Consensus-based Decentralized Multi-agent Reinforcement Learning for Random Access Network Optimization

Myeung Suk Oh, Zhiyao Zhang, FNU Hairi, Alvaro Velasquez, and Jia Liu

Abstract

With wireless devices increasingly forming a unified smart network for seamless, user-friendly operations, random access (RA) multiple access control (MAC) layer design is considered a key solution for handling unpredictable data traffic from multiple terminals. However, it remains challenging to design an effective RA-based MAC layer protocol to minimize collisions and ensure transmission fairness across the devices. While existing multi-agent reinforcement learning (MARL) approaches with centralized training and decentralized execution (CTDE) have been proposed to optimize RA performance, their reliance on centralized training and the significant overhead required for information collection can make real-world applications unrealistic. In this work, we adopt a fully decentralized MARL architecture, where policy learning does not rely on centralized tasks but leverages consensus-based information exchanges across devices. We design our MARL algorithm over an actor-critic (AC) network and propose exchanging only local rewards to minimize communication overhead. Furthermore, we provide a theoretical proof of global convergence for our approach. Numerical experiments show that our proposed MARL algorithm can significantly improve the RA network performance compared to other baselines.

Index Terms

Wireless networks, Random access, Multi-agent reinforcement learning, Average consensus, Actor-critic

M. S. Oh, Z. Zhang, and J. Liu are with The Ohio State University, Columbus, OH, USA (e-mail: oh.746@osu.edu, zhang.15178@osu.edu, liu@ece.osu.edu).

FNU Hairi is with University of Wisconsin-Whitewater, Whitewater, WI, USA (email: hairif@uww.edu).

A. Velasquez is with University of Colorado Boulder, Boulder, CO, USA (email: alvaro.velasquez@colorado.edu).

I. Introduction

1) Background and Motivations: Originating from the ALOHA protocol in the 1970s [1], [2] and going through the subsequent evolutions of carrier sensing multiple access (CSMA) technologies [3], random access (RA) multiple access control (MAC) layer design has been woven into the current fabric of the Internet, becoming an indispensable component of generations of Ethernet and WiFi network standards (e.g., CSMA/CA widely adopted in Wi-Fi and Long Term Evolution Licensed Assisted Access (LTE-LAA) [4] standards). The sustained popularity of RA-based MAC design is primarily due to its simplicity and flexibility in channel utilization when different user devices interact. Specifically, in an RA-based network as shown in Fig. 1, multiple devices share the same communication channel for data transmission. The devices do not rely on a centralized control for data traffic management (e.g., transmission scheduling handled by a server) but make independent decisions on when to transmit their data. This *decentralized* approach allows much simpler channel access management by eliminating the need for allocating a dedicated channel for each device. RA is particularly effective in environments with a large number of devices that are connected with random data transmissions. Moreover, through various emerging networking paradigms (e.g., Internet of Things (IoT) [5], machine-type communications [6], and smart grid communication infrastructure [7]), it is expected that RA-based MAC design will continue to drive the development of future large-scale networks with seamless and user-friendly operations.

However, compared to its controlled access MAC layer counterparts (e.g., TDMA, FDMA, and CDMA), the fundamental challenge in RA-based MAC layer design lies in how to avoid and resolve collisions caused by the contentions of the network devices. As shown in Fig. 1 in the context of WiFi networks, when two or more devices simultaneously attempt for data transmission, a collision occurs. If not treated appropriately, excessive collisions could significantly decay the network performance by wasting scarce network resources and outweigh the benefits of using a RA-based MAC protocol. Over the decades, a significant amount of effort has been dedicated to the optimization of RA-based MAC design. Unfortunately, to date, RA-based MAC layer performances remain far from satisfactory. In the literature, although there exists a large body of works on optimizing RA-based MAC performances (e.g., throughput, delay, and fairness of RA-based networks), these works are either i) too heuristic to provide any optimality guarantee, or ii) too heavily rely on idealized analytical models that often fail to capture real-world complexities (see Section II for more in-depth discussions). Meanwhile,

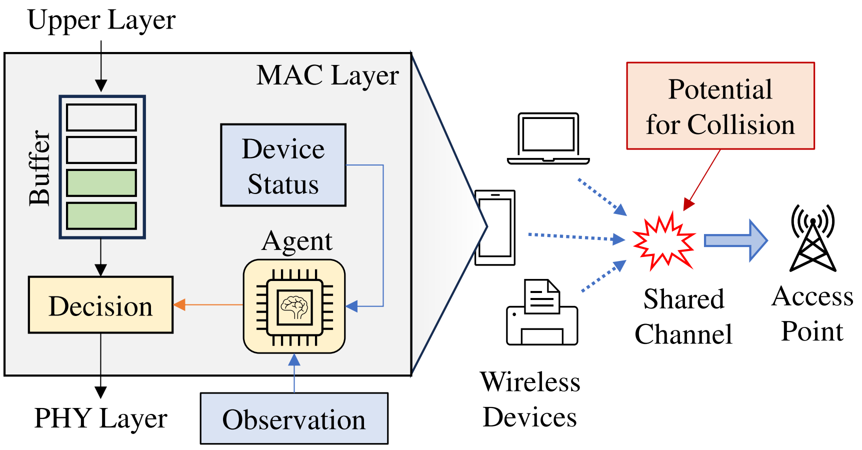


Fig. 1. A visual illustration of random access (RA) network, where multiple wireless devices contend for access to a shared communication channel. Each device employs its own decision-making mechanism to determine when to transmit a packet stored in its buffer.

with the advent of the IoT era with increasingly ubiquitous network access, the importance of RA-based MAC control and optimization is poised to increase for years to come. This widening gap between RA-based MAC optimization research and the rising demand for highly efficient RA-based networks motivates us to revisit this vital topic.

2) The AI/ML Approach and Limitations: With the rapid advancement in fields of artificial intelligence (AI) and machine learning (ML) in recent years, AI/ML optimization strategies have been increasingly explored and applied in RA-based MAC design (e.g., [8], [9], [10]). The rationale for using ML/AI approaches is their potential to address the limitations of idealized analytical models in RA-based MAC research through data-driven insights, while capitalizing on the rapid progress in AI/ML. Specifically, due to the decentralized nature of RA-based networks, an RA-based MAC layer optimization problem can be viewed as a distributed decision-making problem in the form of multi-agent reinforcement learning (MARL). As a subfield of reinforcement learning (RL), MARL focuses on scenarios where multiple agents coexist in a shared environment. These agents learn by making interactions and receiving feedback from the environment, and their ultimate goal is to find a joint policy that maximizes the overall reward.

Inspired by this architectural fit, several MARL-based approaches have been proposed to tackle the RA-based MAC layer optimization problem [11], [12], [13], [14], [15]. These methods commonly adopt the centralized training and decentralized execution (CTDE) framework to securely achieve convergence in MARL. In this framework, a deep neural network (DNN) for policy evaluation is trained centrally, where information from each device is aggregated to form a global representation. Once trained, the network coordinates with locally distributed decision-making DNNs, helping each device to make independent decisions on its action.

While CTDE-based MARL has demonstrated its effectiveness in RA-based MAC layer optimization and has successfully enhanced network performances, its dependence on centralized tasks can render the framework rather impractical for many real-world applications that do not assume the presence of a central entity. Even in networks that support such an entity, collecting the necessary information (e.g., local observations and actions) from all devices for centralized training can incur significant communication overhead as the network scales. Moreover, some RA-based networking scenarios may prioritize data privacy and network security, and thus do not prefer involving centralized tasks due to the risk of having increased vulnerability to potential attacks and reduced fault tolerance.

3) Our Approach and Contributions: To overcome the challenges mentioned above, we take an **decentralized** MARL approach to solve the RA-based MAC layer optimization problem. Specifically, we consider a fully decentralized MARL architecture, where policy learning is executed without the aid of centralized tasks. In particular, we propose to leverage the average consensus mechanism, in which devices locally communicate with their neighbors for information exchange to facilitate global convergence in MARL. We specifically design our fully decentralized MARL algorithm over actor-critic (AC) learning, which has been widely applied in RL tasks with various architectures (e.g., advantage actor-critic (A2C) [16], deep deterministic policy gradient (DDPG) [17], and proximal policy optimization (PPO) [18]). We show that our consensus-based fully decentralized MARL framework achieves a provable finite-time convergence rate guarantee as the CTDE approaches, while avoiding the limitations of CTDE approaches in practice. Our main contributions are summarized as follows.

- We formulate a new fully decentralized MARL framework for RA-based MAC layer optimization problem, where we carefully design our reward function with simple device parameters such that maximizing the reward naturally improves the total network throughput while ensuring fairness across all agents.
- Different from existing works that adopt a CTDE approach, our consensus-based fully decentralized MARL approach does not require centralized procedures but instead relies only on local information exchanges between neighboring devices to achieve global convergence. The proposed algorithm is thus applicable in RA-based MAC scenarios where (i) centralized

controller is not available; and (ii) scalability, privacy and security become critical aspects.

- We present a theoretical analysis demonstrating that our fully decentralized AC algorithm with local reward sharing can converge to a fixed point. Our analysis provides finite-time convergence rates for both the actor and critic. A key distinction from existing analyses is that our analysis reflects on consensus solely applied to the local rewards.
- We conduct extensive numerical experiments to evaluate the performance of our consensus-based fully decentralized MARL algorithm. Through comparisons with baseline methods, we show that our algorithm significantly improves RA-based MAC layer performances while ensuring fairness across devices.

The remainder of this paper is organized as follows. Section II reviews related works on RA-based MAC optimization. In Section III, we present the system model for our RA network. Section IV provides implementation details of our consensus-based decentralized MARL algorithm. We conduct numerical experiments in Section V, and Section VI concludes the paper.

II. Related Work

In this section, we provide a high-level overview of i) traditional RA-based MAC layer optimization and ii) state of the art of recent AI/ML-based RA-based MAC layer optimization approaches.

1) Traditional RA-based MAC Layer Optimization: RA techniques can generally be categorized into two types: sensing-free and sensing-based. In sensing-free RA, devices do not monitor the channel before transmitting data. Protocols like ALOHA [1] and its variants (e.g., slotted ALOHA) belong to this category. Since sensing-free RA allows a device to initiate data transmission even when the channel is already in use, the chance of collision remains high. Nonetheless, sensing-free RA has been recognized as a promising strategy for satellite communications [19], [20] and multi-hop mobile networks [21]. In contrast, sensing-based RA employs the listen-before-talk (LBT) mechanism, where each device monitors the channel for idleness before attempting data transmission. Sensing-based RA thus significantly reduces the collision rate compared to sensing-free cases. A widely used sensing-based RA protocol nowadays is CSMA/CA [3]. In addition to LBT mechanism, CSMA/CA incorporates a random backoff time that delays each transmission attempt for further preventing collisions. A heuristic way of generating backoff times in practice is binary exponential backoff (BEB), in which the size of the contention window (i.e., the range from which the backoff time is randomly selected) doubles after each collision.

While CSMA/CA is effective at reducing collisions, its heuristic way of setting backoff times can lead to increased transmission delays and less efficient channel usage and it is unclear whether CSMA has any theoretical performance guarantee Rather surprisingly, the seminal work in [22] that CSMA can be *"throughput-optimal"* if one can adjust the backoff time and contention window size appropriately based on the queueing backlog at each node. Since then, early 2010s have witnessed an intensive line of research on "queue-length-based adaptive CSMA" for RA-based MAC layer optimization (see, e.g., [22], [23], [24], [25], [26], [27] and their follow-ups). However, many of these theoretical studies relied on somewhat idealistic analytical models that often fail to capture real-world complexities. Moreover, many of these algorithms suffer from poor delay and fairness as the network size increases.

2) AI/ML Approaches for RA-Based MAC Optimization: To address the challenges above in adaptive CSMA/CA, several RL-based approaches have been proposed. For example, authors in [28] proposed Q-learning-based contention window selection algorithms for each cooperative and non-cooperative setting to maximize the total throughput while satisfying the fairness constraints. In [29], deep RL based on soft actor-critic (SAC) and long short-term memory (LSTM) models was utilized to dynamically adjust the device waiting time and optimize the network throughput. These approaches have been proven effective in enhancing RA performance. However, as CSMA/CA fundamentally depends on probabilistic transmissions of each device, performance improvement is still limited by its nature of stochastic operation.

Another approach in RA optimization is to develop a deterministic transmission policy for each participating device, for which several MARL-based strategies have been proposed. In [11], a deep Q-network was adopted to make transmission decisions for each RA device with an aim to maximize the generalized α-fairness objective. This approach was later extended to account for an imperfect wireless channel in which feedback signals for information collection can be corrupted [12]. The work in [13] employed a federated learning framework to implement distributed policy learning in RA networks, where each device is equipped with a DNN for decision-making. Furthermore, QMIX and multi-agent PPO algorithms were explored in [14] and [15], respectively, to implement MARL-based RA and improve network performance. Although these methods have shown promising results in RA optimization, they utilize the CTDE framework, which requires the existence of a central entity capable of handling large communication overhead for information collection. This requirement may not be practical in many real-world RA scenarios, especially where security and data privacy are of great concern. This motivates us to consider a fully decentralized architecture and develop a consensus-aided MARL algorithm that performs RA optimization in a more scalable and robust manner.

III. System Model and Problem Formulation

We consider a time-slotted RA scenario over a finite time horizon T, where each time slot is indexed by $t = 1, 2, \dots, T$. There are N devices in the network contending for channel access to transmit data packets either to an access point (AP) or to its intended receiver. To model a fully decentralized network, we do not assume the presence of an entity that is responsible for tasks like global information collection and centralized training. Instead, each device is aware of nearby devices and capable of exchanging information with them. We use a graph $\mathcal{G} = (\mathcal{N}, \mathcal{E})$ to represent the network, where $\mathcal{N} = \{1, 2, \dots, N\}$ is the set of devices and $\mathcal{E}$ denotes the edge set [30], [31], [32].

Each device $i \in \mathcal{N}$ is equipped with an internal packet buffer of size $Q_{\max}$. When a packet arrives at the MAC layer, it is first queued in the buffer and transmitted on a first-come first-served basis. Here, we use $0 \leq q_i^{(t)} \leq Q_{\max}$ to represent the number of packets in the buffer of device i at time slot t. Similar to CSMA/CA, our RA network operates on a LBT mechanism, where each device first checks on the channel status before transmitting any packets. If there are packets in a device's buffer, i.e., $q_i^{(t)} > 0$, the device enters the clear channel assessment (CCA) phase to check if the channel is idle. If the channel remains idle for an amount of time that is sufficiently long, the device considers the channel to be clear and decides whether to transmit its packet. If the device decides to *wait*, it simply returns to the CCA phase without further action; otherwise, if the device chooses to *transmit*, a single packet is transmitted over the channel.

Upon a successful transmission (i.e., no collision occurs and the AP securely receives the packet), an acknowledgment (ACK) is sent by the intended receiver after some prefixed delay to confirm the successful transmission. The device then returns to the CCA phase to prepare for transmitting the next packet. In the event of collision, the intended receiver does not send an ACK, indicating a failed transmission. Then, the device waits for the next opportunity to retransmit the packet. In Figure 2, we provide a flowchart summarizing the overall RA procedure.

To quantify the performance of our RA network, we define $r_i^{(t)}$ as the amount of reward for successfully transmitting packets from device i by time slot t. We also define $l_i^{(t)}$ as the number

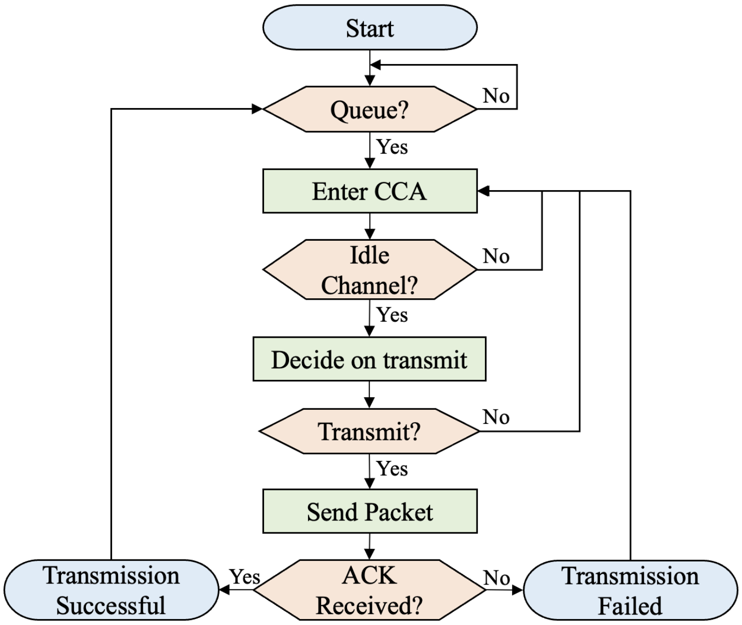


Fig. 2. A flowchart showing the overall RA steps done by each device. The procedure follows the LBT mechanism. The success of transmission is determined based on the reception of an ACK packet.

of time slots elapsed since device i's last successful transmission [14], which can be interpreted as the anticipated packet delay if device i successfully transmits a packet at time slot i. Our goal is to maximize the total network reward, which is given by

$$\frac{1}{T}\sum_{i=1}^{N} r_i^{(T)}, \tag{1}$$

while ensuring fairness across the devices. Note that maximizing throughput can be simply achieved by assigning higher transmission priorities to particular devices such that collisions never occur. However, such a policy may lead to significant imbalances in fairness. The problem becomes non-trivial when both throughput and fairness must be considered simultaneously, i.e., maximizing (1) while keeping $l_i^{(t)}$ at a similar level for all i.

IV. The Proposed Consensus-Based Fully Decentralized MARL Approach

In this section, we first present the fundamentals of MARL to provide background information. Then, we define the MDP formulation for our RA-based MAC layer optimization. Lastly, we will introduce our consensus-based decentralized MARL algorithm for RA-based MAC layer optimization.

A. MARL: A Primer

In MARL, each agent interacts with the environment through actions and observes the state and reward signals that represent the quality of the taken action. Here, the actions performed by the agents are coupled and jointly impact the next state. An MARL problem can be mathematically described by a Markov decision process (MDP) characterized by a 4-tuple: $\{\mathcal{S}, \{\mathcal{A}_i\}_{i\in\mathcal{N}}, P, \{R_i\}_{i\in\mathcal{N}}\}$, where $\mathcal{S}$ is the global state space and $\mathcal{A}_i$ is the action set for agent i. $P : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \to [0, 1]$ is the state transition probability, where $\mathcal{A} = \prod_{i\in\mathcal{N}} \mathcal{A}_i$ is the joint action set of all agents. Lastly, $R_i : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$ is the local reward function for agent i.

In time t, each agent i takes an action $a_i^{(t)} \in \mathcal{A}_i$ based on a policy π_{θ_i} parameterized by θ_i. Since the actions of agents are coupled, the state is transitioned to $s^{(t+1)}$ based on $P(s^{(t+1)}|s^{(t)}, \boldsymbol{a}^{(t)})$ where $\boldsymbol{a}^{(t)} = [a_1^{(t)}, a_2^{(t)}, \dots, a_N^{(t)}]$ is the joint action. Similarly, the instantaneous local reward of agent i for the action taken at time t can be expressed as $r_i^{(t)} = R_i(s^{(t)}, \boldsymbol{a}^{(t)})$. Let $\theta = [\theta_1^\top, \theta_2^\top, \dots, \theta_N^\top]^\top$be the joint weight vector of all N actors, the objective of MARL is to find an optimal θ to maximize the expected infinite-time discounted global reward, which can be written as:

$$J(\theta) := \mathbb{E}_{\pi_\theta}\left[\sum_{t=0}^{\infty} \gamma^t \bar{r}^{(t)}\right], \tag{2}$$

where $\bar{r}^{(t)} = \frac{1}{N}\sum_{i=1}^{N} r_i^{(t)}$ is the global averaged reward and $\gamma \in [0, 1]$ is the discount factor. In MARL, a state value function is commonly used to evaluate a policy given by θ, which can be defined as:

$$V_{\pi_\theta}(s) = \mathbb{E}_{\pi_\theta}\Big[\sum_{t=0}^{\infty} \gamma^t \bar{r}^{(t)}\Big| s^{(0)} = s\Big]. \tag{3}$$

Since $V_{\pi_\theta}(s)$ is typically unknown, in the RL literature [16], $V_\theta(s)$ is often estimated through a temporal differential (TD) learning process referred to as "critic." Specifically, let $V_{w_i}(s)$ be the state value approximation function where w_i is the parameter of agent i's critic model. Then, the TD learning is bootstrapped by using the Bellman optimality principle as follows:

$$V_{w_i}(s) = \mathbb{E}_{\pi_\theta}\Big[\bar{r} + \gamma V_{w_i}(s')\Big]. \tag{4}$$

In RL and MARL, the policy improvement (referred to as "actor") can be facilitated by updates based on the policy gradient:

$$\nabla_{\theta_i} J(\theta) = \mathbb{E}_{s,a}[\nabla_{\theta_i} \log \pi_{\theta_i}^i (a_i|s) \cdot \mathrm{Adv}_\theta(s, a)], \tag{5}$$

where $\text{Adv}_\theta(s,a) = \bar{r} + \gamma V_\theta(s') - V_\theta(s)$ is the advantage function. Once the actor takes an action according to the current policy, the critic evaluates the policy based the acquired reward. Upon evaluation, the policy is improved using (5) so that the actor is able to take actions that lead to improved expected reward.

Compared to single-agent RL, the key differences in MARL are as follows [33]. First, since multiple agents make independent decisions simultaneously, the environment is never seen as stationary to an action of each individual agent. Second, due to the decentralized architecture, each agent may observe only a part of information available in the environment. Therefore, careful design of MARL framework is essential to achieve performance comparable to that of centralized learning.

B. The MARL Problem Formulation for RA-Based MAC Optimization

With the MARL preliminaries, we are now in a position to formulate our RA-based MAC layer optimization as an MARL problem. Toward this end, we first define the state in each time slot as:

$$s^{(t)} = \left(\left\{ \bar{q}_i^{(t)} \right\}_{i=1}^{N}, \left\{ \bar{l}_i^{(t)} \right\}_{i=1}^{N}, c^{(t)} \right), \tag{6}$$

where $\bar{q}_i^{(t)} = q_i^{(t)}/Q_{\text{max}}$ is the normalized packet queue of device i at time slot t, $\bar{l}_i^{(t)} = \lambda_i l_i^{(t)}$ is the normalized time delay since device i's last successful transmission, and $c^{(t)} \in \{0,1\}$ is the channel usage indicator. We formulate our state such that the entire status of our RA network is accurately perceived. In practice, each agent i can only observe the part of state $s^{(t)}$, which we denote using $o_i^{(t)} \in s^{(t)}$ as follows:

$$o_i^{(t)} = \left\{ \bar{l}_i^{(t)}, \left\{ \bar{l}_j^{(t)} \right\}_{j \in \mathcal{N} \setminus i}, c^{(t)} \right\}, \tag{7}$$

which strictly follows $o_i^{(t)} \in S^{(t)}$. Note that, for each device i, the first two components $\bar{q}_i^{(t)}$ and $\bar{l}_i^{(t)}$ are local information. As assumed in [11], [14], [15], we consider the time delays from other network devices, i.e., $\left\{ \bar{l}_j^{(t)} \right\}_{j \in \mathcal{N} \setminus i}$ to be observable since they are traceable by listening to ACK packets broadcast by the intended receiver. The last parameter $c^{(t)}$ is easily observable from monitoring the channel during the CCA phase.

For all devices, we use a discrete action space $\mathcal{A}_i = \{0,1\}$, where 0 represents the action of *wait* and 1 represents the action of *transmit*. In other words, we consider each device i can take one of the two discrete actions at time t, i.e., $a_i^{(t)} \in \{0,1\}$.

Similar to [11], [14], [15], we assume that each agent can store the M latest observations and actions and use them as a set of action-observation history. Let $\tilde{t}_{i,m}$ be the time when the m-th latest action was taken by device i. Then, the observation-action history of length M for device i can be formed as

$$\eta_{M,i}^{(t)} = \left\{ o_i^{(\tilde{t}_{i,M})}, a_i^{(\tilde{t}_{i,M})}, o_i^{(\tilde{t}_{i,M-1})}, a_i^{(\tilde{t}_{i,M-1})}, \ldots, o_i^{(\tilde{t}_{i,1})}, a_i^{(\tilde{t}_{i,1})} \right\}. \tag{8}$$

We aim to utilize (8) as an additional information in feeding both actor and critic to make their learning process to reflect the dynamics of RA environment upon deciding and evaluating the action.

We define our instantaneous local reward for device i at time slot t to be

$$r_i^{(t)} = -\left(\omega_1 \bar{l}_i^{(t)} + \omega_2 \bar{q}_i^{(t)} \right), \tag{9}$$

where ω_1 and ω_2 are scaling factors. We strictly define our reward function to be local, i.e., $r_i^{(t)}$ is not a function of the information from other agents, to reflect the condition of fully decentralized MARL. For the consensus step, we let $r_i^{(t)}$ to be exchanged across the devices through local communication links defined by $\mathcal{G}$.

As outlined in (2), the objective of our MARL is to maximize the long-term discounted global reward. When we consider (9) for the reward function in (2), we observe that our MARL is designed to focus on minimizing both the packet delays and the packet queues within the RA network. Unlike the rewards defined in [11], [14], [15], which assume through CTDE framework that either (i) devices have access to the global reward or (ii) the global reward is directly computed by the central entity, we avoid incorporating action-dependent scores (e.g., assigning negative values upon collision) but instead use status-dependent scores to formulate our reward. This is to prevent inefficient learning that may result from correlating locally exchanged rewards to each device's local action only in a simplistic manner. We rather aim for our global reward $\bar{r}^{(t)}$ to focus on reflecting the condition of the RA network, which can directly be interpreted as the state value in our AC learning framework.

Moreover, we adopt our reward design in (9) for the following reasons. Considering a finite time-horizon T, let $\mathcal{X}_i(T)$ denote the set of time slots at which device i successfully transmits a packet over the T time slots. Let $v_i^{(t)}$ represent the instantaneous throughput of node i in time slot t. Then, the network throughput as given in (1) can be rewritten as

$$\frac{1}{T} \sum_{i=1}^{N} v_i^{(T)} = \sum_{i=1}^{N} \frac{|\mathcal{X}_i(T)|}{\sum_{t \in \mathcal{X}_i(T)} l_i^{(t)}}. \tag{10}$$

Note that the denominator is fixed at T regardless of how often successful transmissions occur. Therefore, the only factor influencing throughput is the numerator, which can be maximized by increasing the number of successful transmissions. Second, we include packet queues as part of the reward to ensure that our MARL reflects the need to prioritize devices with high transmission urgency and thus prevent packet loss due to buffer saturation. Since the queue size is not included in the observation $o_i^{(t)}$, we allow the MARL to reflect each device's urgency in a stochastic manner for decision-making without directly correlating it with a deterministic action.

C. The Proposed Consensus-Based Decentralized MARL Algorithm

The overall architecture of our consensus-based decentralized MARL for RA network optimization is illustrated in Fig. 3. Each device $i \in \mathcal{N}$ updates its own actor-critic (AC) models trained using local experiences. To store and use the action-observation history at each learning step, each device maintains a history buffer to records the past observation-action pairs. As described in Section III, connected devices can exchange of their local information with each other. In our consensus-based decentralized MARL algorithmic design, we allow each node to share local rewards with its neighbors.

Our proposed consensus-based fully decentralized MARL for the RA-based MAC layer of a N-device network is summarized in Algorithm 1. As discussed in Section III, each device obeys the LBT mechanism (Fig. 2) and constantly monitors the channel. Once the channel is assessed to be clear, the device takes an observation $o_i^{(t)}$ and makes a transmission decision $a_i^{(t)}$ using its local policy conditioned on the stored action-observation history and current observation **(Lines 6 - 11)**. Depending on the transmission result (whether or not collision has occurred), each device updates its status. The above step is repeated for the span of T time slots.

Since most of time slots are occupied by the RA protocol steps such as CCA, packet transmission, and waiting upon ACK, the actual time slots related to MARL procedure are confined. Hence, we consider our MDP to only progress over time slots where an action is taken by the devices.

Each time the devices acquire enough information to perform the MARL step (i.e., weight parameter update using gradient descent), the *consensus process* **(Lines 19 - 22)** first starts from an initial step. Each consensus step consists of G rounds of communication, where weight-based averaging is performed in each round. Then, each device updates its actor and critic parameters

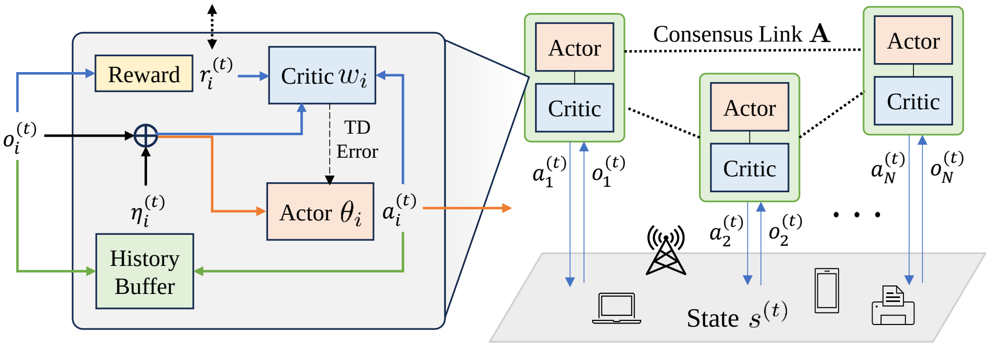


Fig. 3. A visual representation of our fully decentralized MARL framework. Both the actor and critic are trained in a decentralized manner. There exist local communication links among devices that allow local reward sharing for consensus.

after computing the TD error (**Lines 24 - 29**). As a final step, both the current observation and action are stored in the history buffer.

A key **novelty** of our algorithm is that we only exchange the local reward $r_i^{(t)}$ during the consensus step. Note that this differs from many existing decentralized MARL algorithms (e.g., [31], [34], [32], [35]), where the entire weights of critic, i.e., w_i are exchanged during each consensus step to ensure convergence. Given that DNNs are often employed to estimate the value function, the number of parameters that need to be exchanged between agents is huge. In RA networks, this can be an unrealistic requirement since resource allocated for each established link can often be much limited. By contrast, in our algorithm, each agent only needs to exchange its reward value in each time slot, which is only a *scalar*. Thus, our approach significantly reduces the amount of information exchanged across the network and renders our approach more practical. In Section IV-D, we will rigorously show the convergence performance of Algorithm 1. We also note that our proposed algorithm could also of independent interest in the MARL literature.

To further demonstrate the above advantage of our algorithm, we compare our algorithm with one of CTDE in the amount of information exchanged for each consensus step. In CTDE, a central entity must collect information from all participating devices to train the centralized critic, which usually takes a set of history, currently taken observation-action pair, and reward from every device to compute the gradient and update its weights. Let us define D_o, D_a, and D_r to be the dimensions of observation, action, and reward, respectively, for each given algorithm. Then, for centralized critic, the total number of parameters that must be collected for each learning step is given by $N[M(D_o + D_a) + D_r]$. On the other hand, our consensus-based approach requires

Algorithm 1: The Consensus-based Fully Decentralized MARL for the RA-based MAC Layer Optimization.

```
Input: agent set $\mathcal{N}$, neighbor sets $\{\mathcal{N}_i\}_{i\in\mathcal{N}}$, time horizon length $T$, history length $M$, consensus weight
    matrix $\mathbf{A}$, consensus iteration count $G$, actor rate $\alpha$, critic rate $\beta$
Initialize: actor weights $\theta_i$, critic weights $w_i$, transmit flag $\mathrm{f}_t = \text{False}$, and ready-for-update status
    $u_i = \text{False}$ for all $i \in \mathcal{N}$
for $t = 1, 2 \ldots, T$ do
    for $i \in \mathcal{N}$ do
        Update $q_i^{(t)}$, $l_i^{(t)}$, $\mathrm{f}_i$, $c^{(t)}$
        if $q_i^{(t)} > 0$ & $c^{(t)} = 0$ then
            Acquire observation $o_i^{(t)}$ and reward $r_i^{(t)}$
            Select $a_i^{(t)} \sim \pi_{\theta_i}(\,\cdot\,|\{\eta_{M,i}^{(t)}, o_i^{(t)}\})$
            if $a_i^{(t)} = 1$ then
                $\mathrm{f}_t \leftarrow \text{True}$
            $u_i \leftarrow \text{True}$
    for $i \in \mathcal{N}$ do
        if $f_t = \textit{True}$ then
            $c^{(t)} \leftarrow 1$; Transmit a packet
        if ACK received then
            $l_i^{(t)} \leftarrow 0$
    if $u_i = \textit{True}, \forall i \in \mathcal{N}$ then
        $\tilde{r}_{i,0} \leftarrow r_i^{(t)}$ for all $i \in \mathcal{N}$
        for $g = 1, 2 \ldots, G$ do
            $\tilde{r}_{i,g} \leftarrow \sum_{j\in\mathcal{N}_i} a_{ij}\tilde{r}_{j,g-1}$ for all $i \in \mathcal{N}$
        $\tilde{r}_i^{(t)} \leftarrow \tilde{r}_{i,G}$ for all $i \in \mathcal{N}$
        for $i \in \mathcal{N}$ do
            $\delta_i \leftarrow \tilde{r}_i^{(t)} + \gamma V_{w_i}(\{\eta_{M,i}^{(t)}, o_i^{(t)}\}) - V_{w_i}(\eta_{M+1,i}^{(t)})$
            $w_i \leftarrow w_i - \beta\delta_i \cdot \nabla V_{w_i}(\eta_{M+1,i}^{(t)})$
            $\delta_i \leftarrow \tilde{r}_i^{(t)} + \gamma V_{w_i}(\{\eta_{M,i}^{(t)}, o_i^{(t)}\}) - V_{w_i}(\eta_{M+1,i}^{(t)})$
            $\theta_i \leftarrow \theta_i + \alpha\delta_i \cdot \nabla \log \pi_{\theta_i}(a_i^{(t)}|\eta_{M+1,i}^{(t)})$
            Store $o_i^{(t)}$ and $a_i^{(t)}$ in the history buffer
            $u_i \leftarrow \text{False}$
Output: $\theta_i$ for all $i \in \mathcal{N}$
```

TABLE I

DIMENSION OF MDP PARAMETERS FOR DIFFERENT CTDE-BASED RA ALGORITHMS.

Algorithm	D_o	D_a	Dr
Guo [14]	$N+2$	1	2
Yu [12]	1	1	NM

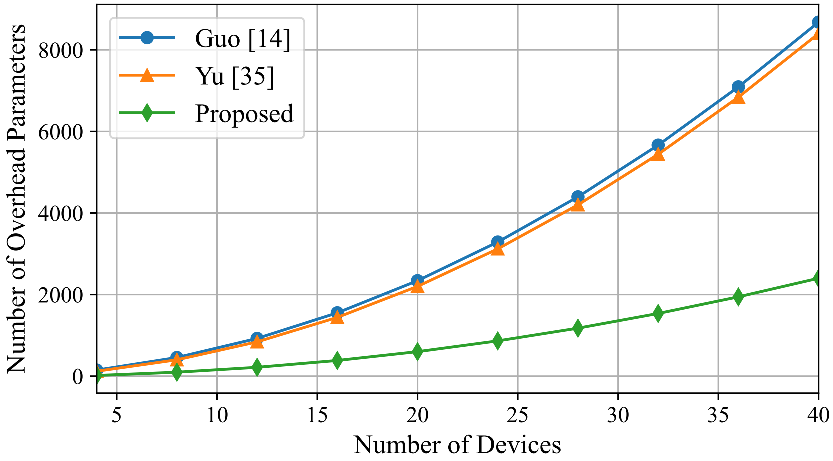


Fig. 4. A plot comparing the number of overhead parameters required to perform MARL. CTDE approaches require more parameters to be exchanged than our consensus-aided decentralized method.

each agent to exchange information with its neighboring devices, for which a total number of $GN\bar{d}D_r$ parameters need to be exchanged, where $\bar{d}$ is the average number of links per device. We can specifically set $G = \lceil \frac{0.5 \log \epsilon^{-1}}{\log \lambda_2^{-1}(A)} \rceil$ to achieve the normalized root mean squared error (RMSE) of ϵ for a given consensus weight matrix A [30]. We use two of the existing CTDE works [14], [12], where we provide the values of D_o, D_a, D_r in Table I, to compare the total number of required overhead to support a given number of devices. According to the result shown in Fig. 4, we can see that the fully decentralized case with local information exchange requires much *less* overhead to conduct MARL.

D. Theoretical Performance Analysis

Before providing our theoretical results regarding Algorithm 1, we provide some necessary assumptions in the following.

Assumption 1 (Bounded Reward)**.** *There exists a positive constant r_{max} such that $r_i^{(i)} \in [0, r_{max}]$ for any $t \geq 0, i \in \mathcal{N}$.*

Assumption 2 (Mixing Time)**.** *There exist a stationary distribution ζ for (s, a), and positive constants κ and $\rho \in (0, 1)$, such that $\sup_{s \in \mathcal{S}} \|P(s^{(t)}, a^{(t)}|s_0 = s) - \zeta(\theta)\|_{TV} \leq \kappa \rho^t, \forall t \geq 0$.*

Assumption 3 (Lipschitz Continuity)**.** *$J(\theta)$ is L_J-Lipschitz continuous w.r.t. θ, respectively, i.e., there exist some positive constant L_J such that, for any θ and θ', we have $|J(\theta) - J(\theta')| \leq L_J \|\theta - \theta'\|_2$.*

Assumption 4 (Consensus Matrix)**.** *The consensus weight matrix $\mathbf{A}$ is doubly stochastic. Additionally, for all $i, j \in \mathcal{N}$, there exists a positive constant $\nu > 0$ such that (i) $a_{ii} \geq \nu$ and (ii) $a_{ij} \geq \nu$ whenever devices i and j are connected.*

Now, we are ready to state our main theoretical results.

Theorem 1 (Finite-Time Critic Convergence Rate)**.** *Consider the iterative updates on $w_i^{(t)}, \forall i \in \mathcal{N}$ in Algorithm 1. For any given policy π_θ and $i \in \mathcal{N}$, it holds that*

$$\mathbb{E}\Big[\|w_i^{(t)} - w_i^{(t)*}\|^2\Big] \leq 2\left(\frac{c_1}{c_3}\right)^2 e^{-2c_2 G + 2c_4 t} + 2c_5(1 - c_6\beta)^{t-\tau(\beta)} + 2c_7\tau(\beta)\beta, \tag{11}$$

where $c_1 = 2\beta N r_{max}(1 + \nu^{-(N-1)})$, $c_2 = \ln(1 - \nu^{N-1})^{-1}$, $c_3 = (1+\gamma)\beta$, $c_4 = \ln(1 + (1+\gamma)\beta)$, $\{c_5, c_6, c_7\}$ are constants independent of step size β, and $\tau(\beta) = \mathcal{O}(\log(\beta^{-1}))$ is mixing time.

Proof Sketch. To derive the bound on the difference between $w_i^{(t)}$ and $w_i^{(t)*}$, we separate it into $w_i^{(t)} - \bar{w}_i^{(t)}$ and $\bar{w}_i^{(t)} - w_i^{(t)*}$, where $\bar{w}_i^{(t)} \triangleq \frac{1}{N}\sum_{i \in \mathcal{N}} w_i^{(t)}$. For the first term, we derive the iterative expression to show that it is a function of the initial consensus error on both local reward and critic parameter. We derive a bound in this term by showing that the error magnitude does not diverge with a careful selection of learning parameters. The analysis of the second term follows the approach provided in [36], which handles the case of single-agent AC. The full proof is provided in Appendix A. ■

We remark that the above result proves the convergence to TD fixed points for all agents even only sharing the rewards instead of critic parameters as related works [31], [34], [32]. From the first term in (11), it requires the communication rounds G for sharing rewards need to be sufficiently large. We further remark that for sample complexity, we ignore the log terms for simplicity and introduce the canonical $\tilde{O}(\cdot)$ notation that ignores log terms. For the RHS of (11) to be $\mathcal{O}(\epsilon)$, for any target threshold $\epsilon > 0$, we require $\beta = \tilde{\mathcal{O}}(\epsilon)$, $t = \tilde{\mathcal{O}}(\epsilon^{-1})$ and $G = \tilde{\mathcal{O}}(\epsilon^{-1})$. As a result, the sample complexity for the critic is $t = \tilde{\mathcal{O}}(\epsilon^{-1})$ while the sample complexity is $tG = \tilde{\mathcal{O}}(\epsilon^{-2})$. The sample complexity is on the same order of that in [37], [34], [32]. On the communication results of $\mathcal{O}(\epsilon^{-2})$, it may look worse than aforementioned references. For example, [34] claims to have $O(\log(\epsilon^{-1}))$. However, recall that these literature share d-dimensional critic parameter whereas we only share scalar rewards. So the scalar communication complexity of [34] requires $O(d\log(\epsilon^{-1}))$ vs $O(\epsilon^{-2})$. If $d = \Omega(\epsilon^{-2})$, then our proposed algorithm actually performs more efficiently. For instance, suppose an MDP with huge $|\mathcal{S}|$, it often requires d to be high-dimensional. If we choose $d = 10^3$ then with $\epsilon = 0.1$, our proposed algorithm provides an order-wise more communication efficient approach.

Theorem 2 (Convergence Rate of Decentralized MARL Algorithm with Local Reward Consensus)**.** *Consider the AC algorithm in Algorithm 1. With step-size set as* $\alpha = \frac{1}{4L_J}$*, it holds that*

$$\mathbb{E}\left[\|\nabla_\theta J(\theta^{(\hat{T})})\|^2\right] \leq \frac{16L_J r_{\mathit{max}}}{T(1-\gamma)} + 18N(1+\gamma)^2 \frac{\sum_{t=1}^{T} \|w_i^{(t)} - w_i^{(t)*}\|^2}{T}$$

$$+ 72N^3 r_{\mathit{max}}^2 \left((1+\nu^{-(N-1)})(1-\nu^{N-1})^G\right)^2 + 18(1+\gamma)^2 \xi_{\mathit{approx}}^{\mathit{critic}} + 72N(r_{\mathit{max}} + (1+\gamma)R_w)^2, \quad (12)$$

where $\hat{T}$ *is uniformly sampled from* $\{1, \cdots, T\}$ *and* R_w *is a constant that is independent of* T*.*

Proof Sketch. We first express the gradient step into an inequality form using descent lemma according to Lipschitz property. After rearranging the terms, we derive the bound on the $\|\nabla_\theta J(\theta^{(t)})\|^2$ and decompose it into several error terms. We derive the upper bound of each term and show that its magnitude can be efficiently controlled to converge over the given time steps T. For the error that arises from using sub-optimal critic, we apply Theorem 1 to show its convergence. For the error that is due to imperfect global reward consensus, we show its convergence based on the doubly stochastic property of consensus weight matrix. The full proof is provided in Appendix B. ■

Based on Theorem 2, we ensure that the output policy of our Algorithm 1 converges to the neighborhood of some stationary point at a rate of $\mathcal{O}(1/T)$.

V. Numerical Experiments

A. Experimental Settings

We conduct a set of numerical experiments to evaluate our proposed MARL algorithm to optimize the IEEE 802.11 CSMA/CA MAC layer. We consider $N = 4$ devices participating in RA over $T = 600$ time slots. As mentioned in Section III, all devices operate under the LBT mechanism and hence follow the RA steps shown in Fig. 2. Following the IEEE 802.11 protocol, we set SIFS and DIFS to take 2 and 4 time slots, respectively, where each time slot is assumed to be 9 us long. In addition, we assume that each data packet size is $1,500$ bytes and it takes 10 time slots for all devices to transmit the packet over the wireless channel. We set the ACK signal transmission to be 4 time slots. Moreover, we consider $\lambda_i = \frac{1}{30}$ for all $i \in \mathcal{N}$.

For the topology of graph $\mathcal{G}$, we use the Watts-Strogatz graph model [38] where each device connected to one neighboring device with no rewiring probability. In generating the consensus weight matrix $\mathbf{A}$, we apply the equal weight for each device's established links, i.e., for each device $i \in \mathcal{N}$, $a_{ij} = \frac{1}{|\mathcal{N}_i|}$ for all $j \in \mathcal{N}_i$.

For our proposed consensus-based MARL algorithm, both actor and critic use a multi-layer perceptron (MLP) network with the width of 128 and depth of 5. While we set each layer of the actor network to use ReLU activation, we do not apply any activation function on the critic network to satisfy our linear approximation on the state value function, i.e., $V_{w_i}(s^{(t)}) = \phi^\top(s^{(t)})w_i, \forall i \in \mathcal{N}$. We set the length of the observation-action history to $M = 4$ to allow our network to collect a moderate amount of past information. We use stochastic gradient descent (SGD) for both actor and critic weight updates with learning rates $\alpha = 0.004$ and $\beta = 0.003$, respectively. We train our network over 1200 episodes and take the average over 20 independent runs. For performance comparison, we consider the following RA baselines:

- **RA with a fixed transmission probability (RA-P)**: As considered in [11], [15], we consider a legacy RA protocol where each device transmits its packet based on a fixed transmission probability. According to the analysis in [39], the maximum throughput is achieved when the probability is set to $\frac{1}{N}$, which we set for our experiments.
- **RA with a fixed contention window (RA-FCW)**: Each device uses a backoff time randomly generated from a contention window of fixed size W_{cw}. We assume that the optimal value of W_{cw} is previously found via experiments, i.e., we set $W_{\text{cw}} = 16$ for our RA scenario of $M = 4$ devices.

- **RA with an adaptive contention window (RA-ACW)**: Each device employs the BEB mechanism, where the size of the contention window doubles after each collision. For each device, we set the initial size of contention window as $W_{\text{cw}} = 1$.
- **RA with an adaptive contention window (RA-CTDE)**: We consider a MARL algorithm with CTDE architecture. For the AC framework, CTDE is realized through a central critic, which collects information on observations, actions, and rewards from every device, and distributed actors who take local actions. For fair comparison, we use the same actor model on each device and a proportionally scaled critic model for centralized training.

For the given algorithms, we measure the following metrics for performance evaluation: the number of successfully transmitted packets (Pkt-T), the number of collisions occurred (Pkt-C), the number of lost packets due to buffer saturation (Pkt-L), total network throughput (TPut), the time delay between each successfully transmitted packet (Delay), and normalized gap to measure the fairness across the devices (N-Gap), which is defined as:

$$\text{N-Gap} = \frac{\max(\{x\}) - \min(\{x\})}{\max(\{x\})} \tag{13}$$

We consider throughput and packet delay for fairness evaluation.

B. Results and Discussion

1) Comparison of average performance: In Table II, we present a comparison of the average network performance for different RA algorithms. We observe that the non-RL methods (i.e., RA-P, RA-ACW, and RA-FCW) result in degraded performance across all metrics. This is primarily due to their probabilistic approach of optimizing performance. In contrast, both RA-CTDE and our algorithm show significantly improved performance, greatly reducing the number of collisions. Considering the simulation noise, their performance is nearly identical. This indicates that our decentralized MARL with local reward sharing is as effective as the CTDE method. Moreover, our algorithm is far more efficient in terms of overhead complexity and practical applicability.

Fig. 5 shows the dynamic change in throughput as the learning episode progresses. It is important to note that non-RL approaches display consistent results throughout the episode as they do not incorporate a process of learning. We can see that both RA-CTDE and our algorithm improve throughput in a similar manner. A significant improvement is observed during the first 200 episodes, primarily due to the reduction in transmission probability to avoid collisions.

TABLE II

AVERAGE NETWORK PERFORMANCE COMPARISON OVER DIFFERENT RA ALGORITHMS. THE PROPOSED ALGORITHM ACHIEVES BETTER PERFORMANCE THAN THE BASELINES.

Algorithm	Pkt-T	Pkt-C	Pkt-L	TPut (Mbps)	Delay (ms)
RA-P	4.52	5.79	5.12	40.216	1.160
RA-ACW	5.13	4.52	4.69	45.622	1.241
RA-FCW	5.62	1.56	4.10	49.969	0.948
RA-CTDE	**6.22**	0.49	3.62	**55.389**	**0.812**
Proposed	6.16	**0.32**	**3.49**	54.733	0.817

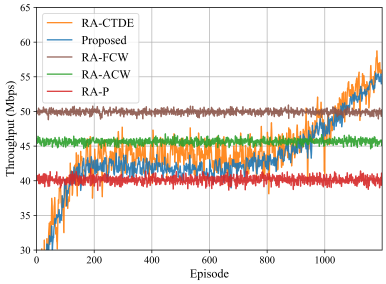


Fig. 5. A total throughput versus episode plot for different RA algorithms. While both MARL-based approaches display similar learning pattern, RA-CTDE exhibits greater variance in learning.

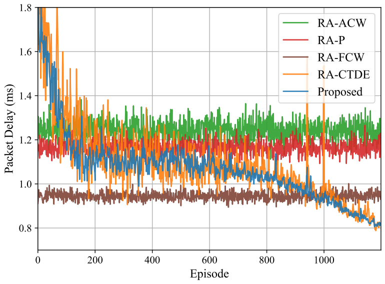


Fig. 6. A packet delay versus episode plot for different RA algorithms. RA-ACW yields the worst performance as it avoids collisions by increasing transmission delays.

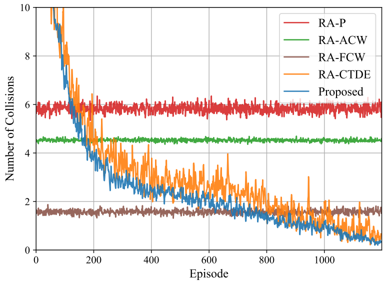


Fig. 7. A packet collision frequency versus episode plot for different RA algorithms. Both MARL-based approaches can reduce the rate of collision almost to zero.

After another approximately 600 episodes of steady performance, the devices converge to a deterministic policy that completely avoids collisions, resulting a steep increase in throughput. This confirms that the effectiveness of our consensus-based decentralized MARL framework in RA network optimization. Also, our result verifies that the average consensus on local rewards is sufficient for discovering optimal policies. Although the overall learning pattern is similar, it is worth noting that the variance in learning is higher for CTDE. This is due to the increased learning complexity of the centralized critic, which processes features of much larger dimensionality. Figs. 6 and 7 show the changes in packet delay and collision frequency as learning episodes progress, respectively. The trends in these figures follow a pattern similar to that in Fig. 5, where RA-CTDE and our algorithm show similar improvements. Similarly, the variance is greater for RA-CTDE in both metrics.

TABLE III

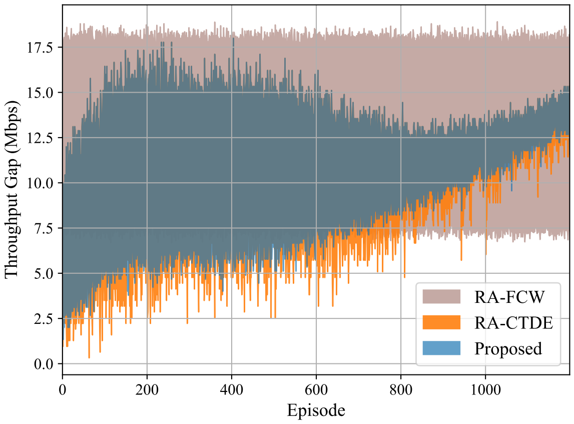
NETWORK FAIRNESS COMPARISON OVER DIFFERENT RA ALGORITHMS. THE PROPOSED ALGORITHM ACHIEVES BETTER FAIRNESS THAN THE BASELINES.

	TPut (Mbps)			Delay (ms)		
Algorithm	Min	Max	N-Gap	Min	Max	N-Gap
RA-P	5.536	14.953	0.629	0.689	1.882	0.634
RA-ACW	4.496	20.214	0.778	0.551	2.369	0.767
RA-FCW	7.409	17.990	0.588	0.598	1.476	0.595
RA-CTDE	12.778	14.921	0.144	0.755	0.876	0.138
Proposed	**12.867**	**14.767**	**0.129**	**0.765**	**0.872**	**0.123**

2) Comparison of fairness: In Table III, we provide the maximum and minimum values for throughput and packet delay, along with the normalized gap for each RA algorithm. Note that a lower gap indicates better fairness. As shown in the table, non-RL approaches exhibit a large gap in both throughput and packet delay. RA-ACW has the largest gap (more than a four-time difference in both throughput and delay) likely due to the adaptive contention window introducing unfair delays across devices. Meanwhile, both RA-CTDE and our algorithm greatly reduce the gap, achieving a much higher fairness level by the end of the learning process. Note that the improvement is approximately five times compared to the non-RL methods. The result highlights that, in addition to enhancing throughput performance, our approach can find a policy that maximizes fairness without relying on centralized tasks.

Fig. 8 presents a plot showing the change in throughput gap across the learning episodes. For clarity, we only include RA-FCW from the non-RL approaches. Although RA-FCW achieves the best fairness among the non-RL algorithms, the throughput gap remains significantly large, with a difference of approximately 10.5 Mbps. For both MARL-based approaches, the gap is initially similar to that of RA-FCW in the early episodes. However, after the 600th episode, the network begins to learn how to improve fairness. Starting from the 800th episode, both the maximum and minimum throughput start to improve together. Although RA-FCW achieves the highest absolute maximum throughput (around 18 Mbps), it does so at the expense of sacrificing throughput from other devices. In contrast, our algorithm learns to prioritize devices with larger packet delays, leading to an overall improvement in total throughput.

We show the change in delay gap across learning episode for different RA algorithms in Fig. 9. We observe that, for RA-FCW, the gap remains constant throughout across the episodes



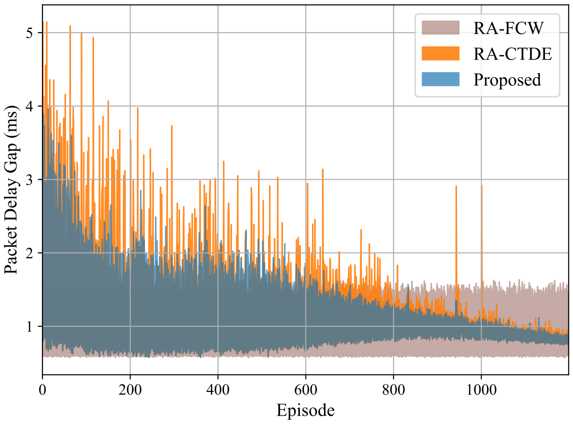


Fig. 8. A plot showing the gap between maximum and minimum throughputs. A smaller gap indicates better fairness.

Fig. 9. A plot showing the gap between maximum and minimum delays. A smaller gap indicates better fairness.

as no learning is involved. In the early stages of learning, both RA-CTDE and our algorithm exhibit significant worst-case delays. It is worth noting that the maximum delays in RA-CTDE are noticeably longer than those in our algorithm. By the end of the learning process, both approaches successfully reduce overall packet delays while also improving fairness. Despite lacking centralized training, our algorithm is able to successfully discover polices that result in a substantial improvement in RA network performance.

VI. Conclusion

In this paper, we considered the RA-based MAC layer optimization problem and proposed a fully decentralized MARL framework to find RA policies that minimize collisions and ensure transmission fairness across the devices. After carefully designing MARL parameters, we developed and theoretically analyzed the AC learning performance of our algorithm. Instead of leveraging centralized training, our algorithm uses the average consensus mechanism to achieve convergence in learning. Unlike many decentralized MARL strategies, our algorithm only exchanges local rewards, providing a significant advantage in overhead reduction. We showed through theoretical and numerical analysis that our proposed approach can attain network performance comparable to those of CTDE.

Appendix A
Proof of Theorem 1

For the convenience of notation, we use $\eta_i^{(t)}$ and $\eta_i^{(t+1)}$ to represent $\eta_{M+1,i}^{(t)}$ and $\{\eta_{M,i}^{(t)}, o_i^{(t)}, a_i^{(t)}\}$ in Algorithm 1, respectively. We first make the following assumption as in [31], [32].

Assumption 5. *Each agent $i \in \mathcal{N}$ uses a linear function to approximate its value function, i.e., $V_{w_i}(\eta_i^{(t)}) = \phi(\eta_i^{(t)})^\top w_i$ where $\phi(\eta_i^{(t)})$ is the uniformly bounded feature associated with $\eta_i^{(t)}$, i.e., $\|\phi(\eta_i^{(t)})\| \leq 1$.*

Let us define $w_i^{(t)*}$ to be the optimal weights for agent i's critic. To derive the upper bound on the difference between $w_i^{(t)}$ and $w_i^{(t)*}$, i.e., $\|w_i^{(t)} - w_i^{(t)*}\|$, for a given time index t, we separate the difference into $w_i^{(t)} - \bar{w}_i^{(t)}$ and $\bar{w}_i^{(t)} - w_i^{(t)*}$, where $\bar{w}_i^{(t)} \triangleq \frac{1}{N}\sum_{i\in\mathcal{N}} w_i^{(t)}$ and derive the bound on each of them.

We start on the first part of our theorem. Recall that each gradient update step yields

$$w_i^{(t)} = w_i^{(t-1)} + \beta\left(\tilde{r}_i^{(t-1)} + \gamma\phi^\top(\eta_i^{(t)})w_i^{(t-1)} - \phi^\top(\eta_i^{(t-1)})w_i^{(t-1)}\right)\phi(\eta_i^{(t-1)}) \tag{14}$$

$$\bar{w}_i^{(t)} = \bar{w}_i^{(t-1)} + \beta\left(\bar{r}^{(t-1)} + \gamma\phi^\top(\eta_i^{(t)})\bar{w}_i^{(t-1)} - \phi^\top(\eta_i^{(t-1)})\bar{w}_i^{(t-1)}\right)\phi(\eta_i^{(t-1)}) \tag{15}$$

where $\tilde{r}_i^{(t)} = [\mathbf{A}^G]_i r^{(t)}$, $\bar{r}^{(t)} = \frac{1}{N}\mathbf{1}^\top r^{(t)}$, and $r^{(t)} = [r_1^{(t)}, \cdots, r_N^{(t)}]^\top$. Note that $[\mathbf{A}^G]_i$ denotes the i-th row of matrix $\mathbf{A}^G$. We get the consensus error vector given by

$$\begin{aligned}
e_i^{(t)} &= w_i^{(t)} - \bar{w}_i^{(t)} && (16)\\
&= (w_i^{(t-1)} - \bar{w}_i^{(t-1)}) + \beta\phi(\eta_i^{(t-1)})\left(\left[[\mathbf{A}^G]_i - \frac{1}{N}\mathbf{1}^\top\right]r^{(t-1)}\right)\\
&\qquad + \beta\phi(\eta_i^{(t-1)})[\gamma\phi(\eta_i^{(t)}) - \phi(\eta_i^{(t-1)})]^\top(w_i^{(t-1)} - \bar{w}_i^{(t-1)}) && (17)\\
&= (w_i^{(t-1)} - \bar{w}_i^{(t-1)}) + \beta\phi(\eta_i^{(t-1)})r^{(t-1)\top}\left[[\mathbf{A}^G]_i^\top - \frac{1}{N}\mathbf{1}\right]\\
&\qquad + \beta\phi(\eta_i^{(t-1)})[\gamma\phi(\eta_i^{(t)}) - \phi(\eta_i^{(t-1)})]^\top(w_i^{(t-1)} - \bar{w}_i^{(t-1)}) && (18)\\
&= e_i^{(t-1)} + \beta C_i^{(t-1)}\left[[\mathbf{A}^G]_i^\top - \frac{1}{N}\mathbf{1}\right] + \beta B_i^{(t-1)}e_i^{(t-1)} && (19)\\
&= (\mathbf{I} + \beta B_i^{(t-1)})e_i^{(t-1)} + \beta C_i^{(t-1)}\left[[\mathbf{A}^G]_i^\top - \frac{1}{N}\mathbf{1}\right], && (20)
\end{aligned}$$

where $B_i^{(t)} = \phi(\eta_i^{(t)})[\gamma\phi(\eta_i^{(t+1)}) - \phi(\eta_i^{(t)})]^\top$ and $C_i^{(t)} = \phi(\eta_i^{(t)})r^{(t)\top}$. Note that (20) is a function of $e_i^{(t-1)}$. Hence, we can express $e_i^{(t)}$ in an iterative form:

$$e_i^{(t)} = \left[\prod_{x=0}^{t-1}(\mathbf{I} + \beta B_i^{(x)})\right]e_i^{(0)} + \beta\sum_{x=0}^{t-1}\left[\prod_{y>x}^{t-1}(\mathbf{I} + \beta B_i^{(y)})\right]C_i^{(x)}\left[[\mathbf{A}^G]_i^\top - \frac{1}{N}\mathbf{1}\right]. \tag{21}$$

Since $e_i^{(0)}$ is zero due to $w_i^{(0)} = \bar{w}_i^{(0)}$, we can express the norm of $e_i^{(t)}$ as

$$\|e_i^{(t)}\| = \Big\|\beta \sum_{x=0}^{t-1} \Big[\prod_{y>x}^{t-1}(\mathbf{I} + \beta B_i^{(y)})\Big] C_i^{(x)} \Big[[\mathbf{A}^G]_i^\top - \frac{1}{N}\mathbf{1}\Big]\Big\| \tag{22}$$

$$\leq \beta \sum_{x=0}^{t-1} \Big\| \prod_{y>x}^{t-1}(\mathbf{I} + \beta B_i^{(y)}) \Big\| \cdot \|C_i^{(x)}\| \cdot \Big\|[\mathbf{A}^G]_i^\top - \frac{1}{N}\mathbf{1}\Big\|. \tag{23}$$

We bound each term in (23) as follows. For the first term, we have

$$\Big\| \prod_{y>x}^{t-1}(\mathbf{I} + \beta B_i^{(y)}) \Big\| \leq \prod_{y>x}^{t-1} \|\mathbf{I} + \beta B_i^{(y)}\| \leq \prod_{y>x}^{t-1} \Big(\|\mathbf{I}\| + \|\beta B_i^{(y)}\| \Big)$$

$$\leq \prod_{y>x}^{t-1} (1 + \beta(1+\gamma)) \tag{24}$$

$$= (1 + \beta(1+\gamma))^{t-1-x}, \tag{25}$$

where the last inequality is due to Assumption 5. For the second term, using Assumptions 1 and 5, we have

$$\|C_i^{(t)}\| = \big\|\phi(\eta_i^{(t)})[r_1^{(t)}, \cdots, r_N^{(t)}]\big\| \tag{26}$$

$$\leq \|\phi(\eta_i^{(t)})\| \cdot \big\|[r_1^{(t)}, \cdots, r_N^{(t)}]\big\| \tag{27}$$

$$\leq \big\|[r_1^{(t)}, \cdots, r_N^{(t)}]\big\| \tag{28}$$

$$\leq \sqrt{N} r_{\max}. \tag{29}$$

For the third term, we get

$$\Big\|[\mathbf{A}^G]_i^\top - \frac{1}{N}\mathbf{1}\Big\| \leq 2\sqrt{N} \frac{(1 + \frac{1}{\nu^{(N-1)}})}{1 - \nu^{N-1}} (1 - \nu^{N-1})^{G+1} \tag{30}$$

$$= 2\sqrt{N}(1 + \nu^{-(N-1)})(1 - \nu^{N-1})^G. \tag{31}$$

By combining each term, we obtain

$$\beta \sum_{x=0}^{t-1} \Big\| \prod_{y>x}^{t-1}(\mathbf{I} + \beta B_y) \Big\| \cdot \|C_x\| \cdot \Big\|[\mathbf{A}^G]_i^\top - \frac{1}{N}\mathbf{1}\Big\|$$

$$\leq \beta \sum_{x=0}^{t-1} (1 + \beta(1+\gamma))^{t-1-x} \cdot \sqrt{N} r_{\max} \cdot 2\sqrt{N}(1 + \nu^{-(N-1)})(1 - \nu^{N-1})^G \tag{32}$$

$$= 2N r_{\max} \beta (1 + \nu^{-(N-1)})(1 - \nu^{N-1})^G \sum_{x=0}^{t-1} (1 + \beta(1+\gamma))^{t-1-x} \tag{33}$$

Thus, the upper bound on $\|w_i^{(t)} - \bar{w}_i^{(t)}\|$ becomes

$$\|w_i^{(t)} - \bar{w}_i^{(t)}\| \leq 2Nr_{\max}\beta(1+\nu^{-(N-1)})(1-\nu^{N-1})^G \sum_{x=0}^{t-1}(1+\beta(1+\gamma))^{t-1-x} \tag{34}$$

To further understand (34), we simplify its right-hand-side (RHS). Let us first denote $c_1 \triangleq 2Nr_{\max}\beta(1+\nu^{-(N-1)})$ and consider $(1-\nu^{N-1})^G = e^{-c_2 G}$ with $c_2 \triangleq \ln(1-\nu^{N-1})^{-1} > 0$. By denoting $c_3 \triangleq (1+\gamma)\beta > 0$, we also have

$$\sum_{x=0}^{t-1}(1+(1+\gamma)\beta)^{t-1-x} = \sum_{x=0}^{t-1}(1+c_3)^x = \frac{(1+c_3)^t - 1}{c_3} \leq \frac{(1+c_3)^t}{c_3}.$$

Furthermore, we have $(1+c_3)^t = e^{t\ln(1+c_3)} = e^{c_4 t}$ where $c_4 \triangleq \ln(1+c_3)$. Combining the points made above, (34) becomes

$$\|w_i^{(t)} - \bar{w}_i^{(t)}\| \leq \frac{c_1}{c_3} e^{-c_2 G + c_4 t}. \tag{35}$$

Note that $\frac{c_1}{c_2}$ is a constant independent of step size β. The inequality in (35) indicates that if the exponent $-c_2 G + c_4 t$ remains a sufficiently large negative number, the consensus error should be sufficiently small.

We now work on the second part of our theorem, which is on the convergence of average parameter. Using [36], we have

Lemma 1. *(Theorem 7 of [36]) For any $t > \tau(\beta)$ and for sufficiently small constant step size β, the finite-time convergence bound for average parameter is*

$$\mathbb{E}\Big[\|\bar{w}_i^{(t)} - w_i^{(t)*}\|^2\Big] \leq c_5(1-c_6\beta)^{t-\tau(\beta)} + c_7\tau(\beta)\beta \tag{36}$$

where c_5, c_6, c_7 are constants independent of step size β, and $\tau(\beta) = \mathcal{O}(\log(\frac{1}{\beta}))$ is mixing time.

By Remark 1 in [36], $\beta\tau(\beta) \to 0$ as $\beta \to 0$.

We are now ready to bound $\|\bar{w}_i^{(t)} - w_i^{(t)*}\|$. For $i \in \mathcal{N}$, we have

$$\begin{aligned} \mathbb{E}\left[\|\bar{w}_i^{(t)} - w_i^{(t)*}\|^2\right] &\leq 2\mathbb{E}\left[\|w_i^{(t)} - \bar{w}_i^{(t)}\|^2\right] + 2\mathbb{E}\left[\|\bar{w}_i^{(t)} - w_i^{(t)*}\|^2\right] && (37) \\ &\leq 2\left(\frac{c_1}{c_3}\right)^2 e^{-2c_2 G + 2c_4 t} + 2c_5(1-c_6\beta)^{t-\tau(\beta)} + 2c_7\tau(\beta)\beta. && (38) \end{aligned}$$

Note that, in the finite time result above, the only constant that is dependent on β is c_4.

APPENDIX B
PROOF OF THEOREM 2

For the ease of notation, we define $v_i^{(t)}(w_i^{(t)}) = \tilde{\delta}_i^{(t)} \cdot \psi_i^{(t)}$ and $h_i^{(t)}(w_i^{(t)}) = \delta_i^{(t)*} \cdot \psi_i^{(t)}$, where $\psi_i^{(t)} = \nabla \log \pi_{\theta_i}(a_i^{(t)}|\eta_i^{(t)})$, $\tilde{\delta}_i^{(t)}$ is the TD error computed using $\tilde{r}_i^{(t)} = [\mathbf{A}^G]_i[r_1^{(t)}, \cdots, r_N^{(t)}]^\top$, and $\delta_i^{(t)*}$ is the TD error computed using $\bar{r}^{(t)} = \frac{1}{N}\mathbf{1}^\top[r_1^{(t)}, \cdots, r_N^{(t)}]^\top$. We also define $w^{(t)} = [w_1^{(t)}, \ldots, w_N^{(t)}]$, $v^{(t)}(w^{(t)}) = [v_1^{(t)}(w_1^{(t)}), \ldots, v_N^{(t)}(w_N^{(t)})]$, and $h^{(t)}(w^{(t)}) = [h_1^{(t)}(w_1^{(t)}), \ldots, h_N^{(t)}(w_N^{(t)})]$. Lastly, we define

$$\mathrm{Adv}_{w^{(t)}}(s^{(t)}, \boldsymbol{a}^{(t)}) = \mathbb{E}_{s' \sim P(\cdot|s,\boldsymbol{a}), r \sim d_r(s,\boldsymbol{a})}[\delta_{w^{(t)}}(s, \boldsymbol{a}, s')|s = s^{(t)}, \boldsymbol{a} = \boldsymbol{a}^{(t)}] \tag{39}$$

$$= \mathbb{E}_{s' \sim P(\cdot|s,\boldsymbol{a}), r \sim d_r(s,\boldsymbol{a})}[r + \gamma V_{w^{(t)}}(s') - V_{w^{(t)}}(s)|s = s^{(t)}, \boldsymbol{a} = \boldsymbol{a}^{(t)}] \tag{40}$$

and

$$g(w^{(t)}, \theta^{(t)}) = \mathbb{E}[\mathrm{Adv}_{w^{(t)}}(s^{(t)}, \boldsymbol{a}^{(t)})\psi_{\theta^{(t)}}(s^{(t)}, \boldsymbol{a}^{(t)})] \tag{41}$$

We now make the following assumption on $\psi_i^{(t)}$.

Assumption 6. *For any policy parameter θ_i, the score function $\psi_i^{(t)}$ is uniformly bounded, i.e., $\|\psi_i^{(t)}\|^2 \leq 1$.*

Since $J(\theta)$ is L_J-Lipschitz continuous from Assumption 3, we can apply descent lemma to obtain the following result:

$$J(\theta^{(t+1)}) \geq J(\theta^{(t)}) + \langle \nabla_\theta J(\theta^{(t)}), \theta^{(t+1)} - \theta^{(t)} \rangle - \frac{L_J}{2}\|\theta^{(t+1)} - \theta^{(t)}\|^2 \tag{42}$$

$$= J(\theta^{(t)}) + \alpha\langle \nabla_\theta J(\theta^{(t)}), v^{(t)}(w^{(t)}) - \nabla_\theta J(\theta^{(t)}) + \nabla_\theta J(\theta^{(t)}) \rangle - \frac{L_J\alpha^2}{2}\|v^{(t)}(w^{(t)})\|^2 \tag{43}$$

$$= J(\theta^{(t)}) + \alpha\|\nabla_\theta J(\theta^{(t)})\|^2 + \alpha\langle \nabla_\theta J(\theta^{(t)}), v^{(t)}(w^{(t)}) - \nabla_\theta J(\theta^{(t)}) \rangle - \frac{L_J\alpha^2}{2}\|v^{(t)}(w^{(t)}) - \nabla_\theta J(\theta^{(t)}) + \nabla_\theta J(\theta^{(t)})\|^2 \tag{44}$$

$$\geq J(\theta^{(t)}) + \left(\frac{1}{2}\alpha - L_J\alpha^2\right)\|\nabla_\theta J(\theta^{(t)})\|^2 - \left(\frac{1}{2}\alpha + L_J\alpha^2\right)\|v^{(t)}(w^{(t)}) - \nabla_\theta J(\theta^{(t)})\|^2, \tag{45}$$

where the last inequality is due to

$$\langle \nabla_\theta J(\theta^{(t)}), v^{(t)}(w^{(t)}) - \nabla_\theta J(\theta^{(t)}) \rangle \geq -\frac{1}{2}\|\nabla_\theta J(\theta^{(t)})\|^2 - \frac{1}{2}\|v^{(t)}(w^{(t)}) - \nabla_\theta J(\theta^{(t)})\|^2 \tag{46}$$

and

$$\|v^{(t)}(w^{(t)}) - \nabla_\theta J(\theta^{(t)}) + \nabla_\theta J(\theta^{(t)})\|^2 \leq 2\|v^{(t)}(w^{(t)}) - \nabla_\theta J(\theta^{(t)})\|^2 + 2\|\nabla_\theta J(\theta^{(t)})\|^2. \tag{47}$$

Taking the expectation on (45) and rearranging the terms, we have:

$$\begin{aligned}&\left(\frac{1}{2}\alpha - L_J\alpha^2\right)\mathbb{E}\left[\|\nabla_\theta J(\theta^{(t)})\|^2\right] \\ &\qquad\leq \mathbb{E}\left[J(\theta^{(t+1)})\right] - J(\theta^{(t)}) + \left(\frac{1}{2}\alpha + L_J\alpha^2\right)\mathbb{E}\left[\|v^{(t)}(w^{(t)}) - \nabla_\theta J(\theta^{(t)})\|^2\right],\end{aligned} \tag{48}$$

where the last term in the RHS should be carefully controlled. To this end, we adopt triangle inequality to attain the following inequality:

$$\begin{aligned}\|v^{(t)}(w^{(t)}) - \nabla_\theta J(\theta^{(t)})\|^2 \leq{}& 6\|v^{(t)}(w^{(t)}) - v^{(t)}(w^{(t)*})\|^2 + 6\|v^{(t)}(w^{(t)*}) - h^{(t)}(w^{(t)*})\|^2 \\ &+ 6\|h^{(t)}(w^{(t)*}) - g(w^{(t)*}, \theta^{(t)})\|^2 + 6\|g(w^{(t)*}, \theta^{(t)}) - \nabla_\theta J(\theta^{(t)})\|^2.\end{aligned} \tag{49}$$

We can decompose the first three terms using the following fact: $\|x\|^2 = \sum_{i\in\mathcal{N}}\|x_i\|^2$ for any $x = [x_1, \ldots, x_N]^\top$.

Now, we are ready to control each term in (49). The first term in the RHS of (49) can be bounded as follows:

$$\left\|v^{(t)}(w^{(t)}) - v^{(t)}(w^{(t)*})\right\|^2 \tag{50}$$

$$= \sum_{i\in\mathcal{N}}\left\|v_i^{(t)}(w_i^{(t)}) - v_i^{(t)}(w_i^{(t)*})\right\|^2 \tag{51}$$

$$= \sum_{i\in\mathcal{N}}\left\|\tilde{\delta}_i^{(t)}(w_i^{(t)}) \cdot \psi_i^{(t)} - \tilde{\delta}_i^{(t)}(w_i^{(t)*}) \cdot \psi_i^{(t)}\right\|^2 \tag{52}$$

$$= \sum_{i\in\mathcal{N}}\left\|\left[(\tilde{\delta}_i^{(t)}(w_i^{(t)}) - \tilde{\delta}_i^{(t)}(w_i^{(t)*})\right] \cdot \psi_i^{(t)}\right\|^2 \tag{53}$$

$$\leq \sum_{i\in\mathcal{N}}\|\tilde{\delta}_i^{(t)}(w_i^{(t)}) - \tilde{\delta}_i^{(t)}(w_i^{(t)*})\|^2 \cdot \|\psi_i^{(t)}\|^2 \tag{54}$$

$$\leq \sum_{i\in\mathcal{N}}\left\|(\tilde{r}_i^{(t)} + \gamma\phi^\top(\eta_i^{(t+1)})w_i^{(t)} - \phi^\top(\eta_i^{(t)})w_i^{(t)}) - (\tilde{r}_i^{(t)} + \gamma\phi^\top(\eta_i^{(t+1)})w_i^{(t)*} - \phi^\top(\eta_i^{(t)})w_i^{(t)*})\right\|^2 \tag{55}$$

$$= \sum_{i\in\mathcal{N}}\left\|\left[\gamma\phi^\top(\eta_i^{(t+1)}) - \phi^\top(\eta_i^{(t)})\right](w_i^{(t)} - w_i^{(t)*})\right\|^2 \tag{56}$$

$$\leq \sum_{i\in\mathcal{N}}\|\gamma\phi^\top(\eta_i^{(t+1)}) - \phi^\top(\eta_i^{(t)})\|^2 \cdot \|w_i^{(t)} - w_i^{(t)*}\|^2 \tag{57}$$

$$\le \sum_{i\in\mathcal{N}} (1+\gamma)^2 \|w_i^{(t)} - w_i^{(t)*}\|^2 \tag{58}$$

$$= (1+\gamma)^2 \sum_{i\in\mathcal{N}} \|w_i^{(t)} - w_i^{(t)*}\|^2 \tag{59}$$

where the second inequality is from Assumption 6, and the last inequality is due to Assumption 5.

The second term in the RHS of (49) can be bounded as follows:

$$\left\| v^{(t)}(w^{(t)*}) - h^{(t)}(w^{(t)*}) \right\|^2 \tag{60}$$

$$= \sum_{i\in\mathcal{N}} \left\| v_i^{(t)}(w_i^{(t)*}) - h_i^{(t)}(w_i^{(t)*}) \right\|^2 \tag{61}$$

$$= \sum_{i\in\mathcal{N}} \left\| \tilde{\delta}_i^{(t)}(w_i^{(t)*}) \cdot \psi_i^{(t)} - \delta_i^{(t)*}(w_i^{(t)*}) \cdot \psi_i^{(t)} \right\|^2 \tag{62}$$

$$= \sum_{i\in\mathcal{N}} \left\| \left[\tilde{\delta}_i^{(t)}(w_i^{(t)*}) - \delta_i^{(t)*}(w_i^{(t)*}) \right] \cdot \psi_i^{(t)} \right\|^2 \tag{63}$$

$$\le \sum_{i\in\mathcal{N}} \left\| \tilde{\delta}_i^{(t)}(w_i^{(t)*}) - \delta_i^{(t)*}(w_i^{(t)*}) \right\|^2 \cdot \left\| \psi_i^{(t)} \right\|^2 \tag{64}$$

$$\le \sum_{i\in\mathcal{N}} \left\| (\tilde{r}_i^{(t)} + \gamma \phi^\top(\eta_i^{(t+1)}) w_i^{(t)*} - \phi^\top(\eta_i^{(t)}) w_i^{(t)*}) - (\bar{r}_i^{(t)} + \gamma \phi^\top(\eta_i^{(t+1)}) w_i^{(t)*} - \phi^\top(\eta_i^{(t)}) w_i^{(t)*}) \right\|^2 \tag{65}$$

$$= \sum_{i\in\mathcal{N}} \left| \left([\mathbf{A}^G]_i - \frac{1}{N}\mathbf{1}^\top \right) [r_1^{(t)}, \cdots, r_N^{(t)}]^\top \right|^2 \tag{66}$$

$$\le \sum_{i\in\mathcal{N}} \left\| [\mathbf{A}^G]_i - \frac{1}{N}\mathbf{1}^\top \right\|^2 \cdot \left\| [r_1^{(t)}, \cdots, r_N^{(t)}] \right\|^2 \tag{67}$$

$$\le \sum_{i\in\mathcal{N}} \left\| [\mathbf{A}^G]_i - \frac{1}{N}\mathbf{1}^\top \right\|^2 N r_{\text{max}}^2 \tag{68}$$

$$\le \sum_{i\in\mathcal{N}} 4N \left[(1+\nu^{-(N-1)})(1-\nu^{N-1})^G \right]^2 \cdot N r_{\text{max}}^2 \tag{69}$$

$$= 4N^3 r_{\text{max}}^2 \left((1+\nu^{-(N-1)})(1-\nu^{N-1})^G \right)^2, \tag{70}$$

where the second inequality is due to Assumption 6, and the last inequality is by the property of gossiping technique.

According to the definitions of $h_i^{(t)}$ and (41), the third term in the RHS of (49) can be written as follows:

$$\left\| h^{(t)}(w^{(t)*}) - g(w^{(t)*}, \theta^{(t)}) \right\|^2 = \sum_{i\in\mathcal{N}} \left\| h_i^{(t)}(w_i^{(t)*}) - g(w_i^{(t)*}, \theta_i^{(t)}) \right\|^2 \tag{71}$$

$$= \sum_{i\in\mathcal{N}} \left\|\delta_i^{(t)*}(w_i^{(t)*})\cdot\psi_i^{(t)} - \mathbb{E}[\mathrm{Adv}_{w_i^{(t)*}}(s^{(t)},\boldsymbol{a}^{(t)})\psi_{\theta_i^{(t)}}(s^{(t)},a_i^{(t)})]\right\|^2. \tag{72}$$

Taking expectation over the filtration F_t on both sides of (72), we have:

$$\mathbb{E}\left[\left\|h^{(t)}(w^{(t)*}) - g(w^{(t)*},\theta^{(t)})\right\|^2|\mathcal{F}_t\right] \tag{73}$$

$$= \mathbb{E}\left[\sum_{i\in\mathcal{N}}\left\|h_i^{(t)}(w_i^{(t)*}) - g(w_i^{(t)*},\theta_i^{(t)})\right\|^2|\mathcal{F}_t\right] \tag{74}$$

$$= \mathbb{E}\left[\sum_{i\in\mathcal{N}}\left\|\delta_i^{(t)*}(w_i^{(t)*})\psi_i^{(t)} - \mathbb{E}\big[\mathrm{Adv}_{w_i^{(t)*}}(s^{(t)},\boldsymbol{a}^{(t)})\psi_{\theta_i^{(t)}}(s^{(t)},a_i^{(t)})\big]\right\|^2|\mathcal{F}_t\right] \tag{75}$$

$$= \mathbb{E}\left[\sum_{i\in\mathcal{N}}\left\|\delta_i^{(t)*}(w_i^{(t)*})\psi_i^{(t)} - \mathbb{E}\big[\delta_i^{(t)*}(w_i^{(t)*})\psi_i^{(t)}\big]\right\|^2|\mathcal{F}_t\right] \tag{76}$$

$$= \mathbb{E}\left[\sum_{i\in\mathcal{N}}\left\|\big(\delta_i^{(t)*}(w_i^{(t)*}) - \mathbb{E}[\delta_i^{(t)*}(w_i^{(t)*})]\big)\cdot\psi_i^{(t)}\right\|^2|\mathcal{F}_t\right] \tag{77}$$

$$\leq \mathbb{E}\left[\sum_{i\in\mathcal{N}}\left|\delta_i^{(t)*}(w_i^{(t)*}) - \mathbb{E}[\delta_i^{(t)*}(w_i^{(t)*})]\right|^2\cdot\|\psi_i^{(t)}\|^2|\mathcal{F}_t\right] \tag{78}$$

$$\leq \mathbb{E}\left[\sum_{i\in\mathcal{N}}\left|\delta_i^{(t)*}(w_i^{(t)*}) - \mathbb{E}[\delta_i^{(t)*}(w_i^{(t)*})]\right|^2|\mathcal{F}_t\right] \tag{79}$$

$$\leq \mathbb{E}\left[\sum_{i\in\mathcal{N}}\left|\delta_i^{(t)*}(w_i^{(t)*})\right|^2 + \left|\mathbb{E}[\delta_i^{(t)*}(w_i^{(t)*})]\right|^2|\mathcal{F}_t\right] \tag{80}$$

$$= \sum_{i\in\mathcal{N}}\mathbb{E}\left[\left|\delta_i^{(t)*}(w_i^{(t)*})\right|^2 + \left|\mathbb{E}[\delta_i^{(t)*}(w_i^{(t)*})]\right|^2|\mathcal{F}_t\right] \tag{81}$$

$$\leq \sum_{i\in\mathcal{N}}2\mathbb{E}\left[\left|\delta_i^{(t)*}(w_i^{(t)*})\right|^2|\mathcal{F}_t\right] + 2\mathbb{E}\left[\left|\mathbb{E}[\delta_i^{(t)*}(w_i^{(t)*})]\right|^2|\mathcal{F}_t\right] \tag{82}$$

$$\leq \sum_{i\in\mathcal{N}}4(r_{\max} + (1+\gamma)R_w)^2 \tag{83}$$

$$= 4N(r_{\max} + (1+\gamma)R_w)^2, \tag{84}$$

where the second inequality is from Assumption 6. The last inequality is due to

$$\|\delta_i^{(t)*}(w_i^{(t)})\| = \|\bar{r}_i^{(t)} + \gamma\phi^\top(\eta_i^{(t+1)})w_i^{(t)*} - \phi^\top(\eta_i^{(t)})w_i^{(t)*}\| \tag{85}$$

$$= \|\bar{r}_i^{(t)} + [\gamma\phi^\top(\eta_i^{(t+1)}) - \phi^\top(\eta_i^{(t)})]w_i^{(t)*}\| \tag{86}$$

$$\leq \|\bar{r}_i^{(t)}\| + \|\gamma\phi^\top(\eta_i^{(t+1)}) - \phi^\top(\eta_i^{(t)})\|\cdot\|w_i^{(t)*}\| \tag{87}$$

$$\leq \|\bar{r}_i^{(t)}\| + \Big(\|\gamma\phi^\top(\eta_i^{(t+1)})\| + \|\phi^\top(\eta_i^{(t)})\|\Big)\cdot\|w_i^{(t)*}\| \tag{88}$$

$$\leq r_{\text{max}} + (\gamma + 1)\, R_w, \tag{89}$$

where the last inequality is due to Assumptions 1 and 5 as well as the 2-norm bound on the equilibrium point $w_i^{(t)*}$ [37].

The last term in the RHS of (49) can be bounded as follows:

$$\left\|g(w^{(t)*}, \theta^{(t)}) - \nabla_\theta J(\theta^{(t)})\right\|^2 \tag{90}$$

$$= \left\|\mathbb{E}\left[\text{Adv}_{w^{(t)*}}(s^{(t)}, \boldsymbol{a}^{(t)})\psi_{\theta^{(t)}}(s^{(t)}, \boldsymbol{a}^{(t)})\right] - \mathbb{E}\left[\text{Adv}_{\theta^{(t)}}(s^{(t)}, \boldsymbol{a}^{(t)})\psi_{\theta^{(t)}}(s^{(t)}, \boldsymbol{a}^{(t)})\right]\right\|^2 \tag{91}$$

$$= \left\|\mathbb{E}\left[(\text{Adv}_{w^{(t)*}}(s^{(t)}, \boldsymbol{a}^{(t)}) - \text{Adv}_{\theta^{(t)}}(s^{(t)}, \boldsymbol{a}^{(t)}))\psi_{\theta^{(t)}}(s^{(t)}, \boldsymbol{a}^{(t)})\right]\right\|^2 \tag{92}$$

$$\leq \Big(\mathbb{E}\big[\|(\text{Adv}_{w^{(t)*}}(s^{(t)}, \boldsymbol{a}^{(t)}) - \text{Adv}_{\theta^{(t)}}(s^{(t)}, \boldsymbol{a}^{(t)}))\psi_{\theta^{(t)}}(s^{(t)}, \boldsymbol{a}^{(t)})\|\big]\Big)^2 \tag{93}$$

$$\leq \Big(\mathbb{E}\left[|\text{Adv}_{w^{(t)*}}(s^{(t)}, \boldsymbol{a}^{(t)}) - \text{Adv}_{\theta^{(t)}}(s^{(t)}, \boldsymbol{a}^{(t)})| \cdot \|\psi_{\theta^{(t)}}(s^{(t)}, \boldsymbol{a}^{(t)})\|\right]\Big)^2 \tag{94}$$

$$\leq \Big(\mathbb{E}\left[|\text{Adv}_{w^{(t)*}}(s^{(t)}, \boldsymbol{a}^{(t)}) - \text{Adv}_{\theta^{(t)}}(s^{(t)}, \boldsymbol{a}^{(t)})|\right]\Big)^2 \tag{95}$$

$$= \Big(\mathbb{E}\left[|\mathbb{E}[\gamma V_{w^{(t)*}}(s^{(t+1)})|s^{(t)}, \boldsymbol{a}^{(t)}] - V_{w^{(t)*}}(s^{(t)}) - \mathbb{E}[\gamma V_{\theta^{(t)*}}(s^{(t+1)})|s^{(t)}, \boldsymbol{a}^{(t)}] + V_{\theta^{(t)*}}(s^{(t)})|\right]\Big)^2 \tag{96}$$

$$\leq \Big(\mathbb{E}\left[|\mathbb{E}[\gamma V_{w^{(t)*}}(s^{(t+1)}) - \gamma V_{\theta^{(t)*}}(s^{(t+1)})|s^{(t)}, \boldsymbol{a}^{(t)}]| + |V_{w^{(t)*}}(s^{(t)}) - V_{\theta^{(t)*}}(s^{(t)})|\right]\Big)^2 \tag{97}$$

$$\leq \Big(\mathbb{E}\left[\mathbb{E}[|\gamma V_{w^{(t)*}}(s^{(t+1)}) - \gamma V_{\theta^{(t)*}}(s^{(t+1)})||s^{(t)}, \boldsymbol{a}^{(t)}] + |V_{w^{(t)*}}(s^{(t)}) - V_{\theta^{(t)*}}(s^{(t)})|\right]\Big)^2 \tag{98}$$

$$= \Big(\mathbb{E}[|\gamma V_{w^{(t)*}}(s^{(t)}) - \gamma V_{\theta^{(t)*}}(s^{(t)})|] + \mathbb{E}\left[|V_{w^{(t)*}}(s^{(t)}) - V_{\theta^{(t)*}}(s^{(t)})|\right]\Big)^2 \tag{99}$$

$$\leq (1 + \gamma)^2\Big(\mathbb{E}\left[|V_{w^{(t)*}}(s^{(t)}) - V_{\theta^{(t)*}}(s^{(t)})|\right]\Big)^2 \tag{100}$$

$$\leq (1 + \gamma)^2\mathbb{E}\left[|V_{w^{(t)*}}(s^{(t)}) - V_{\theta^{(t)*}}(s^{(t)})|^2\right] \tag{101}$$

$$\leq (1 + \gamma)^2\xi^{\text{critic}}_{\text{approx}}, \tag{102}$$

where $\xi^{\text{critic}}_{\text{approx}}$ is the error bound on the linear approximation of value function.

Combining everything together, we can upper bound the RHS of (49) as

$$\mathbb{E}\left[\|v^{(t)}(w^{(t)}) - \nabla_\theta J(\theta^{(t)})\|^2\right] \tag{103}$$

$$\leq 6(1 + \gamma)^2\sum_{i\in\mathcal{N}}\|w_i^{(t)} - w_i^{(t)*}\|^2 + 24N^3 r_{\text{max}}^2\left((1 + \nu^{-(N-1)})(1 - \nu^{N-1})^G\right)^2$$
$$+ 24N(r_{\text{max}} + (1 + \gamma)R_w)^2 + 6(1 + \gamma)^2\xi^{\text{critic}}_{\text{approx}}. \tag{104}$$

Therefore, we have:

$$\left(\frac{1}{2}\alpha - L_J\alpha^2\right)\mathbb{E}\left[\|\nabla_\theta J(\theta^{(t)})\|^2\right]$$

$$\leq \mathbb{E}\left[J(\theta^{(t+1)})\right] - \mathbb{E}[J(\theta^{(t)})] + \left(\frac{1}{2}\alpha + L_J\alpha^2\right)\Bigg(6(1+\gamma)^2\sum_{i\in\mathcal{N}}\|w_i^{(t)} - w_i^{(t)*}\|^2 \tag{105}$$

$$+ 24N^3 r_{\text{max}}^2\left((1+\nu^{-(N-1)})(1-\nu^{N-1})^G\right)^2 + 24N(r_{\text{max}} + (1+\gamma)R_w)^2 + 6(1+\gamma)^2\xi_{\text{approx}}^{\text{critic}}\Bigg). \tag{106}$$

By setting step-size $\alpha = \frac{1}{4L_J}$ and dividing both sides of previous equation by $\frac{1}{16L_J}$, we further obtain:

$$\begin{aligned}&\mathbb{E}\left[\|\nabla_\theta J(\theta^{(t)})\|^2\right]\\&\leq 16L_J\mathbb{E}\left[J(\theta^{(t+1)})\right] - 16L_J\mathbb{E}[J(\theta^{(t)})] + 18(1+\gamma)^2\sum_{i\in\mathcal{N}}\|w_i^{(t)} - w_i^{(t)*}\|^2\\&+ 72N^3 r_{\text{max}}^2\left((1+\nu^{-(N-1)})(1-\nu^{N-1})^G\right)^2 + 72N(r_{\text{max}} + (1+\gamma)R_w)^2 + 18(1+\gamma)^2\xi_{\text{approx}}^{\text{critic}}.\end{aligned} \tag{107}$$

Let $\hat{T}$ be a random integer variable uniformly taken from $(1, T)$. If we take summation over $t = \{1, \ldots, T\}$ and divide it by T, we have

$$\mathbb{E}\left[\|\nabla_\theta J(\theta^{(\hat{T})})\|^2\right] = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}[\|\nabla_\theta J(\theta^{(t)})\|^2] \tag{108}$$

$$\begin{aligned}&\leq \frac{16L_J(\mathbb{E}\left[J(\theta^{(T)})\right] - \mathbb{E}\left[J(\theta^{(0)})\right])}{T} + 18(1+\gamma)^2\frac{\sum_{t=1}^{T}\sum_{i\in\mathcal{N}}\|w_i^{(t)} - w_i^{(t)*}\|^2}{T}\\&+ 72N^3 r_{\text{max}}^2\left((1+\nu^{-(N-1)})(1-\nu^{N-1})^G\right)^2 + 72N(r_{\text{max}} + (1+\gamma)R_w)^2 + 18(1+\gamma)^2\xi_{\text{approx}}^{\text{critic}}\end{aligned} \tag{109}$$

$$\begin{aligned}&\leq \frac{16L_J\mathbb{E}\left[J(\theta^{(T)})\right]}{T} + 18(1+\gamma)^2\frac{\sum_{t=1}^{T}\sum_{i\in\mathcal{N}}\|w_i^{(t)} - w_i^{(t)*}\|^2}{T}\\&+ 72N^3 r_{\text{max}}^2\left((1+\nu^{-(N-1)})(1-\nu^{N-1})^G\right)^2 + 72N(r_{\text{max}} + (1+\gamma)R_w)^2 + 18(1+\gamma)^2\xi_{\text{approx}}^{\text{critic}}\end{aligned} \tag{110}$$

$$\begin{aligned}&\leq \frac{16L_J r_{\text{max}}}{T(1-\gamma)} + 18(1+\gamma)^2\frac{\sum_{t=1}^{T}\sum_{i\in\mathcal{N}}\|w_i^{(t)} - w_i^{(t)*}\|^2}{T}\\&+ 72N^3 r_{\text{max}}^2\left((1+\nu^{-(N-1)})(1-\nu^{N-1})^G\right)^2 + 72N(r_{\text{max}} + (1+\gamma)R_w)^2 + 18(1+\gamma)^2\xi_{\text{approx}}^{\text{critic}}.\end{aligned} \tag{111}$$

References

[1] N. Abramson, "The aloha system: Another alternative for computer communications," in *Proceedings of the November 17-19, 1970, fall joint computer conference*, 1970, pp. 281–285.

[2] R. Binder, N. Abramson, F. Kuo, A. Okinaka, and D. Wax, "Aloha packet broadcasting: a retrospect," in *Proceedings of the May 19-22, 1975, national computer conference and exposition*, 1975, pp. 203–215.

[3] L. Kleinrock and F. Tobagi, "Packet switching in radio channels: Part i-carrier sense multiple-access modes and their throughput-delay characteristics," *IEEE transactions on Communications*, vol. 23, no. 12, pp. 1400–1416, 1975.

[4] B. Chen, J. Chen, Y. Gao, and J. Zhang, "Coexistence of lte-laa and wi-fi on 5 ghz with corresponding deployment scenarios: A survey," *IEEE Communications Surveys & Tutorials*, vol. 19, no. 1, pp. 7–32, 2016.

[5] O. Aouedi, T.-H. Vu, A. Sacco, D. C. Nguyen, K. Piamrat, G. Marchetto, and Q.-V. Pham, "A survey on intelligent internet of things: Applications, security, privacy, and future directions," *IEEE Communications Surveys & Tutorials*, 2024.

[6] N. H. Mahmood, S. Böcker, I. Moerman, O. A. López, A. Munari, K. Mikhaylov, F. Clazzer, H. Bartz, O.-S. Park, E. Mercier *et al.*, "Machine type communications: Key drivers and enablers towards the 6g era," *EURASIP Journal on Wireless Communications and Networking*, vol. 2021, no. 1, p. 134, 2021.

[7] J. C. Rodriguez, F. Grijalva, M. García, D. E. Chérrez Barragán, B. A. Acuña Acurio, and H. Carvajal, "Wireless communication technologies for smart grid distribution networks," *Engineering Proceedings*, vol. 47, no. 1, p. 7, 2023.

[8] I. Ahmad, S. Shahabuddin, H. Malik, E. Harjula, T. Leppänen, L. Loven, A. Anttonen, A. H. Sodhro, M. M. Alam, M. Juntti *et al.*, "Machine learning meets communication networks: Current trends and future challenges," *IEEE access*, vol. 8, pp. 223 418–223 460, 2020.

[9] M. Kulin, T. Kazaz, E. De Poorter, and I. Moerman, "A survey on machine learning-based performance improvement of wireless networks: Phy, mac and network layer," *Electronics*, vol. 10, no. 3, p. 318, 2021.

[10] X. Cao, B. Yang, C. Huang, C. Yuen, M. Di Renzo, Z. Han, D. Niyato, H. V. Poor, and L. Hanzo, "Ai-assisted mac for reconfigurable intelligent-surface-aided wireless networks: Challenges and opportunities," *IEEE Communications Magazine*, vol. 59, no. 6, pp. 21–27, 2021.

[11] Y. Yu, T. Wang, and S. C. Liew, "Deep-reinforcement learning multiple access for heterogeneous wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 37, no. 6, pp. 1277–1290, 2019.

[12] Y. Yu, S. C. Liew, and T. Wang, "Multi-agent deep reinforcement learning multiple access for heterogeneous wireless networks with imperfect channels," *IEEE Transactions on Mobile Computing*, vol. 21, no. 10, pp. 3718–3730, 2022.

[13] L. Zhang, H. Yin, Z. Zhou, S. Roy, and Y. Sun, "Enhancing wifi multiple access performance with federated deep reinforcement learning," in *IEEE Vehicular Technology Conference*. IEEE, 2020, pp. 1–6.

[14] Z. Guo, Z. Chen, P. Liu, J. Luo, X. Yang, and X. Sun, "Multi-agent reinforcement learning-based distributed channel access for next generation wireless networks," *IEEE Journal on Selected Areas in Communications*, vol. 40, no. 5, pp. 1587–1599, 2022.

[15] Y. He, X. Gang, and Y. Gao, "Intelligent decentralized multiple access via multi-agent deep reinforcement learning," in *2024 IEEE Wireless Communications and Networking Conference (WCNC)*. IEEE, 2024, pp. 1–6.

[16] R. S. Sutton, A. G. Barto *et al.*, *Reinforcement learning: An introduction*. MIT press Cambridge, 1998, vol. 1, no. 1.

[17] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[18] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *arXiv preprint arXiv:1707.06347*, 2017.

[19] G. Choudhury and S. Rappaport, “Diversity aloha - a random access scheme for satellite communications,” *IEEE Transactions on Communications*, vol. 31, no. 3, pp. 450–457, 1983.

[20] H. Yu, H. Zhao, Z. Fei, J. Wang, Z. Chen, and Y. Gong, “Deep-reinforcement-learning-based noma-aided slotted aloha for leo satellite iot networks,” *IEEE Internet of Things Journal*, vol. 10, no. 20, pp. 17 772–17 784, 2023.

[21] F. Baccelli, B. Blaszczyszyn, and P. Muhlethaler, “An aloha protocol for multihop mobile wireless networks,” *IEEE transactions on information theory*, vol. 52, no. 2, pp. 421–436, 2006.

[22] L. Jiang, D. Shah, and J. Walrand, “Distributed random access algorithm: Scheduling and congestion control,” *IEEE Transactions on Information Theory*, vol. 56, no. 12, pp. 6182–6207, Dec. 2010.

[23] L. Jiang and J. Walrand, “A distributed CSMA algorithm for throughput and utility maximization in wireless networks,” *IEEE/ACM Transactions on Networking*, vol. 18, no. 3, pp. 960–972, Jun. 2010.

[24] L. Jiang, M. Leconte, J. Ni, R. Srikant, and J. Walrand, “Fast mixing of parallel Glauber dynamics and low-delay CSMA scheduling,” *IEEE Transactions on Information Theory*, vol. 58, no. 10, pp. 6541–6555, Oct. 2012.

[25] J. Ghaderi and R. Srikant, “The impact of access probabilities on the delay performance of Q-CSMA algorithms in wireless networks,” *IEEE/ACM Transactions on Networking*, vol. 21, no. 4, pp. 1063–1075, Aug. 2013.

[26] S. C. Liew, C. Kai, J. Leung, and B. Wong, “Back-of-the-envelope computation of throughput distributions in CSMA wireless networks,” in *Proc. IEEE ICC*, Dresden, Germany, Jun. 14-18, 2009, pp. 1–6.

[27] J. Ni, B. Tan, and R. Srikant, “Q-CSMA: Queue-length-based CSMA/CA algorithms for achieving maximum throughput and low delay in wireless networks,” *IEEE/ACM Transactions on Networking*, vol. 20, no. 3, pp. 825–836, Jun. 2010.

[28] M. Han, S. Khairy, L. X. Cai, Y. Cheng, and R. Zhang, “Reinforcement learning for efficient and fair coexistence between lte-laa and wi-fi,” *IEEE Transactions on Vehicular Technology*, vol. 69, no. 8, pp. 8764–8776, 2020.

[29] C. Lee, S. Park, and T. Cheong, “Dynamic-persistent csma: A reinforcement learning approach for multi-user channel access,” *IEEE Access*, 2024.

[30] S. Boyd, A. Ghosh, B. Prabhakar, and D. Shah, “Randomized gossip algorithms,” *IEEE transactions on information theory*, vol. 52, no. 6, pp. 2508–2530, 2006.

[31] K. Zhang, Z. Yang, H. Liu, T. Zhang, and T. Basar, “Fully decentralized multi-agent reinforcement learning with networked agents,” in *International conference on machine learning*. PMLR, 2018, pp. 5872–5881.

[32] F. Hairi, J. Liu, and S. Lu, “Finite-time convergence and sample complexity of multi-agent actor-critic reinforcement learning with average reward,” in proc. iclr, virtual event, april 2022,” *Proc. ICLR*, 2022.

[33] A. Feriani and E. Hossain, “Single and multi-agent deep reinforcement learning for ai-enabled wireless networks: A tutorial,” *IEEE Communications Surveys & Tutorials*, vol. 23, no. 2, pp. 1226–1252, 2021.

[34] Z. Chen, Y. Zhou, R.-R. Chen, and S. Zou, “Sample and communication-efficient decentralized actor-critic algorithms with finite-time analysis,” in *International Conference on Machine Learning*. PMLR, 2022, pp. 3794–3834.

[35] Hairi, Z. Zhang, and J. Liu, “Sample and communication efficient fully decentralized marl policy evaluation via a new approach: Local td update,” in *Proceedings of the 23rd International Conference on Autonomous Agents and Multiagent Systems*, 2024, pp. 789–797.

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

[37] T. Xu, Z. Wang, and Y. Liang, “Improving sample complexity bounds for (natural) actor-critic algorithms,” *Advances in Neural Information Processing Systems*, vol. 33, pp. 4358–4369, 2020.

[38] D. J. Watts and S. H. Strogatz, “Collective dynamics of ‘small-world’networks,” *nature*, vol. 393, no. 6684, pp. 440–442, 1998.

[39] L. Dai and X. Sun, “A unified analysis of ieee 802.11 dcf networks: Stability, throughput, and delay,” *IEEE Transactions on Mobile Computing*, vol. 12, no. 8, pp. 1558–1572, 2012.