This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2025.3555941

Finite-Time Analysis of Over-the-Air Federated TD Learning

Nicolò Dal Fabbro, Aritra Mitra, Robert W. Heath Jr., Luca Schenato, George J. Pappas

***Abstract*—In recent years, federated learning has been widely studied to speed up various *supervised* learning tasks at the wireless network edge. However, there is a lack of theoretical understanding as to whether similar speedups in sample complexity can be achieved for cooperative reinforcement learning (RL) problems subject to communication constraints. To that end, we study a federated policy evaluation problem over wireless fading channels where, to update model parameters, a central server aggregates local temporal difference (TD) update directions from N agents via analog over-the-air computation (OAC). We refer to this scheme as `OAC-FedTD` and provide a rigorous finite-time convergence analysis of its performance. Our analysis reveals the impact of the noisy fading channels on the convergence rate and establishes a linear convergence speedup w.r.t. the number of agents. Notably, this is the first non-asymptotic analysis of a cooperative RL setting under wireless channels that jointly considers linear value function approximation, Markovian sampling, and the OAC channel-induced distortions and noise. Our work develops the theoretical foundations that are key for relevant advancements in the analysis and design of federated reinforcement learning algorithms over wireless networks.**

***Index Terms*—Reinforcement Learning, over-the-air computation, federated learning, finite-time convergence analysis.**

I. Introduction

In recent years, there has been an increasing demand for edge intelligence, where distributed machine learning algorithms are implemented at the wireless network edge on diverse datasets [1], [2]. One particularly popular instance of edge computing is the federated learning (FL) framework [3], where agents periodically coordinate with a central server to *speed up* the process of model-training for *supervised learning tasks*, while keeping raw training data private. The resulting FL problem boils down to stochastic optimization subject to communication constraints: unpredictable and heterogeneous wireless channels [4] that have limited bandwidth [5] and suffer from channel fading effects [6]. In this context, the effect of imperfect channels on the convergence of the underlying optimization procedure is a topic that has been extensively studied in FL [7], [8]. Departing from this line of work on supervised learning/optimization, in this paper, we ask: *Is it possible to achieve speedups in sample-complexity for cooperative reinforcement learning (RL) problems when communication takes place via over-the-air computation (OAC) over noisy wireless channels?*[1] Surprisingly, there is no theoretical understanding of this question, even for the simplest of RL problems, such as policy evaluation. We seek to bridge this gap with our work.

In a standard RL setup, an agent repeatedly interacts with an environment by playing actions based on some policy, receiving observations (rewards), and then improving the policy with the aim of maximizing long-term cumulative returns. This sequential decision-making problem is captured via a Markov Decision Process (MDP), where the agent's goal is to find the optimal policy without knowledge of the transition kernels and reward functions of the MDP. As such, decision-making in RL is based on *sequential* data in the form of observations of state transitions and rewards.

For MDPs with large state and action spaces, RL algorithms typically require many data samples to achieve a desired level of accuracy [9]. In this regard, the emerging paradigm of federated reinforcement learning (FRL) [10], [11], [12], [13], [14], that has recently gained momentum in wireless networks applications [15], [16], [17], seeks to reduce sample-complexity requirements by parallelizing data collection and computation across multiple agents. In FRL, agents share local model parameters or model differentials (i.e., gradient-like update directions) with a central server while keeping their states, rewards, and actions private. However, similar to FL, achieving a *convergence speedup* w.r.t the number of agents in FRL at the wireless network edge requires frequent information exchanges between the agents and the server over a *shared wireless communication medium*. This leads to an interesting tension: *more agents imply more data, and hence, the hope of a greater convergence speedup; however, more agents also imply a larger communication bottleneck.*

In this work, we provide the first principled understanding of this tension in FRL by considering a setting in which N agents cooperate via OAC to solve a policy evaluation problem. We focus on policy evaluation since it is at the heart of value iteration algorithms like Watkin's Q-learning algorithm [18] for finding the optimal policy. In our setup, each agent interacts with the *same* MDP, collects observations by playing the policy to be evaluated, and employs temporal difference (TD) learning [19] with linear function approximation to construct a local TD(0) update direction. The agents' local update directions are then transmitted to a central server over a shared

N. Dal Fabbro (ndf96@seas.upenn.edu) and G. J. Pappas (pappasg@seas.upenn.edu) are with the Electrical and Systems Engineering Department, University of Pennsylvania, USA. A. Mitra (amitra2@ncsu.edu) is with the Electrical and Computer Engineering Department, North Carolina State University, USA. Robert W. Heath Jr. is with the Department of Electrical and Computer Engineering at the University of California, San Diego, USA. Luca Schenato (schenato@dei.unipd.it) is with the department of Information Engineering, University of Padova, Italy. This work was supported by NSF Award 1837253, and the Italian Ministry of Education, University and Research through the PRIN project no. 2017NS9FEY.

[1]We describe the OAC mechanism in detail in Section II.

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 13:05:04 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2025.3555941

wireless channel medium for model updating. To alleviate the communication bottleneck, we consider analog over-the-air computation (OAC) that has recently been advocated to provide large-scale, bandwidth- and energy-efficient up-link communication in FL [20], [21]. In particular, OAC exploits the waveform-superposition property of the wireless multiple access channel (MAC) to enable the receiver (server) to obtain the average of the analog signals transmitted by the agents over the same time-frequency block [22]. Examples of application of this considered OAC FRL setting are in internet of things and internet of vehicles scenarios, where, equipped with onboard intelligence and wireless communication capabilities, intelligent agents (e.g., robots and autonomous vehicles) will be continually learning and evaluating new machine learning algorithms and RL policies [12], [23]. Compared to standard digital transmission, OAC comes with notable gains in up-link bandwidth efficiency. Furthermore, OAC has intrinsic privacy-preserving features [6], [24]. However, analog signals transmitted over the air are subject to fading channel distortion and additive noise at the receiver [6], [8], [7]. This leads to the main question we investigate in this paper: *Is it still possible to achieve collaborative performance gains for federated policy evaluation under such channel distortion and noise?*

Related work. In this work, we provide an answer in the affirmative. Before elaborating on our contribution in this regard, we first survey the most relevant strands of literature.

• *Distributed Learning over Noisy Channels.* A vast body of work has investigated distributed optimization/supervised learning problems over imperfect, noisy channels [25], [2], [26]. In particular, OAC has been considered in, e.g., [6], [8], [7]. While we draw inspiration from these papers to formulate our problem, deriving finite-time convergence rates for our multi-agent RL (MARL) setting is much more challenging since (i) policy evaluation is *not a static optimization problem*; and (ii) unlike the i.i.d. sampling assumption prevalent in supervised learning, the data samples in RL are part of a single Markov chain, and hence, *temporally correlated.*

• *Single-Agent RL.* Even for the single-agent case, while there is a rich theory characterizing the asymptotic performance of RL algorithms [27], [28], finite-time rates of convergence under Markovian sampling have been established only recently [29], [30], [31], [32]. We note that a finer analysis relative to these papers is needed to establish the benefits of collaboration in a MARL setting like ours.

• *Collaborative MARL/FRL.* There is a growing interest in understanding the performance of collaborative MARL algorithms. The initial results in this space work under a restrictive i.i.d. sampling assumption [33], [34]. Recently, linear speedups in sample complexity under Markovian sampling have been established in federated reinforcement learning (FRL); see [11], [35], [36]. Importantly, [11] and [35] do not consider any effects of noisy wireless channels. In contrast, the OAC mechanism we study in this paper introduces additional forms of randomness that necessitate a very careful analysis. While [36] consider the effects of quantization and packet drops, the channel effects we study in this paper are quite different and, as such, need a separate treatment.[2]

In light of the above, our contributions are the following.

Contributions. First, we formulate and study federated policy evaluation under the analog OAC model - a setting that we refer to as `OAC-FedTD`. Ours is the first work to formally study the convergence behavior of a cooperative RL algorithm subject to the channel distortions and noise introduced by OAC, while jointly accounting for linear value function approximation and Markovian sampling.

Second, as our main technical contribution, we provide a comprehensive non-asymptotic convergence analysis of `OAC-FedTD`. In Theorem 1, we prove that the sample-complexity bounds for `OAC-FedTD` exhibit an N-fold linear speedup relative to the vanilla single-agent TD algorithm. In particular, our bounds reveal that increasing the number of agents helps "drown out" both the effect of the channel noise and of the MDP sampling variance. Since RL algorithms are typically data-hungry, our linear speedup result is of considerable theoretical and practical significance.

Notation. We denote vectors and matrices as lower and upper case bold letters, respectively (e.g., vector $\mathbf{a}$ and matrix $\mathbf{A}$). For a positive integer i, $\mathbf{a}(i)$ denotes the i-th element of $\mathbf{a}$, and $\mathbf{A}^\top$ denotes matrix $\mathbf{A}$'s transpose. $\mathbf{I}_d$ denotes a d-dimensional identity matrix, $\mathbf{0}$ a vector of zeros and $\mathbb{E}[.]$ the expectation operator. We denote by $\langle \mathbf{a}, \mathbf{b} \rangle$ the inner product of two vectors $\mathbf{a}$ and $\mathbf{b}$, and by $\|\cdot\|$ the Euclidean norm.

II. System Model and `OAC-FedTD`

In this paper, we examine a framework in which N agents independently interact with replicas of identical Markov Decision Processes (MDPs). We denote an MDP by $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \gamma)$. The finite set $\mathcal{S}$ comprises the n states of the MDP, while $\mathcal{A}$ is the finite set of actions. We denote by $\mathcal{P}$ the set of Markov transition kernels, which are functions of the actions. We let $\mathcal{R}$ be the reward function. The constant $\gamma \in (0,1)$ is the discount factor. The goal of the federation of agents is to collectively evaluate the value function associated with a policy $\mu : \mathcal{S} \to \mathcal{A}$. To do so, the agents communicate with a central aggregator (server) via up-link analog transmission. In what follows, we first briefly illustrate some necessary background on policy evaluation with function approximation. Then, we formally describe the considered OAC framework, with related objectives and technical challenges.

Policy evaluation with linear function approximation. The purpose of the considered *policy evaluation* setting is to evaluate, for each state $s \in \mathcal{S}$, the value function $\mathbf{u}_\mu(s)$, where $\mathbf{u}_\mu$ is a vector of size n storing the value function for each state. The value $\mathbf{u}_\mu(s)$ is the discounted expected cumulative reward obtained by playing policy μ starting from initial state s. The policy μ interacts with the MDP $\mathcal{M}$ to generate a Markov Reward Process (MRP) characterized by a reward function $R_\mu : \mathcal{S} \to \mathbb{R}$, and a Markov chain with transition probability matrix $\mathbf{P}_\mu$. Mathematically, we have

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

[2]For distributed optimization, the effects of quantization, erasures, and OAC have been explored separately in several papers, e.g., [37], [25], [6], [38], [7]

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 13:05:04 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2025.3555941

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

As is common, we consider a linear value function approximation setting in which the n-dimensional value function $\mathbf{u}_\mu$ is approximated by vectors in a linear subspace of $\mathbb{R}^n$ spanned by a set of d basis vectors $\{\boldsymbol{\phi}_\ell\}_{\ell=1}^d$. Resorting to value function approximation is particularly needed when the state space $\mathcal{S}$ is very large, and indeed usually $d \ll n$. Let $\boldsymbol{\Phi} \triangleq [\boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_d] \in \mathbb{R}^{n\times d}$ be the feature matrix. As is standard [29], we assume that the columns of $\boldsymbol{\Phi}$ are independent and that the rows are normalized. The parametric approximation $\hat{\mathbf{u}}_{\boldsymbol{\theta}}$ of $\mathbf{u}_\mu$ is given by $\mathbf{u}(\boldsymbol{\theta}) := \hat{\mathbf{u}}_{\boldsymbol{\theta}} = \boldsymbol{\Phi}\boldsymbol{\theta}$, where $\boldsymbol{\theta} \in \mathbb{R}^m$ is the approximation parameter. Denoting the s-th row of $\boldsymbol{\Phi}$ by $\boldsymbol{\phi}'_s$, the approximation of $\mathbf{u}_\mu(s)$ is given by $\hat{\mathbf{u}}_{\boldsymbol{\theta}}(s) = \langle \boldsymbol{\theta}, \boldsymbol{\phi}'_s \rangle$.

The goal of the federation of agents is to obtain the best possible estimate of $\boldsymbol{\theta}^*$, which represents the best linear approximation of $\mathbf{u}_\mu$ in the span of $\boldsymbol{\Phi}$. For this purpose, we consider a multi-agent version of the well-known TD(0) algorithm [19]. In this setting, agents $i = 1, ..., N$ are initialized from some initial state $s_{i,0} \in \mathcal{S}$ with an initial estimate $\boldsymbol{\theta}_0 \in \mathbb{R}^m$. Next, at each time-step $k \in \mathbb{N}$, the server broadcasts a global parameter vector $\boldsymbol{\theta}_k$ to all agents. Each agent, at that point, takes an action $a_{i,k} = \mu(s_{i,k})$, and observes the next state $s_{i,k+1} \sim \mathbf{P}_\mu(\cdot|s_{i,k})$ and instantaneous reward $r_{i,k} = R_\mu(s_{i,k})$. Note that $s_{i,k}$ is the state of agent i at iteration k. Using the parameter vector $\boldsymbol{\theta}_k$ and the observed tuple $o_{i,k} = (s_{i,k}, r_{i,k}, s_{i,k+1})$, agent i then computes a local TD update direction as follows:

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

In the rest of the paper, we will often write $\mathbf{g}_{i,k}(\boldsymbol{\theta}_k)$ instead of $\mathbf{g}_{i,k}(\boldsymbol{\theta}_k, o_{i,k})$. Note that although all agents are playing and evaluating the same policy μ, while interacting with replicas of identical MDPs, the local observation processes $\{o_{i,k}\}$, in general, differ across different agents, and we assume that these observation processes are *statistically independent* across agents. Note that, however, for each agent i, the observations are *temporally correlated* because they are all part of a single Markov chain. Given that the observation processes are independent across agents, one hopes that communicating and averaging agents' local TD update directions will reduce the variance in the estimation of $\boldsymbol{\theta}^*$. In the wireless FRL framework that we consider in this paper, the exchange of local TD update directions occurs based on *over-the-air* aggregation. In what follows, we describe this scheme in detail.

Over-the-air computation model. We consider the typical OAC channel model that has been adopted, for example, in [22], [6], [24], [7]. In this scheme, N agents, coordinated by a central entity, synchronously transmit their local update directions as analog wireless signals. The central entity then collects the superposition of these signals; hence, the term 'over-the-air.' The analog signals are subject to fading channel distortion and to additive white Gaussian noise at the receiver.

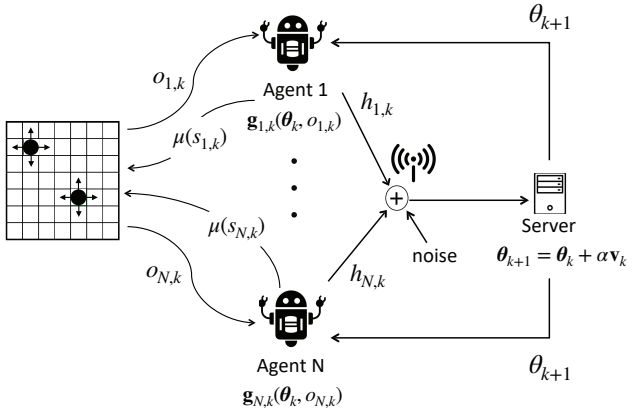


Fig. 1. Illustration of the OAC-FedTD scheme. For a description of the various symbols, see Table I.

Under the assumptions of synchronization and phase compensation [6], [8], [22], the server at iteration k obtains the following *noisy* and *distorted* global TD direction:

$$\mathbf{v}_k = \frac{1}{N}\sum_{i=1}^{N} h_{i,k}\mathbf{g}_{i,k}(\boldsymbol{\theta}_k) + \mathbf{w}_k, \tag{2}$$

where $\mathbf{w}_k \sim \mathcal{N}(\mathbf{0}, \sigma^2_{\mathbf{w}}\mathbf{I}_d)$ and $\sigma^2_{\mathbf{w}} = \tilde{\sigma}^2_{\mathbf{w}}/N^2$, where $\tilde{\sigma}^2_{\mathbf{w}}$ is the additive white noise variance at the receiver. The distortion term $h_{i,k}$ is the *random* channel gain experienced by agent i at iteration k, with mean m_h and variance σ^2_h. We make the standard assumption that the random channel gain process is independent across agents and iterations. We will also assume that the random processes $\{\mathbf{w}_k\}$ and $\{h_{i,k}\}$ related to the channel effects are independent of the Markovian data tuples $\{o_{i,k}\}$. The model in (2) captures different settings of OAC. For example, the model adopted in [22] considers transmitters with adaptive power transmission. In that case, $h_{i,k} = c_{i,k}\sqrt{p_{i,k}}$, where $c_{i,k}$ is the actual channel gain, and $\sqrt{p_{i,k}}$ is the power scaling factor of device i that can be adaptively adjusted to reduce the impact of the channel gain. Due to channel estimation errors [39], even when channel inversion is performed, $h_{i,k}$ is typically a random object. In general, the model considered in this paper captures any OAC framework with phase compensation, as long as the distortion $h_{i,k}$ in (2) admits first and second moments.

Once the server receives $\mathbf{v}_k$, it updates the estimate of the parameter $\boldsymbol{\theta}_k$ according to the following update rule:

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \alpha\mathbf{v}_k, \tag{3}$$

where α is a constant step-size/learning rate, and $\mathbf{v}_k$ is as in (2). We refer to this specific federated TD learning framework as over-the-air federated TD learning (OAC-FedTD).

Objective and challenges. In the rest of the paper, we aim to study the *finite-time* convergence performance of OAC-FedTD. This presents multiple challenges. Indeed, even in a single-agent setting, the non-asymptotic analysis of TD(0) is very demanding due to temporal correlations in the Markov process. To analyze OAC-FedTD, we deal with a multi-agent setting where two different sources of randomness are

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 13:05:04 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2025.3555941

concurrently in place: (i) the randomness due to the time-correlated agents' trajectories, and (ii) the randomness in the up-link communication due to the wireless fading channels. Furthermore, the final objective of OAC−FedTD is to provide a *linear convergence speedup* w.r.t. the number of participants in the federation. This requires a novel and elaborate analysis that we report in the sequel.

III. Main Result

We now present our main contribution, which is the study of the non-asymptotic convergence rate of OAC−FedTD. We start introducing technical machinery that we will use in the remainder of the section. We make the standard assumption that rewards are uniformly bounded. Therefore, there exists a universal constant $\bar{r} > 0$ such that for every state $s \in \mathcal{S}$, $R_\mu(s) \leq \bar{r}$. Thanks to this assumption, we are guaranteed that the value function in (1) is well-defined. We make the following assumption that is pivotal in most finite-time analyses of TD learning and stochastic approximation algorithms with Markovian sampling [29], [30], [27].

Assumption 1. *The Markov chain induced by the policy μ is aperiodic and irreducible.*

A key consequence of the above assumption is that the Markov chain induced by the policy μ admits a unique stationary distribution π [40]. Define the matrix $\mathbf{\Sigma} = \mathbf{\Phi}^\top \mathbf{D} \mathbf{\Phi}$, where $\mathbf{D}$ is a diagonal matrix whose diagonal entries are the probabilities of π. Note that because $\mathbf{\Phi}$ is full column rank, $\mathbf{\Sigma}$ is full rank with smallest eigenvalue $0 < \omega < 1$, which is a key object that will appear in our non-asymptotic rates. Next, we define a mathematical object that plays a crucial role in the analysis. This object is also referred to as the steady-state TD direction, and it is defined as follows:

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

The virtual *deterministic* update rule $\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \alpha \bar{\mathbf{g}}(\boldsymbol{\theta}_k)$ converges with an exponential rate to $\boldsymbol{\theta}^*$ (see [29]), with $\boldsymbol{\theta}^*$ being the unique solution of the projected Bellman equation $\Pi_{\mathbf{D}} \mathcal{T}_\mu(\mathbf{\Phi}\boldsymbol{\theta}^*) = \mathbf{\Phi}\boldsymbol{\theta}^*$. Note that $\Pi_{\mathbf{D}}(\cdot)$ is the projection operator onto the subspace spanned by $\{\boldsymbol{\phi}_\ell\}_{\ell \in [d]}$ with respect to the inner product $\langle \cdot, \cdot \rangle_{\mathbf{D}}$, and $\mathcal{T}_\mu : \mathbb{R}^n \to \mathbb{R}^n$ is the policy-specific Bellman operator [27]. Equipped with the above definition of steady-state TD update direction $\bar{\mathbf{g}}(\boldsymbol{\theta})$, the analysis of OAC−FedTD relies on the notion of mixing time of the Markov chain, that we define as follows.

Definition 1. *Define* $\tau_\epsilon \triangleq \min\{t \geq 1 : \|\mathbb{E}[\mathbf{g}_{i,k}(\boldsymbol{\theta}, o_{i,k})|o_{i,0}] - \bar{\mathbf{g}}(\boldsymbol{\theta})\| \leq \epsilon(\|\boldsymbol{\theta}\| + 1), \forall k \geq t, \forall \boldsymbol{\theta} \in \mathbb{R}^m, \forall i \in N, \forall o_{i,0}\}$.

Note that Assumption 1 implies that the Markov chain induced by μ mixes exponentially fast [41], i.e., there exists some $K \geq 1$, such that $\tau_\epsilon \leq K \log(\frac{1}{\epsilon})$. To establish our theoretical results, in the sequel, we set $\epsilon = \alpha^2$. To lighten the notation, in the remainder of the paper, we omit the subscript of ϵ in τ_ϵ, and call the integer τ the *mixing time*. We define the optimality gap $\delta_k^2 \triangleq \|\boldsymbol{\theta}^* - \boldsymbol{\theta}_k\|^2$, and the quantity $p_h \triangleq \max\{1, m_h^2 + \sigma_h^2\}$ that is directly related to the statistical power of the channel distortion. Furthermore,

TABLE I
List of notations

Notation	Description
$s_{i,k}, \quad r_{i,k}$	state and reward of agent i at iteration k
$o_{i,k} = (s_{i,k}, s_{i,k+1}, r_{i,k})$	observation tuple of agent i at iteration k
$\boldsymbol{\theta} \in \mathbb{R}^d$	value function approximation parameter
$\boldsymbol{\theta}^* \in \mathbb{R}^d$	best approximation parameter
$\delta_k = \|\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\|$	optimality gap at iteration k
$N \in \mathbb{N}$	number of agents
$\mathbf{g}_{i,k}(\boldsymbol{\theta}_k) = \mathbf{g}_{i,k}(\boldsymbol{\theta}_k, o_{i,k}) \in \mathbb{R}^d$	TD update direction of agent i at iteration k
$\bar{\mathbf{g}}(\boldsymbol{\theta}_k) \in \mathbb{R}^d$	steady-state TD direction at iteration k
$\mathbf{\Phi} \in \mathbb{R}^{n \times d}$	feature matrix
$\gamma \in (0, 1)$	discount factor
$\bar{r} > 0$	upper bound on instantaneous reward
$\pi \in \mathbb{R}^n$	stationary distribution of the Markov chain
$\mu : \mathcal{S} \to \mathcal{A}$	policy to be evaluated by OAC−FedTD
$\|\cdot\|_{\mathbf{D}}$	norm induced by $\mathbf{D} = \text{diag}(\pi) \in \mathbb{R}^{n \times n}$
$\omega \in \mathbb{R}$	smallest eigenvalue of $\mathbf{\Sigma} = \mathbf{\Phi}\mathbf{D}\mathbf{\Phi}^\top$
$h_{i,k} \in \mathbb{R}$	channel distortion for agent i at iteration k
$\mathbf{w}_k \in \mathbb{R}^d$	measurement noise at the receiver
$m_h = \mathbb{E}[h_{i,k}]$	mean of channel distortion
$\sigma_h^2 = \text{var}(h_{i,k})$	variance of channel distortion
$\sigma_{\mathbf{w}}^2 \mathbf{I} = \text{Var}(\mathbf{w}_k) \in \mathbb{R}^{d \times d}$	variance of measurement noise
$\alpha > 0$	step-size/learning rate
$\mathbf{v}_k \in \mathbb{R}^d$	global update direction of OAC−FedTD
$\tau = \tau_\epsilon \in \mathbb{N}$	mixing time of the Markov chain
$p_h = \max\{1, m_h^2 + \sigma_h^2\}$	statistical power of channel distortion
$\sigma = \max\{1, \bar{r}, \|\boldsymbol{\theta}^*\|, \delta_0\}$	variance of observation process

we define the "variance" of the observation model as $\sigma \triangleq \max\{1, \bar{r}, \|\boldsymbol{\theta}^*\|, \delta_0\}$.

We now present the main result of the paper, which is the first finite-time convergence result in RL with OAC. Notably, we consider the challenging case in which agents' trajectories follow a Markov process, and show that cooperation between agents provides a linear convergence speedup even under noisy analog communication over wireless fading channels. Our result highlights OAC noise effects on the convergence rate.

Theorem 1. *Consider the update rule of OAC−FedTD in (3). There exists a universal constant $C_0 \geq 1$, such that with $\alpha \leq \frac{m_h \omega (1-\gamma)}{C_0 \tau p_h}$, the following holds for $T \geq 2\tau$:*

$$\mathbb{E}\left[\delta_T^2\right] \leq (1 - m_h \alpha \omega (1-\gamma))^T C_1 + \frac{C_2 p_h \alpha \tau \sigma^2}{m_h \omega (1-\gamma) N} + \frac{C_3 p_h \tau \sigma^2 \alpha^3}{m_h \omega (1-\gamma)} + \frac{C_4 \alpha \tau \tilde{\sigma}_{\mathbf{w}}^2 d}{m_h \omega (1-\gamma) N^2}, \quad (5)$$

where $C_1 = 6\sigma^2 + 2\frac{\tilde{\sigma}_{\mathbf{w}}^2 d}{N^2}$, and C_2, C_3, C_4 are universal constants.

A detailed proof of Theorem 1 is provided in Section IV, where we outline the key technical challenges relative to the centralized analysis in [30].

Discussion: We now discuss the main takeaways from Theorem 1. From (5), we first note that OAC−FedTD converges linearly (in the mean-square sense) to a ball around $\boldsymbol{\theta}^*$ whose radius depends on the second, third and fourth terms in (5). Compared to the case without noise ($m_h = 1$, $\sigma_h = 0$,

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 13:05:04 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2025.3555941

$\tilde{\sigma}_{\mathbf{w}} = 0$), the linear convergence rate worsens due to both the mean distortion m_h, and due to the choice of the step size, which needs to scale inversely with $p_h\tau$. The term p_h also inflates the dominant "variance term", namely the second term in (5). So, given that $\mathbb{E}\left[h_{i,k}^2\right] = m_h^2 + \sigma_h^2$ and recalling that $p_h = \max\{1, m_h^2 + \sigma_h^2\}$, our bound clearly reveals the effect of fading distortion on the rate. This channel effect is consistent with what one observes for analogous settings in FL with OAC [6]. Compared with the effect of noise in FL via OAC, we note that the variance term related to the additive noise at the receiver, i.e., the fourth term in (5), gets scaled by the mixing time τ. Note that the second, third, and fourth terms in (5) have a tight dependence on the mixing time τ. Indeed, we achieve a linear dependence of $O(\tau)$, which is consistent with the centralized setting [29], [30], and has been shown to be minimax optimal [42]. Next, compared to the centralized setting [30, Theorem 7], observe that the second and fourth terms in (5) get scaled down by a factor of N. Moreover, the third term is $O(\alpha^3)$, i.e., it is a higher-order term that is dominated by the second term for small enough α. *Thus, ours is the first work considering MARL/FRL via OAC over wireless fading channels to establish a variance-reduction effect w.r.t. the number of agents*. Notably, the only requirement in terms of assumptions on the channel model is that the fading and measurement noise distribution admit first and second moment. In this sense, our analysis captures the most widely adopted fading models, e.g., Rayleigh, Nakagami and Rician fading [43], [44]. With $\alpha = O(\log(NT)/T)$, we can explicitly show that each of the four terms in (5) is $O(1/NT)$, yielding the linear speedup effect we had hoped for. We explicitly illustrate this result in the following corollary.

Corollary 1. *Consider the update rule of* `OAC-FedTD` *in* (3). *Let the step-size α and the number of iterations T be chosen to satisfy:*

$$\alpha = \frac{\log NT}{\omega(1-\gamma)m_h T}, \;\; \textit{and} \;\; T \geq \frac{2C_0 N\tau \log NT}{\omega(1-\gamma)m_h}, \tag{6}$$

where C_0 is as in Theorem 1. We then have the following bound:

$$\mathbb{E}[\delta_T^2] \leq O\left(\frac{\log(NT)\tau}{m_h^2\omega^2(1-\gamma)^2 NT}\right)\left(\sigma^2 p_h + \frac{\tilde{\sigma}_{\mathbf{w}}^2 d}{N}\right). \tag{7}$$

The above Corollary illustrates explicitly the *sample efficiency* of `OAC-FedTD`, which is established by the presence of NT at the denominator of the first term between brackets. To better illustrate this fact, let us compare it with the centralized case, which was studied for example in [29]. Under Markovian sampling, part (c) of Theorem 3 in [29] establishes that the mean-square error for single-agent TD decays at the following rate:

$$O\left(\frac{G^2\tau\log(T)}{\omega^2(1-\gamma)^2 T}\right), \tag{8}$$

where G, as defined in [29], captures the effect of both the projection radius (in [29], the authors consider a projected version of TD learning) and the noise variance. The term G^2 can be viewed as the analog of σ^2 in our bound. Comparing the above bound with that in (7), we make two immediate observations. First, the term T in the centralized bound gets replaced by NT in our bound. This is precisely what we wanted since in our setting, each agent has access to T samples, yielding a total of NT samples. Essentially, this goes on to show that our algorithm is *sample-efficient* in that it makes use of all the samples from all the agents and achieves a linear speedup w.r.t. the number of agents. At the same time, we can clearly observe the effect of the channel distortion and measurement noise on the rate: note how the variance term σ^2 gets inflated by the statistical power of the distortion p_h and how the variance measurement noise at the receiver $\tilde{\sigma}_{\mathbf{w}}^2$, multiplied by d/N, adds up to the term $p_h\sigma^2$. It is noteworthy that our bound recovers precisely the tight centralized bound [29], [30] in the absence of distortion and noise, namely when $\sigma_{\mathbf{w}}^2 = 0$ and $p_h = m_h = 1$. The proof of the above Corollary follows with some simple algebraic calculations by choosing the number of iterations T and the step size α in (6), and it is omitted. In the next Section, we provide the complete proof of Theorem 1, which is the key result and main technical contribution of this paper.

Remark 2. *Recent work has shown that for certain nonlinear stochastic approximation schemes that subsume variants of Q-learning, finite-time results for TD learning carry over seamlessly to the nonlinear setting, provided the underlying operator satisfies Lipschitzness and contractive properties [32], [45]. As such, while our focus here is on TD learning with linear function approximation, we anticipate that much of our analysis can be extended to nonlinear stochastic approximation schemes with suitable modifications of our techniques.*

IV. Outline of the Proof of the Main Result

Here, we provide a detailed outline of the proof of Theorem 1. We start by introducing some definitions and preliminary results. Then, we show the main steps of the proof, which require some intermediate results, which we present in the form of *auxiliary lemmas*. For convenience, we define, $\forall k \geq \tau, \forall \boldsymbol{\theta} \in \mathbb{R}^d, \forall i \in [N]$,

$$\begin{aligned} \eta_{k,\tau}^{(i)}(\boldsymbol{\theta}) &\triangleq \|\mathbb{E}\left[\mathbf{g}_{i,k}(\boldsymbol{\theta}, o_{i,k})|o_{i,k-\tau}\right] - \bar{\mathbf{g}}(\boldsymbol{\theta})\|, \\ \delta_{k,\tau} &\triangleq \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-\tau}\|. \end{aligned} \tag{9}$$

Both $\eta_{k,\tau}^{(i)}(\boldsymbol{\theta})$ and $\delta_{k,\tau}$ are crucial quantities in the analysis. In particular, $\eta_{k,\tau}^{(i)}(\boldsymbol{\theta})$ quantifies the distance between (i) the expectation of the noisy operator $\mathbf{g}_{i,k}(\boldsymbol{\theta}, o_{i,k})$ conditioned on a past observation $o_{i,k-\tau}$, where recall that τ is the *mixing time* defined in Definition 1, and (ii) the expectation of the operator when the observation $o_{i,k}$ is sampled i.i.d. from the stationary distribution π of the Markov chain $\{o_{i,k}\}$ - see the definition (4) of $\bar{\mathbf{g}}(\boldsymbol{\theta})$. As argued in Section III, Assumption 1 has a direct implication on $\eta_{k,\tau}^{(i)}(\boldsymbol{\theta})$, due to the geometric mixing property. We formalize this implication in the following lemma which follows directly from Assumption 1 and Definition 1.

Lemma 1 (Geometric mixing)**.** *There exists $K > 1$ such that, $\forall \boldsymbol{\theta} \in \mathbb{R}^m, \forall i = 1, ..., N, k \geq \tau$,*

$$\eta_{k,\tau}^{(i)}(\boldsymbol{\theta}) \leq \alpha^2(\|\boldsymbol{\theta}\| + 1), \;\; \textit{with} \;\; \tau \leq K\log\left(\frac{1}{\alpha^2}\right). \tag{10}$$

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 13:05:04 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2025.3555941

As we will illustrate in this section, the above result is key to establish our finite-time analysis and the linear convergence speedup. In order to invoke this property, we will need to deal with terms of the form $\delta_{k,\tau}$, which also plays a relevant role in the analysis. Next, we summarize in one lemma a result from [29] that we will use in our analysis.

Lemma 2. *The following holds* $\forall \boldsymbol{\theta} \in \mathbb{R}^d$*:*

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

This result provides a property for the function $\bar{\mathbf{g}}(\boldsymbol{\theta})$ which resembles strong convexity in optimization. We will also use the fact that the random TD update directions and their steady-state versions are 2-Lipschitz [29], i.e., $\forall i \in [N], \forall k \in \mathbb{N}$, and $\forall \boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d$, we have:

$$\begin{aligned} &\|\bar{\mathbf{g}}(\boldsymbol{\theta}) - \bar{\mathbf{g}}(\boldsymbol{\theta}')\| \leq 2\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|, \text{ and} \\ &\|\mathbf{g}_{i,k}(\boldsymbol{\theta}) - \mathbf{g}_{i,k}(\boldsymbol{\theta}')\| \leq 2\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|. \end{aligned} \tag{11}$$

From [30], we further have

$$\|\mathbf{g}_{i,k}(\boldsymbol{\theta})\| \leq 2\|\boldsymbol{\theta}\| + 2\bar{r}, \forall i \in [N], \forall k \in \mathbb{N}, \forall \boldsymbol{\theta} \in \mathbb{R}^d. \tag{12}$$

We now provide an intuitive outline of the proof, highlighting the challenges and the key technical steps in establishing Theorem 1.

A. Key steps of the proof

The proof relies on analyzing the following recursion, which, in turn, follows directly from the update rule of `OAC-FedTD`:

$$\delta_{k+1}^2 = \delta_k^2 - 2\alpha\langle \mathbf{v}_k, \boldsymbol{\theta}^* - \boldsymbol{\theta}_k \rangle + \alpha^2\|\mathbf{v}_k\|^2. \tag{13}$$

Let us define the following quantities:

$$\begin{aligned} \bar{\mathbf{g}}_N(\boldsymbol{\theta}_k) &\triangleq \frac{1}{N}\sum_{i=1}^N h_{i,k}\bar{\mathbf{g}}(\boldsymbol{\theta}_k), \\ \mathbf{g}_{h,k}(\boldsymbol{\theta}_k) &\triangleq \frac{1}{N}\sum_{i=1}^N h_{i,k}\mathbf{g}_{i,k}(\boldsymbol{\theta}_k) \end{aligned} \tag{14}$$

Taking expectation on both sides of (13), we get

$$\begin{aligned} \mathbb{E}\left[\delta_{k+1}^2\right] = \mathbb{E}\left[\delta_k^2\right] &- 2\alpha\mathbb{E}\left[\langle \bar{\mathbf{g}}_N(\boldsymbol{\theta}_k), \boldsymbol{\theta}^* - \boldsymbol{\theta}_k \rangle\right] \\ &- 2\alpha\mathbb{E}\left[\langle \mathbf{g}_{h,k}(\boldsymbol{\theta}_k) - \bar{\mathbf{g}}_N(\boldsymbol{\theta}_k), \boldsymbol{\theta}^* - \boldsymbol{\theta}_k \rangle\right] \\ &- 2\alpha\mathbb{E}\left[\langle \mathbf{w}_k, \boldsymbol{\theta}^* - \boldsymbol{\theta}_k \rangle\right] + \alpha^2\|\mathbf{v}_k\|^2. \end{aligned} \tag{15}$$

Now note that $\mathbb{E}\left[\langle \mathbf{w}_k, \boldsymbol{\theta}_k - \boldsymbol{\theta}^* \rangle\right] = \langle \mathbb{E}\left[\mathbf{w}_k\right], \mathbb{E}\left[\boldsymbol{\theta}_k - \boldsymbol{\theta}^*\right] \rangle = 0$, using the fact that the measurement noise at iteration k and the iterate $\boldsymbol{\theta}_k$ are independent, and $\mathbb{E}\left[\mathbf{w}_k\right] = \mathbf{0}$. Moreover, using the fact that the distortion $h_{i,k}$ of agent i at iteration k and the parameter $\boldsymbol{\theta}_k$ are independent, we obtain

$$\begin{aligned} \mathbb{E}\left[\langle \bar{\mathbf{g}}_N(\boldsymbol{\theta}_k), \boldsymbol{\theta}^* - \boldsymbol{\theta}_k \rangle\right] &= \frac{1}{N}\sum_{i=1}^N \mathbb{E}\left[h_{i,k}\right]\mathbb{E}\left[\langle \bar{\mathbf{g}}(\boldsymbol{\theta}_k), \boldsymbol{\theta}^* - \boldsymbol{\theta}_k \rangle\right] \\ &= m_h\mathbb{E}\left[\langle \bar{\mathbf{g}}(\boldsymbol{\theta}_k), \boldsymbol{\theta}^* - \boldsymbol{\theta}_k \rangle\right]. \end{aligned} \tag{16}$$

Based on the above discussion, we can write

$$\begin{aligned} \mathbb{E}\left[\delta_{k+1}^2\right] = \mathbb{E}\left[\delta_k^2\right] &- 2\alpha m_h\langle \bar{\mathbf{g}}(\boldsymbol{\theta}_k), \boldsymbol{\theta}^* - \boldsymbol{\theta}_k \rangle \\ &+ 2\alpha\mathbb{E}\left[\langle \mathbf{g}_{h,k}(\boldsymbol{\theta}_k) - \bar{\mathbf{g}}_N(\boldsymbol{\theta}_k), \boldsymbol{\theta}_k - \boldsymbol{\theta}^* \rangle\right] \\ &+ \alpha^2\|\mathbf{v}_k\|^2. \end{aligned} \tag{17}$$

Now define

$$\psi_k \triangleq \langle \mathbf{g}_{h,k}(\boldsymbol{\theta}_k) - \bar{\mathbf{g}}_N(\boldsymbol{\theta}_k), \boldsymbol{\theta}_k - \boldsymbol{\theta}^* \rangle. \tag{18}$$

Using Lemma 2, we then obtain

$$\begin{aligned} \mathbb{E}\left[\delta_{k+1}^2\right] \leq \mathbb{E}\left[\delta_k^2\right] &- 2\alpha m_h(1-\gamma)\omega\mathbb{E}\left[\delta_k^2\right] \\ &+ 2\alpha\mathbb{E}\left[\psi_k\right] + \alpha^2\mathbb{E}\left[\|\mathbf{v}_k\|^2\right]. \end{aligned} \tag{19}$$

The most challenging part of the analysis is in bounding $\mathbb{E}\left[\|\mathbf{v}_k\|^2\right]$ and $\mathbb{E}\left[\psi_k\right]$ while guaranteeing a linear speedup w.r.t. the number of agents. In fact, even without the channel effects, this is highly non-trivial. Let us elaborate on this point. First, in standard stochastic optimization analyses, $\mathbb{E}\left[\psi_k\right]$ would vanish under the unbiasedness assumption of the stochastic gradient oracle. However, in our case, since the Markovian observations are temporally coupled, $\mathbb{E}\left[\psi_k\right]$ does not vanish. To work around this difficulty, the bounding techniques in the centralized setting, like the ones in [29] and [30], use mixing-time arguments in conjunction with equation (26). Unfortunately, directly appealing to such techniques will fail to provide the desired convergence speedup that we seek in our multi-agent setting. The key technical step of our proof is providing a bound for $\mathbb{E}\left[\|\mathbf{v}_k\|^2\right]$ of the following form:

$$\begin{aligned} \mathbb{E}\left[\|\mathbf{v}_k\|^2\right] \leq O\left(p_h\right)\mathbb{E}\left[\delta_k^2\right] + O\left(\frac{\sigma^2 p_h}{N}\right) \\ + O\left(\sigma^2 m_h^2\alpha^4\right) + O\left(\frac{\tilde{\sigma}_{\mathbf{w}}^2 d}{N^2}\right). \end{aligned} \tag{20}$$

We derive this bound by appealing to the Lipschitz properties of $\mathbf{g}_{i,k}(\boldsymbol{\theta}_k)$ and performing some careful manipulations that allow us to exploit the mixing property of the Markov chain which we formally illustrated in Lemma 1. The bound is stated in **Lemma 3**. Leveraging this key result, our next main step is to obtain a bound on $\mathbb{E}\left[\delta_{k,\tau}^2\right]$ of the following form:

$$\begin{aligned} \mathbb{E}\left[\delta_{k,\tau}^2\right] \leq O\left(\alpha^2\tau^2 p_h\right)\mathbb{E}\left[\delta_k^2\right] + O\left(\alpha^2\tau^2\frac{p_h\sigma^2}{N}\right) \\ + O\left(\tau^2\sigma^2\alpha^4\right) + O\left(\alpha^2\tau^2\frac{\tilde{\sigma}_{\mathbf{w}}^2 d}{N^2}\right). \end{aligned} \tag{21}$$

This result, derived in **Lemma 4**, turns out to play an essential role in bounding $\mathbb{E}\left[\psi_k\right]$. In particular, we show that

$$\begin{aligned} \mathbb{E}\left[\psi_k\right] \leq O\left(\alpha\tau p_h\right)\mathbb{E}\left[\delta_k^2\right] + O\left(\frac{\alpha\tau p_h\sigma^2}{N}\right) \\ + O\left(\tau p_h\sigma^2\alpha^3\right) + O\left(\frac{\alpha\tau\tilde{\sigma}_{\mathbf{w}}^2 d}{N^2}\right). \end{aligned}$$

We establish this bound in **Lemma 5**. Combining these bounds allows us to prove Theorem 1. Indeed, with these results, we can re-write (19) as

$$\begin{aligned} \mathbb{E}\left[\delta_{k+1}^2\right] \leq (1 - \alpha m_h(1-\gamma)\omega)\mathbb{E}\left[\delta_k^2\right] + O\left(\alpha^2\tau\frac{p_h\sigma^2}{N}\right) \\ + O\left(\tau p_h\sigma^2\alpha^4\right) + O\left(\alpha^2\tau\frac{\tilde{\sigma}_{\mathbf{w}}^2 d}{N^2}\right). \end{aligned} \tag{22}$$

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 13:05:04 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2025.3555941

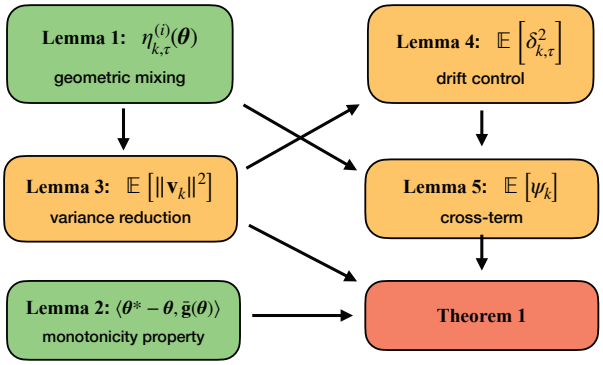


Fig. 2. Flowchart of the proof of Theorem 1 with respect to the five Lemmas presented in Section IV. We mark in green the results from the literature that we use. In orange, our novel results that we present in Section IV-B and prove in the Appendix. In red, Theorem 1, which is the main result of this paper. In each auxiliary block, we write the quantity that the results bound. Arrows connecting blocks signify that a result is used to prove another result.

Unrolling this inequality and with some additional calculations, we can arrive at the main bound (5) of Theorem 1. In the next section, we state and comment Lemma 3, 4 and 5 which are the fundamental milestones to prove the main result. We summarize the outline of the proof in Figure 2.

B. Main Technical Lemmas

We now state three lemmas that we mentioned in the previous section and that are instrumental to the proof of Theorem 1. In particular, these three results allow us to bound the terms $\mathbb{E}\left[\|\mathbf{v}_k\|^2\right]$, $\mathbb{E}\left[\delta_{k,\tau}\right]$ and $\mathbb{E}\left[\psi_k\right]$. We start by providing a bound on $\mathbb{E}\left[\|\mathbf{v}_k\|^2\right]$ of the form illustrated in (20). To that end, we state and prove the following lemma.

Lemma 3. *For $k \geq \tau$, we have*

$$\mathbb{E}\left[\|\mathbf{v}_k\|^2\right] \leq 8p_h\mathbb{E}\left[\delta_k^2\right] + 32\frac{\sigma^2 p_h}{N} + 8\sigma^2 m_h^2\alpha^4 + \frac{\tilde{\sigma}_{\mathbf{w}}^2 d}{N^2}. \quad (23)$$

The key to prove this result is manipulating the explicit expression of the squared norm of $\mathbf{v}_k$ so that we can apply the mixing property which we formally illustrated in Lemma 1. The key takeaway from the lemma is that $\mathbb{E}\left[\|\mathbf{v}_k\|^2\right]$ can be upper bounded by $\mathbb{E}\left[\delta_k^2\right]$ and by variance terms of fading distortion (p_h), measurement noise $\left(\tilde{\sigma}_{\mathbf{w}}^2\right)$, and Markovian sampling $\left(\sigma^2\right)$, scaled down either by the number of agents N or by a higher order function of the step size α.

Our next key result, which builds upon Lemma 3, is a bound on the *drift term* $\mathbb{E}\left[\delta_{k,\tau}^2\right]$.

Lemma 4. *Let $k \geq 2\tau$ and $\alpha \leq \frac{1}{68\tau p_h}$. We then have*

$$\begin{aligned}\mathbb{E}\left[\delta_{k,\tau}^2\right] &\leq 64\alpha^2\tau^2 p_h\mathbb{E}\left[\delta_k^2\right] + 96\alpha^2\tau^2\frac{p_h\sigma^2}{N} \\ &\quad + 4\alpha^4\tau^2\sigma^2 + 4\alpha^2\tau^2\frac{\tilde{\sigma}_{\mathbf{w}}^2 d}{N^2}.\end{aligned} \quad (24)$$

Like Lemma 3, the key takeaway of the above result is that the noise terms are all scaled down by either the number of agents N or by higher-order functions of the step size α. Using jointly the two lemmas above (Lemma 3 and 4), we are now able to provide a bound for $\mathbb{E}\left[\psi_k\right]$, which is the last main ingredient we need to prove Theorem 1.

Lemma 5. *Let $k \geq 2\tau$ and $\alpha \leq \frac{1}{68\tau p_h}$. We then have*

$$\begin{aligned}\mathbb{E}\left[\psi_k\right] &\leq 435\alpha\tau p_h\mathbb{E}\left[\delta_k^2\right] + 657\alpha\tau\frac{p_h\sigma^2}{N} \\ &\quad + 30\tau p_h\sigma^2\alpha^3 + 27\alpha\tau\frac{\tilde{\sigma}_{\mathbf{w}}^2 d}{N^2}.\end{aligned} \quad (25)$$

The proof of the above lemma, in addition to relying on Lemmas 3 and 4, also directly relies on the mixing Markov property as formalized in Lemma 1. Equipped with these bounds on $\|\mathbf{v}_k\|^2$ and on ψ_k, the proof of Theorem 1 follows deriving the inequality in (22) and then unrolling it and computing a further bound on the initial condition. We provide the complete proofs of all Lemmas and the explicit illustration of the final part of the proof of the main result in the Appendix.

V. Simulation Results

In this section, we provide simulation results to validate our theory. We consider (i) synthetic Markov reward processes (MRPs) generated at random in repeated trials and (ii) a more concrete MRP obtained by fixing a policy to be played in the OpenAI-gym Frozen Lake environment [46]. For the synthetic MDP, we consider $|\mathcal{S}| = n = 100$ states and a feature space spanned by $d = 20$ orthonormal basis vectors. For the purpose of our *policy evaluation* simulations, we need to generate a reward function $R_\mu : \mathcal{S} \to \mathbb{R}$ and a transition probability matrix $\mathbf{P}_\mu \in \mathbb{R}^{n\times n}$. We generate the matrix $\mathbf{P}_\mu$ with uniform random numbers - making sure we normalize the rows so that they sum to one, and the reward vector $\mathbf{r}_\mu = [R_\mu(s_1), ..., R_\mu(s_n)]^\top$ at random as well. In the Frozen Lake environment, we consider a random policy over a grid with $|\mathcal{S}| = 25$ states and slippery transitions. We set the discount factor to $\gamma = 0.5$ and the step size to $\alpha = 0.02$. Once $\mathbf{P}_\mu$, $\mathbf{r}_\mu$, and γ are fixed, and once we fix the basis functions $[\boldsymbol{\phi}_1, ..., \boldsymbol{\phi}_d]$, we can compute the best linear function approximation parameter $\boldsymbol{\theta}^*$ in closed form solving $\boldsymbol{\Phi}^\top\mathbf{D}\left(\boldsymbol{\Phi} - \gamma\mathbf{P}_\mu\boldsymbol{\Phi}\right)\boldsymbol{\theta} = \boldsymbol{\Phi}^\top\mathbf{D}\mathbf{r}_\mu$. Note that the matrix $\mathbf{D} = \text{diag}(\boldsymbol{\pi})$, with $\boldsymbol{\pi} = [\pi(s_1), ..., \pi(s_n)]^\top$ - where π is the stationary distribution of the Markov chain, can be computed noting that $\boldsymbol{\pi}^\top\mathbf{P}_\mu = \boldsymbol{\pi}^\top$, and then finding the eigenvector corresponding to the unit eigenvalue of $\mathbf{P}_\mu^\top$. In most experiments, we generate the channel distortion $h_{i,k}$ (with mean m_h and variance σ_h^2) as a Rayleigh random variable, which is the most widely adopted model for fading channels [6], [47]. In Figure 4-(b), instead, we compare different models for the fading distortion, including Nakagami [43] and Rice [44] fading. The focus of the simulations is on two aspects: (i) the effect of the OAC channel distortion on the convergence behavior and (ii) the benefit of cooperation in improving the convergence of `OAC-FedTD`, i.e., convergence speedup w.r.t. N. To this end, we compare the proposed `OAC-FedTD` algorithm with a vanilla version of federated TD learning (i.e., `OAC-FedTD` without channel effects) to which we refer to as `FedTD`. For fairness, in all simulations we set $m_h = 1$, which implies that $\sigma_h^2 \simeq 0.3$, for all fading models, and $\sigma_w^2 = 0.25$, which are values aligned with those of several OAC related

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 13:05:04 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2025.3555941

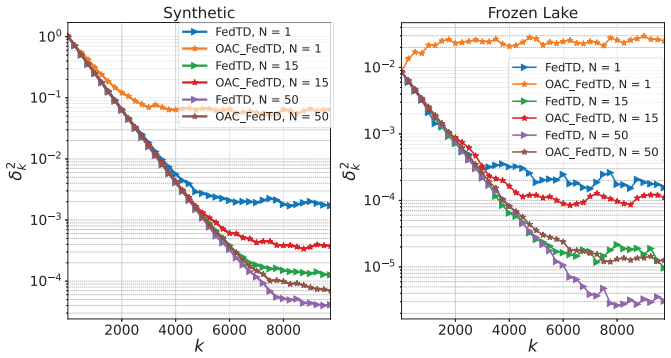


Fig. 3. Comparison between vanilla FedTD and OAC−FedTD, in the single-agent ($N = 1$) and multi-agent ($N = 15, 50$) cases. In the sub-figure to the left, we plot the results with the synthetic MDP, while in the figure to the right we show the results when the MRP is induced by a random policy in the OpenAI-gym Frozen Lake environment.

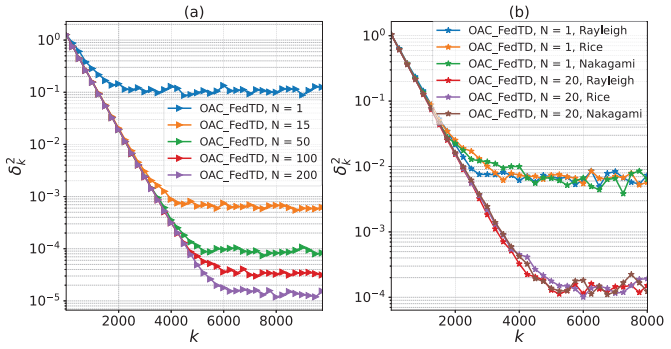


Fig. 4. In sub-figure (a), to the left, to further validate the linear speedup that we show theoretically in Theorem 1, we show the results on OAC−FedTD for several values of the number of agents $N = 1, 15, 50, 100, 200$. In sub-figure (b) to the right, we show the results obtained with different channel distortion models (Rayleigh, Rice, and Nakagami) for $N = 1, 20$.

works, e.g., [6], [48], [49]. From Fig. 3, we note that for the single-agent case ($N = 1$), the convergence error-floor of FedTD is smaller relative to OAC−FedTD with channel distortion, as expected, for both synthetic and Frozen Lake MRPs. For Frozen Lake, in particular, the algorithm is not able to reduce the error for $N = 1$ for the selected step size. Crucially, by increasing the number of agents ($N = 15, 50$), we observe that the error gap can be made much smaller for both MRPs (by two and more orders of magnitude), thus validating the beneficial effect of collaboration and the convergence speedup w.r.t. N in Theorem 1. In Figure 4-(a), to better validate the linear speedup effect with N, we show the performance of OAC−FedTD when the number of agents is $N = 1, 15, 50, 100, 200$. The linear speedup property of OAC−FedTD is effectively validated in the results: indeed, note that the error halves (approximately) when we double the number of agents N. In Figure 4-(b), we show the performance obtained for $N = 1, 15$ when using three different models of fading distortion: Rayleigh, Rician, and Nakagami fading. Note that although slight differences can be observed, the performance is very similar across the different models. This is compliant with our theory which claims that the only relevant quantities for the converegence (in expectation) are

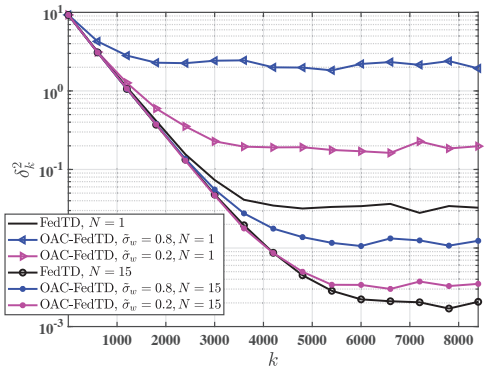


Fig. 5. Performance of OAC−FedTD for different values of the standard deviation of the measurement noise at the receiver ($\tilde{\sigma}_{\mathbf{w}} = 0.2, 0.8$), and for different values of the number of cooperating agents ($N = 1, 15$).

the first and second moment, regardless of the other fading distributions features.

In Figure 5, we show some further results comparing the performance of OAC−FedTD under two different values of the standard deviation of the receiver noise, namely $\tilde{\sigma}_{\mathbf{w}} = 0.2, 0.8$. We show the performance for $N = 1, 15$ agents, with the step size set to $\alpha = 0.2$ and number of states set to $|\mathcal{S}| = 20$. Consistent with the theory, we can see how the convergence ball is inflated by the noise at the receiver. We can note, however, that for $N = 15$, thanks to the speedup guaranteed by cooperation, the noise ball inflation is contained and we still get better performance compared to the single-agent case.

A. A Simple Control Example

Now, to conclude this section, we provide a simple example of applying OAC−FedTD beyond policy evaluation. Specifically, we consider the Q-learning algorithm [18] with linear function approximation applied to solve a shortest-path-to-goal $\ell \times \ell$ grid-world problem [50] ($|\mathcal{S}| = \ell^2$) with starting point and goal in opposite corners. The action space is $\mathcal{A} = \{$"left", "down", "right", "up"$\}$, with a state-action space of size $|\mathcal{S}| \times |\mathcal{A}| = 4\ell^2$. For each action $a \in \mathcal{A}$, we approximate the corresponding state representation using a one-hot encoding of rows and columns and stacking them together. Note that this implies that the parameter $\boldsymbol{\theta}$ we learn is of size $d = 2\ell \times |\mathcal{A}| = 8\ell$, because for both columns and rows the one-hot representation requires a vector of size ℓ. Let us denote the $\boldsymbol{\theta}$-approximated Q-table by $Q_{\boldsymbol{\theta}}(s, a)$. As a behavioral policy, as is common [51], we choose a policy alternating between the action $\arg\max_{a \in \mathcal{A}} Q_{\boldsymbol{\theta}}(s, a)$, that maximizes the current approximation of the Q function, and a random action choice, with probability $1-\epsilon$ and ϵ, respectively. Across learning steps, we linearly decrease the probability of a random action from $\epsilon = 0.9$ to $\epsilon = 0.2$ to encourage exploration of the state-action space at the beginning of the training. In Figure 6, we show the outcome of running the federated Q-learning algorithm using the OAC−FedTD aggregation scheme, over the grid-world environment with $\ell = 13$, for $N = 1, 5, 10$. Note that in this configuration the state-action space has size 676 and the parameter space is in $\mathbb{R}^d$ with $d = 104$. We set the discount factor to $\gamma = 0.95$, the learning rate to $\alpha = 0.02$,

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 13:05:04 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2025.3555941

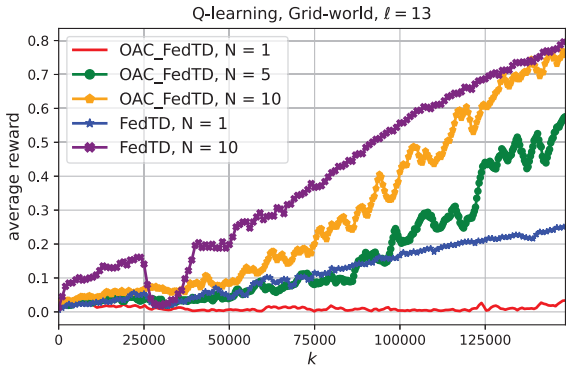


Fig. 6. Illustration of the convergence speedup for federated Q-learning in terms of the average reward across iterations using the OAC−FedTD and FedTD aggregation schemes. We plot the average per-agent reward vs iteration k, varying the number of agents N.

fading channel parameters to $m_h = 1$, $\sigma_h \approx 0.3$, and we set the measurement noise at the receiver to $\tilde{\sigma}_{\mathbf{w}} = 0.2$. We also include the results using the vanilla FedTD digital aggregation scheme for $N = 1, 10$, for comparison. As a performance metric, we plot the average reward over 1000 consecutive steps, multiplied by the Manhattan distance between starting point and goal vs the current iteration k. The resulting average rewards across epochs show a pattern similar to the one obtained in our policy evaluation experiments, which suggests that the convergence speedup with N of OAC−FedTD holds also in the policy control setting of Q-learning with linear function approximation. As we argue next in our concluding remarks, extending our theoretical analysis to this control setting is therefore a promising direction for future work.

VI. Conclusion and Future Work

We studied, for the first time, a federated policy evaluation problem where multiple agents interacting with the same MDP upload local TD(0) update directions over noisy wireless channels that introduce distortions. Our main contribution was to provide a rigorous finite-time analysis of this scheme that we referred to as OAC−FedTD. Our theoretical bounds reveal how the noise and distortion introduced by the channel affects sample-complexity bounds. Moreover, we showed that despite the channel effects, OAC−FedTD achieves a linear convergence speedup w.r.t. the number of agents under Markovian sampling. In particular, our proof of the linear speedup is novel and can be applied to other federated RL problems as well.

There are several interesting questions one can explore based on our work. For instance, the focus of this paper was solely on policy evaluation. The next natural step would be to consider the control problem and study Q-learning and policy gradient algorithms over wireless channels. One can also consider function approximation beyond the linear case to, e.g., neural networks. Our preliminary experiments in this direction suggest that cooperative speedups might occur in that case too, but we leave formally establishing this as future work. One can also ramp up the generality of the wireless channel model and study the effect of interference phenomena [8]. We believe that the analysis framework in our paper provides the theoretical foundation to address these relevant follow-up directions.

Appendix

Here, we provide the proofs of all our theoretical results. First, note that, using equation (12), and given that $(x+y)^2 \leq 2(x^2+y^2), \forall x, y \in \mathbb{R}$, and the definition of σ, we will often use the following:

$$\begin{aligned}\|\mathbf{g}_{i,k}(\boldsymbol{\theta})\|^2 \leq 4(\|\boldsymbol{\theta}\|+\bar{r})^2 &\leq 8(\|\boldsymbol{\theta}\|^2+\bar{r}^2) \\ &\leq 8(\|\boldsymbol{\theta}\|^2+\sigma^2).\end{aligned} \tag{26}$$

In what follows, $\tau = \tau_\epsilon$ with $\epsilon = \alpha^2$.

Proof of Lemma 3

Let us start by noting that the randomness in $\boldsymbol{\theta}_k$ is induced by $\{h_{i,\ell}\}_{i\in[N],\ell\in[k-1]}, \{o_{i,\ell}\}_{i\in[N],\ell\in[k-1]}$, and $\{\mathbf{w}_\ell\}_{\ell\in[k-1]}$. Based on our assumptions on the noise process, $\mathbf{w}_k$ is independent of each of these random variables and also independent of $\{h_{i,k}\}_{i\in[N]}$ and $\{o_{i,k}\}_{i\in[N]}$. Using these observations with the fact that $\mathbb{E}[\mathbf{w}_k] = \mathbf{0}$, we immediately obtain $\mathbb{E}[\langle \mathbf{g}_{h,k}(\boldsymbol{\theta}_k), \mathbf{w}_k\rangle] = \langle \mathbb{E}[\mathbf{g}_{h,k}(\boldsymbol{\theta}_k)], \mathbb{E}[\mathbf{w}_k]\rangle = 0$.
This yields:

$$\mathbb{E}\left[\|\mathbf{v}_k\|^2\right] = \mathbb{E}\left[\|\mathbf{g}_{h,k}(\boldsymbol{\theta}_k)\|^2\right] + \mathbb{E}\left[\|\mathbf{w}_k\|^2\right]. \tag{27}$$

Now, note that in the centralized/single-agent TD analysis, $\|\mathbf{g}_{h,k}(\boldsymbol{\theta}_k)\|^2$ could be bounded using (12), and this would provide a term of the form $O(\delta_k^2) + O(\sigma^2)$. This approach would, however, fail to provide a linear convergence speedup with the number of agents, N. We will show how through a finer analysis, we can establish a tighter bound. We start by writing

$$\begin{aligned}\|\mathbf{g}_{h,k}(\boldsymbol{\theta}_k)\|^2 &= \|\mathbf{g}_{h,k}(\boldsymbol{\theta}_k) - \mathbf{g}_{h,k}(\boldsymbol{\theta}^*) + \mathbf{g}_{h,k}(\boldsymbol{\theta}^*)\|^2 \\ &\leq \frac{2}{N^2}(T_1 + T_2),\end{aligned} \tag{28}$$

where T_1 and T_2 are as follows:

$$\begin{aligned}T_1 &= \|\sum_{i=1}^N h_{i,k}\mathbf{g}_{i,k}(\boldsymbol{\theta}^*)\|^2, \\ T_2 &= \|\sum_{i=1}^N h_{i,k}(\mathbf{g}_{i,k}(\boldsymbol{\theta}_k) - \mathbf{g}_{i,k}(\boldsymbol{\theta}^*))\|^2.\end{aligned} \tag{29}$$

We proceed to bound T_1. We write $T_1 = T_{11} + T_{12}$, with

$$\begin{aligned}T_{11} &= \sum_{i=1}^N h_{i,k}^2 \|\mathbf{g}_{i,k}(\boldsymbol{\theta}^*)\|^2, \text{ and} \\ T_{12} &= \sum_{\substack{i,j=1 \\ i\neq j}}^N h_{i,k}h_{j,k}\langle \mathbf{g}_{i,k}(\boldsymbol{\theta}^*), \mathbf{g}_{j,k}(\boldsymbol{\theta}^*)\rangle.\end{aligned} \tag{30}$$

Using (26) and $\|\boldsymbol{\theta}^*\| \leq \sigma$, we obtain $\|\mathbf{g}_{i,k}(\boldsymbol{\theta}^*)\|^2 \leq 16\sigma^2$, and hence, $T_{11} \leq 16\sigma^2 \sum_{i=1}^N h_{i,k}^2$. Taking expectations,

$$\begin{aligned}\mathbb{E}[T_{11}] &\leq 16\sigma^2 \mathbb{E}\left[\sum_{i=1}^N h_{i,k}^2\right] \\ &= 16\sigma^2 N(m_h^2 + \sigma_h^2) \leq 16N\sigma^2 p_h.\end{aligned}$$

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 13:05:04 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2025.3555941

Next, to bound the cross-terms in T_{12}, we will exploit the mixing property in Definition 1. To that end, we write

$$\begin{aligned}
\mathbb{E}\left[T_{12}\right] &= \sum_{\substack{i,j=1\\ i\neq j}}^{N} \mathbb{E}\left[h_{i,k}h_{j,k}\langle \mathbf{g}_{i,k}(\boldsymbol{\theta}^*), \mathbf{g}_{j,k}(\boldsymbol{\theta}^*)\rangle\right] \\
&\overset{(a)}{=} \sum_{\substack{i,j=1\\ i\neq j}}^{N} \mathbb{E}\left[h_{i,k}h_{j,k}\right]\mathbb{E}\left[\langle \mathbf{g}_{i,k}(\boldsymbol{\theta}^*), \mathbf{g}_{j,k}(\boldsymbol{\theta}^*)\rangle\right] \\
&\overset{(b)}{=} \sum_{\substack{i,j=1\\ i\neq j}}^{N} \mathbb{E}\left[h_{i,k}\right]\mathbb{E}\left[h_{j,k}\right]\left\langle \mathbb{E}\left[\mathbf{g}_{i,k}(\boldsymbol{\theta}^*)\right], \mathbb{E}\left[\mathbf{g}_{j,k}(\boldsymbol{\theta}^*)\right]\right\rangle \\
&\overset{(c)}{=} m_h^2 \sum_{\substack{i,j=1\\ i\neq j}}^{N} \langle \mathbb{E}\left[\mathbb{E}\left[\mathbf{g}_{i,k}(\boldsymbol{\theta}^*)|o_{i,k-\tau}\right] - \bar{\mathbf{g}}(\boldsymbol{\theta}^*)\right], \\
&\qquad \mathbb{E}\left[\mathbb{E}\left[\mathbf{g}_{j,k}(\boldsymbol{\theta}^*)|o_{j,k-\tau}\right] - \bar{\mathbf{g}}(\boldsymbol{\theta}^*)\right]\rangle \\
&\overset{(d)}{\leq} m_h^2 \sum_{\substack{i,j=1\\ i\neq j}}^{N} \|\mathbb{E}\left[\mathbb{E}\left[\mathbf{g}_{i,k}(\boldsymbol{\theta}^*)|o_{i,k-\tau}\right] - \bar{\mathbf{g}}(\boldsymbol{\theta}^*)\right]\| \\
&\quad \times \|\mathbb{E}\left[\mathbb{E}\left[\mathbf{g}_{j,k}(\boldsymbol{\theta}^*)|o_{j,k-\tau}\right] - \bar{\mathbf{g}}(\boldsymbol{\theta}^*)\right]\| \\
&\overset{(e)}{\leq} m_h^2 \sum_{\substack{i,j=1\\ i\neq j}}^{N} \mathbb{E}\left[\underbrace{\|\mathbb{E}\left[\mathbf{g}_{i,k}(\boldsymbol{\theta}^*)|o_{i,k-\tau}\right] - \bar{\mathbf{g}}(\boldsymbol{\theta}^*)\|}_{\eta_{k,\tau}^{(i)}(\boldsymbol{\theta}^*)}\right] \\
&\quad \times \mathbb{E}\left[\underbrace{\|\mathbb{E}\left[\mathbf{g}_{j,k}(\boldsymbol{\theta}^*)|o_{j,k-\tau}\right] - \bar{\mathbf{g}}(\boldsymbol{\theta}^*)\|}_{\eta_{k,\tau}^{(j)}(\boldsymbol{\theta}^*)}\right],
\end{aligned}$$

where (a) follows from the independence between the channel distortion gains and the Markovian tuples; (b) follows from the independence between $h_{i,k}$ and $h_{j,k}$ for $i \neq j$, and between $o_{i,k}$ and $o_{j,k}$ for $i \neq j$; (c) follows from the fact that $\bar{\mathbf{g}}(\boldsymbol{\theta}^*) = \mathbf{0}$ [29]; (d) is a consequence of the Cauchy-Schwarz inequality; and (e) follows from Jensen's inequality. Now observe that:

$$\mathbb{E}\left[\eta_{k,\tau}^{(i)}(\boldsymbol{\theta}^*)\right] \times \mathbb{E}\left[\eta_{k,\tau}^{(j)}(\boldsymbol{\theta}^*)\right] \leq \left(\alpha^2(1+\|\boldsymbol{\theta}^*\|)\right)^2 \leq 4\sigma^2\alpha^4. \tag{31}$$

In the step above, we used the mixing property by noting that $k \geq \tau$. We therefore obtain that $\mathbb{E}\left[T_{12}\right] \leq 4N^2m_h^2\sigma^2\alpha^4$. Combining the bounds for $\mathbb{E}\left[T_{11}\right]$ and $\mathbb{E}\left[T_{12}\right]$ thus yields:

$$\mathbb{E}\left[T_1\right] \leq 16\sigma^2 N p_h + 4N^2m_h^2\sigma^2\alpha^4. \tag{32}$$

Now, using (11), we see that

$$\begin{aligned}
\mathbb{E}\left[T_2\right] &\leq N\sum_{i=1}^{N}\mathbb{E}\left[h_{i,k}^2\|\mathbf{g}_{i,k}(\boldsymbol{\theta}_k) - \mathbf{g}_{i,k}(\boldsymbol{\theta}^*)\|^2\right] \\
&\leq 4N\mathbb{E}\left[\delta_k^2\right]\sum_{i=1}^{N}\mathbb{E}\left[h_{i,k}^2\right] = 4p_hN^2\mathbb{E}\left[\delta_k^2\right].
\end{aligned} \tag{33}$$

Combining all the bounds above, we conclude that

$$\mathbb{E}\left[\|\mathbf{g}_{h,k}(\boldsymbol{\theta}_k)\|^2\right] \leq 8p_h\mathbb{E}\left[\delta_k^2\right] + 32\frac{\sigma^2 p_h}{N} + 8\sigma^2m_h^2\alpha^4. \tag{34}$$

The claim of the lemma then follows from the above bound and by noting that $\mathbb{E}\left[\|\mathbf{w}_k\|^2\right] = \frac{\tilde{\sigma}_{\mathbf{w}}^2 d}{N^2}$.

Proof of Lemma 4

We start by writing

$$\begin{aligned}
\delta_{k+1}^2 &= \delta_k^2 - 2\alpha\langle \mathbf{v}_k, \boldsymbol{\theta}^* - \boldsymbol{\theta}_k\rangle + \alpha^2\|\mathbf{v}_k\|^2 \\
&\leq (1+\alpha)\delta_k^2 + (\alpha+\alpha^2)\|\mathbf{v}_k\|^2 \\
&\leq (1+\alpha)\delta_k^2 + 2\alpha\|\mathbf{v}_k\|^2.
\end{aligned} \tag{35}$$

Now using Lemma 3, we have

$$\begin{aligned}
\mathbb{E}\left[\delta_{k+1}^2\right] &\leq (1+\alpha)\mathbb{E}\left[\delta_k^2\right] \\
&\quad + 2\alpha\left(8p_h\mathbb{E}\left[\delta_k^2\right] + 32\frac{\sigma^2p_h}{N} + 8\sigma^2m_h^2\alpha^4 + \frac{\tilde{\sigma}_{\mathbf{w}}^2 d}{N^2}\right) \\
&\leq (1+17\alpha p_h)\mathbb{E}\left[\delta_k^2\right] \\
&\quad + \underbrace{64\alpha\frac{\sigma^2p_h}{N} + 16\sigma^2m_h^2\alpha^5 + 2\alpha\frac{\tilde{\sigma}_{\mathbf{w}}^2 d}{N^2}}_{B}.
\end{aligned} \tag{36}$$

Iterating this inequality, we can obtain for any $k-\tau \leq k' \leq k$,

$$\mathbb{E}\left[\delta_{k'}^2\right] \leq (1+17\alpha p_h)^{\tau}\mathbb{E}\left[\delta_{k-\tau}^2\right] + B\sum_{\ell=0}^{\tau-1}(1+17\alpha p_h)^{\ell}. \tag{37}$$

Now using the fact that $(1+x) \leq e^x, \forall x \in \mathbb{R}$, observe that $(1+17\alpha p_h)^{\ell} \leq (1+17\alpha p_h)^{\tau} \leq e^{0.25} \leq 2$, for $\alpha \leq 1/(68p_h\tau)$. Using the same argument, we also have $\sum_{\ell=0}^{\tau-1}(1+17\alpha p_h)^{\ell} \leq 2\tau$. This yields:

$$\mathbb{E}\left[\delta_{k'}^2\right] \leq 2\mathbb{E}\left[\delta_{k-\tau}^2\right] + 2B\tau. \tag{38}$$

Next, note that

$$\delta_{k,\tau}^2 \leq \tau\sum_{\ell=k-\tau}^{k-1}\|\boldsymbol{\theta}_{\ell+1} - \boldsymbol{\theta}_{\ell}\|^2 = \tau\alpha^2\sum_{\ell=k-\tau}^{k-1}\|\mathbf{v}_{\ell}\|^2. \tag{39}$$

Taking expectations on both sides of the above equation and applying Lemma 3 and (38), we get

$$\begin{aligned}
\mathbb{E}\left[\delta_{k,\tau}^2\right] &\leq \alpha^2\tau\sum_{\ell=k-\tau}^{k-1}\left(8p_h\mathbb{E}\left[\delta_{\ell}^2\right] + 32\frac{\sigma^2p_h}{N}\right. \\
&\qquad \left. + 8\sigma^2m_h^2\alpha^4 + \frac{\tilde{\sigma}_{\mathbf{w}}^2 d}{N^2}\right) \\
&= 8p_h\alpha^2\tau\sum_{\ell=k-\tau}^{k-1}\mathbb{E}\left[\delta_{\ell}^2\right] + \frac{1}{2}\alpha\tau^2B \\
&\leq 8p_h\alpha^2\tau\sum_{\ell=k-\tau}^{k-1}\left(2\mathbb{E}\left[\delta_{k-\tau}^2\right] + 2B\tau\right) + \frac{1}{2}\alpha\tau^2B \\
&= 16\alpha^2\tau^2p_h\mathbb{E}\left[\delta_{k-\tau}^2\right] + 16\alpha^2\tau^3Bp_h + \frac{1}{2}\alpha\tau^2B.
\end{aligned} \tag{40}$$

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 13:05:04 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2025.3555941

In the above steps, we used the fact that $\ell \geq \tau$ since $k \geq 2\tau$. We now simplify the resulting inequality above as follows:

$$\begin{aligned}\mathbb{E}\left[\delta_{k,\tau}^2\right] &\leq 16\alpha^2\tau^2 p_h \mathbb{E}\left[\delta_{k-\tau}^2\right] \\ &+ 16\alpha^2\tau^3 p_h\left(64\alpha\frac{\sigma^2 p_h}{N} + 16\sigma^2 m_h^2\alpha^5 + 2\alpha\frac{\tilde{\sigma}_{\mathbf{w}}^2 d}{N^2}\right) \\ &+ \frac{1}{2}\alpha\tau^2\left(64\alpha\frac{\sigma^2 p_h}{N} + 16\sigma^2 m_h^2\alpha^5 + 2\alpha\frac{\tilde{\sigma}_{\mathbf{w}}^2 d}{N^2}\right) \\ &= 16\alpha^2\tau^2 p_h \mathbb{E}\left[\delta_{k-\tau}^2\right] \\ &+ 16\alpha^3\tau^3 p_h\left(64\frac{\sigma^2 p_h}{N} + 16\sigma^2 m_h^2\alpha^4 + 2\frac{\tilde{\sigma}_{\mathbf{w}}^2 d}{N^2}\right) \\ &+ \frac{1}{2}\alpha^2\tau^2\left(64\frac{\sigma^2 p_h}{N} + 16\sigma^2 m_h^2\alpha^4 + 2\frac{\tilde{\sigma}_{\mathbf{w}}^2 d}{N^2}\right) \\ &\overset{(a)}{\leq} 16\alpha^2\tau^2 p_h \mathbb{E}\left[\delta_{k-\tau}^2\right] + 48\alpha^2\tau^2\frac{\sigma^2 p_h}{N} \\ &+ 2\alpha^4\tau^2\sigma^2 + 2\alpha^2\tau^2\frac{\tilde{\sigma}_{\mathbf{w}}^2 d}{N^2},\end{aligned} \tag{41}$$

where for (a), we used $\alpha\tau \leq \frac{1}{68p_h}$, and that $\frac{m_h^2}{p_h} \leq 1$, implying $m_h^2\alpha \leq \frac{1}{68\tau}$. Now noting that $\delta_{k-\tau}^2 \leq 2\delta_k^2 + 2\delta_{k,\tau}^2$, we obtain

$$\begin{aligned}\mathbb{E}\left[\delta_{k,\tau}^2\right](1 - 32\alpha^2\tau^2 p_h) &\leq 32\alpha^2\tau^2 p_h \mathbb{E}\left[\delta_k^2\right] \\ &+ 48\alpha^2\tau^2\frac{\sigma^2 p_h}{N} + 2\alpha^4\tau^2\sigma^2 + 2\alpha^2\tau^2\frac{\tilde{\sigma}_{\mathbf{w}}^2 d}{N^2}.\end{aligned} \tag{42}$$

Since $\alpha\tau \leq \frac{1}{68p_h}$, we have that $1 - 32\alpha^2\tau^2 p_h \leq \frac{1}{2}$, and hence

$$\begin{aligned}\mathbb{E}\left[\delta_{k,\tau}^2\right] &\leq 64\alpha^2\tau^2 p_h \mathbb{E}\left[\delta_k^2\right] + 96\alpha^2\tau^2\frac{\sigma^2 p_h}{N} \\ &+ 4\alpha^4\tau^2\sigma^2 + 4\alpha^2\tau^2\frac{\tilde{\sigma}_{\mathbf{w}}^2 d}{N^2}.\end{aligned} \tag{43}$$

Proof of Lemma 5

Recall the definition of $\bar{\mathbf{g}}_N(\boldsymbol{\theta}_k) \triangleq \frac{1}{N}\sum_{i=1}^N h_{i,k}\bar{\mathbf{g}}(\boldsymbol{\theta}_k)$ and $\mathbf{g}_{h,k}(\boldsymbol{\theta}_k) \triangleq \frac{1}{N}\sum_{i=1}^N h_{i,k}\mathbf{g}_{i,k}(\boldsymbol{\theta}_k)$. We write ψ_k as $\psi_k = T_1 + T_2 + T_3 + T_4$, where

$$\begin{aligned}T_1 &= \langle \boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-\tau}, \mathbf{g}_{h,k}(\boldsymbol{\theta}_k) - \bar{\mathbf{g}}_N(\boldsymbol{\theta}_k)\rangle, \\ T_2 &= \langle \boldsymbol{\theta}_{k-\tau} - \boldsymbol{\theta}^*, \mathbf{g}_{h,k}(\boldsymbol{\theta}_{k-\tau}) - \bar{\mathbf{g}}_N(\boldsymbol{\theta}_{k-\tau})\rangle, \\ T_3 &= \langle \boldsymbol{\theta}_{k-\tau} - \boldsymbol{\theta}^*, \mathbf{g}_{h,k}(\boldsymbol{\theta}_k) - \mathbf{g}_{h,k}(\boldsymbol{\theta}_{k-\tau})\rangle, \\ T_4 &= \langle \boldsymbol{\theta}_{k-\tau} - \boldsymbol{\theta}^*, \bar{\mathbf{g}}_N(\boldsymbol{\theta}_{k-\tau}) - \bar{\mathbf{g}}_N(\boldsymbol{\theta}_k)\rangle.\end{aligned} \tag{44}$$

We now bound each of the terms $\mathbb{E}[T_1] - \mathbb{E}[T_4]$ individually. We start by observing that

$$\begin{aligned}\mathbb{E}[T_1] &= \langle \boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-\tau}, \mathbf{g}_{h,k}(\boldsymbol{\theta}_k) - \bar{\mathbf{g}}_N(\boldsymbol{\theta}_k)\rangle \\ &\leq \frac{1}{2\alpha\tau}\mathbb{E}\left[\delta_{k,\tau}^2\right] + \frac{1}{2}\alpha\tau\mathbb{E}\left[\|\mathbf{g}_{h,k}(\boldsymbol{\theta}_k) - \bar{\mathbf{g}}_N(\boldsymbol{\theta}_k)\|^2\right] \\ &\leq \frac{1}{2\alpha\tau}\mathbb{E}\left[\delta_{k,\tau}^2\right] + \alpha\tau\mathbb{E}\left[\|\mathbf{g}_{h,k}(\boldsymbol{\theta}_k)\|^2\right] \\ &+ \alpha\tau\mathbb{E}\left[\|\bar{\mathbf{g}}_N(\boldsymbol{\theta}_k) - \bar{\mathbf{g}}_N(\boldsymbol{\theta}^*)\|^2\right].\end{aligned} \tag{45}$$

Now note that $\mathbb{E}\left[\|\mathbf{g}_{h,k}(\boldsymbol{\theta}_k)\|^2\right]$ can be bounded using the same procedure we used in (28), while for $\mathbb{E}\left[\delta_{k,\tau}^2\right]$ we can invoke Lemma 4. We also have

$$\begin{aligned}&\mathbb{E}\left[\|\bar{\mathbf{g}}_N(\boldsymbol{\theta}_k) - \bar{\mathbf{g}}_N(\boldsymbol{\theta}^*)\|^2\right] \\ &\leq \frac{N}{N^2}\sum_{i=1}^N \mathbb{E}\left[h_{i,k}^2\|\bar{\mathbf{g}}(\boldsymbol{\theta}_k) - \bar{\mathbf{g}}(\boldsymbol{\theta}^*)\|^2\right] \\ &\leq \frac{4}{N}\mathbb{E}\left[\delta_k^2\right]\sum_{i=1}^N \mathbb{E}\left[h_{i,k}^2\right] = 4p_h\mathbb{E}\left[\delta_k^2\right].\end{aligned} \tag{46}$$

Now, combining the above bounds and simplifying,

$$\mathbb{E}[T_1] \leq 44\alpha\tau p_h\mathbb{E}\left[\delta_k^2\right] + 80\alpha\tau\frac{\sigma^2 p_h}{N} + 3\tau\sigma^2\alpha^3 + 2\alpha\tau\frac{\tilde{\sigma}_{\mathbf{w}}^2 d}{N^2}. \tag{47}$$

We now proceed to bound $\mathbb{E}[T_3]$. We will again use the fact that $\delta_{k-\tau}^2 \leq 2\delta_k^2 + 2\delta_{k,\tau}^2$.

$$\begin{aligned}\mathbb{E}[T_3] &= \mathbb{E}\left[\langle \boldsymbol{\theta}_{k-\tau} - \boldsymbol{\theta}^*, \mathbf{g}_{h,k}(\boldsymbol{\theta}_k) - \mathbf{g}_{h,k}(\boldsymbol{\theta}_{k-\tau})\rangle\right] \\ &= \mathbb{E}\left[\left\langle \boldsymbol{\theta}_{k-\tau} - \boldsymbol{\theta}^*, \frac{1}{N}\sum_{i=1}^N h_{i,k}(\mathbf{g}_{i,k}(\boldsymbol{\theta}_k) - \mathbf{g}_{i,k}(\boldsymbol{\theta}_{k-\tau}))\right\rangle\right] \\ &= \mathbb{E}\left[\frac{1}{N}\sum_{i=1}^N h_{i,k}\langle \boldsymbol{\theta}_{k-\tau} - \boldsymbol{\theta}^*, \mathbf{g}_{i,k}(\boldsymbol{\theta}_k) - \mathbf{g}_{i,k}(\boldsymbol{\theta}_{k-\tau})\rangle\right] \\ &\leq m_h\mathbb{E}\left[\delta_{k-\tau}\frac{1}{N}\sum_{i=1}^N \|\mathbf{g}_{i,k}(\boldsymbol{\theta}_k) - \mathbf{g}_{i,k}(\boldsymbol{\theta}_{k-\tau})\|\right] \\ &\leq \frac{\alpha\tau}{2}m_h^2\mathbb{E}\left[\delta_{k-\tau}^2\right] + \frac{2}{\alpha\tau}\mathbb{E}\left[\delta_{k,\tau}^2\right] \\ &\leq \alpha\tau m_h^2\mathbb{E}\left[\delta_k^2\right] + \alpha\tau m_h^2\mathbb{E}\left[\delta_{k,\tau}^2\right] + \frac{2}{\alpha\tau}\mathbb{E}\left[\delta_{k,\tau}^2\right] \\ &\leq \alpha\tau m_h^2\mathbb{E}\left[\delta_k^2\right] + \frac{3}{\alpha\tau}\mathbb{E}\left[\delta_{k,\tau}^2\right],\end{aligned} \tag{48}$$

where we have used that $\alpha\tau \leq \frac{1}{68p_h}$ and $\frac{m_h^2}{p_h} \leq 1$, which imply $m_h^2\alpha\tau \leq 1$. Applying Lemma 4, we can then get

$$\begin{aligned}\mathbb{E}[T_3] &\leq \alpha\tau p_h\mathbb{E}\left[\delta_k^2\right] + \frac{3}{\alpha\tau}\left(64\alpha^2\tau^2 p_h\mathbb{E}\left[\delta_k^2\right]\right. \\ &\left.+96\alpha^2\tau^2\frac{p_h\sigma^2}{N} + 4\alpha^2\tau^2\sigma^2\alpha^2 + 4\alpha^2\tau^2\frac{\tilde{\sigma}_{\mathbf{w}}^2 d}{N^2}\right).\end{aligned} \tag{49}$$

Simplifying the above bound yields:

$$\begin{aligned}\mathbb{E}[T_3] &\leq 193\alpha\tau p_h\mathbb{E}\left[\delta_k^2\right] + 288\alpha\tau\frac{p_h\sigma^2}{N} \\ &+ 12\tau\sigma^2\alpha^3 + 12\alpha\tau\frac{\tilde{\sigma}_{\mathbf{w}}^2 d}{N^2}.\end{aligned} \tag{50}$$

With analogous calculations, we can derive exactly the same bound for $\mathbb{E}[T_4]$.

We now proceed to bound $\mathbb{E}[T_2]$. For ease of notation, let us define $\mathcal{F}_{k,\tau} = (\{o_{i,k-\tau}\}_{i=1}^N, \boldsymbol{\theta}_{k-\tau})$. Observe:

$$\begin{aligned}\mathbb{E}[T_2] &= \mathbb{E}\left[\mathbb{E}[T_2|\mathcal{F}_{k,\tau}]\right] = \mathbb{E}[\langle \boldsymbol{\theta}_{k-\tau} - \boldsymbol{\theta}^*, \\ &\frac{m_h}{N}\sum_{i=1}^N(\mathbb{E}[\mathbf{g}_{i,k}(\boldsymbol{\theta}_{k-\tau}, o_{i,k})|\mathcal{F}_{k,\tau}] - \bar{\mathbf{g}}(\boldsymbol{\theta}_{k-\tau}))\rangle] \\ &\leq \mathbb{E}\left[\delta_{k-\tau}\frac{m_h}{N}\sum_{i=1}^N \eta_{k,\tau}^{(i)}(\boldsymbol{\theta}_{k-\tau})\right] \\ &4 \leq m_h\alpha^2\mathbb{E}\left[\delta_{k-\tau}(1 + \|\boldsymbol{\theta}_{k-\tau}\|)\right].\end{aligned}$$

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 13:05:04 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2025.3555941

Since $\alpha < 1$, we have $\delta_{k-\tau}(\delta_{k-\tau} + 2\sigma) \leq \frac{\delta_{k-\tau}^2}{\alpha} + 2\sigma\delta_{k-\tau} + \alpha\sigma^2 = \left(\frac{\delta_{k-\tau}}{\sqrt{\alpha}} + \sqrt{\alpha}\sigma\right)^2 \leq 2\left(\frac{\delta_{k-\tau}^2}{\alpha} + \alpha\sigma^2\right)$. Based on this observation, Lemma 4, and the fact that $m_h \leq p_h$, we obtain

$$\begin{aligned}\mathbb{E}[T_2] &\leq 2m_h\alpha^2\left(\frac{\delta_{k-\tau}^2}{\alpha} + \alpha\sigma^2\right)\\ &= 2m_h\alpha\delta_{k-\tau}^2 + 2m_h\alpha^3\sigma^2\\ &\leq 4m_h\alpha\delta_k^2 + 4m_h\alpha\delta_{k,\tau}^2 + 2m_h\alpha^3\sigma^2\\ &\leq 5p_h\alpha\tau\mathbb{E}\left[\delta_k^2\right] + \alpha\tau\frac{\sigma^2 p_h}{N}\\ &\quad + 3\tau\sigma^2 p_h\alpha^3 + \alpha\tau\frac{\tilde{\sigma}_{\mathbf{w}}^2 d}{N^2}.\end{aligned} \tag{51}$$

Combining all the terms, we can conclude the proof.

Conclusion of the proof of Theorem 1

Consider the inequality that we derived in (19). For $k \geq 2\tau$, plugging the bounds derived in Lemma 3 and in Lemma 5,

$$\begin{aligned}\mathbb{E}\left[\delta_{k+1}^2\right] &\leq \mathbb{E}\left[\delta_k^2\right] - 2\alpha m_h(1-\gamma)\omega\mathbb{E}\left[\delta_k^2\right]\\ &\quad + 2\alpha\mathbb{E}\left[\psi_k\right] + \alpha^2\mathbb{E}\left[\|\mathbf{v}_k\|^2\right]\\ &\leq \mathbb{E}\left[\delta_k^2\right] - 2\alpha m_h(1-\gamma)\omega\mathbb{E}\left[\delta_k^2\right]\\ &\quad + 2\alpha\left(435\alpha\tau p_h\mathbb{E}\left[\delta_k^2\right] + 657\alpha\tau\frac{p_h\sigma^2}{N}\right.\\ &\quad \left. + 30\tau p_h\sigma^2\alpha^3 + 27\alpha\tau\frac{\tilde{\sigma}_{\mathbf{w}}^2 d}{N^2}\right)\\ &\quad + \alpha^2\left(8p_h\mathbb{E}\left[\delta_k^2\right] + 32\frac{\sigma^2 p_h}{N} + 8\sigma^2 m_h^2\alpha^4 + \frac{\tilde{\sigma}_{\mathbf{w}}^2 d}{N^2}\right)\\ &= \mathbb{E}\left[\delta_k^2\right] - 2\alpha m_h(1-\gamma)\omega\mathbb{E}\left[\delta_k^2\right]\\ &\quad + 878\alpha^2\tau p_h\mathbb{E}\left[\delta_k^2\right] + 1346\alpha^2\tau\frac{p_h\sigma^2}{N}\\ &\quad + 61\tau p_h\sigma^2\alpha^4 + 55\alpha^2\tau\frac{\tilde{\sigma}_{\mathbf{w}}^2 d}{N^2}\\ &= \mathbb{E}\left[\delta_k^2\right] - \alpha\left(2m_h(1-\gamma)\omega - 878\alpha\tau p_h\right)\mathbb{E}\left[\delta_k^2\right]\\ &\quad + 1346\alpha^2\tau\frac{p_h\sigma^2}{N} + 61\tau p_h\sigma^2\alpha^4 + 55\alpha^2\tau\frac{\tilde{\sigma}_{\mathbf{w}}^2 d}{N^2}.\end{aligned} \tag{52}$$

Hence, for $\alpha \leq \frac{m_h(1-\gamma)\omega}{C_0\tau p_h}$, with $C_0 = 878$, we get

$$\begin{aligned}\mathbb{E}\left[\delta_{k+1}^2\right] &\leq (1 - \alpha m_h(1-\gamma)\omega)\mathbb{E}\left[\delta_k^2\right] + 1346\alpha^2\tau\frac{p_h\sigma^2}{N}\\ &\quad + 61\tau p_h\sigma^2\alpha^4 + 55\alpha^2\tau\frac{\tilde{\sigma}_{\mathbf{w}}^2 d}{N^2}.\end{aligned} \tag{53}$$

Unrolling this inequality, we obtain

$$\begin{aligned}\mathbb{E}\left[\delta_T^2\right] &\leq (1 - \alpha m_h(1-\gamma)\omega)^{T-2\tau}\mathbb{E}\left[\delta_{2\tau}^2\right]\\ &\quad + C_2\frac{\alpha\tau p_h\sigma^2}{m_h(1-\gamma)\omega N} + \frac{C_3\tau p_h\sigma^2\alpha^3}{m_h(1-\gamma)\omega}\\ &\quad + \frac{C_4\alpha\tau\tilde{\sigma}_{\mathbf{w}}^2 d}{m_h(1-\gamma)\omega N^2},\end{aligned} \tag{54}$$

with $C_2 = 1346$, $C_3 = 61$ and $C_4 = 55$. To conclude, we proceed to bound $\mathbb{E}\left[\delta_{2\tau}^2\right]$. Note that, for any $k \geq 0$,

$$\mathbb{E}\left[\delta_{k+1}^2\right] \leq (1+\alpha)\mathbb{E}\left[\delta_k^2\right] + 2\alpha\mathbb{E}\left[\|\mathbf{v}_k\|^2\right]. \tag{55}$$

Observe as before (see (27)):

$$\mathbb{E}\left[\|\mathbf{v}_k\|^2\right] = \mathbb{E}\left[\|\mathbf{g}_{h,k}(\boldsymbol{\theta}_k)\|^2\right] + \mathbb{E}\left[\|\mathbf{w}_k\|^2\right]. \tag{56}$$

Note that $\mathbb{E}\left[\|\mathbf{w}_k\|^2\right] = \frac{\tilde{\sigma}_{\mathbf{w}}^2 d}{N^2}$ and that we can bound $\mathbb{E}\left[\|\mathbf{g}_{h,k}(\boldsymbol{\theta}_k)\|^2\right]$ as follows:

$$\begin{aligned}\mathbb{E}\left[\|\mathbf{g}_{h,k}(\boldsymbol{\theta}_k)\|^2\right] &= \mathbb{E}\left[\|\frac{1}{N}\sum_{i=1}^N h_{i,k}\mathbf{g}_{i,k}(\boldsymbol{\theta}_k)\|^2\right]\\ &\leq \frac{N}{N^2}\sum_{i=1}^N h_{i,k}^2\|\mathbf{g}_{i,k}(\boldsymbol{\theta}_k)\|^2\\ &\leq \frac{1}{N}\left(8(\|\boldsymbol{\theta}_k\|^2 + \sigma^2)\sum_{i=1}^N\mathbb{E}\left[h_{i,k}^2\right]\right)\\ &\leq \left(8(2\delta_k^2 + 3\sigma^2)\right)p_h\\ &= 16p_h\mathbb{E}\left[\delta_k^2\right] + 24p_h\sigma^2.\end{aligned} \tag{57}$$

Hence, $\mathbb{E}\left[\|\mathbf{v}_k\|^2\right] \leq 16p_h\mathbb{E}\left[\delta_k^2\right] + 24p_h\sigma^2 + \frac{\tilde{\sigma}_{\mathbf{w}}^2 d}{N^2}$. Thus,

$$\begin{aligned}\mathbb{E}\left[\delta_{k+1}^2\right] &\leq (1+\alpha)\mathbb{E}\left[\delta_k^2\right]\\ &\quad + 2\alpha\left(16p_h\mathbb{E}\left[\delta_k^2\right] + 24p_h\sigma^2 + \frac{\tilde{\sigma}_{\mathbf{w}}^2 d}{N^2}\right)\\ &\leq (1 + 33p_h\alpha)\mathbb{E}\left[\delta_k^2\right] + 48\alpha p_h\sigma^2 + 2\alpha\frac{\tilde{\sigma}_{\mathbf{w}}^2 d}{N^2}.\end{aligned} \tag{58}$$

Iterating this inequality, we obtain

$$\begin{aligned}\mathbb{E}\left[\delta_{2\tau}^2\right] &\leq (1 + 33\alpha p_h)^{2\tau}\delta_0^2\\ &\quad + \left(48\alpha p_h\sigma^2 + 2\alpha\frac{\tilde{\sigma}_{\mathbf{w}}^2 d}{N^2}\right)\sum_{j=0}^{2\tau-1}(1 + 33\alpha p_h)^j.\end{aligned} \tag{59}$$

Now with the same procedure used to obtain (38), we see that if $66\alpha p_h\tau \leq \frac{1}{4}$, then

$$\begin{aligned}\mathbb{E}\left[\delta_{2\tau}^2\right] &\leq 2\mathbb{E}\left[\delta_0^2\right] + 4\tau\left(48\alpha p_h\sigma^2 + 2\alpha\frac{\tilde{\sigma}_{\mathbf{w}}^2 d}{N^2}\right)\\ &\leq 2\mathbb{E}\left[\delta_0^2\right] + 192\alpha\tau\sigma^2 p_h + \frac{8\alpha\tilde{\sigma}_{\mathbf{w}}^2 d\tau}{N^2}\\ &\leq 2\mathbb{E}\left[\delta_0^2\right] + \sigma^2 + \frac{\tilde{\sigma}_{\mathbf{w}}^2 d}{N^2},\end{aligned} \tag{60}$$

where we have used that $\alpha\tau \leq \frac{1}{878p_h}$. With the choice of step size in the statement of the theorem, we have $\alpha\omega(1-\gamma)m_h \leq 1$. This then yields $(1 - \alpha\omega(1-\gamma)m_h)^{-2\tau} \leq (1 - \alpha m_h)^{-2\tau}$. Finally note that since $\alpha\tau \leq \frac{1}{C_0 p_h} \leq \frac{1}{4p_h}$, we have $\alpha\tau m_h \leq \frac{m_h}{4p_h} \leq \frac{1}{4}$, where we used the fact that $\frac{m_h}{p_h} \leq 1$. Based on this discussion and using Bernoulli's inequality, we obtain $(1 - \alpha m_h)^{2\tau} \geq 1 - 2\alpha\tau m_h \geq \frac{1}{2}$; hence, $(1 - \alpha m_h)^{-2\tau} \leq 2$. We thus have $(1 - \alpha\omega(1-\gamma)m_h)^{-2\tau} \leq 2$. Plugging this bound back in (54) completes the proof.

REFERENCES

[1] Y. Shi, K. Yang, T. Jiang, J. Zhang, and K. B. Letaief, "Communication-efficient edge ai: Algorithms and systems," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 4, pp. 2167–2191, 2020.

[2] L. U. Khan, S. R. Pandey, N. H. Tran, W. Saad, Z. Han, M. N. Nguyen, and C. S. Hong, "Federated learning for edge networks: Resource optimization and incentive mechanism," *IEEE Communications Magazine*, vol. 58, no. 10, pp. 88–93, 2020.

[3] B. McMahan, E. Moore, D. Ramage, S. Hampson, and B. A. y Arcas, "Communication-efficient learning of deep networks from decentralized data," in *Artificial Intelligence and Statistics*. PMLR, 2017, pp. 1273–1282.

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 13:05:04 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2025.3555941

[4] Y. Shi, K. Yang, T. Jiang, J. Zhang, and K. B. Letaief, "Communication-efficient edge ai: Algorithms and systems," *IEEE Communications Surveys & Tutorials*, vol. 22, no. 4, pp. 2167–2191, 2020.

[5] A. Reisizadeh, A. Mokhtari, H. Hassani, A. Jadbabaie, and R. Pedarsani, "Fedpaq: A communication-efficient federated learning method with periodic averaging and quantization," in *AISTATS*. PMLR, 2020, pp. 2021–2031.

[6] T. Sery and K. Cohen, "On analog gradient descent learning over multiple access fading channels," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2897–2911, 2020.

[7] G. Zhu, Y. Du, D. Gündüz, and K. Huang, "One-bit over-the-air aggregation for communication-efficient federated edge learning: Design and convergence analysis," *IEEE Transactions on Wireless Communications*, vol. 20, no. 3, pp. 2120–2135, 2020.

[8] H. H. Yang, Z. Chen, T. Q. Quek, and H. V. Poor, "Revisiting analog over-the-air machine learning: The blessing and curse of interference," *IEEE Journal of Selected Topics in Signal Processing*, vol. 16, no. 3, pp. 406–419, 2021.

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

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

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

[12] B. Liu, L. Wang, and M. Liu, "Lifelong federated reinforcement learning: A learning architecture for navigation in cloud robotic systems," *IEEE Robotics and Automation Letters*, vol. 4, no. 4, pp. 4555–4562, 2019.

[13] H. Jin, Y. Peng, W. Yang, S. Wang, and Z. Zhang, "Federated reinforcement learning with environment heterogeneity," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2022, pp. 18–37.

[14] Y. Hu, Y. Hua, W. Liu, and J. Zhu, "Reward shaping based federated reinforcement learning," *IEEE Access*, vol. 9, pp. 67 259–67 267, 2021.

[15] X. Wang, C. Wang, X. Li, V. C. M. Leung, and T. Taleb, "Federated deep reinforcement learning for internet of things with decentralized cooperative edge caching," *IEEE Internet of Things Journal*, vol. 7, no. 10, pp. 9441–9455, 2020.

[16] S. Wang, M. Chen, C. Yin, and H. V. Poor, "Convergence time minimization for federated reinforcement learning over wireless networks," in *2022 56th Annual Conference on Information Sciences and Systems (CISS)*, 2022, pp. 246–251.

[17] M. Krouka, A. Elgabli, C. B. Issaid, and M. Bennis, "Communication-efficient and federated multi-agent reinforcement learning," *IEEE Transactions on Cognitive Communications and Networking*, vol. 8, no. 1, pp. 311–320, 2021.

[18] C. J. Watkins and P. Dayan, "Q-learning," *Machine learning*, vol. 8, pp. 279–292, 1992.

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

[20] M. M. Amiri and D. Gündüz, "Machine learning at the wireless edge: Distributed stochastic gradient descent over-the-air," *IEEE Transactions on Signal Processing*, vol. 68, pp. 2155–2169, 2020.

[21] M. Krouka, A. Elgabli, C. B. Issaid, and M. Bennis, "Communication-efficient and federated multi-agent reinforcement learning," *IEEE Transactions on Cognitive Communications and Networking*, vol. 8, no. 1, pp. 311–320, 2022.

[22] X. Cao, G. Zhu, J. Xu, Z. Wang, and S. Cui, "Optimized power control design for over-the-air federated edge learning," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 1, pp. 342–358, 2021.

[23] V. P. Chellapandi, L. Yuan, C. G. Brinton, S. H. Żak, and Z. Wang, "Federated learning for connected and automated vehicles: A survey of existing approaches and challenges," *IEEE Transactions on Intelligent Vehicles*, 2023.

[24] M. M. Amiri and D. Gündüz, "Federated learning over wireless fading channels," *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, pp. 3546–3557, 2020.

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

[26] F. Pase, M. Giordani, and M. Zorzi, "On the convergence time of federated learning over wireless networks under imperfect csi," in *2021 IEEE International Conference on Communications Workshops (ICC Workshops)*. IEEE, 2021, pp. 1–7.

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

[28] V. S. Borkar and S. P. Meyn, "The ode method for convergence of stochastic approximation and reinforcement learning," *SIAM Journal on Control and Optimization*, vol. 38, no. 2, pp. 447–469, 2000.

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

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

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

[32] A. Mitra, "A simple finite-time analysis of TD learning with linear function approximation," *IEEE Transactions on Automatic Control*, (to appear).

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

[34] R. Liu and A. Olshevsky, "Distributed td (0) with almost no communication," *IEEE Control Systems Letters*, 2023.

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

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

[37] A. Beznosikov, S. Horváth, P. Richtárik, and M. Safaryan, "On biased compression for distributed learning," *Journal of Machine Learning Research*, vol. 24, no. 276, pp. 1–50, 2023.

[38] N. Shlezinger, M. Chen, Y. C. Eldar, H. V. Poor, and S. Cui, "Uveqfed: Universal vector quantization for federated learning," *IEEE Transactions on Signal Processing*, vol. 69, pp. 500–514, 2020.

[39] H. Guo, Y. Zhu, H. Ma, V. K. Lau, K. Huang, X. Li, H. Nong, and M. Zhou, "Over-the-air aggregation for federated learning: Waveform superposition and prototype validation," *Journal of Communications and Information Networks*, vol. 6, no. 4, pp. 429–442, 2021.

[40] R. Durrett, *Probability: theory and examples*. Cambridge university press, 2019, vol. 49.

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

[42] D. Nagaraj, X. Wu, G. Bresler, P. Jain, and P. Netrapalli, "Least squares regression with markovian data: Fundamental limits and algorithms," *Advances in neural information processing systems*, vol. 33, pp. 16 666–16 676, 2020.

[43] M. Nakagami, "The m-distribution—a general formula of intensity distribution of rapid fading," in *Statistical methods in radio wave propagation*. Elsevier, 1960, pp. 3–36.

[44] G. D. Durgin, T. S. Rappaport, and D. A. De Wolf, "New analytical models and probability density functions for fading in wireless communications," *IEEE Transactions on Communications*, vol. 50, no. 6, pp. 1005–1015, 2002.

[45] A. Adibi, N. Dal Fabbro, L. Schenato, S. Kulkarni, H. V. Poor, G. J. Pappas, H. Hassani, and A. Mitra, "Stochastic approximation with delayed updates: Finite-time rates under markovian sampling," in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2024, pp. 2746–2754.

[46] G. Brockman, V. Cheung, L. Pettersson, J. Schneider, J. Schulman, J. Tang, and W. Zaremba, "Openai gym," *arXiv preprint arXiv:1606.01540*, 2016.

[47] S. Jing and C. Xiao, "Federated learning via over-the-air computation with statistical channel state information," *IEEE Transactions on Wireless Communications*, vol. 21, no. 11, pp. 9351–9365, 2022.

[48] K. Yang, T. Jiang, Y. Shi, and Z. Ding, "Federated learning via over-the-air computation," *IEEE Transactions on Wireless Communications*, vol. 19, no. 3, pp. 2022–2035, 2020.

[49] N. Zhang and M. Tao, "Gradient statistics aware power control for over-the-air federated learning," *IEEE Transactions on Wireless Communications*, vol. 20, no. 8, pp. 5115–5128, 2021.

[50] A. Y. Ng, D. Harada, and S. J. Russell, "Policy invariance under reward transformations: Theory and application to reward shaping,"

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 13:05:04 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Wireless Communications. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TWC.2025.3555941

in *Proceedings of the Sixteenth International Conference on Machine Learning*, 1999, pp. 278–287.

[51] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.



Nicolò Dal Fabbro obtained his Master's degree in Telecommunications Engineering and his PhD degree in Information Engineering from the University of Padova, Italy, in 2020 and 2023 , respectively. He is currently a postdoctoral researcher at the University of Pennsylvania, USA, where he is with the department of Electrical and Systems Engineering. He is the recepient of the GTTI (italian telecommunications and information technology group) best 2024 PhD thesis award and of one best dataset award from the IEEE. His research interests are in federated learning, multi-agent reinforcement learning and wireless sensing.



Aritra Mitra received the B.E. degree from Jadavpur University, Kolkata, India, in 2013, the M.Tech. degree from the Indian Institute of Technology Kanpur, Kanpur, India, in 2015, and the Ph.D. degree from Purdue University, West Lafayette, IN, USA, in 2020, all in Electrical engineering. He is currently an Assistant Professor with the Department of Electrical and Computer Engineering, North Carolina State University, Raleigh, NC, USA. He was a Postdoctoral Researcher at the University of Pennsylvania from 2020 to 2022. His current research interests include control theory, optimization, statistical signal processing, machine learning, and distributed algorithms. Dr. Mitra was a recipient of the University Gold Medal at Jadavpur University and the Academic Excellence Award at IIT Kanpur.



Robert W. Heath Jr. (S'96 - M'01 - SM'06 - F'11) is the Charles Lee Powell Chair in Wireless Communications in the Department of Electrical and Computer Engineering at the University of California, San Diego. He is also President and CEO of MIMO Wireless Inc. From 2020-2023 he was the Lampe Distinguished Professor at North Carolina State University and co-founder of 6GNC. From 2002-2020 he was with The University of Texas at Austin, most recently as Cockrell Family Regents Chair in Engineering and Director of UT SAVES. He authored "Introduction to Wireless Digital Communication" (Prentice Hall, 2017) and "Digital Wireless Communication: Physical Layer Exploration Lab Using the NI USRP" (National Technology and Science Press, 2012), and co-authored "Millimeter Wave Wireless Communications" (Prentice Hall, 2014) and "Foundations of MIMO Communication" (Cambridge University Press, 2018).

Dr. Heath has been a co-author of a number award winning conference and journal papers including recently the 2017 Marconi Prize Paper Award, the 2019 IEEE Communications Society Stephen O. Rice Prize, the 2020 IEEE Signal Processing Society Donald G. Fink Overview Paper Award, the 2021 IEEE Vehicular Technology Society Neal Shepherd Memorial Best Propagation Paper Award, and the 2022 IEEE Vehicular Technology Society Best Vehicular Electronics Paper Award. Other notable awards include the 2017 EURASIP Technical Achievement award, the 2019 IEEE Kiyo Tomiyasu Award, the 2021 IEEE Vehicular Technology Society James Evans Avant Garde Award, and the 2025 IEEE/RSE James Clerk Maxwell Medal. In 2017, he was selected as a Fellow of the National Academy of Inventors. In 2024, he was selected as a Fellow of the American Association for the Advancement of Science. He was a member-at-large on the IEEE Communications Society Board-of-Governors (2020-2022) and the IEEE Signal Processing Society Board-of-Governors (2016-2018). He was Editor-in-Chief of IEEE Signal Processing Magazine from 2018-2020. He is also a licensed Amateur Radio Operator, a Private Pilot, a registered Professional Engineer in Texas. He is an elected member of the National Academy of Engineers.



Luca Schenato received the Dr. Eng. degree in electrical engineering from the University of Padova in 1999 and the Ph.D. degree in Electrical Engineering and Computer Sciences from the UC Berkeley, in 2003. He held a post-doctoral position in 2004 and a visiting professor position in 2013-2014 at U.C. Berkeley. Currently he is Full Professor at the Information Engineering Department at the University of Padova. His interests include networked control systems, multi-agent systems, wireless sensor networks, distributed optimisation and synthetic biology. Luca Schenato has been awarded the 2004 Researchers Mobility Fellowship by the Italian Ministry of Education, University and Research (MIUR), the 2006 Eli Jury Award in U.C. Berkeley and the EUCA European Control Award in 2014, and IEEE Fellow in 2017. He served as Associate Editor for IEEE Trans. on Automatic Control from 2010 to 2014 and he is he is currently Senior Editor for IEEE Trans. on Control of Network Systems and Associate Editor for Automatica.



George J. Pappas is the UPS Foundation Professor and Chair of the Department of Electrical and Systems Engineering at the University of Pennsylvania. He also holds a secondary appointment in the Departments of Computer and Information Sciences, and Mechanical Engineering and Applied Mechanics. He is member of the GRASP Lab and the PRECISE Center. He has previously served as the Deputy Dean for Research in the School of Engineering and Applied Science. His research focuses on control systems, robotics, formal methods and machine learning for safe and secure autonomous systems. He has received various awards such as the Antonio Ruberti Young Researcher Prize, the George S. Axelby Award, the O. Hugo Schuck Best Paper Award, the NSF PECASE award, and the George H. Heilmeier Faculty Excellence Award, and numerous best paper awards. He is a Fellow of IEEE and IFAC. More than thirty alumni of his research group are now faculty in leading universities around the world.

Authorized licensed use limited to: University of Pennsylvania. Downloaded on October 29,2026 at 13:05:04 UTC from IEEE Xplore. Restrictions apply.
© 2025 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.