2022 International Symposium on Nonlinear Theory and Its Applications,
NOLTA2022, Virtual, December 12-15, 2022

An Application of Reinforcement Learning to Ground Station Selection in Satellite-Terrestrial Optical Communication

Keigo Makizoe† Atsuhiro Yumoto† Koji Oshima‡ Kenji Suzuki§ and Mikio Hasegawa†

†Graduate School of Engineering, Tokyo University of Science
6-3-1 Niijuku, Katsushika-ku, Tokyo 125-8585, Japan
‡Innovation Design Initiative, National Institute of Information and Communications Technology
4-2-1, Nukui-Kitamachi, Koganei, Tokyo 184-8795, Japan
§Space Communication Systems Laboratory, National Institute of Information and Communications Technology
4-2-1, Nukui-Kitamachi, Koganei, Tokyo 184-8795, Japan
Email: 4322546@ed.tus.ac.jp, 4321559@ed.tus.ac.jp, koji.oshima.jp@ieee.org,
bt_kenji@nict.go.jp, hasegawa@ee.kagu.tus.ac.jp

Abstract–Optical satellite communications enable high-capacity communications, one of the fundamental technologies for a non-terrestrial network in Beyond 5G/6G. It is affected by the interruption of optical communications due to clouds on the communication path. A satellite can mitigate the interruption by switching its destination ground station to another, though it brings additional delays in acquiring the beam. In this study, we propose a ground station selection method using a reinforcement learning algorithm to realize a fast and stable satellite-terrestrial optical communication system. We show its effectiveness through simulation evaluation using pseudo and real data.

1. Introduction

In Beyond 5G/6G, Non-Terrestrial Networks (NTN) that connect space to the ground in three dimensions are expected to be realized [1]. Optical satellite communications [2], which enable broadband and high-capacity communications, have been attracting attention as one of the fundamental technologies for NTNs.

Optical satellite communication is a technology that uses wireless light instead of radio waves to communicate with satellites. Compared with radio frequency (RF) communications, which is conventional wireless communication technology, optical satellite communications have features such as extremely high bandwidth, ease of introduction, license-free frequency allocation, low power consumption (~1/2 that of RF), small size (~1/10 of RF antenna diameter), and improved channel security [2].

In optical satellite communications, when clouds exist on the communication path between a satellite and a ground station, optical communications are blocked by the clouds [2]. Switching the ground station to communicate with can recover the communication in such a situation. However, it brings an additional delay in the acquisition of the beam [3]. Therefore, it is necessary to select an appropriate ground station while avoiding redundant switching. In related research, site diversity, which establishes optical satellite communication lines by combining multiple optical ground stations, has been studied[4]-[7]. In [4], an optical ground station network is optimized by using the correlation of weather conditions among optical ground stations. In [5], the authors assume a downlink optical satellite communication system that aims to communicate with the best optical ground station among multiple sites. In this system, they select and communicate with the ground station with the highest SNR by using a channel model that takes into account fading due to turbulence and atmospheric attenuation due to scattering. In [6], the effect of clouds on free-space optical communication is quantitatively analyzed to determine the appropriate placement of ground stations. In [7], the availability of each optical ground station is determined from environmental data collected in satellite-to-ground optical communications, and efficient site diversity is discussed. These studies focused on the placement of optical ground stations. However, they do not discuss how to autonomously select optical ground stations in response to changing channel conditions.

In this study, we apply a reinforcement learning algorithm to a satellite-to-ground optical communication system and propose a method that a satellite selects ground stations against cloud effects autonomously. We design a reinforcement learning algorithm using channel quality as a state, ground station selection as an action, and throughput as a reward. The performance of the proposed method is evaluated by using real data collected by the Observation System of the Patch of Blue Sky for Optical Communication (OBSOC) [8].

2. System Model

Free-space optical communication requires a LOS connection between the transmitter and receiver. Here, the

ORCID iDs Keigo Makizoe: 0000-0001-9745-8496,
Atuhiro Yumoto: 0000-0001-8448-0627
Koji Oshima: 0000-0001-6878-4794
Kenji Suzuki: 0000-0003-0009-3043
Mikio Hasegawa: 0000-0001-5638-8022

information signal from the transmitter is modulated on an optical carrier, and this modulated signal is propagated through an atmospheric channel or free space toward the receiver [3]. Therefore, the weather affects such optical communications between satellites and the ground.

NICT has deployed an environmental data collection system, OBSOC, at 10 locations [9]. In this system, environmental data acquired at each observation station is stored in a database and analyzed. Each weather station is equipped with an all-sky camera to identify clear areas, a cloud cover/cloud height meter to measure the amount and height of clouds, and various meteorological data. The data collected by these instruments are used to determine whether optical communication is possible, and the data is stored as shown in Figure 1, with 1 indicating that optical communication is possible and 0 indicating that it is not possible.

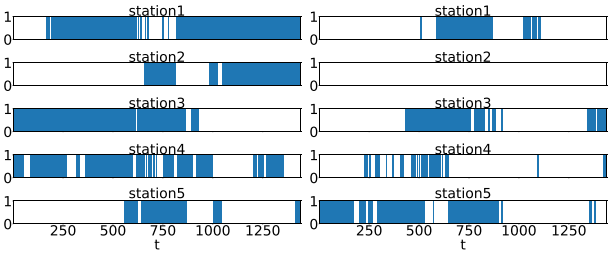


Fig.1. Real data: Channel state.
(communication available: blue, not available: white)

In this study, we assume a downlink satellite-to-ground optical communication system that transmits data from one satellite to K optical ground stations $i = \{1,2,\cdots K\}$ as shown in Figure 2. The quality of the link is modeled as 1 if communication is possible and 0 if communication is not possible. The satellite selects a ground station at every time $t = \{1,2,\cdots T\}$. If there is a cloud between the satellite and the ground station, the optical communication is interrupted. Delay occurs when switching the ground station to be communicated with. Throughput needs to be maximized for fast and stable communication and defined below,

$$\text{Throughput} = M_{total}/\Delta t \quad (1)$$

where M_{total} is the total amount of data received at the ground stations during Δt.

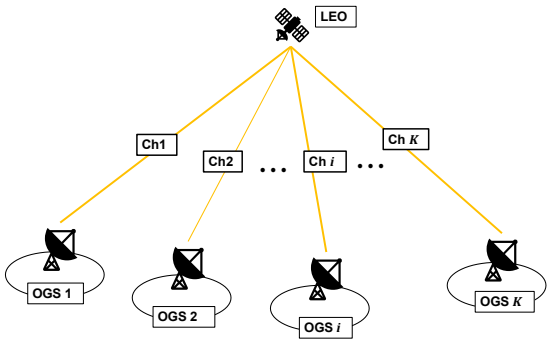


Fig.2. System model.

3. Ground Station Selection Using RL algorithms

We propose a ground station selection method using a reinforcement learning algorithm. The flowchart of the inter-ground station selection is shown in Figure 3. The reward r_i for selecting the ground station i is the throughput at a certain time l [min],

$$r_i(t) = M_l/l \quad (2)$$

where M_l is the amount of data received by all ground stations during l. We examine proposed method in four reinforcement learning algorithms described in the following subsections.

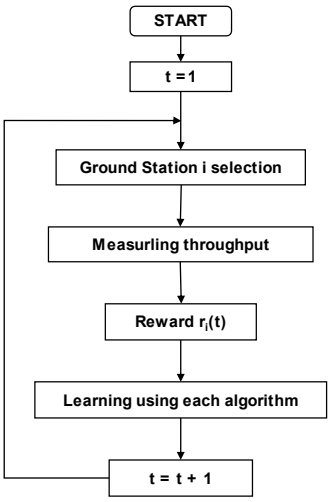


Fig.3. Flowchart of ground station selection.

3.1. ε-greedy

The ε-greedy algorithm is an action selection method that selects one of all actions with probability ε (exploration) and the action with the greatest action value with probability 1-ε (exploitation). The action value is the expected value of the reward for selecting each action. It is used in reinforcement learning to balance exploration and exploitation. It is a simple algorithm and is widely used for solving MAB problems. The ground station to be selected at time t is denoted as

$$a(t) = \begin{cases} \underset{i=1,\ldots,K}{\text{argmax}}(\bar{r}_i(t)) & with\ probability\ 1-\varepsilon \\ random\ select & with\ probability\ \varepsilon \end{cases} \quad (3)$$

where $\bar{r}_i(t)$ denotes the average reward up to t.

3.2. UCB1-Tuned

UCB1-Tuned [9] performs ground station selection by considering the reward probability and confidence interval. The ε-greedy algorithm, which is considered the best performing of the current MAB algorithms, fails to consider the number of times each ground station is selected. In UCB1-Tuned, each ground station is initially selected, and then the ground station $a(t)$ is selected on the t th trial according to the following equation. In UCB1-Tuned, the

ground station is selected on each trial according to the following equation,

$$a(t) = \underset{i=1,\ldots,K}{\operatorname{argmax}} \left(\bar{r}_i(t) + \sqrt{\frac{\ln t}{n_i} \min\left(\frac{1}{4},\ V_i(n_i)\right)} \right) \quad (4)$$

where $\bar{r}_i(t)$ denotes the average reward up to t, min() is a function that adopts the lower value, n_i means the number of times each ground station is selected. $V_i(n_i)$ is expressed by the following equation using the estimated variance.

$$V_i(n_i) = (\hat{r}_i(t))^2 + \sqrt{\frac{2 \ln t}{n_i}} \quad (5)$$

where $\hat{r}_i$ denotes the variance value of the acquired reward up to t.

3.3. Q-Learning

Q-Learning is a type of reinforcement learning, a machine learning algorithm that seeks a strategy to maximize the reward for an agent placed in a certain environment. The past channel quality is the state, and the ground station selection is the action. The equation for updating the Q-value is given below.

$$Q(s,a) \leftarrow Q(s,a) + \alpha\big(r_i(t) + \gamma maxQ(s',a') - Q(s,a)\big) \quad (6)$$

where $Q(s,a)$ is a function that stores the evaluation of action a in state s. α is a parameter that adjusts the amount of modification of the value of the Q function and takes the value $(0 < \alpha < 1)$. Theγ is a parameter that determines how much the expectation of the next state evaluation is considered when updating the Q function and takes the value $(0 < \gamma < 1)$. In Q-Learning, the action value function, which determines the evaluation value of an action in a given state, is defined as a function of TD (temporal difference) error. In Q-Learning, the action value function, which determines the evaluation value of an action in a certain state, is updated so that the TD (temporal difference) error becomes 0. That is, the number of action values is updated so that the value of the Q function, $Q(a)$, is the sum of $r_i(t)$ and the maximum possible value of an action $\gamma maxQ(s',a')$ in the next state s. In this study, past states from time t to j time step before are treated as state s. Table 1 shows an example of a channel state for $K = 5$, where δ is a parameter that determines how much of the past state is referenced.

Table.1. An example of channel state. $(K = 5, j = 5)$

	Ch1	Ch2	Ch3	Ch4	Ch5
$t-5\delta$	1	1	0	1	0
$t-4\delta$	1	1	1	0	0
$t-3\delta$	1	0	0	1	1
$t-2\delta$	1	1	1	0	0
$t-\delta$	1	1	0	0	0
t	?	?	?	?	?

3.4. Deep Q-Network (DQN)

DQN is an algorithm that replaces the Q-table in Q-Learning with a neural network. In the target system, the Q-table is so large that it is difficult to obtain an optimal solution. The satellite, which is an agent, inputs the state of the environment to the neural network, selects from the value of each action outputted, and obtains a reward. In DQN, a neural network w is formed from $Q(s,a)$. In this case, the weight of the neural network is set to w, and $Q(s,a,w) \approx Q(s,a)$. The deep network is trained to minimize the error function. The error function $L(w)$ is expressed by the following equation [10].

$$L(w) = E\left[\big(r(t) + \gamma maxQ(s',a') - Q(s,a)\big)^2\right] \quad (7)$$

Experience Replay is used to ensure stable learning in DQN. The agent's experience at each time step is stored in a dataset, and the deep network is trained on a mini-batch of experiences drawn uniformly at random from the stored samples. We also perform a fixed target deep network; when updating the Q-network, we update it using the network from a certain time step ago. Here, the value of past actions is learned as the target.

4. Evaluation by Simulation

In this evaluation, the effectiveness of the proposed method is verified using pseudo data and real data. As real data, we used the channel state of the optical communication availability judgment at five locations observed and published by OBSOC [8], and used the data collected every minute. In addition, pseudo data was created by imitating the real data. The pseudo data are the channel states (available for communication: 1, not available: 0) that are changed according to the transition probability. As shown in Figure 4, p is the probability of transitioning from 0 to 1 and q is the probability of transitioning from 1 to 0. Table 2 shows the parameters used in the simulations.

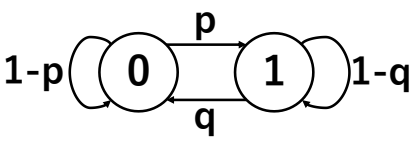


Fig.4. Pseudo data: Channel state.

Table.2. Parameters used in the simulations.

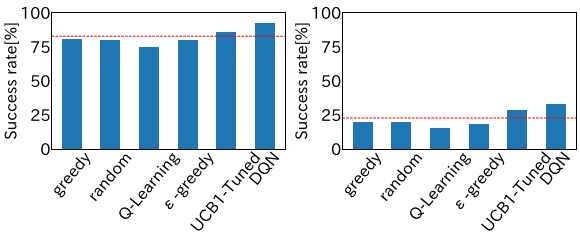
Parameters	Value
Number of ground stations K	5
Number of time step referenced j	5
Availability rate	0.2, 0.8
Data rate	10Mbps
Acquisition delay	120s
Transition probability p, q	0.034, 0.009

We use ε-greedy, UCB1-Tuned, Q-Learning, and DQN as ground station selection methods based on reinforcement learning algorithms, and as comparison methods, we use random selection and greedy. The value for communicating with the ground station with the highest probability of successful communication is optimum. The parameters used in the RL algorithm are listed in Table 3.

Table.3. Parameters used in the RL algorithm.

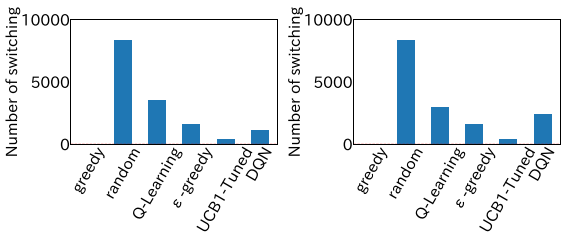
Parameters	Value
l	5 min
δ	11
ε	0.001
γ	0.9

The ratio of time when communication is possible to when communication is not possible is called availability rate, and simulations were performed for availability rate =0.8 and availability rate =0.2. Figure 5 shows the communication success rate, Figure 6 shows the number of switching cycles, and Figure 7 shows the throughput for each method. The dashed line means optimum. It is indicated that DQN can improve the throughput compared to optimum. DQN shows higher throughput than the other methods at both high and low Availability rate, indicating that DQN is able to cope with changes in the environment.



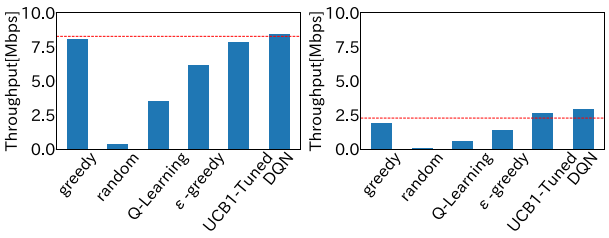
(a)availability rate:0.8 (b)availability rate:0.2

Fig.5. Pseudo data: Communication success rate.



(a)availability rate:0.8 (b)availability rate:0.2

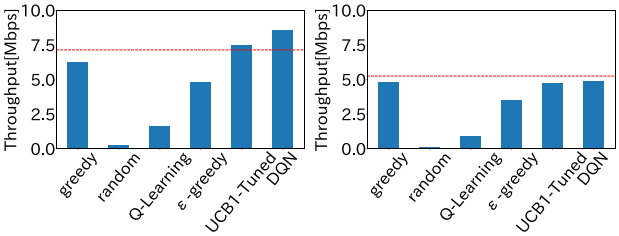
Fig.6. Pseudo data: Number of switching.



(a)availability rate:0.8 (b)availability rate:0.2

Fig.7. Pseudo data: Throughput.

Figure 8 shows the simulation results using real data. The dashed line means optimum. The results show that DQN has the highest throughput in both (a) and (b). In particular, DQN greatly improves the throughput when the availability rate is high as in (a). It is indicated that the reinforcement learning algorithm is effective in selecting ground stations in real environments.



(a) (b)

Fig.8. Real data: Throughput.

5. Conclusions

In this study, we have proposed a ground station selection method based on a reinforcement learning algorithm for high-speed and stable satellite-to-ground optical communications. We have evaluated the proposed scheme using real data collected by OBSOC. We have demonstrated that the throughput can be improved by the proposed scheme that uses the reinforcement learning algorithm for ground station selection. In particular, the proposed scheme using the DQN shows high throughput in various environments simulated based on real data.

References

[1] F. Rinaldi et al., IEEE Access, 8, 165178-165200, 2020.
[2] H. Kaushal and G. Kaddoum, IEEE Communications Surveys & Tutorials, 19, 1, 57-96, 2017,
[3] Hemani Kaushal et al., Free Space Optical Communication, Springer New Delhi,2017
[4] C. Fuchs and F. Moll, JOCN, 7, 12, 1148-1159, 2015.
[5] Randall J. Alliss and Billy Felton, ICSOS, 3-4, 2012
[6] D. R. Kolev et al., PIERS-Toyama, 1680-1685, 2018.
[7] E. Erdogan et al., IEEE Access, 9, 31179-31190, 2021
[8] NICT, OBSOC- Latest Information. Retrieved March 24, 2022, from https://sstg.nict.go.jp/OBSOC/
[9] P. Auer et al., Machine Learning, 47, 235–256, 2002.
[10] Volodymyr Mnih et al., Nature, 518, 529–533, 2015.