Chinese Journal of Electronics
vol. 33, no. 5, pp. 1286–1295, September 2024
https://doi.org/10.23919/cje.2022.00.135

RESEARCH ARTICLE

A Recursive DRL-Based Resource Allocation Method for Multibeam Satellite Communication Systems

Haowei MENG[1], Ning XIN[2], Hao QIN[1], and Di ZHAO[1]

1. *State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, China*
2. *China Academy of Space Technology, Institute of Telecommunication Satellite, Beijing 100094, China*

Corresponding author: Hao QIN, Email: hqin@mail.xidian.edu.cn

Manuscript Received May 17, 2022; Accepted June 20, 2023

Abstract — Optimization-based radio resource management (RRM) has shown significant performance gains on high-throughput satellites (HTSs). However, as the number of allocable on-board resources increases, traditional RRM is difficult to apply in real satellite systems due to its intense computational complexity. Deep reinforcement learning (DRL) is a promising solution for the resource allocation problem due to its model-free advantages. Nevertheless, the action space faced by DRL increases exponentially with the increase of communication scale, which leads to an excessive exploration cost of the algorithm. In this paper, we propose a recursive frequency resource allocation algorithm based on long-short term memory (LSTM) and proximal policy optimization (PPO), called PPO-RA-LOOP, where RA means resource allocation and LOOP means the algorithm outputs actions in a recursive manner. Specifically, the PPO algorithm uses LSTM network to recursively generate sub-actions about frequency resource allocation for each beam, which significantly cuts down the action space. In addition, the LSTM-based recursive architecture allows PPO to better allocate the next frequency resource by using the generated sub-actions information as a prior knowledge, which reduces the complexity of the neural network. The simulation results show that PPO-RA-LOOP achieved higher spectral efficiency and system satisfaction compared with other frequency allocation algorithms.

Keywords — High-throughput satellites, Proximal policy optimization, Deep reinforcement learning, Long-short term memory.

Citation — Haowei MENG, Ning XIN, Hao QIN, *et al.*, "A Recursive DRL-Based Resource Allocation Method for Multibeam Satellite Communication Systems," *Chinese Journal of Electronics*, vol. 33, no. 5, pp. 1286–1295, 2024. doi: 10.23919/cje.2022.00.135.

I. Introduction

With the development of 5G, terrestrial mobile communication networks have achieved significant progress. However, there is a situation that terrestrial networks will become congested or dysfunctional due to emergencies. Multi-beam satellite (MBS) can be regarded as an efficient supplement of the terrestrial communication network [1] with the flexible on-board payloads such as the digital channelizer, which is a critical on-board payload on MBSs, allowing dynamic channel allocations to efficiently exploit the system spectrum resource [2], [3]. As the channel reuse technology may bring about severe co-channel interference (CCI) among beams, suitable and efficient frequency resource allocation algorithms are needed to further improve the performance of satellite systems [4]–[6].

Precoding technology is a means to solve the problem of CCI. Zhang *et al.* proposed a precoding scheme facing phase perturbations for the downlink of MBS communication, which achieves good accuracy and efficiency compared to the benchmarks [7]. Zhang *et al.* proposed a successive convex approximation based algorithm for generating the beamforming vectors at the centralized processor, which is capable of providing feasible solutions within a few iterations [8]. In this paper, we focus on the resource allocation technique, especially frequency resource allocation, which is another way to reduce CCI. Some classical optimization algorithms such as the simulated annealing (SA) algorithm [9], which have

Associate Editor: Prof. Mugen PENG, Beijing University of Posts and Telecommunications.

achieved the desired effect in satellite resource allocation. Kawamoto *et al.* also suggested that optimization techniques are a valid and efficient approach to address the resource allocation problem [10]. However, at a larger scale, the number of resources to be managed, the constraints arising from the system and the massive number of traffic demand situations typically may result in a problem that conventional algorithms cannot solve optimally or are too complex to be implemented. In addition, typical radio resource management (RRM) problem involving carrier allocation can be regarded as an integer combinatorial problem [11]. This means that traditional optimization-based RRM algorithms must cost too numerous iterations and lots of time to reach convergence point.

Deep reinforcement learning (DRL) as a field of machine learning has appeared as a promising alternative for dealing with computationally expensive optimization procedure. DRL utilizes deep neural networks to update strategy with the aim of maximizing long-term benefits through interactions with the environment and is widely used to solve the Markov decision process (MDP) and make optimal decisions. Liu *et al.* adopted deep Q-network (DQN) to solve the problem of dynamic channel allocation [12]. Wang *et al.* proposed a deep deterministic policy gradient (DDPG) algorithm to jointly optimize the power allocation and flight trajectory of unmanned aerial vehicle (UAV) with constrained effective energy to maximize the throughput to ground users [13]. Then, Ma *et al.* used the proximal policy optimization (PPO) algorithm to allocate frequency resource to all beams at once, which met the real-time communication [14]. However, the above algorithms inevitably face complex CCI and huge action space which increases exponentially with the number of beams and channels.

In this paper, we aim to apply reinforcement learning to a larger scale frequency resource allocation problem, so we use the PPO algorithm as the basic framework to extend frequency resource allocation to specific user terminals, which implies a larger action space, and call this algorithm PPO-RA. Furthermore, we propose an improved PPO-RA-LOOP algorithm by using LSTM network structure to allocate frequency resource recursively. Compared with PPO-RA, the PPO-RA-LOOP algorithm significantly reduces the action space thus can effectively handle the complex frequency resource allocation problem faced by MBS. The main contributions of this paper are summarized as follows:

1) Proposing a PPO-based frequency resource allocation algorithm. Different from the way to output the frequency resource allocation schemes of all beams at once, the PPO-RA-LOOP algorithm regards the frequency resource allocation scheme of one beam as a sub-action and then recursively outputs these sub-actions, which significantly reduces the action space.

2) Adopting the LSTM network to predict the next sub-action. Specifically, the previously generated sub-actions, as a prior knowledge, are fed to the LSTM layer to predict the next sub-action, which is beneficial for making better action decisions.

3) Compared with the PPO-RA algorithm, the proposed PPO-RA-LOOP algorithm decreases the exploration cost, accelerates the convergence speed and reduces the complexity of neural network thus achieving better system performance.

II. System Model and Problem Formulation

As shown in Figure 1, the GEO satellite system, forms a coverage area on the Earth surface with N beams, which is represented as $\mathbb{B} = \{n|n = 1, 2, \ldots, N\}$. We assume that there are K users under the satellite coverage area, which are defined as $\mathbb{U} = \{k|k = 1, 2, \ldots, K\}$. The total of system available frequency is denoted as B_{total}. B_{total} is divided into M sub-channels and the available sub-channel set is expressed as $\mathbb{C} = \{m|m = 1, 2, \ldots, M\}$. Thus, the frequency of sub-channel can be calculated by $B_{\text{ch}} = B_{\text{total}}/M$. Similarly, we divide the total transmission power P_{total} equally to each beam and the antenna transmission power of beam n is denoted as P_n. According to [15], the transmission gain of beam n to user k is denoted as $G_{n,k}$ and its maximum value is denoted as $G_n^{\max}$. Then, the channel gain $H_{n,k}$ for user k under beam n can be calculated by

$$H_{n,k} = G_{n,k} G_{\text{r}} L_{\text{path}} \tag{1}$$

where G_{r} represents the receive gain for all user terminals. L_{path} is free space loss of the user k, which is represented as

$$L_{\text{path}} = 4\pi f d_{n,k}/c \tag{2}$$

where f is the frequency and c is the speed of light. $d_{n,k}$ represents the distance between beam n and user k, which is defined as

$$d_{n,k} = \left((x - x_0)^2 + (y - y_0)^2 + (z - z_0)^2\right)^{1/2} \tag{3}$$

where (x_0, y_0, z_0) and (x, y, z) represent the center coordinates of beam n and the coordinates of user k, respectively.

In addition, the bandwidth occupancy of user termi-

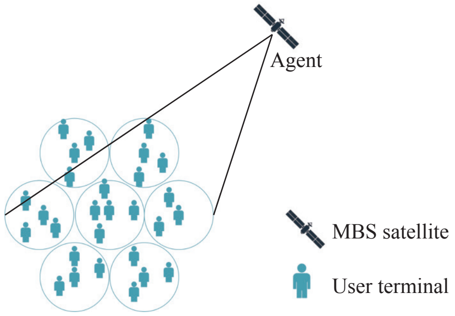


Figure 1 The MBS satellite system.

nal k under beam n at time slot t can be expressed as

$$\boldsymbol{X}_{n,k}^t = [x_{n,k,1}^t, x_{n,k,2}^t, \ldots, x_{n,k,M}^t] \tag{4}$$

$x_{n,k,m}^t = 1$ indicates that sub-channel m is assigned to user terminal k under beam n and 0 means that subchannel m is not allocated to user terminal k. Further, we denote the occupancy of subchannel m under beam n by $b_{n,m}^t$, which is defined as

$$b_{n,m}^t = \sum_{k \in \mathbb{S}_n} x_{n,k,m}^t \tag{5}$$

where $\mathbb{S}_n$ is the set of users served by the beam n. Subchannel m can only be allocated to at most one user at a time slot. Thus, there is $b_{n,k}^t \leq 1$. Based on the above definitions, the signal to interference plus noise ratio (SINR) for user k over sub-channel m under beam n can be calculated by

$$\mathrm{SINR}_{n,k,m}^t = \frac{P_n H_{n,k} x_{n,k,m}^t}{N_0 + \sum_{n' \neq n} P_{n'} H_{n',k} b_{n',m}^t} \tag{6}$$

N_0 is additive white Gaussian noise (AWGN) power. At time slot t, according to Shannon's formula, the throughput of user k can be obtained by

$$T_{n,k}^t = B_{\mathrm{ch}} \sum_{m=1}^{M} \log_2 (1 + \mathrm{SINR}_{n,k,m}^t) \tag{7}$$

To facilitate measuring system satisfaction, the offered capacity and requested capacity of beam n are defined as

$$C_{\mathrm{offer}}^n = \sum_{k \in \mathbb{S}_n} T_{n,k}^t \tag{8a}$$

$$C_{\mathrm{req}}^n = \sum_{k \in \mathbb{S}_n} D_{n,k}^t \tag{8b}$$

where $D_{n,k}^t$ is the traffic demand of user k under beam n waiting for transmission in time slot t, and its unit is bits/s. Further, we define the maximum value of the offered capacity of beam n as $C_{\max}^n$, which can be achieved under the condition that the CCI is zero.

Based on (8), the satisfaction of the beam n is defined as

$$\mathrm{SI}_n = \begin{cases} C_{\mathrm{offer}}^n / C_{\mathrm{req}}^n, & C_{\mathrm{req}}^n \leq C_{\max}^n \text{ and } C_{\mathrm{offer}}^n \leq C_{\mathrm{req}}^n \\ C_{\mathrm{offer}}^n / C_{\max}^n, & C_{\mathrm{req}}^n > C_{\max}^n \text{ and } C_{\mathrm{offer}}^n \leq C_{\mathrm{req}}^n \\ 1, & \text{otherwise} \end{cases} \tag{9}$$

Our goal is to maximize the average system satisfaction, so the frequency resource allocation problem can be modeled as

$$\begin{aligned} &\max_{\boldsymbol{X}} \quad F = \frac{1}{N} \sum_{n=1}^{N} \mathrm{SI}_n \\ &\text{s.t. } \mathrm{C1}: B_{\mathrm{ch}} \sum_{m=1}^{M} b_{n,m}^t \leq B_{\mathrm{total}}, \\ &\qquad \mathrm{C2}: x_{n,k,m}^t = 0, \text{if } D_{n,k}^t = 0, \forall n \in \mathbb{B}, \forall k \in \mathbb{U}, \forall m \in \mathbb{C} \end{aligned} \tag{10}$$

For the optimization problem F in (10), $\boldsymbol{X}$ represents the frequency resource configuration of the MBS system. Constraint C1 based on (5) means the frequency resource allocated by each beam cannot exceed the total frequency. Constraint C2 indicates that sub-channel m can only be allocated to the user k with traffic request under beam n.

III. A DRL-Based Frequency Resource Allocation Solution for MBS

In this section, we first briefly introduce the defined state, action, and reward in the MBS system. Then, we enumerate the issues faced by DRL-based algorithm in the problem of frequency resource allocation. To address the above issues, we propose the PPO-RA-LOOP algorithm.

1. State, action and reward

In the MBS system, the satellite is regarded as an agent, and the traffic demand of user terminals constitutes the state. The state of all beams at time slot t is defined as

$$\boldsymbol{s}_t = (D_{1,1}^t, D_{1,2}^t, \ldots, D_{N,K}^t) \tag{11}$$

$D_{n,k}^t$ is explained in (8). Then, the agent allocates frequency resource according to states. The frequency resource allocation schemes of all beams constitute the action of the agent, defined as

$$\boldsymbol{a}_t = (\boldsymbol{a}_1^t, \boldsymbol{a}_2^t, \ldots, \boldsymbol{a}_N^t) \tag{12}$$

$\boldsymbol{a}_n^t$ is the frequency resource allocation scheme of beam n, which can be obtained by

$$\boldsymbol{a}_n^t = (\mathrm{id}_{n,1}^t, \mathrm{id}_{n,2}^t, \ldots, \mathrm{id}_{n,M}^t) \tag{13}$$

$\mathrm{id}_{n,m}^t \in \{0, 1, \ldots, K\}$, which is used to indicates the specific user terminal number assigned to channel m under beam n. $\mathrm{id}_{n,m}^t = 0$ means the channel m is not assigned.

After executing the action, the agent gets corresponding reward at each time slot. The reward is the same as the optimization objective in (10), which is represented as

$$r_t = \frac{1}{N} \sum_{n=1}^{N} \mathrm{SI}_n \tag{14}$$

SI_n is explained in (9). The goal of agent is to maximize the cumulative rewards.

2. PPO-RA algorithm

As a classic DRL method, the PPO algorithm is able to deal with both discrete and continuous action spaces, which consists of an actor network for generating action and a critic network for evaluating state. The optimization objective of the actor network is written as

$$\theta_{k+1} = \arg\max_{\theta} \frac{1}{T}\sum_{t=1}^{T} L(\boldsymbol{s}_t, \boldsymbol{a}_t, \theta, \theta_k) \tag{15}$$

and $L(\boldsymbol{s}_t, \boldsymbol{a}_t, \theta, \theta_k)$ is defined as

$$\begin{aligned} L(\boldsymbol{s}_t, \boldsymbol{a}_t, \theta, \theta_k) = \min\Big(&\rho_t(\theta, \theta_k) A^{\pi_{\theta_k}}(\boldsymbol{s}_t, \boldsymbol{a}_t), \\ &\text{clip}(\rho_t(\theta, \theta_k), 1-\zeta, 1+\zeta) A^{\pi_{\theta_k}}(\boldsymbol{s}_t, \boldsymbol{a}_t)\Big) \end{aligned} \tag{16}$$

where $\rho_t(\theta, \theta_k) = \pi_\theta(\boldsymbol{a}_t|\boldsymbol{s}_t)/\pi_{\theta_k}(\boldsymbol{a}_t|\boldsymbol{s}_t)$ represents the probability ratio of new and old policies. $A^{\pi_{\theta_k}}(\boldsymbol{s}_t, \boldsymbol{a}_t)$ is the advantage function of policy θ_k. ζ is a (fractional value) hyperparameter. The function clip limits the updated magnitude of the new policy by removing incentives for the policy to change dramatically.

The critic network is used to judge the action generated by actor network through estimating the value $V_\beta(s_t)$. Then, the critic network updates its parameters β by the following equations

$$\begin{aligned} \beta_{k+1} &= \arg\min_{\beta} \frac{1}{T}\sum_{t=1}^{T} (V_\beta(s_t) - \hat{R}_t)^2 \\ \hat{R}_t &= \sum_{t'>t}^{T} \gamma^{t'-t} r_{t'} \end{aligned} \tag{17}$$

where γ is the discount factor. Based on the state, action and reward defined in (11)–(14), utilizing the framework of the PPO algorithm, we derive a frequency resource allocation algorithm, named PPO-RA, which further extends the DBA-PPO algorithm [14] by allocating frequency resource to multiple requesting users.

However, in large-scale satellite systems, with the increment of the number of sub-channels, users and beams, the dimension of the action space will increase dramatically, which is a hard challenge faced by the PPO-RA algorithm. Specifically, the satellite has a total of N beams, and each beam can use all M channels, so there are MN channels available. For each available channel, it can be assigned to one of K users. The size of action space is defined as the total number of all possible frequency resource allocation schemes, so the total action space is K^{MN}, which increases exponentially with the number of beams and sub-channels.

The high-dimensional action space will lead to two problems. Firstly, the high-dimensional action space requires more complex networks to fit and results in high exploration costs. Secondly, the high exploration cost of the optimal policy leads to the slow convergence of the PPO-RA algorithm. To address the above problems, we propose the PPO-RA-LOOP algorithm.

3. Proposed PPO-RA-LOOP algorithm

As shown in Figure 2, different from the way to output the frequency resource allocation schemes of all beams at once, PPO-RA-LOOP algorithm (Algorithm 1) adopts a recursive network structure which splits the action into N sub-actions to recursively output the frequency resource allocation schemes for each beam one by one. Then, the total frequency resource allocation schemes of all beams $\boldsymbol{a}_t$ are obtained after N iterations. Thus, the action space is reduced from K^{MN} to K^M, which avoids the explosive growth of the action space. Correspondingly, the neurons in the output layer of actor network are reduced from KMN to KM, which also reduces the number of parameters of the actor network. This is one reason why we use the recursive structure.

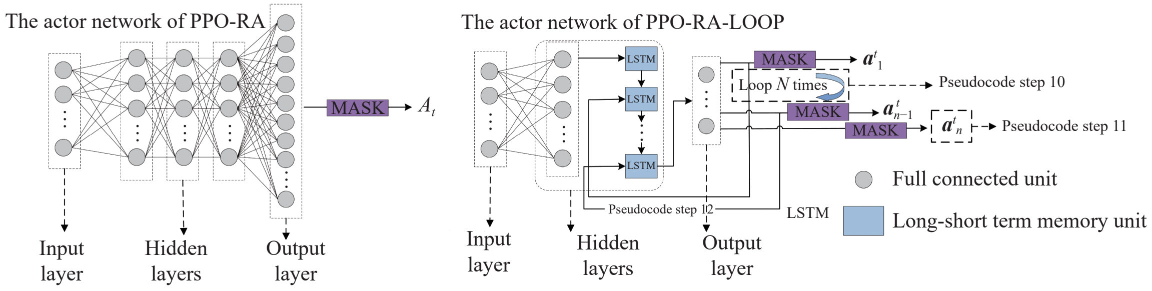


Figure 2 Actor network for two algorithms.

In addition, to reduce CCI in MBS system, we feed back the sub-action generated by each recursion, as a prior knowledge that may reflect CCI in a certain extent, to the input side of the actor network. These sub-actions obtained by recursions are sequentially cascaded and fed back to the input side of actor network, which can be considered as sequential information. Compared with the fully connected neural network, the LSTM network is more efficient in dealing with sequential information. Thus, we replace the hidden layer of the actor network with LSTM. Then, the actor network predicts the frequency resource allocation for the next beam based on the current traffic demand and the sequential action information. Specifically, at time slot t, the frequency allocation scheme $\boldsymbol{a}_n^t$ of beam n can be calculated by

$$\boldsymbol{a}_n^t = \mathrm{LSTM}(\boldsymbol{s}_t^{n-1}) \tag{18}$$

LSTM represents the fitting function composed of LSTM network in the actor network. $\boldsymbol{s}_t^{n-1}$ is the input state information of the LSTM network after $n-1$ recursions at time slot t. This recursive actor network structure fits well with LSTM, which is another advantage for using the recursive network structure.

Algorithm 1 PPO-RA-LOOP

```
Initialize satellite environment;
Initialize policy parameters $\theta$;
Initialize value function parameters $\beta$;
Reset gradients: $\mathrm{d}\theta \leftarrow 0$ and $\mathrm{d}\beta \leftarrow 0$;
for epoch $= 1, 2, \dots, \mathrm{epoch}_{\max}$ do
    Reset Replay Buffer;
    for $t = 1, 2, \dots, \mathrm{step}_{\max}$ do
        Observe the initial state $\boldsymbol{s}_t^0$;
        Input initial $\boldsymbol{s}_t^0$ into the actor network of PPO-RA-LOOP;
        for $n = 1, 2, \dots, N$ do
            Get the $n$th beam sub-action, i.e., $\boldsymbol{a}_n^t = \mathrm{LSTM}(\boldsymbol{s}_t^{n-1})$;
            Add sub-action $\boldsymbol{a}_n^t$ to the current state $\boldsymbol{s}_t^{n-1}$ as prior knowledge and update the next state information as $\boldsymbol{s}_t^n = \{\boldsymbol{s}_t^0, \boldsymbol{a}_1^t, \boldsymbol{a}_2^t, \dots, \boldsymbol{a}_n^t\}$;
        end for
        Based on (18) and current policy $\pi_\theta(\boldsymbol{a}_t|\boldsymbol{s}_t)$, actor network gets the resource allocation schemes $\boldsymbol{a}_t$ of all beams after $N$ iterations;
        Critic network estimates the value of the current state $\mathrm{V}_\beta(\boldsymbol{s}_t)$;
        The agent executes action $\boldsymbol{a}_t$ in satellite environment and gets reward $r_t$ according to (14). Then, agent gets next state $\boldsymbol{s}_{t+1}$;
        Store $\{\boldsymbol{s}_t, \boldsymbol{a}_t, r_t, \boldsymbol{s}_{t+1}\}$ in Replay Buffer;
        if $t$=Buffer size then
            Update policy parameters $\theta$ according to (15);
            Update value function parameters $\beta$ according to (17);
        end if
    end for
end for
```

To improve the exploration efficiency of the agent further, we mask illegal actions (agent allocates sub-channels to users with no traffic requests) through the mask matrix before generating sub-actions, as shown in the MASK module of Figure 2. After the first recursion, the LSTM layer outputs the action $\boldsymbol{a}_1^t = \mathrm{LSTM}(\boldsymbol{s}_t^0)$ of beam 1, and changes $\boldsymbol{s}_t^0$ to $\boldsymbol{s}_t^1$. $\boldsymbol{s}_t^1$ can be represented as

$$\begin{aligned}\boldsymbol{s}_t^1 &= (D_{1,1}^t, D_{1,2}^t, \dots, D_{N,K}^t, \boldsymbol{a}_1^t, \mathbf{0}, \mathbf{0}, \dots, \mathbf{0})\\ \boldsymbol{s}_t^0 &= (D_{1,1}^t, D_{1,2}^t, \dots, D_{N,K}^t, \mathbf{0}, \mathbf{0}, \mathbf{0}, \dots, \mathbf{0})\end{aligned} \tag{19}$$

The input state information of the LSTM network after the $(N-1)$th recursion is written as

$$\boldsymbol{s}_t^{N-1} = (D_{1,1}^t, D_{1,2}^t, \dots, D_{N,K}^t, \boldsymbol{a}_1^t, \boldsymbol{a}_2^t, \dots, \boldsymbol{a}_{N-2}^t, \boldsymbol{a}_{N-1}^t) \tag{20}$$

After N recursions, the action $\boldsymbol{a}_N^t = \mathrm{LSTM}(\boldsymbol{s}_t^{N-1})$ of beam N at time slot t is obtained. At this time, we get the action of all beams, which is explained in (12).

Simulation result shows that the PPO-RA-LOOP algorithm using LSTM as the hidden layer has better performance than that using the fully connected layer as the hidden layer, as shown in Figure 3. Therefore, the PPO-RA-LOOP algorithm adopts LSTM as the hidden layer.

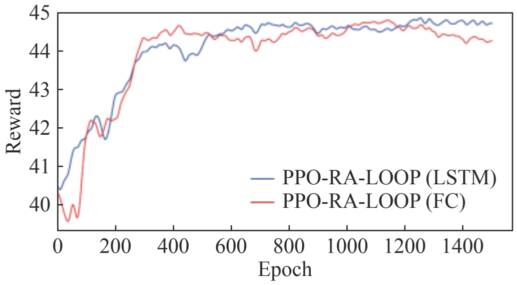


Figure 3 PPO-RA-LOOP (LSTM) indicates that the actor network uses LSTM as the hidden layer with a parameter amount of 1079156, and PPO-RA-LOOP (FC) suggests that the hidden layer is the fully connected layer with a parameter amount of 2860220. The vertical axis represents the cumulative rewards, and the horizontal axis represents the number of iterations.

IV. Simulation Results

In this section, we introduce the traffic model of satellite users and evaluate the performance of the proposed PPO-RA-LOOP algorithm.

1. Simulation settings

There are 324 users in the 27-beam satellite system. The total frequency is equally divided into 12 sub-channels and each beam can use all sub-channels. The main simulation parameters are given in Table 1. The traffic arrivals of each user obey a Poisson distribution with mean λ and the traffic duration follows an exponential distribution with mean μ. Thus, we define the traffic intensity of users as

$$\rho = \lambda/\mu \tag{21}$$

By setting different traffic intensity of users in (21), we obtain two traffic distributions: uniform traffic distribution and non-uniform traffic distribution, as shown in Figures 4(a) and (b).

Table 1 Simulation parameters

Parameter	Value
Number of beams (N)	27
Number of channel (M)	12
Number of user (K)	324
Total transmission power of satellite (P_{total})	2700 W
Bandwidth of each beam (B_{total})	500 MHz
Channel bandwidth (B_{ch})	41.66 MHz
Satellite antenna efficiency (η)	0.5
Maximum transmission antenna gain ($G_n^{\max}$)	50 dBi
Receiver antenna gain (G_{r})	40 dBi
GEO satellite height	35786 km

2. Comparison of algorithm performance

To illustrate the advantages of our proposed algorithm, we compared the following three algorithms:

a) SA: The exponential annealing method, a type of simulated annealing (SA), is employed to calculate the suboptimal value of the fusion control problem of transmit power and beam directivity [16].

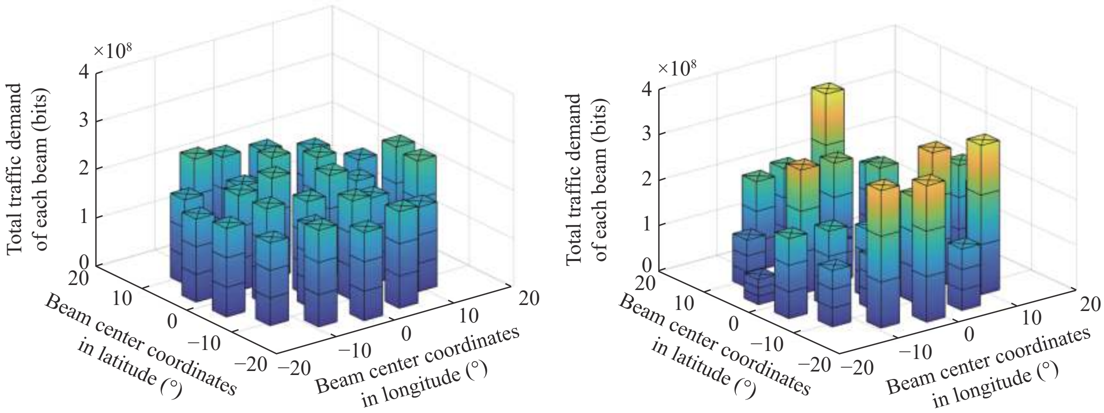
b) PPO-RA: A PPO-based frequency resource allocation algorithm for users in MBS system. There are two PPO-RA algorithms for this: PPO-RA-3 and PPO-RA-5. The numbers indicate that the number of hidden layers of the actor network is 3 and 5, respectively.

c) PPO-RA-LOOP: The proposed algorithm with a recursive LSTM network structure.

We adopt the complexity of neural network, convergence, average system satisfaction and spectral efficiency as indicators to evaluate the above algorithms.

The complexity of neural network can be measured with the number of its parameters. Table 2 summarizes the number of parameters (actor network) in the proposed PPO-RA-LOOP algorithm and the PPO-RA algorithm with different hidden layers settings. It can be seen that the number of parameters of the PPO-RA-LOOP algorithm is reduced by an order of magnitude compared with PPO-RA. To further illustrate the computational complexity of neural network, we used thop toolkit to get the accurate floating point operations per second (FLOPS) result in Table 2. The results show that the computational complexity of the neural network of the PPO-RA-LOOP algorithm is higher than that of the PPO-RA algorithm, which is mainly due to the recursive output sub-actions of the PPO-RA-LOOP algorithm, and the computation increases with the number of recursions N. The specific computational complexity analysis is given in Appendix A.

The convergence of the proposed PPO-RA-LOOP algorithm and the PPO-RA-5 algorithm is illustrated in Figure 5. The proposed PPO-RA-LOOP algorithm achieves the highest rewards with faster convergence. This is because the actor network structure of PPO-RA-LOOP significantly reduces the action space. That is to say, the algorithm performance is improved by rapidly exploring more optimal strategies in a smaller action space. Additionally, the results show that PPO-RA-5 with more hidden layers achieves higher cumulative rewards than PPO-RA-3 in both uniform and non-uniform traffic distributions. This is because an appropriate increase in hidden layers has a better fitting effect and facilitates action decisions. Therefore, in the following algorithm



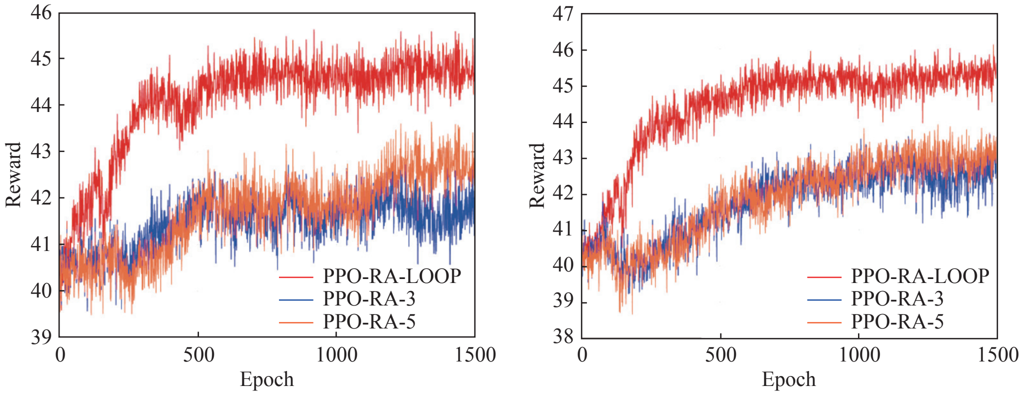
(a) Uniform traffic distribution

(b) Non-uniform traffic distribution

Figure 4 Two traffic distributions. The z axis represents the total traffic demand of each beam, which is explained in equation (8b), and its unit is bits. The x and y axes represent the beam center coordinates in latitude and longitude.

Table 2 The comparison of complexity of action network

Algorithm	Hidden layer	Actor network parameters	FLOPS
PPO-RA-LOOP	FC (56) +LSTM (56,56)	1.079156×10^6	9.19296×10^7
PPO-RA-3	FC (128,256,128)	1.5433044×10^7	3.0654464×10^7
PPO-RA-5	FC (128,256,512,256,128)	1.5695956×10^7	3.1178752×10^7



(a) Uniform traffic distribution

(b) Non-uniform traffic distribution

Figure 5 The comparison of algorithm convergence.

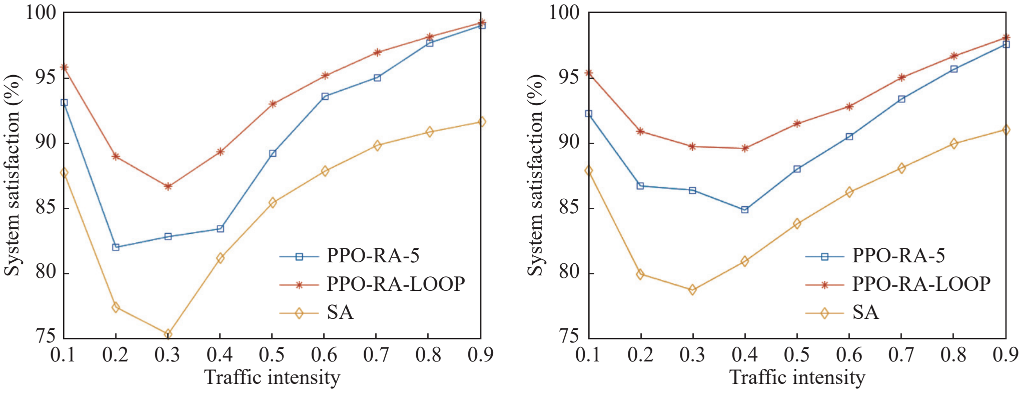
comparisons in terms of system satisfaction and spectral efficiency, we only recorded the algorithm performance of PPO-RA-5. Table 3 gives the running time per epoch of the two algorithms and the number of epochs required for convergence. The PPO-RA-LOOP algorithm recursively outputs sub-actions, so the time per epoch is slightly higher than that of the PPO-RA-5 algorithm. However, the number of epochs required for convergence is significantly less than that of the PPO-RA algorithm. Overall, the total convergence time of PPO-RA-LOOP is smaller than that of PPO-RA-5, so we consider that PPO-RA-LOOP is less time costly.

Table 3 The comparison of time cost of algorithms

ρ	PPO-RA-LOOP			PPO-RA-5		
	Time per epoch (s)	Convergence epochs (approx)	Total time (s)	Time per epoch (s)	Convergence epochs (approx)	Total time (s)
0.1	6.98	600	4188	6.20	700	4340
0.2	12.92	650	8398	12.14	1000	12140
0.3	21.59	400	8636	21.22	900	19098
0.4	37.00	400	14800	34.70	800	27760
0.5	48.58	650	31577	48.39	1200	58068
0.6	62.93	600	37758	60.50	900	54450
0.7	70.40	400	28160	68.76	700	48132
0.8	79.11	500	39555	77.55	800	62064
0.9	84.80	500	42400	83.41	800	66728

Figure 6 shows the average system satisfaction of the three algorithms, which has been defined in as F in (10). When traffic intensity is light ($\rho \in [0.1, 0.3]$), the beam requested capacity C^n_{req} is lower than the upper limit of the beam capacity $C^n_{\max}$. The beam satisfaction, according to (9), is calculated by $\text{SI}_n = C^n_{\text{offer}}/C^n_{\text{req}}$. As traffic intensity increases, the growth rate of the beam offered capacity C^n_{offer} is lower than that of the beam requested capacity C^n_{req}. As a result, the beam satisfaction decreases and so does the average system satisfaction. When traffic intensity is heavy ($\rho \in [0.4, 0.9]$), the beam requested capacity is higher than the upper limit of the beam capacity. At this time, the beam satisfaction is calculated by $\text{SI}_n = C^n_{\text{offer}}/C^n_{\max}$. With the increase of beam offer capacity, the beam satisfaction gradually increases and system satisfaction rises as well. In addition, the PPO-RA-LOOP algorithm outperforms the PPO-RA algorithm by 2.7%–6.9% in uniform traffic distribution and 3.1%–4.7% in non-uniform traffic distribution. In the case of heavy traffic intensity for the two distributions, the system satisfaction achieved by the PPO-RA-LOOP algorithm is slightly higher than that of the PPO-RA algorithm. Then, we compare two DRL-based algorithms with the traditional SA scheduling method. Compared with the SA, the PPO-RA and the PPO-RA-LOOP algorithms have obvious advantages in system satisfaction. This is because the DRL-based algorithms have the ability to efficiently adjust the scheduling scheme according to different distributions of traffic requests.

Table 4 shows the spectral efficiency of the three algorithms. With the advantages of efficient scheduling for available subchannels, the PPO-RA-LOOP algorithm achieves a significant advantage in spectral efficiency compared to the SA algorithm. In the case of uniform traffic distribution, the PPO-RA-LOOP algorithm improves the spectral efficiency by 0.6%–4% at light traffic intensity and 7.9%–8.9% at heavy traffic intensity. Un-



(a) Uniform traffic distribution

(b) Non-uniform traffic distribution

Figure 6 The comparison of system satisfaction.

der non-uniform traffic distribution, the spectral efficiency of the PPO-RA-LOOP algorithm is improved by 2.7%–5.9% at light traffic intensity and 5.7%–8.9% at heavy traffic intensity. In addition, under uniform traffic distribution, compared with the PPO-RA algorithm, PPO-RA-LOOP improves the spectral efficiency by 0.17%–1.49% at light traffic intensity, and 0.26%–5.91% at heavy traffic intensity. From Table 4, it can be seen that under non-uniform traffic distribution, the PPO-RA-LOOP improves 0.36%–1.29% at light traffic intensity and 0.49%–4.26% at heavy traffic intensity. This is because that PPO-RA-LOOP can better avoid CCI through recursive actor network structure with LSTM as hidden layer. In a word, the PPO-RA-LOOP algorithm achieves better performance with lower network complexity compared to PPO-RA.

Table 4 Spectral efficiency of the three algorithms

ρ	Spectral efficiency of algos. in uniform distrib. (bps/Hz)			Spectral efficiency of algos. in non-uniform distrib. (bps/Hz)		
	SA	PPO-RA-5	PPO-RA-LOOP	SA	PPO-RA-5	PPO-RA-LOOP
0.1	0.6298	0.6326	0.6337	0.6262	0.6407	0.6430
0.2	1.2756	1.2867	1.3059	1.2710	1.2789	1.2926
0.3	1.8548	1.9115	1.9281	1.7028	1.7796	1.8026
0.4	2.1451	2.1896	2.3189	2.0104	2.0511	2.1259
0.5	2.2898	2.3974	2.4944	2.1737	2.2700	2.3667
0.6	2.3715	2.5300	2.5744	2.2883	2.4033	2.4659
0.7	2.4425	2.5900	2.6356	2.3841	2.5211	2.5607
0.8	2.4779	2.6667	2.6737	2.4479	2.6007	2.6311
0.9	2.5011	2.7019	2.7089	2.4833	2.6626	2.6756

V. Conclusion

This paper extends the frequency resource allocation scheme for satellite user terminals in the MBS system. On the basis of PPO-RA, we propose a PPO-RA-LOOP algorithm by adopting a recursive neural network structure with LSTM. With the aid of LSTM layers, the proposed algorithm is able to reduce the action space and the complexity of neural network significantly compared with PPO-RA, thus speeding up the convergence of the algorithm. Simulation results demonstrate the applicability of DRL to the resource allocation problem.

Acknowledgements

This work was supported by the National Natural Science Foundation of China (Grant No. 62071354), the Key Research and Development Program of Shaanxi (Grant No. 2022ZDLGY05-08), and the ISN State Key Laboratory.

Appendix A

1. Parameter setting of algorithms

This paper compares three DRL-based algorithms, namely PPO-RA-LOOP, PPO-RA-3, and PPO-RA-5. The hidden layer parameters of the actor network used by these algorithms are given in Table 2. The remaining parameters are shown in Table A-1.

2. Computational complexity analysis of PPO-RA-LOOP

Considering that resource allocation is a sequential decision-making problem, for convenience, we only analyze the complexity of different algorithms when making a decision at one time slot. As stated by Hu *et al.* [17], we assume that multiplication or division between two d-bits needs d^2 basic operations and addition or subtraction needs d basic operations.

The actor network of the PPO algorithm is composed of a fully

Table A-1 Parameters of DRL-based algorithms

Parameter	Value
Training epochs	1500
Steps per episode	50
Buffer size	100
Discount factor	0.99
Clip ratio	0.2
Target KL	0.015
Activation function	tanh
Training iterations per epoch	80
Learning rate of actor	0.0003
Learning rate of critic	0.0001
Optimizer	Adam

connected network, so its computational complexity is equal to $\sum_{e=1}^{E} d^2 h_{\text{in}}^e h_{\text{out}}^e$ where h_{in}^e denotes the input size of the eth layer and h_{out}^e denotes the output size of the eth layer [17].

The proposed algorithm uses LSTM network as the hidden layer. As shown in Figure A-1, the LSTM network includes four nonlinear transformation modules, each of which can be roughly considered as a fully connected network. In addition, to simplify the analysis, we omit the computational complexity of the activation function.

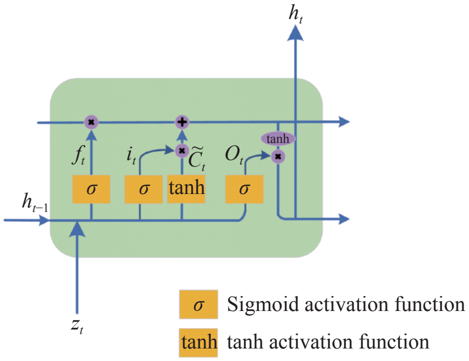


Figure A-1 The network structure of LSTM.

The computational complexity of the LSTM layer of the proposed algorithm can be expressed as

$$\sum_{i=1}^{I} 4Nd^2(h_{t-1}^i + z_t^i)h_t^i \tag{A-1}$$

where 4 denotes the four nonlinear transformation modules (roughly considered as fully connected layers), N denotes the number of algorithm recursions. I represents the number of layers of the LSTM; h_{t-1}^i and z_t^i are shown in Figure A-1, which denote the hidden state dimension and word vector dimension of the ith layer LSTM, respectively.

The total computational complexity of the proposed algorithm can be expressed approximately as

$$O\left(\sum_{i=1}^{I} 4Nd^2(h_{t-1}^i + z_t^i)h_t^i + \sum_{e=1}^{E} d^2 h_{\text{in}}^e h_{\text{out}}^e\right) \tag{A-2}$$

References

[1] F. Fourati and M. S. Alouini, “Artificial intelligence for satellite communication: A review,” *Intelligent and Converged Networks*, vol. 2, no. 3, pp. 213–243, 2021.

[2] H. Yang, J. H. Dang, Y. H. Pan, *et al.*, “A digital channelizer design approach for broadband satellite communications based on frequency domain filter theory,” in *Proceedings 2013 International Conference on Mechatronic Sciences, Electric Engineering and Computer (MEC)*, Shenyang, China, pp. 2986–2990, 2013.

[3] V. K. Singh, W. G. Ho, and R. Gharpurey, “A frequency-folded ADC channelizer with digital equalization and relaxed anti-alias filtering,” *IEEE Transactions on Circuits and Systems I: Regular Papers*, vol. 65, no. 7, pp. 2304–2317, 2018.

[4] K. Kaneko, H. Nishiyama, N. Kato, *et al.*, “An evaluation of flexible frequency utilization in high throughput satellite communication systems with digital channelizer,” in *IEEE International Conference on Communications (ICC)*, Paris, France, pp. 1–6, 2017.

[5] L. Del Consuelo Hernandez Ruiz Gaytan, Z. N. Pan, J. Liu, *et al.*, “Dynamic scheduling for high throughput satellites employing priority code scheme,” *IEEE Access*, vol. 3, pp. 2044–2054, 2015.

[6] A. I. Aravanis, B. M. R. Shankar, P. D. Arapoglou, *et al.*, “Power allocation in multibeam satellite systems: A two-stage multi-objective optimization,” *IEEE Transactions on Wireless Communications*, vol. 14, no. 6, pp. 3171–3182, 2015.

[7] X. Zhang, J. J. Wang, C. X. Jiang, *et al.*, “Robust beamforming for multibeam satellite communication in the face of phase perturbations,” *IEEE Transactions on Vehicular Technology*, vol. 68, no. 3, pp. 3043–3047, 2019.

[8] H. M. Zhang, C. X. Jiang, J. J. Wang, *et al.*, “Multicast beamforming optimization in cloud-based heterogeneous terrestrial and satellite networks,” *IEEE Transactions on Vehicular Technology*, vol. 69, no. 2, pp. 1766–1776, 2020.

[9] G. Cocco, T. De Cola, M. Angelone, *et al.*, “Radio resource management optimization of flexible satellite payloads for DVB-S2 systems,” *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 266–280, 2018.

[10] Y. Kawamoto, T. Kamei, M. Takahashi, *et al.*, “Flexible resource allocation with inter-beam interference in satellite communication systems with a digital channelizer,” *IEEE Transactions on Wireless Communications*, vol. 19, no. 5, pp. 2934–2945, 2020.

[11] F. G. Ortiz-Gomez, L. Lei, E. Lagunas, *et al.*, “Machine learning for radio resource management in multibeam GEO satellite systems,” *Electronics*, vol. 11, no. 7, article no. 992, 2022.

[12] J. H. Liu, B. K. Zhao, Q. Xin, *et al.*, “Dynamic channel allocation for satellite internet of things via deep reinforcement learning,” in *2020 International Conference on Information Networking (ICOIN)*, Barcelona, Spain, pp. 465–470, 2020.

[13] C. W. Wang, Y. L. Cui, D, H, Deng, *et al.*, “Trajectory optimization and power allocation scheme based on DRL in energy efficient UAV-aided communication networks,” *Chinese Journal of Electronics*, vol. 31, no. 3, pp. 397–407, 2022.

[14] S. J. Ma, X. Hu, X. L. Liao, *et al.*, “Deep reinforcement learning for dynamic bandwidth allocation in multi-beam satellite systems,” in *2021 IEEE 6th International Conference on Computer and Communication Systems (ICCCS)*, Chengdu, China, pp. 955–959, 2021.

[15] G. Maral and M. Bousquet, *Satellite Communications Systems: Systems, Techniques and Technology*, 6th ed., John Wiley & Sons, Sussex, UK, 2020.

[16] M. Takahashi, Y. Kawamoto, N. Kato, *et al.*, “DBF-based fusion control of transmit power and beam directivity for flexible resource allocation in HTS communication system toward B5G,” *IEEE Transactions on Wireless Communications*, vol. 21, no. 1, pp. 95–105, 2022.

[17] X. Hu, X. L. Liao, Z. J. Liu, *et al.*, “Multi-agent deep reinforcement learning-based flexible satellite payload for mobile terminals,” *IEEE Transactions on Vehicular Technology*, vol. 69, no. 9, pp. 9849–9865, 2020.

Haowei MENG received the B.S. degree in communication engineering from Zhengzhou University, Zhengzhou, China, 2020. He is currently working towards the M.S. degree at Xidian University, Xi'an, China. His research interests include wireless resource management and reinforcement learning.
(Email: 20011210418@stu.xidian.edu.cn)

Ning XIN received the Ph.D. degree from CAS in 2014, M.S. degree from Naval Aeronautical Engineering Institute in 2007, and B.S. degree from Yantai University in 2004. He is a Researcher with the Institute of Telecommunication Satellite, China Academy of Space Technology. His research interests are spacecraft design and satellite payload design.
(Email: xinning7@sina.com)

Hao QIN received the B.S., M.S., and Ph.D. degrees in communication and information systems from Xidian University in 1996, 1999, and 2004, respectively. He is an Associate Professor with the School of Telecommunications Engineering, Xidian University, Xi'an, China. His research interests include wireless communications and satellite communications.
(Email: hqin@mail.xidian.edu.cn)

Di ZHAO received the B.S. degree in communication engineering from Shandong Normal University, Jinan, China, 2018. She is currently working towards the Ph.D. degree with Xidian University, Xi'an, China. Her research interests include wireless resource management, satellite communications and reinforcement learning in wireless networks.
(Email: dzhao_1@stu.xidian.edu.cn)