Traffic Scheduling in Non-Stationary Multipath Non-Terrestrial Networks: A Reinforcement Learning Approach

Achilles Machumilane*[1], Alberto Gotta[1], Pietro Cassará[1], Claudio Gennaro[1], and Giuseppe Amato[1]

*Department of Information Engineering, University of Pisa-achilles.machumilane@phd.unipi.it
[1]Institute of Information Science and Technologies (ISTI), CNR, Pisa. e-mail:{name.surname}@isti.cnr.it

***Abstract*—In Non-Terrestrial Networks (NTNs), where LEO satellites and User Equipment (UE) move relative to each other, Line-of-Sight (LOS) tracking,and adapting to channel state variations due to endpoint movements are a major challenge. Therefore, continuous LOS estimation and channel impairment compensation are crucial for a UE to access a satellite and maintain connectivity. In this paper, we propose a Actor-Critic (AC)-Reinforcement Learning (RL) framework for traffic scheduling in NTN scenarios where the channel state is non-stationary due to the variability of LOS, which depends on the current satellite elevation. We deploy the framework as an agent in a Multi-Path Routing (MPR) scheme where the UE can access more than one satellite simultaneously to improve link reliability and throughput. We study how the agent schedules traffic on multiple satellite links by adopting the AC version of RL. The agent continuously trains based on variations in satellite elevation angles, handoffs, and relative LOS probabilities. We compare the agent retraining time with the satellite visibility intervals to investigate the effectiveness of the agent's learning rate. We carry out performance analysis considering the dense urban area of Chicago, where high-rise buildings significantly affect the LOS. The simulation results show how the learning agent selects the scheduling policy when it is connected to a pair of satellites. The results also show that the retraining time of the learning agent is up to $0.1\ times$ the satellite visibility time at certain elevations, which guarantees efficient use of satellite visibility.**

***Index Terms*—NTN, Satellites, Link Prediction, Reinforcement Learning, Actor-Critic, Multipath.**

I. Introduction

NTNs, including Low Earth Orbit (LEO) satellite constellations, Unmanned Aerial System (UAS), and High Altitude Platforms (HAPs) have been identified as promising technologies to provide ubiquitous connectivity [1] in the future generation Internet. For this reason, the Third Generation Partnership Project (3GPP) [2] has included NTNs among the supporting technologies for the extension of the terrestrial fifth-generation (5G) into the sixth-generation (6G) mobile networks. NTNs can be exploited to meet the requirements of emerging technologies such as ubiquitous Artificial Intelligence (AI) and Industrial IoT (IIoT) for application use cases like remote monitoring, goods delivery, connected autonomous vehicles (CAVs), and high-speed transportation (trains, aircraft). However, the main challenge in New-Radio-NTN integration is the communication between the UE and the satellite, because it requires the LOS. In dense urban scenarios, high-rise buildings or tall infrastructures can severely affect LOS communication due to signal blockage and reflection phenomena. Communication in LOS between satellites and UE becomes even more challenging in scenarios where the satellite and the UE are moving relative to each other because, in these scenarios, the LOS probability changes with the satellite elevation angle. Therefore, continuous LOS estimation techniques are paramount for the UE to access the satellite and maintain connectivity. This paper proposes an RL-based agent to self-learn link selection in non-terrestrial networks with Multi-Path Routing (MPR) in dense urban scenarios. MPR allows a UE with multiple radios to set up multiple satellite connections to improve reliability and data rate [3], [4] even when the performance of the single link is degraded in terms of LOS. We assume a non-stationary LOS probability due to the continuous variation of the satellite elevation angle. In such scenarios, a reliable LOS estimation model allows the UE to select a link or more links to maximize an objective, such as limiting the End-to-End (E2E) loss while using minimal bandwidth. To this end, we adopt the AC version of the RL, which guarantees good performance with continual learning for non-stationary LOS probability that underlies our system. We analyze the latency of the agent in recovering from an abrupt change in the LOS of one or more links. The changes of LOS are the consequence of the satellite visibility period, which depends on both the satellite elevation angle and the latitude of the UE. In [5], the authors present a theoretical model that estimates the probability of a Cloud-Free LOS (CFLOS) in satellite links based on the elevation angle of the slant link and the altitude of ground stations. Sun et al. [6] propose a Maximum-Likelihood-based technique for Non-Line-of-Sight (NLOS) detection using Global Navigation Satellite System data. In [7], authors

Work supported by the European Union under the Italian National Recovery and Resilience Plan (PNRR) PE00000001 - program "RESTART" and by the HORIZON-CL4-2021-SPACE-01 project "5G+ evoluTion to mutioRbitAl multibaNd neTwORks" (TRANTOR) No. 101081983

propose an empirical model for LOS probability estimation for satellite and HAPs communications. In contrast to these physical and empirical methods, we propose an RL-based model for non-stationary scenarios in which LEO satellites continuously move changing their elevation angles and the relative LOS with the UE. In addition, our model allows a UE to estimate the traffic scheduling policy for multiple satellites, i.e., accounting for multiple parallel transmissions. Regarding MPR techniques, the authors in [8] propose a Deep-Q RL-based scheduler for bandwidth allocation at a WiFi Access Point (AP), to meet the Quality of Service (QoS) requirements of various user applications. Again, in [9] authors propose an AC-based scheduler for Multi-Path air-to-ground multimedia delivery in cellular-assisted-UAV communication. Wu et al. [10] propose Peekaboo, an RL-based MPR scheduler implemented in Multipath-QUIC to address WiFi and cellular channel heterogeneity.

Unlike these works, our proposed model provides link selection and flow protection in NTN, guaranteeing link resilience while avoiding encoding/decoding overhead and retransmission delays, which are introduced when using techniques such as forward error correction (FEC) [11] and automatic repeat request (ARQ). This is of utmost importance because round-trip delay is an unavoidable bottleneck in satellite communications, especially for real-time applications.

The main contributions of our work can be summarized as follows:

- We provide a learning-based framework for selecting an optimal MPR-based policy according to the time-varying satellite elevation angle. We also provide a mechanism for reliable estimation of non-stationary LOS probability.
- Including MPR capabilities in our transmission policy, we allow the UE to transmit on multiple satellite links to improve link availability and data rates and minimize end-to-end (E2E) loss.

The rest of the paper is organized as follows: Section II describes our system model. Section III presents the problem formulation and the AC agent. In Section IV, we describe the simulation setup. Simulation results are presented and discussed in section V. Section VI concludes the paper and sets the direction for future research.

II. System Model

Figure 1 shows the reference scenario considered in this paper. We study LOS estimation and link selection in the presence of dual connectivity in NTNs with simultaneous use of two radios as envisioned by the 3GPP [2]. In this architecture, the LEO satellites are equipped with the gNB-Distributed Unit (DU) [12], while the Centralized Unit (CU) is located on the ground. We consider a scenario in which the satellites and an Unmanned Aerial Vehicle (UAV) equipped with two UEs are moving relative to each other. We consider the StarLink satellite system with a mass constellation of 3,000 LEO satellites. The UE can connect to two satellites simultaneously. According to [13], a satellite in the constellation moves in a circular orbit with inclination ι at an altitude h, and the orbit radius $r_S = r_E + h$, and the satellites move independently of each other. The same authors in [13] defined $\gamma(\theta)$:

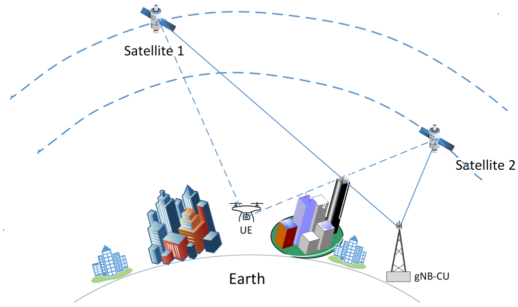

Fig. 1: Reference Scenario: A UE (UAV) accessing two satellites in an NTN in a dense urban environment.

$$\gamma(\theta) = cos^{-1}\left((r_E/r_S)\cdot cos(\theta) - \theta\right), \tag{1}$$

as the central angle between the earth station, the UE in this case, and the locus of the trajectory points of the satellite corresponding to elevation angle θ, with $\theta_{min} \leq \theta \leq \theta_{max}$. For any single point of the satellite's locus, the maximum elevation angle θ_{max} determines the visibility time of the satellite and the distribution of the elevation angles in the visibility region [13]. The visibility region is defined as the smallest angle $\gamma(\theta_{max})$ for which the satellite is visible from the UE along its whole trajectory. Therefore, given the UE latitude ϕ_0, the probability for a satellite in its trajectory to be visible from the UE can be determined from the Probability Density Function (PDF) of θ_{max}, denoted by:

$$f_{\Theta_{max}}(\theta_{max}) = \frac{G(\theta_{max})}{K}\cdot \left(\frac{cos(\phi_0+\gamma(\theta_{max}))}{\pi\sqrt{sin^2(\iota) - sin^2(\phi_0+\gamma(\theta_{max}))}} + \frac{cos(\phi_0-\gamma(\theta_{max}))}{\pi\sqrt{sin^2(\iota) - sin^2(\phi_0-\gamma(\theta_{max}))}}\right) \tag{2}$$

where $\theta_{min} \leq \theta_{max} \leq \frac{\pi}{2}$ and

$$G(\theta_{max}) = \frac{1 + (r_E/r_S)^2 - 2(r_E/r_S)cos(\gamma(\theta_{max}))}{1 - (r_E/r_S)cos(\gamma(\theta_{max}))}$$

$$K = \frac{1}{\pi}\left(sin^{-1}\left(\frac{sin(\phi_0+\gamma(\theta_{min}))}{sin(\iota)}\right)\cdot \cdot sin^{-1}\left(\frac{sin(\phi_0-\gamma(\theta_{min}))}{sin(\iota)}\right)\right)$$

The PDF in (2) may assume different shapes, according to ι, ϕ_0, and $\gamma(\theta_{max})$ as detailed in [13]. For space limitations, we will only account for the PDF of elevation angles considering the points of the satellite's trajectory in the visibility region. The authors in [13] derive this PDF $f_\Theta(\theta)$ as the marginalization of the joint probability $f_{\Theta,\Theta_{max}}(\theta,\theta_{max})$, defined as in the following equation:

$$f_\Theta(\theta) = \int_{\theta}^{\theta_{max}} f_{\Theta,\Theta_{max}}(\theta,\theta_{max}) d\theta_{max} \tag{3}$$

where $\theta_{min} \leq \theta \leq \theta_{max}$ and

$$f_{\Theta,\Theta_{max}}(\theta,\theta_{max}) = \frac{G(\theta) sin(\gamma(\theta))}{\sqrt{cos^2(\gamma(\theta_{max})) - cos^2(\gamma(\theta))}} \cdot \frac{f_{\Theta_{max}}(\theta_{max})}{\displaystyle\int_{\theta_{min}}^{\theta_{max}} f_{\Theta_{max}}(x) \cdot cos^{-1}\left(\frac{cos(\gamma(\theta_{min}))}{cos(\gamma(x))}\right) dx}$$

Therefore, the satellite visibility interval from a UE at a given latitude as the elevation angle varies from θ_i to θ_j, is given by [13] as:

$$T_{\theta_i,\theta_j} = \int_{\theta_i}^{\theta_j} \frac{2}{\omega_S - \omega_E cos(\iota)} cos^{-1}\left(\frac{cos(\gamma(\theta_i))}{cos(\gamma(x))}\right) \cdot f_{\Theta_j}(x) dx \tag{4}$$

The satellites move in different orbits at different speeds. According to 3GPP [2], the LOS probability changes with the changing satellite elevation angle. In general, the LOS probability increases with elevation, reaching a maximum at *Nadir* (90°) when the satellite is above the UE if it is in the orbital plane of the satellite. In dense urban areas, the LOS probability is lower, especially at low altitudes, because the signal is obstructed by and reflected from high-rise buildings. Consequently, the AC agent must learn whether to schedule traffic transmission on any one link, or on both links simultaneously (for redundancy) according to a given Quality of Service (QoS) requirement and according to the estimate of LOS probability model of the two links as the satellites change their elevation angles. The Discrete Markov Chain (DMC) channel model derived in [14], characterizes the LOS/NLOS transition probabilities at specific elevation angles θ of the satellite in the range $\theta \in [15°, 165°]$ with an interval of 10°.

The transition probabilities in [14], were derived using ray-tracing simulations using a map of Chicago downtown with an area of 1070 m x 1070 m, an average building height of 150 m, and a maximum building height of 526m. Using the link state transition probabilities provided in [14], we model the state transition matrices for two independent satellites at selected elevation angles, as explained later in Section IV. We then use these matrices to create Markov link state traces for training our model. Since connectivity with the satellite requires the LOS, we assume successful traffic reception only if there is LOS. We assume that the AC agent receives some feedback reports as in [3], indicating the reception status and, consequently, recording the link state.

III. Problem Formulation

LOS estimation on multiple links can be formulated as a Markov Decision Process (MDP). Specifically, it is modelled as a Partially Observable Markov Decision Process (POMDP) [15] because, to the agent, the environment is not fully observable; the RL agent can only observe the link(s) it has selected out of all the available links. Moreover, the pattern underlying the state variations is unknown to the agent. So, the agent tries to learn the state variation probabilities using its past observations. A POMDP is defined by the tuple $\{\mathcal{S}, \mathcal{A}, P(s_{t+\Delta t}|s_t, a_t), r_t\}$, where $\mathcal{S}$ is the states space of the system, and $\mathcal{A}$ denotes the actions space to achieve the optimal choice. For the easy of notation, we shall use S also to denote the agent's observations. $P(s_{t+\Delta t}|s_t, a_t)$ is the probability of being in state $s_{t+\Delta t} \in \mathcal{S}$ after a time interval Δt conditioned by the action $a_t \in \mathcal{A}$ and the state $s_t \in \mathcal{S}$; and r_t is the immediate reward due to the action a_t that leads to state transition from s_t to $s_{t+\Delta t}$. In the following, we describe the POMDP for our problem, where we assume that the UE can select a subset of the N available satellite links to which it is connected.

1) *States Space*: We assume the state of a link as a binary variable $\{\mathrm{LOS}, \mathrm{NLOS}\}$, and formally define the state of the link $n = 1 \ldots N$ at time t as follows:

$$s_{nt} = \begin{cases} +1 & \text{if } s_{nt} = \mathrm{LOS} \\ -1 & \text{otherwise} \end{cases}$$

The N links dynamically change their states between LOS and NLOS, according to their own transition matrix T_n as defined in [14]. So, we can define the link states space as the set of the vectors $\mathcal{S} = \{\mathbf{s}_t \,|\, \mathbf{s}_t = [s_{1t}, \ldots, s_{Nt}]\}$

2) *Actions*: The action constitutes the choice of the appropriate transmission pattern; that is, a subset of the N links. The actions space is the set of vectors $\mathcal{A} = \{\mathbf{a}_t \,|\, \mathbf{a}_t = [\rho_{1t}, \ldots, \rho_{Nt}]\}$ where $\rho_{nt} = 1$ indicates that the n-th link is selected, and $\rho_{nt} = 0$ otherwise, for $n = 1 \ldots N$. In this case study, we assumed to have a pair of radio interfaces, i.e. $N = 2$, which leads to having an actions space $\mathcal{A} = \{[0,1],[1,0],[1,1]\}$.

3) *Reward* : The immediate reward r_t is defined as a penalty to the agent that is proportional to the loss of the transmitted data as in the following equation:

$$r_t = \begin{cases} \frac{c}{\|\mathbf{a}_t\|_1} & \text{if } \sigma = 0 \\ \frac{1}{\sigma} & \text{otherwise,} \end{cases}$$

where $\|\mathbf{a}_t\|_1$ is the 1-norm of the selected action,which is equal to the number of the selected links and c is a negative constant to provide a penalty to the learning agent. This first term, penalizes more the use of a single link when there are losses and encourages the use of double transmissions to overcome the losses. σ is the total number of received packets. $\sigma = 0$ means that all the traffic sent is lost marking an E2E loss event. The term $\frac{1}{\sigma}$ is meant to conserve bandwidth by discouraging the use of multiple transmissions in favourable link conditions.

The Actor-Critic Agent

Figure 2 shows the architecture of the AC learning agent that achieves the optimization goal through the policy $\pi(a_t|s_t)$. The actor, A with parameters θ_a updates this policy using the feedback from both the environment (observed system) and the critic C with parameters θ_c, which estimates the state-action value. The policy function $\pi(a_t|s_t)$ can be represented as a parameterized function whose parameters can be estimated by using different techniques, such as Neural Networks (NN). The actor updates the optimal policy by maximizing the total reward discounted by the parameter ξ, according to the following optimization function:

$$\pi^*(a_t|s_t) = \arg\max_{a_t} E\Big[\sum_{t=0}^{\infty} \xi^t r_t(s_t, a_t)\Big]. \tag{5}$$

The critic, on the other hand, evaluates the goodness of the updated policy by estimating the state-action function $Q^\pi(s_t, a_t)$. Also, in this case, the action-state function at the critic can be represented as a parameterized model. The estimation of the future state-action values is performed by using the target-critic network as follows:

$$Q^\pi(s_{t+\Delta t}, a_{t+\Delta t}) = E\Big[r_t + \xi\, \hat{Q}^\pi(s_{t+\Delta t}, a_{t+\Delta t})\Big]. \tag{6}$$

The target-critic is used in order to overcome the critic instability caused by frequent updates. The actor can use the feedback coming from the critic to achieve the optimal policy by using the method described in [16]. Instead, we adopt the solution described in [17] for the AC proposed in this work, which involves the evaluation of the Temporal Difference (TD) error, given by:

$$\delta = r_t + \xi \hat{Q}^\pi(s_{t+\Delta t}, a_{t+\Delta t}) - Q^\pi(s_t, a_t) \tag{7}$$

The actor and the critic networks are respectively updated according to the following loss functions:

$$\Delta_A = -\beta\delta \ln \pi(a_t|s_t), \tag{8}$$

$$\Delta_C = \alpha\delta^2, \tag{9}$$

where β and α are learning rates for the actor and the critic respectively. The target critic network is updated with a soft-update method as:

$$\theta^{new}_{c,\,targ.} = \alpha\, \theta^{old}_{c,\,targ.} + (1-\alpha)\theta_c, \tag{10}$$

where θ_c and $\theta_{c,targ}$ are parameters for the critic and the target critic networks respectively.

We implemented the actor and the critic networks using TensorFlow-2 and Keras libraries with ADAM optimizer in a fully connected Multi-layer Perceptron Neural Network (NN) of 3 hidden layers, 64 neurons per hidden layer, learning rates $\beta = 10^{-4}$ for the actor and $\alpha = 5 \cdot 10^{-4}$ for the critic and the discount factor $\xi = 0.96$.

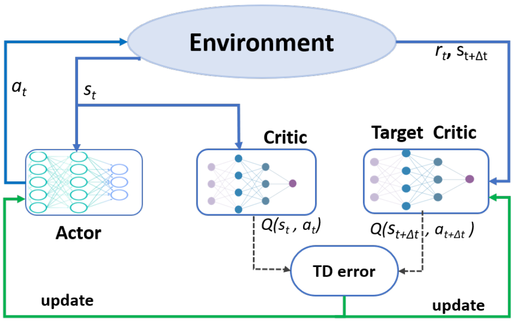

Fig. 2: Architecture of the Actor-Critic Learning Agent.

IV. Simulation Setup

We simulated the UAV-Satellite transmission with dual connectivity, where a UAV uses two UEs to connect to two different satellites. The goal of the simulation is to train our learning agent, to estimate the LOS model of the two satellites and the optimal policy for selecting appropriate links (transmissions policy) while tracking the elevation angles of the satellites. The channel model used to create the dataset to train our learning agent was obtained using the DMC reported in [14] in which transition probabilities are given at specific elevation angles θ, which account for an interval of $\pm 5°$ in the range $\theta \in [15°, 165°]$ with a step of $10°$. We use only the probabilities for dense urban from [14] for this work. As explained in Section I, the LOS probability changes according to the elevation angle. Therefore, the learning agent must continuously track the variation of the LOS of the two satellites as a function of the elevation angles. To this end, using a satellite tracker[1], two pairs of Starlink satellites visible from Chicago downtown in a given

[1] https://satellitemap.space

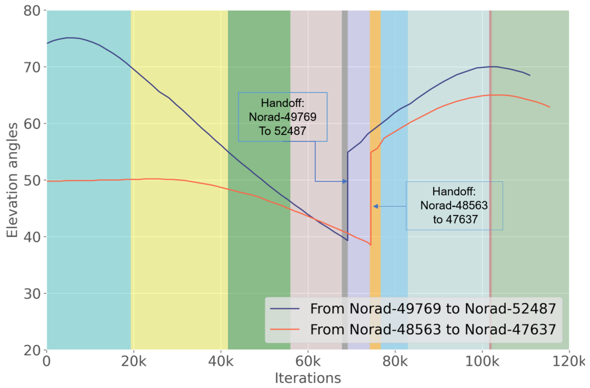

Fig. 3: Elevation angles of the pair of satellite connected to the UE in different contexts.

period of a day were selected: (NORAD-49769, NORAD-48563) and (NORAD-52487, NORAD-47637). Then we selected from [14] the relative channel models associated with the relative elevation angles (θ_1, θ_2) for each pair of the selected satellites. These satellites were selected because they provide two clear handoff events, i.e. from NORAD-49769 to NORAD-48563 and from NORAD-52487 to NORAD-47637, each showing an abrupt change that forces the learning agent to retrain its model. Note that the two handoffs don't happen simultaneously, since the two radio interfaces are independent of the other. The selected pairs of angles and the two handoff events are shown in Figure 3. Figure 4 shows the probability mass function of the satellite visibility at given elevation angles, evaluated using equation (3) as compared to the empirical values achieved from the real dataset collected by the satellite tracker during a window of 15 mins (the maximum allowed).

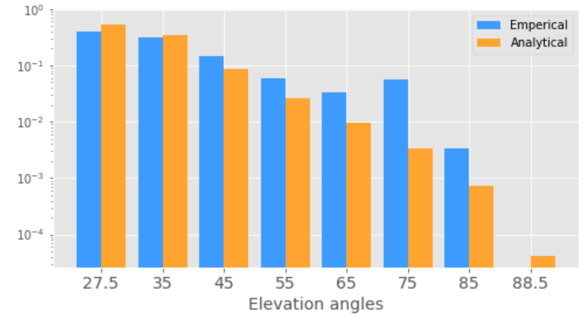

Fig. 4: Probability that a satellite is visible from a UE at given satellite elevation angles at Chicago's latitude.

Using the selected pairs of angles, and the corresponding transition probabilities in [14], we constructed state transition matrices for each satellite for each pair of elevation angles, obtaining a total of 10 transitions of *context*, i.e., at each range of angles, we transition to a different channel model. The dataset contains 120K records, with each of the selected elevation angles having records proportional to the satellite visibility time at that angle.

V. Performance Evaluation

This Section provides a proof of concept of how the RL agent performs and can address a non-stationary channel model by retraining its parameters after either moderate angle elevation variations or abrupt changes, e.g., due to handoffs. We do not wish to go into the merits of how the satellite sequences are selected to perform the handoffs, which may, instead, be the subject of future studies to determine an optimal policy that addresses the handoffs to mitigate the abrupt change and reduce the learning phases.

Figure 5 shows the total discounted rewards achieved by the agent in a sequence of contexts characterized by different colours. We present the median and the 25 and 75 percentiles of the reward. The smooth semi-plateaus within the coloured strips show the steady states achieved by the model within the context. These results show how the agent dynamically detects the change of the satellite elevation angle, which triggers the change of the LOS probability. The relative movements of the respective satellites have been plotted in Figure 3 that, analogously to Figure 5, shows the sequence of contexts, the relative elevation angles, and two handoff events.

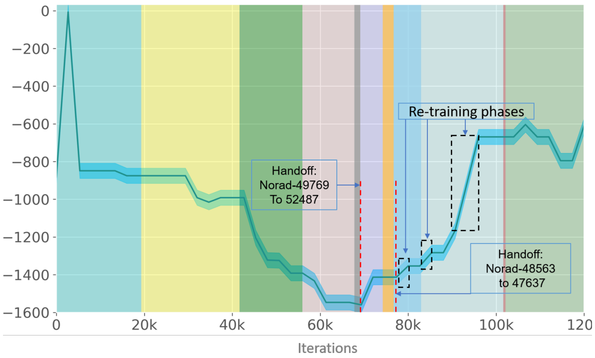

Fig. 5: Total discounted rewards achieved at different elevation angles.

Figure 6 shows the categorical distribution of the AC agent and the optimal policy for transmitting with satellite 1, 2, and with both satellites $1, 2$ in the different contexts; that is, $P(1)$, $P(2)$, and $P(1,2)$, respectively. According to the channel model provided by 3GPP for dense urban areas [18] and [14], the higher the elevation angle, the higher the LOS probability. It is evident from Figure 6 that in all the contexts, when the two satellites are not at the same elevation angle, link 1 connects to satellites at higher elevation angles compared to link 2, which means that link 1 has a higher LOS probability. It can be seen that our model can detect this pattern and transmit more on link 1 than on link 2. However, since the elevation angles in Chicago are not so high for the selected satellites, the learning agent probes significantly two links simultaneously *w.r.t.* a single link to overcome the NLOS probability. We compare the categorical distributions of the AC agent to those obtained by the optimal policy in which the system knows in advance the channel model, and thus, the relative prediction is optimal. It can be seen that even in non-stationary conditions, the AC agent is able to achieve a quasi-optimal scheduling policy without any modeling. Figure 5 also shows the average time it takes the agent to retrain the NN model after a change of elevation angle juxtaposed to the relative context duration. As already said, this is of utmost importance to optimally utilize the satellite visibility time. It was found out that, on average, to update the NN parameters and achieve a local steady state of the reward function, the RL agent requires $2K$ iterations of the DMC which is equivalent to 0.1x the satellite visibility time at a given elevation angle, which guarantees efficient

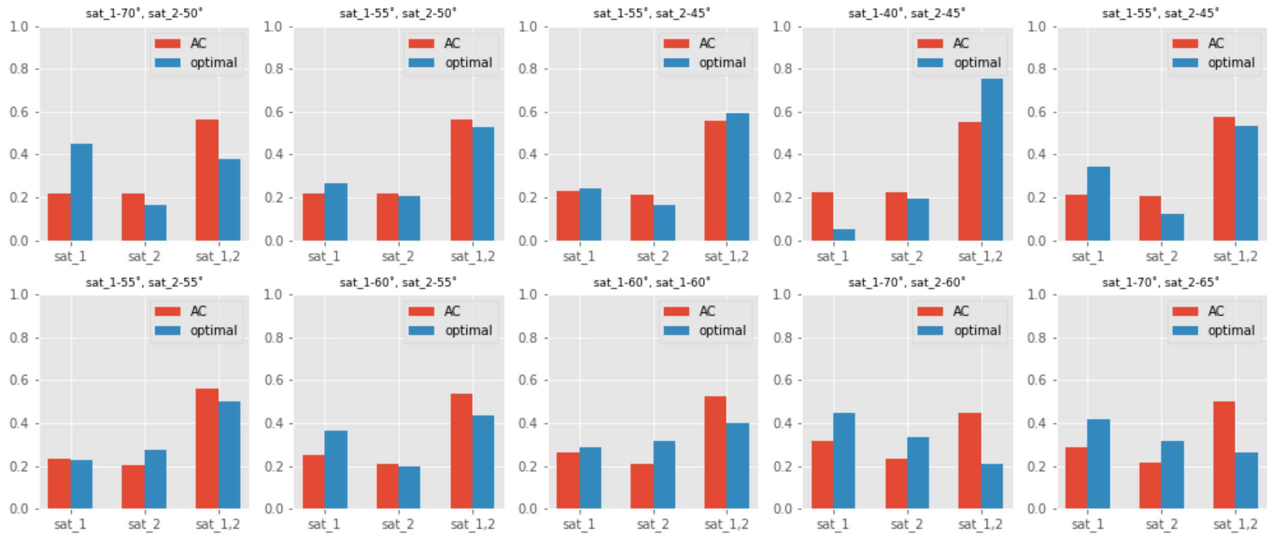

Fig. 6: Categorical distribution for multi link scheduling with AC vs. optimal scheduling at different elevation angles.

utilization of satellite visibility.

VI. Conclusion

In this work, we have presented an Actor-Critic RL agent for LOS estimation in non-stationary conditions relying on multi-link NTNs in dense urban environments. Simulation results have shown that the learning agent has a performance similar to an optimal policy with total knowledge of the channel model in estimating the LOS probabilities of multiple satellite links and in selecting the suitable scheduling policy for the selection of the links. The use of multiple links is for increasing resilience to E2E loss, reliability, data rate, and throughput, and thus, to improve QoS. In this work, we outlined the handoffs between LEO satellites with real traces from the Starlink constellation that lead to an abrupt change in the elevation angles w.r.t. the user equipment. In future research, we plan to deepen the analysis of the handoffs policies and look at the integration of both ground and terrestrial segments.

References

[1] M. Bacco, F. Davoli, G. Giambene, A. Gotta, M. Luglio, M. Marchese, F. Patrone, and C. Roseti, "Networking challenges for non-terrestrial networks exploitation in 5g," in *IEEE 2nd 5G World Forum (5GWF)*. 10.1109/5GWF.2019.8911669, 2019, pp. 623–628.

[2] 3GPP, "Technical specification group radio access network; solutions for nr to support non-terrestrial networks (ntn): Tr 38.821 v16.1.0 (2021-05), (release 16)."

[3] A. Machumilane, A. Gotta, P. Cassarà, and M. Bacco, "A path-aware scheduler for air-to-ground multipath multimedia delivery in real time," *IEEE Communications Magazine*, vol. 60, no. 9, pp. 54–58, 2022.

[4] M. Bacco, P. Cassará, A. Gotta, and V. Pellegrini, "Real-Time Multipath Multimedia Traffic in Cellular Networks for Command and Control Applications," in *2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall)*, 2019, pp. 1–5.

[5] A. Badr, A. Khisti, W.-T. Tan, and J. Apostolopoulos, "Perfecting protection for interactive multimedia: A survey of forward error correction for low-delay interactive applications," *IEEE Signal Processing Magazine*, vol. 34, no. 2, pp. 95–113, 2017.

[6] Y. Sun and L. Fu, "Stacking ensemble learning for non-line-of-sight detection of global navigation satellite system," *IEEE Transactions on Instrumentation and Measurement*, vol. 71, pp. 1–10, 2022.

[7] V. Aydın, İ. H. Çavdar, and Z. HASIRCI, "Line of sight (los) probability prediction for satellite and haps communication in trabzon, turkey," *International Journal of Applied Mathematics Electronics and Computers*, no. Special Issue-1, pp. 155–160, 2016.

[8] Q. Wang, T. Nguyen, and B. Bose, "Towards adaptive packet scheduler with deep-q reinforcement learning," in *2020 International Conference on Computing, Networking and Communications (ICNC)*, 2020, pp. 118–123.

[9] A. Machumilane, A. Gotta, P. Cassará, C. Gennaro, and G. Amato, "Actor-critic scheduling for path-aware air-to-ground multipath multimedia delivery," in *2022 IEEE 95th Vehicular Technology Conference:(VTC2022-Spring)*. IEEE, 2022, pp. 1–5.

[10] H. Wu, Ö. Alay, A. Brunstrom, S. Ferlin, and G. Caso, "Peekaboo: Learning-based multipath scheduling for dynamic heterogeneous environments," *IEEE Journal on Selected Areas in Communications*, vol. 38, no. 10, pp. 2295–2310, 2020.

[11] A. Gotta and P. Barsocchi, "Experimental video broadcasting in dvb-rcs/s2 with land mobile satellite channel: a reliability issue," in *2008 IEEE International Workshop on Satellite and Space Communications*. IEEE, 2008, pp. 234–238.

[12] F. Granelli, "Network slicing," in *Computing in Communication Networks*. Elsevier, 2020, pp. 63–76.

[13] S.-Y. Li and C. Liu, "An analytical model to predict the probability density function of elevation angles for leo satellite systems," *IEEE Communications Letters*, vol. 6, no. 4, pp. 138–140, 2002.

[14] E. Juan, I. Rodriguez, M. Lauridsen, J. Wigard, and P. Mogensen, "Time-correlated geometrical radio propagation model for leo-to-ground satellite systems," in *2021 IEEE 94th Vehicular Technology Conference (VTC2021-Fall)*. IEEE, 2021, pp. 1–5.

[15] G. E. Monahan, "State of the art—a survey of partially observable markov decision processes: theory, models, and algorithms," *Management science*, vol. 28, no. 1, pp. 1–16, 1982.

[16] V. R. Konda and J. N. Tsitsiklis, "Actor-Critic Algorithms ," in *Int. Conf. NIPS*. MIT Press, 1999, pp. 1008–1014.

[17] I. Grondman, L. Busoniu, G. A. D. Lopes, and R. Babuska, "A survey of actor-critic reinforcement learning: Standard and natural policy gradients," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 42, no. 6, pp. 1291–1307, 2012.

[18] 3GPP, "Study on new radio (nr) to support non-terrestrial networks," 2019.