3TO: THz-Enabled Throughput and Trajectory Optimization of UAVs in 6G Networks by Proximal Policy Optimization Deep Reinforcement Learning

†Sheikh Salman Hassan, †Yu Min Park, †Yan Kyaw Tun, ‡Walid Saad, ††Zhu Han, and †Choong Seon Hong
†Department of Computer Science and Engineering, Kyung Hee University, Yongin, 17104, Republic of Korea
‡Wireless@VT, Bradley Department of Electrical and Computer Engineering, Virginia Tech, VA, 24061, USA
††Department of Electrical and Computer Engineering, University of Houston, Houston, TX 77004-4005, USA
Email: {salman0335, yumin0906, ykyawtun7, cshong}@khu.ac.kr, walids@vt.edu, zhan2@uh.edu.

ICC 2022 - IEEE International Conference on Communications | 978-1-5386-8347-7/22/$31.00 ©2022 IEEE | DOI: 10.1109/ICC45855.2022.9838676

***Abstract*—Next-generation networks need to meet ubiquitous and high data-rate demand. Therefore, this paper considers the throughput and trajectory optimization of terahertz (THz)-enabled unmanned aerial vehicles (UAVs) in the sixth-generation (6G) communication networks. In the considered scenario, multiple UAVs must provide on-demand terabits per second (TB/s) services to an urban area along with existing terrestrial networks. However, THz-empowered UAVs pose some new constraints, e.g., dynamic THz-channel conditions for ground users (GUs) association and UAV trajectory optimization to fulfill GU's throughput demands. Thus, a framework is proposed to address these challenges, where a joint UAVs-GUs association, transmit power, and the trajectory optimization problem is studied. The formulated problem is mixed-integer non-linear programming (MINLP), which is NP-hard to solve. Consequently, an iterative algorithm is proposed to solve three sub-problems iteratively, i.e., UAVs-GUs association, transmit power, and trajectory optimization. Simulation results demonstrate that the proposed algorithm increased the throughput by up to 10%, 68.9%, and 69.1% respectively compared to baseline algorithms.**

***Index Terms*—Sixth generation (6G) networking, unmanned aerial vehicles (UAVs), terahertz (THz) communication, proximal policy optimization (PPO), deep reinforcement learning (DRL).**

I. Introduction

Wireless communication systems have grown exponentially in the past several decades due to the need for high-speed data connections wherever and whenever needed. Therefore, sixth-generation (6G) networks must be multi-dimensional and capable of providing ubiquitous and diverse services by integrating current terrestrial network specifications with space and air-based information networks [1]–[3]. 6G networks will likely be cell-free, which means that users will be able to smoothly and automatically switch from one network to another to seek the most appropriate and qualified communication without the need for human administration and settings [4].

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (No. No. 2020R1A4A1018607) and by the Institute of Information and Communications Technology Planning and Evaluation (IITP) Grant funded by the Korea Government (MSIT) (Artificial Intelligence Innovation Hub) under Grant 2021-0-02068. *Dr. CS Hong is the corresponding author.

Furthermore, the terahertz (THz) band, with frequencies ranging from 0.1 to 10 THz, is a viable frequency range for the next generation of ultra-dense wireless networks [5], [6]. The THz channel can provide rates in the order of terabits per second (Tbps) that are suitable to support emerging applications, including high-quality video streaming, virtual and augmented reality, and chip-based wireless networks. Moreover, due to the restricted bandwidth at millimeter-wave (mmW), i.e., 30 gigahertz (GHz) carrier frequencies, it is impossible to reach Tbps data speeds.

In particular, utilizing THz-enabled unmanned aerial vehicles (UAVs) communications to offer seamless coverage and provide high bandwidth to ground users could be an efficient wireless network integration. Many mobile carriers in the United States have tested UAV-mounted LTE base stations, including AT&T and Verizon [7]. Aerial base stations may be attached to UAVs at low altitudes, making them more cost-effective, quick, and adaptable than terrestrial communication infrastructures [8]. Also, UAV communications benefit from better line-of-sight links with ground users due to their high altitudes. However, there are significant problems in terms of deployment, trajectory design, and network resource optimization when using UAVs for wireless communications.

To fully exploit the design degrees of freedom for THz-enabled UAV communications, it is crucial to investigate the UAV's mobility and its network resource management in three-dimensional space. The existing works [9]–[11] consider the UAVs as aerial base stations. The authors in [9] studied the problem of minimization of uplink power through user association in the presence of a single UAV base station. In [10], the authors investigated the problems of optimal trajectory design and transmit power optimization to maximize data rate in the presence of energy harvesting constraints. Similarly, the authors in [11] studied a three-dimensional (3D) coverage maximization problem for UAV networks. Moreover, the authors in [12] analyzed the coverage probability of UAVs with THz communication but did not optimize the UAVs' deployment. The authors in [13] and [14] investigated the problem of UAV deployment and transmitting power for THz communication.

However, they used traditional optimization techniques to solve the proposed problem. Moreover, none of the prior research takes into account the usage of the THz spectrum for UAVs with deployment and trajectory optimization.

Hence, to fill the knowledge gap, we investigate the fundamental study of optimal UAV trajectory deployment design and network resource management at THz frequencies. Nonetheless, given the significant degree of uncertainty in higher frequency bands such as THz, it is critical to provide additional degrees of freedom and control in network management. Therefore, given the ability of UAVs to fly, they are suitable for providing line-of-sight (LoS) communication links to ground users (GUs) [15] and [16]. To reap the benefits of deploying THz-enabled UAVs, it is important to optimize the locations of the UAVs to offer continuous LoS connections to GUs. Therefore, to handle the mobility of UAVs in our proposed dynamic network, we consider deep reinforcement learning-based UAVs location optimization with a low-complexity algorithm for GU association and transmit power optimization. The key contributions of our proposed work can be summarized as follows:

- We propose a THz-enabled UAVs communication network architecture by considering the quality-of-service (QoS) parameters for the GUs and the UAV's optimal trajectory deployment.
- Our objective is to maximize overall throughput between the UAVs and the GU by jointly optimizing the operational UAV's trajectory deployment and GU association, as well as minimizing the transmitting power of the UAVs.
- To tackle this optimization problem, we propose an iterative algorithm that separates the original optimization problem into three subproblems: A GU association subproblem is handled by balanced K-means clustering (BKMC), a power control subproblem is solved by successive convex approximation (SCA), and a trajectory deployment optimization subproblem is tackled by proximal policy optimization deep reinforcement learning (PPO-DRL) that is solved iteratively.
- Numerical results show that the proposed algorithm can increase throughput by up to 10%, 68.9%, and 69.1% respectively when compared to a baseline that optimizes only transmit power with static UAVs, random transmit power with static UAVs and optimizes only UAVs trajectories with random transmit power.

The rest of the paper is organized as follows. The system model and the problem formulation are given in Section II-A and II-C respectively. Then, the proposed algorithm is presented in Section III. Section IV describes the implementation and simulation results, and Section V concludes the paper.

II. System Model & Problem Formulation

A. Network Model

As shown in Fig. 1, we consider a THz-enabled multi-UAV wireless communication network. This system model consists of a set $\mathcal{K}$ of K UAVs that seek to provide communication services to a set $\mathcal{M}$ of M GUs distributed according to a homogeneous Poisson point process (HPPP). The three-dimensional (3D) coordinates of UAV k can be represented as $\boldsymbol{q_k}(n)=\big(x_k(n), y_k(n), z_k(n)\big)$, and the two-dimensional (2D) coordinates of GU m can defined as $\boldsymbol{o_m}(n)=\big(x_m(n), y_m(n)\big)$ respectively.

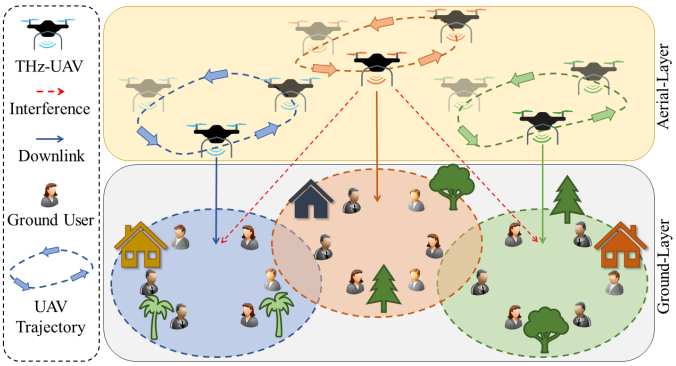

Fig. 1: Illustration of THz-enabled UAVs Network.

B. Channel Model & Link Analysis

The presence of this molecular absorption loss distinguishes the terahertz channel. This loss is produced by molecules in the air, such as H_2O vapor, which each have their own absorption spectrum, making the wireless channel frequency selective. Thus, the THz communication channel between UAV k and GU m can be modeled as follows [13]:

$$h_{k,m} = d_{k,m}^{-2} e^{-a(f)d_{k,m}}, \quad \forall k \in \mathcal{K},\ \forall m \in \mathcal{M}, \tag{1}$$

where $d_{k,m} = \|\boldsymbol{q}_k - \boldsymbol{o}_m\|$ is the distance between UAV k and GU m, $e^{-a(f)d_{k,m}}$ represents the channel path-loss due to the molecular absorption, and $a(f)$ denotes the molecular absorption coefficient, which depends on the network operating frequency (i.e., THz frequency) and the concentration of water vapor molecules in the air[1]. To efficiently use the network resources, we consider each UAV to utilize the same frequency spectrum. Therefore, the network experiences interference from non-associated UAVs to the GUs in each time slot, which can be defined as:

$$\psi_{k,m}(n) = \sum_{\forall k' \neq k} \sum_{\forall m' \neq m} p_{k',m'}(n) h_{k',m'}(n), \quad \forall k \in \mathcal{K},\ \forall m \in \mathcal{M}, \tag{2}$$

where $p_{k',m'}$ is the transmit power, k' is the non-associated UAVs, and m' represents the non-associated GUs, respectively. Therefore, the signal-to-interference-plus-noise ratio (SINR) from each UAV k to GU m in each time slot n can be formulated as:

$$\gamma_{k,m}(n) = \frac{p_{k,m}(n) h_0}{\psi_{k,m} + \alpha_{k,m} B_{k,m} d_{k,m}^2 e^{a d_{k,m}} \sigma^2}, \quad \forall k \in \mathcal{K},\ m \in \mathcal{M}, \tag{3}$$

where h_0 is the channel gain at a reference distance $d_0 = 1$m, $p_{k,m}$ is the transmit power from the UAV k to their associated GU m, $\alpha_{k,m}$ proportion of the channel bandwidth from the UAV

[1]Hereinafter, for the sake of simplicity, we will write a instead of $a(f)$.

k to GU m, $B_{k,m}$ is the total channel bandwidth, and σ^2 is the additive white Gaussian noise (AWGN) power. The achievable throughput from UAV k to GU m in each time slot n can be obtained based on the SINR $\gamma_{k,m}$ as:

$$R_{k,m}(n) = \alpha_{k,m} B_{k,m} \log_2 (1 + \gamma_{k,m}) . \tag{4}$$

C. Problem Formulation

Our goal is to maximize the total throughput from all the deployed UAVs while satisfying the QoS and trajectory constraints of each GU and UAV, respectively, in the network. Therefore, the throughput maximization problem can be defined as:

$$\mathbf{P1:} \max_{\boldsymbol{\alpha},\boldsymbol{p},\boldsymbol{q}} \quad R_k^{\text{lo}}(n) \tag{5a}$$

$$\text{s.t.} \quad \sum_{n=1}^{N} \sum_{m=1}^{M} \alpha_{k,m} R_{k,m}(n) \geq R_k^{\text{lo}}(n), \ \forall k \in \mathcal{K}, \tag{5b}$$

$$\alpha_{k,m} R_{k,m}(n) \geq R^{\min}, \ \forall m \in \mathcal{M}, k \in \mathcal{K}, n \in \mathcal{N}, \tag{5c}$$

$$\alpha_{k,m} \in \{0,1\}, \sum_{m=1}^{M} \alpha_{k,m} = 1, \ \forall k \in \mathcal{K}, m \in \mathcal{M}, \tag{5d}$$

$$\sum_{m=1}^{M} p_{k,m}(n) \leq P_k^{\max}, \forall k \in \mathcal{K}, n \in \mathcal{N}, \tag{5e}$$

$$0 \leq p_{k,m}(n) \leq P_k^{\max}, \forall k \in \mathcal{K}, n \in \mathcal{N}, \tag{5f}$$

$$\|\mathbf{q}_i(n) - \mathbf{q}_j(n)\|_2^2 \geq D_{\min}^2, \forall i \neq j \in \mathcal{K}, \forall n \in \mathcal{N}, \tag{5g}$$

$$\frac{\|\mathbf{q}_k(n+1) - \mathbf{q}_k(n)\|}{t^{\text{mov}}} \leq V^{\max}, \forall k \in \mathcal{K}, n \in \mathcal{N}, \tag{5h}$$

where (5b) represents the optimization objective, which is the sum-rate of each UAV at each time slot n. (5c) ensures the QoS constraint of each user from the associated UAV, (5d) presents that each GU can be associated with at most one UAV, (5e) and (5f) ensures the total transmit power of the UAV have to be less than the maximum transmit power $P_k^{\max}$ of the UAV, (5g) guarantees that the distance between UAVs is not as close as the minimum distance $D_{\min}$, and (5h) is the UAVs speed constraint.

III. Proposed Algorithm

P1 is a MINLP, which is NP-hard and difficult to solve. Therefore, we will decompose it into three subproblems and propose an iterative algorithm that is composed of BKMC, SCA, and PPO-DRL in the following subsections, respectively.

A. Balanced K-means Clustering

Given the initial UAVs' transmit power $\boldsymbol{p}$ and trajectory $\boldsymbol{q_m}$, **P1** is transformed into **P1.1** for optimizing GU association $\boldsymbol{\alpha}$, which is integer linear programming. Mathematically, it can be defined as:

$$\mathbf{P1:1} \max_{\boldsymbol{\alpha}} \quad R_k^{\text{lo}}(n) \tag{6a}$$

$$\text{s.t.} \quad (5b), (5c), \text{ and } (5d). \tag{6b}$$

Algorithm 1 BKMC for GUs Association

1: **Input:** the GU locations $\{\mathbf{o}_m\}_{m\in\mathcal{M}}$, the initial UAV locations $\{\mathbf{q}_k\}_{k\in\mathcal{K}}$.
2: **Initialize:** Initialize centroid locations C^0 to UAV locations $\{\mathbf{q}_k\}_{k\in\mathcal{K}}$.
3: $t \leftarrow 0$
4: **repeat**
5: Calculate distances between GUs and UAVs.
6: Solve an assignment problem by Hungarian algorithm.
7: Calculate new centroid locations C^{t+1}.
8: **until** the positions of the centroids do not change
9: **Output:** Optimal user association. $\boldsymbol{\alpha}^*$

GU association problems, in general, can be solved by K-means clustering. Although K-means clustering poses an unfairness issue that can lead to any cluster expanding excessively. Thus, we propose the use of BKMC, which can calculate the size of each cluster in advance [17].

We follow the same steps as K-means, while the assignment phase is different. We define pre-allocated slots according to the total number of GU M, instead of choosing the closest UAV k, each GU m needs to be associated with these slots i.e., M/K slots per cluster. Assuming that $\lfloor M/K \rfloor = \lceil M/K \rceil = M/K$, this will compel all clusters to have the same size. Otherwise, there will be (M mod K) clusters of size $\lfloor M/K \rfloor$, and $K-(M$ mod $K)$ clusters of size $\lceil M/K \rceil$. To find an assignment that minimizes mean square error (MSE), we solve an assignment problem using the Hungarian algorithm [18].

In contrast to a traditional assignment problem with fixed weights, the weights in this problem vary dynamically after each K-means iteration based on the newly determined centroids. The Hungarian method is then used to obtain the minimum weight pairing. The update step for choosing new UAV locations is similar to normal K-means centroids, where the new location are calculated as the means of each GU location: $\boldsymbol{q}_m$ assigned to each cluster i:

$$C_i^{(t+1)} = \frac{1}{n_i} \sum_{\boldsymbol{o}_m^i \in C_i^{(t)}} \boldsymbol{o}_m^i, \tag{7}$$

where $\boldsymbol{o}_m^i$ represents GU m location in cluster i and C_i denotes the UAV (centroid) location. The edge weight is the distance between the UAV and GU, which is updated following the update step in each iteration. The description of BKMC is given in Algorithm 1.

B. Successive Convex Approximation

Given the optimal GUs association $\boldsymbol{\alpha}^*$ and UAVs trajectory $\boldsymbol{q_m}$, **P1** is transformed into **P1.2a** for optimizing transmit power $\boldsymbol{p}$, which is non-convex programming. Mathematically, it can be defined as:

$$\mathbf{P1.2a:} \max_{\boldsymbol{p}} \quad R_k^{\text{lo}}(n) \tag{8a}$$

$$\text{s.t.} \quad (5b), (5e), \text{ and } (5f). \tag{8b}$$

To transform the non-convex objective, we apply SCA, which is based on the first-order Taylor approximation, knowing that

this provides the global upper bound for the concave function. To develop the SCA-based algorithm, we rewrite the log-term in objective (5a) in the form of the difference of two concave (D.C.) functions, which can be defined as:

$$R(\boldsymbol{p}) = \sum_{k=1}^{K} R_k^{lo} = \log_2\left(1+\gamma_{k,m}\right) = l(\boldsymbol{p}) - h(\boldsymbol{p}), \tag{9}$$

where

$$l(\boldsymbol{p}) = \sum_{k=1}^{K}\sum_{m=1}^{M} \log_2\left(p_{k,m}(n)h_0\right), \tag{10}$$

and

$$h(\boldsymbol{p}) = \sum_{k=1}^{K}\sum_{m=1}^{M} \log_2\left(\psi_{k,m} + B_{k,m}d_{k,m}^2(n)e^{ad_{k,m}}\sigma^2\right). \tag{11}$$

The functions in (10) and (11) are concave, but the difference between them, captured in (9), is neither convex nor concave. Therefore, we presented the SCA, which can calculate the concave lower bound, i.e., the surrogate function for the non-concave objective given in (9), by providing a feasible solution $\boldsymbol{p'}$ of the problem (5). We design its lower bound with substituting $h(\boldsymbol{p})$ by their first-order Taylor approximation which can be defined as:

$$R(\boldsymbol{p}, \boldsymbol{p'}) = l(\boldsymbol{p}) - \tilde{h}((\boldsymbol{p}, \boldsymbol{p'})), \tag{12}$$

where

$$\tilde{h}((\boldsymbol{p}, \boldsymbol{p'})) \triangleq h(\boldsymbol{p'}) - \nabla h(\boldsymbol{p'})(\boldsymbol{p} - \boldsymbol{p'}), \tag{13}$$

where (13) is the first-order Taylor's expansion of $h(\boldsymbol{p})$ near the given point $\boldsymbol{p'}$ in the feasible region of the solution space, and $\nabla h(\boldsymbol{p'})$ denotes the gradient of the $h(\boldsymbol{p})$ at $\boldsymbol{p'}$. The gradient for UAV k can be given as:

$$\nabla_k h(\boldsymbol{p'}) = \frac{\partial h(\boldsymbol{p'})}{\partial p_k'} = \frac{1}{\ln 2} \sum_{\forall k' \neq k}\sum_{\forall m' \neq m} \frac{h_0}{\psi_{k,m} + \alpha_{k,m}B_{k,m}d_{k,m}^2 e^{-ad_{k,m}n}\sigma^2}. \tag{14}$$

The surrogate function is concave given in (13). Additionally, we can find the upper bound of function $h(\boldsymbol{p})$ by the first-order Taylor's expansion as:

$$h(\boldsymbol{p}) \leq h(\boldsymbol{p'}) + \nabla h(\boldsymbol{p'})(\boldsymbol{p} - \boldsymbol{p'}). \tag{15}$$

By analysing (9), (12), and (15), we can deduce the following observations:

$$\begin{aligned} R(\boldsymbol{p}) &= l(\boldsymbol{p}) - h(\boldsymbol{p}) \\ &\geq l(\boldsymbol{p}) - \{h(\boldsymbol{p'}) + \nabla h(\boldsymbol{p'})(\boldsymbol{p} - \boldsymbol{p'})\} \\ &\geq l(\boldsymbol{p}) - h(\boldsymbol{p'}) - \nabla h(\boldsymbol{p'})(\boldsymbol{p} - \boldsymbol{p'}) \\ &= R(\boldsymbol{p}, \boldsymbol{p'}). \end{aligned} \tag{16}$$

where (16) represents that the surrogate function provide the lower bound of the original function. Therefore, these two functions are tangent at point $\boldsymbol{p'}$, i.e., $R(\boldsymbol{p}, \boldsymbol{p'})|_{p=p'}$=$R(\boldsymbol{p'})$. Thus, the function in (16) can provide the lower bound for the original objective function in (5). Therefore, we replace the non-concave objective in (5) with its surrogate function given (12). Afterward, we modify the surrogate objective function and found the convex problem which can be given as:

$$\textbf{P1.2b:} \min_{\boldsymbol{p}} \quad -R_k^{\text{lo}}(n) \tag{17a}$$

$$\text{s.t.} \quad (5b), (5e), \text{ and } (5f). \tag{17b}$$

Algorithm 2 SCA for Transmit Power Optimization (17)

1: **Input:** $p_{k,m}^{\max}$, $\boldsymbol{p}^0$, iteration $j = 0$, tolerance χ, stopping criterion $e = 1$.
2: $j \leftarrow 0$
3: **while** e $\geq \chi$ **do**
4: Designed $R(\boldsymbol{p}, \boldsymbol{p'}) = l(\boldsymbol{p}) - \tilde{h}((\boldsymbol{p}, \boldsymbol{p'}))$ based on (12).
5: Solve (17) and find the $\boldsymbol{p}^{j+1}$.
6: Calculate the stopping criterion $e = |R(\boldsymbol{p}^{j+1}) - R(\boldsymbol{p}^j)|$.
7: Update the iteration counter i.e., $j = j + 1$.
8: **end while**
9: **Output:** Optimal transmit power $\boldsymbol{p}^*$.

Hence, problem (17) is convex, and we can solve it using optimization solvers, i.e., CVXPY. We observe that the optimal solution $\boldsymbol{p}^*$ in each iteration provides the new reference point for $\boldsymbol{p'}$. By updating the value of $\boldsymbol{p'}$, we can define a new surrogate function and find the new approximated convex problem. This procedure will converge iteratively till the convergence criteria value χ is met. The details of SCA-approximation are shown in Algorithm 2. Step 3 represents the convex approximation and step 4 obtains the successive update based on the newly obtained solution respectively in Algorithm 2.

C. Proximal Policy Optimization

The UAV trajectory optimization $\boldsymbol{q}$ problem is still non-convex by giving the optimal user association $\boldsymbol{\alpha}^*$ and transmit power $\boldsymbol{p}^*$. Mathematically, this problem can be defined as:

$$\textbf{P1.3:} \max_{\boldsymbol{q}} \quad R_k^{\text{lo}}(n) \tag{18a}$$

$$\text{s.t.} \quad (5b), (5c), (5g), \text{ and } (5h). \tag{18b}$$

To address the challenge of dynamic UAV trajectory optimization, we propose a PPO-based DRL. PPO-DRL is based on the substitution of flexible constraints for hard constraints, which are seen as penalties. The new, more manageable constraints are utilized to solve a first-order differential problem that approximates the second-order optimization differential equation. The Kullback–Leibler (KL) divergence is used to calculate policy changes at each time step [19]. PPO seeks to compute an update that minimizes the cost function while keeping the divergence from the prior policy to a minimum at each time step. Trust region policy optimization (TRPO) recommends optimizing surrogate loss and limiting the amount of the update using KL. The PPO objective implements a method for updating the trust region that is consistent with stochastic gradient descent (SGD) and simplifies the algorithm by eliminating the KL penalty and the necessity for adaptive updates. Thus, PPO offers the advantages of being simple to apply and having a reduced level of complexity, which is appropriate for our highly dynamic environment. The surrogate objective function in TRPO can be maximized subject to a size limit on the policy update, which can be described as [20]:

$$L^{\text{TRPO}}(\theta) = \hat{E}_t\left[\frac{\pi_\theta(a_t|s_t)}{\pi_{\text{old}}(a_t|s_t)}\hat{A}_t\right] = \hat{E}_t\left[r_t(\theta)\hat{A}_t\right], \tag{19}$$

where π_θ is a stochastic policy, $\hat{A}_t$ is an estimator of the advantage function at step t, θ_{old} is the vector of policy parameters before the update, $r_t(\theta)$ denotes the probability ratio, which can be define as:

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\text{old}}(a_t|s_t)}, \tag{20}$$

and $\hat{E}_t[...]$ indicates an expectation of the empirical average over a finite batch of samples in the algorithm that alternates between sampling and optimization. At this point, learning fails or performance suffers as a result of an excessive rise in the probability ratio $r_t(\theta)$, and in the case of TRPO, the problem is avoided by utilizing KL-Divergence to impose a penalty. However, TRPO has the drawback of being conceptually difficult to grasp, and hence difficult to apply. PPO, like TRPO, optimizes the surrogate objective function via stochastic gradient ascent (SGA). Thus, PPO used computationally efficient penalties and avoided unnecessary policy changes by employing a technique known as the clipped surrogate objective. The clip is a function that determines the minimum and maximum values of a given variable. The probability ratio $r_t(\theta)$ can be trimmed by PPO. The following is the objective function to which the clipping is applied:

$$L^{\text{PPO}}(\theta) = \hat{E}_t \left[\min(r_t(\theta)\hat{A}_t, \text{clip}(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}_t) \right], \tag{21}$$

where ϵ is a hyper parameters. The function in (21) compares the objectives used in the existing TRPO with the objectives to which clipping is employed and takes a smaller value.

Another feature of the PPO method is that, unlike many other DRL algorithms, the PPO-Actor-Critic networks contain a type of parameter sharing, which allows each objective to be combined into a single goal function and optimized at the same time. The methods for power allocation mentioned above are added to the function for computing rewards once the actor-critic network has been updated and learned. With this, it was possible to effectively separate power allocation and UAV trajectory and proceed with optimization. Next, the state, compensation, and behavior of the agent used in learning will be described. In learning step t, the state $s_t(n)$ is defined as:

$$s_t(n) = \{\{\mathbf{q}_k(n)\}_{k\in\mathcal{K}}, \{\mathbf{o}_m\}_{m\in\mathcal{M}}\}, \tag{22}$$

where $\mathbf{q}_k(n)$ and $\mathbf{o}_m(n)$ denotes UAV k and GU m locations, respectively, at time slot n. Therefore, agent takes optimal action through network with all location information input without any other information. The action in learning step t at time slot n is the speed and the moving direction as follow:

$$a_t(n) = \{\{v_k(n), \phi_k(n)\}_{k\in\mathcal{K}}\}. \tag{23}$$

Specifically, we restricted the direction $\phi_k(n)$ between $[-\pi/3, \pi/3]$ to prevent UAVs from turning sharply. Lastly, the reward in learning step t at time slot n is divided into three and can be expressed as follows:

$$r_t(n) = \begin{cases} 2, & \text{if } t = \text{max step}, \\ -2, & \text{if } \exists\ i,j \in \mathcal{K} \\ & \text{s.t. } \|\mathbf{q}_i(n) - \mathbf{q}_j(n)\| < D_{\min}, \\ \sum_{k=1}^{K} R_k^{\text{lo}}(n), & \text{otherwise.} \end{cases} \tag{24}$$

Algorithm 3 PPO-DRL for UAVs Trajectory Optimization (18)

for episode$= 1, 2, ..., E$ **do**
  Initialize randomly each GU's positions
  GUs Association $\boldsymbol{\alpha}$ by **Algorithm 1**
  for actor$= 1, 2, ..., A$ **do**
    for time slot$= 1, 2, ..., N$ **do**
      Run policy $\pi_{\theta_{\text{old}}}$ in environment
      Optimal Power Allocation $\boldsymbol{\mathcal{P}}$ by **Algorithm 2**
      Save (s_n, a_n, r_n, s_{n+1}) in Trajectory memory
    end for
    Compute advantage estimates $\hat{A}_1, ..., \hat{A}_N$
  end for
  Optimize surrogate L^{PPO} wrt θ, with minibatch from Trajectory memory
  $\theta_{\text{old}} \leftarrow \theta$
end for
Output: The optimal PPO network $\pi_{\theta_{\text{opt}}}$

The episode finishes when learning step t reaches the maximum step with a reward of 2 and if the distance between UAVs is as near to the minimum distance then the penalty will be -2. Otherwise, it receives the total of the minimum data rate given by each UAV with the proposed power allocation and GU association algorithm.

The pseudocode for the PPO learning process for UAV trajectory is depicted in Algorithm 3. We proceed with the learning of the E number of episodes. The episode begins by randomly arranging the locations of GUs. Algorithm 1 determines the optimal GU association $\boldsymbol{\alpha}^*$ based on the location of the arranged GUs. In subsequent learning, A actors simultaneously generate data on the environment. Each actor acts under prior policy $\pi_{\theta_{\text{old}}}$ and employs Algorithm 2 to determine optimal power $\boldsymbol{p}^*$ and derive rewards from it. Trajectory memory stores the state, action, reward, and future state of time step n. When the time step is completed, we calculate the advantage estimates $\hat{A}_n$ for each time step n. Trajectory memory is used to extract minibatch, which is then optimized for the target function L^{PPO}. The learning process is repeated by substituting the updated parameter θ with the old value θ_{old}. As a result, we can obtain a PPO network $\pi_{\theta_{\text{opt}}}$ for UAVs trajectory.

D. Algorithms Complexities and Convergences

Based on the definition mentioned above, the detailed process of resource allocation via BKMC and SCA is proposed as shown in Algorithm 2 and Algorithm 3 respectively. The computational complexity of BKMC and SCA is $O(K^3)$ and $O(K)$ respectively. Moreover, after obtaining the optimal resource allocation, we execute PPO-DRL as mentioned in Algorithm 3 to get the optimal UAVs trajectories policies and its complexity is $O(Ka^2)$, where a represents the number of actions. It can be observed that the proposed algorithms converged to the sub-optimal solutions.

TABLE I: Simulation Parameters

Parameter	Value	Parameter	Value
Bandwidth	B=0.1 THz	Channel gain at ref.	h_0=−40 dBm
Noise power	σ^2=−174 dBm/Hz	Max. transmit power	$P^{\max}$=2 W
Minimum rate	$R^{\min}$=0.02 Tbps	Absorption coefficient	$a(f)$=0.005
Episodes	E=5e + 5	Batch size	120
Discount factor	γ =0.99	Learning rate	0.0003
Clipping ϵ	0.2	Regularizer parameter	$\lambda = 0.95$
Epochs	3	Hidden layer's units	128
Hidden layers	2	Carrier Frequency	f=1.2 THz [13]

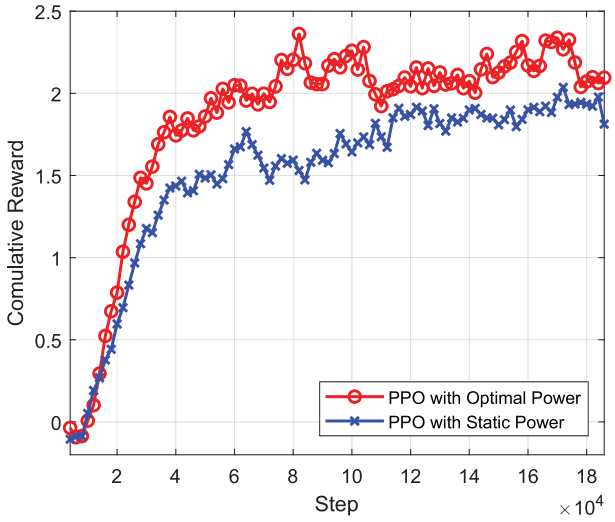

Fig. 2: PPO learning results (reward).

IV. Simulation Results

In our network configuration, we consider 3 UAVs and 36 GUs distributed by following homogeneous PPP within a 200 m × 200 m area in each episode. Moreover, UAVs are assumed to be hovering at a fixed altitude of z_k=20 m. The maximum speed of the UAVs is 5 m/s. Other parameters are listed in Table I. All statistical findings are averaged after several simulation runs. To assess the performance of our proposed algorithm, we consider four benchmark algorithms as follows:

- **SU with RP**: The algorithm which considers static UAVs (SU) positions with the random power (RP) allocation.
- **OU with RP**: The algorithm uses the optimal UAVs (OU) trajectory with the random power (RP) allocation.
- **SU with PP**: The algorithm assumes the static UAVs (SU) positions with the proposed power (PP) allocation.
- **OU with PP (proposed method)**: The algorithm considers the optimal UAV (OU) trajectory with the proposed power (PP) allocation.

Fig. 2 shows the convergence performance of the cumulative reward with two power allocation schemes, i.e., optimal and static. Initially, both schemes provide fewer rewards, but after the convergence, the proposed schemes provide better results than the static schemes.

Fig. 3 provides a comparison of the proposed algorithm (OU with PP) with three benchmarks, i.e., SU with RP, OU with RP, and SU with PP in terms of the average achievable rate for GUs.

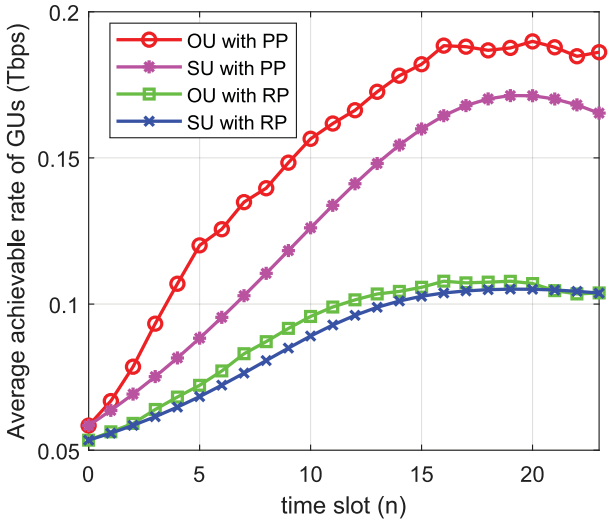

Fig. 3: Achievable rate with benchmarks schemes.

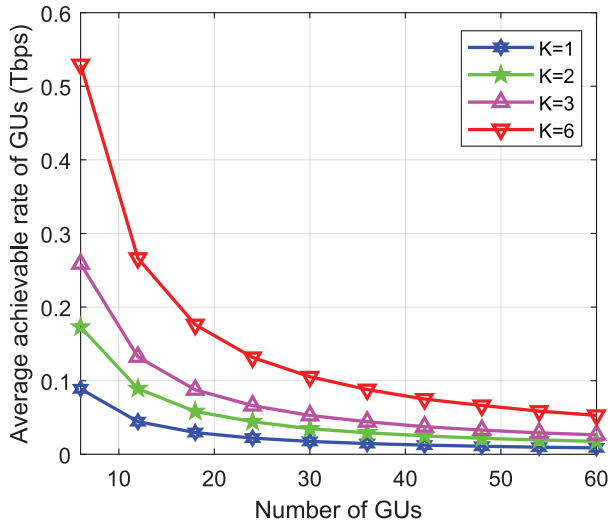

Fig. 4: Achievable rate with UAVs.

It can be observed that OU with PP outperforms the rest of the three algorithms just after a few time slots.

Moreover, Fig. 4 depicts the average achievable rate of GUs by varying the number of deployed UAVs in the area. Since all UAVs need to fulfill the QoS of GUs, and therefore, as the number of GUs increases in the area, network performance degrades. But, it can be observed that as the number of UAVs increases, the average performance increases.

The resultant UAV trajectory with the proposed power and GU allocation algorithm is shown in Fig. 5. It is clear that the dots represent the GUs, and their clusters are separated by distinct colors. Each UAV follows the PPO-DRL algorithm to optimize its trajectory from an initialized location to the endpoint in time slot $n = 1$ to $n = 25$. Each UAV follows a distinct trajectory while providing optimal network resources to the GU.

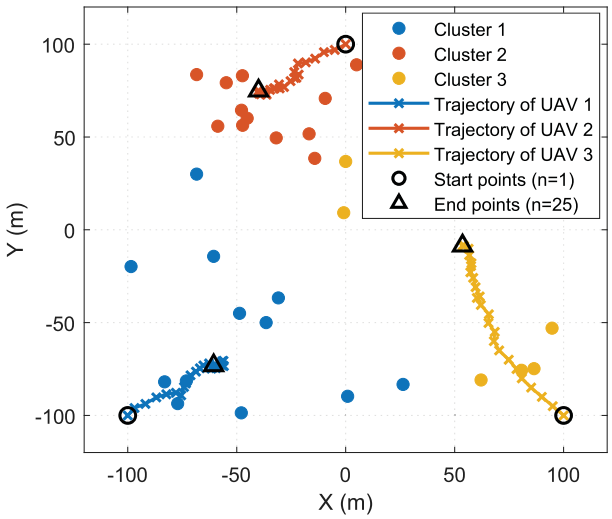

Fig. 5: UAVs trajectory obtained by PPO-DRL.

V. Conclusions

In this article, we have explored THz-enabled UAVs to facilitate ubiquitous 6G mobile communication networks. The molecular absorption effect has been explicitly incorporated in the THz-enabled UAV channel gain model. Then, we have formulated an optimization problem to optimize the average throughput of deployed UAVs by enhancing UAV-GU association, transmit power, and trajectories while satisfying the GU's demands. To address this problem, we have proposed an iterative algorithm that separates the original problem into three subproblems. Firstly, to tackle the UAVs-GUs association problem, we have employed the BKMC algorithm. To deal with the optimal transmit power, we have utilized the SCA-based algorithm. To handle dynamic UAV trajectories optimization, a PPO-DRL-based algorithm is has been designed, which can make quick decisions in the given environment owing to its low complexity. Based on the experience replay and target networks, the PPO method has efficiently learned the optimal trajectory with fast convergence speed. The simulation results have shown that our proposed algorithms outperform the other baselines.

References

[1] W. Saad, M. Bennis, and M. Chen, "A vision of 6G wireless systems: Applications, trends, technologies, and open research problems," *IEEE Network*, vol. 34, no. 3, pp. 134–142, June 2020.

[2] S. S. Hassan, Y. K. Tun, W. Saad, Z. Han, and C. S. Hong, "Blue data computation maximization in 6G space-air-sea non-terrestrial networks," in *proceedings of the IEEE Global Communications Conference (GLOBECOM)*, Madrid, Spain, Dec. 2021, pp. 1–6.

[3] S. S. Hassan, D. H. Kim, Y. K. Tun, N. H. Tran, W. Saad, and C. S. Hong, "Seamless and energy efficient maritime coverage in coordinated 6G space-air-sea non-terrestrial networks," *arXiv preprint arXiv:2201.08605*, 2022.

[4] Z. Zhang, Y. Xiao, Z. Ma, M. Xiao, Z. Ding, X. Lei, G. K. Karagiannidis, and P. Fan, "6G wireless networks: Vision, requirements, architecture, and key technologies," *IEEE Vehicular Technology Magazine*, vol. 14, no. 3, pp. 28–41, July 2019.

[5] K. M. S. Huq, J. Rodriguez, and I. E. Otung, "3D network modeling for THz-enabled ultra-fast dense networks: A 6G perspective," *IEEE Communications Standards Magazine*, vol. 5, no. 2, pp. 84–90, June 2021.

[6] C. Chaccour, M. N. Soorki, W. Saad, M. Bennis, and P. Popovski, "Can Terahertz provide high-rate reliable low latency communications for wireless VR?" 2021.

[7] M. Mozaffari, W. Saad, M. Bennis, Y.-H. Nam, and M. Debbah, "A tutorial on UAVs for wireless networks: Applications, challenges, and open problems," *IEEE Communications Surveys Tutorials*, vol. 21, no. 3, pp. 2334–2360, 2019.

[8] S. S. Hassan, Y. M. Park, and C. S. Hong, "On-Demand MEC empowered UAV deployment for 6G time-sensitive maritime internet of things," in *proceedings of the 22nd Asia-Pacific Network Operations and Management Symposium (APNOMS)*, Tainan, Taiwan, Sep. 2021, pp. 386–389.

[9] Z. Yang, C. Pan, M. Shikh-Bahaei, W. Xu, M. Chen, M. Elkashlan, and A. Nallanathan, "Joint altitude, beamwidth, location, and bandwidth optimization for UAV-enabled communications," *IEEE Communications Letters*, vol. 22, no. 8, pp. 1716–1719, June 2018.

[10] J.-M. Kang and C.-J. Chun, "Joint trajectory design, Tx power allocation, and Rx power splitting for UAV-enabled multicasting SWIPT systems," *IEEE Systems Journal*, vol. 14, no. 3, pp. 3740–3743, Jan. 2020.

[11] W. Wang, H. Dai, C. Dong, X. Cheng, X. Wang, P. Yang, G. Chen, and W. Dou, "Placement of unmanned aerial vehicles for directional coverage in 3D space," *IEEE/ACM Transactions on Networking*, vol. 28, no. 2, pp. 888–901, Mar. 2020.

[12] X. Wang, P. Wang, M. Ding, Z. Lin, F. Lin, B. Vucetic, and L. Hanzo, "Performance analysis of terahertz unmanned aerial vehicular networks," *IEEE Transactions on Vehicular Technology*, vol. 69, no. 12, pp. 16 330–16 335, 2020.

[13] L. Xu, M. Chen, M. Chen, Z. Yang, C. Chaccour, W. Saad, and C. S. Hong, "Joint location, bandwidth and power optimization for THz-enabled UAV communications," *IEEE Communications Letters*, vol. 25, no. 6, pp. 1984–1988, Mar. 2021.

[14] Z. Yuan and G.-M. Muntean, "AirSlice: A network slicing framework for UAV communications," *IEEE Communications Magazine*, vol. 58, no. 11, pp. 62–68, Nov. 2020.

[15] H. Wang, J. Wang, G. Ding, J. Chen, Y. Li, and Z. Han, "Spectrum sharing planning for full-duplex UAV relaying systems with underlaid D2D communications," *IEEE Journal on Selected Areas in Communications*, vol. 36, no. 9, pp. 1986–1999, Aug. 2018.

[16] J. Hu, H. Zhang, L. Song, Z. Han, and H. V. Poor, "Reinforcement learning for a cellular internet of uavs: Protocol design, trajectory control, and resource management," *IEEE Wireless Communications*, vol. 27, no. 1, pp. 116–123, Mar. 2020.

[17] M. I. Malinen and P. Fränti, "Balanced k-means for clustering," in *Joint IAPR International Workshops on Statistical Techniques in Pattern Recognition (SPR) and Structural and Syntactic Pattern Recognition (SSPR)*. Springer, 2014, pp. 32–41.

[18] R. Burkard, M. Dell'Amico, and S. Martello, *Assignment Problems. Revised reprint.* SIAM - Society of Industrial and Applied Mathematics, 2012, 393 Seiten.

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

[20] J. Schulman, S. Levine, P. Abbeel, M. Jordan, and P. Moritz, "Trust region policy optimization," in *Proceedings of the 32nd International Conference on Machine Learning*, F. Bach and D. Blei, Eds., vol. 37. Lille, France: PMLR, 07–09 Jul 2015, pp. 1889–1897. [Online]. Available: https://proceedings.mlr.press/v37/schulman15.html