Maximum Entropy Deep Reinforcement Learning Based Power Allocation for NOMA Maritime Network

Jiayi He[1], Yakai Zhang[1], and Zhiyong Liu[1,2,3(✉)]

[1] School of Information Science and Engineering of HIT, Weihai 264209, China
lzyhit@hit.edu.cn
[2] Shandong Provincial Key Laboratory of Marine Electronic Information and Intelligent Unmanned Systems, Weihai 264209, China
[3] Key Laboratory of Cross Domain Synergy and Comprehensive Support for Unmanned Marine Systems of Ministry of Industry and Information Technology, Weihai 264209, China

Abstract. In order to address the challenges of high propagation delays and limited service capabilities in maritime satellite communications, unmanned aerial vehicles have been proposed as an airborne backhaul solution to enhance communications between satellites and maritime base stations. The non-orthogonal multiple access (NOMA) framework can solve the user sparsity problem in maritime networks. In this paper, a deep reinforcement learning algorithm is used to solve the nonconvex power allocation problem under NOMA. In order to mitigate the risk of overestimation of Q values and local optimal convergence of Deep Q Network (DQN) algorithm, we propose an algorithm called Soft Agent Critical Ocean Satellite Communication Power Allocation (SAC-OSCPA) based on the idea of maximum entropy and compare it with the traditional DQN algorithm. The main goal of this research is to maximize network throughput in scenarios with randomly distributed users. Simulation results show that the average system throughput is improved by 13.18% with the SAC-OSCPA algorithm, and the average throughput of the worst performing user is significantly improved by 41.59%. These results demonstrate the efficacy of the proposed algorithm in optimizing the communication performance of maritime satellite networks.

Keywords: maritime network · deep reinforcement learning · power allocation · non-orthogonal multiple access (NOMA)

This work was supported in part by the National Natural Science Foundation of China under Grant 61871148; in part by the Major Scientifi and Technological Innovation Project of Shandong Province of China under Grant 2020CXGC010705, under Grant 2021ZLGX05 and under Grant 2022ZLGX04, in part by Strategic Rocketry Innovation Fund Project under Grant ZH2022007.

H.-H. Chen and W. Meng (Eds.): WiSATS 2024, LNICST 606, pp. 215–229, 2025.
https://doi.org/10.1007/978-3-031-86203-8_18

1 Introduction

Inmarsat is responsible for maritime communications, but its data rates are extremely low and inefficient for instantaneous applications. To overcome this problem, especially to provide high quality communication services to maritime users, the use of satellites and UAVs (Unmanned aerial vehicles) to relay communication signals is a potential solution [1]. Reference [2,3] shows that existing fifth generation land based cellular networks can be extended to maritime networks. UAVs have been tested with airborne base stations, demonstrating their ability to provide line of sight (LoS) services to terrestrial user equipment. Reference [4,5] has explored the potential of UAVs as an adjunct and relay for LEO satellite maritime communications. LEO satellite constellations provide services with the assistance of UAVs. In addition, in maritime communication scenarios, due to wide coverage and sparse user distribution, it is very difficult and costly to improve resource efficiency [6]. In China's offshore areas, the average number of users per square kilometer is about 0.2 [7], which is much smaller than on land. In most cases, the path loss in maritime networks is close to the direct path loss. Due to the complexity of maritime Internet of Things (IoT), the transmission of messages between different users can be flexibly and efficiently carried out through a non-orthogonal mode of operation [8], which proposed the application of the NOMA framework to nearshore maritime networks. This approach can achieve better spectral efficiency and increase system throughput than OMA [9].

1.1 Related Work

For the power allocation problem in maritime network communication scenarios, it can be solved by various methods, such as heuristic search algorithms. A particle swarm algorithm based power allocation strategy for multibeam satellite systems was proposed in [10]. In addition, a joint power and bandwidth allocation strategy for multibeam satellite systems based on genetic algorithms was proposed in [11].

One of the promising methods is artificial intelligence based methods such as reinforcement learning. Based on the form of Q tables, [12] applied RL to the problem of power allocation in satellite earth communications utilizing LEO satellites, extending the battery life of LEOs. Furthermore, [13] applied Q learning to the satellite user pairing problem and consequently power allocation. However the above works do not consider the characteristics of maritime networks. Reference [4] proposes a Q learning based power allocation algorithm for maritime satellites.

In response to the difficulty of traditional Q table based reinforcement learning algorithms to solve the power allocation problem in real maritime network environments due to its complexity and high dimensionality, [14] proposed a deep reinforcement learning algorithm using deep neural networks to estimate the Q values of the table entries, i.e. DQN. DQN has been utilized in a variety of contexts including vehicular internetworking and satellite terrestrial integrated

networks, and it has the advantage of better performance compared to traditional Q table based reinforcement learning algorithms [15]. [5] used DQN for resource allocation in maritime networks with good results. [16] suggested the use of DQN in time frequency two dimensional resource allocation algorithms. In [17], an algorithm mDQN extended by DQN to the multi-agent case was proposed. In the UAV path planning and power allocation problems in urban scenarios, mDQN achieves better results compared to DQN. However, DQN has the problem of being prone to overestimation of the Q value and falling into locally optimal strategies. To address this problem, [18] proposed a Soft Actor Critic (SAC) algorithm, which is an offline reinforcement learning algorithm with higher stability compared to DQN. The SAC algorithm was proposed in 2018, and has now become one of the most effective model free reinforcement learning algorithms. However, the soft behavioral critique algorithm does not use the display policy function but exploits the Boltzmann distribution of the function, which makes it difficult to apply in continuous spaces such as maritime network communication scenarios. To solve this problem, we improve the SAC algorithm by suggesting the use of actors instead of Boltzmann distribution to represent the policy function, and define the action space, state space and reward function adapted to maritime network communication scenarios, and propose the SAC-OSCPA algorithm. Since it has been shown in [4] that DQN outperforms Q learning and Water filling, this paper only compares with DQN in the simulation experiments. Additionally, as the communication scenario utilized in [8] with traditional methods differs from ours, no comparison is made with it. Our research goal is to improve the throughput of the network when users are randomly distributed.

1.2 Our New Contribution

The following is a summary of our primary contributions for this system proposal:

- To accommodate the extensive sparsity of communication users at sea, we construct a comprehensive integrated network of LEO satellites, UAVs, users at sea, and shipboard base stations based on the NOMA architecture to improve the throughput of the communication network.
- We use an improved maximum entropy deep reinforcement learning algorithm to solve the nonconvex power allocation optimization problem and compare it with classical methods such as DQN.

2 System Model and Problem Formulation

2.1 System Model

In the maritime satellite communication scenario studied in this paper, we consider a NOMA based satellite UAV sea surface network at sea. As shown

in Fig. 1, the maritime network consists of near Earth orbit (LEO) satellites $\mathcal{R} = \{1, 2, \cdots, r_k, \cdots, R\}$ and a series of UAVs $S = \{1, 2, \cdots, s_i, \cdots, S\}$ that provide network services to a range of surface users $\mathcal{U} = \{1, 2, \cdots, u_j, \cdots, U\}$ within their coverage area. Deploying satellite communications directly at sea is challenging due to the huge line of sight (LoS) and propagation delay. Therefore, UAVs are deployed as an airborne backhaul medium between users and satellites to ensure throughput and latency requirements. At sea, tethered UAVs are dynamically dispatched and form a virtual cluster with the TBSs. In forming the virtual cluster, we assume that the position of the UAVs is fixed and the UAVs provide services to the maritime end users based on the channel gain. We assume that the signal received at the receiver consists of delayed, attenuated, scattered and reflected signals. Where the attenuation is caused by variations in the received signal and can be represented by Riley's Beyond Line of Sight (BLoS) model and LoS model. The model has been shown to be applicable to maritime scenarios. To accommodate the extensive sparsity of the maritime network, we adopt the NOMA scheme. The cluster has multiple users, and orthogonal physical resource blocks (PRBs) are assigned to different pairs of users. Since clustering occurs only among users sharing the same PRB, we only consider a single PRB and its corresponding user. Edge servers are used for collaboration and powering of UAVs.

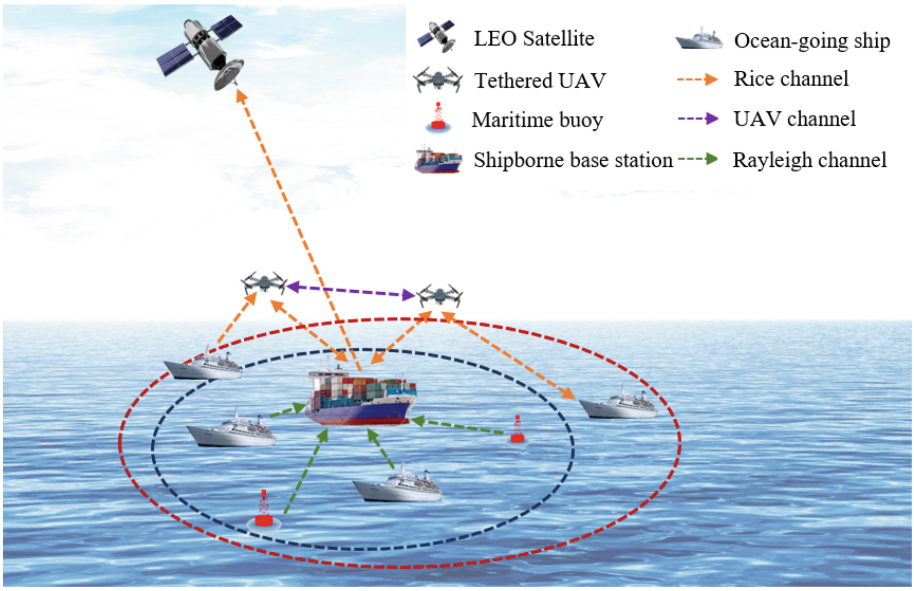

Fig. 1. NOMA maritime network framework.

Consider the Rayleigh channel model between the satellite and the backhaul UAV, as this model is commonly used in over the horizon communications. The

definition of the Rayleigh distribution model is as follows:

$$f(r) = \frac{r}{\sigma^2} \exp^{\left(-\frac{r^2}{2\sigma^2}\right)}, \quad r \geq 0. \tag{1}$$

In (1), r is a random variable, σ is the distributional mode and the root mean square of the signal that was received, $r^2/2$ is the current power, and σ^2 is the average power of the signal.

In addition, we consider the Rice fading channel model for maritime communication between UAVs and mobile vessels. The model is similar to the Rayleigh fading model, but in the Rayleigh fading model, the line of sight component between the UAV and the vessel plays a large dominant role. Therefore, the scattered wave is weaker than the signal wave. This Rice distribution model is defined as follows:

$$f(r) = \frac{r}{\sigma^2} \exp^{\left(-\frac{r^2+s^2}{2\sigma^2}\right)} I_0 \frac{rs}{\sigma^2}, \quad r \geq 0, \tag{2}$$

where I_0 denotes the first type of zero order modified Bessel function, and s represents the field strength of the LoS component. Ships in or close to coastal areas are also linked to coastal base stations (BSs). The next generation cellular network's extension, known as maritime BS, can offer maritime users in some waters broadband Internet access. These radio networks can only serve coastal marine users due to their limited coverage. In addition to helping with marine communication within a limited range, island base stations can offer return transit from satellites through drone relay. They can also offer superior communication management to nearby ships, vessels, and vessels.

For UAVs and maritime users, the system model is a downlink LEO satellite NOMA cellular system with significant sparsity. We assume that every satellite will utilize the same available bandwidth in order to maximize system spectrum efficiency. We think that network data like SINR and transmit power can be gathered by a central controller. The user's movement angle and speed at time t are specified as:

$$0 \leq \nu_{m,t} \leq \nu_{\max}, \quad 0 \leq \phi_{m,t} \leq 2\pi. \tag{3}$$

Describing the geographic distribution of satellites inside a constellation using a Poisson point process model. N orthogonal subcarriers are completely reused by each satellite. Our goal is to improve the utilization of system resources.

Therefore, we make the following basic assumptions: Each LEO can only provide services to one shipborne base station at a time. The transmission resources allocated for each task meet the visibility requirements of low orbit satellites and shipborne base stations. The data transmission process for each task will not be interrupted.

$$s.t.C_1 : \sum_{k=1}^{K} \sum_{i,k}^{|\mathrm{TW}_{i,k}|} X_{i,k}^{l} \leq 1, \forall 1 \leq i \leq |\mathrm{ME}|, \tag{4a}$$

$$C_2 : \sum_{l=1}^{|\mathrm{TW}_{i,k}|} X_{i,k}^{l} z_i \leq d_i, \forall 1 \leq i \leq |\mathrm{ME}|, \tag{4b}$$

$$\begin{aligned} &C_3 : X_{i,k}^{l}(s_i - st_{i,k}^{l})(et_{i,k}^{l} - z_i) \geq 0 \\ &\forall 1 \leq i \leq |\mathrm{ME}|, 1 \leq k \leq |R|, 1 \leq l \leq |\mathrm{TW}_{i,k}|, \end{aligned} \tag{4c}$$

$$\begin{aligned} &C_4 : (a_j - p_{ik} - a_i) \sum_{l=1}^{|\mathrm{TW}_{i,k}|} X_{i,k}^{l} \sum_{m=1}^{|\mathrm{TW}_{j,k}|} X_{j,k}^{m}(Y_{i,j,k} - 1) \\ &+ (a_i - p_{jk} - a_j) \sum_{l=1}^{|\mathrm{TW}_{i,k}|} X_{i,k}^{l} \sum_{m=1}^{|\mathrm{TW}_{j,k}|} X_{j,k}^{m}(1 - Y_{i,j,k}) \geq 0, \\ &\forall 1 \leq i \leq |\mathrm{ME}|, 1 \leq k \leq |R|. \end{aligned} \tag{4d}$$

In the next section, we create an optimization problem based on the previous analysis.

2.2 Problem Formulation

Let $\mu_{s,u}^{n}$ symbolize the signal to noise ratio received by the n^{th} subcarrier $\left(n \in \mathcal{N} = \left\{1, ..., N\right\}\right)$ from the s^{th} UAV service device u at time t , which is ascertainable as:

$$\mu_{s,u}^{n} = \frac{x_{s,u}\mathcal{X}_{s',u}^{n} h_{s,u}^{n} p_{s,u,t}^{n}}{\sum_{s' \neq s} h_{s',u}^{n'} p_{s',t}^{n'} + \sigma^2}. \tag{5}$$

where $h_{s,u}^{n}$ and $h_{s',u}^{n'}$ stand for device u's channel gain on the n^{th} subcarrier of the s^{th} and s'^{th} UAVs, respectively. Likewise, on the n^{th} subcarrier, $p_{s,u,t}^{n}$ and $p_{s',t}^{n'}$ stand for the total transmit power of the s^{th} and s'^{th} UAVs. Where σ^2 denotes the degree of Gaussian white noise and $\chi_{s,u}^{n}$ denotes whether UAV s is the subcarrier n assigned to device u, i.e., $\chi_{s,u}^{n} \in \left[0, 1\right]$. Lastly, $x_{s,u}$ defines the satellite device association, i.e. $x_{s,u} = \left\{1, 0\right\}$. Where, 1 means device u is connected to UAV s, otherwise it is 0.

Kilobyte (kb) is the unit of measurement for the system's overall throughput. The capacity that satellite has attained on the relevant subcarrier n equipment at time t is:

$$C_{s,t}^{n} = \frac{B}{N} \log_2 \left(1 + \sum_{u=1}^{U} \mu_{s,u}^{n}\right). \tag{6}$$

Therefore, the system throughput can be defined as: $R_t = \sum_{s=1}^{S} \sum_{n=1}^{N} C_{s,t}^{n}$. Our goal is to modify the transmit power of the satellite on the subcarrier, i.e.

$p_{s,t} = \left[p_{s,t}^1, ..., p_{s,t}^n, ..., p_{s,t}^N\right]$ to optimize network throughput. The optimization issue is stated mathematically as follows:

$$\text{P1:} \min_{p} R_t, \text{s.t.} p_{s,t}^n \geq p^{\min}, \forall s \in S, n \in N, \tag{7a}$$

$$\sum_{s_1}^{S} \sum_{n_1}^{N} p_{s,t}^n \leq p^{\max}, \forall s \in S, n \in N, p_{s,\mathrm{u},t}^n \geq 0, \forall s \in S, \forall u \in U, \forall n \in N. \tag{7b}$$

where $p^{\max}$ is the satellite's highest transmit power and $p^{\min}$ is the subcarrier's lowest transmit power. A multifaceted non-convex optimization challenge is what the aforementioned issue is. A popular approach to this issue's solution is the heuristic search algorithm. Nevertheless, the majority of these algorithms lack real time online modification capabilities and are inefficient. In the following part, we will create a deep reinforcement learning technique to address this issue.

3 Proposed SAC-OSCPA

3.1 SAC

The unpredictability of the random variable X is denoted by the entropy H. Specifically, if it is a random variable with probability density function $p(\cdot)$, its entropy is defined as follows: $H(X) = \mathbb{E}_{x\sim p}[-\log p(x)]$.

$H(\pi(\cdot \mid s))$ can be used in reinforcement learning to represent the strategy $\pi(\cdot)$'s random degree in the state s. The goal of maximum entropy reinforcement learning is to increase the stochasticity of the strategy based on optimizing the cumulative reward. Therefore, the universal objective function of maximum entropy reinforcement learning can be formulated as follows:

$$\pi^* = \arg\max_{\pi} \mathbb{E}_{\pi}[\sum_{t} r(s_t, a_t) + \alpha H(\pi(\cdot|s_t))]. \tag{8}$$

where α is a regularized variable that regulates the entropy's significance, s_t denotes the state of the environment at moment t, a_t denotes the action taken at t, and $r(\cdot)$ denotes the reward at t. Since (8) is a universal objective function, we need to correct it in the next step. Entropy regularization improves the degree to which the reinforcement learning algorithm explores unknown domains in the environment, and to some extent ameliorates the problem of employing greedy strategies in reinforcement learning which makes the agent fall into a locally optimal strategy. the larger the value of α, the faster the agent learns a new strategy in the next step, and the less likely it is to perform a suboptimal local optimization. With the modification of the objective function, certain additional definitions in the maximum entropy reinforcement learning model have changed.

First, the soft Bellman equation is shown below:

$$Q(s_t, a_t) = r(s_t, a_t) + \gamma \mathbb{E}_{s_{t+1}}[V(s_{t+1})]. \tag{9}$$

where we define the function that determines the value of the state return (state value function) as follows:

$$V(s_t) = \mathbb{E}_{a_t \sim \pi}[Q(s_t, a_t) - \alpha \log \pi(a_t \mid s_t)] = \mathbb{E}_{a_t \sim \pi}[Q(s_t, a_t)] + H(\pi(\cdot \mid s_t)). \quad (10)$$

Thus, in the case where both the state and action spaces are constrained, soft strategy evaluation can be combined with a soft Q function for strategy π, as shown in the soft Bellman equation. The following soft strategy facilitation equation can then be utilized to enhance the strategy:

$$\pi_{new} = \arg\min_{\pi'} D_{KL}\left(\pi'(\cdot \mid s), \exp(\frac{1}{\alpha} Q^{\pi_{old}}(s, \cdot))/Z^{\pi_{old}}(s, \cdot)\right). \quad (11)$$

where Z is a normalization function that does not contribute to the strategy gradient and is therefore negligible.

The alternate repeated iterative application of soft policy evaluation and policy boosting can help agent to converge the final policy to the optimal answer to the maximum entropy reinforcement learning objective. However, the soft policy iterative method is only applicable to the case of setting up Q value tables in traditional reinforcement learning, i.e., in environments with limited action and state spaces. In a continuous space such as the NOMA maritime satellite communication network, we need to improve the above process. Specifically we approximate such iterations by parameterizing the function Q and the policy π. To this end, we propose SAC-OSCPA in the next section to address the above problem.

3.2 SAC-OSCPA

SAC-OSCPA is an offline policy algorithm, in order to adapt to the special characteristics of satellite communication where it is difficult to implement the deployment and regulation of satellites, the agent can be trained offline on the ground before deploying it to LEO. In SAC-OSCPA algorithm, we model two action value functions Q (parameters ω_1 and ω_2 respectively) and a strategy function π (parameter θ). SAC employs two Q networks based on the concept of Double DQN. However, each time SAC-OSCPA selects that network with a lower Q value to minimize the problem of overestimation. For every function Q, the loss function is:

$$\begin{aligned} L_Q(\omega) &= \mathbb{E}_{(s_t, a_t, r_t, s_{t+1}) \sim R}\left[\frac{1}{2}\Big(Q_\omega(s_t, a_t) - \Big(r_t + \gamma V_{\omega^-}(s_{t+1})\Big)\Big)^2\right] \\ &= \mathrm{E}_{(s_t, a_t, r_t, s_{t+1}) \sim R, a_{t+1} \sim \pi_\theta(\cdot \mid s_{t+1})} \\ &\left[\frac{1}{2}\left(Q_\omega(s_t, a_t) - \left(r_t + \gamma\left(\min_{j=1,2} Q_{w_j^-}(s_{t+1}, a_{t+1}) - \alpha \log \pi(s_{t+1}, a_{t+1})\right)\right)\right)^2\right]. \end{aligned} \quad (12)$$

where R is the data already acquired by the current policy, since SAC-OSCPA is an offline policy approach. $Q_w(\cdot)$ denotes that the parameter is an action valued

function of $\mathcal{W}$, $Q_{w_j^-}$ denotes the target Q-network, and γ denotes the discount factor, which is a constant and takes values in the range $[0, 1)$. The closer γ is to 1 indicates that the strategy focuses more on long term cumulative rewards, and conversely more on short term rewards. In addition, $V_{w^-}(\cdot)$ denotes the state value function based on the strategy π in the Markov decision process, $V_{w^-} = \mathbb{E}_{a_t \sim \pi}\left[Q(s_t, a_t) - \alpha \log \pi(a_t \mid s_t)\right]$, $\mathbb{E}_{a_t \sim \pi}(\cdot)$ denotes the computational expectation.

In addition, we improve the stability of SAC-OSCPA training by utilizing the target Q network Q_{w^-}. Here again, two target Q networks Q_{w^-} are used, corresponding one by one to the two Q networks. The target Q networks in SAC-OSCPA are updated in the same way as DDPG. The difference between the probability distributions of the two Q networks is quantified by the KL dispersion, and the loss function of the strategy $\pi(\cdot)$ can be obtained, which can be simplified as:

$$L_\pi(\theta) = \mathbb{E}_{s_t \sim R, a_t \sim \pi_\theta}\left[\alpha \log(\pi_\theta(a_t \mid s_t)) - Q_\omega(s_t, a_t)\right]. \tag{13}$$

where (13) can be understood as maximizing the function $V(\cdot)$, since: $V(s_t) = \mathbb{E}_{a_t \sim \pi}\left[Q(s_t, a_t) - \alpha \log \pi(a_t \mid s_t)\right]$.

For the NOMA maritime communications network environment with a continuous action space, the SAC-OSCPA strategy yields Gaussian distributed means and standard deviations. However, it is not possible to sample actions directly from Gaussian distributions in the NOMA maritime communications network environment because the process is not derivable. Therefore, SAC-OSCPA employs the reparameterization technique to address this issue. The reparameterization is done by first sampling from a unit Gaussian distribution $\mathcal{N}$ and then multiplying the sampled values by the standard deviation and adding the mean. This can be viewed as sampling from a strategy Gaussian distribution and the strategy function corresponding to this sampling process is derivable, so we can derive the strategy function and denote it as $a_t = f_\theta(\varepsilon_t; s_t)$, where ε_t denotes the noisy random variable. The loss function of the rewrite strategy considers two Q functions:

$$L_\pi(\theta) = \mathbb{E}_{s_t \sim R, \varepsilon_t \sim N}\left[\alpha \log\left(\pi_\theta\left(f_\theta(\varepsilon_t; s_t) \mid s_t\right)\right) - \min_{j=1,2} Q_{w_j}\left(s_t, f_\theta(\varepsilon_t; s_t)\right)\right]. \tag{14}$$

In the SAC-OSCPA algorithm, the choice of the entropy regularity term coefficients is crucial. SAC-OSCPA rewrites the universal objective function in (8) as an optimization problem with the constraints (4a)-(4d), which allows it to modify the entropy regularity term automatically:

$$\max_\pi \mathbb{E}_\pi\left[\sum_t r(s_t, a_t)\right] s.t.\ \mathrm{E}_{(s_t, a_t) \sim \rho_\pi}\left[-\log(\pi_t(a_t \mid s_t))\right] \geq \mathcal{H}_0, \tag{15}$$

i.e., optimize the expected return whilst requiring the entropy mean value to be larger than $\mathcal{H}_0$. Following simplification by some mathematical tricks, the loss

function for α is obtained:

$$L(\alpha) = \mathbb{E}_{s_t \sim R, a_t \sim \pi(\cdot|s_t)} \left[-\alpha \log \pi(a_t \mid s_t) - \alpha H_0 \right]. \tag{16}$$

That is, in the process of minimizing the loss function $L_\pi(\theta)$ as previously described, the training objective $L(\alpha)$ increases the value of a when the strategy's entropy is lower than the target value $\mathcal{H}_0$, thereby increasing the significance of the corresponding component of the strategy's entropy; conversely, when the strategy's entropy is higher than the target value $\mathcal{H}_0$, the training objective $L(\alpha)$ decreases the value of α, thereby increasing the focus of the strategy training on value enhancement. (16) is well suited for solving the power allocation problem under NOMA maritime networks.

In order for the proposed SAC-OSCPA algorithm to be able to solve the power allocation problem under NOMA maritime networks, we need to proceed to define its state space, action space, and reward function in the maritime network communication environment as follows:

Action space: In order to provide user u access to the network, the agent must choose which seafaring base station to utilize. The action space is the same as the action space of the Agent. Let $a_u(t)$ indicate the action space at moment t as follow:

$$a_u(t) = \{\mathrm{A}_u^{com}(t)\}, \tag{17a}$$

$$ComR_u(t) = [ComR_u^1(t), ..., ComR_u^A(t)]. \tag{17b}$$

where $ComR_u^a(t), \forall a \in \mathcal{A}$ indicates whether the AP provides network availability for user u. $ComR_u^a(t) \in \{0, 1\}$. 1 indicates that network access is provided and 0 indicates that it is not provided.

State space: the state space is defined as a one dimensional matrix based on the system's wireless channel resources, i.e., $S(t) = \left(h_u^1(t), h_u^2(t), \cdots, h_u^A(t)\right)$. where $h_u^i(t)$ represents the channel quality state at moment t between the smart body and the i shipboard base station.

The reward function $R_u(t)$ measures the utility of a unit of resource and is calculated as the ratio of revenue generated from the provision of services to expenditures related to the rental of communications resources. A higher reward value indicates a more efficient use of resources. Define the reward function as follows:

$$\begin{aligned} R_u(t) &= \sum_{a \in \mathcal{A}} R_{u,a}^{Comm}(t) = \sum_{a \in \mathcal{A}} ComA_u^a(t) \left(\frac{\tau_u ComR_u^a(t)}{\delta_a B_u^a(t)} \right) \\ &= \sum_{a \in \mathcal{A}} ComA_u^a(t) \left(\frac{\tau_u B_u^a(t) \nu_u^a(t)}{\delta_a B_u^a(t)} \right) = \sum_{a \in \mathcal{A}} ComA_u^a(t) \left(\frac{\tau_u \nu_u^a(t)}{\delta_a} \right). \end{aligned} \tag{18}$$

where τ_u denotes the cost per unit of resource in communication, δ_a denotes the utility of the system for the shipboard base station to access the network service in (\$/Hz), and denotes the transmission rate of user u at time t.

At this point, we have finished introducing the overall idea of the SAC-OSCPA algorithm, and its specific algorithmic flow is as follows:

Algorithm 1. Proposed SAC-OSCPA.

Input: $Q_{w_1}(s,a)$, $Q_{w_2}(s,a)$, $\pi_\theta(s)$, w_1, w_2, θ
Initialization: $w_1^- \leftarrow w_1$, $w_2^- \leftarrow w_2$, $Q_{w_1^-}$, $Q_{w_2^-}$
for list $e = 1 \rightarrow E$ **do**
  Get the initial state s_1 from the environment
  for t $= 1 \rightarrow T$ **do**
    Select action $a_t = \pi_\theta(s_t)$ based on current policy
    Execute a_t, obtain r_t, $s_t \rightarrow s_{t+1}$
    Put $(s_t, a_t, \mathrm{r}_t, s_{t+1})$ in to playback pool R
    for number of training rounds $k = 1 \rightarrow K$ **do**
      Sample N tuples from R $\{(s_i, a_i, r_i, s_{i+1})\}_{i=1,\ldots,N}$
      For each tuple, calculate with the target network
      $y_i = r_i + \gamma \min_{j=1,2} Q_{w_i^-}(s_{i+1}, a_{i+1}) - \alpha \log \pi_\theta(a_{i+1} \mid s_{i+1})$,
      among $a_{i+1} \sim \pi_\theta(\cdot \mid s_{i+1})$
      Update the two Critic networks:
      for j=1,2, minimization the loss function:
      $L = \frac{1}{N} \sum_{i=1}^{N} (\mathrm{y}_i - Q_{w_j}(s_i, a_i))^2$
      Update $\pi_\theta(s)$ with the loss function:
      $$L_\pi(\theta) = \frac{1}{N} \sum_{i=1}^{N} \left(\alpha \log \pi_\theta(\tilde{a}_i \mid s_i) - \min_{j=1,2} Q_{w_j}(s_i, \tilde{a}_i) \right)$$
      Update entropy regular term coefficient α
      Update target network: $w_1^- \leftarrow r w_1 + (1-\tau) w_1^- w_2^- \leftarrow \tau w_2 + (1-\tau) w_2^-$
    end for
  end for
end for

4 Simulation Result

In this section, simulation results are provided to verify the effectiveness of the proposed SAC-OSCPA algorithm, and the conventional DQN is selected as the benchmark algorithm for comparison. In the simulation, the users are randomly distributed in the service area, and the UAV is deployed near the service area boundary at an initial moment with a height of 100 m. One LEO, three UAVs and six users are set up. Each UAV has 7 actions and each user has 3 power levels. The neural network used has 3 layers and 40 nodes in the hidden layer. The greedy behavior strategy is used and the Adam optimizer is used to train the neural network. It can be found that SAC-OSCPA has excellent convergence performance in maritime communication network environments, and the throughput gain curve is more stable than that of DQN.

As shown in the NOMA-DQN and OMA-DQN curves in Fig. 2, the system throughput of the NOMA framework is slightly lower than that of the OMA in

the first 50 frames, and the system throughputs of the NOMA framework are basically the same from 60 frames to 80 frames, and the performance of the NOMA framework is significantly better than that of the OMA framework in the first 80 frames, where the system peak throughput of the NOMA framework is 4264.04 MB higher than that of the OMA framework. After 80 frames, the performance of NOMA is significantly better than that of OMA. The peak throughput of NOMA is 4264.04 MB higher than that of OMA, which is 40.27% optimized compared to OMA, and the average throughput of NOMA is 7.75% higher than that of OMA. Therefore, from Fig. 2, it can be seen that NOMA performs better than OMA. This is because the number of users supported by NOMA is not strictly limited by orthogonal time frequency resources, and different users can reuse both time and frequency domain resources. In addition, as shown in Fig. 2, the SAC-OSCPA algorithm proposed in this paper improves the average throughput of the system by 13.18% compared with DQN under the NOMA framework.

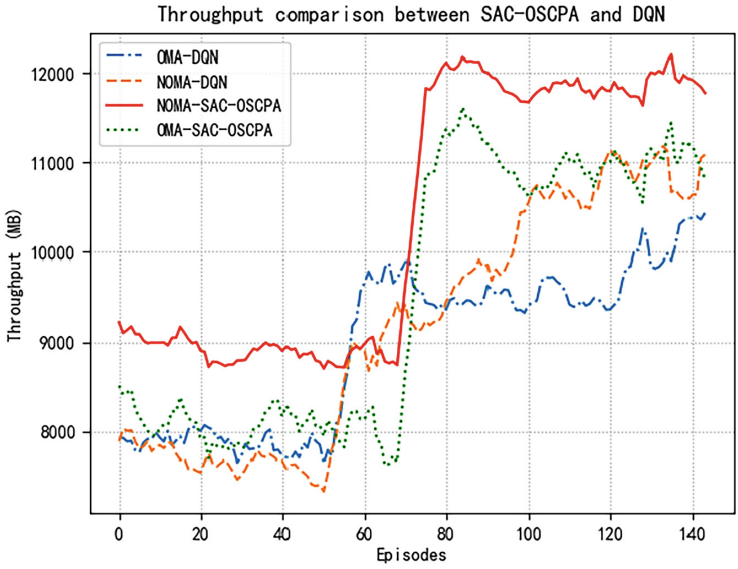

Fig. 2. Throughput comparison between SAC-OSCPA and DQN.

As shown in Fig. 3, the SAC-OSCPA algorithm proposed in this paper also outperforms DQN in terms of worst user throughput. The average throughput of worst user is improved by 41.59% compared to DQN. In Fig. 3, DQN slightly outperforms SAC-OSCPA at the beginning due to the higher complexity of SAC-OSCPA compared to DQN and the fact that SAC-OSCPA is more conservative in taking actions upfront. Interestingly, we can observe that the worst user throughput of DQN drops sharply at 60 episodes and then shows an

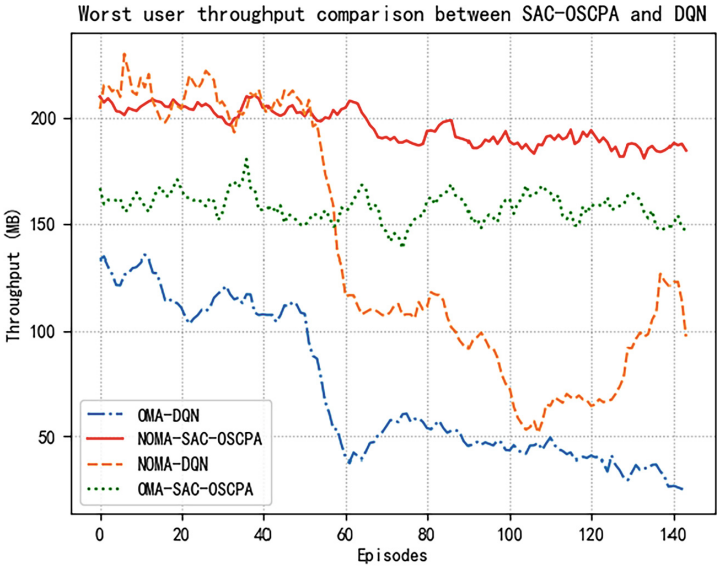

Fig. 3. Worst user throughput comparison between SAC-OSCPA and DQN.

upward trend. The reason for this phenomenon could be that the DQN generates an overestimation of the Q value and thus adopts a wrong strategy. The performance of the DQN improved again after it was trained more adequately afterwards.

In addition, since the method can avoid the problem of overestimating the Q value of DQN, it is more stable during the training process as can be seen in Fig. 2 and Fig. 3. As can be seen from Fig. 2 and Fig. 3, the SAC-OSCPA algorithm is superior to the DQN algorithm. This superiority can be attributed to the use of entropy regularization and reparameterization techniques in the SAC-OSCPA algorithm, which prioritizes the optimization of cumulative rewards while enhancing the stochastic nature of the policy, thus reducing the risk of falling into local optimums.

5 Conclusions

Aiming to serve diverse users in a flexible manner, NOMA technology has been implemented to adjust to the vast sparsity of marine networks. This paper examines a hybrid network of marine satellite UAV sea surface networks based on NOMA. We offer a power distribution strategy to mitigate the difficult interference between various users, groups, and network segments. Firstly, based on this architecture, an optimization problem for communication network resources was designed. Next, we use deep reinforcement learning techniques to address the aforementioned optimization difficulties in order to tackle the non-convex problem. A maximum entropy deep reinforcement learning power allocation algo-

rithm (SAC-OSCPA) has been proposed. Finally, we conducted a comparative analysis between the SAC-OSCPA algorithm proposed in this paper and the DQN algorithm, demonstrating the effectiveness and reliability of the SAC-OSCPA algorithm. The outcomes of the simulation indicate that the NOMA based sea satellite UAV sea hybrid network has the potential to improve the throughput of sea communication systems.

References

1. Fourati, F., Alouini, M.-S.: Artificial intelligence for satellite communication: a review. Intell. Converged Netw. **2**(3), 213–243 (2021)
2. Hassan, S.S., Kim, D.H., Tun, Y.K., Tran, N.H., Saad, W., Hong, C.S.: Seamless and energy efficient maritime coverage in coordinated 6G space-air-sea non-terrestrial networks. IEEE Internet Things J **10**, 4749–4769 (2022)
3. Zhang, L., Liang, Y.-C., Niyato, D.: 6g visions: mobile ultrabroadband, super internet-of-things, and artificial intelligence. China Commun. **16**(8), 1–14 (2019)
4. Salman Hassan, S., Park, S.-B., Huh, E.-N., Seon Hong, C.: Seamless and intelligent resource allocation in 6G maritime networks framework via deep reinforcement learning. In: 2023 International Conference on Information Networking (ICOIN), Bangkok, Thailand, 2023, pp. 505–510 (2023)
5. Xu, F., Yang, F., Zhao, C., Wu, S.: Deep reinforcement learning based joint edge resource management in maritime network. China Commun. **17**(5), 211–222 (2020)
6. Li, Y., Su, L., Wei, T., Zhou, Z., Ge, N.: Location-aware dynamic beam scheduling for maritime communication systems. In: 2018 10th International Conference on Communications, Circuits and Systems (ICCCAS), Chengdu, China, 2018, pp. 265–268 (2018)
7. Wei, T., Feng, W., Wang, J., Ge, N., Lu, J.: Exploiting the shipping lane information for energy-efficient maritime communications. IEEE Trans. Veh. Technol. **68**(7), 7204–7208 (2019)
8. Fang, X., et al.: NOMA-based hybrid satellite-UAV-terrestrial networks for 6G maritime coverage. IEEE Trans. Wirel. Commun. **22**(1), 138–152 (2023)
9. Ding, Z., Adachi, F., Poor, H.V.: The application of MIMO to non-orthogonal multiple access. IEEE Trans. Wirel. Commun. **15**(1), 537–552 (2016)
10. Pachler, N., Luis, J.J.G., Guerster, M., Crawley, E., Cameron, B.: Allocating power and bandwidth in multibeam satellite systems using particle swarm optimization. In: IEEE Aerospace Conference. Big Sky, MT, USA 2020, pp. 1–11 (2020)
11. Paris, A., Del Portillo, I., Cameron, B., Crawley, E.: A genetic algorithm for joint power and bandwidth allocation in multibeam satellite systems. In: IEEE Aerospace Conference. Big Sky, MT, USA 2019, pp. 1–15 (2019)
12. Tsuchida, H., et al.: Efficient power control for satellite-borne batteries using Q-learning in low-earth-orbit satellite constellations. IEEE Wirel. Commun. Lett. **9**(6), 809–812 (2020)
13. Zhao, B., Dong, X., Ren, G., Liu, J.: Optimal user pairing and power allocation in 5G satellite random access networks. IEEE Trans. Wireless Commun. **21**(6), 4085–4097 (2022)
14. Mnih, V., Kavukcuoglu, K., Silver, D. et al.: Human-level control through deep reinforcement learning. Nature **518**, 529–533 (2015). https://doi.org/10.1038/nature14236

15. De Santis, E., Giuseppi, A., Pietrabissa, A., Capponi, M., Delli Priscoli, F.: Satellite integration into 5G: deep reinforcement learning for network selection. Mach. Intell. Res. **19**(2), 127–137 (2022)
16. He, Y., Sheng, B., Yin, H., Yan, D., Zhang, Y.: Multi-objective deep reinforcement learning based time-frequency resource allocation for multi-beam satellite communications. China Commun. **19**(1), 77–91 (2022)
17. Zhong, R., Liu, X., Liu, Y., Chen, Y.: Multi-agent reinforcement learning in NOMA-aided UAV networks for cellular offloading. IEEE Trans. Wireless Commun. **21**(3), 1498–1512 (2022)
18. Duan, J., Guan, Y., Li, S.E., Ren, Y., Sun, Q., Cheng, B.: Distributional soft actor-critic: off-policy reinforcement learning for addressing value estimation errors. IEEE Trans. Neural Netw. Learn. Syst. **33**(11), 6584–6598 (2022)