Hybrid Policy Learning for Decentralized Uplink Spectrum and Power Management in ISTNs

Seoyeong Park[1], Soyi Jung[2]
[1]*Dept. Aritificial Intelligence Convergence Network, Ajou Univerity, Suwon, 16499, South Korea*
[2]*Dept. Electrical and Computer Engineering, Ajou University, Suwon, 16499, South Korea*
{syjm0819, sjung}@ajou.ac.kr

***Abstract*—Centralized control for uplink (UL) interference is impractical in large-scale integrated satellite-terrestrial networks (ISTNs). We propose a decentralized multi-agent deep reinforcement learning (MADRL) solution where each user equipment (UE), as an agent, learns a hybrid policy for joint channel selection and power control using only local observations. Simulation results validate that our approach significantly improves the system's performance compared to a random baseline, proving its scalability and effectiveness for UL co-existence.**

***Index Terms*—Integrated satellite-terrestrial networks, multi-agent deep reinforcement learning, uplink, decentralized learning**

I. Introduction

The integration of terrestrial network (TN) and non-terrestrial network (NTN) is crucial for future connectivity, but introduces severe uplink (UL) interference from numerous co-channel user equipments (UEs) [1]. The dynamic nature of NTNs makes traditional centralized management impractical due to high signaling overhead and scalability issues [2], [3]. To address this, we propose a decentralized framework where each UE acts as an agent that learns a hybrid policy for joint spectrum selection and power control. Our objective is to maximize the system's average signal-to-interference-plus-noise ratio (SINR) and coverage probability (CP).

II. System Model of ISTN

A. Co-existence scenario case

Fig. 1 shows the co-existence system model considered in this paper. The 3rd generation partnership project (3GPP) has defined six TN-NTN co-existence cases, and the case adopted in this paper is based on the frequency division duplex (FDD) scenario where both the TN and NTN operate in the UL within the S-band [4]. In this scenario, the NTN UL acts as the interference victim, while the TN UL acts as the interference aggressor.

B. Channel modeling

The channel model for this UL scenario is based on the 3GPP TR 38.821 specifications [5]. The received power of the UL signal is formulated as

$$y_{UL} = P_t + G_r + G_t - PL, \tag{1}$$

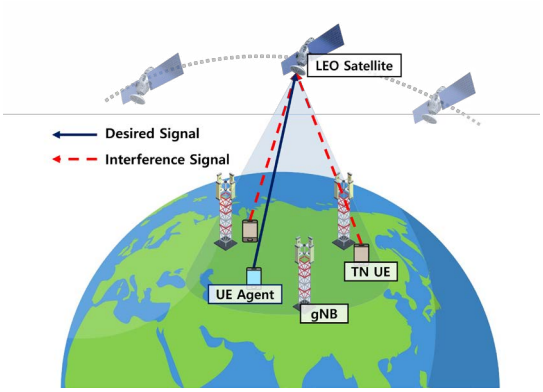

Fig. 1. System model

where P_t is the transmit power of the transmitter, G_r and G_t are the antenna gains of the receiver and transmitter, respectively, and PL denotes the path loss.

The SINR is calculated based on the desired signal power, co-channel and adjacent channel interference, and the noise power. The NTN UL SINR, which includes interference from both TN ULs and other NTN ULs, is calculated as

$$\gamma_{c,e} = \frac{y_{c,e}}{\sum_{e'=1}^{N_{e'}} y_{c,e'} + \sum_{u=1}^{N_u} y_{c,u} + N_0}, \tag{2}$$

where c is the satellite beam that receives the signal, e' denotes other NTN UEs excluding the UE transmitting the desired signal, and $y_{c,e'}$ includes co-channel and adjacent channel interference from the NTN UL. The term of u is a TN UE, and $y_{c,u}$ includes co-channel and adjacent channel interference from the TN UL. N_0 is the noise signal power.

III. Proposed Management Framework

This section details the proposed framework, which is a fully decentralized multi-agent deep reinforcement learning (MADRL)-based management scheme.

A. Proposed hybrid MADRL algorithm

To address the joint spectrum and power management problem in a decentralized manner, we model the system as a multi-agent Markov decision process (MDP). The key components of our MDP formulation are defined as follows:

TABLE I. Environmental setup parameters

Parameter	Value
Altitude of the LEO satellite	600 km
The number of LEO beams and gNBs	19, 20
The number of NTN and TN UEs	20, 200
SINR threshold γ_{th}	1 dB

TABLE II. RL simulation parameters

Parameter	Value
Discount factor γ	0.99
Learning rate α	0.001
Batch size $\mathscr{B}$	128
Buffer size M	10,000

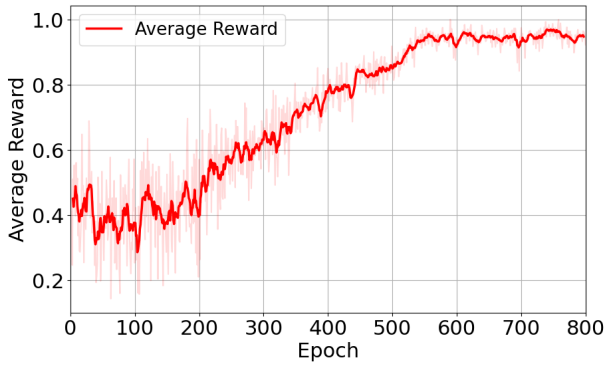

Fig. 2. Reward convergence per training epoch

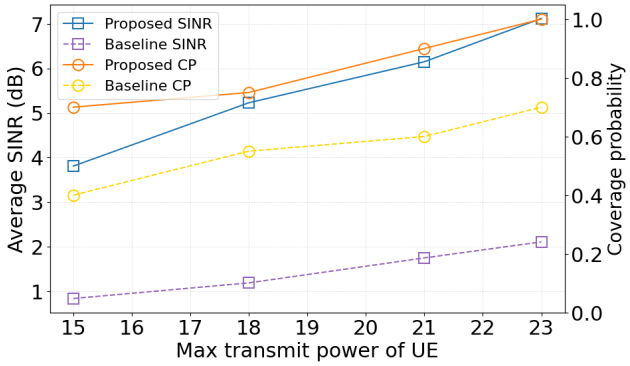

Fig. 3. Average SINR and CP as increasing of UE P_{max}

*1) **State**:* The system state is represented as a *local observation vector* for each UE agent e. The state vector is expressed as

$$s_e(t) = \{U_e(t), C_e(t), R_e(t)\}, \tag{3}$$

where $U_e(t)$ consists of the agent's own information, such as its location, SINR state, and power level information. $C_e(t)$ represents the location information of satellite beam connected with UE e.

*2) **Action**:* Given the observed local state $s_e(t)$, each agent e independently determines its spectrum selection and transmission power at time step t. The overall action of agent e is defined as

$$a_e(t) = \{a_e^S(t), a_e^P(t)\}, \tag{4}$$

where the spectrum selection action $a_e^S(t)$ is performed by the DQN algorithm, and the transmission power control action $a_e^P(t)$ is performed by the DDPG algorithm. $a_e^S(t)$ is represented as a binary variable, and once spectrum allocation is completed, $a_e^P(t)$ is assigned with continuous values.

*3) **Reward**:* The reward function $r(t)$ is designed to align with the objectives of maximizing the average SINR and CP of the overall system. It is formulated as

$$r(t) = w_{\text{SINR}} \cdot \left(\frac{1}{N_c} \sum_{c=1}^{N_c} \gamma_c \right) + w_{\text{cov}} \cdot P_{\text{cov}}, \tag{5}$$

where γ_c is the SINR of satellite beam c, and P_{cov} is the CP defined as

$$P_{\text{cov}} = \frac{1}{N_e} \sum_{e=1}^{N_e} \begin{cases} 1, & \text{if } \gamma_{c,e} \geq \gamma_{th}, \\ 0, & \text{if } \gamma_{c,e} < \gamma_{th}. \end{cases} \tag{6}$$

It represents the proportion of the total UEs that satisfy the QoS requirement.

B. DQN and DDPG framework

This section describes the two core components of the proposed hybrid learning framework.

1) DQN for spectrum selection: For discrete spectrum selection, we use DQN, which approximates Q-values via a neural network. The agent selects the action a_t that maximizes the Q-function $Q(s_t, a; \theta)$. To ensure stable learning in dynamic interference conditions, DQN utilizes experience replay and a target network.

2) DDPG for Power Control: For continuous power control, we adopt DDPG, which employs an actor-critic architecture. The actor network $\mu(s|\theta^\mu)$ determines a deterministic action, while the critic network $Q(s, a|\theta^Q)$ evaluates its value. DDPG uses target networks and adds noise to the actor's output for stable learning and exploration, enabling precise and adaptive power adjustments.

IV. Simulation Results and Analysis

The performance of the proposed algorithm is compared against a baseline that allocates resources randomly. The simulation is conducted in an environment where the satellite moves at each time step. The detailed parameters for the simulation environment are summarized in Table I, and the hyperparameters used for training are provided in Table II.

Fig. 2 illustrates the convergence trend of the average reward according to the training epoch for the proposed algorithm. This convergence pattern demonstrates that the proposed algorithm successfully learns a stable optimal policy that maximizes both the SINR and the CP in the given environment.

Fig. 3 shows the average SINR and CP of the system as a function of the UE's maximum transmit power. While the performance of both algorithms improves with increased available power, the proposed algorithm exhibits superior performance across all power ranges. This demonstrates that the proposed algorithm goes beyond simply increasing transmit power, instead, each UE agent intelligently manages spectrum and power by being aware of the surrounding interference conditions, thereby utilizing network resources far more efficiently.

V. Conclusion

This paper introduced a decentralized MADRL framework for uplink co-existence in ISTNs, where each UE learns a hybrid policy for joint spectrum and power control from local observations. Simulation results validated its superior performance in SINR and CP over a baseline, confirming its scalability without centralized control. Future research will address user mobility to enhance robustness.

References

[1] 3GPP TR 38.811 v16.1.0, "Study on new radio (NR) to support non-terrestrial networks (Release 15)," 3rd Generation Partnership Project (3GPP), Technical Report 38.811, September 2020.

[2] M. Vaezi, A. Azari, S. R. Khosravirad, M. Shirvanimoghaddam, M. M. Azari, D. Chasaki, and P. Popovski, "Cellular, wide-area, and non-terrestrial IoT: A survey on 5G advances and the road toward 6G," *IEEE Communications Surveys Tutorials*, vol. 24, no. 2, pp. 1117–1174, Secondquarter 2022.

[3] Y. Wu, L. Xiao, J. Zhou, M. Feng, P. Xiao, and T. Jiang, "Large-scale MIMO enabled satellite communications: Concepts, technologies, and challenges," *IEEE Communications Magazine*, vol. 62, no. 8, pp. 140–146, August 2024.

[4] 3GPP TR 38.863 v16.1.0, "Solutions for NR to support non-terrestrial networks (NTN): NTN related RF and co-existence aspects (Release 18)," 3rd Generation Partnership Project (3GPP), Technical Report 38.863, June 2024.

[5] 3GPP TR 38.821 v16.1.0, "Solutions for NR to support non-terrestrial networks (NTN) (Release 16)," 3rd Generation Partnership Project (3GPP), Technical Report 38.821, May 2021.