ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY
Volume 28, Number 4, 2025, 327–340
doi: 10.59277/ROMJIST.2025.4.02

Deep Reinforcement Learning and Metaheuristic Approaches to Maximize Downlink Sum-Rate for Internet of Things Systems in Non-Orthogonal Multiple Access-based Space-Air-Ground Integrated Networks

Sothearath MENG[1], Kimchheang CHHEA[2], Sengly MUY[2], and Jung-Ryun LEE[1,3,*]

[1]Department of Intelligent Energy Industry Convergence, Chung-Ang University, South Korea
[2]Department of Intelligent Energy and Industry, Chung-Ang University, South Korea
[3]School of Electrical and Electronics Engineering, Chung-Ang University, South Korea
Email: sothearath@cau.ac.kr, chheangkim@cau.ac.kr, muysengly@cau.ac.kr, jrlee@cau.ac.kr*
* Corresponding author

Abstract. Internet of Things (IoT) device has significantly increased the need for reliable and efficient communication systems. Space–Air–Ground Integrated Network (SAGIN) addresses this need through its hierarchical structure, by integrating Low Earth Orbit (LEO) satellites, High Altitude Platforms (HAPs), and Unmanned Aerial Vehicles (UAVs). This paper focuses on maximizing the downlink sum-rate in a Non-Orthogonal Multiple Access (NOMA)-based SAGIN-IoT system by jointly optimizing the geographical locations and transmit powers of HAPs and UAVs, bandwidth allocation ratio, and link selection between a LEO satellite and IoT devices. The problem is formulated as a complex joint optimization task involving both discrete and continuous variables, reflecting the dynamic and large-scale nature of the SAGIN network. To solve this, two solution algorithms are employed: a deep reinforcement learning (DRL) algorithm and an Alternative Optimization (AO) algorithm. The proposed DRL framework leverages a deep Q-learning (DQL) architecture to efficiently navigate the high-dimensional and dynamic environments of SAGIN. The AO algorithm, on the other hand, decomposes the original optimization problem into two subproblems, iteratively solving them using Differential Annealing (DA). The performance of the proposed DQL and AO algorithms is compared with that of Gradient Search (GS). Simulation results demonstrate that DQL achieves superior performance in terms of overall sum-rate optimization with lower computational complexity. While the AO algorithm provides competitive results, it requires higher computational complexity than both DQL and GS.

Key-words: Alternative Optimization (AO), Deep Reinforcement Learning (DRL); High-Altitude Platform (HAP); Internet of Things (IoT); link selection; Unmanned Aerial Vehicle (UAV).

1. Introduction

The ongoing fast development in information and communication technology has established the Internet of Things (IoT) as a driver to support diverse intelligent applications ranging from virtual reality to smart agriculture and remote environmental monitoring. These types of IoT applications increasingly rely on high data rate communications to transmit continuous streams of sensitive data and support real-time decision-making. Consequently, ensuring high downlink data rates is essential to support the large-scale data requirements and responsiveness of modern IoT systems. Moreover, IoT applications are expanding into infrastructure-less environments, where connectivity is limited by geographical and economic restrictions [1]. As the number of IoT devices is expected to grow to approximately 30.9 billion by 2025 [2], traditional terrestrial IoT technologies may present challenges in keeping up with the required spectrum efficiency and quality of service (QoS) in future IoT networks [3]. To overcome these limitations, satellite technologies are rapidly developing [4], expanding network coverage [5] and connectivity beyond conventional terrestrial networks [6]. A strong example is the Starlink project, which demonstrates how global coverage and seamless connections can be achieved, even in remote regions [7]. The primary solution involves employing satellites as a backhaul for terrestrial networks [8]. However, direct satellite communication has inherent limitations in communication resources, which restricts its ability to meet high-data rate requirements for IoT applications. To overcome this problem, high-altitude platforms (HAPs) and Unmanned Aerial Vehicle (UAV) are introduced as potential solutions. HAPs operate in the stratosphere at altitudes of 17-22 kilometers, offering cost-effective, high-capacity communication services and enhancing terrestrial networks by providing over-the-horizon coverage. Meanwhile, UAVs have become a major focus in research and industry due to their adaptable positioning and direct line-of-sight communication with ground devices. Together, these aerial platforms enable more flexible, high-capacity communication links between satellites and ground IoT devices. In this context, a space-air-ground integrated network (SAGIN) has emerged as a promising solution for seamless communication coverage, integrating LEO satellites, HAPs, and multiple UAVs. This multi-layered architecture combines the global coverage of satellites with the flexibility and capacity of aerial platforms to deliver reliable, high-speed communication services.

In a SAGIN architecture, variations in path loss and transmission delay occur differently due to the different altitudes of LEO satellites, HAPs, and UAVs. Therefore, selecting the proper communication links between ground IoT devices and LEO satellite, whether directly or through HAPs and/or UAVs, is an important control parameter for enhancing the performance of the SAGIN system. Additionally, the deployment location of the HAP and UAVs is important in determining the likelihood of maintaining line-of-sight connections to ground IoT devices, which can further affect system performance. Although HAP, and especially UAVs are flexible in terms of deployment location and mobility, their operations are constrained by limited onboard energy. This restricts the maximum transmit power available for signal transmission, highlighting the need for efficient power management strategies [9]. In this context, Non-Orthogonal Multiple Access (NOMA) is an effective technology for improving spectrum efficiency and overall system capacity. By allowing multiple users to share the same frequency resources with different power levels, NOMA achieves better spectrum utilization compared to traditional orthogonal access

methods [10]. Integrating NOMA into SAGIN architectures has gained attention as a promising algorithm to further enhance system performance in next-generation IoT networks. Apart from energy constraints, wireless bandwidth is another critical constraint in SAGIN's multilayer platform architecture. Hence, optimizing the bandwidth allocation policy is required to efficiently distribute resources across different network layers. The details of the related work are provided in [33], and the references cited in [33] correspond to the list given in the References section of this paper.

The main contributions of this paper are summarized as follows:

1. System Design: A SAGIN communication system is designed, where an HAP and multiple UAVs are deployed to serve IoT devices randomly distributed at ground level. An optimization model is formulated for NOMA-based SAGIN-IoT systems to jointly optimize multiple control parameters, including the deployment locations of the HAP and UAVs, bandwidth allocation ratio, link selection, and transmit power. The objective is to maximize the downlink sum-rate of the IoT system while satisfying all system constraints.
2. Solution Methodology: Two practical solutions are presented: (1) a DQL framework that efficiently handles the dynamic and complex SAGIN-IoT environment, and (2) an AO algorithm that decomposes the original optimization model into subproblems, each solved using DA, a probabilistic global optimization technique suitable for non-convex problems.
3. Performance Evaluation: The simulation results of the proposed DQL and AO algorithms are compared with the Gradient Search (GS) algorithm. Results demonstrate that the primary proposed DQL algorithm outperforms both AO and GS in terms of convergence speed, scalability, and overall downlink sum-rate, while maintaining lower computational complexity.

The structure of the paper is as follows: Section 2 describes the proposed system model, which includes the network scenario and the communication channel model. Section 3 presents the optimization problem. Section 4 explains the DQL architecture and the training procedure. Section 5 introduces an alternative optimization algorithm. Section 6 discusses the simulation results, and Section 6 concludes the paper.

2. System Model

2.1. Network scenario

In this network scenario, a downlink SAGIN-IoT communication system is considered, where a LEO satellite, an HAP, and multiple UAVs are employed to support multiple IoT sensor devices (SD), as illustrated in Fig. 1 of the supplementary file provided in [33]. The set of SDs is denoted by $\mathcal{I} = \{1, ..., i, ..., I\}$ and the set of UAVs by $\mathcal{U} = \{1, .., k, .., K\}$. Considering 3-dimensional Cartesian coordinates, the location of the i-th SD, the k-th UAV, the HAP, and LEO satellite are denoted as $u_i = \{x_i, y_i, z_i\}, u_k = \{x_k, y_k, z_k\}, u_h = \{x_h, y_h, z_h\}$ and $u_0 = \{x_0, y_0, z_0\}$, respectively. The total duration of the operation is set to $\mathcal{T} = \{1, ..., t, ...T\}$. To maximize spectral efficiency while managing cross-user interference, non-orthogonal multiple access (NOMA) technology is employed, as it allows multiple users to simultaneously access common channel resources, including time slots, frequencies, and codes. It is noted that NOMA works by superimposing signals at different power levels and decoding them using successive interference cancellation (SIC) at the receiver.

2.2. Communication channel model

The communication channels between space-air platforms and SDs are influenced by both small-scale and large-scale fading. Small-scale fading occurs due to multi-path propagation, where the signal reaches the receiver through multiple paths, causing rapid fluctuations in signal strength over short distances. On the other hand, large-scale fading is modeled using free-space path loss [23], where the channel quality is affected by the long distance between transmitter j and SD i. It can be formulated as:

$$PL_{j,i}[t] = \left(\frac{c}{4\pi f d_{j,i}^{\beta}} \right)^2, \ j \in \{k, h, 0\}, \tag{1}$$

where h, k, and 0 correspond to the k-th UAV, HAP and LEO satellite, respectively. f, c, $d_{j,i}$, and β denote the carrier frequency in hertz (Hz), the speed of light (approximately $\approx 3 \times 10^8 m/s$), the distance between the transmitter j and SD i, and the path loss coefficient, respectively. To accurately represent the Line-of-Sight (LoS) propagation between transmitter j and SD i, small-scale fading is characterized using Rician fading, given by [24]:

$$g_{j,i}[t] = \sqrt{\frac{\zeta}{\zeta+1}} g_{j,i}^{LOS}[t] + \sqrt{\frac{1}{\zeta+1}} g_{j,i}^{NLOS}[t], j \in \{k, h, 0\}, \tag{2}$$

where $g_{j,i}^{LOS}$ and $g_{j,i}^{NLOS}$ represent the LoS component and the NLoS component, respectively. The LoS component follows a complex normal distribution $g_{j,i}^{NLOS} \sim \mathcal{CN}(0, 1)$, while the NLoS component is modeled with the Rician fading parameter ζ.

In the NOMA downlink, the transmitter simultaneously serves multiple SDs. Each SD employs the SIC technique to decode its signal while managing interference from other signals. The order of SIC decoding follows the order of decreasing channel gain relative to noise and inter-cell interference, expressed as $|H_{j,1}[t]|^2 \geq |H_{j,2}[t]|^2 \geq ... \geq |H_{j,i}[t]|^2$, where the channel gain is defined as $H_{j,i}[t] = g_{j,i}[t]\sqrt{PL_{j,i}[t]}$ [25]. In this sequence, each SD i successfully decodes the signals of all SDs with higher decoding priority. For example, if the channel gain of SD i is lower than SD i', the SIC process for SD i' begins by decoding the signal from SD i. Then SD i' subtracts the SD i's signal component from its received signal, enabling it to decode its own signal without interference from i. Therefore, assuming that all SDs $i < i'$, the signal i decodes its own signal while treating the stronger signal from i' as interference. Consequently, the signal-to-interference-plus-noise ratio (SINR) between transmitter j and SD i can be formulated as:

$$\Gamma_{j,i}[t] = \frac{p_{j,i}[t] \left| H_{j,i}[t] \right|^2}{\sigma^2 + \sum_{i' \neq i} p_{j,i'}[t] \left| H_{j,i}[t] \right|^2}, j \in \{k, h, 0\}, \tag{3}$$

where $p_{j,i}$ represents the transmit power allocated to SD i, $H_{j,i}$ is the channel gain for SD i, and σ^2 is the Additive White Gaussian Noise. Using the Shannon-Hartley theorem, the data rate for a single link between transmitter j to SD i is calculated as:

$$\Upsilon_{j,i}[t] = b_j[t] B \log_2 \left(1 + \Gamma_{j,i}[t]\right), j \in \{k, h, 0\}, \tag{4}$$

where B represents the total bandwidth, and b_j is the bandwidth allocation ratio for all SDs connected to transmitter j. Because of the limited bandwidth resource, the UAVs, HAP and LEO

satellite need to share the resource. Thus, the sum bandwidth allocation ratio of the UAVs, HAP and LEO satellite is given as:

$$b_0[t] + b_h[t] + \sum_{k \in \mathcal{K}} b_k[t] = 1, \tag{5}$$

where b_0, b_h, and b_k represent the bandwidth ratio of the LEO Satellite, HAP, and k-th UAV, respectively. From (4), the total achievable data rate can be determined as:

$$\Upsilon_{total}[t] = \sum_{i \in \mathcal{I}} \left(\sum_{k \in \mathcal{K}} \Upsilon_{k,i}[t] + \Upsilon_{h,i}[t] + \Upsilon_{0,i}[t] \right). \tag{6}$$

3. Problem Formulation

This paper aims to maximize the total achievable data rate for the NOMA-based SAGIN-IoT system by controlling location of HAP, location of UAV $u_j, \forall j \in \{k, h\}$, bandwidth allocation ratio $b_j, \forall j \in \{k, h, 0\}$, transmit power $p_{j,i}, \forall j \in \{k, h, 0\}$, and link selection $l_{j,i}$. Here, the link selection of transmitter j and SD i is defined as a binary variable, where $l_{j,i} = 1$ when transmitter j is chosen for SD i, otherwise $l_{j,i} = 0$. Hence, the optimization problem can be expressed as:

$$\max_{u_j, b_j, p_{j,i}, l_{j,i}} \sum_{t \in \mathcal{T}} \Upsilon_{total}[t], \tag{7}$$

$$\begin{aligned}
\text{s.t.} \quad & \text{C1} : P_{\min}^{UAV} \leq p_k[t] \leq P_{\max}^{UAV}, && \text{(7a)} \\
& \text{C2} : P_{\min}^{HAP} \leq p_h[t] \leq P_{\max}^{HAP}, && \text{(7b)} \\
& \text{C3} : P_{\min}^{LEO} \leq p_0[t] \leq P_{\max}^{LEO}, && \text{(7c)} \\
& \text{C4} : x_{\min}^{UAV} \leq x_k \leq x_{\max}^{UAV}, && \text{(7d)} \\
& \text{C5} : y_{\min}^{UAV} \leq y_k \leq y_{\max}^{UAV}, && \text{(7e)} \\
& \text{C6} : x_{\min}^{HAP} \leq x_h \leq x_{\max}^{HAP}, && \text{(7f)} \\
& \text{C7} : y_{\min}^{HAP} \leq y_h \leq y_{\max}^{HAP}, && \text{(7g)} \\
& \text{C8} : d_{k,k\prime}[t] \geq d_{\min}, \quad \forall k, k\prime \in \mathcal{U}, k \neq k\prime, && \text{(7h)} \\
& \text{C9} : \Upsilon_{j,i}[t] \geq \Upsilon_{j,\min}, \quad \forall j \in \{k, h, 0\}, && \text{(7i)} \\
& \text{C10} : b_j \in [0, 1], \quad \forall j \in \{k, h, 0\}, && \text{(7j)} \\
& \text{C11} : b_0[t] + b_h[t] + \sum_{k \in \mathcal{K}} b_k[t] = 1, && \text{(7k)} \\
& \text{C12} : l_{j,i} \in \{0, 1\}, \quad \forall j \in \{k, h, 0\}, && \text{(7l)} \\
& \text{C13} : \sum_{j} l_{j,i}[t] = 1, \quad \forall j \in \{k, h, 0\}, && \text{(7m)}
\end{aligned}$$

where C1, C2, and C3 are the transmit power allocation constraints for the UAVs, HAP, and LEO satellite, respectively. These constraints ensure that the power levels stay within the predefined minimum and maximum values. Constraints C4-C7 define the deployment boundaries for UAVs and the HAP, and ensure that their positions remain within the specified spatial limits. Constraint

C9 ensures the minimum data rate requirement for the UAVs, HAP, and LEO to guarantee service quality, where $\Upsilon_{j,\min}$ is the minimum required rate predefined for the UAVs, HAP, and the LEO satellite. Constraint C8 ensures the minimum distance between UAV k and its neighbor $k\prime$ to prevent collisions. Constraints C10 and C11 are for the bandwidth allocation ratio, which ensures that the total allocation ratio does not exceed the available bandwidth. Constraint C12 serves as the link selection indicator, while constraint C13 ensures that each SD i can select only one link in any given time slot. Given the non-convexity of the formulated mixed-integer programming (MIP) problem, achieving an optimal solution via analytical approaches is considerably challenging. To address this, two solution algorithms are proposed: a Deep Q-Learning (DQL) algorithm under a Deep Reinforcement Learning (DRL) framework, and an Alternative Optimization (AO) algorithm using Differential Annealing (DA). The DQL algorithm approximates the optimal resource deployment by considering the locations of HAP and UAVs, the bandwidth allocation ratio, link selection, and transmit power. It operates by interacting with the environment, using the downlink sum-rate as the reward to guide learning. On the other hand, the AO algorithm decomposes the original problem into subproblems and solves them iteratively using DA. This algorithm deterministically updates parameters at each step, aiming to maximize the sum-rate. Both algorithms aim to achieve this objective by iteratively updating system parameters to converge toward the maximum sum-rate for the downlink SAGIN-IoT network.

4. Proposed Deep Reinforcement Learning Algorithm

This section begins with a brief description of Q-Learning and DRL. The proposed algorithm is next presented, which uses DQL to allow transmitter j to learn from the network environment and adjust its control parameters to obtain the maximum sum-rate for the network. Finally, the state space, action space, and reward function are defined for the proposed algorithm.

4.1. Overview of Q-learning

Reinforcement Learning (RL) has gained increasing attention in wireless communication due to its ability to solve complex problems. Q-Learning is a powerful technique for finding the optimal policy π, which defines a set of strategic guidelines that allow an agent to maximize cumulative rewards over time. In this algorithm, agents interact with their environment to achieve their specific objectives and effectively determine the best possible actions. The Q-value for the corresponding state-action pair is updated in a Q-table. The Q-Learning algorithm refines these Q-values through repeated interactions, based on the Bellman equation given as [26]:

$$\mathbb{Q}\left(s\left[t\right],a\left[t\right]\right)=\mathbb{Q}\left(s\left[t\right],a\left[t\right]\right)+\alpha\left[r\left[t+1\right]+\gamma\underset{a}{\max}\mathbb{Q}\left(s\left[t+1\right],a\right)-\mathbb{Q}\left(s\left[t\right],a\left[t\right]\right)\right], \quad (8)$$

where $r\left[t+1\right]$ is the future feedback reward, α denotes the learning rate with $(0<\alpha\leq 1)$, and γ is the learning discount factor $(0<\gamma\leq 1)$.

Q-learning performs well when the state-action space is small enough to be represented fully in a Q-table. However, in this network scenario, where data is transmitted through dynamic space-air links to SDs, the system must simultaneously control the deployment locations of both the HAP and multiple UAVs, link selection, transmit power allocation, and bandwidth allocation ratio. As the number of UAVs and SDs increases, the state and action spaces grow exponentially, making the problem even more challenging. Consequently, traditional Q-learning is not suitable

for such scenarios. To overcome this limitation, the DQL algorithm enhances Q-learning by employing a deep neural network to approximate the Q-values, replacing the traditional Q-value table. This extension allows the algorithm to manage environments with large state-action spaces and handle high-dimensional input, which is well-suited to the case study treated here.

In the environment setup, the agent determines the best action by choosing either a random action or the action from the action space A with the highest Q-value based on current observations. This decision-making process uses the Epsilon ϵ-greedy model [26]. After performing the selected action, the agent receives a reward $r[t]$, and moves to the next state $s\left[t+1\right]$. To ensure effective and stable training, the DQL incorporates replay memory, which allows the agent to store past experiences $e\left[t\right] = \left(s\left[t\right], a\left[t\right], r\left[t\right], s\left[t+1\right]\right)$ in a memory buffer. Replay memory improves the performance of the DQL algorithm by improving its ability to approximate the Q value $\mathbb{Q}\left(s\left[t\right], a\left[t\right], \omega\right)$ with the weight of the network ω. Once the replay memory reaches its capacity, the agent randomly selects a batch of experiences for training. During training, the agent relies on two networks: the policy network and the target network. The target network is a duplicate of the policy network, sharing the same weights and biases. The target network, defined as: $\mathbb{Q}\prime\left(s\prime\left[t\right], a\prime\left[t\right], \omega\prime\right)$, is used to compute the target Q-value, which is given as:

$$y\left[t\right] = r\left[t\right] + \gamma \max_{a\prime} \mathbb{Q}\prime\left(s\prime, a\prime, \omega\prime\right). \tag{9}$$

The DQN architecture, proposed in [33] and illustrated in Fig. 2, is designed as a bias-free fully connected architecture consisting of three layers: an input layer, L hidden layers, and an output layer. In each time step t, the agent input observes the state $s[t]$ of the SAGIN-IoT system and uses this information to compute the Q-values for possible actions. The Q-network is trained in a way to minimize the loss function. The loss function is typically defined using the Mean Squared Error (MSE), which is given as:

$$\mathcal{L}\left(\omega\right) = \frac{1}{2}\left(y\left[t\right] - \mathbb{Q}\left(s, a, \omega\right)\right)^2. \tag{10}$$

The weight ω of the policy network is updated iteratively using Stochastic Gradient Descent (SGD) optimizer. During this process, transmitter devices send data over the wireless channel through space to the ground. The action is selected according to $arg\max_a \mathbb{Q}\left(s\left[t\right], a\left[t\right], \omega\right)$, and the system receives feedback in the form of reward and next-state information, which is used to guide agent training in the SAGIN-IoT environment.

4.2. Proposed DQL algorithm

In the proposed algorithm, each transmitter device j is treated as an agent that interacts with the environment in discrete time steps. At each time step t, each agent j observes the current states $s_j[t]$, performs an action $a_j\left[t\right]$, receives a reward $r_j\left[t\right]$, and moves to the next state $s_j\left[t+1\right]$. Hence, the state $s_j\left[t\right]$, action $a_j\left[t\right]$, reward $r_j\left[t\right]$ for each agent $j \in \{k, h, 0\}$ at time t are defined as follows. The state space of agent j at time t is defined as:

$$s_j\left[t\right] = \left\{\hat{u}_j\left[t\right], \hat{b}_j\left[t\right], \hat{p}_{j,i}\left[t\right], \hat{H}\left[t\right], \hat{l}_{j,i}\left[t\right]\right\}. \tag{11}$$

Here, to ensure stability and efficient learning, all variables in the state $s_j\left[t\right]$ are normalized to the range [0,1], which makes them suitable for input into the Q-Network. The coordinates of the location of the agent are normalized as $\hat{u}_j\left[t\right] = \left[\frac{x_j}{x_{max}-x_{min}}, \frac{y_j}{y_{max}-y_{min}}\right]$, where $x_{max}, x_{min}, y_{max},$

and y_{min} represent the values of the maximum and minimum position range of the agent j [27]. The set $H\left[t\right]=[H_{j,1},H_{j,2},...,H_{j,i}]$ represents the channel gains of the set between agent j and SD i at various links. Hence, the channel gains are normalized as $\hat{H}\left[t\right]=\left[\frac{H_{j,1}[t]-H_{min}^{j,1}[t]}{H_{max}^{j,1}[t]-H_{min}^{j,1}[t]},...,\right.$ $\left.\frac{H_{j,i}[t]-H_{min}^{j,i}[t]}{H_{max}^{j,i}[t]-H_{min}^{j,i}[t]}\right]$. $\hat{b}_j\left[t\right]$ is the current bandwidth allocation ratio for agent j. Link selection is transformed into a unique vector using one-hot encoding, where each variable is mapped to a vector with a single '1' and the rest '0's. For example, if agent$j$ selects the second link for i, it is represented as $l_{j,2}=[0,0,1,0,0,0,0,0,0,0]$.

The agent j in this network scenario makes decisions for moving to its optimal location, adjusting the bandwidth allocation ratio, setting the transmit power for each SD, and selecting the most efficient connection link for each SD, based on its current state $s_j\left[t\right]$. The action space is expressed as:

$$a_j\left[t\right]=\left\{u_j\left[t\right],b_j\left[t\right],p_{j,i}\left[t\right],l_{j,i}\left[t\right]\right\}, \tag{12}$$

where the set of movement direction $u_j\left[t\right]\in\{(x_j,y_j+L),(x_j,y_j-L),(x_j-L,y_j),$ $(x_j+L_j,y),(x_j,y_j)\}$ represents upward, downward, leftward, rightward, and stationary movements, respectively. Here, $L=\frac{1}{\mathbb{A}-1}2R$, where $\mathbb{A}$ is the quantization level with the cell radius R. The bandwidth allocation ratio for agent j, $b_j\left[t\right]\in\left\{0,\frac{1}{M},\frac{2}{M},...,1\right\}$ corresponds to the quantization levels $(M+1)$. The transmit power $p_{j,i}\left[t\right]\in\left\{0,\frac{P_{\max}^{j}}{N},\frac{2P_{\max}^{j}}{N},...,P_{\max}^{j}\right\}$ is quantized in levels $(N+1)$. $l_{j,i}\left[t\right]\in\{l_{1,1},l_{1,2},...,l_{j,i}\}$, where each element represents a specific link selection choice.

The reward of the model is defined to maximize sum-rate for SD system given as:

$$r_j\left[t\right]=\Upsilon_{total}[t], \tag{13}$$

when constraints C8, and C9 are satisfied; otherwise $r_j\left[t\right]=0$.

Algorithm 1 outlines the DQL process with experience replay for training the agent, where the complete pseudocode process is provided in the supplementary file in [33].

5. Proposed AO Algorithm Using Differential Annealing (DA)

The optimization problem in the case study involves controlling multiple parameters. However, solving this problem directly as an optimization model formulation is computationally infeasible due to the high dimensional search space and strong interdependencies among variables. To address this challenge, an AO algorithm is adopted, which decomposes the optimization process into two subproblems, each focusing on a specific set of control parameters. This decomposition simplifies the complex optimization model while maintaining the relationships between resource allocation and network topology. The optimization variables are categorized into two groups. The first subproblem $(P1)$ involves optimizing the bandwidth allocation ratio and transmit power $(b_j,p_{j,i})$, which is defined as:

$$(P1)\quad \max_{p_{j,i},b_j}\ \Upsilon_{\text{total}}[t] \tag{14}$$

subject to the constraints $C1$, $C2$, $C3$, $C10$, and $C11$. This subproblem focuses on optimizing continuous variables that directly impact spectral efficiency while keeping UAV/HAP locations

and link selection fixed. Since optimal transmit power allocation depends on available bandwidth, their joint optimization is essential for maximizing spectral efficiency. Both parameters determine the achievable sum-rate based on the Shannon capacity formula, as presented in equations (3) and (4). The second subproblem $(P2)$ involves optimizing the location of the transmitter devices $u_j, \forall j \in \{h, k\}$ and link selection $l_{j,i}$ while keeping other variables fixed, which is given as:

$$(P2) \quad \max_{u_j, l_{j,i}} \ \Upsilon_{\text{total}}[t] \tag{15}$$

subject to constraints $C4$–$C8$ and $C12$–$C13$. These variables, which depend on the network topology, shape network connectivity and should be jointly optimized under fixed transmit power allocation and bandwidth allocation ratio. UAV/HAP positioning directly affects channel quality, while binary link selection decisions determine which transmitter serves each SD device. Their joint optimization ensures efficient traffic distribution, preventing congestion at specific network nodes. By solving these highly coupled subproblems in an alternating manner, the AO framework effectively explores the solution space. The AO-based algorithm is detailed in Algorithm 2, where the pseudocode is provided in [33].

6. Performance Evaluation

6.1. Simulation results

In this simulation, a single LEO satellite is deployed within a cell radius of $2000 \times 2000\,\mathrm{m}^2$, positioned at an altitude of 340 km. The feasible positions for the HAP and UAVs are considered at altitudes of $25\ \mathrm{km}$ [23] and $300\ \mathrm{m}$ [28], respectively. To ensure the safety and efficiency of aerial entities, the closest allowed proximity between UAVs $d_{k,k\prime}$ is set to $20\ \mathrm{m}$ [29]. Each SD i is required to maintain a minimum data rate $r_{j,min} = 1\,\mathrm{Mbps}$ [30]. For communication purposes, Gaussian noise with a power spectral density of $\sigma^2 = -174\,\mathrm{dBm}$ is assumed. Additional system parameters are summarized in Table 1 provided in [33].

As shown in Fig. 2, which is given in [33], the proposed DQL framework consists of three components: input, hidden, and output layers. It employs two Deep Q-Networks: one for the policy network and one for the target network, each comprising four fully connected layers with 256 ReLU activation units per layer. The model is trained using the SGD optimizer with a learning rate of $\alpha = 0.001$, a discount factor of $\gamma = 0.99$, along with an ϵ-greedy exploration strategy where $\epsilon = 0.1$. The replay memory size is set to $\mathcal{D} = 5000$, and a mini-batch size of 128 is used for each update [27]. The convergence behavior of the proposed DQL algorithm is evaluated based on the system sum-rate (in Mbps), which serves as the reward function.

As shown in Fig. 3 of [33], the training is performed over $T_{ep} = 1000$ episodes, each consisting of up to 200 steps. Due to the exploration strategy and the stochastic nature of DQL, the reward initially fluctuates. However, convergence is achieved after approximately 500 episodes (50% of the training), with the reward stabilizing around 220 Mbps. These early fluctuations are expected in DQL and reflect the trade-off between exploration and exploitation. The reward function, defined in Eq. (13), is designed to guide the agent toward maximizing the overall downlink sum-rate. For performance evaluation, the average downlink sum-rate is measured over 1000 independent test episodes after the training phase to ensure the reliability and stability of the learned policy. Fig. 4 in [33] illustrates the convergence behavior of the AO algorithm over 1000 iterations, demonstrating its balance between exploration and exploitation. In the

early iterations, variability in the sum-rate is observed due to the stochastic nature of the DA algorithm, which combines global exploration with local refinement. This randomness enables the algorithm to escape local optima by accepting worse solutions, a behavior controlled by a temperature-dependent acceptance probability. Between iterations 200 and 600, the sum-rate increases from approximately 50 to 150 Mbps. During this stage, frequent fluctuations suggest that the algorithm is still actively exploring the solution space. This phase marks a transition period in which the algorithm gradually shifts from broad exploration to more focused exploitation. As the temperature decreases, the algorithm becomes more selective, reducing the acceptance of suboptimal solutions while still allowing stochastic jumps to avoid fast convergence. From iterations 600 to 1000, the curve stabilizes within the range of 160–175 Mbps, indicating a phase of refined local search and convergence. This stage reflects the algorithm's ability to fine-tune solutions within a favorable region, handling the complexities of the optimization landscape effectively. Overall, the convergence pattern showcases the DA algorithm's strengths, its capacity for global exploration and local refinement, consistent convergence with significant improvement over the initial values, and a well-balanced exploration–exploitation trade-off. The final sum-rate of approximately 175 Mbps confirms the algorithm's effectiveness for SAGIN-IoT optimization problems.

To evaluate the effectiveness of the proposed DQL and AO algorithms, their performance is compared with the benchmark algorithm, GS. Although DQL is learning-based while AO and GS are optimization-based, all algorithms are evaluated under the same system model, input data, performance metrics, and network conditions. The hyperparameters for training the DQL agent are chosen based on experimental validation and fine-tuned to ensure stable convergence. Meanwhile, the AO algorithm iteratively updates solutions using key parameters, including the acceptance parameter and temperature schedule, which are set through practical tuning and literature guidance to balance convergence speed and the ability to maximize the system sum-rate. The GS algorithm employs a tolerance threshold, ϵ^{-2}, as a convergence criterion to efficiently terminate the exhaustive search while maintaining solution accuracy. This parameter tuning across all algorithms ensures a balanced and fair performance comparison. The average performance of all algorithms is measured over 1000 independent test episodes to ensure reliability. Fig. 4, given in [33], presents the impact of increasing the maximum UAV transmit power $P_{\max}^{UAV}$ on the sum-rate performance of the three algorithms: the proposed DQL and AO, and the benchmark GS. As $P_{\max}^{UAV}$ increases, all algorithms achieve higher sum-rates due to improved signal strength at the receivers. However, the proposed DQL consistently achieves the highest performance, demonstrating its superior learning capability and adaptability compared to the other algorithms. AO, also proposed in this work, performs better than the benchmark GS but remains below DQL, demonstrating a moderate improvement through optimization technique.

To further assess scalability, Fig. 5 in [33] compares the sum-rate performance as the number of SDs increases. The proposed DQL algorithm starts at approximately 70 Mbps for 5 SDs and scales up to around 1200 Mbps for 45 SDs, significantly outperforming AO and GS. This trend highlights DQL's strong scalability and efficiency in managing increasing network load. Similarly, AO consistently outperforms GS across all network sizes but does not achieve the performance level of DQL. In contrast, GS shows relatively slower growth, indicating limited scalability. Overall, the results confirm that while both proposed algorithms improve network performance compared to GS, DQL delivers the highest overall performance and adapts most effectively to growing network sizes, making it the most robust and scalable algorithm among those evaluated.

Fig. 6 (a) in [33] illustrates the impact of a low minimum rate constraint on network performance. When the minimum rate is set to $\Upsilon_{j,\min} = 1$ Mbps, most SDs tend to connect directly to the LEO satellite despite its high path loss. This is because the low threshold allows SDs to easily meet the rate requirement. However, this leads to congestion at the LEO satellite's transmitter, ultimately reducing the achievable data rates. In contrast, Fig. 6 (b) in [33] shows the average data rates for SD groups when the minimum rate is increased to $\Upsilon_{j,\min} = 10$ Mbps. Under this stricter constraint, SDs are more evenly distributed across the LEO satellite, HAP, UAV1, and UAV2 to meet the requirement. The results demonstrate that all SDs satisfy the constraint, regardless of which node they connect to. Among the algorithms, the proposed DQL algorithm achieves the highest average SD data rate of approximately 15 Mbps, followed by AO with about 13 Mbps, and GS with around 10 Mbps. These findings further highlight the superior adaptability and performance of the proposed DQL algorithm under varying network constraints.

Fig. 7 (a) in [33] illustrates the initial stage of the dynamic link selection process, where SDs connect randomly to available nodes. During training, the proposed DQL agent explores various link options and evaluates their performance based on the received signal strength from nearby transmitters j. This exploration enables SDs achieve the highest possible data rate. The training continues until the DQL agent converges to a stable policy that maximizes the expected sum-rate of the network. As shown in Fig. 7 (b) in [33], the learned policy demonstrates intelligent link selection across the network. SDs located in edge areas primarily connect to LEO satellites, benefiting from their wide coverage despite higher path loss. Meanwhile, UAVs are deployed in densely populated area, where favorable line-of-sight conditions and proximity to SDs enable stronger signal quality. Due to their limited coverage and capacity, HAPs complement UAVs by serving area beyond UAV reach, thereby enhancing overall network coverage and capacity. These results clearly demonstrate the DQL algorithm's ability to dynamically optimize link selection, leading to a more efficient and high-performing SD network.

To validate the efficiency and accuracy of the proposed algorithms, they are evaluated using simulation-based performance metrics, including downlink sum-rate, convergence behavior, and scalability. A comparative analysis is conducted with a benchmark algorithm, GS, under the same conditions to ensure fairness. The evaluation includes the convergence curve shown in Fig. 3 of [33], sum-rate compared to transmit power in Fig. 4 of [33], scalability with an increasing number of SDs in Fig. 5 of [33], and the computational complexity and execution time summarized in Table 2, which is provided in [33]. Together, these results validate the accuracy, learning effectiveness, and scalability of the proposed algorithm in an SAGIN-IoT environment.

6.2. Computational analysis

In this section, the computational efficiency of three algorithms, AO, GS, and DQL, is analyzed. The detailed complexity analysis is provided in the Supplementary Material [33].

7. Conclusions

This paper addressed the problem of maximizing the sum-rate in the NOMA-based SAGIN-IoT system. To achieve this, an optimization problem is formulated by jointly controlling the deployment locations of HAP and UAV, bandwidth allocation ratio, link selection, and transmit power. To solve this problem, the DQL and DA with AO framework are proposed. The DQL

algorithm enables efficient decision-making in a highly dynamic and high-dimensional environment, while the AO algorithm divides the original problem into two subproblems and solves these subproblems using DA. The computational complexity of three algorithms is analyzed and compare the simulation results of the proposed DQL and AO algorithms with GS algorithm. Results show that the proposed DQL algorithm achieves superior performance both AO and GS in terms of overall sum-rate optimization with lower computational complexity. While the AO algorithm provides competitive performance but with higher computational complexity compared to DQL and GS.

Future research will be focused on further improving energy efficiency, in order to achieve superior network performance while conserving power consumption. This strategic direction reflects authors' commitment to advancing IoT technology, ensuring sustainable and high-performance solutions for future applications.

Acknowledgments. This work was supported in part by the Ministry of Science and ICT (MSIT), Korea through Information Technology Research Center (ITRC) Support Program under Grant IITP-RS-2024-00436248 Supervised by the Institute for Information & Communications Technology Planning & Evaluation(IITP), in part by the Human Resources Development under Grant 20214000000280 of the Korea Institute of Energy Technology Evaluation and Planning (KETEP) Grant Funded by the Korea Government Ministry of Trade, Industry and Energy, and in part by Chung-Ang University Young Scientist Scholarship in 2025 (2023).

References

[1] Z. SHEN, J. JIN, C. TAN, A. TAGAMI, S. WANG, Q. LI, Q. ZHENG and J. YUAN, *A survey of next-generation computing technologies in space-air-ground integrated networks*, ACM Computing Surveys **56**(1), 2023, pp. 1–40.

[2] H. ABOU-ZEID, F. PERVEZ, A. ADINOYI, M. ALJLAYL and H. YANIKOMEROGLU, *Cellular V2X transmission for connected and autonomous vehicles standardization, applications, and enabling technologies*, IEEE Consumer Electronics Magazine **8**(6), 2019, pp. 91–98.

[3] K. SHAFIQUE, B. A. KHAWAJA, F. SABIR, S. QAZI and M. MUSTAQIM, *Internet of Things (IoT) for next-generation smart systems: A review of current challenges, future trends and prospects for emerging 5G-IoT scenarios*, IEEE Access **8**, 2020, pp. 23022–23040.

[4] A. DOHR, R. MODRE-OPSRIAN, M. DROBICS, D. HAYN and G. SCHREIER, *The Internet of Things for ambient assisted living*, Proceedings of Seventh International Conference on Information Technology: New Generations, Las Vegas, NV, USA, 2010, pp. 804–809.

[5] D. LE-PHUOC, A. POLLERES, M. HAUSWIRTH, G. TUMMARELLO and C. MORBIDONI, *Rapid prototyping of semantic mash-ups through semantic web pipes*, Proceedings of 18th International Conference on World Wide Web, Madrid, Spain, 2009, pp. 581–590.

[6] Y. LI, M. HOU, H. LIU and Y. LIU, *Towards a theoretical framework of strategic decision, supporting capability and information sharing under the context of Internet of Things*, Information Technology and Management **13**, 2012, pp. 205–216.

[7] Z. NIU, X. S. SHEN, Q. ZHANG and Y. TANG, *Space-air-ground integrated vehicular network for connected and automated vehicles: Challenges and solutions*, Intelligent and Converged Networks **1**(2), 2020, pp. 142–12169.

[8] P. QIN, H. ZHAO, Y. FU, S. GENG, Z. CHEN, H. ZHOU and X. ZHAO, *Energy-efficient resource allocation for space–air–ground integrated industrial power Internet of Things network*, IEEE Transactions on Industrial Informatics **20**(4), 2023, pp. 5247-5284.

[9] D. HU, Q. ZHANG, Q. LI and J. QIN, *Joint position, decoding order, and power allocation optimization in UAV-based NOMA downlink communications*, IEEE Systems Journal **14**(2), 2019, pp. 2949–2960.

[10] C. WANG, M. PANG, T. WU, F. GAO, L. ZHAO, J. CHEN, W. WANG, D. WANG, Z. ZHANG and P. ZHANG, *Resilient massive access for SAGIN: A deep reinforcement learning approach*, IEEE Journal on Selected Areas in Communications **43**(1), 2024, pp. 297–313.

[11] S. LIU, H. DAHROUJ and M.-S. ALOUINI, *Joint user association and beamforming in integrated satellite-HAPS-ground networks*, IEEE Transactions on Vehicular Technology **73**(4), 2023, pp. 5162–5178.

[12] H. JIA, Y. WANG and W. WU, *Dynamic resource allocation for remote IoT data collection in SAGIN*, IEEE Internet of Things Journal, **11**(11), 2024, pp. 20575–20589.

[13] E. M. MOHAMED, M. A. ALNAKHLI and M. M. FOUDA, *Joint UAV trajectory planning and LEO-sat selection in SAGIN*, IEEE Open Journal of the Communications Society **5**, 2024, pp. 1624–1638.

[14] Q. GAO, M. JIA, Q. GUO, X. GU and L. HANZO, *Jointly optimized beamforming and power allocation for full-duplex cell-free NOMA in space-ground integrated networks*, IEEE Transactions on Communications **71**(5), 2023, pp. 2816–2830.

[15] Z. LIN, M. LIN, J.-B. WANG, T. DE COLA and J. WANG, *Joint beamforming and power allocation for satellite-terrestrial integrated networks with non-orthogonal multiple access*, IEEE Journal of Selected Topics in Signal Processing **13**(3), 2019, pp. 657–670.

[16] X. WANG, H. CHEN and F. TAN, *Joint terminal-AP association and power allocation for NOMA-enabled space–air–ground integrated networks*, Physical Communication **58**, 2023, paper 102020.

[17] L. YANG, H. RAO, M. LIN, Y. XU and P. SHI, *Optimal sensor scheduling for remote state estimation with limited bandwidth: A deep reinforcement learning approach*, Information Sciences **588**, 2022, pp. 279–292.

[18] I. A. ZAMFIRACHE, R.-E. PRECUP and E. M. PETRIU, *Adaptive reinforcement learning-based control using proximal policy optimization and slime mould algorithm with experimental tower crane system validation*, Applied Soft Computing **160**, 2024, paper 111687.

[19] J. HAN, C. YANG, C. C. LIM, C. C. ZHOU and P. SHI, *Stackelberg game approach for robust optimization with fuzzy variables*, IEEE Transactions on Fuzzy Systems **30**(1), 2020, pp. 258–269.

[20] L. U. ZHENG-LIANG and U. H. LOK, *Dimension-reduced modeling for local volatility surface via unsupervised learning*, Romanian Journal of Information Science and Technology **27**(3–4), 2024, pp. 255–266.

[21] U. KILIC, E. S. ESSIZ and M. K. KELES, *Binary anarchic society optimization for feature selection*, Romanian Journal of Information Science and Technology **26**(3–4), 2023, pp. 351–364.

[22] R.-E. PRECUP, E.-L. HEDREA, R.-C. ROMAN, E. M. PETRIU, A.-I. SZEDLAK-STINEAN and C.-A. BOJAN-DRAGOS, *Experiment-based approach to teach optimization techniques*, IEEE Transactions on Education **64**(2), 2021, pp. 88–94.

[23] J. LIU, Y. SHI, Z. M. FADLULLAH and N. KATO, *Space-air-ground integrated network: A survey*, IEEE Communications Surveys & Tutorials **20**(4), 2018, pp. 2714–2741.

[24] M. SAMIR, S. SHARAFEDDINE, C. M. ASSI, T. M. NGUYEN and A. GHRAYEB, *UAV trajectory planning for data collection from time-constrained IoT devices*, IEEE Transactions on Wireless Communications **19**(1), 2019, pp. 34–46.

[25] I. CUMALI, B. OZBEK, G. K. KURT and H. YANIKOMEROGLU, *User selection and codebook design for NOMA-based high altitude platform station (HAPS) communications*, IEEE Transactions on Vehicular Technology **72**(3), 2022, pp. 3636–3646.

[26] Y. LIU, L. JIANG, Q. QI and S. XIE, *Energy-efficient space–air–ground integrated edge computing for Internet of Remote Things: A federated DRL approach*, IEEE Internet of Things Journal **10**(6), 2022, pp. 4845–4856.

[27] A. H. ARANI, P. HU and Y. ZHUH, *HAPS-UAV-enabled heterogeneous networks: A deep reinforcement learning approach*, IEEE Open Journal of the Communications Society **4**, 2023, pp. 1745–1760.

[28] C. YOU and R. ZHANG, *3D trajectory optimization in Rician fading for UAV-enabled data harvesting*, IEEE Transactions on Wireless Communications **18**(6), 2019, pp. 3192–3207.

[29] M. D. NGUYEN, L. B. LE and A. GIRARD, *Integrated computation offloading, UAV trajectory control, edge-cloud and radio resource allocation in SAGIN*, IEEE Transactions on Cloud Computing **12**(1), 2023, pp. 100–115.

[30] K. FAN, B. FENG, X. ZHANG and Q. ZHANG, *Network selection based on evolutionary game and deep reinforcement learning in space-air-ground integrated network*, IEEE Transactions on Network Science and Engineering **9**(3), 2022, pp. 1802–1812.

[31] N. OKATI, T. RIIHONEN, D. KORPI, I. ANGERVUORI and R. WICHMAN, *Downlink coverage and rate analysis of low Earth orbit satellite constellations using stochastic geometry*, IEEE Transactions on Communications **68**(8), 2020, pp. 5120–5134.

[32] C. CARTIS, N. I. M. GOULD and P. L. TOINT, *On the complexity of steepest descent, Newton's and regularized Newton's methods for nonconvex unconstrained optimization problems*, SIAM Journal on Optimization **20**(6), 2010, pp. 2833–2852.

[33] S. MENG, K. CHHEA, S. MUY and JR. LEE, Supplementary material of the paper Sothearath MENG, Kimchheang CHHEA, Sengly MUY and Jung-Ryun LEE, *Deep Reinforcement Learning and Metaheuristic Approaches to Maximize Downlink Sum-Rate for IoT Systems in NOMA-based Space-Air-Ground Integrated Networks, Romanian Journal of Information Science and Technology*. [Online]. Available here.