DQN-Driven Beam-Hopping Pattern Optimization for LEO Satellite Communication Systems

Donghyeon Kim*, Haejoon Jung*, and In-Ho Lee†

* Department of Electronics and Information Convergence Engineering, Kyung Hee University, Yongin, 17104, South Korea
† School of Electronic and Electrical Engineering, Hankyong National University, Anseong, 17579, South Korea
Emails: {dhkim3988, haejoonjung}@khu.ac.kr and ihlee@hknu.ac.kr

***Abstract*—This paper addresses the problem of beam-hopping pattern design (BHPD) in low Earth orbit (LEO) satellite communication systems, with the objective of maximizing energy efficiency (EE) while satisfying the traffic demands of ground users. Specifically, the focus is placed on the joint optimization of BHPD and transmit power allocation, a task that is inherently challenging due to the mixed-integer nonlinear programming structure of the problem and the computational constraints typically encountered in LEO satellite systems. Conventional heuristic algorithms, such as matching-based approaches, have been extensively explored to tackle these issues. However, their performance remains suboptimal and is accompanied by significant computational overhead. To overcome these limitations, this work proposes a deep Q-network-based method for optimizing BHPD. Simulation results demonstrate that the proposed approach substantially improves both EE and outage performance when compared to existing methods.**

***Index Terms*—LEO satellite communications, beam-hopping pattern, deep learning**

I. Introduction

With the rapid advancement of communication technologies, user expectations for high-quality, ubiquitous communication services have significantly increased. Despite continuous improvements in terrestrial communication networks, coverage gaps persist in remote and challenging environments, thereby positioning satellite communication systems as an essential complementary infrastructure [1]. Among various satellite architectures, low Earth orbit (LEO) satellites have garnered substantial attention due to their inherent advantages, such as reduced round-trip latency and lower signal attenuation compared to geostationary orbit satellites [2]. In this context, beam-hopping (BH) has emerged as a key enabling technology for LEO satellite communication systems, attracting considerable interest from both academia and industry owing to its adaptability and implementation efficiency [3]. BH enhances resource utilization by activating a selected subset of beams in each time slot based on a predefined illumination pattern, which is periodically repeated across BH time windows. Furthermore, to alleviate inter-beam interference resulting from simultaneous beam activations, various resource allocation strategies have been developed [4].

In order to effectively mitigate inter-beam interference in LEO satellite systems, the joint optimization of BH pattern design (BHPD) and transmit power allocation has been extensively studied [5], [6]. Furthermore, given the limited onboard energy resources and the growing demand for high-capacity communication, energy efficiency (EE) maximization, while simultaneously meeting the traffic requirements of ground users, has emerged as a critical design objective in LEO resource management frameworks [7], [8]. However, the joint optimization of BHPD and transmit power allocation constitutes a mixed-integer nonlinear programming (MINLP) problem, which is inherently difficult to solve optimally due to its combinatorial and non-convex nature. To address this challenge, existing studies typically adopt either decomposition-based methods [7] or direct joint optimization approaches [8]. The former often relies on problem relaxations and approximations, which may compromise optimality, while the latter, although capable of achieving globally optimal solutions in theory, faces significant scalability issues owing to its exponentially expanding solution space.

To tackle the aforementioned challenge, the authors in [9] introduce a decomposition method that splits the original MINLP problem into two tractable subproblems: an integer programming component and a nonlinear programming component. Notably, this decomposition does not involve any form of relaxation or approximation, thereby preserving full equivalence with the original problem. Nonetheless, even after decomposition, solving the BHPD subproblem remains computationally intractable, particularly in large-scale LEO satellite networks where the solution space is substantially large. To tackle this issue, various studies have employed matching algorithms for BHPD optimization in LEO systems [10]. However, conventional matching approaches, which typically rely on stochastic or random reassignment of candidate matches, often struggle to guarantee optimality. To overcome these limitations, a greedy search-based matching (GSM) algorithm has been introduced in [11]. By iteratively selecting matchings that yield the most favorable improvement in the objective function, the GSM method demonstrates improved convergence properties and enhanced performance relative to traditional matching techniques.

Although the GSM algorithm demonstrates improved performance over traditional matching methods, it still faces challenges in achieving optimal solutions due to its inherently greedy and iterative nature. Also, the iterative search process imposes a high computational burden, which limits its practicality in large-scale scenarios. To address the computational

complexity associated with conventional heuristic and iterative approaches, deep neural network (DNN)-based resource allocation frameworks have garnered increasing attention, as in [12]. In particular, deep Q-network (DQN)-based strategies, as explored in [8], [13], have shown promise for solving the integer programming component, such as the BHPD optimization problem. Compared to GSM-based approaches, DQN techniques have the potential to overcome local optima by maximizing long-term cumulative rewards. However, due to the extensive action space associated with BHPD optimization, multi-agent approaches have been explored, as in [8]. Nevertheless, multi-agent DQN frameworks continue to encounter challenges related to their distributed optimization method and often exhibit suboptimal convergence behavior.

Motivated by these insights, this paper proposes a DQN-based BHPD optimization algorithm for LEO satellite communication systems, with the objective of maximizing EE while ensuring that ground user traffic demands are satisfied. Unlike the existing DQN method [8], the proposed centralized approach sequentially optimizes each BHPD index to reduce the action space and mitigate the complexity associated with multi-agent frameworks. Extensive simulation results demonstrate that the proposed scheme significantly outperforms conventional methods in terms of EE and outage performance, underscoring its potential as a viable and efficient solution for next-generation LEO satellite communication networks.

II. System Model

We consider a downlink LEO satellite communication system in which a satellite is equipped with M_{fix} Earth-fixed beams. Among these, the satellite can simultaneously activate M_{act} spot beams to illuminate M_{act} distinct ground positions, as illustrated in Fig. 1. Let $\mathbb{G} = \{1, 2, ..., G\}$ denote the set of beam groups and $\mathbb{J} = \{1, 2, ..., M_{fix}\}$ denote the set of fixed beam indices. The total number of groups G is given by $G = \lceil M_{fix}/M_{act} \rceil$, where each group corresponds to a unique configuration of M_{act} active beams. We define $x_{g,j} \in \{0, 1\}$ as the BHPD indicator variable, subject to the constraints $\sum_{g\in\mathbb{G}} x_{g,j} = 1$ and $\sum_{j\in\mathbb{J}} x_{g,j} = M_{act}$, ensuring that each beam is assigned to exactly one group and each group comprises exactly M_{act} active beams.

Furthermore, each beam serves K users, who are allocated orthogonal channels via frequency division multiple access (FDMA), with the total system bandwidth B equally partitioned across the users. All beams are assumed to have an identical coverage radius, which is determined by the 3 dB beam-width angle, denoted as θ_{3dB}. Let T_g represent the number of time slots allocated to group g for the operation of its M_{act} active beams, where $g \in \mathbb{G}$. The total time slot allocation must satisfy the constraint $\sum_{g\in\mathbb{G}} T_g \leq T_{max}$, where T_{max} denotes the total number of time slots within a BH window. We assume that the duration of the BH window is fixed and that all beam hopping operations are completed within this interval. Additionally, for analytical tractability, we consider uniform time slot allocation across groups, i.e., $T_g = T_{max}/G$ for all $g \in \mathbb{G}$.

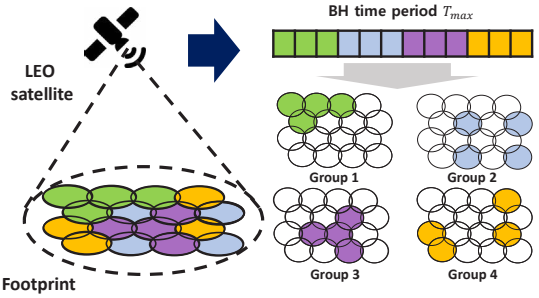

Fig. 1. An example of LEO satellite system, where $M_{fix} = 16$, $M_{act} = 4$, and $G = 4$.

We define the channel gain experienced by user k, located within the coverage area of beam $j \in \mathbb{J}$, from a neighboring interfering beam $l \in \mathbb{J}\backslash\{j\}$, as follows [14]:

$$h_{j,l,k} = \frac{\alpha_{UE}\alpha_{LEO}}{L_k}\beta_{j,l,k}, \tag{1}$$

where G_{UE} denotes the antenna gain of the ground user, and $G_{LEO} = \zeta(70\pi/\theta_{3dB})^2$ represents the peak antenna gain of the LEO satellite, where ζ is the antenna efficiency parameter, and θ_{3dB} is the 3 dB beam-width angle [15]. Upon determination of the BHPD index, let $\mathbb{M}_g$ denote the set of M_{act} active beams assigned to group g, and let $\mathbb{K}_{g,m}$ represent the set of users served by beam $m \in \mathbb{M}_g$. The channel gain between user k and beam m in group $g \in \mathbb{G}$ is denoted by $h_{g,m,k}$. The overall channel loss L_k experienced by user k includes both free-space path loss and shadowing effects at the carrier frequency f_c, where the shadowing component is modeled as a log-normal random variable with standard deviation σ_s. Let $\theta_{g,m,k}$ denote the angular separation between the central axis of beam $m \in \mathbb{M}_g$ and user $k \in \mathbb{K}_{g,m}$. The beam pattern gain for user k from beam m in group g is denoted by $\beta_{g,m,k}$ and is defined as [16]:

$$\beta_{g,m,k} = \left(\frac{J_1(\eta_{g,m,k})}{2\eta_{g,m,k}} + \frac{36J_3(\eta_{s,n,u})}{\eta_{s,n,u}^3}\right)^2, \tag{2}$$

where $\eta_{g,m,k} = 2.07123\sin(\theta_{g,m,k})/\sin(\theta_{3dB})$, and $J_1(\cdot)$ and $J_3(\cdot)$ are the Bessel functions of the first kind of order one and three, respectively. Throughout the BH window, the channel gains are assumed to remain quasi-static, i.e., constant over the scheduling duration.

The achievable capacity for user k is expressed as

$$\Gamma_{g,m,k} = \frac{T_g B}{T_{max} K}\log_2\left(1 + \frac{\tilde{h}_{g,m,k}p_{g,m,k}}{\sum_{i\in\mathbb{M}_g\backslash\{m\}}\tilde{h}_{g,i,k}p_{g,i,k} + 1}\right), \tag{3}$$

where $p_{g,m,k}$ denotes the transmit power assigned to user k, and $\tilde{h}_{g,m,k} = h_{g,m,k}/(\sigma_n B/K)$ represents the normalized channel gain, interpreted as the effective transmit signal-to-noise ratio (SNR), with σ_n being the noise power spectral density. Subsequently, the EE of the system is defined as [17]

$$EE = \frac{\sum_{g\in\mathbb{G}}\sum_{g\in\mathbb{M}_g}\sum_{k\in\mathbb{K}_{g,m}}\Gamma_{g,m,k}}{\sum_{g\in\mathbb{G}}\frac{T_g}{T_{max}}\sum_{g\in\mathbb{M}_g}\sum_{k\in\mathbb{K}_{g,m}}p_{g,m,k} + P_c}, \tag{4}$$

where P_c denotes the constant circuit power consumption of the satellite payload.

Accordingly, in LEO satellite communication systems, we reformulate the resource allocation problem as a joint optimization of BHPD and transmit power, as follows:

$$\mathcal{P}_0: \max \quad EE, \tag{5a}$$

$$\text{s.t.} \sum_{g\in\mathbb{G}} x_{g,j} = 1,\ \forall j \in \mathbb{J}, \tag{5b}$$

$$\sum_{j\in\mathbb{J}} x_{g,j} = M_{act},\ \forall g \in \mathbb{G}, \tag{5c}$$

$$\sum_{g\in\mathbb{G}} \frac{T_g}{T_{max}} \sum_{g\in\mathbb{M}_g} \sum_{k\in\mathbb{K}_{g,m}} p_{g,m,k} \le P_{max}, \tag{5d}$$

$$\Gamma_{g,m,k} \ge D_{g,m,k},\ \forall k \in \mathbb{K}_{g,m}, \forall m \in \mathbb{M}_g, \forall g \in \mathbb{G}. \tag{5e}$$

Here, constraints (5b) and (5c) enforce the structure of the BHPD index, ensuring that each beam is allocated to one and only one group, and each group contains a fixed number of active beams. Constraint (5d) imposes an upper bound on the total transmit power, such that it does not exceed the satellite's maximum power budget $P_{\max}$. Furthermore, constraint (5e) guarantees that the data rate of each user meets the corresponding traffic demand requirement $D_{g,m,k}$.

III. Proposed BHPD Optimization for LEO Satellite Systems

In this section, we introduce the proposed DQN-based BHPD optimization scheme to address the problem defined in (5). Prior to presenting the detailed algorithmic framework, we first decompose the original joint optimization problem into two tractable subproblems: an integer programming subproblem corresponding to BHPD optimization and a nonlinear programming subproblem corresponding to power allocation. These subproblems are equivalent in structure to the original formulation.

Let the total transmission power associated with the EE maximization problem $\mathcal{P}_0$ be defined as $P_{tot} = \sum_{g\in\mathbb{G}} \frac{T_g}{T_{max}} \sum_{m\in\mathbb{M}g} \sum_{k\in\mathbb{K}_{g,m}} p_{g,m,k}$. For any fixed value of P_{tot}, the original problem can be equivalently reformulated as follows:

$$\mathcal{P}_1: \max \frac{\sum_{g\in\mathbb{G}} \sum_{g\in\mathbb{M}_g} \sum_{k\in\mathbb{K}_{g,m}} \Gamma_{g,m,k}}{P_{tot} + P_c}, \tag{6a}$$

$$\text{s.t.} \sum_{g\in\mathbb{G}} x_{g,j} = 1,\ \forall j \in \mathbb{J}, \tag{6b}$$

$$\sum_{j\in\mathbb{J}} x_{g,j} = M_{act},\ \forall g \in \mathbb{G}, \tag{6c}$$

$$\sum_{g\in\mathbb{G}} \frac{T_g}{T_{max}} \sum_{g\in\mathbb{M}_g} \sum_{k\in\mathbb{K}_{g,m}} p_{g,m,k} = P_{tot}, \tag{6d}$$

$$\Gamma_{g,m,k} \ge D_{g,m,k},\ \forall k \in \mathbb{K}_{g,m}, \forall m \in \mathbb{M}_g, \forall g \in \mathbb{G}. \tag{6e}$$

Since the denominator $P_{tot}+P_c$ in the EE expression becomes constant under fixed P_{tot}, the objective function can be equivalently represented as $\sum_{g\in\mathbb{G}} \sum_{g\in\mathbb{M}_g} \sum_{k\in\mathbb{K}_{g,m}} \Gamma_{g,m,k}$. Moreover, since $\sum_{g\in\mathbb{G}} \sum_{m\in\mathbb{M}_g} \sum_{k\in\mathbb{K}_{g,m}} D_{g,m,k}$ is constant, the objective function can be equivalently expressed as $\sum_{g\in\mathbb{G}} \sum_{g\in\mathbb{M}_g} \sum_{k\in\mathbb{K}_{g,m}} (\Gamma_{g,m,k} - D_{g,m,k})$. This formulation reveals that the structure of the objective function aligns with the aggregate form of the traffic demand constraints in (6e), thereby providing insights into the tractability of the decomposed optimization problem.

By reformulating the constraint in (6e), we derive the following expression

$$\begin{aligned}&\Gamma_{g,m,k} \ge D_{g,m,k}\\ &\equiv \frac{\tilde{h}_{g,m,k}}{\Lambda_{g,m,k}} p_{g,m,k} - \sum_{i\in\mathbb{M}_g\setminus\{m\}} \tilde{h}_{g,i,k} p_{g,i,k} \ge 1,\end{aligned}$$

where $\Lambda_{g,m,k} = 2^{KD_{g,m,k}T_{max}/(BT_g)} - 1$. Let us define $\boldsymbol{\Xi}_{g,k}$ as an $M_{act} \times M_{act}$ matrix whose diagonal elements are given by $\tilde{h}_{g,m,k}/\Lambda_{g,m,k}$ and off-diagonal elements are $-\tilde{h}_{g,i,k}$. Using this definition, the above inequality can be compactly rewritten as

$$\begin{aligned}&\boldsymbol{\Xi}_{g,k}\boldsymbol{p}_{g,k} \ge \boldsymbol{e}\\ &\equiv p_{g,m,k} \ge p^{min}_{g,m,k}, \forall k \in \mathbb{K}_{g,m}, \forall m \in \mathbb{M}_g, \forall g \in \mathbb{G},\end{aligned}$$

where $\boldsymbol{p}_{g,k} = [p_{g,m,k}]_{1\times M_{act}}$ is the transmit power vector associated with user k in group g, and $\boldsymbol{e} = [1]_{1\times M_{act}}$ is the all-one vector. Accordingly, the minimum transmit power vector $\boldsymbol{p}^{min}_{g,k}$ required to satisfy the traffic demand $D_{g,m,k}$ can be obtained by solving the matrix equation $\boldsymbol{\Xi}^{-1}_{g,k}\boldsymbol{e}$, as previously demonstrated in [11]. This can also be equivalently expressed as

$$p^{min}_{g,m,k} = \frac{\Lambda_{g,m,k}\left(\sum_{i\in\mathbb{M}_g\setminus\{m\}} \tilde{h}_{g,i,k} p^{min}_{g,i,k} + 1\right)}{\tilde{h}_{g,m,k}}, \tag{7}$$

From the preceding analysis, it is evident that the summation in the final constraint of problem $\mathcal{P}_1$ structurally corresponds to the objective function itself. Consequently, the original problem can be equivalently reformulated as the following optimization problem:

$$\begin{aligned}&\mathcal{P}_2: \max \sum_{g\in\mathbb{G}} \frac{T_g}{T_{max}} \sum_{m\in\mathbb{M}_g} \sum_{k\in\mathbb{K}_{g,m}} p_{g,m,k} - p^{min}_{g,m,k}\\ &\text{s.t.(6b)} - \text{(6e)}.\end{aligned}$$

By exploiting the second constraint, the objective function can be rewritten as $P_{tot} - \sum_{g\in\mathbb{G}} \frac{T_g}{T_{max}} \sum_{m\in\mathbb{M}_g} \sum_{k\in\mathbb{K}_{g,m}} p^{min}_{g,m,k}$. Since P_{tot} is considered a fixed constant, the objective function can be further simplified to $-\sum_{g\in\mathbb{G}} \frac{T_g}{T_{max}} \sum_{m\in\mathbb{M}_g} \sum_{k\in\mathbb{K}_{g,m}} p^{min}_{g,m,k}$. Moreover, given that the second and third constraints correspond to the power allocation phase subsequent to BHPD, the original problem can be decomposed accordingly. Thus, the BHPD optimization subproblem can be expressed independently as

$$\mathcal{P}_{BHPD}: \min \sum_{g\in\mathcal{G}} \frac{T_g}{T_{max}} \sum_{m\in\mathcal{M}_g} \sum_{k\in\mathcal{K}_{g,m}} p^{min}_{g,m,k} \tag{8a}$$

$$\text{s.t.} \sum_{g\in\mathbb{G}} x_{g,j} = 1,\ \forall j \in \mathbb{J}, \tag{8b}$$

$$\sum_{j\in\mathbb{J}} x_{g,j} = M_{act},\ \forall g \in \mathbb{G}. \tag{8c}$$

Correspondingly, the remaining power allocation subproblem can be formulated as

$$\mathcal{P}_{PA}: \max \quad EE \tag{9a}$$

$$\text{s.t.} \sum_{g\in\mathbb{G}} \frac{T_g}{T_{max}} \sum_{g\in\mathbb{M}_g} \sum_{k\in\mathbb{K}_{g,m}} p_{g,m,k} \le P_{max}, \tag{9b}$$

$$\Gamma_{g,m,k} \ge D_{g,m,k},\ \forall k \in \mathbb{K}_{g,m}, \forall m \in \mathbb{M}_g, \forall g \in \mathbb{G}. \tag{9c}$$

This decomposition confirms that the BHPD optimization problem $\mathcal{P}_{BHPD}$ is consistently derived from the original problem $\mathcal{P}_0$ for any given P_{tot}. Therefore, the optimal solution to the BHPD component can be effectively obtained by solving $\mathcal{P}_{BHPD}$, obviating the need to address the joint optimization problem $\mathcal{P}_0$ directly.

A. Proposed DQN-Based BHPD Optimization

The proposed BHPD optimization algorithm utilizes a DQN-based approach to address the optimization problem $\mathcal{P}_{BHPD}$. Within this framework, each episode ι corresponds to a unique realization of the BH window, implying that a new episode commences whenever the channel conditions undergo variation, where the total number of episodes is denoted by ι_{max}. Each episode comprises a fixed number of discrete decision-making intervals, indexed by t, with a maximum of t_{max} steps. In contrast to conventional multi-agent learning schemes, the proposed method incrementally optimizes the BHPD index at each time step. This sequential design aims to mitigate the exponential growth of the BHPD index search space, thereby improving computational tractability. The underlying decision-making framework is modeled as a Markov Decision Process (MDP), which is defined as follows:

- *State*: Within the DQN framework, the LEO satellite continuously observes the system state at each time step t. This state comprises multiple components: the normalized transmit SNR, the ratio of the desired signal to interference and the users' traffic demand, and the current BHPD allocation. Formally, the system state is expressed as

$$\boldsymbol{s}(t) = [\boldsymbol{\Psi}, \boldsymbol{\Pi}, \boldsymbol{\Upsilon}], \tag{10}$$

  where $\boldsymbol{\Psi} = \log_{10}(1+\tilde{h}_{j,l,k}), \forall j \in \mathbb{J}, \forall l \in \mathbb{J}$ represents the normalized channel gain. Moreover, $\boldsymbol{\Pi} = \log_2(1 + \tilde{h}_{g,m,k}/\sum_{i\in\mathbb{M}_g\setminus\{m\}} \tilde{h}_{g,i,k})/D_{g,m,k}, \forall k \in \mathbb{K}_{g,m}, \forall m \in \mathbb{M}_g, \forall g \in \mathbb{G}$ and $\boldsymbol{\Upsilon} = x_{g,j}, \forall j \in \mathbb{J}, \forall g \in \mathbb{G}$.
- *Action*: The action space for the BHPD optimization is formally defined as follows:

$$\boldsymbol{a}(t) \in \{\{\Delta_j, \Delta_{\bar{j}}\}, \varnothing\}, \tag{11}$$

  where the notation $\{\Delta_j, \Delta_{\bar{j}}\}$ denotes a pairwise exchange of the BHPD index between beam $j \in \mathbb{J}$ and another beam $\bar{j} \in \mathbb{J} \setminus \{j\}$. The action $\varnothing$ represents a no-operation (i.e., the current BHPD configuration remains unchanged). Consequently, the total size of the action space in the proposed DQN-based framework is given by ${}_{M_{fix}}\mathrm{C}_2 + 1$, where ${}_{M_{fix}}\mathrm{C}_2 = M_{fix}!/((M_{fix}-2)!2!)$ denotes the number of distinct beam-pair combinations.
- *Reward*: To align with the objective of problem $\mathcal{P}_{BHPD}$, the reward structure is designed to reflect changes in the cost function Ω, which is defined as

$$\Omega(t) = \sum_{g\in\mathbb{G}} \frac{T_g}{T_{max}} \sum_{g\in\mathbb{M}_g} \sum_{k\in\mathbb{K}_{g,m}} p^{DQN}_{g,m,k}. \tag{12}$$

  In calculating $p^{DQN}_{g,m,k}$, the transmit power is estimated based on the expression in (7), assuming an equal power allocation strategy given by $p_{g,m,k} = P_{max}/(M_{fix}KT_{max})$. Accordingly, the reward at time step t is determined by

$$\boldsymbol{r}(t) \in \{\pm 1, 0\}. \tag{13}$$

  where the reward is $\boldsymbol{r}(t) = 1$ if $\Omega(t) < \Omega(t-1)$, which indicates a decrease in the objective function. Conversely, when $\Omega(t) > \Omega(t-1)$, the reward is set to -1. If the cost function is same, that is, $\Omega(t) = \Omega(t-1)$ (corresponding to $\boldsymbol{a}(t) = \varnothing$), the reward is assigned a value of 0.

Algorithm 1: Training Mechanism for Proposed DQN-Based BHPD Optimization

```
Initialize Q-network Q(Θ) and target network Q(Θ').
Initialize the replay buffer, batch size O, epsilon greedy
  parameter ϵ, max time step t_max, discount factor γ, and
  target network update period τ_max.
for ι = 1 : ι_max do
    Initialize s(0) and Ω(0).
    for t = 1 : t_max do
        Select action a(t) based on ϵ-greedy as
        a(t) = { Random ∈ {{Δ_j, Δ_j̄}, ∅},          if ϵ,
               { arg max_{a(t)} Q(s(t), a(t); Θ),    if 1 − ϵ.
        Conduct a(t) and obtain s(t + 1).
        Calculate Ω(t) and r(t).
        Store the sample (s(t), a(t), r(t), s(t + 1)) in the
          replay buffer.
        if replay buffer size ≥ O then
            Obtain O random samples in the replay buffer
              and calculate L in (14).
            Trainable parameters (Θ) are updated.
        end
    end
    if mod (ι, ι_max) = 0 then
        Target network is updated by (Θ' ← Θ)
    end
end
```

To facilitate the training of the proposed model, the agent concurrently updates the Q-network $Q(\boldsymbol{\Theta})$ and the target network $Q(\boldsymbol{\Theta}')$, where $\boldsymbol{\Theta}$ and $\boldsymbol{\Theta}'$ represent the respective sets of trainable parameters for each network. Both networks receive the state vector defined in (10) as input and output the corresponding action values for the defined action space. At each discrete time step, the agent selects an action according to the ϵ-greedy strategy, which balances the trade-off between exploration and exploitation. The chosen action transitions the

TABLE I
SIMULATION PARAMETERS

System parameters	Values
Number of active beams (M_{act})	4
Number of fixed beams (M_{fix})	16
Number of users (K)	2-6
Number of time slots (T_{max})	32
Maximum transmission power (P_{max})	53 [dBm]
The 3 dB beam-width angle (θ_{3dB})	3°
Total bandwidth (B)	10 [MHz]
Ground user antenna gain (α_{UE})	23 [dB]
Antenna efficiency (ζ)	0.5
Carrier frequency (f_c)	2 [GHz]
Standard deviation of shadowing (σ_s)	4 [dB]
Elevation angle	$\{10°, 20°, 30°, 40°, 50°, 60°, 70°, 80°, 90°\}$
Earth radius	6,371 [km]
Satellite altitude	600 [km]
Noise spectral density (σ_n)	-174 [dBm/Hz]
Circuit power consumption (P_c)	30 [dBm]
Average traffic demand	0.4-0.65 [Mbps]
Number of data instances	9,000
DQN parameters	**Values**
Number of episodes (ι_{max})	9,000
Number of time steps (t_{max})	32
Number of hidden layers	2
Number of hidden nodes	512
Capacity of replay buffer	1,000
Batch size (O)	32
Discount factor (γ)	0.2
Target network update period (τ_{tar})	10
Learning rate	0.001

system to a new state $\boldsymbol{s}(t+1)$, from which the updated cost function $\Omega(t)$ and the corresponding reward $\boldsymbol{r}(t)$ are derived. The resulting tuple $(\boldsymbol{s}(t), \boldsymbol{a}(t), \boldsymbol{r}(t), \boldsymbol{s}(t+1))$ is stored in a replay buffer for experience replay.

Let O denote the mini-batch size used during training. Once the number of samples in the replay buffer exceeds O, the Q-network is updated by minimizing a loss function based on the log-cosh metric, as follows:

$$L = \frac{1}{O}\sum_{o=1}^{O} \ln\left(\cosh(Q\left(\boldsymbol{s}(o), \boldsymbol{a}(o); \boldsymbol{\Theta}\right) - Z(o)\right), \tag{14}$$

where $Z(o)$, the target Q-value, which is defined as

$$Z(o) = \boldsymbol{r}(o) + \gamma \max_{\boldsymbol{a}(\hat{o})} Q(\boldsymbol{s}(\hat{o}), \boldsymbol{a}(\hat{o}); \boldsymbol{\Theta}^{'}). \tag{15}$$

Here, $\hat{o}$ represents the subsequent time step following o, and γ denotes the discount factor that modulates the importance of future rewards. To enhance training stability, the parameters of the target network are periodically updated. Specifically, if the condition $\mod(\iota, \tau_{tar}) = 0$ is satisfied, the target network's weights are softly synchronized with those of the main Q-network. Here, τ_{tar} indicates the update interval of the target network. The complete training protocol for the proposed DQN-based BHPD optimization algorithm is summarized in Alg. 1.

IV. SIMULATION RESULTS

This section provides a thorough performance evaluation of the proposed scheme in terms of EE, outage rate, and execution time. The simulation parameters, adopted from the configurations in [11], [14], [16], are summarized in Table I. Each simulation instance reflects a distinct user deployment scenario. A total of 1000 independent data instances are considered, where each instance includes 9 variations in elevation angle ranging from 10° to 90°. It is important to note that the shadowing effects are also varied in accordance with the elevation angle. Consequently, the overall performance evaluation encompasses 1000×9 episodes. A separate, non-overlapping dataset comprising another 1000×9 episodes is utilized for training the proposed model. For the DQN training, the exploration rate ϵ is initialized to 1, and decays multiplicatively by a factor of 0.995 until reaching a minimum value of 0.01. The DNN employed within the DQN framework consists of two hidden layers, each comprising 512 neurons, with Sigmoid functions used as activation mechanisms.

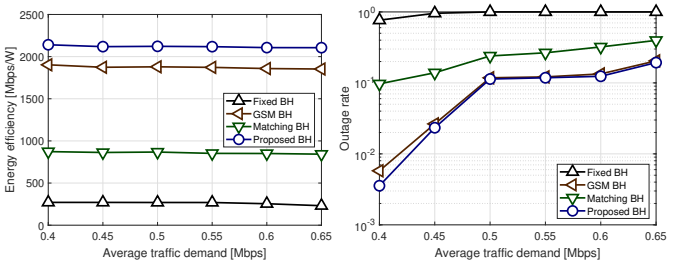

Fig. 2. EE and outage rate performance under varying average traffic demand, with a fixed total user count of 48

For benchmarking purposes, the performance of the proposed algorithm is compared against the following baseline schemes:

1) *Fixed BH:* This baseline utilizes a static BHPD without dynamic adaptation. As a result, it lacks the ability to effectively manage inter-beam interference or optimize EE and outage probability.
2) *GSM BH:* This approach is based on the GSM algorithm proposed in [11]. While it improves over fixed methods, it remains suboptimal due to its greedy iterative nature and incurs high computational complexity.
3) *Matching BH:* This method employs a matching algorithm for BHPD optimization as introduced in [16]. However, because it relies on random pairwise exchanges for beam matching, it cannot guarantee convergence to the global optimum.
4) *Proposed BH:* This refers to the DQN-based BHPD optimization approach introduced in Section III-A. All the aforementioned schemes utilize the optimal algorithm from [18] to solve the nonlinear power allocation problem $\mathcal{P}_{PA}$ optimally.

Fig. 2 illustrates the impact of varying average traffic demand on EE and outage rate under a fixed total user count of 48. An outage event is defined as a scenario wherein any user fails to meet its assigned traffic requirement. As shown, the fixed BH scheme exhibits the poorest performance in terms of both EE and outage rate, primarily due to its static BHPD. The matching BH approach shows moderate improvement over the fixed BH scheme; however, its performance remains

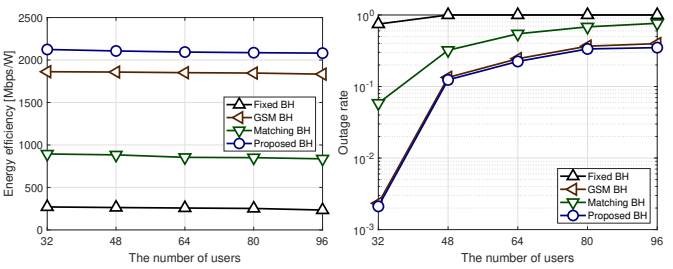

Fig. 3. EE and outage rate performances under varying total number of users, with a fixed average traffic demand of 0.6 Mbps.

constrained by the inherent limitations of its random swapping mechanism. The GSM BH method outperforms both fixed and matching BH schemes, yet its greedy convergence behavior imposes a ceiling on achievable performance. In contrast, the proposed DQN-based scheme is specifically designed to maximize cumulative rewards, thereby effectively overcoming the limitations of the GSM method and achieving enhanced EE and reduced outage probability.

Fig. 3 presents the relationship between the total number of users, $M_{fix}K$, and the resulting EE and outage rate, with the average traffic demand fixed at 0.6 Mbps. Across all user count regions, the proposed BH scheme consistently demonstrates superior performance in both metrics. Additionally, Fig. 4 evaluates the computational time required by each scheme as the number of users increases. Notably, both the matching and GSM algorithms exhibit substantial execution time due to their iterative BHPD optimization processes. In contrast, the proposed DQN-based BH scheme incurs significantly lower computational complexity, owing to its ability to perform BHPD optimization through a pre-trained DNN.

V. Conclusions

In this paper, we have proposed a DQN-driven BHPD optimization framework that aims to enhance the EE of LEO satellite communication systems while guaranteeing that the traffic demands of ground users are met. Given the difficulty of achieving optimal solutions for the considered EE maximization problem, we have first decomposed the original MINLP formulation into two tractable subproblems: BHPD and transmit power optimizations. To further address the computational burden and subpar EE performance of conventional techniques, the DQN-based optimization method was introduced for the BHPD. Numerical evaluations confirmed that the proposed algorithm achieves significant gains in EE while simultaneously reducing the outage probability. In the future, we plan to comprehensively address practical issues in LEO satellites, such as the Doffler effect and outdated channels.

Acknowledgement

The work was supported by the MSIT, Korea, in part under the National Research Foundation of Korea grant (RS-2022-NR069055), in part under the ITRC support programs (IITP-2025- RS-2021-II212046), and in part under the Convergence security core talent training business support program (IITP-2023-RS-2023-00266615) supervised by the IITP.

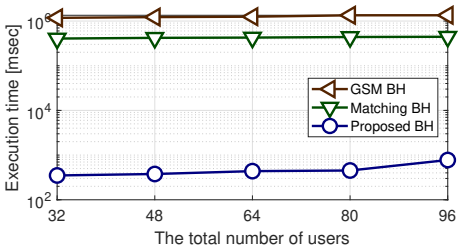

Fig. 4. Computational time variation with respect to the total number of users.

References

[1] J. Liu, Y. Shi, Z. M. Fadlullah, and N. Kato, "Space-air-ground integrated network: A survey," *IEEE Commun. Surv. Tutor.*, vol. 20, no. 4, pp. 2714–2741, 4th Quart. 2018.

[2] F. Rinaldi, H.-L. Maattanen, J. Torsner, S. Pizzi, S. Andreev, A. Iera, Y. Koucheryavy, and G. Araniti, "Non-terrestrial networks in 5G & beyond: A survey," *IEEE Access*, vol. 8, pp. 165 178–165 200, Sep. 2020.

[3] J. Wang, C. Qi, S. Yu, and S. Mao, "Joint beamforming and illumination pattern design for beam-hopping LEO satellite communications," *IEEE Trans. Wireless Commun.*, vol. 23, no. 12, pp. 18 940–18 950, Dec. 2024.

[4] H. Deng, K. Ying, D. Feng, L. Gui, Y. He, and X.-G. Xia, "Satellites beam hopping scheduling for interference avoidance," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 12, pp. 3647–3658, Oct. 2024.

[5] L. Lei, E. Lagunas, Y. Yuan, M. G. Kibria, S. Chatzinotas, and B. Ottersten, "Deep learning for beam hopping in multibeam satellite systems," in *Proc. 2020 IEEE 91st Veh. Technol. Conf. (VTC2020-Spring)*, May 2020, pp. 1–5.

[6] Z. Lin, Z. Ni, L. Kuang, C. Jiang, and Z. Huang, "Satellite-terrestrial coordinated multi-satellite beam hopping scheduling based on multi-agent deep reinforcement learning," *IEEE Trans. Wireless Commun.*, vol. 23, no. 8, pp. 10 091–10 103, Aug. 2024.

[7] H. Yang, D. Yang, Y. Li, and J. Kuang, "Cluster-based beam hopping for energy efficiency maximization in flexible multibeam satellite systems," *IEEE Commun. Lett.*, vol. 27, no. 12, pp. 3300–3304, Dec. 2023.

[8] Y. Ran, F. Tan, S. Chen, J. Lei, and J. Luo, "Towards beam hopping and power allocation in multi-beam satellite systems with parameterized reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 73, no. 9, pp. 14 050–14 055, Sep. 2024.

[9] D. Kim, H. Jung, I.-H. Lee, and D. Niyato, "Novel resource allocation algorithm for IoT networks with multicarrier NOMA," *IEEE Internet of Things J.*, vol. 11, no. 18, pp. 30 354–30 367, Sep. 2024.

[10] G. Cui, X. Xin, L. Xu, W. Wang, and X. Tang, "Joint beam hopping and precoding for dense LEO satellite communication systems," *IEEE Internet Things J.,* early access, Jun. 2025.

[11] D. Kim, H. Jung, I.-H. Lee, and D. Niyato, "Multibeam management and resource allocation for LEO satellite-assisted IoT networks," *IEEE Internet Things J.*, vol. 12, no. 12, pp. 19 443–19 458, Jun. 2025.

[12] B. Mao, F. Tang, Y. Kawamoto, and N. Kato, "Optimizing computation offloading in satellite-UAV-served 6G IoT: A deep learning approach," *IEEE Netw.*, vol. 35, no. 4, pp. 102–108, Aug. 2021.

[13] D. Kim, H. Jung, and I.-H. Lee, "DQN-based scheduling algorithm for beam-hopping LEO satellite communication systems," *IEEE Wireless Commun. Lett.*, vol. 14, no. 8, pp. 2401–2405, Aug. 2025.

[14] L. Lei, A. Wang, E. Lagunas, X. Hu, Z. Zhang, Z. Wei, and S. Chatzinotas, "Spatial–temporal resource optimization for uneven-traffic LEO satellite systems: Beam pattern selection and user scheduling," *IEEE J. Sel. Areas Commun.*, vol. 42, no. 5, pp. 1279–1291, May 2024.

[15] A. Wang, L. Lei, X. Hu, E. Lagunas, A. I. Pérez-Neira, and S. Chatzinotas, "Adaptive beam pattern selection and resource allocation for NOMA-based LEO satellite systems," in *Proc. GLOBECOM 2022 - 2022 IEEE Global Commun. Conf.*, Dec. 2022, pp. 674–679.

[16] H.-H. Choi, G. Park, K. Heo, and K. Lee, "Joint optimization of beam placement and transmit power for multibeam LEO satellite communication systems," *IEEE Internet Things J.*, vol. 11, no. 8, pp. 14 804–14 813, Apr. 2024.

[17] A. Zappone and E. Jorswieck, *Energy Efficiency in Wireless Networks via Fractional Programming Theory*. Now Foundations and Trends, 2015.

[18] R. H. Byrd, M. E. Hribar, and J. Nocedal, "An interior point algorithm for large-scale nonlinear programming," *SIAM J. Optim.*, vol. 9, no. 4, pp. 877–900, Sep. 1999.