

# Battery Lifetime-Aware Satellite Activation via Branching Dueling DQN for LEO Satellite Constellations

Jiseok Jang<sup>1</sup> and Soyi Jung<sup>2</sup>

<sup>1</sup>Dept. Artificial Intelligence Convergence Network, Ajou University, Suwon, 16499, South Korea

<sup>2</sup>Dept. Electrical and Computer Engineering, Ajou University, Suwon, 16499, South Korea

{star12191254, sjung}@ajou.ac.kr

**Abstract**—Low Earth orbit satellite constellations must satisfy quality-of-service requirements while operating on energy-limited on-board batteries that degrade through repeated charge and discharge cycles. This paper studies a battery-lifetime-aware satellite activation problem that jointly considers per-region quality-of-service satisfaction and battery degradation. The problem is formulated as a constrained optimization over binary activation decisions and battery dynamics, and is then reformulated as a Markov decision process. To address the combinatorial joint action space, we adopt a shared Branching Dueling Deep Q-Network, which decomposes the control decision into satellite-wise binary branches. Simulation results show that the proposed method extends the average battery lifetime while maintaining the quality-of-service satisfaction ratio above the required threshold, and remains effective under off-training transmission-power conditions.

**Index Terms**—LEO satellite, battery lifetime, reinforcement learning, Branching Dueling DQN, quality of service.

## I. INTRODUCTION

Recent advances in low Earth orbit (LEO) satellite constellations have accelerated the deployment of global and low-latency communication services, and the 3rd Generation Partnership Project (3GPP) has standardized supporting frameworks for non-terrestrial networks (NTNs) in New Radio (NR) [1], [2]. However, unlike terrestrial infrastructure, LEO satellites operate under strict on-board energy constraints and repeatedly experience sunlit and eclipse periods along their orbital trajectories. These orbital conditions induce frequent charge and discharge cycles, making battery lifetime a critical factor for reliable and sustainable constellation operation.

A fundamental challenge is the trade-off between service quality and battery preservation. Activating more satellites can increase the served traffic and improve the quality-of-service (QoS) satisfaction of ground regions, but it also increases energy consumption and accelerates battery degradation. Conversely, conservative satellite activation can reduce battery wear but may lead to unmet traffic demand. Therefore, constellation control should jointly consider traffic demand, satellite visibility, battery state, and long-term degradation [3].

Recent studies on LEO constellation operation have mainly focused on networking-level problems, such as energy-efficient

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2026-RS2024-00436887). (Corresponding authors: Soyi Jung)

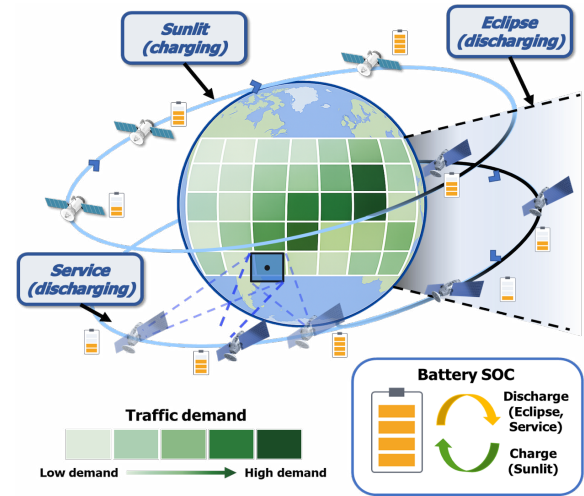


Fig. 1: System model of LEO satellite operation considering traffic demand and battery states.

routing, adaptive routing under dynamic topologies, hierarchical satellite routing, and NTN-TN spectrum coordination [4]–[8]. However, these works primarily address routing or spectrum management, while satellite-level activation control under joint QoS and battery-lifetime objectives remains insufficiently explored.

To address this issue, this paper proposes a battery-lifetime-aware learning-based control framework for LEO satellite constellations. The key contribution is a unified satellite activation formulation that couples per-region QoS requirements with battery degradation dynamics. We formulate the problem as a constrained sequential decision-making problem and recast it as an MDP. To handle the exponentially growing binary action space, we adopt a shared branching dueling deep Q-network (BDQ), which decomposes the joint activation decision into satellite-wise binary sub-actions. Simulation results against representative RL baselines show that the proposed method improves battery lifetime while satisfying the QoS requirement.

## II. SYSTEM MODEL

We consider a LEO satellite constellation that serves multiple ground regions with spatially and temporally varying traffic

demand, as shown in Fig. 1. Since each satellite experiences time-varying battery conditions along its orbital trajectory, the operation model should capture both service requirements and battery-state evolution. In this section, we define the main components of the proposed system model, including the traffic demand model, region-satellite service relationship, satellite control action, QoS model, battery energy dynamics, and battery degradation model.

Let  $\mathcal{R}$  and  $\mathcal{S}$  denote the sets of ground service regions and satellites, respectively. The system is observed over discrete time steps indexed by  $t$ , and  $D_r(t)$  denotes the traffic demand of region  $r \in \mathcal{R}$  at time  $t$ . The set of satellites visible from region  $r$  is denoted by  $\mathcal{S}_r(t) \subseteq \mathcal{S}$ , and satellite  $i$  provides a service capacity  $C_i(t)$  when it is visible to the region.

Each satellite  $i \in \mathcal{S}$  is controlled by a binary action  $x_i(t) \in \{0, 1\}$ , where  $x_i(t) = 1$  indicates active operation and  $x_i(t) = 0$  indicates a low-power sleep state. The joint satellite control action is given by

$$\mathbf{x}(t) = [x_i(t)]_{i \in \mathcal{S}}. \quad (1)$$

Given the visible satellites and their activation states, the served traffic of region  $r$  is modeled as

$$S_r(t) = \min \left\{ D_r(t), \sum_{i \in \mathcal{S}_r(t)} x_i(t) C_i(t) \right\}. \quad (2)$$

This model indicates that the served traffic is limited by both the regional demand and the aggregate capacity of the activated visible satellites.

The QoS satisfaction ratio of region  $r$  is defined as

$$\eta_r(t) = \begin{cases} \frac{S_r(t)}{D_r(t)}, & D_r(t) > 0, \\ 1, & D_r(t) = 0. \end{cases} \quad (3)$$

A minimum QoS requirement  $\eta_{\min}$  is considered to ensure service reliability, which is incorporated into the control objective and constraints in the problem formulation.

Let  $E_i(t)$  denote the battery energy of satellite  $i$  at time  $t$ , where  $0 \leq E_i(t) \leq B_{\max}$  and  $B_{\max}$  is the battery capacity. The energy consumption term of satellite  $i$  depends on its operation mode and is modeled as

$$P_i(t) = P_{\text{slp}} + (P_b + P_{\text{tx}} + P_{\text{rx}})x_i(t), \quad (4)$$

where  $P_{\text{slp}}$  is the sleep-mode consumption term,  $P_b$  is the active base consumption term, and  $P_{\text{tx}}$  and  $P_{\text{rx}}$  are the transmission and reception consumption terms, respectively. These consumption terms are assumed to be constant over the considered horizon.

Let  $z_i(t) \in \{0, 1\}$  indicate whether satellite  $i$  is in the sunlit phase, where  $z_i(t) = 1$  represents the sunlit phase and  $z_i(t) = 0$  represents the eclipse phase. Let  $P_{\text{sol}}$  denote the constant solar charging term. Following the satellite battery energy balance model in [9], the battery energy evolves as

$$E_i(t+1) = \Pi_{[0, B_{\max}]}(E_i(t) + z_i(t)P_{\text{sol}} - P_i(t)), \quad (5)$$

where  $\Pi_{[0, B_{\max}]}(\cdot)$  projects the battery energy onto the feasible range. Here,  $P_{\text{slp}}$ ,  $P_b$ ,  $P_{\text{tx}}$ ,  $P_{\text{rx}}$ , and  $P_{\text{sol}}$  are expressed as energy terms.

The state of charge (SoC) and depth of discharge (DoD) are defined as

$$\rho_i(t) = \frac{E_i(t)}{B_{\max}}, \quad \delta_i(t) = 1 - \rho_i(t). \quad (6)$$

To capture battery aging, we adopt the cycle-based degradation model for satellite-borne batteries [10]. Since LEO satellites repeatedly traverse sunlit and eclipse phases along their orbits, one discharge cycle is naturally identified between consecutive eclipse entries and exits, which avoids the need for offline rainfall counting. In this work, the simulation horizon is set to correspond to one discharge cycle. Let  $\mathcal{T} = \{1, \dots, T\}$  denote the set of time indices in the considered discharge cycle. The peak DoD reached within this cycle is

$$\hat{\delta}_i = \max_{t \in \mathcal{T}} \delta_i(t). \quad (7)$$

In the cited cycle-life model, the allowable number of charge-discharge cycles decreases as the discharge depth grows. Following a linear damage-accumulation principle, the wear of a single cycle is taken as the fraction of total cycle life it consumes and is normalized so that a full-depth ( $\hat{\delta}_i = 1$ ) cycle contributes unit wear. The equivalent wear of satellite  $i$  over the considered discharge cycle is therefore modeled as

$$W_i(T) = 10^{A(\hat{\delta}_i - 1)}, \quad (8)$$

where  $A > 0$  is the DoD-sensitivity coefficient of the degradation model. A full-depth cycle yields  $W_i(T) = 1$ , whereas a shallow cycle ( $\hat{\delta}_i \ll 1$ ) yields exponentially smaller wear, so that  $W_i(T)$  measures the equivalent number of full-depth cycles consumed over the considered discharge cycle.

With  $L_C$  the rated cycle life and  $\tau$  the real-time span of the simulation horizon expressed in years, the battery lifetime of satellite  $i$  is estimated as

$$L_i(T) = \frac{L_C \tau}{W_i(T)}, \quad W_i(T) > 0, \quad (9)$$

that is, the rated cycle life divided by the equivalent wear consumed over the considered discharge cycle and converted to an annual scale. The end-of-horizon value  $L_i(T)$  is the realized battery lifetime, and since  $L_i(T)$  decreases monotonically with the equivalent wear  $W_i(T)$ , extending the lifetime is equivalent to suppressing battery wear.

### III. PROBLEM FORMULATION AND PROPOSED METHOD

This section formulates the satellite control problem as a constrained optimization, discusses the difficulty of solving it directly, and then introduces the proposed RL-based method. The MDP formulation and the shared BDQ policy are described in turn.

#### A. Optimization Problem Formulation

The operational objective is to maximize the average battery lifetime of the constellation while satisfying the per-region QoS requirement:

$$\max_{\{x_i(t)\}} \frac{1}{|\mathcal{T}|} \sum_{i \in \mathcal{S}} L_i(T) \quad (10a)$$

$$\text{s.t. } \eta_r(t) \geq \eta_{\min}, \quad \forall r \in \mathcal{R}, \forall t \in \mathcal{T}. \quad (10b)$$

The decision variables are the binary activations  $x_i(t) \in \{0, 1\}$  defined in (1), and the battery energy evolves according to the update in (5), whose projection  $\Pi_{[0, B_{\max}]}(\cdot)$  inherently enforces the energy bounds  $0 \leq E_i(t) \leq B_{\max}$  while coupling consecutive time slots through the SoC and the accumulated wear.

Problem (10) is a sequential mixed-integer constrained optimization. The time-varying  $D_r(t)$ ,  $\mathcal{S}_r(t)$ , and  $z_i(t)$ , the temporal coupling of  $w_i(t)$  through the battery dynamics, and the combinatorial joint action  $\mathbf{x}(t)$  make a direct solution impractical and motivate a data-driven sequential decision-making approach.

### B. MDP Formulation

The optimization problem in (10) is reformulated as an MDP because each activation decision affects not only the current QoS satisfaction but also the future battery state and degradation. Activating more satellites can improve the served traffic in the current slot, whereas it increases energy consumption and battery wear. Therefore, the controller needs to learn a sequential policy that balances service quality and battery-lifetime preservation.

We define the MDP as  $(\mathcal{X}, \mathcal{A}, P, r, \gamma)$ , where  $\mathcal{X}$  and  $\mathcal{A}$  denote the state and action spaces,  $P$  is the transition probability induced by traffic variation, satellite visibility, orbital illumination, and battery dynamics,  $r$  is the reward function, and  $\gamma$  is the discount factor.

The **state** observed by the controller of region  $r$  at time step  $t$  is defined as

$$s_r(t) = \{D_r(t), \{\rho_i(t), z_i(t), w_i(t)\}_{i \in \mathcal{S}_r(t)}, \bar{\eta}_r(t), t/T\}, \quad (11)$$

where  $D_r(t)$  is the local traffic demand,  $\rho_i(t)$  is the SoC of satellite  $i$ ,  $z_i(t)$  is the sunlit indicator,  $w_i(t)$  represents the battery wear state,  $\bar{\eta}_r(t)$  is the recent QoS satisfaction history, and  $t/T$  is the normalized time index.

The **action** of the controller is the binary activation decision over the satellites visible from region  $r$ , given by

$$a_r(t) = \{x_i(t)\}_{i \in \mathcal{S}_r(t)}, \quad x_i(t) \in \{0, 1\}. \quad (12)$$

Since the joint action space grows as  $2^{|\mathcal{S}_r(t)|}$ , directly estimating a Q-value for every joint action becomes inefficient as the number of visible satellites increases. This motivates the use of a branching action representation, where each branch determines one binary satellite activation decision.

The **reward** is designed to jointly reflect QoS satisfaction, battery degradation, and activation cost:

$$r(t) = -\alpha \sum_{r \in \mathcal{R}} [\eta_{\min} - \eta_r(t)]^+ - \beta \sum_{i \in \mathcal{S}} w_i(t) - \lambda \sum_{i \in \mathcal{S}} x_i(t), \quad (13)$$

where  $[a]^+ \triangleq \max(a, 0)$  and  $\alpha, \beta, \lambda \geq 0$  are weighting coefficients. The first term penalizes QoS shortfall based on (3), the second term accounts for battery wear in (8), and the third term discourages unnecessary satellite activations.

---

### Algorithm 1 Shared BDQ-Based Battery-Lifetime-Aware Control

---

```

1: Initialize online BDQ network  $Q(\cdot; \theta)$  and target network  $Q(\cdot; \bar{\theta})$ 
2: Initialize replay buffer  $\mathcal{D}$ 
3: for episode = 1 to  $N_{\text{episode}}$  do
4:   Reset environment; initialize demand, visibility, and battery states
5:   for time step  $t = 1$  to  $T$  do
6:     Observe regional states  $\{s_r(t)\}_{r \in \mathcal{R}}$ 
7:     for each region  $r \in \mathcal{R}$  do
8:       Select binary sub-actions from BDQ branches using  $\epsilon$ -greedy policy
9:       Form regional action  $a_r(t) = \{x_i(t)\}_{i \in \mathcal{S}_r(t)}$ 
10:    end for
11:    Apply the joint satellite control action  $\mathbf{x}(t)$ 
12:    Update served traffic, QoS satisfaction, battery energy, and wear cost
13:    Compute reward  $r(t)$  and observe next states  $\{s_r(t+1)\}_{r \in \mathcal{R}}$ 
14:    Store transitions  $(s_r(t), a_r(t), r(t), s_r(t+1))$  in  $\mathcal{D}$ 
15:    Update  $\theta$  using a minibatch from  $\mathcal{D}$  with Double-DQN targets
16:    if  $t \bmod K = 0$  then
17:      Update target network  $\bar{\theta} \leftarrow \theta$ 
18:    end if
19:  end for
20:  Decay exploration probability  $\epsilon$ 
21: end for

```

---

### C. Shared Branching Dueling DQN

A conventional DQN estimates Q-values over the entire joint action space. In the considered satellite control problem, the action is a binary activation vector over multiple visible satellites. If  $M_r(t)$  satellites are visible in region  $r$ , the number of possible activation combinations becomes  $2^{M_r(t)}$ . Therefore, directly applying DQN becomes inefficient as the number of visible satellites increases.

The proposed shared BDQ structure is suitable for this problem because the joint activation vector can be decomposed into multiple binary sub-actions. Each branch corresponds to the activation decision of one visible satellite, while the shared encoder extracts a common representation from the regional state  $s_r(t)$ . In this structure, the action-evaluation scale is reduced from  $2^{M_r(t)}$  in conventional DQN to  $2M_r(t)$ . Compared with a multi-agent formulation, which may require separate policy modules with repeated encoder parameters, the shared BDQ uses a single encoder and multiple lightweight branches. Thus, it preserves linear action evaluation while reducing the model complexity required for coordinated satellite activation. The branch-wise Q-value is computed as

$$Q_m(s_r, a_m) = V(s_r) + \left( A_m(s_r, a_m) - \frac{1}{|\mathcal{A}_m|} \sum_{a' \in \mathcal{A}_m} A_m(s_r, a') \right), \quad (14)$$

where  $\mathcal{A}_m = \{0, 1\}$  is the binary action set of branch  $m$  [11].

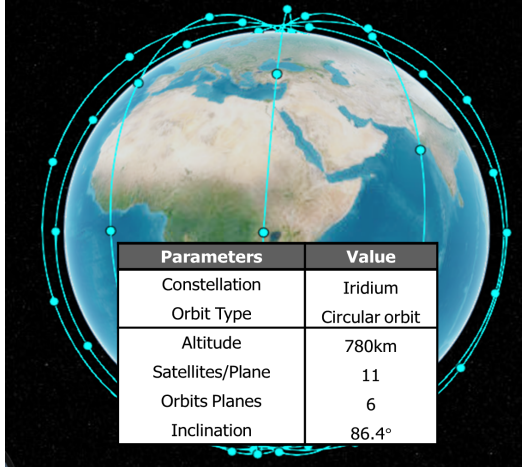


Fig. 2: Simulation environment.

Each branch greedily selects its binary sub-action according to  $Q_m(s_r, a_m)$ , and the selected sub-actions are concatenated to form the joint action  $a_r(t)$ . In terms of action evaluation, conventional DQN requires  $O(2^{M_r(t)})$  outputs, whereas the proposed BDQ requires  $O(M_r(t)|\mathcal{A}_m|) = O(M_r(t))$  because  $|\mathcal{A}_m| = 2$ . The policy is trained using experience replay and a Double-DQN target. The overall training procedure is summarized in Algorithm 1.

#### IV. PERFORMANCE ANALYSIS

This section evaluates the proposed method. The simulation environment, the parameter settings, and the considered baselines are described in Section IV-A. Section IV-B then analyzes the obtained results.

##### A. Simulation Setup

The simulation environment follows the system model in Section II. Each episode consists of  $T$  time slots with duration  $\Delta t$ , where the orbital geometry determines the visible satellite set  $\mathcal{S}_r(t)$  and the illumination state  $z_i(t)$ . The LEO satellite constellation and orbital configuration used in the simulation are illustrated in Fig. 2. The battery capacity, solar charging power, and power-consumption parameters are configured based on the satellite battery energy model in [9], while the rated cycle life  $L_C$  used for the wear cost follows the satellite-borne battery characterization in [10]. The main system parameters are summarized in Table I.

The proposed BDQ agent and the RL baselines are trained under the same environment and hyperparameter settings. Each training run consists of 400 episodes, where each episode contains 95 time slots. The discount factor is set to  $\gamma = 0.99$ , and the networks are optimized using Adam with a learning rate of  $10^{-4}$ . Experience replay with a buffer size of  $10^5$  and a minibatch size of 256 is used. We adopt a 3-step return, Huber loss, gradient clipping with a maximum norm of 10, and a soft target-network update with  $\tau = 5 \times 10^{-3}$ . The exploration rate  $\epsilon$  is linearly decayed from 1.0 to  $10^{-3}$ . The BDQ network consists of a shared encoder with hidden dimensions (256, 128), followed by branch heads with 64

TABLE I: Simulation parameters [9].

| Symbol           | Description                      | Value           |
|------------------|----------------------------------|-----------------|
| $ \mathcal{R} $  | Number of service regions        | 18              |
| $N_{\max}$       | Max. slots per region            | 8               |
| $T$              | Simulation step                  | 200 slots       |
| $C$              | Per-satellite service capacity   | 75 MB/min       |
| $\eta_{\min}$    | QoS threshold                    | 0.95            |
| $B_{\max}$       | Battery capacity                 | 32.5 Wh         |
| $p^{\text{sol}}$ | Solar charging power             | 20 W            |
| $P_{\text{slp}}$ | Sleep-mode power                 | 0 W             |
| $P_b$            | Active base power                | 10 W            |
| $P_{\text{tx}}$  | Transmission power               | 5–25 W          |
| $P_{\text{rx}}$  | Reception power                  | 4 W             |
| $L_C$            | Rated cycle life (cycles)        | $6 \times 10^4$ |
| $A$              | Wear DoD-sensitivity coefficient | 0.8             |

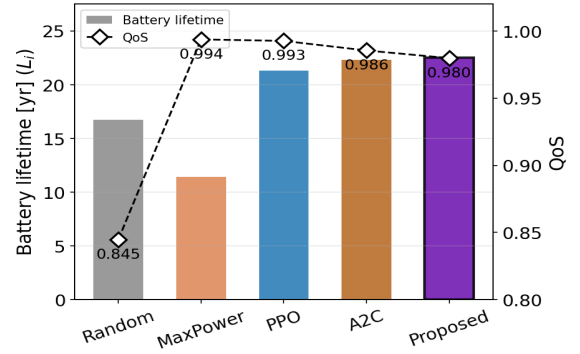


Fig. 3: Battery lifetime and QoS satisfaction comparison under different policies.

hidden units. The policy checkpoint with the best validation score is used for evaluation.

##### B. Simulation Results

The proposed method is compared against four representative baselines under an identical evaluation protocol. The Random and MaxPower policies represent two operational extremes. Random independently activates each visible satellite with probability 0.5, whereas MaxPower always activates all visible satellites. The PPO and A2C agents are policy-gradient learners that use the same per-satellite Bernoulli action representation as the proposed agent. The Proposed agent corresponds to the proposed satellite activation policy described in Section III-C. All RL agents are trained in the same environment with the same parameter set, and their best validation checkpoints are used for evaluation. We report two metrics: the average per-satellite battery lifetime [yr] and the QoS satisfaction ratio averaged over all regions and time slots.

Fig. 3 shows the lifetime–QoS trade-off achieved by each policy under the default environment. The two non-learning baselines illustrate the main trade-off between service quality and battery degradation. MaxPower attains the highest QoS of 0.994 but yields a battery lifetime of only 11.41 yr, since continuous activation increases the per-slot wear cost. Random reduces the activation ratio but achieves a QoS of 0.845, thereby failing to satisfy the QoS constraint. The RL-based

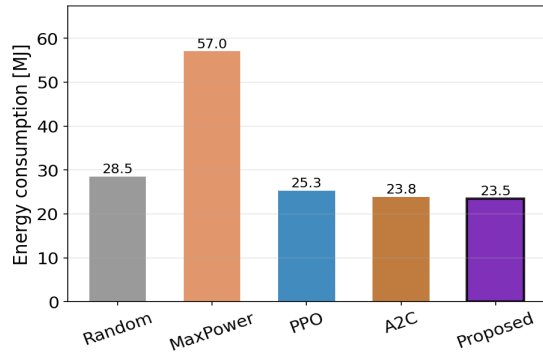


Fig. 4: Constellation energy consumption and QoS satisfaction under different policies.

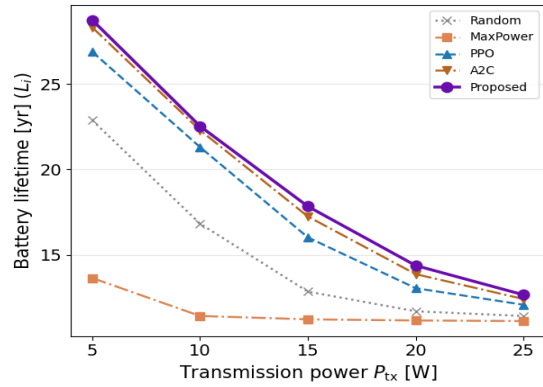


Fig. 5: Robustness of average battery lifetime under different transmission powers.

baselines, PPO and A2C, both satisfy the QoS constraint while extending the battery lifetime to 21.31 yr and 22.30 yr, respectively. The Proposed method achieves the longest lifetime of 22.54 yr while keeping the QoS at 0.980, which remains above the QoS threshold. Compared with the always-on MaxPower baseline, the proposed method extends the battery lifetime by approximately 97.5% while still satisfying the QoS constraint. This result indicates that the joint reward design in (13) guides the policy toward a favorable lifetime–QoS operating point.

Fig. 4 compares the constellation energy consumed over the evaluation horizon. MaxPower consumes the most (57.0 MJ) by activating every visible satellite, while Random uses less (28.5 MJ) but violates the QoS requirement (0.845). Among the QoS-satisfying policies, the proposed method consumes the least energy (23.5 MJ), about 59% less than MaxPower, with PPO and A2C at a comparable level (25.3 and 23.9 MJ). This shows that the proposed policy activates only the satellites needed to meet the demand, reducing energy consumption as a co-benefit of its lifetime-aware control.

Fig. 5 evaluates the robustness to environment shifts, where all learning-based policies are trained only at  $P_{tx} = 10$  W and tested across 5–25 W. As  $P_{tx}$  increases, each satellite activation causes deeper eclipse-phase discharge, and thus the battery lifetime decreases for all policies. Among the QoS-feasible policies, the proposed method achieves the longest lifetime at every operating point, including unseen power conditions,

while A2C and PPO show slightly shorter lifetimes. Although the Random policy exhibits a higher lifetime curve, it is infeasible because it violates the QoS requirement. The advantage over MaxPower is largest at low power, with 28.8 years compared with 13.6 years at 5 W, and becomes smaller at high power, with 12.6 years compared with 11.1 years at 25 W, because QoS-mandated activations dominate battery wear. These results indicate that the proposed SoC- and wear-aware state representation generalizes well across different power conditions.

## V. CONCLUSION

This paper presented a battery-lifetime-aware control framework for LEO satellite constellations under time-varying traffic demand, orbital illumination, and battery dynamics. The problem was formulated as a constrained constellation control problem and addressed using a shared branching dueling DQN policy to handle the combinatorial binary satellite activation space. Simulation results showed that the proposed method achieves the longest average battery lifetime while satisfying the QoS threshold, and maintains its advantage under off-training transmission-power conditions. Future work will extend the framework to bursty traffic, inter-satellite link constraints, and joint routing-control optimization.

## 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] 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.
- [3] H. Tsuchida, Y. Kawamoto, N. Kato, K. Kaneko, S. Tani, M. Hangai, and H. Aruga, “Improvement of battery lifetime based on communication resource control in low-Earth-orbit satellite constellations,” *IEEE Transactions on Emerging Topics in Computing*, vol. 10, no. 3, pp. 1388–1398, July–September 2022.
- [4] G. Xu, L. Wang, S. Chen, L. Zhu, M. Guizani, and L. Shi, “Mpaee: A multipath adaptive energy-efficient routing scheme for low Earth orbit-based industrial internet of things,” *IEEE Internet of Things Journal*, vol. 12, no. 17, pp. 34 793–34 805, September 2025.
- [5] P. Zhang, Z. Xian, M. Liao, H. Huang, and J. Yang, “Adaptive routing mechanism for LEO satellite network based on control domain partition,” *IEEE Transactions on Green Communications and Networking*, vol. 9, no. 1, pp. 70–82, March 2025.
- [6] J. Jang, I.-S. Cho, M. Shin, J. Kim, and S. Jung, “Gateway-assisted neural angular routing for hierarchical satellite networks,” *IEEE Internet of Things Journal*, vol. 13, no. 6, pp. 12 322–12 339, January 2026.
- [7] J. Jang, J. Kim, J. Kim, and S. Jung, “Joint interference approximation and guard-band management for spectrum-efficient integrated NTN–TN networks,” *IEEE Internet of Things Journal*, vol. 12, no. 15, pp. 32 220–32 236, August 2025.
- [8] J. Kim, J. Lee, J. Jang, J.-H. Kim, and S. Jung, “Synergistic power optimization in satellite-terrestrial coexistence networks via attention-driven deep reinforcement learning,” in *2025 IEEE 101st Vehicular Technology Conference (VTC2025-Spring)*, Oslo, Norway, 2025, pp. 1–6.
- [9] R. do N. Mota Macambira, C. B. Carvalho, and J. F. de Rezende, “Energy-efficient routing in leo satellite networks for extending satellites lifetime,” *Computer Communications*, vol. 195, pp. 463–475, November 2022.
- [10] H. Tsuchida, Y. Kawamoto, N. Kato, K. Kaneko, S. Tani, S. Uchida, and H. Aruga, “Efficient power control for satellite-borne batteries using q-learning in low-earth-orbit satellite constellations,” *IEEE Wireless Communications Letters*, vol. 9, no. 6, pp. 809–812, January 2020.
- [11] A. Tavakoli, F. Pardo, and P. Kormushev, “Action branching architectures for deep reinforcement learning,” in *Proc. AAAI Conference on Artificial Intelligence*, vol. 32, no. 1, New Orleans, LA, USA, February 2018.