

Title	Sum rate maximization in downlink HAP-RSMA-based THz systems: A generative diffusion model enabled RL approach
Authors	Le, Mai;Pham, Quoc-Viet;O'Sullivan, Barry;Nguyen, Hoang D.
Publication date	2025
Original Citation	Le, M., Pham, Q.-V., O'Sullivan, B. and Nguyen, H. D. (2025) 'Sum rate maximization in downlink HAP-RSMA-based THz systems: A generative diffusion model enabled RL approach', IEEE Global Communications Conference, Taipei, Taiwan, 8-12 December 2025.
Type of publication	Conference item
Rights	© 2025, the Authors. For the purpose of Open Access, a CC BY licence applies to any Author Accepted Manuscript from this submission. - https://creativecommons.org/licenses/by/4.0/
Download date	2026-10-27 12:58:51
Item downloaded from	https://hdl.handle.net/10468/18352

Sum Rate Maximization in Downlink HAP-RSMA-based THz Systems: A Generative Diffusion Model enabled RL Approach

Mai Le*, Quoc-Viet Pham[†], Barry O'Sullivan*, Hoang D. Nguyen*

*Insight Centre for Data Analytics, School of Computer Science & IT, University College Cork, Ireland

[†]School of Computer Science and Statistics, Trinity College Dublin, Dublin 2, D02 PN40, Ireland

E-mails: mle@ucc.ie, viet.pham@tcd.ie, {b.osullivan, hn}@cs.ucc.ie

Abstract—This paper investigates the maximization of the achievable rate for users served by a high-altitude platform (HAP) acting as a flying base station in the downlink of rate-splitting multiple access (RSMA)-based terahertz (THz) communication systems. Considering the dynamic and uncertain environment caused by user mobility and molecular absorption effects, we propose a generative diffusion model (DM)-based deep reinforcement learning approach to address this challenge. The problem is formulated as a Markov decision process, aiming to maximize the long-term achievable rate for all users by jointly optimizing power allocation and the common rate splitting ratio. Moreover, the generative DM significantly improves the decision-making capabilities of a deep reinforcement learning algorithm, namely the deep deterministic policy gradient (DDPG). Experimental simulations demonstrate the effectiveness of the proposed DM-DDPG algorithm compared to alternative schemes.

Index Terms—Deep learning, generative diffusion model, high-altitude platform, resource allocation, rate spilling multiple access, reinforcement learning, THz communications.

I. INTRODUCTION

Nowadays, the rapid proliferation of new applications and services necessitates a significant upgrade of wireless network communication capabilities to meet service demands effectively [1]–[3]. The non-terrestrial network has been investigated to overcome these challenges in providing new services, reducing the burden on the terrestrial base stations (BSs), and offering a new set of intelligent services from the sky. Specifically, the high altitude platform (HAP) potentially provides the extension coverage with powerful of large-array antenna, light weight infrastructure [3].

The resource scarcity under ever-increasing requirements of new services and applications in sixth-generation (6G) networks necessitates the usage of a higher frequency band and new multiple access schemes. As such, terahertz (THz) communications and rate split multiple access (RSMA) methods have emerged as key enablers of 6G [1], [2]. By dividing user messages into common parts intended for all users and private parts only intended for the corresponding users, RSMA enables efficient resource sharing and superior interference management compared to conventional multiple access methods. This leads to significant improvements in spectral and energy efficiency, user fairness, system complexity, and overall quality of service, thanks to its flexible rate-splitting and successive interference cancellation (SIC) techniques. Meanwhile, THz communications attracted considerable research interest for providing ultra-high data rates in future 6G networks [2].

Some works on RSMA/THz in non-terrestrial communication networks have been carried out. For instance, authors in [4] investigated the RSMA scheme in aerial access networks, achieving the highest performance among benchmarks. Le *et al.* [5] explored the non-orthogonal multiple access method in HAP-based THz communications and demonstrated the network capability to provide massive connectivity to user equipments (UEs) with higher data rate.

The communication environment is characterized by the uncertainties of a highly dynamic wireless setting, including the mobility and the molecular absorption effects in THz communications. To overcome these challenges, deep reinforcement learning (DRL)-based algorithms have been introduced to make intelligent decisions, catching up with changes in the dynamic communication systems [6]–[8]. However, in THz non-terrestrial networks, the methods proposed by the existing studies aim to either learn an optimal policy that generates actions according to the input data distribution [9], [10] or analyze the network performance [11]–[14]. As a result, the decision-making process of these methods often relies on unimodal policy distributions, leading to suboptimal results and hindering capturing of environmental changes and intricate situations [15]. Generative diffusion model (DM)-based DRL has been regarded as an effective solution in addressing these challenges for complex resource allocation problems, thanks to the generation of diverse data types [4], [6], [8]. However, such an effective approach has been ignored in highly dynamic and complex HAP-RSMA enabled THz networks, which is vital for performance improvement.

In this work, we investigate the use of the DM to tackle the aforementioned issues in HAP-RSMA enabled THz communications. We leverage the advantages of generative DM-based DRL, especially its ability to balance exploration and exploitation phases [16], to solve the resource allocation problem by maximizing the achievable data rates of UEs in the downlink of HAP-RSMA THz networks. The main contributions and features of this paper are summarized below.

- We propose to optimize the long-term sum achievable data rate maximization problem in HAP-RSMA-based THz networks by jointly optimizing the power allocation and common rate splitting ratio, subject to the SIC and the minimum rate requirement of each user under the dynamic environment. We then reformulate the problem as a Markov decision process (MDP) framework.

- We develop a generative DM-based deep deterministic policy gradient (DM-DDPG) method to solve the long-term, non-convex optimization problem. Specifically, the DM can act as an direct alternative actor network in DDPG to dynamically generate the optimal variables, thereby significantly enhancing the network performance.
- We evaluate the performance of our proposed algorithm through numerical simulations. Notably, the results indicate that our DM-DDPG method achieves larger rewards and outperforms the benchmark schemes in terms of data rate in the setting of different power budgets.

II. SYSTEM MODELS AND PROBLEM FORMULATION

We consider a downlink HAP-RSMA-based THz communications system, consisting of an HAP as a flying BS serving multiple UEs. The HAP is equipped with N antennas, indexed by $n \in \mathcal{N} = \{1, \dots, N\}$, and serves K single-antenna UEs, indexed by $k \in \mathcal{K} = \{1, \dots, K\}$. The system operates in a horizon comprised of T time slots with equal duration indexed by $t \in \mathcal{T} = \{1, \dots, T\}$. Assuming HAP flies at the center of the serving area and at a fixed altitude, as illustrated in Fig. 1.

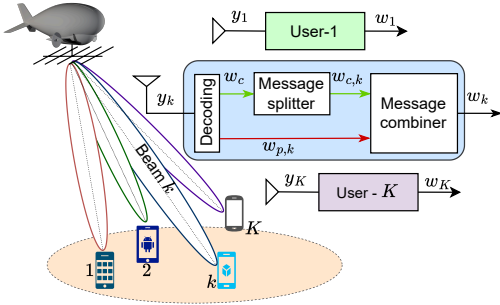


Fig. 1. The system model illustration.

In the RSMA scheme, the intended message of the UE k at each time slot t , denoted as $w_{k,t}$ is divided into the common part $w_{c,k,t} = \{w_{c,1,t}, \dots, w_{c,K,t}\}$ encoded into a common stream $s_{c,t}$ and private parts $w_{p,k,t}$ encoded into a independent private stream $s_{p,k,t}$. The transmit signal model at the HAP can be written as

$$s_t = \sum_{k \in \mathcal{K}} \sqrt{p_{k,t}} s_{p,k,t} + \sqrt{p_{c,t}} s_{c,t}, \quad (1)$$

where $p_{c,t}$ and $p_{k,t}$ are the transmit power assigned to the common and private streams, respectively. The power control at the HAP at time step t is defined as $\mathbf{P}_t = [p_{c,t}, p_{1,t}, \dots, p_{K,t}]$ under constrained of $P_{\text{use},t} = p_{c,t} + \sum_{k=1}^K p_{k,t} \leq P_{\text{max}}$, where P_{max} is the maximum transmit power budget of the HAP.

The achievable rate of the UE k for the common and private streams at a time step t is respectively given as follows:

$$R_{c,k,t} = \log_2 \left(1 + \frac{p_{c,t} |\mathbf{h}_{k,t}^H|^2}{\sum_{i=1}^K p_{i,t} |\mathbf{h}_{k,t}^H|^2 + \sigma^2} \right), \quad (2)$$

$$R_{p,k,t} = \log_2 \left(1 + \frac{p_{k,t} |\mathbf{h}_{k,t}^H|^2}{\sum_{i=1, i \neq k}^K p_{i,t} |\mathbf{h}_{k,t}^H|^2 + \sigma^2} \right), \quad (3)$$

with $\mathbf{h}_{k,t} \in \mathbb{C}^{N \times 1}$ being the channel link between BS and k -th UE and σ^2 referring to the noise power density. Each channel element $[h_{k,t}]_n, \forall n \in \mathcal{N}$ that is calculated as [5]

$$[h_{k,t}]_n = G_{\text{tx}} G_{\text{rx}} \left(\frac{c}{4\pi f} \right)^2 e^{-\varsigma_{\text{abs}} d_{n,k,t}} d_{n,k,t}^{-2}. \quad (4)$$

Here, G_{tx} and G_{rx} represent antenna gains at the transmitter and receiver, respectively, and c is the speed of light. In addition, ς_{abs} is the molecular absorption coefficient depending on the working frequency f and $d_{n,k,t}$ is the distance between the n -th antenna element of the BS and the UE k .

In order to ensure that the common stream $s_{c,t}$ is successfully decoded at all UEs, the actual common rate, denoted as $R_{c,t}$, should not exceed the achievable rate of the weakest UE, i.e., $R_{c,t} \leq \min\{R_{c,k,t}\}, \forall k \in \mathcal{K}$. Let $C_{k,t}$, denote the portion of common rate $R_{c,t}$ allocated for the user k to decode the common signal $W_{c,k,t}$, i.e., $\sum_{k=1}^K C_{k,t} = R_{c,t}$. Therefore, the total achievable of the system at one time step is as $R_t = \sum_{i=1}^K (C_{i,t} + R_{p,i,t})$.

Furthermore, the power allocation different to successfully implement SIC method for decoding the common stream which must satisfy the following constraint [17]:

$$p_{c,t} |\mathbf{h}_{k,t}^H|^2 - \sum_{i=1}^K p_{i,t} |\mathbf{h}_{k,t}^H|^2 - \sigma^2 \geq 0, \forall k, t. \quad (5)$$

In this paper, our objective is to maximize the long-term total achievable rates by jointly optimizing the transmit power allocation $\mathbf{P}_t$ and the portion of the common rate of each UE $\mathbf{C}_t$. Formally, the optimization problem with objective function is redefined to the expected value of discounted cumulative reward with a discount factor $\gamma \in (0, 1)$ over time horizon T , which can be formulated as

$$\max_{\mathbf{P}, \mathbf{C}} \mathbb{E} \left(\sum_{t=1}^T \gamma^{t-1} R_t \right), \quad (6a)$$

$$\text{s.t.} \quad P_{\text{use},t} \leq P_{\text{max}}, \quad (6b)$$

$$p_{c,t} > 0, p_{k,t} > 0, \forall k, t, \quad (6c)$$

$$\sum_{i=1}^K C_{i,t} \leq R_{c,t}, \forall t, \quad (6d)$$

$$C_{k,t} > 0, \forall k, t, \quad (6e)$$

$$C_{k,t} + R_{p,k,t} \geq R_{\min}, \forall k, t, \quad (6f)$$

$$p_{c,t} |\mathbf{h}_{k,t}^H|^2 - \sum_{i=1}^K p_{i,t} |\mathbf{h}_{k,t}^H|^2 - \sigma^2 \geq 0, \forall k, t. \quad (6g)$$

where $\mathbf{C}_t = [C_{1,t}, \dots, C_{K,t}]$ and $R_{\min}$ denotes the minimum rate requirement of UEs. Constraints in (6b) and (6c) ensures the power allocation that is not over the maximum power at the BS. Constraint in (6f) is to ensure the minimum rate requirement of each UE. Due to the non-convexity of the problem and the dynamics and uncertainties of the HAP-RSMA THz communications environment, we employ the generative DM enabled DRL approach to solve the optimization problem in (6), as detailed in the following section.

III. PROPOSED SOLUTION

A. Markov Decision Process Formulation

The problem in (6) is reformulated as an MDP through the action space, state space, and reward function defined below.

1) *State*: We define the system state at one time step as s_t , consisting of the current channel estimation and the common and private rates for all UEs at the previous time step as

$$s_t = [|h_{1,t}^H|^2, \dots, |h_{K,t}^H|^2, R_{c,1,t-1}, \dots, R_{c,K,t-1}, R_{p,1,t-1}, \dots, R_{p,K,t-1}]. \quad (7)$$

2) *Action*: The action, denoted as a_t , is to maximize the cumulative reward at any step t , which can be selected through the interaction between the learning agent with the surrounding environment. The action is constructed by the transmit power P_t and the portion of the achievable rate of the common message C_t . Thus, the action is given by

$$a_t = [P_t, C_t] = [p_{c,t}, p_{1,t}, \dots, p_{K,t}, C_{1,t}, \dots, C_{K,t}]. \quad (8)$$

3) *Reward*: The reward function generally considers both the objective function and constraints of the optimization problem (6). Considering the constraint (6g) that indicates the necessary condition of power allocation for successful SIC operation at the receivers, we first reformulate the common achievable rate for the k -th UE as

$$\bar{R}_{c,k,t} = \begin{cases} R_{k,t}, & \text{diff} \geq 0, \\ 0, & \text{diff} < 0, \end{cases} \quad (9)$$

where diff refers to the left-hand side part of the inequality (5). The actually achievable common rate of the UE k is defined as $R_{c,t} \leq \min(\bar{R}_{c,k,t}), \forall k \in \mathcal{K}$. Accordingly, the reward function is defined as follows:

$$r_t = \sum_{k=1}^K (C_{k,t} + R_{p,k,t}) \times \chi_{\text{pow}} \times \chi_{\text{com}} \times \chi_{\text{qos}}, \quad (10)$$

where

$$\begin{aligned} \chi_{\text{pow}} &= \begin{cases} 1, & P_{\text{use},t} - P_{\text{max}} \leq 0, \\ 0, & P_{\text{use},t} - P_{\text{max}} > 0. \end{cases} \\ \chi_{\text{com}} &= \begin{cases} 1, & \sum_{k=1}^K C_{k,t} - R_{c,t} \leq 0, \\ 0, & \sum_{k=1}^K C_{k,t} - R_{c,t} > 0. \end{cases} \\ \chi_{\text{qos}} &= \begin{cases} 1, & (C_{k,t} + R_{p,k,t}) - R_{\min} \geq 0, \forall k, \\ 0, & (C_{k,t} + R_{p,k,t}) - R_{\min} < 0, \forall k. \end{cases} \end{aligned}$$

Here, χ_{pow} , χ_{com} and χ_{qos} represent the penalties for actions not meeting the constraints in (6b), (6d) and (6f), respectively. When all the constraints are satisfied, the reward value is represented by the total achievable rates.

In RL, the Q function is used to evaluate the long-term accumulative reward of the system returning from the state s taking action a under the policy π . With γ regarding to the discount factor in the range $[0, 1]$, it has a form as follows:

$$Q(s, a) = \mathbb{E} \left[\sum_{t=1}^K \gamma^{t-1} r_t | s_1 = s, a_1 = a \right]. \quad (11)$$

With the aim of finding an optimal policy π^* mapping directly from states to actions can be found by maximizing the optimal Q function over all the states to maximize the long-term reward, i.e., $\pi^*(s) = \arg\max_a \{Q(s, a)\}$. The optimal Q function is given as

$$Q^*(s, a) = \mathbb{E}(r_t + \gamma \max_{a'} Q^*(s', a')), \quad (12)$$

where s' and a' are the next state and action after agent takes action a given by state s , respectively.

B. The DDPG-based algorithm

From the formulation of the MDP above, we observe that the decision variables are continuous values. Hence, the optimization problem may have a large state and action space due to massive connectivity in future network systems. In this context, DDPG is an efficient algorithm dealing with the high-dimensional issue and continuous action spaces [18]. In general, DDPG is an extended version of an actor-critic network, including two main deep neural networks (DNNs).

- An actor network, parameterized by θ_a , make an action given a input state s , as $a_t = \pi(s_t; \theta_a)$.
- A critic network, parameterized by θ_c , refers to a learned function, which evaluates how good the action a is. The $Q(s_t, a_t; \theta_c)$ function is used to formulate this network.

Moreover, DDPG uses two more DNNs, a target actor $\hat{\pi}(s_t; \theta'_a)$ and a target critic $\hat{Q}(s_t, a_t; \theta'_c)$, to improve the stability of learning process, with θ'_a and θ'_c being the target actor and target critic networks, respectively. At each time slot t , a tuple (s_t, a_t, r_t, s'_t) experience including the current action a_t given a state s_t with received reward r_t , and the next state s'_t is stored in a replay buffer with a limited size R . The oldest sample will be discarded if the buffer is full and the new data is collected. A random mini-batch of B experiences from the buffer is sampled to train the actor and critic.

During the learning process, the actor is trained to maximize the objective function as

$$J(\theta_a) = Q(s_t, a_t = \pi(s_t, \theta_a), \theta_c). \quad (13)$$

The network parameters are updated using a stochastic gradient descent method [18]

$$\nabla_{\theta_a} J(\theta_a) = \nabla_{\theta_c} Q(s_t, a_t, \theta_c) \nabla_{\theta_a} \pi(s_t, \theta_a). \quad (14)$$

The critic parameter is iteratively updated to minimize the loss function $L(\theta_c) = (Y_t - Q(s_t, a_t, \theta_c))^2$. Here, Y_t can be calculated based on target networks and the past experiences stored in the replay buffer as $Y_t = r_t + \gamma Q(s'_t, \hat{\pi}(s'_t; \theta'_a), \theta'_c)$. Two parameters of the target networks are softly updated as

$$\theta'_a \leftarrow (1 - \tau)\theta_a + \tau\theta'_a, \theta'_c \leftarrow (1 - \tau)\theta_c + \tau\theta'_c, \quad (15)$$

where a soft-updating coefficient $\tau \ll 1$ ensures the target values change slowly, thus improving the learning stability.

C. Diffusion Model-based Actor Network

Next, we embed the DM as a denoising diffusion probabilistic model into the learning policy on continuous action. The integration of DM into the DDPG algorithm significantly enhances decision-making by providing a more diverse set of potential actions. Let $a_0 = \pi_{\theta_a}(s)$ be the distribution of joint optimal variables of power allocation and common rates given the environment state s . The forward process generates a series of samples $a_1, \dots, a_m, \dots, a_M$ by systematic addition of Gaussian noise to the initial sample a_0 over M period.

Specifically, at each time step m , the noisy sample a_m is sampled from the distribution $p(a_m|a_{m-1})$. The transition process is represented as

$$p(a_m|a_{m-1}) = \mathcal{N}(a_m; \sqrt{1 - \beta_m}a_{m-1}, \beta_m \mathbf{I}), \quad (16)$$

where $\mathbf{I}$ is an identity matrix and the forward process variance $\beta_m = 1 - e^{-\frac{\beta_{\min}}{M} - \frac{2m-1}{2M^2}(\beta_{\max} - \beta_{\min})}$ is controlled by the variational posterior (VP) scheduler with $\beta_{\max}$ and $\beta_{\min}$ being the constant parameters that control the noise level [16]. The joint distribution of $a_1, \dots, a_M$ conditioned on a_0 can be expressed as follows:

$$p(a_1, \dots, a_M|a_0) = \prod_{m=1}^M p(a_m|a_{m-1}). \quad (17)$$

It is noted that when the execution time M is large, the computational complexity increases significantly [16]. Let $\alpha_m = 1 - \beta_m$ and $\bar{\alpha}_m = \prod_{j=1}^m \alpha_j$, the relationship between a_m and a_0 simply express as

$$a_m = \sqrt{\bar{\alpha}_m}a_0 + \sqrt{1 - \bar{\alpha}_m}\epsilon_0, \quad (18)$$

where $\epsilon_0, \dots, \epsilon_{m-1} \in \mathcal{N}(0, \mathbf{I})$. The forward diffusion process iteratively injects the noise into the optimal value (*i.e.*, the original data), which is impractical for our considerable scenarios, where we aim to discover the relationship between the generated output action and the input state of the DDPG network. Therefore, we only deploy the reverse diffusion process to improve the quality of making decisions by actor networks of DDPG. Meanwhile, the reverse diffusion process aims to obtain the original data $a_0 \sim \pi_{\theta_a}(s)$ from the Gaussian noisy data a_M by iteratively removing the noise through M steps. Let $p(a_{m-1}|a_m)$ be the transition probability from a_m to a_{m-1} , which is unknown yet. It has been proven that the reverse process can be modeled as a Gaussian distribution with actor parameterized model θ_a , and expressed as [16]

$$p_{\theta_a}(a_{m-1}|a_m) = \mathcal{N}(a_{m-1}, \mu_{\theta_a}(a_m, m, s), \tilde{\beta}_m \mathbf{I}), \quad (19)$$

$$\tilde{\beta}_m = \frac{1 - \bar{\alpha}_{m-1}}{1 - \bar{\alpha}_m} \beta_m.$$

Here, μ_{θ_a} is the mean value at the diffusion step that can be predicted by DNNs and is calculated as

$$\mu_{\theta_a}(a_m, m, s) = \frac{1}{\sqrt{\alpha_m}} \left(a_m - \frac{\beta_m}{\sqrt{1 - \bar{\alpha}_m}} \epsilon_{\theta_a}(a_m, m, s) \right), \quad (20)$$

where ϵ_{θ_a} is the deep model (*i.e.*, actor network) parameterized by θ_a generating denoising noises based on the observation s . Based on the mean and variance values and from [16], we can formulate the generative distribution $p_{\theta_a}(a_0)$ as

$$p_{\theta_a}(a_0) = p(a_M) \prod_{m=1}^M p_{\theta_a}(a_{m-1}|a_m), \quad (21)$$

where $p(a_M)$ is the Gaussian probability density function.

In the DM-DDPG algorithm, the reverse diffusion network is utilized to generate an output action a based on input state s instead of a traditional multilayer perceptron network. However, the significant challenge when integrating the DM is that the random variables can not be back-propagated in the

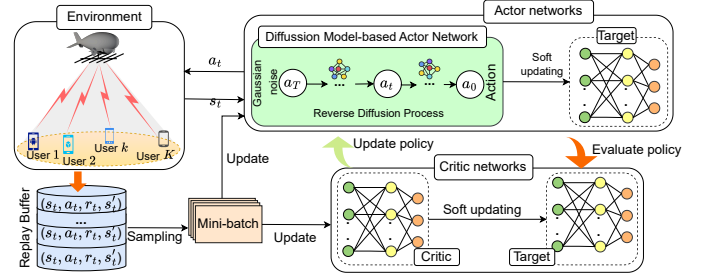


Fig. 2. Illustration of the proposed DDPG-DM algorithm.

operation of sampling to update network parameters during the training process. We employ parameterization to tackle this issue with the distribution a_{m-1} at the diffusion step $m - 1$ at the time slot t follows to

$$a_{m-1,t} = \mu_{\theta_a}(a_{m,t}, m, s_t) + \left(\tilde{\beta}_m / 2 \right)^2 \odot \epsilon, \quad (22)$$

where $\epsilon \sim \mathcal{N}(0, \mathbf{I})$. From the reverse process, we can obtain all $a_m, \forall m, 0 \leq m \leq M$, particularly, the sample original a_0 . As a result, we can sample action a_t according to the distribution a_0 .

Algorithm 1 The proposed DM-DDPG algorithm

- 1: Initialize θ_a , θ_c , and $\theta'_a \leftarrow \theta_a$ and $\theta'_c \leftarrow \theta_c$ for the main actor, critic, target actor and target critic, respectively.
 - 2: **for** each episode $e = 1, \dots, E$ **do**
 - 3: Reset the network environment.
 - 4: **for** each time slot $t = 1, \dots, T$ **do**
 - 5: Observe state s_t .
 - 6: Generate a random Gaussian noise $a_M \sim \mathcal{N}(0, \mathbf{I})$.
 - 7: **for** the denoising steps $m = M$ to 1 **do do**
 - 8: Utilize the actor network to predict the noise distribution in reverse diffusion process with mean μ_{θ_a} in (20) and variance $\tilde{\beta}_t$.
 - 9: Calculate the distribution a_{m-1} according to the distribution in (22).
 - 10: **end for**
 - 11: Execute a_t based on the distribution of a_0 , receive the reward r_t , and move to the next state s'_t .
 - 12: Store (s_t, a_t, r_t, s'_t) in the replay buffer.
 - 13: Sample a random mini-batch of B from R .
 - 14: Update the actor network θ_a using (14).
 - 15: Update the critic parameters θ_c using $L(\theta_c)$.
 - 16: Update target networks θ'_a and θ'_c using (15).
 - 17: **end for**
 - 18: **end for**
-

The detailed implementation of the proposed algorithm is described in Algorithm 1 and pictorially illustrated in Fig. 2.

D. Complexity Analysis

We use a fully-connected architecture for learning networks to achieve optimal values in the optimization problem (6). The computational complexity includes (i) the computational

TABLE I
SIMULATION PARAMETERS.

Parameter	Value
Maximum transmit power of the HAP $P_{\max}$	30 dBm
Antenna gains $G_{\text{tx}}, G_{\text{rx}}$	32, 25 dBi
Noise power spectrum density σ^2	-174 dBm/Hz
The operating frequency f	140 GHz
Molecular absorption coefficient ς_{abs}	0.5 dB/km
The minimum data rate $R_{\min}$	0.001 bps/Hz
Number of episodes T	200
Number of time steps of each episode E	200
Discount factor γ	0.99
Soft-updating coefficient τ	0.005
Replay buffer R	10000
Batch size B	64
Learning rate of the actor network	10^{-4}
Learning rate of the critic network	10^{-3}
Denoising steps for the DM M	5

complexity of the generation action-reverse DM is calculated as $\mathcal{O}(ETM|\theta_a|)$, where $|\theta_a|$ represents the number of parameters in the diffusion-enabled actor network; (ii) the second part involved in updating process of DNNs, calculated as $\mathcal{O}(EB(|\theta_a| + |\theta_c|)) + E(|\theta'_a| + |\theta'_c|)$. Because the parameters of two target networks are of the same size as original networks, the computational complexity of the second part can be rewritten as $\mathcal{O}(E(B+1)(|\theta_a| + |\theta_c|))$. Therefore, the computation complexity of DM-DDPG algorithm is $\mathcal{O}(ETM|\theta_a| + E(B+1)(|\theta_a| + |\theta_c|))$.

IV. SIMULATION RESULTS

We consider a network with one HAP flying at an altitude at $h = 17$ km with $N = 2$ antennas serving $K = 4$ single-antenna UEs. Assuming UEs are randomly distributed in the coverage area with a radius of 1700 m. According to [4], [5], the simulation parameters and values are listed in Table I.

To evaluate the effectiveness of the proposed DM-DDPG algorithm, we consider the following alternative schemes.

- **DDPG**: refers to a scheme using the traditional DDPG-based algorithm without embedding the DM to address the considerable optimization problem (6).
- **Greedy**: refers to a scheme using the DM-DDPG algorithm with the learning rate of neural networks as 0, i.e., one-shot optimization is considered for the problem (6).
- **Random**: refers to a scheme in which the actions are performed randomly without considering the dynamic environments.

In the first evaluation, we investigate the convergence of the proposed algorithm with varying learning rates of the actor network $\{10^{-6}, 10^{-5}, 5 \times 10^{-4}, 10^{-4}, 10^{-3}\}$. As shown in Fig. (3), the proposed algorithm achieves a higher reward and converges to its optimum with the learning rate of 10^{-4} . In case of a too small or too high learning rate (e.g., $10^{-6}, 10^{-3}$), the algorithm cannot reach the global optimum since it may be stuck at local optimal values or jump over the optimum. Hence, the learning rate of 10^{-4} is chosen as the learning rate of the actor network for the following evaluations.

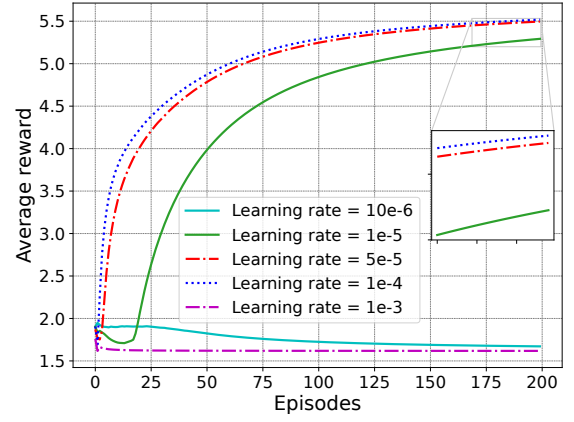


Fig. 3. Convergence reward versus the actor learning rate.

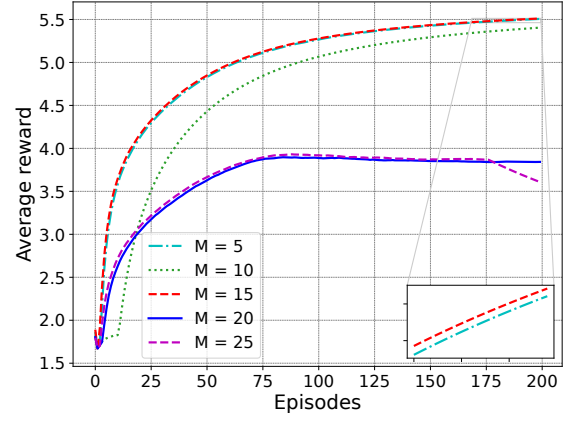


Fig. 4. Convergence reward versus the denoising step.

In Fig. 4, it depicts the convergence rate of the proposed DM-DDPG algorithm with varying denoising steps M in the range $[5, 10, 15, 20, 25]$. Obviously, the average reward increases when the denoising step $M = \{5, 10\}$ is small enough, but decreases significantly when the denoising step is sufficiently large $M = \{20, 25\}$. The main reason for this is that the higher denoising step will leave more noise behind in the sample and even decrease the DM's ability to locate the best action. In our simulation setting, the proposed algorithm achieves the best performance with $M = 15$.

Next, Fig. 5 illustrates the reward training process of DM-DDPG compared to the other three benchmarks, given the learning rate of 10^{-4} and the diffusion step of 15. Clearly, our proposed algorithm is superior to all of the baseline schemes (i.e., **Greedy** and **Random**). It also proves the ability of DM in enhancing the performance of DRL algorithms, thus showing the effectiveness of the proposed DM-DDPG algorithm in solving the long-term, non-convex optimization problem (6).

Finally, Fig. 6 depicts the reward value of all methods under different maximum powers at the HAP. It is observed that the total achievable rate of our proposed algorithm increases proportionally with the increase of $P_{\max}$ values. When the transmit power allocation for private streams increases, the interference from private message into common message also

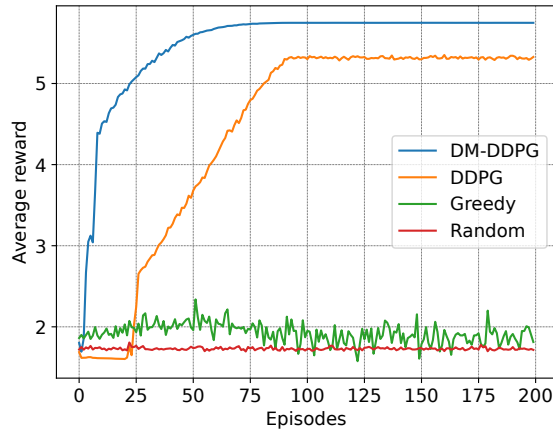


Fig. 5. Convergence reward of the proposed algorithm and benchmarks.

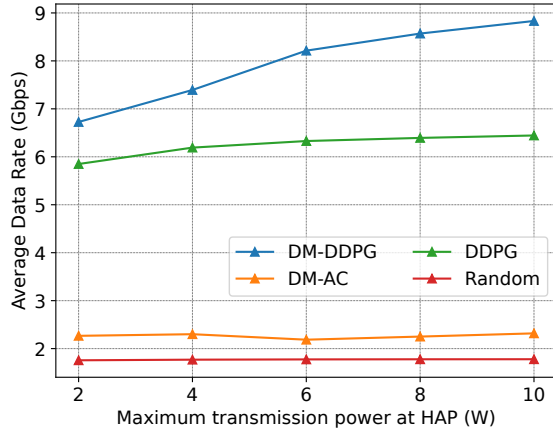


Fig. 6. Performance comparison under different transmit powers.

increases, it generally impacts to data rate of each user. However, the proposed algorithm with the ability to provide a more diverse set of actions can potentially allocate the optimal values to satisfy the power tolerance constraints and attain the highest performance among all schemes. The **Greedy** scheme generally outperforms the **Random** scheme by leveraging per-slot optimization, but is lower than both DDPG-based schemes across all variations in terms of the total transmit power.

V. CONCLUSION

This paper has investigated the resource allocation problem in HAP-RSMA-based THz communications downlink networks. We have introduced the denoising DM based actor network of the DDPG algorithm to address the non-convex nature of the problem under dynamic environments, power tolerance, and minimum rate requirement, aiming to maximize the total achievable rate of the system. We have demonstrated that our proposed algorithm outperforms other benchmark algorithms in terms of convergence and under varying parameter settings.

ACKNOWLEDGMENTS

This publication has emanated from research supported in part by a grant from Research Ireland under Grant number [12-RC-2289-P2] which is co-funded under the European Regional Development Fund. The work of Quoc-Viet Pham is supported

in part by Research Ireland under the European Innovation Council CHIST-ERA SHIELD project (Project No. 216449, Award No. 19226). For the purpose of Open Access, a CC BY licence applies to any Author Accepted Manuscript from this submission.

REFERENCES

- [1] B. Clerckx, Y. Mao, Z. Yang, M. Chen, A. Alkhateeb, L. Liu, M. Qiu, J. Yuan, V. W. S. Wong, and J. Montojo, "Multiple access techniques for intelligent and multifunctional 6G: Tutorial, survey, and outlook," *Proc. IEEE*, vol. 112, no. 7, pp. 832–879, 2024.
- [2] W. Jiang *et al.*, "Terahertz communications and sensing for 6G and beyond: A comprehensive review," *IEEE Commun. Surveys Tuts.*, vol. 26, no. 4, pp. 2326–2381, 2024.
- [3] Y. Xiao, Z. Ye, M. Wu, H. Li, M. Xiao, M.-S. Alouini, A. Al-Hourani, and S. Cioni, "Space-air-ground integrated wireless networks for 6G: Basics, key technologies, and future trends," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 12, pp. 3327–3354, 2024.
- [4] X. Wang, K. Wang, D. Zhang, J. Li, M. Zhou, and T. Hämäläinen, "Generative diffusion model-based deep reinforcement learning for uplink rate-splitting multiple access in LEO satellite networks," in *ISCC*, 2024, pp. 1–8.
- [5] M. Le, Q.-V. Pham, Q. Vinh Do, Z. Han, and W.-J. Hwang, "Resource allocation in THz-NOMA-enabled HAP systems: A deep reinforcement learning approach," *IEEE Trans. Consum. Electron.*, vol. 70, no. 4, pp. 6808–6816, 2024.
- [6] H. Du, R. Zhang, Y. Liu, J. Wang, Y. Lin, Z. Li, D. Niyato, J. Kang, Z. Xiong, S. Cui, B. Ai, H. Zhou, and D. I. Kim, "Enhancing deep reinforcement learning: A tutorial on generative diffusion models in network optimization," *IEEE Commun. Surveys Tuts.*, vol. 26, no. 4, pp. 2611–2646, 2024.
- [7] K.-T. Tran, D. Dao, M.-D. Nguyen, Q.-V. Pham, B. O'Sullivan, and H. D. Nguyen, "Multi-agent collaboration mechanisms: A survey of LLMs," 2025. [Online]. Available: <https://arxiv.org/abs/2501.06322>
- [8] P. Ning, H. Wang, T. Zhang, J. Zhang, H. Du, D. Niyato, and F. Richard Yu, "Diffusion-based deep reinforcement learning for resource management in connected construction equipment networks: A hierarchical framework," *IEEE Trans. Wireless Commun.*, vol. 24, no. 4, pp. 2847–2861, 2025.
- [9] S. S. Hassan, Y. M. Park, Y. K. Tun, W. Saad, Z. Han, and C. S. Hong, "SpaceRIS: LEO satellite coverage maximization in 6G sub-THz networks by MAPPO DRL and whale optimization," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 5, pp. 1262–1278, 2024.
- [10] Z. Hu, C. Han, W. Gerstacker, and I. F. Akyildiz, "Tera-SpaceCom: GNN-based deep reinforcement learning for joint resource allocation and task offloading in terahertz band space networks," *arXiv preprint arXiv:2409.07911*, 2024.
- [11] X. Wang, N. Deng, and H. Wei, "Coverage and rate analysis of LEO satellite-to-airplane communication networks in terahertz band," *IEEE Trans. Wireless Commun.*, vol. 22, no. 12, pp. 9076–9090, 2023.
- [12] Y. Gu, R. Wang, D. Wu, Y. Cui, P. He, and B. Yang, "Multi-dimensional modeling and connectivity analysis for THz space-air-ground integrated network," *IEEE Trans. Wireless Commun.*, 2025.
- [13] W. Gao, C. Han, Z. Chen, Y. Chen, Y. He, and W. Zhang, "Terahertz aerospace communications: Enabling technologies and future directions," *arXiv preprint arXiv:2502.17808*, 2025.
- [14] A. M. Benaya, M. I. AlHajri, M. H. Ismail, M. S. Hassan, and T. Landolsi, "Outage analysis of THz-enabled IRS-assisted multi-user space-air-ground integrated networks," *IEEE Open J. Commun. Soc.*, vol. 6, pp. 2599–2615, 2025.
- [15] L. Yang, Z. Huang, F. Lei, Y. Zhong, Y. Yang, C. Fang, S. Wen, B. Zhou, and Z. Lin, "Policy representation via diffusion probability model for reinforcement learning," 2023. [Online]. Available: <https://arxiv.org/abs/2305.13122>
- [16] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," *NeurIPS*, vol. 33, pp. 6840–6851, 2020.
- [17] M. S. Ali, E. Hossain, A. Al-Dweik, and D. I. Kim, "Downlink power allocation for CoMP-NOMA in multi-cell networks," *IEEE Trans. Commun.*, vol. 66, no. 9, pp. 3982–3998, 2018.
- [18] 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," in *ICLR*, San Juan, Puerto Rico, 2016.