Machine Intelligence Research
www.mi-research.net

19(2), April 2022, 127-137
DOI: 10.1007/s11633-022-1326-3

Satellite Integration into 5G: Deep Reinforcement Learning for Network Selection

Emanuele De Santis Alessandro Giuseppi Antonio Pietrabissa
Michael Capponi Francesco Delli Priscoli

Department of Computer, Control and Management Engineering "Antonio Ruberti", University of Rome La Sapienza, Rome 00185, Italy

Abstract: This paper proposes a deep-Q-network (DQN) controller for network selection and adaptive resource allocation in heterogeneous networks, developed on the ground of a Markov decision process (MDP) model of the problem. Network selection is an enabling technology for multi-connectivity, one of the core functionalities of 5G. For this reason, the present work considers a realistic network model that takes into account path-loss models and intra-RAT (radio access technology) interference. Numerical simulations validate the proposed approach and show the improvements achieved in terms of connection acceptance, resource allocation, and load balancing. In particular, the DQN algorithm has been tested against classic reinforcement learning one and other baseline approaches.

Keywords: Network selection, HetNet, deep reinforcement learning, deep-Q-network (DQN), 5G communications.

Citation: E. De Santis, A. Giuseppi, A. Pietrabissa, M. Capponi, F. Delli Priscoli. Satellite integration into 5G: Deep reinforcement learning for network selection. *Machine Intelligence Research*, vol.19, no.2, pp.127–137, 2022. http://doi.org/10.1007/s11633-022-1326-3

1 Introduction

The exponential increase in bandwidth, coverage, and data rate demands, along with the diversification of use cases that are planning to use cellular radio access networks (RANs) to provide connectivity, has prompted the development of the fifth-generation (5G) radio access technology (RAT). Through the support for higher mobile bandwidths complemented with low latency and more reliable communications, the 5G RAT is expected to address the significant increase in data rate demands that network operators are expecting and to support the diversification of services required by user equipment (UE) during the coming years. Moreover, the 5G specifications, starting with [1], will include other RATs in the 5G environment, such as 4G long term evolution (LTE) and satellite access points (APs). In this system, where the connection demand continues to increase, appropriate network resources management is required since an optimal allocation of those resources will guarantee better performances and will help to ensure user requirements in terms of quality of experience (QoE) without overloading the network.

In this paper, a network selection technique relying on Markov decision processes (MDPs) and deep-Q-network (DQN) algorithm[2] has been studied. A centralized controller will take care of allocating requests in the best way coming from UE analyzing the network state in terms of APs load and UE perceived transmission power. The goal of this study is to show the effectiveness of the proposed deep reinforcement learning approach by simulations with a realistic multi-RAT (5G/4G/Satellite) network scenario. Moreover, several classes of user requests have been modeled in order to represent different connection service requirements in terms of downlink bitrate, quality of service (QoS) requirements, and QoE profiles.

The remainder of the paper is organized as follows. Section 2 provides an overview of the state of the art and the main contributions of the paper. In Section 3, a sketch of the control algorithm is presented, while in Section 4, some preliminaries on MDPs and DQN are introduced. In Section 5, the problem modelling is discussed. Section 6 reports the simulation results and the validation of the proposed algorithm. Finally, Section 7 draws the conclusions and highlights future works.

2 State of the art, innovations, and limitations of the proposed approach

Network selection plays a fundamental role in providing stable connections with an adequate level of QoS. Hence, network operators and providers commonly exploit several advanced techniques to select the best AP to allocate new connections. Among the various techniques

Research Article
Manuscript received November 22, 2021; accepted March 1, 2022
Recommended by Associate Editor Mao-Guo Gong
Colored figures are available in the online version at https://link.springer.com/journal/11633

proposed in the literature, multiple attribute decision making (MADM) proved to be one of the most flexible solutions to capture user preferences and QoE-related aspects in the decision process[3–7]. In MADM solutions, the information characterizing the decision-making is made by the so-called attribute values and attribute weights: The first ones describe characteristics, qualities, and performances of different alternatives, whereas the latter ones are used to measure the relevance of attributes.

By modeling the network selection problem as an MADM, it is then possible to decide the trade-off among service QoS requirements, user preferences, and overall network congestion.

A similar approach is followed in the present work, in which a different QoE profile is associated with the various connections, depending on its specific service characteristics.

Among other solutions, we mention fuzzy logic approaches[8–11], a methodology that allows fast decision making, but relies heavily on the operator′s knowledge and best practices, and game theory[12–16].

In game theory based approaches, the problem is modelled as a set of players/agents coupled with a set of network states and possible agent actions, commonly utilizing the MDP framework[17]. The main idea behind this method is that the player′s actions are influenced by the choices and actions of the other players. The interaction among the players can either be adversarial, i.e., each agent tries to maximize its performance, or cooperative, when agents share a common objective.

The approaches mentioned so far are typically employed in scenarios in which the controller is provided with a model of the network and user behavior, such as a statistical distribution of the incoming connection requests and QoE profiles, like in [18, 19], where the authors studied how to maximize QoE/QoS for specific services (e.g., video streaming applications). On the contrary, this work employs reinforcement learning (RL)[17], a model-free control methodology that allows the network controller to automatically acquire the knowledge on the system by interacting with it and experiencing its response to different control policies.

RL has been extensively applied in the network control domain[20–25] and has become particularly appealing over the last few years due to the innovations bought by its deep learning based variant, namely deep reinforcement learning (DeepRL)[2], that allowed RL-based controllers to address previously challenging problems due to their complexity and high dimensionality[26]. Deep RL has also been used for network selection and radio resource assignment, respectively[21, 27]. This paper differs from these two works because it aims at maximizing the user′s perceived QoE in a multi-RAT environment, where multiple radio access technologies are available at the same time. For multi-RAT network control, deep learning approaches (e.g., using long short-term memory (LSTM)) have been used in [28], which focuses on the cloud-edge computation offloading in satellite-UAV-served 6G networks.

The main contributions of this work are:

1) The design of a two-step network control algorithm based on deep reinforcement learning for the problem of network selection and optimal resource management in the heterogeneous 5G networks setting also envisaging the presence of satellite communication systems.

2) The inclusion in such a control framework of QoE maximization by considering three different service types with different QoS-QoE relations.

3) The development of an open-source network simulator[29] able to model several different radio access technologies, including satellite systems, in terms of network resource usage.

3 Sketch of the control algorithm

The algorithm designed in this work is a 2-step process: First, the controller that governs the RAN receives a connection request and determines which available AP it should be allocated. The AP reserves the allocation of the network resources needed to satisfy the connection minimum QoS requirements to guarantee service provision. Then, the distributed controllers that oversee the various APs distribute the remaining network resources to the connections they sustain to improve the QoE of their users. Fig. 1 reports a functional diagram of the proposed control scheme, highlighting the flowchart of the algorithm and the related data flow.

The first part of the proposed control algorithm will be based on a deep reinforcement learning agent, whereas the network resource allocation will distribute the available resources over the various connections according to their priority.

Section 4 provides the reader with the necessary background information on MDP and DeepRL.

4 Markov decision process, Q-learning, and deep-Q-network

An MDP is defined as the tuple $\{S, A, T, R, \Sigma, \gamma\}$, where S and A are the (continuous or discrete) finite state and action set, respectively, T is the transition probability function $T : S \times A \times S \to [0, 1]$, with $T(s, a, s')$ denoting the probability that the next state is s' when the current state is s and the chosen action is a, and with $\sum_{s' \in S} T(s, a, s') = 1$, R is the one-step reward function $R : S \times A \times S \to \mathbf{R}$, Σ is the in initial state distribution, and $\gamma \in (0, 1)$ is the discount factor that weights future rewards against immediate ones. The set of actions might be state-dependent as not all the actions might be avail-

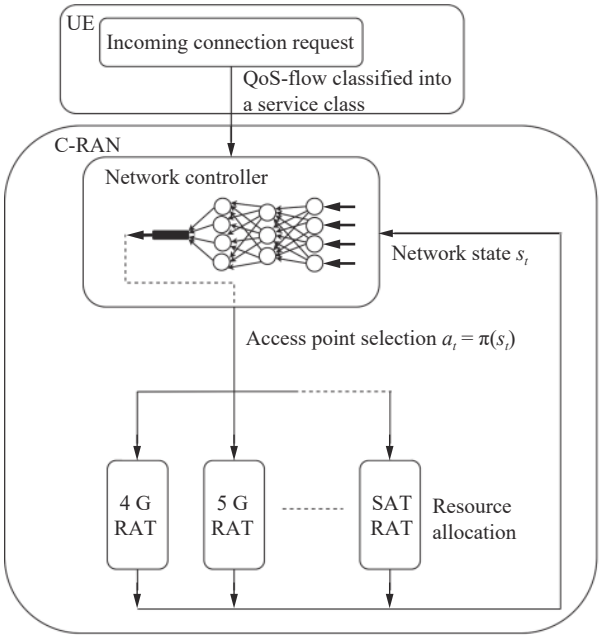

Fig. 1 Flow-chart of the control algorithm

able at each state, the set of actions available at a given state $s \in S$ will be denoted by $A(s) \subseteq A$.

A deterministic policy $\pi : S \to A$ selects one action for each state. Let Π be the set of feasible policies π such that $\pi(s) \in A(s)$ for all $s \in S$. The expected discounted reward obtained by starting from state s and following policy π thereafter is represented by the state-value function, defined as

$$V_\pi(s) = \mathrm{E}_\pi\left(\sum_t \gamma^t R(s_t, a_t, s_{t+1}) \,|s_0 = s\right) \tag{1}$$

where E_π is the expected value under policy π and s_t, and a_t represents the state and action at time t. Similarly, the state-action-value function

$$Q_\pi(s, a) = \mathrm{E}_\pi\left(\sum_t \gamma^t R(s_t, a_t, s_{t+1}) \,|s_0 = s, a_0 = a\right) \tag{2}$$

represents the expected discounted reward obtained by following the policy π when starting from state s and taking action $a \in A(s)$.

Solving the MDP means to find the optimal policy π^* that maximizes the expected cumulative discounted reward, i.e., $\pi^* = \mathrm{argmax}_{\pi\in\Pi} V_\pi(s)$. Dynamic programming approaches[17] can be used to determine exactly π^*. However, they typically require the complete knowledge of the MDP dynamics – in particular of T and R – and their computing time exponentially increases with the dimensions of state and action sets.

Conversely, RL algorithms, such as Q-learning, aim to obtain an estimate of the optimal state-action-value function Q_{π^*} based on the experience the controller gathers by interacting with the environment.

The standard update rule for Q-learning is

$$Q(s_t, a_t) = (1 - \alpha_t) Q(s_t, a_t) + \alpha_t(r_t + \gamma \max_{a\in A(s_{t+1})} Q(s_{t+1}, a)) \tag{3}$$

where $r_t = R(s_t, a_t, s_{t+1})$ is the measured reward obtained at time t and $\alpha_t > 0$ is the learning rate, which, in order to assure convergence, is subject to the conditions $\sum_{t=1}^{\infty} \alpha_t = \infty$ and $\sum_{t=1}^{\infty} \alpha_t^2 < \infty$.

The balancing between exploration and exploitation is controlled by the parameter $\varepsilon_t \in [0, 1]$ in the so-called ε-greedy policies: At any time t, the agent chooses a random action with probability ε_t, whereas it chooses the action that maximizes the state-action-value function (i.e., $\mathrm{argmax}_{a\in A(s)} Q(s, a)$) with probability $(1 - \varepsilon_t)$.

It is worth noting that in standard RL approaches, the Q function is updated only for the visited state-action pairs. Thus, in order to have a complete estimation of the optimal Q function, it is needed to visit at least once every state-action pair. This implies that the state space S and the action space A must be finite and discrete, and if their dimensions increase, RL algorithms also incur the so-called curse of dimensionality.

To address these issues, the DQN algorithm was proposed in [2] as a deep learning solution for function approximation-based Q-learning[17]. DQN approximates the Q function by means of a deep neural network able to approximate high-dimensional functions with a low-dimensional representation. The training process for the neural network is detailed in [2], and despite having included some technical solutions to address the neural network limitations, such as the target network and memory buffers, conceptually it remains the same as in the standard Q-learning, with (3) replaced by the neural network training process and in particular by the weight updates.

The main advantage of using DQN is its ability to cope with continuous state spaces, and it proved capable of solving complex problems, such as playing video games. Note that DQN still considers discrete action sets, actor-critic solutions such as the deterministic deep policy gradient (DDPG) should be used when dealing with continuous actions.

5 Problem modelling

This section presents the modelling of the network selection problem as an MDP. In particular, Sections 5.1–5.3 formulate the sets and functions required for the MDP formalism, while Sections 5.4 and 5.5 detail the physical processes that allow the conversion of network resources into bitrate provision.

Let I be the set of UEs connected within a RAN constituted by a set P of APs. Each UE $i \in I$ is connected to

an AP $p \in P$ of the RAN, characterized by a certain amount W_p of physical resource blocks (PRBs) available. In addition, let $P^i \subseteq P$ be the set of APs available at UE i, depending on its position and antennas. Moreover, let K be the set of different service types considered, each one characterized by a different minimum bitrate $B_k, k \in K$. Finally, let n_{pk} be the number of requests of type k allocated to an AP p.

Three different types of services are considered here, as in [30], namely: elastic, non-elastic, and multi-codec, each characterized by a different QoE profile.

Let b^i_{pk} be the bitrate allocated on AP p for the service k requested by the UE i. We can model the three QoE profiles as the functions $r^i_{pk}\left(b^i_{pk}\right)$ depicted in Figs. 2–4. In particular:

1) Elastic services have a linear QoE behavior with respect to the allocated bitrate, starting from a minimum level b^1_k up to a maximum bitrate b^2_k, where the perceived quality is saturated, as depicted in Fig. 2. This service captures applications such as web surfing and file downloading.

2) Non-elastic services have a threshold-like behavior with respect to the allocated bitrate. Thus, if the bitrate is less than b^1_k, the perceived quality is 0; otherwise it is maximal, as depicted in Fig. 3. This service type represents well real-time applications with guaranteed bitrate requirements.

3) Multi-codec services have a stair-like QoE profile, as the perceived quality has different thresholds corresponding to the utilized codec, depending on the amount of bitrate allocated b^1_k, b^2_k, b^3_k, as reported in Fig. 4. This service type represents multi-codec video and audio streaming.

The proposed modelling of the services is compliant with the 5G standards, as the so-called QoS-flows that constitute the various connections can be associated with one of the three service types introduced above depending on their QoS requirements and characteristics.

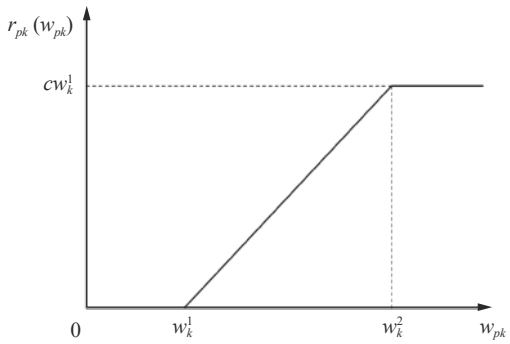

Fig. 2 QoE profile of elastic services (k=1)

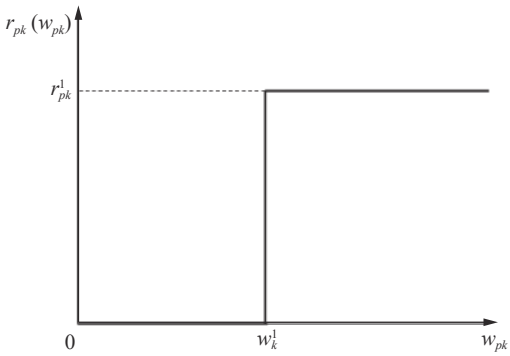

Fig. 3 QoE profile of non-elastic services (k=2)

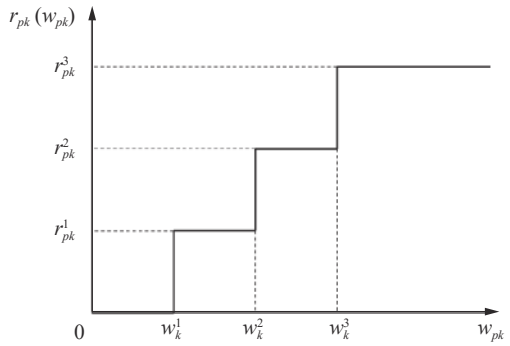

Fig. 4 QoE profile of multi-coded services (k=3)

5.1 State space definition

As already introduced, each AP is characterized by the number of its physical resources available for allocation, denoted as W_p, $p \in P$.

To allow the controller to take an optimal decision on the allocation of a new incoming connection request from a given UE, the state of the network should contain information regarding: 1) the congestion level of the physical resources over the various APs; 2) the coverage quality that the APs provide to the UE; 3) the service class, to infer its associated QoE profile, and its bitrate requirements.

In this sense, the minimum quantity of physical resources that need to be allocated to sustain a single QoS-flow i of type k on a given access point p is denoted as w^i_{pk}, with $i \in I_{pk}$, where I_{pk} is defined as the set of QoS-flows of type k related to AP p. Note that, referring to Fig. 2–4, this quantity represents the number of resources needed to provide the UE with a connection with an associated bitrate b^1_k.

Let $\eta^1_p(t)$ denote the number of resources allocated at time t to sustain the allocated services (i.e., the number of physical resources required to support the on-going QoS-flows at their minimum bitrate level). By definition,

$$\eta^1_p(t) = \sum_{k \in K} \sum_{i \in I_{pk}} w^i_{pk}(t), \quad p \in P. \tag{4}$$

Let $l_p(t)$ be the load level of an AP p, defined as the allocated physical resources over the total available ones:

$$l_p(t) = \frac{\eta_p^1(t)}{W_p}, \quad p \in P. \tag{5}$$

Given a UE $i \in I$ requesting a service of type $k \in K$, the state space is then given by the following three quantities:

1) The load level related to each AP $p \in P$;

2) The reference signals received power (RSRP) value $\mathcal{P}_{i,p}$ for each AP p, measured by the UE itself;

3) The minimum amount of bitrate required for the requested service class B_k (b_k^1 in Fig. 2–4).

The state set can then be defined as

$$S = \left\{ s = \left\{ (l_p)_{p \in P'} \, (\mathcal{P}_{i,p})_{i \in I,\, p \in P'} \, (B_k)_{k \in K} \right\} \right\}. \tag{6}$$

The resulting state $s \in S$ is a vector with $2|P|+1$ elements. With little abuse of notation, we will denote by $l_p(s), \mathcal{P}_{i,p}(s)$, and $B_k(s)$ as the load level of AP p, the RSRP value, and the minimum required amount of bitrate in state s, respectively.

5.2 Action space definition

When a new connection request arrives to the network controller, there are two possible outcomes: 1) The controller accepts the request and allocate it to (exactly) one AP p. 2) The connection is rejected as there are no APs that can handle it due to insufficient resources. The RAN controller is then required to act as an advanced connection and admission controller (CAC).

Now we define the action set similarly to [30]. Let δ_p be a vector with $2|P|+1$ values, i.e., the same dimension of the state vector $s \in S$, where all the values are zeros, but the element associated with the AP p. The single non-zero element in δ_p represents the extra load that would be added to access point p in case the new connection request is accepted. It follows that, in each state s, a request service may be allocated on AP p if and only if $s + \delta_p \in S$, i.e., by allocating the new request to the AP p, the newly generated state still belongs to S.

The action set available in a state $s \in S$ is then defined as

$$A(s) = \left\{ (\zeta_1, \zeta_2, \ldots, \zeta_{|P|}) \,\middle|\, \sum_{j \in 1, \cdots, |P|} \zeta_j = 1, \zeta_j \in \{0,1\}, \forall j \right\} \bigcup \mathbf{0} \tag{7}$$

where $\mathbf{0}$ is a P-vector of zeroes, and the action is a vector whose only non-zero element is equal to one and indicates which AP has been selected for the allocation. The special case in which $a_i = \mathbf{0}$ represents a condition in which the connection request must be rejected due to a lack of network resources, as no AP can allocate the incoming request assuring its minimum required bitrate.

In the simulation in Section 6, we will assume that requests of the type of service $k \in K$ for each UE arrive according to a Poisson distribution in time with mean value v_k and that their termination rates follow an exponential distribution with mean termination frequency μ_k.

5.3 Reward function definition

In the presented definition of the states and actions, it was assumed that the network controller only allocates the network resources needed to satisfy the minimum amount of bitrate required by the various connections. As introduced in Section 3, the network control algorithm follows a two-step procedure: Firstly, it selects which AP will serve the incoming connection request. Then, each AP distributes its remaining resources $\eta_p^1(s)$ over its connections, according to some prioritization order that may take into account the user tariff or operator preferences.

In our simulations, the APs will firstly distribute their available resources uniformly to the multi-codec services so that each connection receives a bitrate up to b_3^3. Afterward, the remaining resources are uniformly distributed to the elastic services up to a bitrate of b_1^1. Non-elastic services, due to their threshold-like behavior, are always given a bitrate of b_2^1.

To define the reward function, we have to introduce S_{pi} as the amount of additional bitrate that the AP p is able to provide to the connection i using a share of its remaining resources. This quantity is directly linked to the QoE profile associated to the connection. As it is possible to notice from Figs. 2–4, the QoE obtained by the allocation depends on the minimum bitrate allocated by the DQN algorithm $b_k^1 + \mathcal{S}_{pi}$, i.e., the total bitrate available to the service i of class k.

The reward function shall then capture three cases:

1) The connection request is rejected (i.e., no AP allocates the connection).

2) The connection is allocated on an AP with a low resource usage.

3) The connection is allocated on an AP that is already providing several other connections.

To capture those three cases, the reward $r_t(s_t, a_t, s_{t+1})$ obtained by the controller when allocating a connection i of class k of AP p can be defined as

$$r_t(s_t, a_t, s_{t+1}) = \begin{cases} -r^0 < 0, & \text{if } a_t = \mathbf{0} \\ r_{pk}\left(b_k^1 + \mathcal{S}_{pi}\right), & \text{if } l_p(t+1) \le 0.5 \\ r_{pk}\left(b_k^1 + \mathcal{S}_{pi}\right) - r^{sat}, & \text{if } l_p(t+1) > 0.5. \end{cases} \tag{8}$$

The negative reward $-r^0$ represents a penalty given to the agent if the allocation is rejected to capture the cost incurred by the network operator in failing to provide a connection. The term $r_{pk}\left(b_k^1 + \mathcal{S}_{pi}\right)$ is a positive reward, shaped depending on k as in Fig. 2–4, that captures the QoE of the new user, and the term $-r^{sat}$ is a negative reward subtracted from $r_{pk}\left(b_k^1 + \mathcal{S}_{pi}\right)$ in case the new alloc-

ation is destined to an AP whose saturation level is higher than the desired threshold (50% in our case).

The long-term maximization of this reward allows the network controller to maximize the overall QoE of its users while keeping the connection rejection rate minimized.

5.4 5G NR and 4G LTE resource allocation description

In order to relate the physical resources that appear in the state definition with the transmission bitrate needed by the reward function to estimate the QoE level, it is now necessary to detail their relationships and how one translates into the other for both terrestrial and satellite APs.

5G new radio (NR) APs have a limited set of resources[31], both in terms of frequency bandwidth and time to allocate UE requests. The minimum allocation unit for a 5G NR AP is the PRB, each composed of 12 frequency subcarriers with a $2^{\mu}\times15$ kHz bandwidth and a time duration of $2^{-\mu}\times1$ ms, where $\mu\in\{0,1,2,3,4\}$ is the parameter called numerology defined by 5G NR standards. The number of PRBs available on AP p depends on the total available bandwidth on the AP and its numerology, as defined by 5G NR standards[31].

For 4G LTE APs, the definition of PRB still stands, but the numerology parameter is constrained to $\mu=0$, so there is no flexibility on using less/more subcarrier bandwidths and more/less time slot durations. Even if 4G LTE will likely to be replaced by 5G NR in the next few years, it has been considered in this work since it is currently the predominant radio access technology for mobile devices, and its seamless integration in the multi-connectivity framework allows for more stable and broadly available connectivity.

The receiving power, or RSRP, $\mathcal{P}_{i,p}$ that appears in the states of (6) represents the transmission power measured by the UE $i\in I$ between itself, and the AP $p\in P$ is computed as follows:

$$\mathcal{P}_{i,p}=\mathcal{P}_p\times G_p\times L_p\times L_{i,p} \tag{9}$$

where $\mathcal{P}_p$ is the AP's antenna power, G_p is the AP's antenna gain, L_p is the AP's feeder losses, and $L_{i,p}$ is the path loss between UE i and AP p.

In our simulations, the path loss $L_{i,p}$ is computed through the COST-HATA model[32] which is a statistical model that considers many factors as the building density (rural, suburban, urban), the carrier frequency used for the communications, and the relative heights of UE and AP.

In order to estimate the number of resource blocks to be allocated by the AP $p\in P$ for the communication with the UE $i\in I$, the signal-over-interference-plus-noise-ratio (SINR) has to be computed. The thermal noise part can be computed according to:

$$\mathcal{N}_p=k_bT^{env}B_p\Theta_p \tag{10}$$

$$\Theta_p(t)=\frac{\sum_{\tau\in(t-T,t)}\sum_{j\in I\setminus i}C_{j,p}(\tau)N_{j,p}(\tau)}{T\times\#R_p} \tag{11}$$

where $\Theta_p(t)$ is the resource blocks utilization ratio (RBUR) of AP p at time t, k_b is the Boltzmann constant, T^{env} is the environmental temperature, B_p is the total bandwidth for the AP p, T is the length of the moving average, $C_{j,p}(t)$ is equal to 1 if UE j is connected to AP p at time t and 0 otherwise, and $N_{j,p}(t)$ is the number of PRB allocated by AP p to UE j, and $\#R_p$ is the total number of resource blocks of AP p.

The interference part is computed as follows:

$$\mathcal{J}_{i,p}=\sum_{p'\neq p}F_{p,p'}P_{i,p'}\times\Theta_{p'}(t) \tag{12}$$

where $F_{p,p'}$ is 1 if AP p and p' share the same carrier frequency and 0 otherwise.

Using (10) and (12), it is possible to compute the SINR, and so it is possible to estimate the data rate that can be transmitted by allocating one PRB to UE i using the Shannon formula:

$$r_{i,p}=2^{-\mu}10^{-3}B_{PRB}\log_2(1+SINR_{i,p}) \tag{13}$$

where B_{PRB} is the bandwidth of a single PRB, and it can be computed as $B_{PRB}=12\times2^{\mu}15$ kHz.

Now, given a certain bitrate request b_p^i from UE i, it is possible to compute the number of resource blocks to be allocated by AP p to satisfy the request: $n_{i,p}^{PRB}=\lceil(b_{pk}^i/r_{i,p})\rceil$.

5.5 Satellite resource allocation description

Contrary to ground APs, the satellite APs use time division multiple access (TDMA) in order to serve multiple UEs at the same time. In this case, the minimum allocation unit is a block of symbols that occupies a certain time slot in the satellite time frame.

The receiving power $P_{i,p}$ can be still computed as (9), but in this case, the path loss function will be the free space path loss:

$$L_{i,p}^{FSPL}=\left(\frac{4\pi d_{i,p}f}{c}\right)^2 \tag{14}$$

where $d_{i,p}$ is the Euclidean distance between UE i and AP p, f is the carrier frequency used, c is the speed of light, and $FSPL$ represents the free space path loss.

The thermal noise can be computed as (10), and the

interference can be computed as (12). Using the Shannon formula (considering that this time the bandwidth is the total bandwidth of the satellite AP since the TDMA utilizes all the bandwidth only for a certain amount of time), one has that the bitrate obtainable by a single block of symbols is

$$r_{i,p} = bB\log_2(1 + SINR_{i,p}) \tag{15}$$

where b is the ratio between the number of symbols in a single block and the total number of symbols of the satellite AP. The number of blocks to be allocated for a requested bitrate b^i_{pk} from UE i is then computed as $n^{blocks}_{i,p} = \left(b^i_{pk}/r_{i,p}\right)$.

6 Simulation results and validation

In order to demonstrate the effectiveness of the proposed approach, a simulative environment has been built up according to the model definition previously introduced.

6.1 Scenario definition

We developed a scenario consisting of four terrestrial access points (NR1 and NR2 are 5G NR APs and the remaining two are 4G LTE APs) and a satellite access point in a 2.5 × 2.5 km^2 area, as shown in Fig 5.

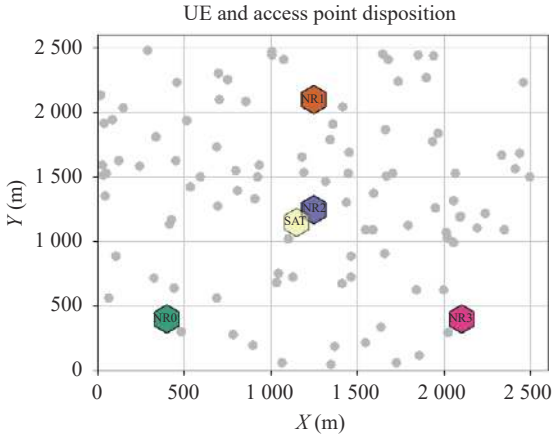

Fig. 5 Considered network scenario

In particular, for 5G NR access points, we considered a carrier frequency of 1.7 GHz (band n66) with numerology $\mu = 2$, while for 4G LTE access points, we considered a carrier frequency of 800 MHz (band 20). All the terrestrial APs have 20 dB power, 16 dB antenna gain, and 3 dB feeder losses. For the satellite access point, we considered the Inmarsat implementation from Example 6.6.2 of [33]. A total of 100 UEs has been considered in the given area. Each of them follows a Poisson distribution for requesting data with a certain service type and for the duration of such request; the parameters for each service type are described in Table 1. Moreover, we considered $\gamma = 0.9$, $\varepsilon = 1$, ε-decay = 0.999 5 and ε-min = 0.01. As for the DQN parameters, we considered a replay buffer of 2 000 tuples, a batch size of 64 tuples, and the update of target network weights every 50 steps. Finally, the DNN hidden layers have a tanh activation function, the learning rate of the DNN is 10^{-4}, and the network performs 4×10^4 training steps before finishing the training. The training process and its testing using the proposed radio access network simulator run on an Intel Core i7 6700HQ machine with 16 GB RAM. No dedicated GPU has been used for the training process since the small size of the processed data makes the training step faster than copying such data from RAM to VRAM. Most of the computation complexity is, of course, in the training process of the four hidden layers of the DQN network; once trained, DQN has $O(1)$ computational cost to compute the best action.

Table 1 Service type requests

	Elastic	Non-elastic	Multi-codec
Bitrate (Mbps)	10	200	100
Arrival rate (s)	2	6	4
Dwelling time (s)	30	120	90

6.2 Simulation results

The results displayed in Figs. 6–12 will focus on the performance of the controller in terms of QoS-flows allocation and their management. In order to validate the results of the proposed DQN algorithm, a set of other approaches have been simulated. In particular, a classical, tabular, Q-learning (QL in Figs. 6–11) approach has been simulated, together with a least loaded (LL in Figs. 6–11) approach, where a new request will be allocated to the least loaded AP, and a Max-RSRP (MR) approach, where a new request will be allocated to the AP with the maximum receiving power. The Q-learning approach shares the same MDP representation as the one presented for the DQN, except for the fact that the state-space needed to be discretized so that the AP loads and the RSRP values contained in the states in (6) were uniformly quantized into four levels.

The various controllers have been tested on the same scenarios to obtain fair performance results. Moreover, to ensure more balanced experiments, the results are the average between ten different scenarios, each tested by all the different controllers. Finally, both the DQN and the QL controllers have been trained before executing the simulations. Several metrics are shown to understand better the performances of the controllers with respect to each other.

As it emerges from Fig. 6, the DQN controller outperforms the other controllers in terms of rejection rate, even

if both the Max-RSRP (MR) one and the Q-learning (RL) one have similar results. This behavior is not surprising since the Max-RSRP approach allocates requests to the AP with minimum path-loss, so the number of requested physical resources will be in general lower, and the Q-learning approach has a similar behavior w.r.t. the DQN approach, since the only difference is in its finite state space.

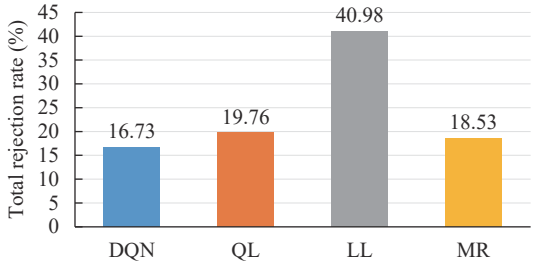

Fig. 6 Overall rejection rates

Fig. 7 reports the rejection rate of each controller divided by service type. In Fig. 7, we can note how all controllers allocate a lower percentage of the non-elastic service requests, whereas the LL controller shows a significantly higher rejection rate for the elastic services.

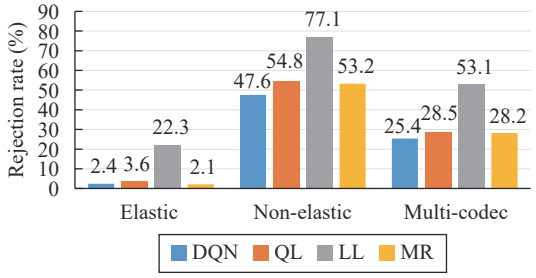

Fig. 7 Rejection rates divided by service type

In terms of bitrate, the DQN approach is found to be the best one, allocating around 48 Gbit over the accepted incoming requests.

Fig. 8 details the allocated bitrate percentage with respect to the total requested bitrate divided by service type.

The result demonstrates that DQN behaves almost in the same way as MR for what concerns the elastic services, while it allocates about 6% more than the other approaches for what regards non-elastic traffic and about 3% for what regards the multi-codec requests.

In addition, from Fig. 9, which represents the average percentage of successful allocations in each AP on all the requests made by the UEs, it is evident that the least-loaded controller is the one that better balances the load between the APs. Despite its limited performances according to the other metrics presented, due to its definition, it allocates requests to the least used AP at the given time instant, resulting in an overall reasonable balance among all the APs.

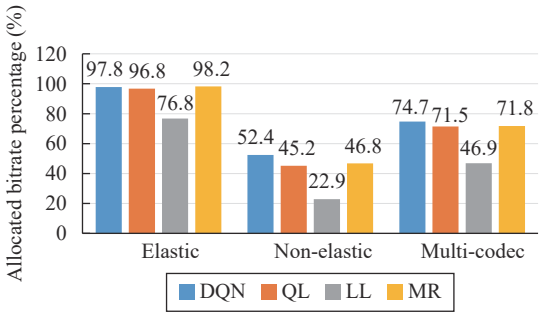

Fig. 8 Allocated bitrate percentage divided by service type

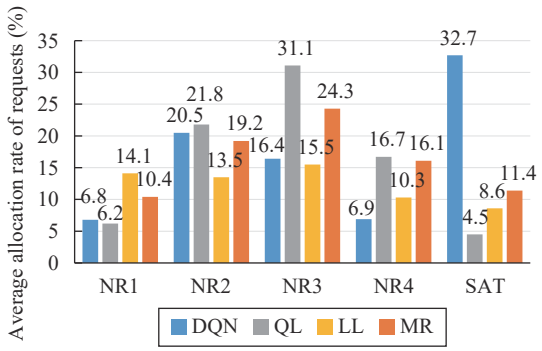

Fig. 9 Load distribution among each AP

The other controllers appear to be less balanced when allocating resources, with one or two base stations being exploited more than the others. In particular, the DQN controller relies heavily on the satellite base station to allocate incoming requests, allocating about 30% of requests to this AP. DQN is hence the only approach that fully exploits satellite resources, as the others tend to utilize mainly the NR base stations.

Fig. 10 represents the QoE collected by each of the controllers. The values for each controller are computed by summing the QoE gained by each request according to the QoE profiles defined in Section 3 and then normalized on the result obtained by the DQN controller. As expected, the Q-learning controller has similar performances with respect to the DQN one, reaching the highest QoE level. The performance gap increases when comparing a learning-based agent against the other approaches.

Finally, Fig. 11 represents the QoE collected by each controller in case the number of UEs is less than 100. From Fig. 11, it is possible to notice, as expected, that the DQN controller is able to achieve better performances compared to the competitor controllers when the number of UEs is smaller, while, if the number of UEs increases, the network is more likely saturated, making the DQN, Q-learning and MR approaches gain similar levels of QoE. In fact, the rejection rates of all approaches increase as the network overload increases. However, the DQN and the Q-learning controllers continue to prefer rejecting non-elastic traffic in favor of multi-codec and

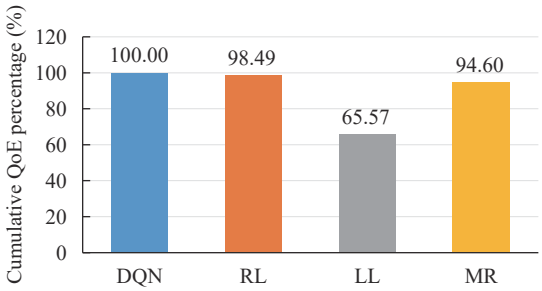

Fig. 10 Cumulative QoE gained by each of the controllers with respect to DQN controller

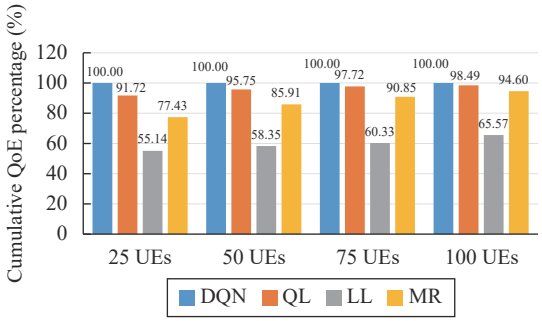

Fig. 11 Cumulative QoE with different numbers of user equipment, normalized on the corresponding DQN performance

elastic service classes. As shown in Fig. 12, the cumulative QoE of the DQN approach, normalized with the QoE of the case with 100 UEs, still increases (in a sub-linear way) as the number of UEs increases. This is because, even if the network is going towards saturation, the controller is still able to allocate some more UEs w.r.t. the cases with fewer UEs.

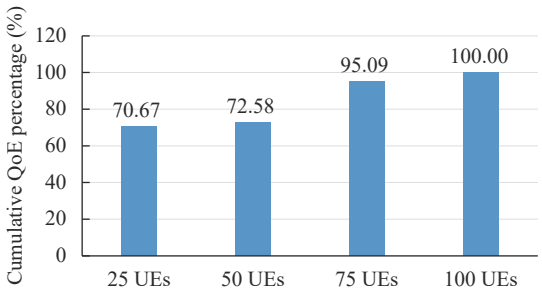

Fig. 12 Cumulative QoE percentage for DQN approach with different numbers of user equipment, normalized on the DQN performance with 100 UEs

7 Conclusions

The paper proposed a network controller based on deep reinforcement learning to enable the integration of satellite systems into 5G heterogeneous networks. The proposed controller dealt with the problem of network selection by formulating it as a Markov decision process and was compared to several standard benchmark algorithms. The proposed solution proved to be able to cope with large-scale scenarios involving 100 different UEs.

For validation purposes, the authors developed an open-source network simulator[29] that realistically captures the network resource usage of different radio technologies, including satellite connections.

Overall, the proposed controller improved the performance of the network, increasing the connection-flow acceptance rate and providing better resource management compared to the other methods tested.

Future works are related to the introduction of other unmodeled complexities in the simulator, such as user and access point mobility. Actor-critic algorithms[26] will also be explored to enable the split of QoS-flows and multi-connectivity, allocating a single flow over different access points at the same time.

Acknowledgements

This work was supported by the European Commission in the framework of the H2020 EU-Korea project 5G-ALLSTAR (5G AgiLe and fLexible integration of SaTellite And cellulaR, www.5g-allstar.eu) (No. 815323). The authors acknowledge all their colleagues of the Consortium for the Research in Automation and Telecommunication (CRAT) team working on the project for their fruitful discussions and confrontations.

References

[1] 3GPP TR 38.811 Study on New Radio (NR) to Support Non-terrestrial Networks, Technical Report. 3GPP, France 2017.

[2] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, M. Riedmiller. Playing Atari with deep reinforcement learning, [Online], Available: https://arxiv.org/abs/1312.5602v1, 2013.

[3] K. S. S. Anupama, S. S. Gowri, B. P. Rao. A comparative study of outranking MADM algorithms in network selection. In *Proceedings of the 2nd International Conference on Computing Methodologies and Communication*, IEEE, Erode, India, pp. 904–907, 2018. DOI: 10.1109/ICCMC.2018.8487931.

[4] Y. F. Zhong, H. Q. Wang, H. W. Lv. A cognitive wireless networks access selection algorithm based on MADM. *Ad Hoc Networks*, vol. 109, Article number 102286, 2020. DOI: 10.1016/j.adhoc.2020.102286.

[5] S. Radouche, C. Leghris, A. Adib. MADM methods based on utility function and reputation for access network selection in a multi-access mobile network environment. In *Proceedings of International Conference on Wireless Networks and Mobile Communications*, IEEE, Rabat, Morocco, 2017. DOI: 10.1109/WINCOM.2017.8238177.

[6] Q. Y. Song, A. Jamalipour. Network selection in an integrated wireless LAN and UMTS environment using mathematical modeling and computing techniques. *IEEE Wire-*

less Communications, vol. 12, no. 3, pp. 42–48, 2005. DOI: 10.1109/mwc.2005.1452853.

[7] T. Ding, L. Liang, M. Yang, H. Q. Wu. Multiple attribute decision making based on cross-evaluation with uncertain decision parameters. *Mathematical Problems in Engineering*, vol. 2016, Article number 4313247, 2016. DOI: 10.1155/2016/4313247.

[8] R. K. Goyal, S. Kaushal, A. K. Sangaiah. The utility based non-linear fuzzy AHP optimization model for network selection in heterogeneous wireless networks. *Applied Soft Computing*, vol. 67, pp. 800–811, 2018. DOI: 10.1016/j.asoc.2017.05.026.

[9] X. Y. Yan, P. Dong, T. Zheng, H. K. Zhang. Fuzzy and utility based network selection for heterogeneous networks in high-speed railway. *Wireless Communications and Mobile Computing*, vol. 2017, Article number 4967438, 2017. DOI: 10.1155/2017/4967438.

[10] M. M. R. Mou, M. Z. Chowdhury. Service aware fuzzy logic based handover decision in heterogeneous wireless networks. In *Proceedings of International Conference on Electrical, Computer and Communication Engineering*, IEEE, Cox's Bazar, Bangladesh, pp. 686-691, 2017. DOI: 10.1109/ECACE.2017.7912992.

[11] A. Wilson, A. Lenaghan, R. Malyan. Optimising wireless access network selection to maintain QoS in heterogeneous wireless environments. In *Proceedings of International Symposium on Wireless Personal Multimedia Communications*, Aalborg, Denmark. 1236−1240, 2005.

[12] R. Trestian, O. Ormond, G. M. Muntean. Game theory-based network selection: Solutions and challenges. *IEEE Communications Surveys & Tutorials*, vol. 14, no. 4, pp. 1212–1231, 2012. DOI: 10.1109/surv.2012.010912.00081.

[13] J. Antoniou, A. Pitsillides. 4G converged environment: Modeling network selection as a game. In *Proceedings of the 16th IST Mobile and Wireless Communications Summit*, IEEE, Budapest, Hungary, 2007. DOI: 10.1109/IST-MWC.2007.4299242.

[14] T. Rahman, M. Z. Chowdhury, Y. M. Jang. Radio access network selection mechanism based on hierarchical modelling and game theory. In *Proceedings of International Conference on Information and Communication Technology Convergence*, IEEE, Jeju, Korea , pp. 126−131, 2016. DOI: 10.1109/ICTC.2016.7763451.

[15] L. Rajesh, K. B. Bagan, B. Ramesh. User demand wireless network selection using game theory. In *Proceedings of International Conference on Nano-electronics, Circuits & Communication Systems*, Jharkhand, India, pp.39−53, 2017. DOI: 10.1007/978-981-10-2999-8_4.

[16] Meenakshi, N. P. Singh. A comparative study of cooperative and non-cooperative game theory in network selection. In *Proceedings of International Conference on Computational Techniques in Information and Communication Technologies*, IEEE, New Delhi, India, pp. 612−617, 2016. DOI: 10.1109/ICCTICT.2016.7514652.

[17] R. S. Sutton, A. G. Barto, *Reinforcement Learning:* An Introduction, Cambridge, UK: MIT Press, 1998.

[18] Z. H. Zhang, X. F. Jiang, H. S. Xi. Optimal content placement and request dispatching for cloud-based video distribution services. *International Journal of Automation and Computing*, vol. 13, no. 6, pp. 529–540, 2016. DOI: 10.1007/s11633-016-1025-z.

[19] F. S. Lin, B. Q. Yin, J. Huang, X. M. Wu. Admission control with elastic QoS for video on demand systems. *International Journal of Automation and Computing*, vol. 9, no. 5, pp. 467–473, 2012. DOI: 10.1007/s11633-012-0668-7.

[20] Z. Y. Du, C. X. Wang, Y. M. Sun, G. F. Wu. Context-aware indoor VLC/RF heterogeneous network selection: Reinforcement learning with knowledge transfer. *IEEE Access*, vol. 6, pp. 33275–33284, 2018. DOI: 10.1109/access.2018.2844882.

[21] Y. Yang, Y. Wang, K. Y. Liu, N. Zhang, S. S. Gu, Q. Y. Zhang. Deep reinforcement learning based online network selection in CRNs with multiple primary networks. *IEEE Transactions on Industrial Informatics*, vol. 16, no. 12, pp. 7691–7699, 2020. DOI: 10.1109/tii.2020.2971735.

[22] D. D. Nguyen, H. X. Nguyen, L. B. White. Reinforcement learning with network-assisted feedback for heterogeneous RAT selection. *IEEE Transactions on Wireless Communications*, vol. 16, no. 9, pp. 6062–6076, 2017. DOI: 10.1109/twc.2017.2718526.

[23] F. Liberati, A. Giuseppi, A. Pietrabissa, V. Suraci, A. Di Giorgio, M. Trubian, D. Dietrich, P. Papadimitriou, F. Delli Priscoli. Stochastic and exact methods for service mapping in virtualized network infrastructures. *International Journal of Network Management*, vol. 27, no. 6, Article number e1985, 2017. DOI: 10.1002/nem.1985.

[24] X. W. Wang, J. D. Li, L. X. Wang, C. G. Yang, Z. Han. Intelligent user-centric network selection: A model-driven reinforcement learning framework. *IEEE Access*, vol. 7, pp. 21645–21661, 2019. DOI: 10.1109/access.2019.2898205.

[25] K. S. Shin, G. H. Hwang, O. Jo. Distributed reinforcement learning scheme for environmentally adaptive IoT network selection. *Electronics Letters*, vol. 56, no. 9, pp. 462–464, 2020. DOI: 10.1049/el.2019.3891.

[26] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, D. Wierstra. Continuous control with deep reinforcement learning. In *Proceedings of the 4th International Conference on Learning Representations*, San Juan, Puerto Rico, 2016.

[27] Y. B. Zhou, Z. M. Fadlullah, B. M. Mao, N. Kato. A deep-learning-based radio resource assignment technique for 5G ultra dense networks. *IEEE Network*, vol. 32, no. 6, pp. 28–34, 2018. DOI: 10.1109/MNET.2018.1800085.

[28] B. M. Mao, F. X. Tang, Y. Kawamoto, N. Kato. Optimizing computation offloading in satellite-UAV-served 6G IoT: A deep learning approach. *IEEE Network*, vol. 35, no. 4, pp. 102–108, 2021. DOI: 10.1109/MNET.011.2100097.

[29] E. De Santis. Trunk96/wireless-network-simulator, [On-

line], Available: https://github.com/trunk96/wireless-network-simulator, 2022.

[30] F. D. Priscoli, A. Giuseppi, F. Liberati, A. Pietrabissa. Traffic steering and network selection in 5G networks based on reinforcement learning. In *Proceedings of European Control Conference*, IEEE, St. Petersburg, Russia, pp. 595–601, 2020. DOI: 10.23919/ECC51009.2020.9143837.

[31] 5G; NR; Physical Channels and Modulation, ETSI TS 138 211 v15.2.0. 3GPP, 2018.

[32] Final report for COST Action 231, [Online], Available: http://www.lx.it.pt/cost231/final_report.htm, 2022.

[33] G. Maral, M. Bousquet, Z. L. Sun. *Satellite Communications Systems: Systems, Techniques and Technology*. 6th ed., Hoboken, USA: Wiley, 2020. DOI: 10.1002/9781119673811.

Emanuele De Santis received the B. Sc. degree in automatic control and M. Sc. degree in engineering in computer science with specialization in communication networks control from Sapienza University of Rome, Italy in 2017 and 2019, where he is currently a Ph. D. degree candidate in automatic control. He participated in the H2020 projects 5G-ALLSTAR and 5G-Solutions and in the European Space Agency (ESA) project ARIES. He is a student member of IEEE.

His research interests include power and communication network control, artificial intelligence, and optimal control.

E-mail: edesantis@diag.uniroma1.it (Corresponding author)
ORCID iD: 0000-0003-1011-9737

Alessandro Giuseppi received the B. Sc. degree in computer and automation engineering, the M. Sc. degree in control engineering and the Ph. D. degree in automatica from University of Rome La Sapienza, Italy in 2014, 2016 and 2019, respectively, where he is currently a postdoctoral researcher in automatic control. Since 2016, he has participated in five European and national research projects. He is a member of IEEE.

His research interests include network control and intelligent systems.

E-mail: giuseppi@diag.uniroma1.it
ORCID iD: 0000-0001-5503-8506

Antonio Pietrabissa received the M. Sc. degree in electronics engineering and the Ph. D. degree in systems engineering from Sapienza University of Rome, Italy in 2000 and 2004, respectively. He is an associate professor at Sapienza University of Rome, Italy. He has participated in about 20 European and national research projects. He is a senior member of IEEE.

His research interests include the application of systems and control theory to the analysis and control of networks.

E-mail: pietrabissa@diag.uniroma1.it
ORCID iD: 0000-0003-0188-3346

Michael Capponi received the M. Sc. degree in communication networks control from Sapienza University of Rome, Italy in 2020. Now, he works in a company in the field of computer science.

His research interests include reinforcement learning applications to communication networks.

E-mail: michaelcapponi96@gmail.com
ORCID iD: 0000-0002-5610-9422

Francesco Delli Priscoli received the M. Sc. degree in electronics engineering and the Ph. D. degree in systems engineering from University of Rome, Italy in 1986 and 1991, respectively. From 1986 to 1991, he was with Telespazio, Italy. Since 1991, he has been with University of Rome, Italy, where, at present, he is a full professor of automatic control, control of autonomous multiagent systems, and control of communication and energy networks. He is an Associate Editor of *Control Engineering Practice* and a Member of the IFAC Technical Committee on Networked Systems. He was/is the Scientific Responsible with University of Rome, for 40 projects funded by the European Union and by the European Space Agency. He is a member of IEEE.

His research interests include closed-loop multiagent learning techniques in advanced communication and energy networks.

E-mail: dellipriscoli@diag.uniroma1.it
ORCID iD: 0000-0001-6140-3661