-increasing, and satisfies $\sum_{t=0}^{\infty} \alpha_t = \infty$ and $\sum_{t=0}^{\infty} \alpha_t^2 < \infty$.

To state our first main result, we need the following concepts.

Definition 1. A graph sequence $\{\mathbb{G}_t\}$ is uniformly strongly connected if there exists a positive integer L such that for any $t \geq 0$, the union graph $\cup_{k=t}^{t+L-1} \mathbb{G}_k$ is strongly connected. If such an integer exists, we sometimes say that $\{\mathbb{G}_t\}$ is uniformly strongly connected by sub-sequences of length L.

Remark 2. Two popular joint connectivity definitions in consensus literature are "B-connected" (Nedić, Olshevsky, Ozdaglar, & Tsitsiklis, 2009) and "repeatedly jointly strongly connected" (Cao, Morse, & Anderson, 2008). A graph sequence $\{\mathbb{G}_t\}$ is B-connected if there exists a positive integer B such that the union graph $\cup_{t=kB}^{(k+1)B-1} \mathbb{G}_t$ is strongly connected for each integer $k \geq 0$. Although the uniformly strongly connectedness looks more restrictive compared with B-connectedness at first glance, they are in fact equivalent. To see this, first it is easy to see that if $\{\mathbb{G}_t\}$ is uniformly strongly connected, $\{\mathbb{G}_t\}$ must be B-connected; now supposing $\{\mathbb{G}_t\}$ is B-connected, for any fix t, the union graph $\cup_{k=t}^{t+2B-1} \mathbb{G}_k$ must be strongly connected, and thus $\{\mathbb{G}_t\}$ is uniformly strongly connected by sub-sequences of length $2B$. Thus, the two definitions are equivalent. It is also not hard to show that the uniformly strongly connectedness is equivalent to "repeatedly jointly strongly connectedness" provided the directed graphs under consideration all have self-arcs at all vertices, with "repeatedly jointly strongly connectedness" being defined upon "graph composition" (Cao et al., 2008). □

Definition 2. Let $\{W_t\}$ be a sequence of stochastic matrices. A sequence of stochastic vectors $\{\pi_t\}$ is an absolute probability sequence for $\{W_t\}$ if $\pi_t^\top = \pi_{t+1}^\top W_t$ for all t.

This definition was first introduced by Kolmogorov who proved that every sequence of stochastic matrices has an absolute probability sequence (Kolmogoroff, 1936). An alternative proof of this fact was given by Blackwell (1945). In general, a sequence of stochastic matrices may have more than one absolute probability sequence; when the sequence of stochastic matrices is "ergodic", it has a unique absolute probability sequence (Nedić & Liu, 2017). It is easy to see that when W_t is a fixed irreducible stochastic matrix W, π_t is simply the normalized left eigenvector of W for eigenvalue one. More can be said.

Lemma 1 (*Lemma 5.8 in Touri (2012)*)**.** *Suppose that Assumption 1 holds. If $\{\mathbb{G}_t\}$ is uniformly strongly connected, then there exists a unique absolute probability sequence $\{\pi_t\}$ for the matrix sequence $\{W_t\}$ and a constant $\pi_{\min} \in (0, 1)$ such that $\pi_t^i \geq \pi_{\min}$ for all i and t.*

Let $\langle\theta\rangle_t = \sum_{i=1}^N \pi_t^i \theta_t^i$, which is a column vector and convex combination of all θ_t^i. It is easy to see that $\langle\theta\rangle_t = (\pi_t^\top \Theta_t)^\top = \Theta_t^\top \pi_t$. From Definition 2 and (3), we have $\pi_{t+1}^\top \Theta_{t+1} = \pi_{t+1}^\top W_t \Theta_t + \alpha_t \pi_{t+1}^\top W_t \Theta_t A(X_t)^\top + \alpha_t \pi_{t+1}^\top B(X_t) = \pi_t^\top \Theta_t + \alpha_t \pi_t^\top \Theta_t A(X_t)^\top + \alpha_t \pi_{t+1}^\top B(X_t)$, which implies that

$$\langle\theta\rangle_{t+1} = \langle\theta\rangle_t + \alpha_t A(X_t)\langle\theta\rangle_t + \alpha_t B(X_t)^\top \pi_{t+1}. \tag{4}$$

Asymptotic performance of (1) with any uniformly strongly connected neighbor graph sequence is characterized by the following two theorems.

Theorem 1. *Suppose that Assumption 1, 2 and 5 hold. Let $\{\theta_t^i\}$, $i \in \mathcal{V}$, be generated by (1). If $\{\mathbb{G}_t\}$ is uniformly strongly connected, then $\lim_{t\to\infty} \|\theta_t^i - \langle\theta\rangle_t\|_2 = 0$ for all $i \in \mathcal{V}$.*

Theorem 1 only shows that all the sequences $\{\theta_t^i\}$, $i \in \mathcal{V}$, generated by (1) will finally reach a consensus, but not necessarily convergent or bounded. To guarantee the convergence of the sequences, we further need the following assumption, whose validity is discussed in Remark 3.

Assumption 6. The absolute probability sequence $\{\pi_t\}$ for the stochastic matrix sequence $\{W_t\}$ has a limit, i.e., there exists a stochastic vector π_∞ such that $\lim_{t\to\infty} \pi_t = \pi_\infty$.

Theorem 2. *Suppose that Assumptions 1–6 hold. Let $\{\theta_t^i\}$, $i \in \mathcal{V}$, be generated by (1) and θ^* be the unique equilibrium point of the ODE*

$$\dot\theta = A\theta + b, \quad b = \sum_{i=1}^N \pi_\infty^i b^i, \tag{5}$$

where A and b^i are defined in Assumption 2 and π_∞ is defined in Assumption 6. If $\{\mathbb{G}_t\}$ is uniformly strongly connected, then all θ_t^i will converge to θ^* both with probability 1 and in mean square.

Remark 3. Though Assumption 6 may look restrictive at first glance, simple simulations show that the sequences $\{\theta_t^i\}$, $i \in \mathcal{V}$, do not converge if the assumption does not hold (e.g., even when W_t changes periodically). It is worth emphasizing that the existence of π_∞ does not imply the existence of $\lim_{t\to\infty} W_t$, though the converse is true. Indeed, the assumption subsumes various cases including (a) all W_t are doubly stochastic matrices, and (b) all W_t share the same left eigenvector for eigenvalue 1, which may arise from the scenario when the number of in-neighbors of each agent does not change over time (Olshevsky & Tsitsiklis, 2013). An important implication of Assumption 6 is when the consensus interaction among the agents, characterized by $\{W_t\}$, is replaced by resilient consensus algorithms such as LeBlanc et al. (2013), Vaidya et al. (2012) in order to attenuate the effect of unknown malicious agents, the resulting dynamics of non-malicious agents, in general, will not converge, because the resulting interaction stochastic matrices among the non-malicious agents depend on the state values transmitted by the malicious agents, which can be arbitrary, and thus the resulting stochastic matrix sequence, in general, does not have a convergent absolute probability sequence; of course, in this case, the trajectories of all the non-malicious agents will still reach a consensus as long as the step-size is diminishing, as implied by Theorem 1. Further discussion on Assumption 6 can be found in Appendix B. □

We now study the finite-time performance of the proposed distributed linear stochastic approximation (1) for both fixed and time-varying step-size cases. Its finite-time performance is characterized by the following theorem.

Let $\eta_t = \|\pi_t - \pi_\infty\|_2$ for all $t \geq 0$. From Assumption 6, η_t converges to zero as $t \to \infty$.

Theorem 3. *Let the sequences $\{\theta_t^i\}$, $i \in \mathcal{V}$, be generated by (1). Suppose that Assumptions 1–4, 6 hold and $\{\mathbb{G}_t\}$ is uniformly strongly connected by sub-sequences of length L. Let q_t and m_t be the unique integer quotient and remainder of t divided by L, respectively. Let δ_t be the diameter of $\cup_{k=t}^{t+L-1} \mathbb{G}_k$, $\delta_{\max} = \max_{t\geq 0} \delta_t$, and*

$$\epsilon = \left(1 + \frac{2b_{\max}}{A_{\max}} - \frac{\pi_{\min}\beta^{2L}}{2\delta_{\max}}\right)(1 + \alpha A_{\max})^{2L} - \frac{2b_{\max}}{A_{\max}}(1 + \alpha A_{\max})^L, \tag{6}$$

where $0 < \alpha < \min\{K_1, \frac{\log 2}{A_{\max}\tau(\alpha)}, \frac{0.1}{K_2\gamma_{\max}}\}$.

(1) Fixed step-size: *Let $\alpha_t = \alpha$ for all $t \geq 0$. For all $t \geq T_1$,*

$$\begin{aligned}
&\sum_{i=1}^N \pi_t^i \mathbf{E}\left[\left\|\theta_t^i - \theta^*\right\|_2^2\right] \\
&\leq 2\epsilon^{q_t} \sum_{i=1}^N \pi_{m_t}^i \mathbf{E}\left[\left\|\theta_{m_t}^i - \langle\theta\rangle_{m_t}\right\|_2^2\right] + C_1\left(1 - \frac{0.9\alpha}{\gamma_{\max}}\right)^{t-T_1} \\
&\quad + C_2 + \frac{\gamma_{\max}}{\gamma_{\min}} 2\alpha\zeta_4 \sum_{k=0}^{t-T_1} \eta_{t+1-k}\left(1 - \frac{0.9\alpha}{\gamma_{\max}}\right)^k.
\end{aligned} \tag{7}$$

(2) Time-varying step-size: *Let $\alpha_t = \frac{\alpha_0}{t+1}$ with $\alpha_0 \geq \frac{\gamma_{\max}}{0.9}$. For all $t \geq LT_2$,*

$$\begin{aligned}
&\sum_{i=1}^N \pi_t^i \mathbf{E}\left[\left\|\theta_t^i - \theta^*\right\|_2^2\right] \\
&\leq 2\epsilon^{q_t - T_2} \sum_{i=1}^N \pi_{LT_2+m_t}^i \mathbf{E}\left[\left\|\theta_{LT_2+m_t}^i - \langle\theta\rangle_{LT_2+m_t}\right\|_2^2\right] \\
&\quad + C_3\left(\alpha_0 \epsilon^{\frac{q_t-1}{2}} + \alpha_{\lceil\frac{q_t-1}{2}\rceil L}\right) \\
&\quad + \frac{1}{t}\left(C_4 \log^2\left(\frac{t}{\alpha_0}\right) + C_5 \sum_{k=LT_2}^{t} \eta_k + C_6\right).
\end{aligned} \tag{8}$$

Here $T_1, T_2, K_1, K_2, C_1 - C_6$ are finite constants whose definitions are given in Appendix A.1.

Since π_t^i is uniformly bounded below by $\pi_{\min} \in (0, 1)$ from Lemma 1, it is easy to see that the above bound holds for each individual $\mathbf{E}[\|\theta_t^i - \theta^*\|_2^2]$. To better understand the theorem, we provide the following remark.

Remark 4. In Appendix D.2.1, we show that both ϵ and $(1 - \frac{0.9\alpha}{\gamma_{\max}})$ lie in the interval $(0, 1)$. It is easy to show that ϵ is monotonically increasing for $\delta_{\max}$ and L, monotonically decreasing for β and $\pi_{\min}$. Also,

$$\begin{aligned}
&\lim_{t\to\infty} \sum_{k=0}^{t-T_1} \eta_{t+1-k}\left(1 - \frac{0.9\alpha}{\gamma_{\max}}\right)^k \\
&= \lim_{t\to\infty} \sum_{l=0}^{\lfloor\frac{t-T_1}{2}\rfloor} \eta_{T_1+1+l}\left(1 - \frac{0.9\alpha}{\gamma_{\max}}\right)^{t-T_1-l} \\
&\quad + \sum_{l=\lceil\frac{t-T_1}{2}\rceil}^{t-T_1} \eta_{T_1+1+l}\left(1 - \frac{0.9\alpha}{\gamma_{\max}}\right)^{t-T_1-l} \\
&\leq \lim_{t\to\infty} \frac{\gamma_{\max}}{0.9\alpha}\left(\left(1 - \frac{0.9\alpha}{\gamma_{\max}}\right)^{\frac{t-T_1}{2}} \max_{l=0,\ldots,\lceil\frac{t-T_1}{2}\rceil} \eta_{T_1+1+l} \right. \\
&\quad \left. + \max_{l=\lceil\frac{t-T_1}{2}\rceil,\ldots,t-T_1+1} \eta_l\right) = 0.
\end{aligned}$$

Therefore, the summands in the finite-time bound (7) for the fixed step-size case are exponentially decaying except for the constant C_2, which implies that $\limsup_{t\to\infty}\sum_{i=1}^N \pi_t^i \mathbf{E}[\|\theta_t^i - \theta^*\|_2^2] \leq C_2$, providing a constant limiting bound. From Appendix A, C_2 is monotonically increasing for $\gamma_{\max}, \delta_{\max}, b_{\max}$ and L, and monotonically decreasing for $\gamma_{\min}, \pi_{\min}$ and β. In Appendix D.2.2, we show that $\lim_{t\to\infty}\frac{1}{t}\sum_{k=1}^t \eta_k = 0$, which implies that the finite-time bound (8) for the time-varying step-size case converges to zero as $t \to \infty$. We next comment on 0.1 in the inequality defining α. Actually, we can replace 0.1 with any constant $c \in (0,1)$, which will affect the value of ϵ and the feasible set of α, with the latter becoming $0 < \alpha < \min\{K_1, \frac{\log 2}{A_{\max}\tau(\alpha)}, \frac{c}{K_2\gamma_{\max}}\}$. Thus, the smaller the value of c is, the smaller is the feasible set of α, though the feasible set is always nonempty. For convenience, we simply pick $c = 0.1$ in this paper; that is why we also have 0.9 in (7). Lastly, we comment on α_0 in the time-varying step-size case. We set $\alpha_0 \geq \frac{\gamma_{\max}}{0.9}$ for the purpose of getting a cleaner expression of the finite-time bound. For $\alpha_0 < \frac{\gamma_{\max}}{0.9}$, our approach still works, but will yield a more complicated expression. The same is true for Theorem 5. □

Technical Challenge and Proof Sketch As described in the introduction, the key challenge of analyzing the finite-time performance of the distributed stochastic approximation (1) lies in the condition that the consensus-based interaction matrix is time-varying and stochastic (not necessarily doubly stochastic). To tackle this, we appeal to the absolute probability sequence π_t of the time-varying interaction matrix sequence and introduce the quadratic Lyapunov comparison function $\sum_{i=1}^N \pi_t^i \mathbf{E}[\|\theta_t^i - \theta^*\|_2^2]$. Then, using the inequality $\sum_{i=1}^N \pi_t^i \mathbf{E}[\|\theta_t^i - \theta^*\|_2^2] \leq 2\sum_{i=1}^N \pi_t^i \mathbf{E}[\|\theta_t^i - \langle\theta\rangle_t\|_2^2] + 2\mathbf{E}[\|\langle\theta\rangle_t - \theta^*\|_2^2]$, the next step is to find the finite-time bounds of $\sum_{i=1}^N \pi_t^i \mathbf{E}[\|\theta_t^i - \langle\theta\rangle_t\|_2^2]$ (Lemmas 4, 7) and $\mathbf{E}[\|\langle\theta\rangle_t - \theta^*\|_2^2]$ (Lemmas 5, 8), respectively. The latter term is essentially the "single-agent" mean-square error. Our main analysis contribution here is to bound the former term for both fixed and time-varying step-size cases.

3. Push-SA

The preceding section shows that the limiting state of consensus-based distributed stochastic approximation depends on π_∞, which leads to a convex combination of the local equilibria of all the agents in the absence of communication, but the convex combination is in general "uncontrollable". Note that this convex combination will correspond to a convex combination of the network-wise accumulative rewards in applications such as distributed TD learning. In an important case when the convex combination is desired to be the straight average, the existing literature e.g. Doan et al. (2019, 2021) relies on doubly stochastic matrices whose corresponding $\pi_\infty = (1/N)\mathbf{1}_N$. As mentioned in the introduction, doubly stochastic matrices implicitly require bi-directional communication between any pair of neighboring agents; see e.g. gossiping (Boyd, Ghosh, Prabhakar, & Shah, 2006; Liu, Mou, Morse, Anderson, & Yu, 2011) and the Metropolis algorithm (Xiao, Boyd, & Lall, 2005). A popular method to achieve the straight average target while allowing uni-directional communication between neighboring agents is to appeal to the idea so-called "push-sum" (Kempe, Dobra, & Gehrke, 2003), which was tailored for solving the distributed averaging problem over directed graphs and has been applied to distributed optimization (Nedić & Olshevsky, 2015). In this section, we will propose a push-based distributed stochastic approximation algorithm tailored for uni-directional communication and establish its finite-time error bound.

Each agent i has control over three variables, namely y_t^i, $\tilde{\theta}_t^i$ and θ_t^i, in which y_t^i is scalar-valued with initial value 1, $\tilde{\theta}_t^i$ can be arbitrarily initialized, and $\theta_0^i = \tilde{\theta}_0^i$. At each time $t \geq 0$, each agent i sends its weighted current values $\hat{w}_t^{ji} y_t^i$ and $\hat{w}_t^{ji}(\tilde{\theta}_t^i + \alpha_t A(X_t)\theta_t + \alpha_t b^i(X_t))$ to each of its current out-neighbors $j \in \mathcal{N}_t^{i-}$, and updates its variables as follows:

$$\begin{cases} y_{t+1}^i = \sum_{j\in\mathcal{N}_t^i} \hat{w}_t^{ij} y_t^j, \quad y_0^i = 1, \\ \tilde{\theta}_{t+1}^i = \sum_{j\in\mathcal{N}_t^i} \hat{w}_t^{ij}\left[\tilde{\theta}_t^j + \alpha_t\left(A(X_t)\theta_t^j + b^j(X_t)\right)\right], \\ \theta_{t+1}^i = \frac{\tilde{\theta}_{t+1}^i}{y_{t+1}^i}, \quad \theta_0^i = \tilde{\theta}_0^i, \end{cases} \tag{9}$$

where $\hat{w}_t^{ij} = 1/|\mathcal{N}_t^{j-}|$. It is worth noting that the algorithm is distributed yet requires that each agent be aware of the number of its out-neighbors.

Asymptotic performance of (9) with any uniformly strongly connected neighbor graph sequence is characterized by the following theorem.

Theorem 4. *Suppose that Assumptions 2–5 hold. Let $\{\theta_t^i\}$, $i \in \mathcal{V}$, be generated by (9) and $\theta^* \in \mathbb{R}^d$ be the unique equilibrium point of the ODE*

$$\dot{\theta} = A\theta + \frac{1}{N}\sum_{i=1}^N b^i, \tag{10}$$

where A and b^i are defined in Assumption 2. If $\{\mathbb{G}_t\}$ is uniformly strongly connected, then θ_t^i will converge to θ^* in mean square for all $i \in \mathcal{V}$.

In this section, we define $\langle\tilde{\theta}\rangle_t = \frac{1}{N}\sum_{i=1}^N \tilde{\theta}_t^i$ and $\langle\theta\rangle_t = \frac{1}{N}\sum_{i=1}^N \theta_t^i$. To help understand these definitions, let $\hat{W}_t$ be the $N \times N$ matrix whose ij-th entry equals $\hat{w}_t^{ij}$ if $j \in \mathcal{N}_t^i$, otherwise equals zero. It is easy to see that each $\hat{W}_t$ is a column stochastic matrix whose diagonal entries are all positive. Then, $\pi_t = \frac{1}{N}\mathbf{1}_N$ for all $t \geq 0$ can be regarded as an absolute probability sequence of $\{\hat{W}_t\}$. Thus, the above two definitions are intuitively consistent with $\langle\theta\rangle_t$ in the previous section.

Finite-time performance of (9) with any uniformly strongly connected neighbor graph sequence is characterized by the following theorem.

Let $\mu_t = \|A(X_t)(\langle\theta\rangle_t - \langle\tilde{\theta}\rangle_t)\|_2$. In Appendix D.3, we show that $\|\langle\theta\rangle_t - \langle\tilde{\theta}\rangle_t\|_2$ converges to zero as $t \to \infty$, so does μ_t.

Theorem 5. *Suppose that Assumptions 2–4 hold and $\{\mathbb{G}_t\}$ is uniformly strongly connected by sub-sequences of length L. Let $\{\theta_t^i\}$, $i \in \mathcal{V}$, be generated by (9) with $\alpha_t = \frac{\alpha_0}{t+1}$ and $\alpha_0 \geq \frac{\gamma_{\max}}{0.9}$. Then, there exists a nonnegative $\bar{\epsilon} \leq (1 - \frac{1}{N^{NL}})^{\frac{1}{L}}$ such that for all $t \geq \bar{T}$,*

$$\begin{aligned} &\sum_{i=1}^N \mathbf{E}\left[\left\|\theta_{t+1}^i - \theta^*\right\|_2^2\right] \\ &\leq C_7\bar{\epsilon}^t + C_8\left(\alpha_0\bar{\epsilon}^{\frac{t}{2}} + \alpha_{\lceil\frac{t}{2}\rceil}\right) + C_9\alpha_t \\ &+ \frac{1}{t}\left(C_{10}\log^2\left(\frac{t}{\alpha_0}\right) + C_{11}\sum_{k=\bar{T}}^t \mu_k + C_{12}\right), \end{aligned} \tag{11}$$

where $\bar{T}$ and $C_7 - C_{12}$ are finite constants whose definitions are given in Appendix A.2.

In Appendix D.3, we show that $\lim_{t\to\infty}\frac{1}{t}\sum_{k=1}^t \mu_k = 0$, which implies that the finite-time bound (11) converges to zero as $t \to \infty$. It is worth mentioning that the theorem does not consider the fixed step-size case, as our current analysis approach cannot be directly applied for this case.

Proof Sketch and Technical Challenge Using the inequality for any i

$$\mathbf{E}[\|\theta^i_{t+1}-\theta^*\|_2^2] \le 2\mathbf{E}[\|\theta^i_{t+1}-\langle\tilde{\theta}\rangle_t\|_2^2] + 2\mathbf{E}[\|\langle\tilde{\theta}\rangle_t-\theta^*\|_2^2],$$

our goal is to derive the finite-time bounds of $\mathbf{E}[\|\theta^i_{t+1}-\langle\tilde{\theta}\rangle_t\|_2^2]$ (Lemma 12) and $\mathbf{E}[\|\langle\tilde{\theta}\rangle_t-\theta^*\|_2^2]$ (Lemma 14), respectively. Although this looks similar to the proof of Theorem 3, the derivation is quite different. First, the iteration of $\langle\tilde{\theta}\rangle_t$ is a single-agent stochastic approximation (SA) plus a disturbance term $\langle\theta\rangle_t-\langle\tilde{\theta}\rangle_t$, so we cannot directly apply the existing single-agent SA finite-time analyses to bound $\mathbf{E}[\|\langle\tilde{\theta}\rangle_t-\theta^*\|_2^2]$; instead, we have to show that $\langle\theta\rangle_t-\langle\tilde{\theta}\rangle_t$ will diminish and quantify the diminishing "speed". Second, both the proof of showing diminishing $\langle\theta\rangle_t-\langle\tilde{\theta}\rangle_t$ and derivation of bounding $\sum_{i=1}^N \mathbf{E}[\|\theta^i_{t+1}-\langle\tilde{\theta}\rangle_t\|_2^2]$ involve a key challenge: to prove the sequence $\{\theta^i_t\}$ generated from the Push-SA (9) is bounded almost surely (Lemma 11). To tackle this, we introduce a novel way to constructing an absolute probability sequence for the Push-SA as follows. From (9), $\theta^i_{t+1}=\sum_{j=1}^N \tilde{w}^{ij}_t[\theta^j_t+\alpha_t A(X_t)\frac{\theta^j_t}{y^j_t}+\alpha_t\frac{b^j(X_t)}{y^j_t}]$, where $\tilde{w}^{ij}_t=(\hat{w}^{ij}_t y^j_t)/(\sum_{k=1}^N \hat{w}^{ik}_t y^k_t)$. We show that each matrix $\tilde{W}_t=[\tilde{w}^{ij}_t]$ is stochastic, and there exists a unique absolute probability sequence $\{\tilde{\pi}_t\}$ for the matrix sequence $\{\tilde{W}_t\}$ such that $\tilde{\pi}^i_t\ge\tilde{\pi}_{\min}$ for all $i\in\mathcal{V}$ and $t\ge 0$, with the constant $\tilde{\pi}_{\min}\in(0,1)$. Most importantly, we show two critical properties of $\{\tilde{W}_t\}$ and $\{\tilde{\pi}_t\}$ in Lemma 10, namely $\lim_{t\to\infty}(\Pi^t_{s=0}\tilde{W}_s)=\frac{1}{N}\mathbf{1}_N\mathbf{1}_N^\top$ and $\frac{\tilde{\pi}^i_t}{y^i_t}=\frac{1}{N}$ for all $i,j\in\mathcal{V}$ and $t\ge 0$, which have never been reported in the literature though push-sum-based distributed algorithms have been extensively studied.

Remark 5. It is worth mentioning that the approach for analyzing push-SA here can be leveraged to establish a better convergence rate for the subgradient-push algorithm proposed in Nedić and Olshevsky (2015); see a much more comprehensive development of the novel push-sum based analysis tool and its application in analyzing subgradient-push in Lin and Liu (2022). □

4. Concluding remarks

In this paper, we have established both asymptotic and non-asymptotic analyses for a consensus-based distributed linear stochastic approximation algorithm over uniformly strongly connected graphs, and proposed a push-based variant for coping with uni-directional communication. Both algorithms and their analyses can be directly applied to TD learning. One limitation of our finite-time bounds is that they involve quite a few constants which are well defined and characterized but whose values are not easy to compute. Future directions include leveraging the analyses for resilience in the presence of malicious agents and extending the tools to more complicated RL.

Appendix A. List of constants

In this appendix, we list all the constants used in our main results, Theorems 3 and 5. They are finite and their expressions do not affect the understanding of the theorems. Since their expressions are quite long and complicated, we begin with the following set of constants, based on which we will be able to present the constants used in the theorems and the proofs of the theorems in an easier way. We hope that this way can also help the readers to better understand and follow our results and analyses.

The first constant ζ_1 is defined as follows. Recall that ϵ is given in (6) as

$$\epsilon=\left(1+\frac{2b_{\max}}{A_{\max}}-\frac{\pi_{\min}\beta^{2L}}{2\delta_{\max}}\right)(1+\alpha A_{\max})^{2L}-\frac{2b_{\max}}{A_{\max}}(1+\alpha A_{\max})^{L}.$$

ζ_1 is defined as the unique solution for which $\epsilon=1$ if $\alpha=\zeta_1$. The following remark shows why ζ_1 uniquely exists.

Remark 6. From (6), it is easy to see that ϵ is monotonically increasing for $\alpha>0$. Define the corresponding monotonic function as

$$f(\alpha)=\left(1+\frac{2b_{\max}}{A_{\max}}-\frac{\pi_{\min}\beta^{2L}}{2\delta_{\max}}\right)(1+\alpha A_{\max})^{2L}-\frac{2b_{\max}}{A_{\max}}(1+\alpha A_{\max})^{L}.$$

Note that $0<f(0)<1$ and $f(+\infty)=+\infty$. Thus, $f(\alpha)=1$ has a unique solution ζ_1. □

The other constants are defined as follows:

$$\zeta_2=\frac{4b^2_{\max}}{A^2_{\max}}\left[(1+\alpha A_{\max})^L-1\right]^2+2b_{\max}\frac{(1+\alpha A_{\max})^L-1}{A_{\max}}(1+\alpha A_{\max})^L \tag{A.1}$$

$$\zeta_3=\left(144+4A^2_{\max}+912\tau(\alpha)A^2_{\max}+168\tau(\alpha)A_{\max}b_{\max}\right)\|\theta^*\|_2^2+2+2b^2_{\max}+4\|\theta^*\|_2^2+\frac{48b^2_{\max}}{A^2_{\max}}+\tau(\alpha)A^2_{\max}\left[152\left(\frac{b_{\max}}{A_{\max}}+\|\theta^*\|_2\right)^2+\frac{48b_{\max}}{A_{\max}}\left(\frac{b_{\max}}{A_{\max}}+1\right)^2+\frac{87b^2_{\max}}{A^2_{\max}}+\frac{12b_{\max}}{A_{\max}}\right] \tag{A.2}$$

$$\zeta_4=\sqrt{N}b_{\max}\left(2+\frac{12b^2_{\max}}{A^2_{\max}}+38\|\theta^*\|_2^2\right) \tag{A.3}$$

$$\zeta_5=144+916A^2_{\max}+168A_{\max}b_{\max} \tag{A.4}$$

$$\zeta_6=4b^2_{\max}\alpha L^2(1+\alpha A_{\max})^{2L-2}+2b_{\max}L(1+\alpha A_{\max})^{2L-1} \tag{A.5}$$

$$\zeta_7=(148+916A^2_{\max}+168A_{\max}b_{\max})\|\theta^*\|_2^2+2+\frac{48b^2_{\max}}{A^2_{\max}}+152\left(b_{\max}+A_{\max}\|\theta^*\|_2\right)^2+12A_{\max}b_{\max}+89b^2_{\max}+48A_{\max}b_{\max}\left(\frac{b_{\max}}{A_{\max}}+1\right)^2 \tag{A.6}$$

$$\zeta_8=144+916A^2_{\max}+168A_{\max}b_{\max}+144A_{\max}\mu_{\max} \tag{A.7}$$

$$\zeta_9=2+(4+\zeta_8)\|\theta^*\|_2^2+48\frac{(b_{\max}+\mu_{\max})^2}{A^2_{\max}}+152\left(b_{\max}+\mu_{\max}+A_{\max}\|\theta^*\|_2\right)^2+12A_{\max}b_{\max}+48A_{\max}(b_{\max}+\mu_{\max})\left(\frac{b_{\max}+\mu_{\max}}{A_{\max}}+1\right)^2+89(b_{\max}+\mu_{\max})^2 \tag{A.8}$$

Here $\mu_{\max}=(N+1)A_{\max}C_\theta$, where C_θ is a finite number defined in Lemma 11 which can be regarded as an upper bound of 2-norm of each agent i's state θ^i_t generated by the Push-SA algorithm (9).

A.1. Constants used in Theorem 3

$$K_1=\min\left\{\zeta_1,\ \frac{\gamma_{\max}}{0.9}\right\}$$

$$K_2=144+4A^2_{\max}+912\tau(\alpha)A^2_{\max}+168\tau(\alpha)A_{\max}b_{\max} \tag{A.9}$$

$$C_1 = \frac{\gamma_{\max}}{\gamma_{\min}}\left(8\exp\{2\alpha A_{\max}T_1\}+4\right)\mathbf{E}\left[\|\langle\theta\rangle_0-\theta^*\|_2^2\right] + 8\frac{\gamma_{\max}}{\gamma_{\min}}\exp\{2\alpha A_{\max}T_1\}\left(\|\theta^*\|_2+\frac{b_{\max}}{A_{\max}}\right)^2$$

$$C_2 = \frac{2\zeta_2}{1-\epsilon}+\frac{\gamma_{\max}}{\gamma_{\min}}\cdot\frac{2\alpha\zeta_3\gamma_{\max}}{0.9}$$

$$C_3 = \frac{2\zeta_6}{1-\epsilon}$$

$$C_4 = 2\zeta_7\alpha_0 C\frac{\gamma_{\max}}{\gamma_{\min}}$$

$$C_5 = 2\alpha_0\zeta_4\frac{\gamma_{\max}}{\gamma_{\min}}$$

$$C_6 = 2LT_2\frac{\gamma_{\max}}{\gamma_{\min}}\mathbf{E}\left[\|\langle\theta\rangle_{LT_2}-\theta^*\|_2^2\right]$$

T_1 is any positive integer such that for all $t \geq T_1$, there hold $t \geq \tau(\alpha)$ and $36\sqrt{N}b_{\max}\eta_{t+1}\gamma_{\max}+K_2\alpha\gamma_{\max}\leq 0.1$.

Remark 7. We show that T_1 must exist. From $0<\alpha<\min\{K_1, \frac{\log 2}{A_{\max}\tau(\alpha)}, \frac{0.1}{K_2\gamma_{\max}}\}$, it is easy to see that the feasible set of α is nonempty and $K_2\alpha\gamma_{\max}<0.1$. Since $\lim_{t\to\infty}\eta_t=0$ by Lemma 6 and $\tau(\alpha)\leq -C\log\alpha$ by Assumption 3, there exists a time instant $T\geq -C\log\alpha$ such that for any $t\geq T$, there hold $t\geq\tau(\alpha)$ and $\eta_{t+1}\leq(0.1-K_2\alpha\gamma_{\max})/(36\sqrt{N}b_{\max}\gamma_{\max})$, which implies that T_1 must exist. □

T_2 is any positive integer such that for all $t\geq LT_2$, there hold $\alpha_t\leq\alpha$, $2\tau(\alpha_t)\leq t$, $\tau(\alpha_t)\alpha_{t-\tau(\alpha_t)}\leq\min\{\frac{\log 2}{A_{\max}}, \frac{0.1}{\zeta_5\gamma_{\max}}\}$ and

$$\zeta_5\alpha_{t-\tau(\alpha_t)}\tau(\alpha_t)\gamma_{\max}+36\sqrt{N}b_{\max}\eta_{t+1}\gamma_{\max}\leq 0.1.$$

Remark 8. We explain why T_2 must exist. Since $\alpha_t=\frac{\alpha_0}{t+1}$ is monotonically decreasing for t and $\tau(\alpha_t)\leq -C\log\alpha_t = -C\log\alpha_0+C\log(t+1)$ from Assumption 3, there exists a positive S_1 such that for any $t\geq S_1$, we have $\alpha_t\leq\alpha$ and $t\geq 2\tau(\alpha_t)$ for any constant $0<\alpha<\min\{K_1, \frac{\log 2}{A_{\max}\tau(\alpha)}, \frac{0.1}{K_2\gamma_{\max}}\}$. Moreover, it is easy to show that

$$\lim_{t\to\infty} t-\tau(\alpha_t)\geq\lim_{t\to\infty} t+C\log\alpha_0-C\log(t+1) = +\infty,$$

$$\lim_{t\to\infty}\tau(\alpha_t)\alpha_{t-\tau(\alpha_t)}\leq\lim_{t\to\infty}\frac{-C\alpha_0\log\alpha_0+C\alpha_0\log(t+1)}{t-\tau(\alpha_t)+1} = 0.$$

Then, there exists a positive S_2 such that for any $t\geq S_2$, we have $\tau(\alpha_t)\alpha_{t-\tau(\alpha_t)}\leq\min\{\frac{\log 2}{A_{\max}}, \frac{0.1}{\zeta_5\gamma_{\max}}\}$. In addition, since $\lim_{t\to\infty}\eta_t=0$ from Lemma 6, when $\tau(\alpha_t)\alpha_{t-\tau(\alpha_t)}\leq\frac{0.1}{\zeta_5\gamma_{\max}}$, there exists a positive S_3 such that for any $t\geq S_3$, we have $\eta_{t+1}\leq(0.1-\zeta_5\alpha_{t-\tau(\alpha_t)}\tau(\alpha_t)\gamma_{\max})/(36\sqrt{N}b_{\max}\gamma_{\max})$. Thus, T_2 must exist as we can set $T_2=\max\{S_1,S_2,S_3\}$. □

A.2. Constants used in Theorem 5

$$C_7 = \frac{16}{\epsilon_1}\mathbf{E}\left[\left\|\sum_{i=1}^N\tilde{\theta}_0^i+\alpha_0 A(X_0)\tilde{\theta}_0^i+\alpha_0 b^i(X_0)\right\|_2\right]$$

$$C_8 = \frac{16}{\epsilon_1}\cdot\frac{A_{\max}C_\theta+b_{\max}}{1-\bar{\epsilon}}$$

$$C_9 = 2A_{\max}C_\theta+2b_{\max}$$

$$C_{10} = 2N\zeta_9\alpha_0 C\frac{\gamma_{\max}}{\gamma_{\min}}$$

$$C_{11} = 2\alpha_0 N\frac{\gamma_{\max}}{\gamma_{\min}}$$

$$C_{12} = 2\bar{T}N\frac{\gamma_{\max}}{\gamma_{\min}}\mathbf{E}\left[\|\langle\tilde{\theta}\rangle_{\bar{T}}-\theta^*\|_2^2\right]$$

Here ϵ_1 is a positive constant defined as

$$\epsilon_1 = \inf_{t\geq 0}\min_{i\in\mathcal{V}}(\hat{W}_t\cdots\hat{W}_0\mathbf{1}_N)^i.$$

From Corollary 2 (b) in Nedić and Olshevsky (2015) and the fact that each $\hat{W}_t$ is column stochastic, $\epsilon_1\in[\frac{1}{N^{NL}},1]$. See Lemma 12 for more details.

$\bar{T}$ is any positive integer such that for all $t\geq\bar{T}$, there hold $2\tau(\alpha_t)\leq t$, $\mu_t+\tau(\alpha_t)\alpha_{t-\tau(\alpha_t)}\zeta_8\leq\frac{0.1}{\gamma_{\max}}$ and $\tau(\alpha_t)\alpha_{t-\tau(\alpha_t)}\leq\min\{\frac{\log 2}{A_{\max}}, \frac{0.1}{\zeta_8\gamma_{\max}}\}$.

Remark 9. From Lemma 13, $\lim_{t\to\infty}\mu_t=0$. Then, using the similar arguments as in Remark 8, we can show the existence of $\bar{T}$. □

Appendix B. Discussion on Assumption 6

We contend that Assumption 6 has more general applications than the previously known case.

First, as mentioned in Remark 3, there are at least two cases which satisfy Assumption 6, yet cannot be directly handled by the existing analysis tool, which was developed only for doubly stochastic matrices. Case 1 is when the number of in-neighbors of agents is unchanged over time. This case has an interesting behavioral interpretation in fish biology, and has been adopted in bio-inspired distributed algorithm design (Abaid & Porfiri, 2010). Case 2 is when the interaction matrix changes arbitrarily over time during an initial period, after which it finally becomes fixed. As we describe below, Case 2 occurs naturally in certain multi-agent systems.

Case 1 is mathematically equivalent to the situation when all stochastic matrices share the same left dominant eigenvector, which subsumes doubly stochastic matrices as a special case; thus it could be analyzed by carefully choosing a fixed norm. There may be different choices: one choice is to apply our time-varying quadratic Lyapunov comparison function $\sum_{i=1}^N\pi_t^i\mathbf{E}[\|\theta_t^i-\theta^*\|_2^2]$ to the time-invariant case (i.e., π_t^i does not change over time), which leads to the weighted Frobenius norm defined in the appendix.

The extension to Case 1 just described may be straightforward, but Case 2 is not. As we proved in Theorems 2 and 3, when the interaction matrix arbitrarily changes over time for an initial period, say of length T, and finally becomes a fixed matrix or enters Case 1, all agents' trajectories determined by (1) will converge in mean square. Also, recall that the corresponding finite-time error bounds in this case were derived using the "absolute probability sequence" technique. Note that the existing techniques can only be applied to analyze (1) after time T; when T is very large, such an analysis is undesirable, since the focus and challenge here are for "finite" time.

It is important to note that Case 2 provides a realistic model for certain systems. Consider scenarios in which some agents do not function stably and thus they communicate with their neighbors sporadically for a certain period, leading to a time-varying stochastic matrix. Such scenarios occur naturally when there is unstable communication due to environmental changes or movement of agents (e.g., robots or UAVs may need to move into a new formation while continuing computation). After this unstable period, which could be long, the whole system then enters a stable operation status. This satisfies Case 2 and our finite-time analysis can be applied to the whole process, no matter how long the unstable period could be, as long as it is finite. In addition to this example, Case 2 and our analysis can be applied to certain scenarios in the presence of malicious agents. Suppose the system is aware that a small subset of agents have

potentially been attacked and are thus behaving maliciously. To protect the system, the consensus interaction among the agents can switch to resilient consensus algorithms such as LeBlanc et al. (2013), Vaidya et al. (2012) in order to attenuate the effect of malicious agents. In this situation, the resulting dynamics of the non-malicious agents are in general characterized by a time-varying stochastic matrix. After identifying and/or fixing the malicious agents, which could be a very slow process, the system can switch back to normal operation status. This example again satisfies Case 2, and our analysis can be applied to the whole procedure. As we mentioned in Remark 3, if some malicious agents always exist, the non-malicious agents in general will not converge, and thus a finite-time analysis is probably meaningless. The non-convergence issue will be further explained in the next subsection.

Appendix C. Distributed TD learning

In this section, we apply our distributed stochastic approximation finite-time analyses to distributed TD learning, as TD(λ) is a special cases of stochastic approximation. To this end, we first introduce the following multi-agent MDP tailored for distributed TD.

The multi-agent MDP can be defined by a tuple $(\mathcal{S}, \{\mathcal{U}^i\}_{i\in\mathcal{V}}, \{R^i\}_{i\in\mathcal{V}}, \bar{P}, \gamma, \{\mathbb{G}_t\}_{t\geq 0})$. Here $\mathcal{S} = \{1, \ldots, S\}$ is the finite set of S states and $\mathcal{U}^i$ is the set of control actions for agent i. For each agent i, $R^i : \mathcal{S}\times\mathcal{U}\times\mathcal{S} \to \mathbb{R}$ is the local reward function, where $\mathcal{U} = \prod_{i=1}^N \mathcal{U}^i$ is the joint control action space. In addition, $\bar{P} : \mathcal{S}\times\mathcal{U}\times\mathcal{S} \to [0,1]$ denotes the state transition probability matrix of the MDP, and $\gamma \in (0,1)$ is the discount factor. Given a fixed policy, let $\bar{P}$ be of size $S\times S$ for convenience, and thus its ij-th entry $\bar{p}^{ij}$ equals the probability from state i to state j under the given policy. The multi-agent MDP then evolves as follows. At each time $t \geq 0$, each agent i observes the current state $s_t \in \mathcal{S}$, takes action $u_t^i = \mu^i(s_t) \in \mathcal{U}^i$, and receive a corresponding reward $R^i(s_t, u_t, s_{t+1})$, where $\mu^i : \mathcal{S} \to \mathcal{U}^i$ is a function mapping a state to a control action in $\mathcal{U}^i$ and $u_t = \prod_{i=1}^N u^i \in \mathcal{U}$. It is worth emphasizing that in such a multi-agent setting, each agent's rewards and reward function are private information, and thus cannot be shared with any other agents.

The discounted accumulative reward $J : \mathcal{S} \to \mathbb{R}$ associated with the above multi-agent MDP is defined for each $s \in \mathcal{S}$ as

$$J(s) = \mathbf{E}\Big[\sum_{t=0}^{\infty}\gamma^t \sum_{i\in\mathcal{V}} c^i R^i(s_t, u_t, s_{t+1}) \mid s_0 = s\Big], \tag{C.1}$$

which satisfies the Bellman equation (Sutton & Barto, 2018), i.e.,

$$J(s) = \sum_{s'=1}^{S} \bar{p}^{ss'}\Big[\sum_{i\in\mathcal{V}} c^i R^i(s,s') + \gamma J(s')\Big], \quad s \in \mathcal{S},$$

where $c^i > 0$, $i \in \mathcal{V}$, is a set of convex combination weights. The existing distributed RL algorithms all set $c^i = 1/N$ for all $i \in \mathcal{V}$ (e.g., Doan et al. 2019, Zhang, Yang, Liu, et al. 2018), and this is why they require interaction matrices all be doubly stochastic. We will show that $c^i = \pi_\infty^i$ for all $i \in \mathcal{V}$ for general stochastic matrix sequences. Since for any doubly stochastic matrix sequence, its absolute probability sequence is $\pi_t = (1/N)\mathbf{1}_N$, i.e., $\pi_\infty^i = 1/N$ for all $i \in \mathcal{V}$, our results generalize the existing results, e.g. Doan et al. (2019, 2021). In Section 3, we will show how to achieve the straight average reward, i.e., $c^i = 1/N$ for all $i \in \mathcal{V}$, without requiring doubly stochastic matrices.

When the number of the states is very large, the computation of exact J may be intractable. To get around this, as did in Tsitsiklis and Van Roy (1997), we use a low-dimensional linear function $\hat{J}$ to approximate J. Specifically, the linear function approximator $\hat{J}$ takes the form $\hat{J}(s,\theta) = \sum_{k=1}^K \theta^k \phi_k^s$, $s \in \mathcal{S}$, where each ϕ_k^s is a fixed scalar function defined on the state space $\mathcal{S}$, each θ^k is the associated weight, and $K \ll S$. In other words, $\hat{J}$ is parameterized by $\theta \in \mathbb{R}^K$, with θ^k being the kth entry of θ. To proceed, let $\phi_k \in \mathbb{R}^S$ be the vector whose jth entry is ϕ_k^j for all $k \in \{1,\ldots,K\}$, let $\phi(s) \in \mathbb{R}^K$ be the vector whose jth entry is ϕ_j^s for all $s \in \mathcal{S}$, and let $\Phi \in \mathbb{R}^{S\times K}$ be the matrix whose ith row is the row vector $\phi(i)^\top$ and whose jth column is the vector ϕ_j, i.e., $\Phi = [\phi_1, \ldots, \phi_K] = [\phi(1), \ldots, \phi(S)]^\top \in \mathbb{R}^{S\times K}$, which implies $\hat{J} = \Phi\theta$. The goal for the multi-agent network is to find an optimal θ^* with which the distance between $\hat{J}$ and J is minimized, under the following standard assumptions adopted in e.g. Doan et al. (2019), Srikant and Ying (2019).

Assumption 7. All the rewards are uniformly bounded, i.e., there exists a positive constant R such that $|R^i(s,s')| \leq R$ for all $i \in \mathcal{V}$ and $s, s' \in \mathcal{S}$.

Assumption 8. The vectors $\phi_1, \ldots, \phi_K$ are linearly independent, i.e., Φ has full column rank, and $\|\phi(s)\|_2 \leq 1$ for all $s \in \mathcal{S}$.

Assumption 9. The Markov chain that evolves according to the transition probability matrix $\bar{P}$ is irreducible and aperiodic.

Under Assumption 9, let $d \in \mathbb{R}^S$ be the unique stationary distribution associated with $\bar{P}$, i.e., $d^\top\bar{P} = d^\top$.

C.1. Distributed TD(λ)

In this subsection, we make use of TD(λ) to estimate θ^* in a distributed manner. Note that TD(0) can be applied in a similar manner. Each agent $i \in \mathcal{V}$ updates its own estimator of θ^*, θ_t^i, for all time $t \in \{0, 1, 2, \ldots\}$ as follows:

$$\theta_{t+1}^i = \sum_{j\in\mathcal{N}_t^i} w_t^{ij}\theta_t^j + \alpha_t\Big(A(X_t)\sum_{j\in\mathcal{N}_t^i} w_t^{ij}\theta_t^j + b^i(X_t)\Big), \tag{C.2}$$

where $X_t = (s_t, s_{t+1}, z_t)$ is the Markov chain, with $z_t = \sum_{k=0}^t (\gamma\lambda)^{t-k}\phi(s_k)$, and

$$A(X_t) = z_t(\gamma\phi(s_{t+1}) - \phi(s_t))^\top, \quad b^i(X_t) = r_t^i z_t, \tag{C.3}$$

with r_t^i being the reward for agent i at time t. It is worth emphasizing that the proposed TD(λ) algorithm is different from that in Doan et al. (2021).

In the sequel, we will show that the update (C.2) with (C.3) is a special case of (1) so that our analysis for (1) can be applied here. To this end, let $D = \mathrm{diag}(d) \in \mathbb{R}^{S\times S}$, where d is given right after Assumption 9,

$$A = \Phi^\top D(U - I)\Phi, \quad U = (1-\lambda)\sum_{t=0}^{\infty}\lambda^t(\gamma\bar{P})^{t+1},$$

$$b^i = \Phi^\top D\sum_{t=0}^{\infty}(\gamma\lambda\bar{P})^t r^i, \quad i \in \mathcal{V}, \tag{C.4}$$

where $r^i \in \mathbb{R}^S$ whose kth entry is $r^{ik} = \sum_{s=1}^S \bar{p}^{ks}R^i(k,s)$, and set $A_{\max} = \frac{1+\gamma}{1-\gamma\lambda}$ and $b_{\max} = \frac{R}{1-\gamma\lambda}$, where R is given in Assumption 7.

Lemma 2. *Let the sequences $\{\theta_t^i\}$, $i \in \mathcal{V}$, be generated by (C.2) with (C.3). If Assumptions 7–9 hold, so do Assumptions 2–4.*

Lemma 2 implies that our analysis for (1) can be applied here. From the proof of Theorem 1 in Tsitsiklis and Van Roy (1997), A in (C.4) is a negative definite matrix, which implies that $A + A^\top$ is a symmetric negative definite matrix. From Theorem 7.11 in Rugh (1996), A is a Hurwitz matrix. Let $\sigma_{\min} > 0$ be the smallest

eigenvalue of $-\frac{1}{2}(A+A^\top)$. Thus, we can also choose $P=I$ in Assumption 4 and use the Lyapunov function $V(\langle\theta\rangle_t)=\|\langle\theta\rangle_t-\theta^*\|_2^2$ in the analysis, where θ^* here is the limiting point of (C.2). Using the same argument as in Theorem 2, we can show that θ^* is the unique equilibrium point of the ODE (5) with A and b^i being defined in (C.4).

C.2. Push-TD(λ)

In this subsection, we propose a push-based distributed TD(λ) algorithm and provide its finite-time error bounds. Note that push-based distributed TD(0) can be applied in the similar manner. Each agent $i\in\mathcal{V}$ updates its variables at each time $t\geq 0$ as follows:

$$\begin{cases} y_{t+1}^i=\sum_{j\in\mathcal{N}_t^i}\hat{w}_t^{ij}y_t^j, \quad y_0^j=1,\\ \hat{\theta}_{t+1}^i=\sum_{j\in\mathcal{N}_t^i}\hat{w}_t^{ij}\hat{\theta}_t^j+\alpha_t\Big(A(X_t)\hat{w}_t^{ij}\theta_t^j+b^i(X_t)\Big),\\ \theta_{t+1}^i=\frac{\hat{\theta}_{t+1}^i}{y_{t+1}^i}, \end{cases}$$

where $\hat{w}_t^{ij}=1/|\mathcal{N}_t^j|$, $X_t=(s_t,s_{t+1},z_t)$ is the Markov chain, with $z_t=\sum_{k=0}^t(\gamma\lambda)^{t-k}\phi(s_k)$, $A(X_t)$ and $b^i(X_t)$ are given in (C.3). Using the same argument as in Theorem 4, we can show that θ^* is the unique equilibrium point of the ODE (10) with A and b^i being defined in (C.4).

It is not hard to verify that Theorems 3 and 5 can be applied to Distributed TD(λ) and Push-TD(λ) to obtain their finite-time performance bounds, respectively.

Appendix D. Analysis and some proofs

In this appendix, we provide the analysis of our two algorithms, (1) and (9), and the proofs of all the assertions in the paper. We begin with some notation used in the analysis.

D.1. Notation

We use $\mathbf{0}_n$ to denote the vector in $\mathbb{R}^n$ whose entries all equal to 0's. For any vector $x\in\mathbb{R}^n$, we use $\mathrm{diag}(x)$ to denote the $n\times n$ diagonal matrix whose ith diagonal entry equals x^i. We use $\|\cdot\|_F$ to denote the Frobenius norm. For any positive diagonal matrix $W\in\mathbb{R}^{n\times n}$, we use $\|A\|_W$ to denote the weighted Frobenius norm for $A\in\mathbb{R}^{n\times m}$, defined as $\|A\|_W=\|W^{\frac{1}{2}}A\|_F$. It is easy to see that $\|\cdot\|_W$ is a matrix norm. We use $\mathbf{P}(\cdot)$ to denote the probability of an event and $\mathbf{E}(X)$ to denote the expected value of a random variable X.

D.2. Distributed stochastic approximation

In this subsection, we analyze the distributed stochastic approximation algorithm (1) and provide the proofs of the results in Section 2. We begin with the asymptotic performance.

Proof of Theorem 1. Without loss of generality, let $\{\mathbb{G}_t\}$ be uniformly strongly connected by sub-sequences of length L. Note that for any $i\in\mathcal{V}$,

$$\begin{aligned} 0&\leq\pi_{\min}\|\theta_t^i-\langle\theta\rangle_t\|_2^2\leq\pi_{\min}\sum_{j=1}^N\|\theta_t^j-\langle\theta\rangle_t\|_2^2\\ &\leq\sum_{j=1}^N\pi_t^j\|\theta_t^j-\langle\theta\rangle_t\|_2^2, \end{aligned}\tag{D.1}$$

where $\pi_{\min}$ is defined in Lemma 1. From Lemma 7,

$$\begin{aligned} &\lim_{t\to\infty}\sum_{i=1}^N\pi_t^i\|\theta_t^i-\langle\theta\rangle_t\|_2^2\\ &\leq\lim_{t\to\infty}\hat{\epsilon}^{q_t-T_4^*}\sum_{i=1}^N\pi_{T_4^*L+m_t}^i\|\theta_{T_4^*L+m_t}^i-\langle\theta\rangle_{T_4^*L+m_t}\|_2^2\\ &\quad+\lim_{t\to\infty}\frac{\zeta_6}{1-\hat{\epsilon}}\Big(\alpha_0\hat{\epsilon}^{\frac{q_t-1}{2}}+\alpha_{\lceil\frac{q_t-1}{2}\rceil L}\Big)=0. \end{aligned}\tag{D.2}$$

Combining (D.1) and (D.2), it follows that for all $i\in\mathcal{V}$, $\lim_{t\to\infty}\pi_{\min}\|\theta_t^i-\langle\theta\rangle_t\|_2^2=0$. Since $\pi_{\min}>0$ by Lemma 1, $\lim_{t\to\infty}\|\theta_t^i-\langle\theta\rangle_t\|_2=0$ for all $i\in\mathcal{V}$. ■

Proof of Theorem 2. From Theorem 1, all θ_t^i, $i\in\mathcal{V}$, will reach a consensus with $\langle\theta\rangle_t$ and the update of $\langle\theta\rangle_t$ is given in (4), which can be treated as a single-agent linear stochastic approximation whose corresponding ODE is (5). From Kushner (1983), Kushner and Yin (1987),[1] we know that $\langle\theta\rangle_t$ will converge to θ^* w.p.1, which implies that θ_t^i will converge to θ^* w.p.1. In addition, from Theorem 3-(2) and Lemma 6, $\lim_{\to\infty}\sum_{i=1}^N\pi_t^i\mathbf{E}[\|\theta_t^i-\theta^*\|_2^2]=0$. Since π_t^i is uniformly bounded below by $\pi_{\min}>0$, as shown in Lemma 1, it follows that θ_t^i will converge to θ^* in mean square for all $i\in\mathcal{V}$. ■

We now analyze the finite-time performance of (1). In the sequel, we use K to denote the dimension of each θ_t^i, i.e., $\theta_t^i\in\mathbb{R}^K$ for all $i\in\mathcal{V}$.

D.2.1. Fixed step-size

We first consider the fixed step-size case and begin with validation of two "convergence rates" in Theorem 3.

Lemma 3. *Both ϵ and $(1-\frac{0.9\alpha}{\gamma_{\max}})$ lie in the interval $(0,1)$.*

Lemma 4. *Suppose that Assumptions 1 and 2 hold and $\{\mathbb{G}_t\}$ is uniformly strongly connected by sub-sequences of length L. Then, when $\alpha\in(0,\zeta_1)$, we have for all $t\geq\tau(\alpha)$,*

$$\sum_{i=1}^N\pi_t^i\|\theta_t^i-\langle\theta\rangle_t\|_2^2\leq\epsilon^{q_t}\sum_{i=1}^N\pi_{m_t}^i\|\theta_{m_t}^i-\langle\theta\rangle_{m_t}\|_2^2+\frac{\zeta_2}{1-\epsilon},$$

where ζ_1 is defined in Appendix A, ϵ and ζ_2 are defined in (6) and (A.1), respectively.

Lemma 5. *Suppose that Assumptions 2–4 and 6 hold. Then, when $0<\alpha<\min\{\frac{\log 2}{A_{\max}\tau(\alpha)},\ \frac{0.1}{K_2\gamma_{\max}}\}$, we have for any $t\geq T_1$,*

$$\begin{aligned} &\mathbf{E}[\|\langle\theta\rangle_{t+1}-\theta^*\|_2^2]\\ &\leq\Big(1-\frac{0.9\alpha}{\gamma_{\max}}\Big)^{t-T_1}\frac{\gamma_{\max}}{\gamma_{\min}}\mathbf{E}\left[\|\langle\theta\rangle_{T_1}-\theta^*\|_2^2\right]+\frac{\alpha\zeta_3\gamma_{\max}^2}{0.9\gamma_{\min}}\\ &\quad+\frac{\gamma_{\max}}{\gamma_{\min}}\alpha\zeta_4\sum_{k=0}^{t-T_1}\eta_{t+1-k}\Big(1-\frac{0.9\alpha}{\gamma_{\max}}\Big)^k\\ &\leq\Big(1-\frac{0.9\alpha}{\gamma_{\max}}\Big)^{t+1-T_1}\frac{C_1}{2}+\frac{\alpha\zeta_3\gamma_{\max}^2}{0.9\gamma_{\min}}\\ &\quad+\frac{\gamma_{\max}}{\gamma_{\min}}\alpha\zeta_4\sum_{k=0}^{t-T_1}\eta_{t+1-k}\Big(1-\frac{0.9\alpha}{\gamma_{\max}}\Big)^k. \end{aligned}$$

where C_1, ζ_3, ζ_4 and K_2 are defined in Appendix A.1, (A.2), (A.3) and (A.9), respectively.

[1] On page 1289 of Kushner and Yin (1987), it says that the idea in Kushner (1983) can be adapted to get the w.p.1 convergence result.

We are now in a position to prove the fixed step-size case in Theorem 3.

Proof of Case (1) in Theorem 3. From Lemmas 4 and 5, we have for any $t \geq T_1$,

$$\begin{aligned}
&\sum_{i=1}^{N} \pi_t^i \mathbf{E}[\|\theta_t^i - \theta^*\|_2^2] \\
&\leq 2\sum_{i=1}^{N} \pi_t^i \mathbf{E}[\|\theta_t^i - \langle\theta\rangle_t\|_2^2] + 2\mathbf{E}[\|\langle\theta\rangle_t - \theta^*\|_2^2] \\
&\leq 2\epsilon^{q_t} \sum_{i=1}^{N} \pi_{m_t}^i \mathbf{E}[\|\theta_{m_t}^i - \langle\theta\rangle_{m_t}\|_2^2] \\
&\quad + \frac{2\zeta_2}{1-\epsilon} + \frac{2\alpha\zeta_3\gamma_{\max}^2}{0.9\gamma_{\min}} + \left(1 - \frac{0.9\alpha}{\gamma_{\max}}\right)^{t-T_1} C_1 \\
&\quad + \frac{\gamma_{\max}}{\gamma_{\min}} 2\alpha\zeta_4 \sum_{k=0}^{t-T_1} \eta_{t+1-k}\left(1 - \frac{0.9\alpha}{\gamma_{\max}}\right)^k \\
&\leq 2\epsilon^{q_t} \sum_{i=1}^{N} \pi_{m_t}^i \mathbf{E}[\|\theta_{m_t}^i - \langle\theta\rangle_{m_t}\|_2^2] + C_1\left(1 - \frac{0.9\alpha}{\gamma_{\max}}\right)^{t-T_1} \\
&\quad + C_2 + \frac{\gamma_{\max}}{\gamma_{\min}} 2\alpha\zeta_4 \sum_{k=0}^{t-T_1} \eta_{t+1-k}\left(1 - \frac{0.9\alpha}{\gamma_{\max}}\right)^k,
\end{aligned}$$

where C_1 and C_2 are defined in Appendix A.1. This completes the proof. ■

D.2.2. Time-varying step-size

In this subsection, we consider the time-varying step-size case and begin with a property of η_t.

Lemma 6. *Suppose that Assumption 6 holds. Then,* $\lim_{t\to\infty} \eta_t = 0$ *and* $\lim_{t\to\infty} \frac{1}{t+1}\sum_{k=0}^{t} \eta_k = 0$.

To prove the theorem, we need the following lemmas.

Lemma 7. *Suppose that Assumptions 1 and 2 hold and* $\{\mathbb{G}_t\}$ *is uniformly strongly connected by sub-sequences of length L. Given* α_t *and* T_2 *defined in Theorem 3, for all* $t \geq T_2 L$,

$$\begin{aligned}
\sum_{i=1}^{N} \pi_t^i \|\theta_t^i - \langle\theta\rangle_t\|_2^2 &\leq \frac{\zeta_6}{1-\epsilon}\left(\epsilon^{\frac{q_t-1}{2}}\alpha_{m_t} + \alpha_{\lceil\frac{q_t-1}{2}\rceil L + m_t}\right) \\
&\quad + \epsilon^{q_t - T_2} \sum_{i=1}^{N} \pi_{T_2 L + m_t}^i \|\theta_{T_2 L + m_t}^i - \langle\theta\rangle_{T_2 L + m_t}\|_2^2,
\end{aligned}$$

which implies that

$$\begin{aligned}
\sum_{i=1}^{N} \pi_t^i \|\theta_t^i - \langle\theta\rangle_t\|_2^2 &\leq \frac{\zeta_6}{1-\epsilon}\left(\alpha_0\epsilon^{\frac{q_t-1}{2}} + \alpha_{\lceil\frac{q_t-1}{2}\rceil L}\right) \\
&\quad + \epsilon^{q_t - T_2} \sum_{i=1}^{N} \pi_{T_2 L + m_t}^i \|\theta_{T_2 L + m_t}^i - \langle\theta\rangle_{T_2 L + m_t}\|_2^2,
\end{aligned}$$

where ϵ *and* ζ_6 *are defined in (6) and (A.5), respectively.*

Lemma 8. *Under Assumptions 1–6, when the*

$$\tau(\alpha_t)\alpha_{t-\tau(\alpha_t)} \leq \min\{\frac{\log 2}{A_{\max}}, \frac{0.1}{\zeta_5\gamma_{\max}}\},$$

we have for any $t \geq T_2 L$,

$$\begin{aligned}
\mathbf{E}\left[\|\langle\theta\rangle_t - \theta^*\|_2^2\right] &\leq \frac{T_2 L}{t}\frac{\gamma_{\max}}{\gamma_{\min}}\mathbf{E}[\|\langle\theta\rangle_{T_2 L} - \theta^*\|_2^2] \\
&\quad + \frac{\zeta_7\alpha_0 C\log^2(\frac{t}{\alpha_0})}{t}\frac{\gamma_{\max}}{\gamma_{\min}} + \alpha_0\zeta_4\frac{\gamma_{\max}}{\gamma_{\min}}\frac{\sum_{l=T_2 L}^{t}\eta_l}{t},
\end{aligned}$$

where T_2 *is defined in Appendix A.1, and* ζ_4, ζ_5, ζ_7 *are defined in (A.3), (A.4), (A.6), respectively.*

We are now in a position to prove the time-varying step-size case in Theorem 3.

Proof of Case (2) in Theorem 3. From Lemmas 7 and 8, for any $t \geq T_2 L$, we have

$$\begin{aligned}
&\sum_{i=1}^{N} \pi_t^i \mathbf{E}[\|\theta_t^i - \theta^*\|_2^2] \\
&\leq 2\sum_{i=1}^{N} \pi_t^i \mathbf{E}[\|\theta_t^i - \langle\theta\rangle_t\|_2^2] + 2\mathbf{E}[\|\langle\theta\rangle_t - \theta^*\|_2^2] \\
&\leq 2\epsilon^{q_t - T_2} \sum_{i=1}^{N} \pi_{T_2 L + m_t}^i \mathbf{E}[\|\theta_{T_2 L + m_t}^i - \langle\theta\rangle_{T_2 L + m_t}\|_2^2] \\
&\quad + \frac{2T_2 L}{t}\frac{\gamma_{\max}}{\gamma_{\min}}\mathbf{E}[\|\langle\theta\rangle_{T_2 L} - \theta^*\|_2^2] + 2\alpha_0\zeta_4\frac{\gamma_{\max}}{\gamma_{\min}}\frac{\sum_{l=T_2 L}^{t}\eta_l}{t} \\
&\quad + \frac{2\zeta_7\alpha_0 C\log^2(\frac{t}{\alpha_0})}{t}\frac{\gamma_{\max}}{\gamma_{\min}} + \frac{2\zeta_6}{1-\epsilon}(\alpha_0\epsilon^{\frac{q_t-1}{2}} + \alpha_{\lceil\frac{q_t-1}{2}\rceil L}) \\
&\leq 2\epsilon^{q_t - T_2} \sum_{i=1}^{N} \pi_{LT_2 + m_t}^i \mathbf{E}\left[\left\|\theta_{LT_2 + m_t}^i - \langle\theta\rangle_{LT_2 + m_t}\right\|_2^2\right] \\
&\quad + C_3\left(\alpha_0\epsilon^{\frac{q_t-1}{2}} + \alpha_{\lceil\frac{q_t-1}{2}\rceil L}\right) \\
&\quad + \frac{1}{t}\left(C_4\log^2\left(\frac{t}{\alpha_0}\right) + C_5\sum_{k=LT_2}^{t}\eta_k + C_6\right),
\end{aligned}$$

where $C_3 - C_6$ are defined in Appendix A.1. This completes the proof. ■

Remark 10. For distributed SA algorithms, finite-time performance analysis essentially boils down to two parts, namely bounding the consensus error and bounding the "single-agent" mean-square error. For the case when consensus interaction matrices are all doubly stochastic, the consensus error bound can be derived by analyzing the square of the 2-norm of the deviation of the current state of each agent from the average of the states of the agents. With consensus in the presence of doubly stochastic matrices, the average of the states of the agents remains invariant. Thus, it is possible to treat the average value as the state of a fictitious agent to derive the mean-square consensus error bound with respect to the limiting point. More formally, this process relies on two properties of a doubly stochastic matrix W, namely that (1) $\mathbf{1}^\top W = \mathbf{1}^\top$, and (2) if $x_{t+1} = Wx_t$, then $\|x_{t+1} - (\mathbf{1}^\top x_{t+1})\mathbf{1}\|_2 \leq \sigma_2(W)\|x_t - (\mathbf{1}^\top x_t)\mathbf{1}\|_2$ where $\sigma_2(W)$ denotes the second largest singular value of W (which is strictly less than one if W is irreducible). Even if the doubly stochastic matrix is time-varying (denoted by W_t), property (1) still holds and property (2) can be generalized as in Nedić, Olshevsky, and Rabbat (2018). Thus, the square of the 2-norm $\|x_t - (\mathbf{1}^\top x_t)\mathbf{1}\|_2^2$ is a quadratic Lyapunov function for the average consensus processes. Doubly stochastic matrices in expectation can be treated in the same way by looking at the expectation. This is the core on which all the existing finite-time analyses of distributed RL algorithms are based. However, if each consensus interaction matrix is stochastic, and not necessarily doubly stochastic, the above two properties may not hold. In fact, it is well known that quadratic Lyapunov functions for general consensus processes $x_{t+1} = S_t x_t$, with S_t being stochastic, do not

exist (Olshevsky & Tsitsiklis, 2008). Here we appeal to the idea of quadratic comparison functions for general consensus processes. This was first proposed in Touri (2012) and makes use of the concept of absolute probability sequences. We provide a general analysis method and results that subsume the existing finite-time analyses for single-timescale distributed linear stochastic approximation (Lemmas 4, 5, 7 and 8) and TD learning as special cases. □

D.3. Push-SA

In this subsection, we analyze the push-based distributed stochastic approximation algorithm (9) and provide the proofs of the results in Section 3.

Let $\hat{W}_t$ be the matrix whose ij-th entry is $\hat{w}_t^{ij}$. Then, from (9),

$$\begin{aligned}\theta_{t+1}^i &= \frac{\tilde{\theta}_{t+1}^i}{y_{t+1}^i} = \frac{\sum_{j=1}^N \hat{w}_t^{ij}(\tilde{\theta}_t^j + \alpha_t A(X_t)\theta_t^j + \alpha_t b^j(X_t))}{y_{t+1}^i} \\ &= \sum_{j=1}^N \frac{\hat{w}_t^{ij} y_t^j}{\sum_{k=1}^N \hat{w}_t^{ik} y_t^k}\left[\frac{\tilde{\theta}_t^j}{y_t^j} + \alpha_t A(X_t)\frac{\theta_t^j}{y_t^j} + \alpha_t \frac{b^j(X_t)}{y_t^j}\right] \\ &= \sum_{j=1}^N \tilde{w}_t^{ij}\left[\theta_t^j + \alpha_t A(X_t)\frac{\theta_t^j}{y_t^j} + \alpha_t \frac{b^j(X_t)}{y_t^j}\right], \end{aligned} \tag{D.3}$$

where $\tilde{w}_t^{ij} = \frac{\hat{w}_t^{ij} y_t^j}{\sum_{k=1}^N \hat{w}_t^{ik} y_t^k}$ and $\tilde{W}_t = [\tilde{w}_t^{ij}]$ is a row stochastic matrix, i.e., $\sum_{j=1}^N \tilde{w}_t^{ij} = \frac{\sum_{j=1}^N \hat{w}_t^{ij} y_t^j}{\sum_{k=1}^N \hat{w}_t^{ik} y_t^k} = 1$, for all i.

Let $\Theta_t = [\theta_t^1, \dots, \theta_t^N]^\top$. Then (D.3) can be written as

$$\begin{aligned}\Theta_{t+1} = \tilde{W}_t\Bigg[\Theta_t + \alpha_t \begin{bmatrix} (\theta_t^1)^\top/y_t^1 \\ \cdots \\ (\theta_t^N)^\top/y_t^N \end{bmatrix} A(X_t)^\top \\ + \alpha_t \begin{bmatrix} (b^1(X_t))^\top/y_t^1 \\ \cdots \\ (b^N(X_t))^\top/y_t^N \end{bmatrix}\Bigg]. \end{aligned} \tag{D.4}$$

Since each matrix $\tilde{W}_t = [\tilde{w}_t^{ij}]$ is stochastic, from Lemma 1, there exists a unique absolute probability sequence $\{\tilde{\pi}_t\}$ for the matrix sequence $\{\tilde{W}_t\}$ such that $\tilde{\pi}_t^i \geq \tilde{\pi}_{\min}$ for all $i \in \mathcal{V}$ and $t \geq 0$, with the constant $\tilde{\pi}_{\min} \in (0, 1)$.

Lemma 9. *Suppose that $\{\mathbb{G}_t\}$ is uniformly strongly connected. Then, $\Pi_{s=0}^t \hat{W}_s$ will converge to the set $\{v\mathbf{1}_N^\top : v \in \mathbb{R}^N\}$ exponentially fast as $t \to \infty$.*

Lemma 10. *Suppose that $\{\mathbb{G}_t\}$ is uniformly strongly connected. Then, $(\Pi_{l=s}^t \tilde{W}_l)^{ij} = \frac{y_s^j}{y_{t+1}^i}(\Pi_{l=s}^t \hat{W}_l)^{ij}$ and $\frac{\tilde{\pi}_s^i}{y_s^i} = \frac{1}{y_s^i}\lim_{t\to\infty}(\Pi_{l=s}^t \tilde{W}_l)^{ji} = \frac{1}{N}$ for all $i, j \in \mathcal{V}$ and $s \geq 0$.*

Lemma 11. *The sequence $\{\Theta_n\}$ generated by (D.4) is bounded almost surely, i.e., $C_\theta = \sup_n \|\Theta_n\|_F < \infty$ almost surely.*

From (9), by using the definition of $\langle\tilde{\theta}\rangle_t = \frac{1}{N}\sum_{i=1}^N \tilde{\theta}_t^i$ and $\langle\theta\rangle_t = \frac{1}{N}\sum_{i=1}^N \theta_t^i$, we have

$$\begin{aligned}\langle\tilde{\theta}\rangle_{t+1} &= \langle\tilde{\theta}\rangle_t + \alpha_t A(X_t)\langle\theta\rangle_t + \frac{\alpha_t}{N}\sum_{i=1}^N b^i(X_t) \\ &= \langle\tilde{\theta}\rangle_t + \alpha_t A(X_t)\langle\tilde{\theta}\rangle_t + \frac{\alpha_t}{N}\sum_{i=1}^N b^i(X_t) + \alpha_t\rho_t, \end{aligned} \tag{D.5}$$

where $\rho_t = A(X_t)\langle\theta\rangle_t - A(X_t)\langle\tilde{\theta}\rangle_t$. From Lemma 11, we have $\|\langle\theta\rangle_t\|_2 \leq \max_{i\in\mathcal{V}}\|\theta_t^i\|_2 \leq C_\theta$ for all $t \geq 0$, which implies that $\|\langle\tilde{\theta}\rangle_t\|_2 \leq NC_\theta$ and $\mu_t = \|\rho_t\|_2 = \left\|A(X_t)\langle\theta\rangle_t - A(X_t)\langle\tilde{\theta}\rangle_t\right\|_2 \leq \mu_{\max}$, where $\mu_{\max} = (N+1)A_{\max}C_\theta$.

Lemma 12. *Suppose that Assumptions 2 and 5 hold and $\{\mathbb{G}_t\}$ is uniformly strongly connected by sub-sequences of length L. Let $\epsilon_1 = \inf_{t\geq 0}\min_{i\in\mathcal{V}}(\hat{W}_t\cdots\hat{W}_0\mathbf{1}_N)^i$. For all $t \geq 0$ and $i \in \mathcal{V}$,*

$$\begin{aligned}&\|\theta_{t+1}^i - \langle\tilde{\theta}\rangle_t\|_2 \\ &\leq \frac{8}{\epsilon_1}\bar{\epsilon}^t\|\sum_{i=1}^N \tilde{\theta}_0^i + \alpha_0 A(X_0)\theta_0^i + \alpha_0 b^i(X_0)\|_2 + \alpha_t b_{\max} \\ &\quad + \frac{8}{\epsilon_1}\frac{A_{\max}C_\theta + b_{\max}}{1-\bar{\epsilon}}\left(\alpha_0\bar{\epsilon}^{t/2} + \alpha_{\lceil\frac{t}{2}\rceil}\right) + \alpha_t A_{\max}C_\theta, \end{aligned}$$

where $\epsilon_1 > 0$ and $\bar{\epsilon} \in (0, 1)$ satisfy $\epsilon_1 \geq \frac{1}{N^{NL}}$ and $\bar{\epsilon} \leq (1 - \frac{1}{N^{NL}})^{1/L}$.

Lemma 13. *$\lim_{t\to\infty}\mu_t = \lim_{t\to\infty}\|\rho_t\|_2 = 0$ and $\lim_{t\to\infty}\frac{\sum_{k=0}^t \mu_k}{t+1} = \lim_{t\to\infty}\frac{\sum_{k=0}^t \|\rho_k\|_2}{t+1} = 0$.*

Lemma 14. *Suppose that Assumptions 2–4 hold and $\alpha_t = \frac{\alpha_0}{t+1}$. When $\mu_t + \tau(\alpha_t)\alpha_{t-\tau(\alpha_t)}\zeta_8 \leq \frac{0.1}{\gamma_{\max}}$ and $\tau(\alpha_t)\alpha_{t-\tau(\alpha_t)} \leq \min\{\frac{\log 2}{A_{\max}}, \frac{0.1}{\zeta_8\gamma_{\max}}\}$, we have for $t \geq \bar{T}$,*

$$\begin{aligned}&\mathbf{E}[\|\langle\tilde{\theta}\rangle_{t+1} - \theta^*\|_2^2] \\ &\leq \frac{\bar{T}}{t+1}\frac{\gamma_{\max}}{\gamma_{\min}}\mathbf{E}[\|\langle\tilde{\theta}\rangle_{\bar{T}} - \theta^*\|_2^2] + \frac{\zeta_9\alpha_0 C\log^2(\frac{t+1}{\alpha_0})}{t+1}\frac{\gamma_{\max}}{\gamma_{\min}} \\ &\quad + \alpha_0\frac{\gamma_{\max}}{\gamma_{\min}}\frac{\sum_{l=\bar{T}}^{t+1}\mu_l}{t+1}, \end{aligned}$$

where $\bar{T}$ is defined in Appendix A.2, ζ_8 and ζ_9 are defined in (A.7) and (A.8), respectively.

We are now in a position to prove Theorem 5.

Proof of Theorem 5. Note that $\sum_{i=1}^N \mathbf{E}[\|\theta_{t+1}^i - \theta^*\|_2^2] \leq 2\sum_{i=1}^N \mathbf{E}[\|\theta_{t+1}^i - \langle\tilde{\theta}\rangle_t\|_2^2] + 2N\mathbf{E}[\|\langle\tilde{\theta}\rangle_t - \theta^*\|_2^2]$. From Lemmas 12 and 14, we have for any $t \geq \bar{T}$,

$$\begin{aligned}&\sum_{i=1}^N \mathbf{E}\left[\|\theta_{t+1}^i - \theta^*\|_2^2\right] \\ &\leq \frac{16}{\epsilon_1}\bar{\epsilon}^t\mathbf{E}[\|\sum_{i=1}^N \tilde{\theta}_0^i + \alpha_0 A(X_0)\theta_0^i + \alpha_0 b^i(X_0)\|_2] \\ &\quad + 2\alpha_t A_{\max}C_\theta + 2\alpha_t b_{\max} + \frac{2\bar{T}N}{t}\frac{\gamma_{\max}}{\gamma_{\min}}\mathbf{E}[\|\langle\tilde{\theta}\rangle_{\bar{T}} - \theta^*\|_2^2] \\ &\quad + \frac{16}{\epsilon_1}\frac{A_{\max}C_\theta + b_{\max}}{1-\bar{\epsilon}}\left(\alpha_0\bar{\epsilon}^{t/2} + \alpha_{\lceil\frac{t}{2}\rceil}\right) \\ &\quad + \frac{2N\zeta_9\alpha_0 C\log^2(\frac{t}{\alpha_0})}{t}\frac{\gamma_{\max}}{\gamma_{\min}} + 2\alpha_0 N\frac{\gamma_{\max}}{\gamma_{\min}}\frac{\sum_{l=\bar{T}}^t \mu_l}{t} \\ &\leq C_7\bar{\epsilon}^t + C_8\left(\alpha_0\bar{\epsilon}^{\frac{t}{2}} + \alpha_{\lceil\frac{t}{2}\rceil}\right) + C_9\alpha_t \\ &\quad + \frac{1}{t}\left(C_{10}\log^2\left(\frac{t}{\alpha_0}\right) + C_{11}\sum_{l=\bar{T}}^t \mu_l + C_{12}\right). \end{aligned}$$

This completes the proof. ■

We next show the asymptotic performance of (9).

Proof of Theorem 4. From Lemma 12, since $\bar{\epsilon} \in (0, 1)$ and $\alpha_t = \frac{\alpha_0}{t}$, it follows that $\lim_{t\to\infty}\|\theta_{t+1}^i - \langle\tilde{\theta}\rangle_t\|_2 = 0$, which implies that all θ_{t+1}^i, $i \in \mathcal{V}$, will reach a consensus with $\langle\tilde{\theta}\rangle_t$. The update of $\langle\tilde{\theta}\rangle_t$ is given in (D.5), which can be treated as a single-agent linear

stochastic approximation whose corresponding ODE is (10). In addition, from Theorem 5 and Lemma 13, $\lim_{\to\infty} \sum_{i=1}^{N} \mathbf{E}[\|\theta_{t+1}^i - \theta^*\|_2^2] = 0$, it follows that θ_{t+1}^i will converge to θ^* in mean square for all $i \in \mathcal{V}$. ■

Remark 11. Finite-time analysis for such a push-based distributed algorithm is challenging. Almost all, if not all, the existing push-based distributed optimization works build on the analysis in Nedić and Olshevsky (2015); however, that analysis assumes that a convex combination of the entire history of the states of each agent (and not merely the current state of the agent) is being calculated. This assumption no longer holds in our case. To obtain a direct finite-time error bound without this assumption, we appeal to a new approach to analyze our push-based SA algorithm by leveraging our consensus-based analyses to establish direct finite-time error bounds for stochastic approximation. Specifically, we tailor an absolute probability sequence for the push-based stochastic approximation algorithm and exploit its properties (Lemma 10). ■

References

Abaid, N., & Porfiri, M. (2010). Consensus over numerosity-constrained random networks. *IEEE Transactions on Automatic Control*, *56*(3), 649–654.

Bhandari, J., Russo, D., & Singal, R. (2018). A finite time analysis of temporal difference learning with linear function approximation. In *Proceedings of the 31st conference on learning theory* (pp. 1691–1692).

Bianchi, P., Fort, G., & Hachem, W. (2013). Performance of a distributed stochastic approximation algorithm. *IEEE Transactions on Information Theory*, *59*(11), 7405–7418.

Blackwell, D. (1945). Finite non-homogeneous chains. *Annals of Mathematics*, *46*(4), 594–599.

Borkar, V. S., & Meyn, S. P. (2000). The ODE method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, *38*(2), 447–469.

Borkar, V. S., & Pattathil, S. (2018). Concentration bounds for two time scale stochastic approximation. In *Proceedings of the 56th annual allerton conference on communication, control, and computing* (pp. 504–511).

Boyd, S., Ghosh, A., Prabhakar, B., & Shah, D. (2006). Randomized gossip algorithms. *IEEE Transactions on Information Theory*, *52*(6), 2508–2530.

Cao, M., Morse, A. S., & Anderson, B. D. O. (2008). Reaching a consensus in a dynamically changing environment: A graphical approach. *SIAM Journal on Control and Optimization*, *47*(2), 575–600.

Chen, S., Devraj, A. M., Bušić, A., & Meyn, S. (2020). Explicit mean-square error bounds for Monte-Carlo and linear stochastic approximation. In *Proceedings of the 23rd international conference on artificial intelligence and statistics* (pp. 4173–4183).

Chen, Z., Maguluri, S. T., Shakkottai, S., & Shanmugam, K. (2020). Finite-sample analysis of contractive stochastic approximation using smooth convex envelopes. *Advances in Neural Information Processing Systems*, *33*.

Chen, J., & Sayed, A. H. (2012). Diffusion adaptation strategies for distributed optimization and learning over networks. *IEEE Transactions on Signal Processing*, *60*(8), 4289–4305.

Dalal, G., Szörényi, B., Thoppe, G., & Mannor, S. (2018). Finite sample analyses for TD(0) with function approximation. In *Proceedings of the 32nd AAAI conference on artificial intelligence* (pp. 6144–6160).

Dalal, G., Thoppe, G., Szörényi, B., & Mannor, S. (2018). Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning. In *Proceedings of the 31st conference on learning theory* (pp. 1199–1233).

Dayan, P. (1992). The convergence of TD(λ) for general λ. *Machine Learning*, *8*(3–4), 341–362.

Doan, T. T., Maguluri, S. T., & Romberg, J. (2019). Finite-time analysis of distributed TD(0) with linear function approximation on multi-agent reinforcement learning. In *Proceedings of the 36th international conference on machine learning* (pp. 1626–1635).

Doan, T. T., Maguluri, S. T., & Romberg, J. (2021). Finite-time performance of distributed temporal-difference learning with linear function approximation. *SIAM Journal on Mathematics of Data Science*, *3*(1), 298–320.

Gupta, H., Srikant, R., & Ying, L. (2019). Finite-time performance bounds and adaptive learning rate selection for two time-scale reinforcement learning. In *Advances in neural information processing systems, vol. 32*.

Huang, M. (2012). Stochastic approximation for consensus: A new approach via ergodic backward products. *IEEE Transactions on Automatic Control*, *57*(12), 2994–3008.

Jadbabaie, A., Lin, J., & Morse, A. S. (2003). Coordination of groups of mobile autonomous agents using nearest neighbor rules. *IEEE Transactions on Automatic Control*, *48*(6), 988–1001.

Kempe, D., Dobra, A., & Gehrke, J. (2003). Gossip-based computation of aggregate information. In *44th IEEE symposium on foundations of computer science* (pp. 482–491).

Kolmogoroff, A. (1936). Zur theorie der markoffschen ketten. *Mathematische Annalen*, *112*(1), 155–160.

Kushner, H. J. (1983). An averaging method for stochastic approximations with discontinuous dynamics, constraints, and state dependent noise. In M. H. Rizvi, J. S. Rustagi, & D. Siegmund (Eds.), *Recent advances in statistics* (pp. 211–235). Academic Press.

Kushner, H. J., & Yin, G. (1987). Asymptotic properties of distributed and communicating stochastic approximation algorithms. *SIAM Journal on Control and Optimization*, *25*(5), 1266–1290.

Kushner, H. J., & Yin, G. G. (1997). *Stochastic approximation algorithms and applications*. Springer, New York.

Lakshminarayanan, C., & Szepesvari, C. (2018). Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *Proceedings of the 21st international conference on artificial intelligence and statistics* (pp. 1347–1355).

LeBlanc, H. J., Zhang, H., Koutsoukos, X., & Sundaram, S. (2013). Resilient asymptotic consensus in robust networks. *IEEE Journal on Selected Areas in Communications*, *31*(4), 766–781.

Lin, Y., Gupta, V., & Liu, J. (2021). Finite-time error bounds for distributed linear stochastic approximation. arXiv preprint arXiv:2111.12665.

Lin, Y., & Liu, J. (2022). Subgradient-push is of the optimal convergence rate. In *Proceedings of the 61st IEEE conference on decision and control* (pp. 5849–5856).

Lin, Y., Zhang, K., Yang, Z., Wang, Z., Başar, T., Sandhu, R., et al. (2019). A communication-efficient multi-agent actor-critic algorithm for distributed reinforcement learning. In *Proceedings of the 58th IEEE conference on decision and control* (pp. 5562–5567).

Liu, J., Mou, S., Morse, A. S., Anderson, B. D. O., & Yu, C. (2011). Deterministic gossiping. *Proceedings of the IEEE*, *99*(9), 1505–1524.

Ma, S., Chen, Z., Zhou, Y., & Zou, S. (2021). Greedy-GQ with variance reduction: Finite-time analysis and improved complexity. In *Proceedings of the 10th international conference on learning representations*.

Ma, S., Zhou, Y., & Zou, S. (2020). Variance-reduced off-policy TDC learning: Non-asymptotic convergence analysis. In *Advances in neural information processing systems, vol. 33* (pp. 14796–14806).

Nedić, A., & Liu, J. (2017). On convergence rate of weighted-averaging dynamics for consensus problems. *IEEE Transactions on Automatic Control*, *62*(2), 766–781.

Nedić, A., & Olshevsky, A. (2015). Distributed optimization over time-varying directed graphs. *IEEE Transactions on Automatic Control*, *60*(3), 601–615.

Nedić, A., Olshevsky, A., Ozdaglar, A., & Tsitsiklis, J. N. (2009). On distributed averaging algorithms and quantization effects. *IEEE Transactions on Automatic Control*, *54*(11), 2506–2517.

Nedić, A., Olshevsky, A., & Rabbat, M. G. (2018). Network topology and communication-computation tradeoffs in decentralized optimization. *Proceedings of the IEEE*, *106*(5), 953–976.

Olfati-Saber, R., Fax, J. A., & Murray, R. M. (2007). Consensus and cooperation in networked multi-agent systems. *Proceedings of the IEEE*, *95*(1), 215–233.

Olshevsky, A., & Tsitsiklis, J. N. (2008). On the nonexistence of quadratic Lyapunov functions for consensus algorithms. *IEEE Transactions on Automatic Control*, *53*(11), 2642–2645.

Olshevsky, A., & Tsitsiklis, J. N. (2013). Degree fluctuations and the convergence time of consensus algorithms. *IEEE Transactions on Automatic Control*, *58*(10), 2626–2631.

Qu, G., & Wierman, A. (2020). Finite-time analysis of asynchronous stochastic approximation and Q-learning. In *Proceedings of the 33rd conference on learning theory* (pp. 3185–3205).

Robbins, H., & Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 400–407.

Rugh, W. J. (1996). *Linear System theory* (2nd ed.). USA: Prentice-Hall, Inc..

Srikant, R., & Ying, L. (2019). Finite-time error bounds for linear stochastic approximation and TD learning. In *Proceedings of the 32nd conference on learning theory* (pp. 2803–2830).

Stanković, M. S., Ilić, N., & Stanković, S. S. (2016). Distributed stochastic approximation: Weak convergence and network design. *IEEE Transactions on Automatic Control*, *61*(12), 4069–4074.

Stanković, M. S., & Stanković, S. S. (2016). Multi-agent temporal-difference learning with linear function approximation: Weak convergence under time-varying network topologies. In *Proceedings of the 2006 American control conference* (pp. 167–172).

Stanković, S. S., Stanković, M. S., & Stipanović, D. (2010). Decentralized parameter estimation by consensus based stochastic approximation. *IEEE Transactions on Automatic Control*, *56*(3), 531–543.

Sun, J., Wang, G., Giannakis, G. B., Yang, Q., & Yang, Z. (2020). Finite-time analysis of decentralized temporal-difference learning with linear function approximation. In *Proceedings of the 23rd international conference on artificial intelligence and statistics* (pp. 4485–4495).
Suttle, W., Yang, Z., Zhang, K., Wang, Z., Başar, T., & Liu, J. (2020). A multi-agent off-policy actor-critic algorithm for distributed reinforcement learning. In *Proceedings of the 21st IFAC world congress*.
Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: an introduction*. MIT Press.
Touri, B. (2012). *Product of random stochastic matrices and distributed averaging*. Springer Science & Business Media.
Tsitsiklis, J. N., & Van Roy, B. (1997). An analysis of temporal-difference learning with function approximation. *IEEE Transactions on Automatic Control*, *42*(5), 674–690.
Vaidya, N. H., Tseng, L., & Liang, G. (2012). Iterative approximate Byzantine consensus in arbitrary directed graphs. In *Proceedings of the 2012 ACM symposium on principles of distributed computing* (pp. 365–374).
Wang, Y., Chen, W., Liu, Y., Ma, Z., & Liu, T. (2017). Finite sample analysis of the GTD policy evaluation algorithms in Markov setting. In *Proceedings of the 31st conference on neural information processing systems* (pp. 5504–5513).
Wang, G., Li, B., & Giannakis, G. B. (2019). A multistep Lyapunov approach for finite-time analysis of biased stochastic approximation. arXiv:1909.04299.
Wang, G., Lu, S., Giannakis, G. B., Tesauro, G., & Sun, J. (2020). Decentralized TD tracking with linear function approximation and its finite-time analysis. In *Advances in neural information processing systems, vol. 33* (pp. 13762–13772).
Wang, Y., & Zou, S. (2020). Finite-sample analysis of greedy-GQ with linear function approximation under Markovian noise. In *Proceedings of the 36th conference on uncertainty in artificial intelligence* (pp. 11–20).
Weng, W., Gupta, H., He, N., Ying, L., & Srikant, R. (2020). The mean-squared error of double Q-learning. In *Advances in neural information processing systems, vol. 33* (pp. 6815–6826).
Wu, Y., Zhang, W., Xu, P., & Gu, Q. (2020). A finite time analysis of two time-scale actor critic methods. In *Proceedings of the 34th conference on neural information processing systems*.
Xiao, L., Boyd, S., & Lall, S. (2005). A scheme for robust distributed sensor fusion based on average consensus. In *Proceedings of the 4th international conference on information processing in sensor networks* (pp. 63–70).
Xu, P., & Gu, Q. (2020). A finite-time analysis of Q-learning with neural network function approximation. In *Proceedings of the 37th international conference on machine learning* (pp. 10555–10565).
Xu, T., Zou, S., & Liang, Y. (2019). Two time-scale off-policy TD learning: Non-asymptotic analysis over Markovian samples. In *Advances in neural information processing systems, vol. 32*.
Zeng, S., Doan, T. T., & Romberg, J. (2020). Finite-time analysis of decentralized stochastic approximation with applications in multi-agent and multi-task learning. arXiv:2010.15088.
Zhang, K., Yang, Z., & Başar, T. (2018). Networked multi-agent reinforcement learning in continuous spaces. In *Proceedings of the 57th IEEE conference on decision and control* (pp. 2771–2776).
Zhang, K., Yang, Z., & Başar, T. (2021). Multi-agent reinforcement learning: A selective overview of theories and algorithms. In K. G. Vamvoudakis, Y. Wan, F. L. Lewis, & D. Cansever (Eds.), *Handbook of reinforcement learning and control. studies in systems, decision and control, vol. 325*. Springer, Cham.
Zhang, K., Yang, Z., Liu, H., Zhang, T., & Başar, T. (2018). Fully decentralized multi-agent reinforcement learning with networked agents. In *Proceedings of the 35th international conference on machine learning* (pp. 5872–5881).
Zhang, K., Yang, Z., Liu, H., Zhang, T., & Başar, T. (2021). Finite-sample analysis for decentralized batch multi-agent reinforcement learning with networked agents. *IEEE Transactions on Automatic Control*.
Zhang, Y., & Zavlanos, M. M. (2019). Distributed off-policy actor-critic reinforcement learning with policy consensus. In *Proceedings of the 58th IEEE conference on decision and control* (pp. 4674–4679).
Zou, S., Xu, T., & Liang, Y. (2019). Finite-sample analysis for SARSA with linear function approximation. In *Advances in neural information processing systems, vol. 32*.

Yixuan Lin received the B.S. degree in Mathematics from Fudan University, Shanghai, China, in 2017, and the M.S degree in Applied Mathematics and Statistics from Stony Brook University, Stony Brook, NY, USA, in 2020. She is currently pursuing her Ph.D. degree in Applied Mathematics and Statistics at Stony Brook University. Her research interests include reinforcement learning, optimization, and federated learning.

Vijay Gupta is in the Elmore Family School of Electrical and Computer Engineering at the Purdue University. He received his B. Tech degree at Indian Institute of Technology, Delhi, and his M.S. and Ph.D. at California Institute of Technology, all in Electrical Engineering. He received the 2018 Antonio J Rubert Award from the IEEE Control Systems Society, the 2013 Donald P. Eckman Award from the American Automatic Control Council and a 2009 National Science Foundation (NSF) CAREER Award. His research interests are broadly at the interface of communication, control, distributed computation, and human decision making.

Ji Liu received the Ph.D. degree from Yale University, New Haven, CT, USA. He is currently an Assistant Professor in the Department of Electrical and Computer Engineering at Stony Brook University, Stony Brook, NY, USA. His current research interests include distributed control and optimization, distributed reinforcement learning, and resiliency of distributed algorithms.