Journal of Computer and Knowledge Engineering, Vol. , No.. 2023.
DOI:

Advancing Over-the-Air Federated Learning through Deep Reinforcement Learning in UAV-Assisted Networks with Movable Antennas*

Mohsen Ahmadzadeh[1], Saeid Pakravan [2], Ghosheh Abed Hodtani [3]

***Abstract*-- This paper investigates the deployment of over-the-air federated learning (OTA-FL), leveraging the dynamic repositioning and line-of-sight communication capabilities of unmanned aerial vehicles (UAVs) and movable antennas to enhance network efficiency. A closed-form expression is derived to quantify the optimality gap between the actual federated learning (FL) model and its theoretical ideal, accounting for the capabilities of movable antennas to show the diverse relationship between Mean Square Error (MSE) and the optimality gap. Then An MSE minimization problem is then formulated, involving the joint optimization of moveable antenna position vectors, and the beamforming vector at the UAV. This complex non-convex problem is reformulated as a Markov Decision Process (MDP) and solved using the Twin Delayed Deep Deterministic Policy Gradient (TD3) algorithm within the deep reinforcement learning (DRL) framework. Numerical results demonstrate that the proposed algorithm outperforms benchmarks such as Advantage Actor-Critic(A2C) and Soft Actor-Critic (SAC).**

***Index* Over-the-air federated learning, Deep reinforcement learning, Unmanned aerial vehicles, Movable Antenna.**

I. INTRODUCTION

Federated Learning (FL) is a secure method for collaboratively building a unified model across various participants. Yet, its application in practice often encounters difficulties caused by restricted data exchange capabilities [1] [2]. To tackle the challenge of achieving minimal delay and broad connectivity in IoT-driven Federated Learning, an innovative solution known as over-the-air FL (OTA-FL) has been developed [3] [4]. This technique maximizes the efficient use of bandwidth by leveraging the inherent combining feature of wireless access networks through analog signaling. OTA-FL achieves model integration by utilizing the overlapping characteristics of wireless signals, where updates from edge devices are merged into a collective representation. This process employs over the air computation (AirComp) to perform direct aggregation, avoiding the step of separately processing each parameter. Consequently, it decreases delays and boosts resource efficiency by operating within common time and frequency allocations.

Unmanned aerial vehicles (UAVs) are increasingly playing a pivotal role in modern wireless communication systems, owing to their cost-effectiveness, high mobility, and versatile capabilities. These vehicles are capable of operating as aerial base stations, relays, or access points, which significantly extend coverage and ensure reliable line-of-sight (LoS) connectivity for data transmission across various environments [5].

The evolution of communication systems has led to the widespread adoption of Multiple-Input Multiple-Output (MIMO) technology, characterized by the use of multiple antennas [6, 7]. MIMO systems are primarily designed to improve channel capacity, boost data transmission rates, and optimize various performance parameters of communication networks. [8, 9]. Traditional fixed-position antennas (FPAs) often face limitations in achieving optimal beamforming gains within dynamic environments. To address this limitation, we propose integrating movable antennas (MAs) into OTA-FL systems, allowing for real-time adaptation to varying wireless channel conditions [4, 10]. Leveraging MAs at the receiving server enhances OTA-FL performance by utilizing spatial degrees of freedom (DoF). Unlike FPAs, MAs offer the capability to reconfigure the wireless environment dynamically, introducing extra DoFs that substantially boost the efficiency and effectiveness of OTA-FL systems [4, 11].

Reinforcement Learning (RL) is a promising approach for autonomous decision-making, where an agent learns by interacting with its environment, taking actions, and adjusting its strategy based on rewards to optimize performance [12]. However, RL struggles with large-scale environments due to its high demand for computational resources. To address this, Deep Reinforcement Learning (DRL) leverages deep neural networks, enabling more efficient learning in complex, high-dimensional environments. DRL methods are particularly useful for modern networks with high computational complexity. Additionally, while centralized RL can create significant signalling overhead, DRL allows for decentralized multi-agent systems, where agents make independent decisions, reducing overhead and improving scalability, especially in applications like UAV-assisted networks [13, 14].

[1] Department of Electric and Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran, Email: m.ahmadzadehbolghan@mail.um.ac.ir

[2] Department of Electric and Computer Engineering, Laval University, Quebec City, Canada, Email: saeid.pakravan.1@ulaval.ca

[3] Department of Electric and Computer Engineering, Ferdowsi University of Mashhad, Mashhad, Iran, Email: hodtani@um.ac.ir

This study introduces a UAV-assisted MA-assisted framework that OTA-FL. The main contributions of this paper are outlined as follows:

1- System Design: This paper introduces a UAV-enabled MA-architecture that incorporates OTA-FL in the AP.

2- Optimality Gap Analysis: We perform a comprehensive optimality gap evaluation, deriving a closed-form expression to quantify the gap between the achieved and optimal loss. This analysis reveals how MSE impact the convergence behaviour of the OTA-FL algorithm.

3- Performance Assessment: We conduct extensive simulations to evaluate the effectiveness of the learning network. The results show that the TD3 method surpasses all single-agent techniques, including Advantage Actor-Critic(A2C) and Soft Actor-Critic (SAC).

A. Related work

Numerous research efforts have explored the use of UAVs and MA in OTA-FL. In the following section, we present an in-depth analysis of these studies.

To improve the efficiency of OTA-FL, various research efforts have utilized UAV to address challenges related to magnitude alignment during model aggregation at the edge server. In [15], the Fog-aided Internet of Drones framework employs machine learning to analyse data collected by drones at fog nodes, offering various services. FL is utilized to enhance data privacy by enabling local drone training and sharing model parameters instead of raw data. However, privacy risks persist due to potential eavesdropping on uploaded parameters. The study focuses on optimizing drone power control to maximize the FL system's security rate while meeting battery and quality of service (QoS) constraints. A non-linear programming approach is proposed, and simulations validate the algorithm's effectiveness. In [16], a federated learning-based framework, Aerial Edge, is proposed for orchestrating aerial edge computing systems using UAVs. The approach employs multi-output regression to optimize resource allocation and execution time, selecting UAVs with suitable resources and flight time. A bin-packing optimization variant is introduced for efficient task scheduling, achieving fast execution and improved resource utilization, validated with real-world data. In [17], UAV swarms leveraging FL are studied to enable edge intelligence while addressing bandwidth and energy limitations. To minimize training energy consumption, the study jointly optimizes convergence thresholds, iterations, resource, and bandwidth allocation under accuracy and latency constraints. A fairness-focused variant minimizes maximum energy consumption across UAVs. Simulations demonstrate superior energy efficiency compared to baseline approaches. In [18], UAVs are employed in CR networks to leverage their high mobility and LoS transmission. However, spectrum sharing can cause interference, reducing the throughput of secondary users. RIS are utilized to mitigate this interference by reconstructing propagation links. The study focuses on maximizing SU throughput while ensuring primary user interference constraints are met, through joint optimization of UAV trajectory, RIS passive beamforming, and UAV power allocation. The problem is divided into three subproblems: beamforming, power allocation, and trajectory design, and an alternating iterative optimization algorithm is proposed. Numerical results demonstrate significant throughput improvement. In [19], a joint subchannel assignment and power allocation algorithm is proposed for NOMA-enabled cognitive satellite-UAV-terrestrial networks to optimize the sum rate of the secondary network under imperfect channel state information. The problem, constrained by interference temperature for primary users, minimum secondary user rates, UAV power limits, and subchannel capacity, is formulated as a mixed-integer non-linear programming task. It is addressed by decoupling into subchannel assignment and power allocation subproblems, solved using heuristic and successive convex approximation methods, respectively. Simulations demonstrate the algorithm's superior performance in large-scale networks compared to benchmarks. In [4], the authors study an OTA-FL system with MAs at the AP to enhance

Table 1: summaries of all parameters

parameter	Definition
N	The total number of FL clients.
K	The number of MA on the UAV.
v_t	The global model at timeslot t
γ	Denotes the learning rate parameter.
q	The dimensionality of the model parameter space.
$\nabla G(v_t, S_n)$	Denotes the gradient of the local loss function.
S_n	signifies the local dataset for n-th UE
D	Refers to a one-dimensional segment of the MA antenna length.
D_0	The minimum spacing between adjacent antennas to prevent coupling.
$g_n[d]$	Represents the wireless channel between the n-th user equipment and the UAV.
ℓ_0	Denotes the path loss at the reference distance.
λ	Denotes the wavelength.
α	Represents the path loss exponent.
x_n	Denotes the distance between the MA antennas and the n-th UE.
θ_n	the AoA of the LoS path.
p_n	Represents the transmission power factor for the n-th UE.
z	Denotes an AWGN matrix, where each element follows a complex normal distribution.
W	Represents the beamforming vector at the UAV.
η	Denotes the scaling factor used for aligning signal amplitude.
$V_{\max}$	Represents the speed of the UAV in meters per second (m/s).
$I[t]$	ꞁepresents the location of the UAV at timeslot t
δ	Denotes the flying time between two consecutive timeslots.
r	Parameter representing the assumption of smoothness of model

μ	Parameter related to the PL inequality.
Λ	Denotes the upper limit of the model parameter.
S_t	Represents the states of the wireless environment for the MDP problem.
a_t	Denotes the action space of the MDP agent.
$reward_t$	Represents the reward function in the MDP framework.
π_φ	Denotes the policy function in DRL.
ϑ_1 , ϑ_2	Represents the parameters of the critic network in DRL.

learning performance. They derive the optimality gap to evaluate FA mobility's impact and propose a nonconvex optimization framework to jointly optimize FA positions and beamforming. The problem is modeled as a MDP and solved using the recurrent deep deterministic policy gradient algorithm. Simulations show the FA-assisted OTA-FL system outperforms fixed-antenna systems, with RDPG surpassing existing methods. In [20], we explore the application of DRL techniques to design MA for OTA-FL in UAV networks, aiming to enhance the overall network performance by optimizing the antenna positions. By considering UAVs as FL clients, we demonstrate the efficacy of this approach in improving the communication and learning capabilities within the network. In [21], the authors propose an OTA-FL framework using movable antennas (MAs) and UAVs for IoT support. They minimize MSE via joint antenna and beamforming optimization, modeled as an MDP and solved with TD3. Simulations show TD3-based MA systems outperform FPA and other DRL methods, achieving higher rewards and better performance.

B. Organization:

This paper is structured as follows: Section II. provides a detailed explanation of the system architecture, covering OTA-FL techniques and the UAV-enabled communication framework. In Section III. , the convergence behavior of the OTA-FL method is analyzed. Section IV. formulates the optimization problem, with a focus on the optimality gap. Section V. presents a DRL-based framework for optimization. The simulation setup, experimental scenarios, and comparative results with existing benchmarks are discussed in Section VI. to validate the proposed approach. The paper concludes with Section VII.

II. SYSTEM MODEL

We focus on the upload phase of an OTA-FL framework, which involves N single-antenna user equipment (UE) devices denoted as $UE_n, \forall n = [1, ..., N]$, referred to as FL clients. These clients are randomly distributed across a designated area to collect local datasets, train local models, and collaboratively optimize a global model. The training of the global model is coordinated by a UAV equipped with K movable antennas (MA-UAV). The UAV moves randomly within the area of interest to facilitate communication and coordination with the FL clients.

We analyze the OTA-FL framework, where full participation involves performing sequential tasks in each training round. The process of OTAFL is as follows: The UAV transmits the updated global model, $v_t \in \mathrm{R}^q$ to all UEs, with q representing the size of the model parameter space. Each UE_n updates its local model using the gradient descent method, described as:

$$v_{n,t} = v_t - \gamma \nabla G(v_t, S_n), \quad (1)$$

here, γ represents the learning rate, $\nabla G(v_t, S_n)$ denotes the gradient of the local loss function, and S_n signifies the local dataset for UE_n , $|S_n| = S$, with $|S_n|$ indicating its size. Each UE sends its updated local model back to the UAV, which aggregates these models by averaging them to update the global model, expressed as:

$$v_{t+1} = \frac{1}{N} \sum_{n=1}^{N} v_{n,t}. \quad (2)$$

This process is repeated iteratively until the predefined maximum number of outer iterations is achieved.

The UAV is equipped with a K MAs, which can be adjusted along a one-dimensional segment of length D with $[0, D]$. Each MA's position is restricted to the interval $[0, D]$ maintaining a minimum spacing of D_0 between adjacent antennas to prevent coupling. The positions of the K MAs are represented by the vector $d = [d_1, ..., d_K]$, with their movement confined to a single dimension as defined by $d_1 < d_2 ... < d_K$.

Under the assumption of line-of-sight (LoS) propagation conditions, the channel between the n-th UE and the UAV denoted as $g_n[d] \in C^{K \times 1}$, is expressed as:

$$g_n[d] = \sqrt{\frac{\ell_0}{x_n^{\alpha}}} [e^{j\frac{2\pi}{\lambda} d_1 \cos(\theta_n)}, ..., e^{j\frac{2\pi}{\lambda} d_K \cos(\theta_n)}]^T, \quad (3)$$

here, ℓ_0 represents the path loss at the reference distance, λ denotes the wavelength, and α is the path loss exponent. Additionally, x_n and θ_n correspond to the distance between the MAs and the n-th UE, and the angle of arrival (AoA) of the LoS path, respectively. These values are determined based on the UAV locations during each training round.

In this context, it is assumed that UAV operates within a predefined area and transmits global model parameters from a fixed position. Moreover, because the signal path length is substantially greater than the extent of MA movement, the MA field condition between the UAV and UEs is presumed to hold. Consequently, θ_n and x_n are treated as constants during the transmission phase.

During the t-th training round, the UAV receives the local model parameters from all UEs, expressed as:

$$y = \sum_{n=1}^{N} p_n g_n[d] v_n + z, \qquad (4)$$

here, p_n represents the transmission power factor for the n-th UE, and $z \in C^{q \times N}$ denotes an additive white Gaussian noise (AWGN) matrix, where each element follows a complex normal distribution $CN(0, \sigma^2)$. The aggregated model parameter vector $\hat{v}$ in the t-th training round is obtained by applying post-processing to the received signal at the UAV expressed as:

$$\begin{aligned} \hat{v}_{t+1} &= \frac{1}{N}\left(\frac{1}{\sqrt{\eta}} W^H y\right) \\ &= \frac{1}{N}\sum_{n=1}^{N} \frac{1}{\sqrt{\eta}} W^H p_n g_n[d] v_n + \frac{w^H z}{N\sqrt{\eta}}, \end{aligned} \qquad (5)$$

here, $W \in$ [illegible] $^{N \times 1}$ represents the beamforming vector at the UAV, and η denotes the scaling factor used for aligning the signal amplitude.

As outlined in [4], maximum power factor in each UEs should satisfies:

$$\frac{1}{q} p_n^{\,2} E[|v_n|^2] \le P_{\max}, \quad \forall n \in [1,...,N]. \qquad (6)$$

It is assumed that the UAV starts and concludes the FL process at the same location, with its maximum allowable speed represented as $V_{\max}$ in meters per second (m/s). The UAV's movement is subject to the following constraints:

$$\begin{aligned} &|l[t+1] - l[t]|^2 \le V_{\max}\delta, \\ &l[0] = [0,0,0], \\ &l[T] = [0,0,0], \end{aligned} \qquad (7)$$

here, $l[t]$ represents the location of the UAV at time slot t, and δ denotes the flying time between two consecutive time slots. In order to make easy to read and follow equation in Table 1 is summarized all character.

III. CONVERGENCE ANALYSIS

To support our convergence analysis, we adopt the following widely accepted assumptions as outlined in [4], [17]: the global loss function is r-smooth, meaning that for any given model parameters, there exists a nonnegative constant r for any given model parameters $v_1, v_2 \in R^q$, such that:

$$\begin{aligned} &G(v_1) - G(v_2) \le \\ &(v_1 - v_2)^T \nabla G(v_1) + \frac{r}{2} \| v_1 - v_2 \|^2 . \end{aligned} \qquad (8)$$

Where r is a measure of how smooth the function is, a smaller value indicates a smoother function. Additionally, the loss function satisfies the Polyak–Łojasiewicz (PL) inequality, ensuring that:

$$|G(v)|^2 \ge 2\mu[G(v) - G(v^*)] \qquad (9)$$

Where, if $G(v^*)$ represents the optimal global loss value and where $\mu \ge 0$ is the PL constant. Lastly, the model parameters for each UE are bounded by an upper limit, ensuring that for $\Lambda \ge 0$ we have:

$$E[|v|^2] \le \Lambda \qquad (10)$$

Theorem 1: Under the conditions specified in above assumptions, and by setting the learning rate $\frac{1}{r}$, the optimality gap after T rounds of training is bounded by:

$$\begin{aligned} &E[G(v_{T+1}) - G(v^*)] \le \\ &(1 - \frac{\mu}{r})^T (E[G(v_1)] - E[G(v^*)]) \\ &+ \sum_{t=1}^{T} (1 - \frac{\mu}{r})^{T-t} MSE_t . \end{aligned} \qquad (11)$$

To further refine the bounds, MSE_t can be expressed as follows:

$$MSE_t = \frac{r\sigma^2\Lambda}{2N^2 P_{\max}} \max_{n \in [1,...,N]} \frac{|w_t|^2}{|w_t g_t[d]|^2} . \qquad (12)$$

Proof: See Appendix.

IV. PROBLEM FORMULATION

According to Theorem 1, the optimality gap, which reflects the learning performance of OTAFL, can be expressed in terms of the MSE in each communication round. This is determined based on the relationship between the model updates, the aggregation error introduced by wireless communication, and the cumulative effect of these errors over multiple rounds. To enhance the learning performance, we can formulate an optimization problem aimed at jointly optimizing key system parameters, such as $w = [w_1, ...w_N]^T$ and $d = [d_1, ..., d_N]$, with the objective of minimizing the MSE as follow:

$$\begin{aligned} &\min_{d,w} \ \mathrm{MSE}_t \\ &\text{S.t.} \quad C_1 : 0 \le d_n \le D \qquad n \in [1,...,N], \\ &\qquad C_2 : d_1 \le d_2 ... \le d_N , \\ &\qquad C_3 : d_n - d_{n-1} \ge D_0 \quad n \in [1,...,N], \end{aligned}$$

In this context, several constraints are defined to regulate the behavior of the MAs. These constraints include limiting the valid range within which the positions of the MAs can be located in C_1 , determining the sequence in which the MAs are positioned in C_2 and enforcing a minimum required distance between neighboring Mas in C_3 . The complexity of the problem is further heightened by the inherent nonconvexity present in the objective function, categorizing it as a nonconvex optimization problem.

Traditional mathematical optimization techniques, commonly referenced in the literature, encounter significant difficulties when applied to such problems. This is primarily due to the high dimensionality of the optimization variables and the dynamic, unpredictable nature of the underlying system characteristics.

To address these challenges effectively, we propose leveraging a DRL algorithm, which offers greater flexibility and adaptability to accommodate the varying configurations and demands of the system.

V. PROPOSED DRL ALGORITHM

In this segment, we begin by reinterpreting the optimization challenge as a MDP, laying the foundation for addressing it using the TD3 algorithm. To tackle this problem, a DRL agent is implemented at the UAV, aiming to develop an optimal decision-making strategy that significantly boosts the efficiency of OTA-FL. The proposed framework leverages the MDP structure, enabling systematic problem-solving. The specifics of the MDP's state representation, action space, and reward mechanism are elaborated below for clarity and completeness.

State Space: The state space at time slot t captures the key environmental and system parameters that influence decision-making. It includes the distances between the MAs and clients, as well as the AoA for the LoS paths associated with these entities. These parameters collectively describe the dynamic state of the system at time t. Mathematically, the state space is represented as:

$$S_t = [[x_1, ..., x_n], [\theta_1, ..., \theta_n]]. \quad (13)$$

Action Space: The action space at a given time slot t defines the set of controllable variables that the agent can adjust to optimize system performance. It encompasses two critical components: the beamforming vector, which directs the transmitted signal's power and phase, and the spatial locations of the MAs. These actions collectively determine the system's ability to adapt to environmental changes and maximize efficiency. Mathematically, the action space at time slot t is represented as:

$$a_t = [[d_1, \ldots, d_N], [w, \ldots, w_N]]. \quad (14)$$

These variables collectively constitute the optimization parameters that require iterative refinement at each time step.

Reward Function: The reward function is designed to align with the optimization objective while adhering to system constraints. It evaluates the agent's actions by penalizing deviations from the desired outcome, ensuring that the learning process encourages improvements in system performance. The reward is mathematically defined as:

$$reward_t = -MSE_t \times r\_tune, \quad (15)$$

where MSE_t represents a function inversely related to the MSE, effectively penalizing higher error values, and r_tune is a constant parameter that can be fine-tuned during simulations to facilitate faster convergence of the learning algorithm. This formulation ensures that the reward reflects the system's performance, driving the agent toward minimizing the MSE while achieving the desired trade-offs between speed and accuracy.

A. TD3 ALGORITHM

In this study, we explore the utilization of the TD3 algorithm, a model-free and policy-based DRL approach, to handle the complexities and dynamic nature of the environment. Our proposed framework leverages the TD3 algorithm to efficiently manage continuous state and action spaces, addressing challenges inherent in such systems.

The proposed TD3-based solution incorporates six distinct neural networks, each playing a specific role in the decision-making process. These networks work collaboratively to optimize performance and ensure stability in training. The roles and functionalities of these networks are elaborated as follows:

Actor Network: The actor network, often referred to as the policy network, is a key component responsible for generating actions based on the current state of the environment. It is parameterized π_ϕ and serves as the decision-making entity within the TD3 framework. The actor network maps a given state S_t to a corresponding action a_t ,enabling the agent to interact with the environment effectively. This process can be mathematically expressed as:

$$a_t = \pi_\phi(S_t) + \zeta, \quad (16)$$

where π_φ represents the policy function parameterized by φ and ζ denotes a random noise process introduced to encourage exploration of the action space during training. This exploration ensures that the agent does not converge prematurely to suboptimal policies and thoroughly explores the environment.

Two Critic Networks: The TD3 algorithm incorporates two critic networks, often referred to as Q-networks, to evaluate the quality of actions taken by the agent. These networks, parameterized by ϑ_1 and ϑ_2 ,estimate the Q-value for a given action a_t and state S_t providing a measure of expected future rewards. The Q-value predictions from the critic networks can be expressed as: $Q_{\vartheta_1}(S_t, a_t; \vartheta_1)$ and $Q_{\vartheta_2}(S_t, a_t; \vartheta_2)$, where $Q_{\vartheta_1}(S_t, a_t; \vartheta_1)$ and $Q_{\vartheta_1}(S_t, a_t; \vartheta_2)$ are the outputs of the two critic networks. Using two critics helps mitigate the overestimation bias commonly found in Q-learning algorithms, as the TD3 framework selects the minimum of the two Q-values during training to compute the target. This design enhances the stability and reliability of the learning process.

Target Actor Network: The Target Actor Network serves as a prior version of the actor network, distinguished by ϕ' as: s. It produces an output with added noise ζ to stabilize the value estimation. This output is clipped to ensure it remains within a defined target range. The network's parameters are periodically refreshed using a soft update strategy, governed by a coefficient, as described below:

$$\phi' \leftarrow \tau\phi + (1-\tau)\phi'. \quad (17)$$

Two Target Critic Networks: These are the earlier iterations of the critic networks, parameterized by $\vartheta_1', \vartheta_2'$. they compute the Q-value $Q_{\vartheta_i'}(S_{t+1}, a_{t+1}; \vartheta_i')$. The parameters are gradually updated over time using a soft update process defined as:

$$\vartheta_1' \leftarrow \tau\vartheta_1 + (1-\tau)\vartheta_1'. \quad (18)$$

$$\vartheta_2' \leftarrow \tau\vartheta_2 + (1-\tau)\vartheta_2'.$$

The actor network is optimized to maximize the objective function through a policy gradient approach, which adjusts the actor's parameters using the following update rule:

$$\nabla_\phi J(\phi) = E[\nabla_{a_t} Q_\vartheta(S_t, a_t)|_{a_t=\pi(S_t)} \nabla_\phi \pi(S_t)]. \quad (19)$$

Simultaneously, the critic networks are trained to minimize the loss function relative to the target value. This process is expressed as:

$$L(\vartheta_1) = E((Y - Q_{\vartheta_1}(S_t, a_t))^2), \quad (20)$$

$$L(\vartheta_2) = E((Y - Q_{\vartheta_2}(S_t, a_t))^2).$$

The target function is defined as:

$$Y = reward + \quad (21)$$

$$\iota(\min(Q_{\vartheta_1}(S_{t+1}, \pi'(S_{t+1})), Q_{\vartheta 2}(S_{t+1}, \pi'(S_{t+1}))) + \zeta)).$$

Here, ι denotes the discount factor. Choosing the smaller Q-value from the critics mitigates Q-value overestimation, while adding the noise term to the target policy helps reduce overfitting.

VI. Simulation Results

In this section, we present numerical results to demonstrate how integrating MA arrays with the TD3 algorithm can significantly enhance the performance of OTA-FL. The simulation setup assumes that the distances between c ents and $l[0]$, uniformly distributed within the range of [20, 1( ] meters, while the AoAs are uniformly distributed over $[-\pi/2, \pi/2]$ radians. The parameters for the MA array are set as $D_0 = 0.5\lambda$ and $D = 8\lambda$, where λ is the wavelength.

For the TD3, SAC, A2C algorithm, the configuration includes a learning rate of 0.0005, a replay buffer size of 10^4, a batch size of 64, a soft update rate of 0.001, and a discount factor of 0.9.

To assess the performance, we conduct two comparisons. First, we compare the MA-based system with a FPA approach, where a fixed location vector $d = [\frac{D}{N+1}, \ldots, \frac{ND}{N+1}]^T$ is used. This comparison helps us evaluate the effectiveness of the MA array in improving OTA-FL performance over a more traditional method. Second, we compare the proposed TD3 algorithm with two other reinforcement learning algorithms: SAC and A2C. This comparison is aimed at showcasing the advantages of the TD3 algorithm over these alternative approaches in optimizing the system's performance.

The learning performance is measured by calculating the average rewards over 10 episodes. The average reward for episode b is computed as:

$$reward_{avg}(e) = 0.1 \times \sum_{i=b-10}^{b} reward_i \quad (22)$$

where $reward_i$ represents the reward obtained in episode i and b denotes the total number of episodes. This method allows for a thorough evaluation of the TD3 algorithm's performance in optimizing OTA-FL, both when compared to fixed strategies and other reinforcement learning methods. In this simulation, the Baseline3 library in Python is utilized to model and analyze the system dynamics under various operational conditions.

Figure 1 presents the trend of average rewards for three DRL algorithms, showing a steady increase followed by eventual convergence after 200 training episodes. The SAC algorithm follows a similar convergence pattern to TD3 but achieves lower average rewards when trained with the same learning rate, highlighting the superior performance of TD3. On the other hand, the A2C algorithm converges at a later stage and exhibits consistently lower average rewards throughout the training, suggesting that its performance is inferior compared to both TD3 and SAC. This comparison underscores the effectiveness of the TD3 algorithm in optimizing the learning process.

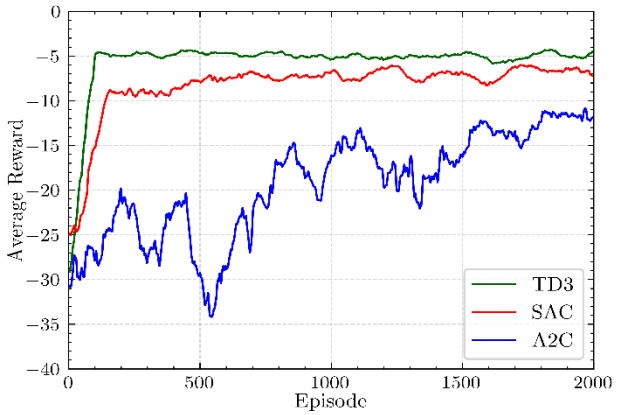

Figure 1: Comparison of different DRL algorithms versus episodes

Figure 2 illustrates the relationship between the number of FL clients and the average reward achieved using the TD3 and SAC algorithms in both FPA and MA scenarios. The results show that, for the same number of clients, the proposed TD3 algorithm consistently outperforms SAC across both FPA and MA setups. Additionally, the data indicates that the MA configuration consistently yields better performance compared to the FPA setup, highlighting the advantages of using MA in optimizing the learning process in federated learning environments.

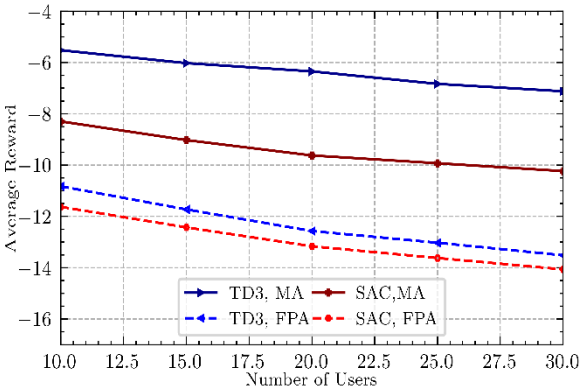

Figure 2: Comparison of algorithms with different numbers of FL clients

Figure 3 demonstrates how the effectiveness of the algorithm changes with varying values of K for both SAC and TD3, under the FPA and MA frameworks. In both algorithms, the MA framework consistently shows superior performance compared to the FPA setup. Additionally, TD3 outperforms SAC across

both frameworks. As the number of elements K increases, performance initially improves for both algorithms. However, this improvement begins to level off and even diminish as K continues to grow, which is likely due to the degradation in channel quality

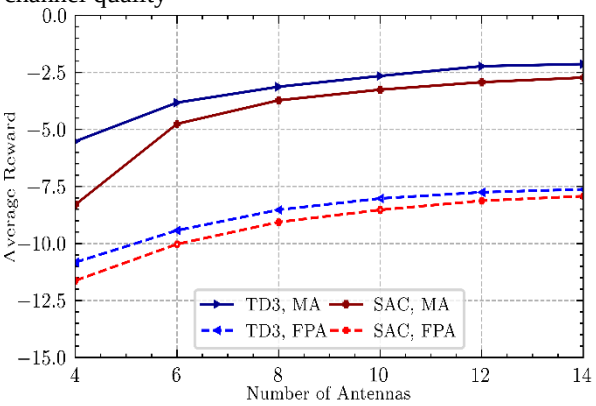

Figure 3: Comparison of algorithms for different numbers of antennas

For the same values of k, the proposed TD3 algorithm consistently achieves better results than SAC. The diminishing performance enhancement for large k values can be attributed to the limited physical space available when the number of antennas increases while keeping the antenna array length constant. As the number of antennas grows, the Spacing between them decreases, reducing the diversity and spatial resolution of the system. This limitation leads to a smaller overall performance gain, making the performance difference between the FPA and MA frameworks less pronounced as k becomes large.

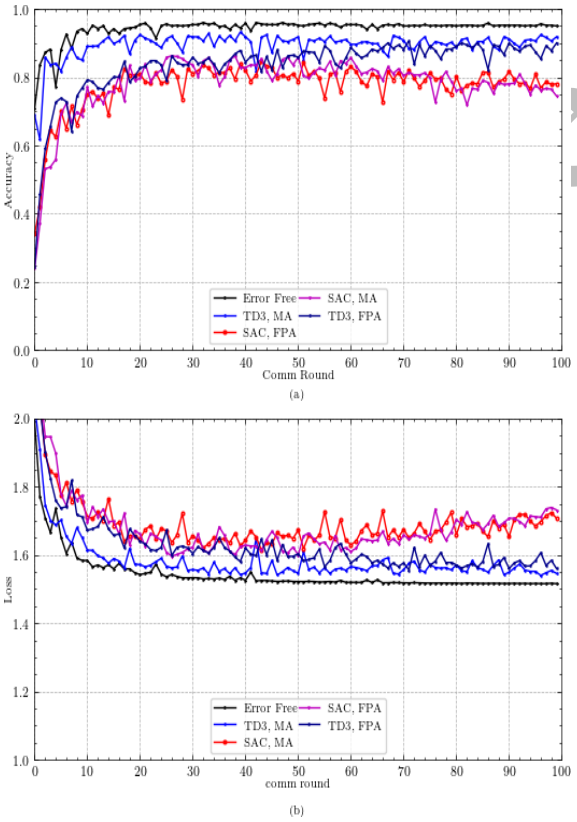

Figure 4: Comparison of DRL Algorithms on MNIST

To showcase the efficiency of our proposed approach for tackling FL tasks, we trained image classification models on the MNIST dataset. A fully participatory FL framework with five FL clients (N=5), representing UEs, was used. The MNIST dataset was divided into training (90%) and testing (10%) subsets. The training subset was then split into five non-overlapping shards, which were distributed among the clients without replacement. Each client implemented a local model based on a feedforward neural network. The architecture included an input layer with 200 neurons and a ReLU activation function, followed by a hidden layer also containing 200 neurons and a ReLU activation function. The output layer had neurons equal to the number of classes and used a softmax activation function for classification. An ideal FL scenario was considered, without communication noise or interference temperature constraints, to serve as a benchmark.

Error! Reference source not found. illustrates that the TD3 method surpasses SAC in both training loss and test accuracy. As the communication rounds progress, the performance gap between the TD3 method and SAC increases. Additionally, in both algorithms, MA consistently performs better than FPA. This underscores the importance of optimizing the RIS configuration, where phase shift optimization plays a pivotal role in improving overall performance.

VII. CONCLUSION

This paper introduced an innovative OTA-FL framework that incorporates MAs at the UAV and UEs as FL clients to significantly enhance system learning efficiency. A non-convex optimization problem was formulated with the goal of minimizing MSE by jointly optimizing antenna placement and beamforming vectors. To account for dynamic environments, this optimization problem was reformulated as a MDP and solved using the TD3 algorithm. The simulation results demonstrated that TD3 outperforms traditional stationary antenna configurations, illustrating the superiority of the proposed MA-based approach. Comparisons between MA arrays and FPA, as well as with alternative algorithms like SAC and A2C, consistently showed that TD3 outperforms FPA, especially as the number of antennas and clients increases. This leads to higher average rewards and significantly better system performance, confirming the effectiveness of the proposed method in enhancing OTA-FL performance.

VIII. APPENDIX

In the t-th communication round, the global model update is expressed in the following form:

$$\hat{v}_{t+1} = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{\sqrt{\eta}}W^H p_n g_n[d]v_n + \frac{w^H z}{N\sqrt{\eta}}, \qquad (23)$$

In order to minimize MSE, by applying channel inverse power allocation [22] [4], the power for each client is now determined as follows:

$$p_n = \frac{\sqrt{\eta}}{W^H g_n[d]}, \quad \forall n \in [1,...,N]. \tag{24}$$

By substituting equation (24) into equation (23) we obtain:

$$\begin{aligned} \hat{v}_{t+1} &= \frac{1}{N}\sum_{n=1}^{N} v_{n,t} + \frac{w^H z}{N\sqrt{\eta}}, \\ &= \frac{1}{N}\sum_{n=1}^{N} (v_t - \gamma \nabla G(v_t, S_n)) + \frac{w^H z}{N\sqrt{\eta}}, \\ &= v_t - \frac{\gamma}{N}\sum_{n=1}^{N} (\nabla G(v_t, S_n)) + \frac{w^H z}{N\sqrt{\eta}}, \\ &= v_t - \gamma(\nabla G(v_t)) - \varepsilon_t). \end{aligned} \tag{25}$$

Here, $\nabla G(v_t) = \frac{1}{N}\sum_{n=1}^{N} \nabla G(v_t, S_n)$ represents the global gradient, while $\varepsilon_t = \frac{w^H z}{\gamma N\sqrt{\eta}}$ refers to the aggregation error.

By incorporating equation (8) and setting $\gamma = \frac{1}{r}$ and taking expectation, we obtain the following expression:

$$\begin{aligned} &\mathbf{E}(G(v_{t+1})) - \mathbf{E}(G(v_t)) \\ &\leq \frac{-1}{2r} \| \nabla G(v_t) \|^2 + \frac{1}{2r}\mathbf{E}(|\varepsilon_t|^2), \end{aligned} \tag{26}$$

Now, with some straightforward mathematical simplifications, we obtain:

$$\begin{aligned} &\mathbf{E}(G(v_{t+1})) \leq \mathbf{E}(G(v_t)) - \\ &\frac{1}{2r} \| \nabla G(v_t) \|^2 + \frac{q|m|^2}{2N^2\eta}\mathbf{E}(|z|^2), \end{aligned} \tag{27}$$

By applying the upper bound of the above equation using equation (10) and equation (9), we derive:

$$\begin{aligned} &\mathbf{E}(G(v_{t+1})) \leq \mathbf{E}(G(v_t)) \\ &-\frac{\mu}{r}(\mathbf{E}(G(v_t)) - \mathbf{E}(G^*(v_t))) + \frac{q|m|^2}{2N^2\eta}\sigma^2, \end{aligned} \tag{28}$$

Based on the maximum power constraint for each client and equation (24), η can be calculated as follows:

$$\begin{aligned} &\frac{1}{q}\frac{\eta}{|W^H g_n[d]|^2}|E[|v_n|^2]|^2 \leq P_{\max}, \\ &\eta \leq \frac{P_{\max} q |W^H g_n[d]|^2}{\Lambda |w|^2}, \\ &\eta = \min_n \frac{P_{\max} q}{\Lambda}\frac{|W^H g_n[d]|^2}{|w|^2}. \end{aligned} \tag{29}$$

By applying the minimum value of η and performing some simple mathematical simplifications, we obtain:

$$\begin{aligned} &\mathbf{E}(G(v_{t+1})) - \mathbf{E}(G(v^*{}_t)) \\ &\leq (1-\frac{\mu}{r})(\mathbf{E}(G(v_t)) - \mathbf{E}(G(v_t^*))) \\ &+ \frac{r\sigma^2\Lambda}{2N^2 P_{\max}} \max_n \frac{|w|^2}{|W^H g_n[d]|^2}. \end{aligned} \tag{30}$$

where the second part of the above equation corresponds to the MSE in this scenario:

$$\begin{aligned} \mathrm{MSE}_t &= E[|\hat{v}_{t+1} - v_{t+1}|^2] \\ &= \frac{1}{N^2}\sum_{n=1}^{N} |1 - \frac{1}{\sqrt{\eta_t}} w_t^H p_{n,t} g_{n,t}[d]|^2 \; E[|v_{n,t}|^2] \\ &+ \frac{q|w|^2\sigma^2}{N^2\eta_t}, \\ &= \frac{q|w|^2\sigma^2}{N^2\eta}, \\ &= \frac{r\sigma^2\Lambda}{2N^2 P_{\max}} \max_{n\in[1,...N]} \frac{|w_t|^2}{|w_t^H g_{n,t}[d]|^2}, \end{aligned} \tag{31}$$

Thus, the equation (30) can be rewritten as follows:

$$\begin{aligned} &\mathbf{E}(G(v_{t+1})) - \mathbf{E}(G^*(v_t)) \\ &\leq (1-\frac{\mu}{r})(\mathbf{E}(G(v_t)) - \mathbf{E}(G^*(v_t))) \\ &+ \boldsymbol{MSE_t}. \end{aligned} \tag{32}$$

The optimality gap quantifies the discrepancy between the global model's state at the t -th communication round and its initial state, relative to the optimal model. To compute this, we systematically and iteratively apply recursive operations based on the structure outlined in (32). Additionally, by integrating the detailed definition of the MSE provided in (31), we obtain a comprehensive and precise expression for the cumulative optimality gap, which can be represented as follows:

$$\begin{aligned} &\mathbf{E}(G(v_{t+1})) - \mathbf{E}(G(v^*)) \\ &\leq (1-\frac{\mu}{r})(\mathbf{E}(G(v_t)) - \mathbf{E}(G(v^*))) \\ &+ \boldsymbol{MSE_t}, \\ &\leq (1-\frac{\mu}{r})(((1-\frac{\mu}{r}) \\ &(\mathbf{E}(G(v_{t-1}) - \mathbf{E}(G(v^*)) \\ &+ \boldsymbol{MSE_{t-1}}) - \mathbf{E}(G(v^*))) + \boldsymbol{MSE_t}, \\ &\leq (1-\frac{\mu}{r})(((1-\frac{\mu}{r})(((1-\frac{\mu}{r}) \\ &(\mathbf{E}(G(v_{t-2})) - \mathbf{E}(G(v^*))) + \boldsymbol{MSE_{t-2}}) \\ &-\mathbf{E}(G(v^*)) + \boldsymbol{MSE_{t-1}}) - \mathbf{E}(G(v^*))) \\ &+ \boldsymbol{MSE_t}, \\ &\leq \ldots \\ &\leq (1-\frac{\mu}{l})^T (E[G(v_1)] - E[G(v^*)]) \\ &+ \sum_{t=1}^{T} (1-\frac{\mu}{r})^{T-t} \boldsymbol{MSE_t}. \end{aligned} \tag{33}$$

This concludes the proof of Theorem 1.

IX. REFERENCES

[1] Ahmadzadeh M, Pakravan S, Hodtani GA, Zeng M, Chouinard JY. Deep Reinforcement Learning for Robust RIS-Aided Over-the-Air Federated Learning in Cognitive Radio. In2024 IEEE Middle East Conference on Communications and Networking (MECOM) 2024 Nov 17 (pp. 368-373). IEEE.

[2] M.Pourghasemian, and et al., "Cooperative Partial Task-Offloading for Heterogeneous Industrial Robotic MEC System Using Spectral and Energy-Efficient Federated Learning.," IEEE Global Communications Conference, 2023.

[3] Pakravan S, Ahmadzadeh M, Zeng M, Hodtani GA, Chouinard JY, Rusch LA. Robust Resource Allocation for Over-the-Air Computation Networks with Fluid Antenna Array. InIEEE Globecom Workshops (GC Wkshps) 2024 Sep.

[4] M. Ahmadzadeh, and et al., "Enhancement of Over-the-Air Federated Learning by Using AI-based Fluid Antenna System," arXiv preprint arXiv:2407.03481, 2024.

[5] Y.Wang, and et al., "Learning in the air: Secure federated learning for UAV-assisted crowdsensing," EEE Transactions on Network Science and Engineering, vol. 8, no. 2, pp. 1055-1069, Aug, 2020.

[6] M Bakhshi, and et al., "Enhancing long-range radar (LRR) automotive applications: Utilizing metasurface structures to improve the performance of K-band Longitudinal Slot Array Antennas," AEU - International Journal of Electronics and Communications, 2023.

[7] Pakravan S, Chouinard JY, Li X, Zeng M, Hao W, Pham QV, Dobre OA. Physical layer security for NOMA systems: Requirements, issues, and recommendations. IEEE Internet of Things Journal. 2023 Jul 18;10(24):21721-37.

[8] M.Bakhshi, and et al., "Enhanced 2-port MIMO antenna with composite two-step metasurface for 77 GHz Vehicle-to-Everything applications," AEU - International Journal of Electronics and Communications, 2024.

[9] Pakravan S, Chouinard JY, Zeng M, Li X, Hao W, Dobre OA. Physical-Layer Security of RIS-Assisted Networks Over Correlated Fisher-Snedecor F Fading Channels. IEEE Internet of Things Journal. 2023 Dec 18;11(9):15152-65.

[10] Park S, Seo H. Federated Learning Meets Fluid Antenna: Towards Robust and Scalable Edge Intelligence. arXiv preprint arXiv:2503.03054. 2025 Mar 4.

[11] Shen LH, Chiu YH. RIS-Aided Fluid Antenna Array-Mounted UAV Networks. IEEE Wireless Communications Letters. 2025 Jan 17.

[12] M.Pourghasemian, and et al., "AI-based mobility-aware energy efficient resource allocation and trajectory design for NFV enabled aerial networks," EEE Transactions on Green Communications and Networking, 2022.

[13] M.Pourghasemian , and et al., "Cooperative Partial Task-Offloading for Heterogeneous Industrial Robotic MEC System Using Spectral and Energy-Efficient Federated Learning," IEEE Global Communications Conference, 2023.

[14] S.Sheikhzadeh, and et al., "AI-based secure NOMA and cognitive radio enabled green communications: Channel state information and battery value uncertainties," IEEE Transactions on Green Communications and Networking, 2022.

[15] "Secure Federated Learning by Power Control for Internet of Drones," *IEEE Transactions on Cognitive Communications and Networking*, vol. 7, no. 4, 2021.

[16] J Yao, and et al.,"Secure Federated Learning by Power Control for Internet of Drones," IEEE Transactions on Cognitive Communications and Networking, vol. 7, no. 4, 2021.

[17] W. Pham, and et al., "Energy-efficient federated learning over UAV-enabled wireless powered communications.," IEEE Transactions on Vehicular Technology, 2022

[18] D.Yu, and et al., " Joint trajectory and resource optimization for RIS assisted UAV cognitive radio," IEEE Transactions on Vehicular Technology, 2023.

[19] G. Liu, and et al., " Resource allocation for NOMA-enabled cognitive satellite–UAV–terrestrial networks with imperfect CSI," IEEE Transactions on Cognitive Communications and Networking, pp. 26-58, 2023.

[20] Mhsen ahmadzadeh, saeid pakravan, ghoshe abed hodtani, "Movable Antenna Design for UAV-Aided Federated Learning via Deep Reinforcement Learning," 15th Conference on Information and Knowledge Technology, 2024.

[21] Ahmadzadeh M, Pakravan S, Hodtani GA. Movable Antenna Design for UAV-Aided Federated Learning via Deep Reinforcement Learning. In2024 15th International Conference on Information and Knowledge Technology (IKT) 2024 Dec 24 (pp. 91-95). IEEE.

[22] Cao X, Zhu G, Xu J, Cui S. Transmission power control for over-the-air federated averaging at network edge. IEEE Journal on Selected Areas in Communications. 2022 Jan 14;40(5):1571-86.