Dynamic Resource Allocation for Multi-beam Satellite Communication Systems

Siya Zhang(✉), Rong Chai, Lei Liu, and Guorong Yang

School of Communications and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, People's Republic of China
s220132218@stu.cqupt.edu.cn

Abstract. Dynamic resource allocation is a crucial concern for achieving flexible and efficient data transmission in satellite communication systems. This paper examines the dynamic resource allocation challenge within a satellite communication system with multiple beams. Considering the difference between user requirement and service providing capability, we establish the concept of system cost and cast the integrated problem of beam illumination, sub-channel, and power allocation as a minimization task for the average system cost. To address the defined issue, we first propose two beam scheduling schemes, then we frame the issue of sub-channel and power allocation as a Markov decision process. We introduce two algorithms, namely, a Double Deep Q Learning (DDQN)-based algorithm and an Improved Priority Experience Replay DDQN (PER-DDQN)-based algorithm, for the joint power and sub-channel allocation. The simulation outcomes illustrate the efficacy and superiority of the proposed algorithms.

Keywords: Beaming illumination · Multi-beam satellite · Resource allocation · Double deep Q learning

1 Introduction

Satellite communication technology is considered a pivotal element in future 6G communications. Satellite communication systems equipped with a significant quantity of beams and a versatile on-board payload, emerge as an effective solution to accommodate the growing user traffic [1]. The dynamic characteristics of satellite systems and the conflict between increasing user requirements and limited service providing capability pose challenges and difficulties to the allocation of resources in multi-beam satellite systems [2]. To effectively distribute resources in satellite systems with multiple beams, a comprehensive understanding of system dynamics is required. This includes satellite orbit and channel characteristics, as well as user traffic patterns and service requirements. To tackle these challenges, researchers are exploring innovative techniques, such as machine

F. Gao et al. (Eds.): Chinacom 2023, LNICST 590, pp. 348–363, 2024.
https://doi.org/10.1007/978-3-031-67162-3_22

learning algorithms and optimization models, to optimize resource allocation in the context of satellite systems with multiple beams [3].

The matter of allocating resources in satellite communication systems has garnered significant attention in recent years. The power allocation problem has been investigated for satellite systems in [4,5]. The authors in [4] explored balancing the relationship between transmit power and beam directivity, we have introduced a versatile power allocation algorithm aimed at maximizing the rate of traffic accommodation. Literature [5] suggested an energy-aware power allocation algorithm that concurrently minimizes addressing the unfulfilled system capacity and total radiated power through multi-objective optimization. In [6, 7], explored the joint power and frequency-domain resource allocation in a satellite system. Literature [6] suggested a versatile power and carrier allocation method aimed at meeting heterogeneous traffic demands while maximizing satellite resource utilization. Reference [7] suggested a resource allocation scheme focused on optimizing power and frequency efficiency to minimize inter-component interference and maximize throughput for users.

Over the past few years, deep reinforcement learning algorithms have been utilized to address resource allocation challenges in satellite communication systems. In [8], present a deep reinforcement learning algorithm that leverages the Twin Delayed Deep Deterministic Policy Gradient method to collaboratively allocate sub-channels and power to satellite users. Reference [9] proposed a cooperative multi-agent deep reinforcement learning framework for managing radio resources in satellite systems. Additionally, formulated a bandwidth allocation strategy derived from this proposed framework.

While the issue of allocation of resources in satellite systems has been explored in prior research, the joint design of beam scheduling, power allocation and sub-channel allocation has not been investigated extensively. Furthermore, few work considers the gap between user service requirement and the service providing capability of the system. This paper explores the dynamic allocation of radio resources in satellite systems with multiple beams. We frame the integrated problem of beam scheduling, sub-channel, and power allocation as a minimization task for an average cost function. To address the defined problem, we begin by formulating it as a Markov Decision Process (MDP). Subsequently, we introduce an algorithm based on Double Deep Q Learning (DDQN) and an enhanced Priority Experience Replay DDQN (PER-DDQN) algorithm.

2 System Model

2.1 Network Model

In this study, we examine a satellite communication system comprising a geosynchronous orbit (GEO) satellite with multiple beams and several cells. The GEO satellite is furnished with an onboard transceiver, enabling the transmission of data packets to users through service links. To improve the efficiency of data transmission, the satellite is equipped with several spot beams, each covering specific cells. Let K represent the total count of beams, and N represent the total

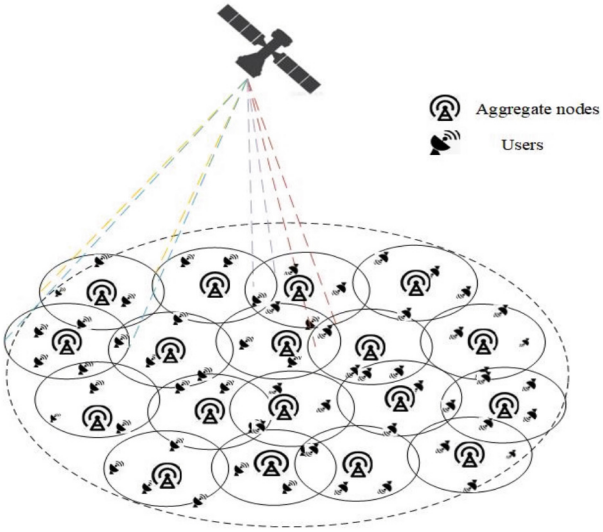

Fig. 1. System model

count of cells. C_n represents the n-th cell, $1 \leq n \leq N$. Assuming the satellite's overall bandwidth is partitioned into F sub-channels, the beams support full frequency multiplexing, where each beam accommodates multiple sub-channels. Let B represent the bandwidth allocated to each sub-channel.

For simplicity, the system time T is divided into consecutive time slots of equal duration, where the duration of each time slot is represented by $\acute{O}$. Without sacrificing generality, we make the assumption that the channel characteristics of the links between the satellite and the cells may change over time, but they remain constant within each time slot. We make the assumption that within a single time slot, a satellite beam exclusively covers one cell. Suppose there might be overlapping regions between adjacent cells, if these cells share same sub-channels, transmission interference may occur.

Considering the similarity of the users in link transmission characteristics and service requirement, we introduce the concept of aggregate nodes (ANs) to represent the users in individual cells. Let AN_n denote the AN of C_n. The location of AN_n is set as the geographical center of C_n and the traffic amount of AN_n is defined as the total traffic aggregation of all users in C_n. Due to the limited beam width, the overlap between adjacent cells. When the neighboring cells are selected by the beam, there will be an overlap between the beams, which can cause interference when using the same sub-channel. To describe the neighbourship between ANs, we introduce a binary variable $\alpha_{n,j} \in \{0, 1\}$. If AN_n and AN_j are neighboring nodes, we set $\alpha_{n,j} = 1$, otherwise $\alpha_{n,j} = 0$. The system model depicted in Fig. 1 is under consideration in this paper.

2.2 Channel Model

Let $h_{t,n,f}$ represent the channel gain of the service link between the satellite and AN_n at time slot t and sub-channel f, which can be represented as

$$h_{t,n,f} = g^{\mathrm{t}} g_n^{\mathrm{r}} L_{n,f} L_t \tag{1}$$

where g^{t} denote the gain of the transmit antenna for satellite beams and can be expressed as

$$g^{\mathrm{t}} = \frac{2\pi - (2\pi - \theta)\delta}{\theta} \tag{2}$$

where θ represents the antenna beam width, $\delta \ll 1$ is a small constant, g_n^{r} denotes the receiving antenna gain of ANn, and Ln, f denotes the free space loss in the link from the satellite to AN_n on sub-channel f, which can be computed

$$L_{n,f} = \left(\frac{c}{4\pi d_n \xi_f}\right)^2 \tag{3}$$

where c denotes the speed of light, d_n denotes the distance between the AN_n and satellite, and ξ_f represents the carrier frequency of sub-channel f, L_t represents the coefficient of rain attenuation in the time slot t, which can be represented as a Gaussian random variable following $N(\mu, \sigma^2)$.

3 Problem Formulation

In this section, we establish system cost function which jointly considers user service requirement and the service providing capability of the system. Then, a joint beam illumination, sub-channel and power allocation problem is formulated as a constrained system cost minimize problem.

3.1 System Cost Function Formulation

We denote ϑ_t as the system cost function in time slot t. Taking into account the collective utilities of all cells in the system, we formulate ϑ_t as

$$\vartheta_t = \sum_{n=1}^{N} \vartheta_{t,n} \tag{4}$$

where $\vartheta_{t,n}$ denotes the cost function of AN_n in time slot t. Addressing the difference between the service requirement and achievable capacity of AN_n, we define $\vartheta_{t,n}$ as

$$\vartheta_{t,n} = |\eta_{t,n} - R_{t,n}| \tag{5}$$

where $\eta_{t,n}$ denotes the amount of data packets to be transmitted to AN_n in time slot t, i.e.,

$$\eta_{t,n} = \max\{\eta_{t-1,n} - R_{t-1,n}, 0\} + A_{t,n} \tag{6}$$

where $A_{t,n}$ represents the service requirement at the commencement of time slot t, $R_{t,n}$ represents the achievable capacity of AN_n in time slot t. $R_{t,n}$ can be computed as

$$R_{t,n} = \sum_{f=1}^{F} x_{t,n,f} R_{t,n,f} \tag{7}$$

where $x_{t,n,f}$ is the sub-channel selection variable, if AN_n is assigned sub-channel f in time slot t, we set $x_{t,n,f} = 1$, otherwise $x_{t,n,f} = 0$, $R_{t,n,f}$ is the achievable data of the link from the satellite to AN_n on sub-channel f in time slot t, which given by

$$R_{t,n,f} = B \log(1 + \gamma_{t,n,f}) \tag{8}$$

where $\gamma_{t,n,f}$ illustrates the ratio of received signal strength to the combined interference and noise level, known as Signal-to-Interference-plus-Noise Ratio (SINR) of AN_n in time slot t, $\gamma_{t,n,f}$ can be computed as

$$\gamma_{t,n,f} = \frac{P_{t,n,f} h_{t,n,f}}{N_0 B + \sum\limits_{\hat{n} \neq n, \hat{n} \in N} x_{t,n,f} \alpha_{n,\hat{n}} \beta h_{t,n,f} P_{t,\hat{n},f}} \tag{9}$$

where $P_{t,n,f}$ represents the transmit power of AN_n which select sub-channel f in time slot t, β represents the channel attenuation coefficient, which describes the interference level received from neighboring cells, N_0 denotes the power spectral density of additive white Gaussian noise (AWGN).

3.2 Optimization Constraints

In this subsection, the optimization constraints which should be considered when designing resource allocation strategy are discussed in detail.

Beam Illumination Constraint. To depict the beam illumination status of the cell, we introduce a binary variable $y_{t,n,k}$. If AN_n is lit up by a beam in time slot t, we set $y_{t,n,k} = 1$, otherwise $y_{t,n,k} = 0$. To avoid inter-beam interference within an individual user group, we assume that an AN can only be lit up by a single satellite beam in a given time slot t, i.e.,

$$\mathrm{C1} : \sum_{k=1}^{K} y_{t,n,k} \leq 1. \tag{10}$$

In each time slot, we assume that one satellite beam can only illuminate one AN, i.e.,

$$\mathrm{C2} : \sum_{n=1}^{N} y_{t,n,k} \leq 1. \tag{11}$$

Time-Frequency Assignment Constraints. It is apparent that only if an AN is illuminated by one beam, time-frequency resource should be assigned to the user. The constraint is expressed as

$$\mathrm{C3}: x_{t,n,f} \leq y_{t,n,k}. \tag{12}$$

The number of sub-channels assigned to one AN cannot exceed F, i.e.,

$$\mathrm{C4}: \sum_{n=1}^{N} x_{t,n,f} \leq F. \tag{13}$$

It is presumed that each AN in time slot t can only occupy one sub-channel, the constraint can be represented as

$$\mathrm{C5}: \sum_{f=1}^{F} x_{t,n,f} \leq 1. \tag{14}$$

Power Allocation Constraint. The power emitted by a satellite beam must not surpass the specified maximum limit. This restriction can be articulated as

$$\mathrm{C6}: P_{t,n,f} \leq P_{\max} \tag{15}$$

where $P_{\max}$ denotes the highest power output of a satellite beam.

3.3 Optimization Problem Formulation

We formulate the integrated problem of joint beam illumination, sub-channel selection, and power allocation as a minimization task for a long-term average system cost function. This can be computed as

$$\min_{x_{t,n,f}, P_{t,n,f}, y_{t,n,k}} \lim_{T \to \infty} \frac{1}{T} E\left[\sum_{t=1}^{T} \vartheta_t\right] \tag{16}$$

$$\text{s.t. } \mathrm{C1} - \mathrm{C6}. \tag{17}$$

4 Solution to the Optimization Problem

The optimization problem formulated in (16) is categorized as an NP-hard problem which obtaining the optimal solution through conventional methods poses challenges. In this paper, we initially introduce a scheme for grouping of cells and scheduling of beams and then propose DDQN-based and PER-DDQN-based joint power and sub-channel allocation algorithms.

4.1 Beam Scheduling Scheme

In this subsection, we propose two beam scheduling schemes, referred to as Scheme 1 and Scheme 2, respectively. In Scheme 1, the cells with the highest service demand are illuminated in each time slot. In particular, we rank the cells based on their service demand, then select the K cells with the highest service demand for beam illumination. Suppose $\eta_{t,n_1} \geq \eta_{t,n_2} \geq \ldots \geq \eta_{t,n_N}$, the first n_K cells are selected, we set $y^*_{t,n_k,k} = 1, 1 \leq k \leq K$.

The basic idea of Scheme 2 is to schedule the interference-free cells with the highest service demand. In particular, we rank the cells based on their service demand. Suppose $\eta_{t,n_1} \geq \eta_{t,n_2} \geq \ldots \geq \eta_{t,n_N}$, the cell with the highest demand, i.e., cell AN_{n_1} is selected, and we set $y^*_{t,n_1,k} = 1$. Then, the cell with the second highest demand, i.e., cell AN_{n_2} is examined. To avoid interference between cells, we check whether cell AN_{n_1} and AN_{n_2} are adjacent cells, if not, i.e., $\alpha_{ni,nj} = 0$, AN_{n_2} is selected for beam illumination and we set $y^*_{t,n_2,k} = 1$; otherwise, it is removed from the rank list and is not illuminated in time slot t. The process repeats until K cells are selected for beam illumination.

4.2 MDP Modeling

Based on the obtained beam illumination strategy, we now design power and sub-channel allocation strategy for the cells illuminated in individual time slots. To this end, we model MDPs for Scheme 1 and Scheme 2, respectively. An MDP can be represented by a triple $\langle s, a, r \rangle$, where s is system state set, a is the action, r is the immediate reward function. Specifically, the states, actions and the reward functions of the MDPs for Scheme 1 and Scheme 2 can be expressed as follows.

MDP Modeling for Scheme 1

State. Let s_t denote the state in time slot t, defined as the set of service requirements and the channel characteristics of the ANs, i.e.,

$$s_t = \{\eta_t, h_t\}, \tag{18}$$

where $\eta_t = \{\eta_{t,n} | \exists k, y^*_{t,n,k} = 1\}$, $h_t = \{h_{t,n,f} | \exists k, y^*_{t,n,k} = 1\}$.

Action. Note that the transmitting power of the satellite is a continuous variable that cannot be optimized using DDQN. To resolve this problem, we apply discretization scheme to convert continuous power to discrete power levels. Specifically, the transmit power is evenly divided into J orders, furthermore, ANs can only be assigned specific power levels. Let $\hat{P}_j$ denote the transmit power of the j-th level. $\hat{P}_j$ can be computed as

$$\hat{P}_j = \frac{P_{\max} j}{J}. \tag{19}$$

Suppose AN_n is exposed to a beam during time slot t, i.e., $\exists k$, such that $y_{t,n,k} = 1$. Let $P_{t,n}$ denote the transmitting power of AN_n in time slot t, we obtain $P_{t,n} \in \{\hat{P}_1, \hat{P}_2, \ldots, \hat{P}_j\}$. To characterize the sub-channel allocation strategy of AN_n in time slot t, we introduce $X_{t,n} \in \{0, 1, ..., F\}$. Specifically, if AN_n is assigned sub-channel f in time slot t, we set $X_{t,n} = f$, $1 \leq f \leq F$. Let $a_t = \{a_{t,n}\}$ denote the action of ANs in time slot t which can be represented as

$$a_{t,n} = \{X_{t,n}, P_{t,n}\}. \tag{20}$$

Reward. At state s_t, the agent takes action a_t and receives an immediate reward. In this paper, aiming to minimize the difference between the service requirement and achievable capacity, the reward function is defined as r_t, which can be modeled as

$$r_t = \frac{1}{1 + \exp(b_1(q_t - c_1))}, \tag{21}$$

where b_1 and c_1 are the weighting parameters, and q_t can be computed as

$$q_t = \rho_1 \vartheta_t + \rho_2 P_t, \tag{22}$$

where ρ_1 and ρ_2 represent weighting parameters.

MDP Modeling for Scheme 2. According to Scheme 2, only the non-adjacent cells receive illumination in each time slot, therefore, the transmission interference among cells is avoided. As a result, within a given time slot, the strategies for allocating sub-channels and power to various ground cells can be independently designed. Consequently, the initial problem of allocating sub-channels and power in the satellite system is streamlined to the resource allocation problem for individual ground cells. Suppose AN_n is illuminated by a beam emitted from a satellite in time slot t, the MDP is modeled as below.

The state of the system environment is defined as $\hat{s}_t = \{\eta_{t,n}, h_{t,n,f}\}$. Let $\hat{a}_{t,n}$ denote the action of ANs in time slot t, which can be computed as

$$\hat{a}_t = \{x_{t,n,f}, P_{t,n}\}, \tag{23}$$

At state $\hat{s}_t$, the agent takes action $\hat{a}_t$ and receives an immediate reward. In this paper, aiming to minimize the difference between the service requirement and achievable capacity, the reward function is defined as $\hat{r}_t$, which can be modeled as

$$\hat{r}_t = \frac{1}{1 + \exp(b_2(\hat{q}_t - c_2))}, \tag{24}$$

where b_2 and c_2 are the weighting parameters. $\hat{q}_t$ can be computed as

$$\hat{q}_t = \rho_3 \vartheta_t + \rho_4 P_t, \tag{25}$$

where ρ_3 and ρ_4 represent weighting parameters.

4.3 DDQN-Based Joint Sub-channel and Power Allocation Algorithm

The DQN algorithm consists of a target network and a prediction network. In particular, the prediction network $Q(s, a, \theta)$ is used to evaluate different actions, and the parameters of the prediction network are updated in real time. The target network $\hat{Q}(s, a, \hat{\theta})$ is utilized to assist the prediction network in updating network parameters, additionally, the target network parameters are updated periodically to minimize the likelihood of oscillation divergence during training. The DQN algorithm achieves Q value stabilization by integrating the target network, maintaining a consistent state. Simultaneously, it diminishes the correlation between the target Q value and the predicted Q value, leading to the training loss value converging towards a stable state. Ultimately, this accelerates the speed at which the training algorithm converges.

Given that the DQN algorithm employs the same Q value for indexing and predicting actions, there is a potential risk of Q value overestimation. To address this issue. The DDQN algorithm is proposed which decouples the selecting actions and evaluating target Q values to mitigate overestimation. The predicted Q value can be expressed as

$$y_t = r_t + \gamma \hat{Q}(s_{t+1}, \arg\max_a Q(s_{t+1}, a; \theta); \hat{\theta}) \tag{26}$$

where γ represents the discount factor, θ and $\hat{\theta}$ denote the parameters of the target network and the prediction network, respectively.

The minimum mean square error (MSE) serves as the loss function for optimizing the parameters of the prediction network. The parameter θ can be trained through the minimization of the loss function, i.e.,

$$\theta = \theta - \alpha \nabla E[y_t - Q(s_t, a_t; \theta)^2] \tag{27}$$

where $\alpha \in [0, 1]$ is the learning rate.

The algorithm utilizing DDQN, as proposed, is outlined in Algorithm 1.

4.4 PER-DDQN Based Joint Beam Illumination, Sub-channel Selection and Power Allocation Algorithm

In the DQN and DDQN algorithms, the classic "experience replay" scheme is applied, which adopts uniform sampling and batch updates. Consequently, the cached experience transitions are replayed with similar probabilities. However, as the time difference error (TD-error) may vary for different training samples, choosing these samples with the same possibilities may result in relatively long convergence time.

In this subsection, we propose a PER-DDQN based joint beam illumination, sub-channel and power allocation algorithm. In PER-DDQN framework, we apply a PER scheme which increases the utilization of samples with larger TD-errors. Specifically, the data with larger TD errors in previous training has

Algorithm 1. DDQN-based Resource Allocation Algorithm

Initialize experience replay pool D
Initialize prediction network Q with weight θ, target network $\hat{Q}$ with weight $\hat{\theta}$
for training step $= 1 : N_{\max}$ **do**
  Giving the initial observed state s_1
  for $t = 1 : T$ **do**
    Choose an action using the ϵ-greedy method from the prediction network $Q(s, a; \theta)$
    Obtain a reward according to (21), system state transits to new state s_{t+1}
    Store transition pair (s_t, a_t, r_t, s_{t+1}) in data set D
    for $j = 1 : J$ **do**
      Sample a random mini-batch of transitions (s_j, a_j, r_j, s_{j+1}) from data set D
      Update y_j according to (26)
    end for
    Update weight θ according to (27)
    Update target network weight $\hat{\theta} = \theta$ after certain steps
  end for
end for

higher probability to be sampled, thus accelerating gradient descent and speeding up convergence.

Without loss of generality, we denote the j-th transition stored in the replay memory D as (s_j, a_j, r_j, s_{j+1}). Let q_j denote the priority of transition j, which can be expressed as

$$q_j = |\delta_j| + \eta \tag{28}$$

where η is a positive constant that guarantees the transition is assigned a non-zero priority, δ_j represents the TD-error of transition j, which can be calculated as

$$\delta_j = r_j + \gamma \hat{Q}(s_{j+1}, \arg\max_a(Q(s_{j+1}, a_i; \theta); \hat{\theta})) - Q(s_j, a_j; \theta) \tag{29}$$

where γ denotes discount factor. Let p_j denote the sampling probability of transition j, which is defined as

$$p_j = \frac{q_j^{\psi}}{\sum_{i=1}^{J}(q_i^{\psi})} \tag{30}$$

where $\psi \in [0, 1]$ is the prioritization exponent. J is the size of mini-batch.

In order to ensure experience transitions with smaller TD-errors to be extracted and achieve the diversity of extracted experience samples, importance-sampling weights are further introduced. Let ω_j denote the weight of transition j, which can be expressed as

$$\omega_j = \frac{(|D|p_j)^{-\phi}}{\max_i(\omega_i)} \tag{31}$$

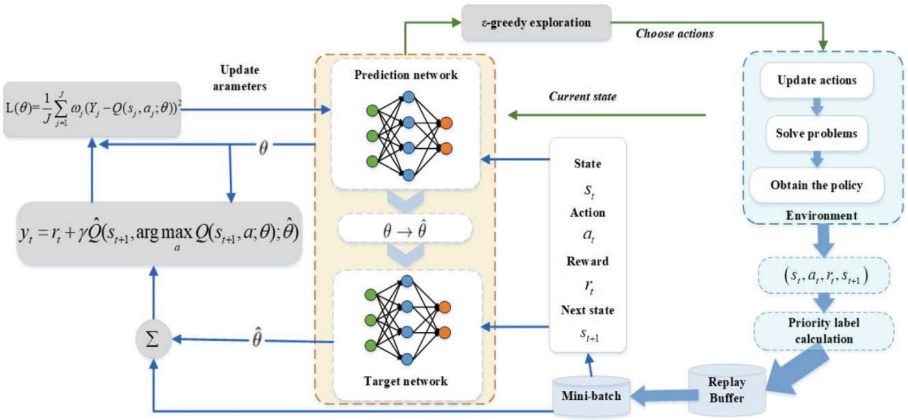

Fig. 2. The structure of the PER-DDQN framework

where ϕ is the correction parameter, $|D|$ is the size of experience replay pool.

We utilize the gradient descent algorithm to update and optimize the parameters of the prediction network, i.e.,

$$\theta \leftarrow \theta + \alpha \nabla \frac{1}{J} \sum_{j=1}^{J} (w_j \delta_j^2) \tag{32}$$

where $\alpha \in [0, 1]$ denotes the learning rate. After a specific number of iterations, we obtain the updated parameter values denoted as θ, which are subsequently used to replace the previous parameter values $\hat{\theta}$

Algorithm 2 shows the PER-DDQN based joint beam illumination, sub-channel selection and power allocation algorithm. Figure 2 shows the structure of the PER-DDQN framework.

5 Simulation Results and Analysis

5.1 System Parameters

In this section, we assess the effectiveness of the proposed algorithm. For comparison, we evaluate the performance of DQN, DDQN and PER-DDQN-based algorithms that simultaneously consider sub-channel selection and power allocation together. Simulation parameters are shown in Table 1. To assess the effectiveness of the proposed algorithms, we initially simulate a satellite communication scenario using Python software. Subsequently, we proceed to implement the proposed algorithm framework based on DDQN and PER-DDQN. We leverage Google TensorFlow-2.0 and configure two identical fully connected neural networks, where one acts as the prediction network and the other as the target network. Each neural network comprises one output layer, one input layer, and

Algorithm 2. PER-DDQN based joint sub-channel selection and power allocation algorithm

Prioritized experience replay pool size $|D|$, mini-batch size J, hyper-parameters.
Initialize prioritized experience replay pool D, prediction network Q with weight θ, target network $\hat{Q}$ with weight $\hat{\theta}$
for training step $= 1 : N_{\max}$ **do**
  Giving the initial observed state s_1
  for $t = 1 : T$ **do**
    Select an action using ϵ-greedy method from the prediction network $Q(s, a; \theta)$
    Execute action a_t, obtains a reward according to (24), system state transits to new state s_{t+1}
    Store transition pair (s_t, a_t, r_t, s_{t+1}) in data set D and set $q_t = \max_{i<t} q_i$
    for $j = 1 : J$ **do**
      Sample transition j with probability p_j in (30)
      Calculate the importance-sampling weight ω_j using (31)
      Calculate $|\delta_j|$ using (29)
      Update transition priority q_j according to $|\delta_j|$
    end for
    Update weight θ according to (32)
    Update target network weight $\hat{\theta} = \theta$ after certain steps
  end for
end for

three hidden layers, with each hidden layer containing 50 neurons. The rectified linear unit function is adopted as the activation function of all hidden layers. Additionally, we use the Adam optimizer during the training of the Deep Neural Network (DNN) to minimize the Mean Squared Error (MSE). Furthermore, We establish a storage container called a prioritized experience replay pool with a capacity of 5000, which is used to store historical experimental data.

5.2 Simulation Results

Figure 3 depicts the relationship between the cumulative reward and the number of algorithm episodes. The cumulative reward is calculated as the sum of rewards obtained in each episode. For comparison, we plot the DQN, DDQN and PER-DDQN-based algorithms. As can be seen from the figure, the algorithms reach convergence as the number of training steps increases, demonstrating the effectiveness of algorithms. Comparing the results obtained from the DQN, DDQN and PER-DDQN algorithms, we can see that PER-DDQN offers the fastest convergence rate.

Figure 4 shows the system cost function versus sub-channel bandwidth. For comparison, we plot the system cost function obtained from our proposed algorithms with different beam illumination schemes. From the figure we can observe that as sub-channel bandwidth increases, system cost decreases first, and starts to increase as the bandwidth reaches to a certain value. The reason is that when

Table 1. System parameters.

Parameter	Value
Satellite altitude (H)	35786 km
Carrier frequency (f_m)	20 GHz
Power spectral density (N_0)	–174 dBm/Hz
Satellite transmit antenna gain (g^{t})	52 dBi
Maximum beam power ($P_{\max}$)	30 W
System bandwidth (B)	24 MHz
Number of cells (N)	18
Number of beams (K)	4
Number of sub-channels (F)	3
Training episode	2000
Replay memory size ($\|D\|$)	5000
Greedy probability (ε)	0.9
Learning rate (α)	0.00001
Discount factor (γ)	0.95

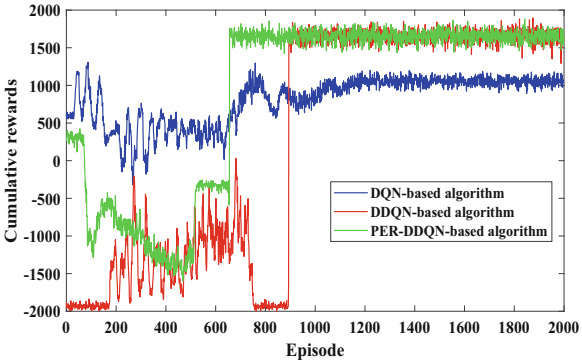

Fig. 3. The cumulative reward vs the number of episodes

the bandwidth is small, increasing bandwidth leads to higher service providing capability, and smaller difference in user requirement and service providing capability. However, as the bandwidth reaches to a certain value, increasing bandwidth results in excessive system service providing capability, and the increase of the difference between user requirement and service providing capability, causing high cost function. Comparing the proposed approaches, it is evident from the results that our proposed Scheme 2 algorithm achieves superior performance. This is because the interference among various cells is avoided, resulting in higher transmission performance.

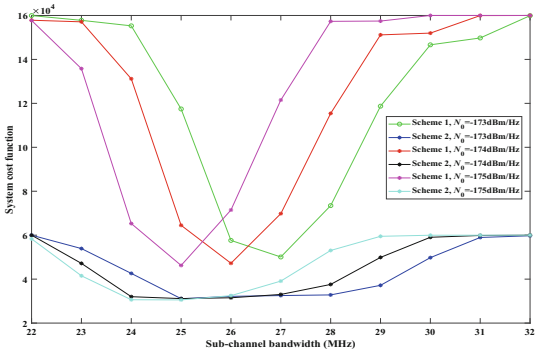

Fig. 4. System cost function vs sub-channel bandwidth.

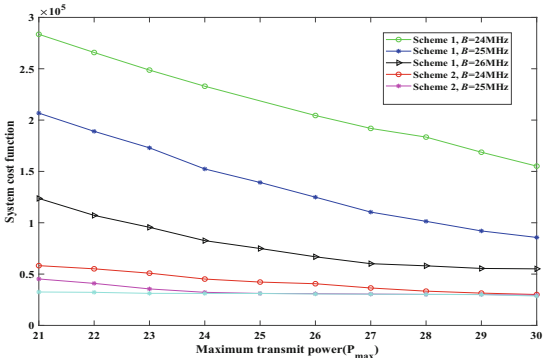

Fig. 5. System cost function vs maximum transmit power (different beam illumination schemes).

Figure 5 shows system cost function versus maximum transmit power obtained from different beam illumination schemes. Observing the figure, it is observed that with an increase in the maximum transmit power, the system cost decreases. The reason for this is that when the maximum transmit power is low, the service providing capability may not achieve user requirement, leading to high difference between user requirement and service providing capability. With the increase in maximum transmit power, the optimal allocation of transmit power can be achieved, leading to a reduced disparity between user requirements and service provisioning capabilities. Consequently, the system cost is minimized. Comparing the proposed approaches, it is evident from the results that our proposed Scheme 2 algorithm achieves superior performance. From the figure we can also observe that higher sub-channel bandwidth yields better system performance.

Figure 6 shows system cost function versus maximum transmit power obtained from different RL algorithms. For comparison, we plot the system cost function obtained from DQN and DDQN-based algorithms. From the figure, we

can be observed that with an increase in the maximum transmit power, the system cost decreases. Our proposed algorithm based on DDQN outperforms the DQN algorithm in terms of performance.

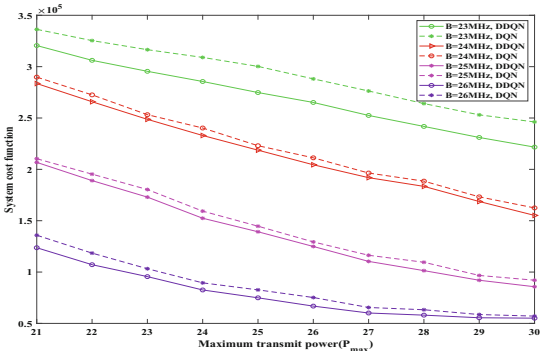

Fig. 6. System cost function vs maximum transmit power (different RL algorithms).

6 Conclusions

In this paper, we have conducted a study on the problem of joint beam scheduling, sub-channel allocation, and power allocation for multi-beam satellite systems and formulated the problem as an average cost function minimization problem. In order to address the issue, we have proposed two beam scheduling schemes. Based on the beam scheduling strategies, we have then proposed a DDQN-based and an improved PER-DDQN algorithm to determine sub-channel and power allocation strategy. Several numerical experiments have been conducted to assess the performance of the proposed algorithms. In particular, both the proposed DDQN and the improved PER-DDQN algorithm are capable of achieving convergence and the improved PER-DDQN algorithm offers faster convergence speed than the DDQN-based algorithm. We have also assessed the effects of sub-channel bandwidth and maximum transmit power on the system cost function. The simulation results have illustrated the effectiveness and superiority of the proposed algorithm.

References

1. Centenaro, M., Costa, C.E., Granelli, F., Sacchi, C., Vangelista, L.: A survey on technologies, standards and open challenges in satellite IoT. IEEE Commun. Surv. Tutor. **23**(3), 1693–1720 (2021)
2. Xiao, A., Wang, X., Wu, S., Jiang, C., Ma, L.: Mobility-aware resource management for integrated satellite-maritime mobile networks. IEEE Netw. **36**(1), 121–127 (2021)

3. Torkzaban, N., Zoulkarni, A., Gholami, A., Baras, J.S.: Capacitated beam placement for multi-beam non-geostationary satellite systems. In: IEEE Wireless Communication on Networking Conference (WCNC), pp. 1–6 (2023)
4. Takahashi, M., Kawamoto, Y., Kato, N., Miura, A., Toyoshima, M.: Adaptive power resource allocation with multi-beam directivity control in high-throughput satellite communication systems. IEEE Wirel. Commun. Lett. **8**(4), 1248–1251 (2019)
5. Efrem, C.N., Panagopoulos, A.D.: Dynamic energy-efficient power allocation in multibeam satellite systems. IEEE Wirel. Commun. Lett. **9**(2), 228–231 (2020)
6. Abdu, T.S., Kisseleff, S., Lagunas, E., Chatzinotas, S.: Flexible resource optimization for GEO multibeam satellite communication system. IEEE Trans. Wirel. Commun. **20**(12), 7888–7902 (2021)
7. Chan, S., Lee, H., Kim, S., Oh, D.: Intelligent low complexity resource allocation method for integrated satellite-terrestrial systems. IEEE Wirel. Commun. Lett. **11**(5), 1087–1091 (2022)
8. Deng, D., Wang, C., Pang, M., Wang, W.: Dynamic resource allocation with deep reinforcement learning in multibeam satellite communication. IEEE Wirel. Commun. Lett. **12**(1), 75–79 (2022)
9. Liao, X., Hu, X., Liu, Z.: Distributed intelligence: a verification for multi-agent DRL-based multibeam satellite resource allocation. IEEE Commun. Lett. **24**(12), 2785–2789 (2020)