Deep Reinforcement Learning Approach for Age of Information Minimization in Multi-Beam LEO Satellite Networks

Khai Doan, Sangmin Han, and Wonjae Shin

School of Electrical Engineering, Korea University, Seoul 02841, Republic of Korea

Emails: {khaidoan, smhan22, wjshin}@korea.ac.kr

Abstract

Multi-beam low earth orbit satellite systems offer a promising solution for ensuring data freshness in time-sensitive applications by directing data beams to various locations. In this work, we propose a beam hopping illumination plan driven by a Deep Reinforcement Learning (DRL) framework for age of information minimization. Due to large state and action spaces, the DRL model's architecture can become extensive. To address this problem, we propose a solution where the DRL model is trained with respect to only a subset of action space. The considered action subset is dynamically optimized during training. Simulation results indicate that our framework performs close to the optimal solutions in ideal scenarios when users' preferences are known.

I. Introduction

The dynamic channel conditions, especially in Low Earth Orbit (LEO) satellite networks, pose challenges in maintaining up-to-date information at receivers [1]-[3]. In ensuring desirable Age of Information (AoI), designing a Beam Hopping Illumination Plan (BHIP) driven by Deep Reinforcement Learning (DRL) appears as a potential and interesting approach. However, infinitely large state and action spaces may require an extensive number of parameters in the DRL model to capture the relationship in training data. To address the issue, we propose a DRL framework in which the training focuses only on a subset of actions. The subset is continuously updated during training, guiding the DRL model towards the optimal solution. We conduct simulations to examine the proposed framework.

The rest of this work is organized as follows. In Section II, we present the system model under consideration. The proposed BHIP framework is outlined in Section III. Subsequently, Section IV presents numerical results examining the proposed method. Finally, we conclude our work in Section V.

II. System Model

Our system model operates in a slotted time horizon. In each time slot, a satellite passes by an area with U mobile users are locating. Each satellite carries I different types of information. The AoI of information i at user u in a time slot t is denoted by $A_{ui}(t) \geq 0$. If t' is the latest time slot that information i is updated at user u, $A_{ui}(t) = t - t'$ for $t > t'$. Users have different request probabilities toward information types.

Satellite can illuminate $B \geq 1$ beams where beams share the same bandwidth and each beam carries a single information type. We divide a time slots into L subslots. Beam illuminated regions can be switched between subslots. We denote by $p_b(l,t)$ the power allocated to beam b in subslot l of time slot t which will be referred to as $beam(b,l,t)$ for the rest of this article. Let $P_{\max}$ be the maximum transmission power which implies that $\sum_{b=1}^{B} p_b(l,t) \leq P_{\max}$. A user u can decode an information i from $beam(b,l,t)$ if its SINR is at least a predefined threshold δ, i.e.,

$$\delta_{ui}^{b}(l,t) = \frac{\sum_{b\in\mathbf{B}_i(l,t)} |h_{ub}(l,t)|^2 p_b(l,t)}{\sum_{b'\in B_j(l,t), j\neq i} |h_{ub'}(l,t)|^2 p_{b'}(l,t) + \sigma^2} \geq \delta,$$

where $\mathbf{B}_i(l,t)$ is the set of beams at subslot l of time slot t that carry information i. σ^2 is the noise power. $h_{ub}(l,t) = G(\theta_{ub}(l,t))\mathrm{PL}_u G_u$ denotes the channel gain at user u with respect to $beam(b,l,t)$. The angle $\theta_{ub}(l,t)$ is between the center axis of $beam(b,l,t)$ and the line connecting the satellite and user u. $G(\theta_{ub}(l,t))$ is the transmitting antenna gain. PL_u is the pathloss component. Readers may refer to [3, Eq. (13)] for the expressions of $G(\theta_{ub}(l,t))$ and PL_u. G_u is the receiving antenna gain of user u. Let us define the set $\mathbf{M}_{ui}(t) = \{(b,l)|\delta_{ui}^{b}(l,t) \geq \delta\}$. The maximum amount of data regarding information i that the satellite can transfer to user u at time slot t is computed by:

$$R_{ui}(t) = \sum_{l=1}^{L} \sum_{(b,l)\in\mathbf{M}_{ui}(t)} \tau W \log_2\left(1 + \delta_{ui}^{b}(l,t)\right).$$

τ and W are the subslot duration and available bandwidth, respectively. Besides, the update of information i at user u is completed when user u sufficiently receives data of size $S_i^{\max}$.

III. Beam Hopping Illumination Plan via Deep Reinforcement Learning

We define system states, beam hopping actions and system costs as follows:

$\mathbf{s}(t) = \{(A_{ui}(t), S_{ui}(t))|u = 1,\dots,U; i = 1,\dots,I\}$,

$\mathbf{a}(t) = \{(x_{bl}(t), y_{bl}(t), i_{bl}(t), p_{bl}(t))|b = 1,\dots,B; l = 1,\dots,L\}$,

$C(\mathbf{s}(t),\mathbf{a}(t)) = \sum_{u=1}^{U}\sum_{i=1}^{I} A_{ui}(t)\mathbf{1}\{R_{ui}(t) < S_{ui}(t)\}X_{ui}(t)$.

$S_{ui}(t)$ is the remaining amount of data for updating information i at user u in time slot t, and computed by:

$$S_{ui}(t) = \begin{cases} R_{ui}(t) - S_{ui}(t), & \text{if } R_{ui}(t) - S_{ui}(t) < 0, \\ S_i^{\max} & \text{, if } R_{ui}(t) - S_{ui}(t) \geq 0. \end{cases}$$

$x_{bl}(t)$ and $y_{bl}(t)$ are the coordinates of the illuminated region's center; $i_{bl}(t)$ and $p_{bl}(t)$ are the carried information and allocated power, respectively. The

random variable $X_{ui}(t)$ takes value 1 if user u requests information i at time slot t, and 0 otherwise. $\mathbf{1}\{\cdot\}$ is the indicator function returning 1 if the enclosed condition is satisfied, and 0 otherwise. Our reinforcement learning framework is summarized in Algorithm 1.

Algorithm 1 DRL-Driven BHIP

Input:
- A deep neural network whose weight set is denoted by $\mathbf{W}$.
- A subset $\mathbf{D}$ of actions that are initially selected uniformly randomly.
- A replay buffer $\mathbf{R}$ which is initially empty.

Output: A trained weight set $\mathbf{W}$.

Initialization: $t \leftarrow 1$, $\bar{C}_T \leftarrow +\infty$, $\bar{C}'_T \leftarrow +\infty$ ▷ *$\bar{C}_T$ is the average cost over the most recent T time slots. $\bar{C}'_T$ serves as copied version of $\bar{C}_T$.*

Form system state $\mathbf{s}(t)$.

Select action $\mathbf{a}(t) \in \mathbf{D}$ uniformly randomly.

Obtain $C(\mathbf{s}(t), \mathbf{a}(t))$ and $\mathbf{s}(t+1)$.

Compute state-action value $V(\mathbf{s}(t), \mathbf{a}(t))$:

$$V(t) = C(\mathbf{s}(t), \mathbf{a}(t)) + \gamma \min_{\mathbf{a}'(t)} \hat{C}(\mathbf{s}(t), \mathbf{a}'(t)|\mathbf{W}).$$

▷ *γ is a predefined discount factor; $\hat{C}(\mathbf{s}(t), \mathbf{a}'(t)|\mathbf{W})$ is a predicted cost given the set of weight $\mathbf{W}$.*

$\mathbf{R} \leftarrow \mathbf{R} \cup \{(\mathbf{s}(t), \mathbf{a}(t), V(t))\}$.

Update $\mathbf{W}$ using data in $\mathbf{R}$ with $\mathbf{s}(t)$ and $\mathbf{a}(t)$ as inputs and $V(t)$ as target output.

Compute $\bar{C}'_T$.

If $\bar{C}'_T$ converged do:

  If $\bar{C}'_T > \bar{C}_T$ do:

    Restore the previous sets $\mathbf{D}$ and $\mathbf{W}$.

  Else do:

    $\bar{C}_T \leftarrow \bar{C}'_T$.

  Record the sets $\mathbf{D}$ and $\mathbf{W}$.

  Randomly replace an action in $\mathbf{D}$ by a new one randomly sampled from the action space.

$t \leftarrow t+1$.

Repeat until $\bar{C}_T$ converges.

IV. Numerical Results

We examine the learning model under different number of users: $U = 8$, 11 and 14. Our result is summarized in Fig. 1. The learning model enters a testing session after every 4×10^3 training time slot. In each session, we run a simulation for 10^3 time slots, compute the total cost, and divide the total cost by 10^3 to compute the average cost per time slot as shown on the y-axis. During testing, we use the best weight set $\mathbf{W}$ and action set $\mathbf{D}$ obtained previously. Therefore, two testing sessions would return the same average cost if the $\mathbf{W}$ and $\mathbf{D}$ are not updated during the training period in between. We compare our model with exhaustive search results in ideal cases where the request probability of users on every information type is known.

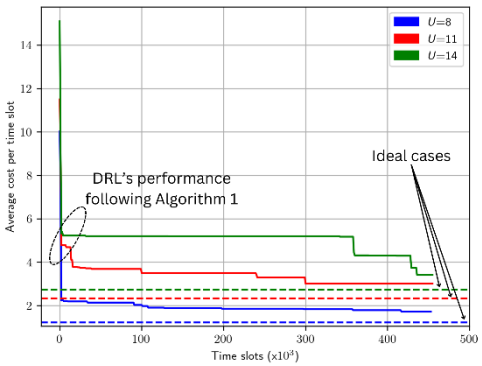

Fig. 1: Average cost per time slot achieved by the proposed BHIP for different number of users, U. The comparison is made against exhaustive search results in ideal cases when the request probability of users on each information type is known.

V. Conclusion

We have presented a DRL-driven BHIP framework to address the AoI minimization problem in a multi-beam LEO satellite system. To overcome the complexity resulted by the infinitely large state and action spaces, we have proposed a method that has restricted the training to a subset of action space. The subset has been dynamically updated during training. Simulation results have demonstrated close gaps between the achieved costs and unreachable bounds obtained by exhaustive searching in ideal scenarios when the request probability of users has been available.

ACKNOWLEDGEMENT

This work was supported in part by the National Research Foundation of Korea (NRF) grants (No.2022R1A2C4002065), in part by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grants (No.2024-00359235, No.2022-0-00704, No.2021-0-00260, and No.2021 0-00467), and in part by the BK21 FOUR program (No.5199991514504).

REFERENCES

[1] Q. Yan, J.Jiao, Y.Wang, L. An, R. Lu, and Q. Zhang, "Age of Information Minimization for Short-Packet Communications RSMA in Satellite-Based IoT," in *Proc. IEEE VTC*, 2023, pp. 1-5.

[2] X. Xia, H. H. Esmat, K. Dyer, B. Lorenzo, and L. Guo, "Cross-Domain Federated Computation Offloading for Age of Information Minimization in Satellite-Airborne-Terrestrial Networks," in *Proc. IEEE PIMRC*, 2023, pp. 1-7.

[3] W. Li, M. Zeng, X. Wang, and Z. Fei, "Dynamic Beam Hopping of Double LEO Multi-Beam Satellite Based on Determinant Point Process," in *Proc. IEEE WCSP*, 2022, pp. 713-718.