IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received September 2, 2021, accepted September 13, 2021. Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2021.3113501

Deep Reinforcement Learning for Energy-Efficient Multi-Channel Transmissions in 5G Cognitive HetNets: Centralized, Decentralized and Transfer Learning Based Solutions

ANASTASIOS GIANNOPOULOS[1], SOTIRIOS SPANTIDEAS[1], NIKOLAOS KAPSALIS[1], PANAGIOTIS KARKAZIS[2], AND PANAGIOTIS TRAKADAS[3]

[1]School of Electrical and Computer Engineering (ECE), National Technical University of Athens (NTUA), 157 80 Athens, Greece
[2]Department of Informatics and Computer Engineering, University of West Attica, 122 43 Athens, Greece
[3]Department of Ports Management and Shipping, National and Kapodistrian University of Athens, 344 00 Athens, Greece

Corresponding author: Anastasios Giannopoulos (angianno@mail.ntua.gr)

This work was supported in part by the Affordable5G Project funded by the European Commission through the Horizon 2020 and 5G-PPP Programs (www.affordable5g.eu/) under Grant H2020-ICT-2020-1 and Grant 957317.

ABSTRACT Energy efficiency (EE) constitutes a key target in the deployment of 5G networks, especially due to the increased densification and heterogeneity. In this paper, a Deep Q-Network (DQN) based power control scheme is proposed for improving the system-level EE of two-tier 5G heterogeneous and multi-channel cells. The algorithm aims to maximize the EE of the system by regulating the transmission power of the downlink channels and reconfiguring the user association scheme. To efficiently solve the EE problem, a DQN-based method is established, properly modified to ensure adequate QoS of each user (via defining a demand-driven rewarding system) and near-optimal power adjustment in each transmission link. To directly compare different DQN-based approaches, a centralized (C-DQN), a multi-agent (MA-DQN) and a transfer learning-based (T-DQN) method are deployed to address whether their applicability is beneficial in the 5G HetNets. Results confirmed that DQN-assisted actions could offer enhanced network-wide EE performance, as they balance the trade-off between the power consumption and achieved throughput (in Mbps/Watt). Excessive performance was observed for the MA-DQN approach (>5 Mbps/Watt), since the decentralized learning supports low-dimensional agents to be coordinated with each other through global rewards. In further comparing the T-DQN against MA-DQN solutions, T-DQN presents beneficial usage for very low or very high inter-cell distances, whereas the usage of MA-DQN is preferred (by a factor of $\sim$1.3) for intermediate inter-cell distances (100-600m), where the power savings are feasible towards achieving increased EE. Furthermore, T-DQN scheme guarantees good EE solutions (above 2 Mbps/Watt), even for densely-deployed macro-cells, with effortless training and memory requirements. On the contrary, MA-DQN offers the best EE solutions at the expense of massive training resources and required training time.

INDEX TERMS 5G network, deep Q-learning, energy efficiency, radio resource management, reinforcement learning, transfer learning.

I. INTRODUCTION

The unstoppable evolution of wireless communication networks intends to provide ubiquitous, reliable and near-instant pervasive connectivity between humans and machines [1], [2]. Driven by an endless need for ever-growing data capacity, 5G cellular networks will be the bridging platform to meet unpreceded user requirements and enable Internet of Things (IoT), massive unmanned mobility, augmented reality (AR), virtual reality (VR) and Industry 4.0 applications [3]. These innovations are coupled with novel technical approaches spanning across all the 5G network layers, such as new physical-layer transmission schemes, MIMO antenna systems, routing algorithms, network

The associate editor coordinating the review of this manuscript and approving it for publication was Fakhrul Alam.

slicing, software-defined radios (SDR) and network function virtualization (NFV). In this context, massive data will be exchanged among numerous interconnected objects, with each one requesting broadband, fast and reliable communication. According to Cisco [4], it is expected that by 2030, 5G networks have to support billions of connected devices, offering coverage and capacity to every corner, while ensuring enhanced Quality of Service (QoS) and data rates [5]. As an unfavorable consequence, it is also expected that 5G networks will consume a thousand times more energy than existing systems [6]. Hence, a successful 5G system has to not only bring high-capacity and complete coverage, but also to guarantee greener and more sustainable deployments [7]–[9].

The surged growth of mobile data traffic results into overloaded communication systems. To deal with inadequacy in the network-wide capacity, there are specific planning options defined by Shannon's capacity formula [10]; the latter implies that the capacity improvement is positively related with bandwidth increase, spectral efficiency melioration and/or frequency reuse. Bandwidth and spectral efficiency increments have been thoroughly identified and addressed from day-1 of 5G deployment, mainly exploiting already known principles like carrier aggregation, cognitive radio, MIMO techniques, interference mitigation, error-correction coding and traffic adaptation [11]. Having noticed near-saturated progress in the abovementioned metrics, 5G clearly targets to increase the frequency reuse factor by densely deploying multi-tier heterogeneous cells [12], ranging from macro-area outdoor to femto-area indoor servers. This high-degree densification not only boosts the complexity of network planning, especially in the physical-layer infrastructure, but also re-poses spectral efficiency and signal-to-noise ratio degradation issues due to severe interferences. Importantly, the energy efficiency (EE) defines the extent to which the network is both throughput-sufficient (i.e. covers the requested services) and cost-efficient (i.e. prefers low-cost links and reduces the power consumption) [13]. Quantitatively, the system EE is strongly reduced when either the interferences significantly degrade the experienced throughput or the power consumption is not restricted. A key conflicting point of 5G is then identified: Towards the significant improvement in system capacity, the EE may be proven a major limiting factor.

In modern wireless systems, the optimization of a specific metric (such as capacity or EE) usually sets upper-bounds or even degrades the performance of other equally important indices. Agreeing in this conflicting groundtruth, along with the tremendous complexity of realistic wireless networks, the non-convexity nature of EE optimization has been proven [14]. Traditionally, several EE optimization algorithms, such as convex optimization, fractional programming and game theory, have been proposed showing significant drawbacks in their applicability. Notably, optimization algorithms search for analytical solutions and thus can be applied under simple system models, usually including unrealistic network parameters and many simplifications to derive solvable expressions, such as perfect channel state indications, limited interferences, low-dimensional environments, static users and so on. Moreover, a specific optimization algorithm usually targets at a single-component optimization, due to dimensionality and complexity increase when considering joint optimization of multiple metrics. Towards overcoming these limitations, machine learning (ML) algorithms offer a unique opportunity to supervise the future wireless networks, primarily due to (i) the massive data that are collected, (ii) the tremendous progress in the computing power of processors, (iii) the ability of learning extremely complex patterns (i.e. function fitting) from data and (iv) the ability to provide predictions or make decisions in near-real time.

With the exponential increase in the number of both access and demand points, mathematical models that define the network structure become inaccurate, thus raising imminent challenges in network resource management. Moving beyond traditional rule-based and brute-force methods for supervising the 5G network resources, recent concepts of self-organizing networks (SONs) and zero-touch optimization (ZTO) have acquired considerable interest in the 5G era, allowing the network configurations to be automatized and eliminating the need for expensive hands-on management. In this context, ML will be the key enabler for cognitive networks as it provides efficient learning without requiring explicit mathematical models. Deep learning (DL) based algorithms have shown promising abilities in 5G resource management given sufficient access to historically collected data [15]. However, the presence of large datasets to train DL networks still remains far away from reality, mainly due to massive storage, privacy and confidential issues among operators. On the contrary, reinforcement learning (RL) models [16] include a software agent interacting with the telecom environment with the aim of finding a (sub) optimal strategy to maximize its long-term rewards. At each training step, the agent observes the system state, takes an action, receives a scalar reward, and moves to the next state. Due to this trial-and-error approach, RL does not require training datasets and thus it is widely studied in wireless networks optimization problems. Over the last years, neural networks have been combined with RL (Deep RL or DRL) as function approximators to estimate the 'quality' (Q-value) of performing a particular action from a given state. As opposed to conventional RL, DRL agents are insensitive to large state-action space and can be applied to high-dimensional problems and under non-stationary conditions.

In this paper, a mobility-aware multi-channel power control scheme is proposed for improving the system-level EE of two-tier 5G heterogeneous cells. The algorithm aims to maximize the EE of the system by regulating the transmission power of the downlink channels and reconfiguring the user association scheme. To efficiently solve the EE problem, a DQN-based method is established, properly modified to ensure adequate QoS of each user

(via defining a demand-driven rewarding system) and near-optimal power adjustment in each transmission link. After showing that the EE optimization follows the principles of non-convex optimization problems, a three-way DQN-based approach is outlined. Specifically, a *centralized* DQN algorithm is initially employed to solve the problem of interest exploiting the global network state. Inspired by the distributed nature of multi-agent methods, the same algorithm was then implemented in a cooperative *decentralized* manner, where each individual agent has partial observability and thus low-dimensionality neural networks, to directly contrast its pros and cons in relation to the first approach. Finally, to further investigate potential complexity and cost reduction possibilities of the EE optimization framework, we explore the performance of the proposed DRL scheme using the principles of *transfer learning*. The proposed framework is tested on 5G-compliant environments with realistic user mobility considering standardized channel models [17] and assuming the simultaneous operation of macro- and micro-cells in urban conditions.

The key contributions of the proposed algorithms may be summarized as follows:

(i) The proposed algorithm is based on the model-free DRL that learns by only reading network measurements, making and correcting unprofitable moves. Such an approach does not require the presence of training datasets which raises massive storage and confidential issues in mobile operators. A trained optimal-policy software agent can be then inferred to supervise the EE of the system.

(ii) The objective of the EE optimization is properly defined so as to jointly ensure that adequate throughput is allocated in the single-user level (considering diverse realistic QoS requirements) and the transmission power is sufficiently reduced in the single-channel and single-station levels.

(iii) This work directly deploys and investigates three DRL approaches to solve the EE optimization problem, namely the centralized, the multi-agent and the transfer learning based solutions. As such, a direct contrast in terms of performance, convergence speed, computational capacity and applicability among the three approaches is provided, as an attempt to identify their limitations in 5G HetNets.

(iv) Extensive simulations are performed to stabilize the learning hyper-parameters and validate the potency of the algorithms in the presence of 5G-compliant channel models and mobility patterns of the mobile users, thus establishing an EE ML-aided optimization framework with increased scalability and generalizability.

(v) The developed methodology may be effortlessly extended to include multi-tier network models (e.g. pico, femto cells and IoT devices) with diverse power constraints, various network topologies and different system parameters (operating frequency, propagation characteristics, 5G numerology, etc.).

The rest of the paper is organized as follows: Section II outlines the related work of the EE network optimization. In Section III, the system model, the interference model and the problem formulation are presented. In Section IV, the three DRL-based algorithms are described, along with the related mathematical principles and background. In Section V, the simulation setup and the hyper-parameter fine-tuning of the training phase are firstly outlined. In addition, several simulations took place in order to investigate various comparative aspects between the proposed DRL and baseline schemes. Finally, Section VI concludes the paper.

II. RELATED WORK

Several studies have highlighted the problem of EE maximization. In [13], the EE optimization is addressed using fractional programming and sequential convex optimization methods for centralized (network-centric) networks. The EE problem is then extended to the multi-agent decentralized (user-centric) case, considering multiple agents being engaged in a non-cooperative game, i.e. each individual agent targets at its own EE maximization. In both cases, minimum throughput requirements of the users and maximum power constraints of the transmitters are considered, which reflect realistic conditions and allow smooth integration of 5G technologies.

Moreover, in [14], the authors demonstrate the solution to the EE optimization problem for downlink two-tier heterogeneous networks (macro- and pico-cells), by jointly considering the beamformer design and power regulation of the transmitters. In this work, diverse sum-rate requirements of the mobile users are considered including video conferencing and online gaming, as well as file transfer and online video. The authors formulate the EE optimization problem as a mixed combinatorial and non-convex optimization problem with multiple inequality constraints by effectively decomposing the original problem into multiple sub-problems with a single inequality constraint. Their approach involves a two-fold resource allocation strategy: an inner-layer is first employed to find the maximum EE considering a specified user-rate, followed by an outer-layer for optimizing the EE via a gradient-based algorithm. However, the problem formulation in [14] includes a single-channel and fixed user association policy, whereas the optimization solutions are obtained for multiple stationary network conditions (user positioning and demand vector). It should be noted also that iterative approaches may be time-consuming during the real-time network operation, especially when time-varying and increased mobility scenarios are considered. This is attributed to the need for multiple iterations before obtaining a solution, which is usually infeasible in the real-time network operation. Thus, a significant drawback of iterative-obtained or heuristic solutions against the DRL methods relies on the fact: once a DQN agent is pre-trained, it can almost effortlessly be inferred (through basic arithmetic/array operations for the neural network inference) to provide real-time suggestions. On the

contrary, the training phase of a DRL scheme is more complex than the conventional approaches.

Furthermore, the authors in [18] investigate the downlink EE optimization problem by jointly performing power allocation and user association, while also considering the minimum sum-rate constraints and the maximum power constraints for the transmitters. An energy-efficient low-complexity algorithm is thus developed and implemented on a downlink massive MIMO system to jointly allocate the optimal transmission power using Newton's methods and configure the user association scheme based on the Lagrange's decomposition methods, ensuring high minimal sum-rate constraints.

More recent research works focus on the implementation of DL and DRL methods in diverse 5G resource allocation problems, showing that RL frequently outperforms analytical, rule-based and heuristic methods [15], [19]. For DL-based approaches, a recent trend is the "learn-to-optimize" approach, according to which a neural network is used to provide approximations of analytical mathematical models after being trained on previously derived analytical solutions [20]. On the other hand, typical applications of RL-based algorithms include power allocation strategies and user-association schemes [21]–[23], dynamic spectrum access [24], beamforming techniques [25] or joint approaches [26]–[28]. These studies are mainly focused on optimizing fundamental physical-layer metrics, such as sum-rate utility, interference management and spectrum utilization.

Towards the specific objective of EE optimization, RL and DRL methods have been also employed. For instance, the authors in [29] propose a multi-agent distributed intuitive online learning energy-efficient power allocation scheme for multi-tier 5G heterogeneous network, also maintaining QoS requirements. Each secondary transmitter (pico, femto and device-to-device transmitters) in the network area of a macro-cell acts as a single agent, speculating the power allocation policies of the other agents by directly interacting with the wireless environment. For purposes of reducing the state space and facilitating the training process and the algorithm convergence, the Q-value of the online learning is approximated as a function of significantly reduced number of parameters.

Finally, the authors in [30] develop a multi-agent DRL framework for jointly optimizing the user association and power control in OFDMA based uplink HetNets. Their decentralized approach manages each user equipment (UE) as a single agent, targeting at maximizing its individual energy efficiency without coordinating with other agents, taking also into consideration the maximum transmit power constraint, as well as the UE's own QoS requirements. Other studies have focused on intelligent control of transmitters' operation (switch on/off) for purposes of eliminating unwanted energy waste [31].

Various limitations may be identified in the existing EE literature. At first, although the DL approaches significantly reduce the time required for solving the EE optimization problem, they inevitably inherit the over-idealized assumptions of analytical solutions. Secondly, although both centralized and distributed learning approaches have been proposed, there is a lack of direct comparison between them for the EE problem in terms of convergence speed, computational capacity, storage requirements, performance and sensitivity related to dimensionality. Furthermore, most of the existing studies have not taken advantage of the transferable learning option across cognitive agents (transfer learning), which can efficiently speed up the learning process, given the spatio-temporal correlation of the involved cells/agents. This technique would allow to considerably reduce the amount of training resources and presumably result into significant system-level EE algorithm enhancements. Finally, simultaneous consideration of mobility-aware and dynamically changing environments to realistically reflect the non-stationary network conditions have been understudied.

III. NETWORK MODEL AND PROBLEM FORMULATION

A. NETWORK MODEL

An urban 5G HetNet area is deployed by considering a large-area macro radio unit (MaRU) overlapped with K small-area micro cells (MiRUs), as depicted in Fig. 1. Each RU has M available physical resource blocks (PRBs) for physical-layer transmissions, depending on the selected operational 5G numerology and the available bandwidth B. The first tier of the cell consists of a single MaRU dedicated to cover a large area. Moreover, a second tier including multiple low-power MiRUs is also deployed for purposes of increasing the network area capacity. Given that the MaRU is deployed to provide fixed coverage in the network area, it simultaneously comprises the major interference source for the second tier mobile users. To that end, when a particular user is located inside a MiRU service area, it experiences the accumulated interference of both MaRU and neighboring MiRUs.

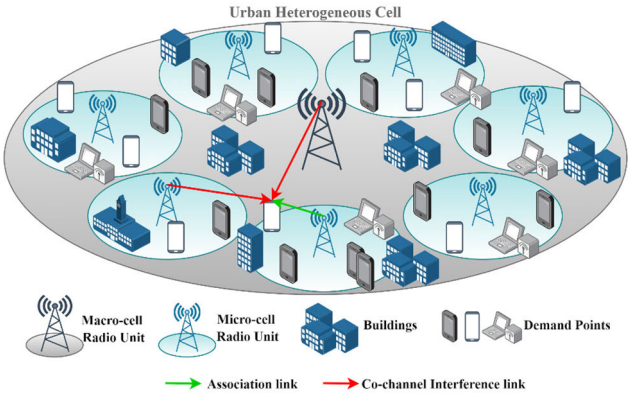


FIGURE 1. Network model consisting of an urban two-tier 5G HetNet cell.

A set of N mobile demand points (DPs) are assumed to be located in the second tier, requesting a specific service that reflects realistic service level agreement (SLA)

requirements in terms of throughput. A binary matrix A is defined to denote the association between a specific DP n and a single PRB m of MiRU k ($A_{k,m,n} = 1$ when the downlink association is established). A demand vector with respective elements D_n denotes the requested QoS of DP n expressed in Mbps. A random walk model is also established for each mobile DP n with velocity v_n (in m/s) and polar direction φ_n (in degrees).

The MaRU transmits over a particular PRB m with a power level of $P_{MaRU,m}$ (in Watts), whereas the transmitted power of MiRU k over PRB m is denoted as $P_{k,m}$. Sum-power constraints are considered separately for MaRU and MiRUs, namely $\sum_{m=1}^{M} P_{MaRU,m} \leq P_{max}^{MaRU}$ and $\sum_{m=1}^{M} P_{k,m} \leq P_{max}^{MiRU}, \forall k$. Moreover, a minimum power level of each PRB is also determined to account for basic signaling and beacon transmissions, i.e. $P_{k,m} \geq P_{min}, \forall k, m$. A complete frequency reuse scheme among RUs is also assumed.

The efficacy of the system-level EE is defined via the extent to which the second tier cognitive network jointly accommodates the QoS requirements and avoids over-consumption of power resources.

Finally, we assume that a single DP occupies a single PRB of a particular MiRU, whereas multiple DPs may be associated with a particular MiRU.

B. INTERFERENCE MODEL

A DP n that establishes a downlink association with a particular PRB m of a MiRU k experiences a signal-to-interference-plus-noise ratio (SINR) γ:

$$\gamma_{k,m,n} = \frac{P_{k,m} \cdot L_{k,m,n}}{P_{MaRU,m} \cdot L_{MaRU,n} + \sum_{k' \neq k} P_{k',m} \cdot L_{k',m,n} + n_0}, \tag{1}$$

where $L_{k,m,n}$ and $L_{MaRU,n}$ stand for the channel losses between the DP n and the MiRUs and MaRU respectively over PRB n and n_0 stands for the noise power density at the location of the receiver. The channel losses reflect the propagation losses of the wireless environment, i.e. shadowing and pathloss (PL) [17]. The PL model due to the operation of the MaRU can be expressed by:

$$PL_{UMa-LOS} = \begin{cases} PL_1, & 10m < d_{2D} < d'_{BP} \\ PL_2, & d'_{BP} < d_{2D} < 5km, \end{cases} \tag{2}$$

where d_{2D} is the 2D distance between the DP and the MaRU, d'_{BP} is the breakpoint distance and the pathloss models PL_1 and PL_2 may be calculated as follows:

$$PL_1 = 28 + 22 \log_{10}(d_{3D}) + 20 \log_{10}(f_c) \tag{3}$$

$$PL_2 = 28 + 40 \log_{10}(d_{3D}) + 20 \log_{10}(f_c) - 9 \log_{10}\left[\left(d'_{BP}\right)^2 + (h_{BS} - h_{UT})^2\right] \tag{4}$$

In the above equations, d_{3D} is the 3D distance between the DP and the MaRU, f_c is the operating frequency of the transmitter, h_{BS} is the height of the transmitter and h_{UT} is the height of the DP. According to [17], the breakpoint distance can be expressed by:

$$d'_{BP} = 4h'_{BS}h'_{UT}f_c/c, \tag{5}$$

where the effective antenna heights are calculated as $h'_{BS} = h_{BS} - h_E$, $h'_{UT} = h_{UT} - h_E$, h_E is the environmental height and c is the speed of light.

Due to the cell heterogeneity, different channel models are taken into account for the MiRUs in the calculation of the SINR γ (Eq. (1)). The pathloss may be expressed by:

$$PL_{UMi-LOS} = \begin{cases} PL_1, & 10m < d_{2D} < d'_{BP} \\ PL_2, & d'_{BP} < d_{2D} < 5km, \end{cases} \tag{6}$$

where the pathloss models PL_1 and PL_2 in this case are computed by:

$$PL_1 = 32.4 + 21 \log_{10}(d_{3D}) + 20 \log_{10}(f_c) \tag{7}$$

$$PL_2 = 32.4 + 40 \log_{10}(d_{3D}) + 20 \log_{10}(f_c) - 9.5 \log_{10}\left[\left(d'_{BP}\right)^2 + (h_{BS} - h_{UT})^2\right]. \tag{8}$$

Following the computation of the SINR for an established downlink association, the transmission data rate R of a DP n with a particular PRB m of a MiRU k can be calculated based on the Shannon formula:

$$R_{k,m,n} = \frac{B}{M} \cdot \log_2(1 + \beta \cdot \gamma_{k,m,n}), \tag{9}$$

where the assumption that each DP may occupy a single PRB has been also taken into consideration. Finally, β depends on the Bit Error Rate (BER) threshold ($\beta = 1$ for $BER = 10^{-6}$).

C. PROBLEM FORMULATION

As already mentioned, the system-level EE jointly examines the second tier transmitting power resources of the K MiRUs that are present in the network area and the requested QoS of the DPs associated with this cell. The EE is formally described by:

$$EE = \frac{\sum_{n=1}^{N} R_n}{\sum_{k=1}^{K} \sum_{m=1}^{M} P_{k,m}}, \tag{10}$$

where R_n is the data transmission rate of DP n once a link has been established between DP n and MiRU k over PRB m. The *EE optimization problem* (P) can be formulated:

$$(P) \quad \max_{P} \left\{ EE = \frac{\sum_{n=1}^{N} min\{D_n, R_n\}}{\sum_{k=1}^{K} \sum_{m=1}^{M} P_{k,m}} \right\} \tag{11}$$

$$s.t.:$$

$$(C1) \quad R_n = max\left\{R_{k,m,n}\right\}, \quad \forall k, \ \forall m \tag{12}$$

$$(C2) \quad \sum_{m=1}^{M} P_{MaRU,m} \leq P_{max}^{MaRU} \tag{13}$$

$$(C3) \quad \sum_{m=1}^{M} P_{k,m} \leq P_{max}^{MiRU}, \quad \forall k \tag{14}$$

$$(C4) \quad P_{k,m} \geq P_{min}, \quad \forall k, m \tag{15}$$

$$(C5) \quad \sum_{k=1}^{K} \sum_{m=1}^{M} A_{k,m,n} \leq 1, \quad \forall n \tag{16}$$

It should be noted that the maximization of the optimization problem involves the limitation of the requested throughput in the calculation of EE, since the objective is to fulfill the DP's requirements and not total network throughput maximization, avoiding the DP over-satisfaction. Moreover, the constraint C1 implies that the DP n will perform the MiRU and PRB association according to the *maximum throughput criterion*. Constraints (C2), (C3) concern the power limitations of the MaRU and the K MiRUs, specifying that the sum-power of each MiRU and MaRU should not exceed a total power consumption threshold (different for MiRU and MaRU) and also that a minimum power level for each PRB is required to enable beacon transmissions (C4). Finally, constraint (C5) reflects that each DP can occupy at most a single PRB of one MiRU.

IV. METHODOLOGY APPROACH

In this section, the DQL principles are described and the proposed DRL framework is outlined considering state, action and rewards spaces. Furthermore, the centralized, the multi-agent and the transfer learning methodologies implemented to solve the EE optimization problem are presented.

A. DEEP Q-LEARNING PRINCIPLES

In general, RL enables an agent interacting with an environment, and getting feedback loops between the learning system and its experiences, in terms of rewards or punishments. The conventional form of RL is the tabular Q-learning method, according to which the RL agents take advantage of the so-called Q-table to become near-optimal predictors of beneficial actions [16]. During the learning process, the agent records its past experiences by continuously filling-and-updating the Q-table. An immediate extension of the tabular Q-learning is to utilize a neural network (DQL) as Q-function approximator, instead of using a memory-inefficient array structure. The DQL has been successfully applied in various non-convex optimization problems due to its capability to cope with enormous state and action spaces [15], [19]. Since a tabular approach requires a significant amount of state-action values, its feasibility is limited to rather small state spaces. To this end, DQL has been widely used in order to overcome the state-action space restrictions and improve the generalizability of RL models.

In principle, the condition of the environment is acknowledged to the DRL agent through the system state $s \in S$ (state space). The agent can then interact with the environment by performing an action $a \in A$ (action space). The learning process loop is completed with the received reward r, involving the feedback (positive, negative or none) of the performed action, as well as the new state of the environment. The concept of the learning method is to train the agent to gradually prefer the actions that return the most profitable rewards. In this manner, the agent is implicitly trained to learn a policy for the objective of the optimization problem, i.e. a sequence of actions for maximizing the long-term rewards. Using the Bellman equation, the agent can estimate the "quality" (the so-called Q-value) of being in state s_t and performing the action a_t [16]:

$$Q_t(s_t, a_t) = (1-\alpha) \cdot Q_{t-1}(s_t, a_t) + \alpha \cdot (r(s_t, a_t) + \gamma \cdot \max_{a'} \{Q(s_{t+1}, a')\}) \tag{17}$$

As readily observed from Eq. (17), the Q-value for a given state/action pair is updated based on (i) its previous value ($Q_{t-1}(s_t, a_t)$), (ii) the immediate reward ($r(s_t, a_t)$) received by performing the action a_t and (iii) the estimated future return $\gamma \cdot \max_{a'} \{Q(s_{t+1}, a')\}$. Moreover, the learning rate $\alpha \in [0, 1]$ is used to weight the previous and learned Q-values, while the discount factor $\gamma \in [0, 1]$ notifies the significance of future long-term rewards.

By repeatedly interacting with the environment, the DQL agent aims to completely gather the information derived by the environment. This may be achieved by minimizing the temporal difference (TD) function, which reflects the difference between the learned value $r(s_t, a_t) + \gamma \cdot \max_{a'} \{Q(s_{t+1}, a')\}$ and the old value $Q_{t-1}(s_t, a_t)$ [16]:

$$TD_t(s_t, a_t) = \left(r_t(s_t, a_t) + \gamma \cdot \max_{a'} \{Q(s_{t+1}, a')\}\right) - Q_{t-1}(s_t, a_t) \approx 0 \tag{18}$$

The key concept of DQL is the utilization of two function approximators (i.e., neural networks) to estimate the current best action (Q-network) and to predict the next best action (target Q–network). The Deep Learning part of the DQL is basically a regression problem in which the objective is to minimize a loss function. This function equals to the difference between the outputs of the Q- (considered as the predicted values) and target Q-networks (considered as the actual values). As a result of the training process, the weights of the Q-network are properly adjusted so as to provide predictions about which action to take from a given state (inference phase).

B. PROPOSED DQL-BASED SOLUTIONS

In general, the DQN schemes are divided into 2 broad categories, namely the classical *centralized/single-agent DQN* vs. the *decentralized/multi-agent DQN*. In this subsection, three approaches to solve the optimization problem (P) are formulated, namely a centralized Deep Q-Network method (*C-DQN*), a multi-agent/decentralized DQL method (*MA-DQN*) and a transfer learning-based DQL (*T-DQN*) technique. The general state, action and reward definitions are firstly described before being specifically adapted to each solution:

1) STATE SPACE

For a sequence of T episodes, the system state can be described through the set $S = \{S_1, \ldots, S_t, \ldots, S_T\}$. At a given time t, the state can be described via a three-fold information for each DP. Specifically, each DP reports to the DRL agent: the ID of the associated RU, the ID of the associated PRB and an indicative parameter about its individual satisfaction status. The satisfaction status SS_n of DP n is discretized in 5 levels:

$$SS_n = \begin{cases} 0, & R_n/D_n < 0 \\ 1, & 1 \le R_n/D_n < 1.2 \\ 2, & 1.2 \le R_n/D_n < 1.5 \\ 3, & 1.5 \le R_n/D_n < 2 \\ 4, & R_n/D_n \ge 2 \end{cases} \tag{19}$$

The SS_n reflects the extent to which the allocated throughput of DP n is above its requested throughput D_n, ranging from under-satisfaction ($SS_n = 0$) to over-satisfaction levels ($SS_n = 4$).

2) ACTION SPACE

The performed actions during T episodes are notated as $\{A_1, \ldots, A_t, \ldots, A_T\}$. At a given step t, the agent can increase, decrease or maintain the power level of a selected PRB on each RU. The action taken at time t is denoted as $A_t = [a_{1,m_1}, \ldots a_{k,m_k}, \ldots, a_{K,m_K}]$, where $a_{k,m_k} \in \{P_{step}, -P_{step}, 0\}$ is the power adjustment on the selected PRB m_k of RU k and P_{step} is the power change value (in Watts). Thus, the power update rule can be described by:

$$P_{k,m}(t) = P_{k,m}(t-1) + a_{k,m_k}(t) \tag{20}$$

3) REWARD

After taking an action, the system transits into a new state thus leading to alternate association scheme and SS levels. The feedback received at time t is expressed by:

$$r_t\,(S_{t-1}, A_{t-1}) = \begin{cases} \dfrac{EE_t - EE_{t-1}}{EE_{t-1}} \times 100, & if\ EE_t > EE_{t-1} \\ 0, & otherwise \end{cases} \tag{21}$$

where the EE_t is the resulting system-level EE. It should be noted that a positive reward quantifies the % EE increment. Following this rewarding system, the agent will learn a policy that gradually improves the network EE.

4) ACTION SELECTION POLICY

In each training episode, the agent selects either a (random) explorative action or an exploitative action (based on the predicted Q-values). For the action selection strategy, we used the ε-*greedy* method, according to which the agent passes smoothly-over-time from an exclusively exploration phase to an exclusively exploitation phase. The ε decaying rule was selected to be linear, starting from 1 and ending to 0 for the first half of the training episodes.

C. CENTRALIZED DQL

The single-agent *C-DQN* algorithm can be described through the agent-environment interaction, as shown in Fig. 2. A centralized controller is acknowledged with the network-wide state and is able to perform an action concerning all MiRUs.

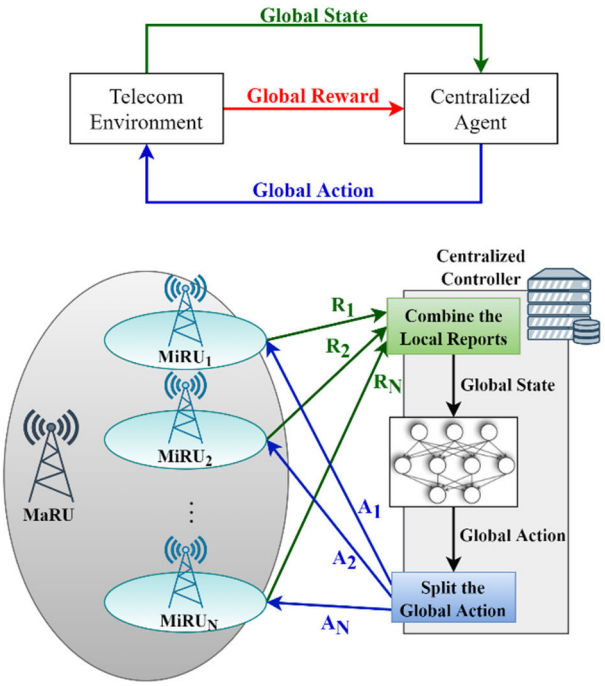


FIGURE 2. C-DQN algorithmic framework. Interaction cycle between the centralized agent and the telecom environment.

To this end, a global network state is constructed by combining N individual reports ($R_1, R_2, \ldots, R_N$) from all MiRUs. Each local report R_i (obtained from MiRU i) is composed of a three-fold information about each DP located in the service area of MiRU i, namely the serving RU ID, the associated PRB ID and the respective SS. Then, the central agent selects either an explorative (randomly) or an exploitative action (based on the Q-network output). The output layer of the C-DQN consists of $N \times M \times 3$ neurons, including the Q-value of increasing, decreasing or maintaining the power level on PRB m ($m = 1, \ldots, M$) of MiRU n ($n = 1, \ldots, K$). By splitting the global action into N separate actions, the MiRU-specific actions are obtained. According to Eq. (21), the global reward, computed as the difference between the current and previous system-level EE, quantifies the impact of the global action on the network state. Fig. 4A shows the architecture of the DQN used for C-DQN solution.

D. MULTI-AGENT DQL

Similar to the C-DQN framework, a multi-agent *MA-DQN* approach was also deployed. As depicted in Fig. 3, one agent per micro-cell is trained based on its partial observability (i.e. the reports of associated DPs only). This approach results into N different DRL models.

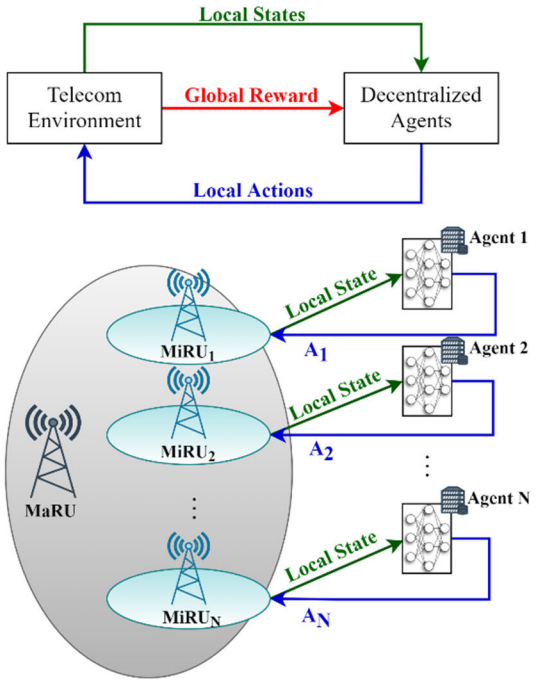


FIGURE 3. MA-DQN algorithmic framework. Each individual agent partially interacts with the telecom environment and collects a global reward.

Specifically, each agent collects the local state vectors of the respective MiRU, concerning the DPs that are presently connected to it. Moreover, each agent can regulate the power levels of its respective MiRU. Although the state and action spaces have significantly lower dimensionality than the C-DQN framework, the selfish behavior of each agent could make the learning convergence infeasible. To obtain a co-operative scheme among the multiple agents, a global reward is returned to each DRL agent. Thus, the global reward allows the agents to sense the others' actions, as the training phase unfolds. Fig. 4B illustrates the architecture of a single-agent DQN.

The outline of the MA-DQN scheme may be summarized in the following steps (for a given episode):

Step 1: Each agent observes only its associated users (their associated PRB and their experienced throughput)

Step 2: Based on its own policy, selects a (local) action.

Step 3: The individual actions selected by each agent are combined to form the global action vector, i.e. the power value of each PRB and cell.

Step 4: Then, the reward is defined similarly to the centralized scheme, defined in Eq. (21). Specifically, the system-level EE increment takes into account the throughput experienced by all users and the power level of all MiRUs (i.e. global reward).

Step 5: In case that the system-level EE was improved, the agents continues to play in the same episode. Otherwise, the reward is zero and another episode is initiated.

Step 6: Steps 1-5 are repeated until convergence.

In this manner, although each individual agent has partial observability of its environment, it is able to "sense"

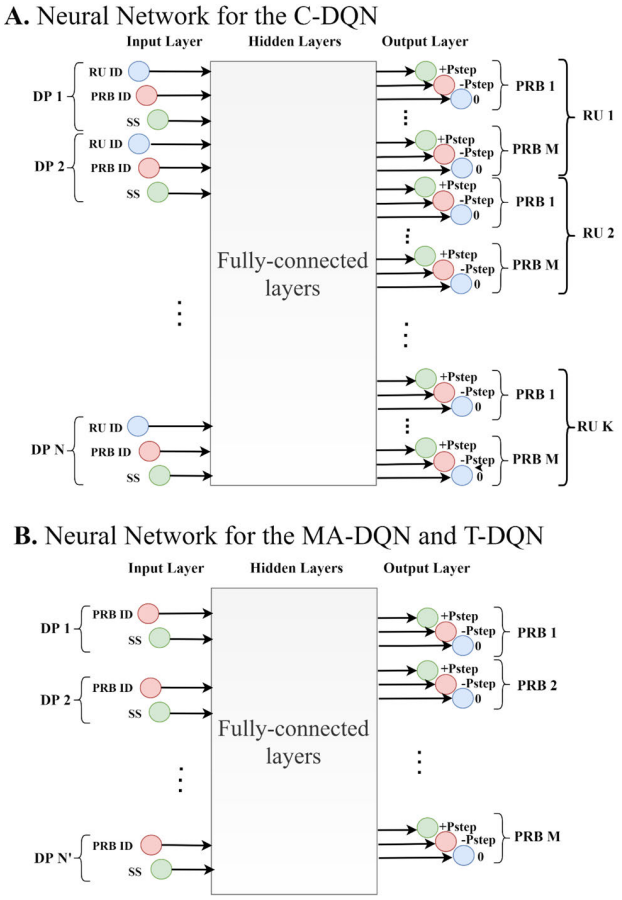


FIGURE 4. Architecture of the neural networks used in the proposed solutions. A. Neural network for the C-DQN approach. B. Neural network for the MA-DQN and T-DQN methods.

the network-wide environment by receiving the system-level EE increment. For example, when a micro-cell agent selects a selfish action based on the observability of its micro-area users, it only contributes in some terms of the global EE formula, as seen from Eq. (10). By taking into account the throughput of each user and the power level of all micro-cells, a global reward is returned to every micro-cell agent, giving insights to each of them about how globally good were their local actions.

E. TRANSFER LEARNING-BASED DQL

In the transfer learning based *T-DQN* approach, we investigated the inheritance capabilities of a single-MiRU model that has pre-trained for different positions inside the MaRU area. This approach exhibits the simplest training approaches in terms of dimensionality, convergence and computational resources, since only one model is trained. In the inference phase, all the existing MiRUs simply inherit the same pre-trained model in order to assist their EE objective. Although the T-DQN has significantly reduced dimensionality in its training phase, the drawback of this approach is that each agent has no information about the interaction with other agents and, thus, multiple selfish models are obtained.

As shown in Fig. 4B, the T-DQN uses the same neural network as a single MA-DQN agent.

V. SIMULATION RESULTS

Several simulations were conducted to illustrate the performance of the developed approaches both for evaluative and comparative purposes. The following simulation setups were implemented in Python 3.8 using Tensorflow 2.4. The training phase of all algorithms ran on a personal PC (CPU i7-8700; 3.2 GHz; RAM 8 GB; no GPU usage). This section is divided into two sub-sections, including (i) the presentation of the individual training procedures for each EE solution (C-DQN, T-DQN and MA-DQN) and (ii) the evaluation of the proposed approaches, along with their comparison with other baseline solutions. The validation procedure includes comparisons in terms of demandingness of requested services, degree of densification and inter-cell distances, as crucial aspects in the EE degradation. All validation results of the DQN-based algorithms were derived by inferring the pre-trained DQN-based models.

In Table 1, we tabulate the simulation parameters both for the configuration of the considered wireless network [17], [32], [33] and the fundamental architectural parameters of the DQN neural networks. Notably, according to the selected 5G numerology 4 (i.e. each available channel is segmented into 6 PRBs), we consider full-capacity scenarios in the training phase (i.e. 6 users, each occupying 1 PRB) [33]. This setup ensures utmost spectrum utilization conditions in the MiRU cells, while also guaranteeing that MaRU unavoidably causes interferences to all PRBs of all MiRUs. All transmitting sites are equipped with omnidirectional antennas, while the mobility of users is characterized by the pedestrian speed of 1 m/s following a random walk model and a random initial positions. Note that, when a DP exits the area of its served micro-cell, an additional DP is randomly initialized to ensure constantly fully-capacity loads in the offloaded micro-cell. This was done to ensure that all PRBs are constantly active and test the EE optimization on the second tier of the cell in worst-case interferences (the MaRU causes interference to all PRBs of all MiRUs). Finally, to include realistic variability in the requested services, three different types of requested SLAs are assumed, namely $SLA_1 = 1$ Mbps, $SLA_2 = 5$ Mbps, $SLA_3 = 10$ Mbps.

It is worth noticing the inter-dependency between the channelization (i.e. channel segmentation) scheme of the system and the achievable throughput. In our simulations, we used the numerology 4 of standardized 5G spectrum guidelines [33], according to which the band (of 20 MHz) is segmented into 2.88 MHz physical resource blocks. It would be possible to apply the presented algorithms in other spectrum configurations (i.e. 5G numerologies). A specific numerology defines how many channels are available within a particular band. In this context, several conceptual alterations have to be noticed when the numerology is modified, especially for those triggered by the relationship between the available channels and the targeted throughput.

TABLE 1. Simulation setup parameters.

System Parameters		DQN Parameters	
Parameter	**Value**	**Parameter**	**Value**
Frequency	6 GHz	Number of hidden layers	3
5G numerology	4	Activation function of input and hidden layers	ReLu
Number of PRBs per channel	6	Activation function of output layer	Linear
PRB bandwidth	2.88 MHz	Memory size	5000
Number of users per MiRU	6	Mini-batch size	64
MaRU/MiRU cell radius	500/100 m	Update target frequency	100
Noise power density	-174 dBm/Hz	Optimizer	Adam
Max power of MaRU/MiRU	80/25 W	ε decay	Linear
Min power per PRB	0.1 W	Loss function	Huber loss

Shannon's formula [10] justifies that the achievable throughput is positively correlated with the available channel bandwidth and the interference mitigation (i.e. the SINR). Therefore, a lower numerology that segments the same band into a higher number of subchannels has the following contradictive consequences:

(i) A higher number of available channels results into higher flexibility among users to occupy different channels or exchange channels between each other. This directly increases the system-level SINR, because the number of common-channel links is reduced.

(ii) On the other hand, a high-degree channel segmentation results into low-bandwidth channels available to each user, thus defining a lower upper-bound in the achievable throughput per user.

To sum up, a throughout maximization scheme has to jointly consider the power regulation strategy (to reduce the interferences), the channel allocation scheme (to intelligently assign the users to channels) and the user association policy (to handle for handovers and cell associations).

A. STABILIZATION OF DQN HYPER-PARAMETERS

The impact of the algorithm hyper-parameters has to be cautiously stabilized, as their values can significantly influence the performance convergence. In general, DQN-based algorithms are characterized by stochastic behavior, since multiple runs of the same model could exhibit variations in the resulted performance. The crucial learning parameters in the proposed algorithms are initially fine-tuned according to whether they present near-optimal convergence values in terms of the achieved EE increment. Specifically, we sequentially examined the number of episodes (directly affecting the exploration duration), the learning rate (a, balances the

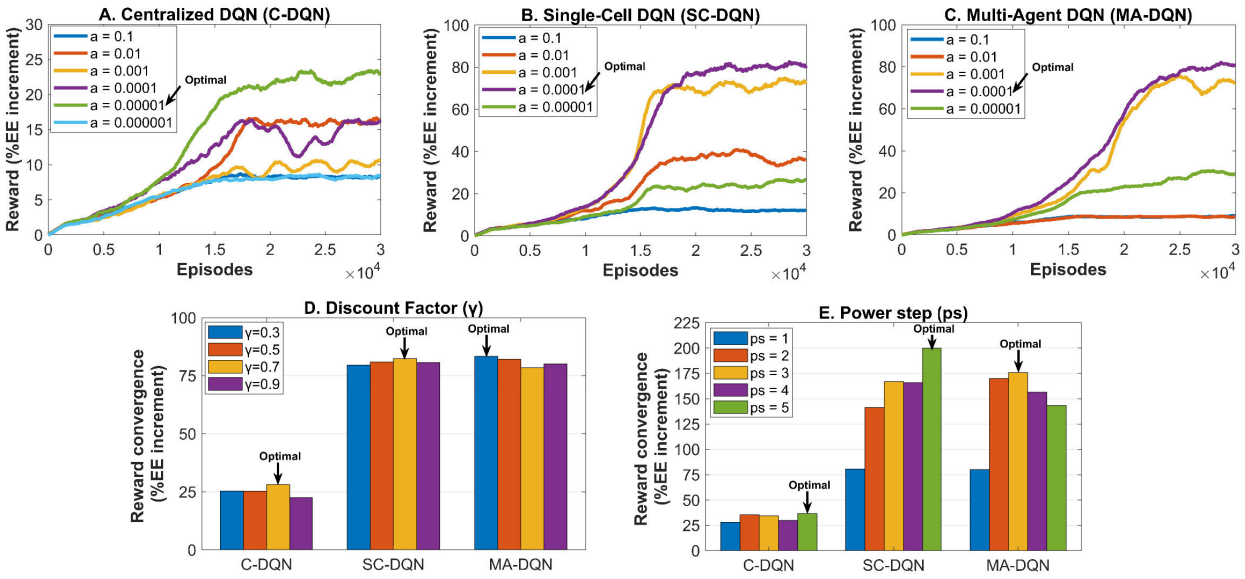


FIGURE 5. Impact of the training hyper-parameters on the proposed approaches. A. Learning curve of the C-DQN solution for varying values of the learning rate. B. Learning curve of the single-cell trained model for varying values of the learning rate. C. Learning curve of the MA-DQN solution for varying values of the learning rate. The performance of the MA-DQN scheme was assessed by individually inferring each agent's model and computing the system-level EE. D. Convergence value of EE increment (average reward across the 100 last episodes) per method for different values of discount factor. E. Convergence value of EE increment (average reward across the 100 last episodes) per method for varying values of power step.

contribution of the new and old Q-value in the Bellman formula), the discount factor (γ, defines the significance of future rewards), as well as the power granularity (ps). This was done for all deployed models, namely the centralized (C-DQN), the multi-agent (MA-DQN) and the single-cell (SC-DQN) model that will be used for the transfer learning-based (T-DQN) scheme evaluation.

The hyper-parameter stabilization was performed following a "brute-force" approach, meaning that the algorithm convergence was tested on multiple values of all the learning parameters. According to this approach, one model per parameter configuration was obtained and stored. Then, the individual models were compared and the parameter-specific model with the highest reward convergence was selected.

For this section, we consider a topology consisting of a MaRU and four fully-occupied MiRUs. The center of the micro-cells was randomly selected at the start of each episode, whereas the inter-cell distances were kept at a minimum of 100m, as recommended in [17] for the micro-cell UMi areas. Fig. 5 depicts the training evaluation of the proposed DQN-based algorithms. In general, a number of 30000 training episodes was proved sufficient for the reward convergence of both SC-DQN and MA-DQN schemes. Evidently, we observed a considerable impact of the learning rate on the performance of all schemes, as it regulates how rapidly or gradually the learning procedure unfolds (Fig. 5A-C). As expected, the C-DQN requires lower learning rate ($\alpha = 10^{-5}$) compared with SC- and MA-DQN approaches ($\alpha = 10^{-4}$), since it takes into consideration larger state-action spaces and monitors the global (high-dimensional) network state. Noteworthy, the SC-DQN model achieved a reward of ∼81%, meaning that it can guarantee ∼81% EE increment relative to the initial system EE. This means that, if we infer the SC-DQN model for purposes of maximizing the system EE, a gain of 81% will be achieved relative to the initial system EE (without SC-DQN assistance). Note also that, the final reward of a given episode is accumulated by summing the individual rewards of the actions (i.e. power regulation steps) taken during this particular episode. Furthermore, although the SC-DQN showed excessive training performance, it refers to a single-cell (selfish) model that will be inherited to multiple cells in the next section and, thus, it does not characterize the system-level EE. Similar performance in the order of 80% EE increment was also observed for the MA-DQN model. This is attributed to the ability of multi-agent schemes to combine the low-dimensional DQN complexity (multiple DQNs identical to the SC-DQN) with the global reward to achieve inter-cell coordination. We also noticed elongated delay (∼500 episodes) required for the convergence stabilization of MA-DQN relative to the SC-DQN model. This delay is attributed to the inherent property of multi-agent RL schemes that enforce multiple selfish agents to be coordinated with each other (by sharing the same global reward) through partial observability of the environment. Finally, we observed that C-DQN approach resulted into ∼3.2 times worse EE performance (∼23% EE increment) compared with MA-DQN.

The significant differences in dimensionality between the C-DQN and MA-DQN models could explain the superiority of the MA-DQN scheme in achieving enhanced system EE. Indicatively, the possible states that can be seen from the C-DQN equal to $(5 \times K \times M)^N$ and the possible actions equal to $(3 \times M)^K$, whereas those for a single-agent is $(5 \times M)^{N/K}$ and $3 \times M$, respectively (5 satisfaction status levels, 3 available actions, K MiRUs, M PRBs, N DPs). Apart from the large dimensionality and the enormous state-action spaces that characterize the centralized approach, the C-DQN model faces additional applicability/technical challenges because it requires high-bandwidth backhaul channels to carry the required information (a centralized controller collects the information of all users in the network area, while also it manipulates all the existing antennas in the network). In many real cases, this is practically infeasible, mainly due to the backhaul overhead and the massive required storage at the controller side. Based on this limitation, the C-DQN is primarily used for comparison purposes, as well as to quantitatively show its inability to converge in an adequate (i.e. comparable to the other schemes) EE solution under the same number of training episodes.

As shown in Fig. 5D, discount factor showed negligible impact on the performance of the algorithms, with the optimal values being highlighted using arrows. Finally, Fig. 5E illustrates the EE convergence values for discrete power granularity levels, ranging from 1 to 5 Watts in steps of 1 Watt. Interestingly, the selfish SC-DQN model shows the optimal reward when considering the maximum power step (5 Watts), as there were no other micro-cells inside the macro-area, allowing for rapid power steps towards EE maximization. On the contrary, MA-DQN scheme exhibited the optimal performance for the median power step (3 Watts) among all agents, indicating the coordinated behavior among micro-cells. This behavior implies that a high-valued power step can potentially enhance the EE of a single-cell, but, contradictorily, can degrade the total system EE due to enhanced inter-cell interferences. Thus, the optimal power configuration towards ensuring high system EE requires the agents to be more conservative in order to eliminate the inter-cell interferences. The impact of power step in the C-DQN model was negligible, showing the best value for power step of 5 Watts.

Regarding the run time required for the training of the algorithms, SC-DQN needed ∼1 hour, C-DQN ∼3 hours and MA-DQN ∼15 hours to accomplish 30000 training episodes, since the latter trains multiple model simultaneously. For the rest of the simulations, the hyper-parameters of the neural networks were set to the optimal values presented in Fig. 5.

B. PERFORMANCE COMPARISON BETWEEN DIFFERENT METHODS

In this section, we evaluate the performance of the developed methodologies and we compare the DQN-based solutions against three other baseline schemes. For this section, we consider a similar network configuration (consisting of a MaRU and four fully-occupied MiRUs) and a fixed topology of the second-tier network (inter-cell distances 500 m placed in squared topology). This was done to ensure fairness in the performance comparisons among the considered schemes. The three DQN-based models (C-DQN, MA-DQN and T-DQN) were contrasted in terms of their final achieved network-wide EE, following the suggested actions of the pre-trained models.

The performance validation of each model was assessed by inferring the respective model for 1000 different validation scenarios (i.e. random user positioning and service requests per scenario). The final performance metric was quantified as the average EE increment across the validation scenarios. Three baseline schemes were additionally used for comparison purposes, namely (i) the fixed average power allocation scheme (AVG), which offers a reasonable trade-off between the consumed power and the achievable throughput, (ii) a random power control policy, which randomly assigns power levels to each PRBs, respecting only the power limitations and (iii) a heuristic Particle Swarm Optimization (PSO) solution, targeting at the system-level EE maximization with identical power constraints.

To account for diverse demanding conditions, validation scenarios with incremental difficulty are considered:

- *Scenario 1:* all requests are SLA_1 services,
- *Scenario 2:* all requests are SLA_2 services,
- *Scenario 3:* random requests of SLA_1, SLA_2, SLA_3 services,
- *Scenario 4:* random requests of SLA_2, SLA_3 services,
- *Scenario 5:* all requests are SLA_3 services.

For each of the six methods, we computed the resulting system-level EE (Mbps/Watt), as illustrated in Fig. 6.

In a general point of view, the performance deviation between the methods becomes more obvious as the difficulty of scenario increases. As readily observed from Fig. 6, AVG and Random methods showed poor EE performance, primarily due to the absence of intelligent power configuration to avoid inter-cell interferences. Given that the rest of the methods involve non-fixed power control strategies, it is shown that a beneficial reduction in the power consumption may result into a gain in the EE performance. C-DQN presented inability to optimally find the best EE solution, since an enormous state space must be visited during the training phase (i.e. exploration phase) in order to gather sufficient knowledge about the possible environment states.

The heuristic PSO-based solution showed median performance among all validation scenarios. It constantly outperforms the C-DQN approach, since it searches a suboptimal solution for each different user positioning instance. This fact raises significant applicability issues of PSO in real 5G deployments, as it would be infeasible to infer PSO-based solutions (in the order of minutes to obtain a solution) in time-varying and increased mobility scenarios. On the contrary, all pre-trained DQN-based algorithms can almost

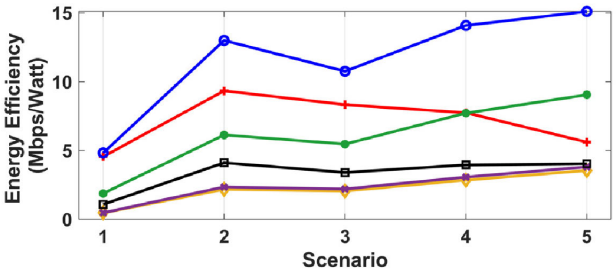


FIGURE 6. Performance comparisons in 5 different demanding scenarios between the six contrasted algorithms in terms of energy-efficiency.

effortlessly be inferred (in the order of milliseconds to obtain the suggested action), given that the neural network weights are already adjusted.

Importantly, MA-DQN method showed systematically the best EE outcomes, since it combines the benefits of decentralized low-dimensional learning with a centralized global reward. The latter allows the discrete agents to sense the whole environment, although the individual state space (that is visible from each agent) offers partial network observability. In other words, following the MA-DQN approach, it is possible for an agent to avoid selfish moves, since actions that can degrade the network-wide EE would return zero rewards. The opposite situation applies to the T-DQN solution, where multiple optimal-but-selfish models are used for the EE monitoring. Specifically, we observed that in the first three scenarios, both T-DQN and MA-DQN exhibit comparable outcomes in terms of EE performance, while in scenarios 4 and 5, the superiority of MA-DQN over T-DQN becomes clearer. When challenging requirements of the users (as in scenarios 4 and 5) are considered, it is expected that the link-specific power levels should be increased in order to ensure acceptable reception. Simultaneously, if all micro-cells use increased power levels to ensure local satisfaction, inter-cell interferences will be the side effect in limiting the targeted throughput. Thus, the underperformance of T-DQN may be attributed to the selfishness of the single-cell trained models, which, although are well-perform in standalone scenarios, do not take into account the presence of the other micro-cells. On the contrary, MA-DQN approach inherently considers the presence of all micro-cell inside the macro-area, while the coordinated behavior among agents is gradually achieved by adopting the global reward function. Moreover, scenario 5 highlights the main drawback of T-DQN, noticing better performance for PSO when considering high-demanding conditions. Again, it should be noted that PSO suffers from significant challenges in the response time to obtain a solution. As intuitively expected, T-DQN may become inefficient in cases that strict demands are considered, since the increased demands require coordinated behavior among the agents. To this end, MA-DQN seems the most applicable option when extreme demanding conditions are realized.

C. IMPACT OF NETWORK DENSIFICATION

Driven by the outcomes of the previous section, we further evaluate the performance of the best EE solvers (i.e. MA-DQN and T-DQN). We have already noted that T-DQN is the optimal method in terms of computational resources needed to train a single-cell model (and then share it across all micro-cells), whereas the MA-DQN is the best EE solver in terms of the achieved EE. To further address whether it is beneficial to use T-DQN over MA-DQN (and vice versa), we examined the impact of network densification on the EE performance. Note that, an increase in the number of MiRU inside the MaRU area results into reduced inter-cell distances, thus increasing the overall system interferences. To that end, five different topologies were realized with increasing intrinsic micro-cells, ranging from 2 to 10 micro-cells. This scenarios allowed us to not only contrast the algorithms in increasing interferences conditions, but also to assess the scalability of the algorithms, since the number of DPs is increased by 6 for each additional micro-cell. Note that, we used almost symmetrical topologies in order to ensure that the MaRU causes near-identical interferences to all MiRUs.

As depicted in Fig. 7, the EE degrades with the degree of densification, independently of the method used for EE maximization. We also observed that the MA-DQN superiority in achieving better EE is more noticeable as the degree of densification increases, whereas for sparse topologies (2 and 4 microcells), where the individual microcells are relatively isolated between each other, the T-DQN is probably beneficial, providing quasi-identical EE with significantly cost-efficient training resources. As generic conclusion, T-DQN scheme guarantees good EE solutions (above 2 Mbps/Watt), even for densely-deployed macro-cells, with effortless training and memory requirements. On the contrary, MA-DQN offers the best EE solutions at the expense of massive training resources and required training time.

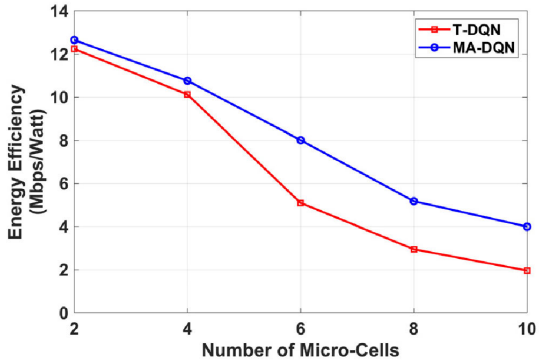


FIGURE 7. Energy efficiency performance derived from T-DQN (red) and MA-DQN (blue) schemes as a function of network densification.

D. IMPACT OF INTER-CELL DISTANCE

In this section, we consider two MiRUs located in equal distances from the MaRU. The training of both T-DQN and MA-DQN models was performed with dynamic topologies

that were based on random center positions of the micro-cells inside the macro-cell area. Then, 6 validation scenarios are set, where the distance between the MiRUs varies between 100 and 600 meters to gradually eliminate the inter-cell interferences. For comparison purposes, both models are inferred for the 6 different topologies, one for each inter-cell distance scenario, considering random SLA requirements (SLA_1-SLA_3).

As depicted in Fig. 8, MA-DQN can clearly present increased EE (by a factor of ∼1.3) compared with T-DQN in cases that the inter-cell distances ranges between 200-400 m. In such cases, we noticed a power saving of ∼3 Watt in MA-DQN relative to T-DQN, at the expense of a throughput degradation of ∼2 Mbps. It is worth also noting that both methods allocate reduced power levels as the inter-cell distances increase. In this way, this behavior results into increased experienced throughput, thus offering enhanced system EE solutions. In summary, EE is positively correlated with the inter-cell distance, because micro-cells can be considered isolated between each other, thus they are able to reduce their power costs. Importantly, the other case in which MA-DQN and T-DQN performances are indistinguishable occurs for the distance of 100 m (both EE ∼2.5 Mbps/Watt). This is directly attributed to the generic inability of both methods to optimally allocate the power levels in order to achieve EE. In such case, both algorithms show a peak power consumption (∼16 Watt) as an attempt to overcome the powerful interferences.

To sum up, T-DQN presents beneficial usage for very low (<100 m) or very high (>600 m) inter-cell distances, whereas the usage of MA-DQN is preferred for intermediate inter-cell distances, where the power savings are feasible towards achieving increased EE. In other words, key conclusions include: (i) when the micro-cells are isolated enough, the selfish models can present adequate EE due to eliminated interferences, and (ii) when the micro-cells are very close, there are no alternations in EE performance either by using MA-DQN or T-DQN.

E. IMPACT OF MOBILITY SPEED

For the previous sections, we have used the constant user speed of 1m/s (all pedestrians). However, a sufficiently trained DQN agent should exhibit stable performance independently of the user mobility speed and the spatio-temporal traffic fluctuations. To verify this aspect, the MA-DQN model, as the dominant EE maximization scheme, was further validated for varying user speed scenarios. Five scenarios were conducted by allowing 3 different types of users, namely pedestrians (1 m/s), low-speed vehicles (30 km/h) and high-speed vehicles (80 km/h). The scenarios used the same topology as in Section V.B and were parameterized as:

- *Scenario 1:* all users are pedestrians,
- *Scenario 2:* all users are low-speed vehicles,
- *Scenario 3:* all are randomly assigned as pedestrians, low-speed or high-speed vehicles,
- *Scenario 4:* half of users are low-speed vehicles and half are high-speed vehicles,
- *Scenario 5:* all users are high-speed vehicles.

Note that, scenarios 1-5 are sorted in increasing mobility order. For each scenario, the achieved system-level EE and the total allocated throughput were extracted as the average EE across 1000 validation episodes. As readily observed from Fig. 9, the DQN-assisted EE solution is independent

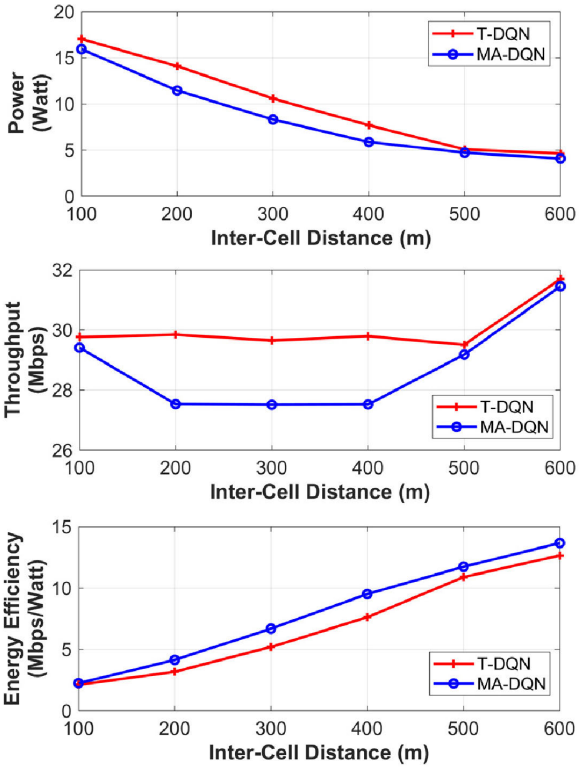


FIGURE 8. Power consumption (A), Achieved throughput (B) and corresponding energy efficiency (C) performance derived from T-DQN (red) and MA-DQN (blue) schemes as a function of the inter-cell distance (in m).

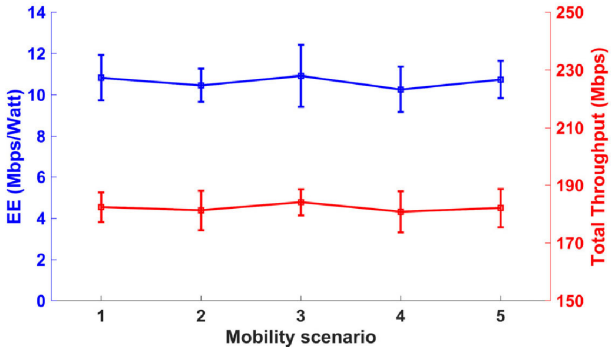


FIGURE 9. MA-DQN performance for 5 different mobility speed scenarios. In the left y-axis, the mean EE performance (blue) across 1000 validation scenarios is depicted. In the right y-axis, the mean throughput performance (red) across 1000 validation scenarios is depicted. Error bars represent the standard error of the mean.

of the user velocity, showing stability both in the EE (with small fluctuations around 10 Mbps/Watt) and throughput (∼180Mbps) performance even for diverse mobility scenarios ranging from slow- to fast-changing spatio-temporal traffic patterns. Based on these simulations, it is concluded that a mobility-aware DQN scheme can generalize for multiple user mobility scenarios.

F. IMPACT OF THROUGHPUT DEGRADATION TOLERANCE

As implied by the definition of EE, there is a contradictive relationship between the total allocated throughput and the power consumptions. This section introduces a parameter that can tune the trade-off between a user-centric and an operator-centric (or environmental-centric) EE power allocation policy. This parameter is the Throughput Degradation Tolerance (η) which defines the extent (0-100%) to which a throughput degradation is acceptable towards EE improvements. So far, the η has been deterministically set to 30% as a reasonable value for degrading the throughput to obtain a power-efficient solution. This means that the presented results can degrade the total requested throughput at most 30% for purposes of improving the system EE. Formally, this can be written as:

$$\sum_{u\in U} \min(R_u, D_u) \geq (1-\eta)\cdot \sum_{u\in U} D_u \qquad (22)$$

where U is the set of users, R_u is the allocated throughput (in Mbps) at user $u \in U$ and D_u is the demand (in Mbps) of user $u \in U$. The *min* operation ensures that the calculation of the allocated throughput is upper-bounded by the demanded throughput to avoid over-estimation in the allocated throughput values. Eq. (22) implies that every EE power allocation scheme resulted from the algorithm has to guarantee at least $1-\eta$ user satisfaction, thus low values of η can be used for user-centric policies and high values of η can be used for operator- or environmental-centric policies.

In this section, we investigate the impact of η in the proposed MA-DQN scheme, as the latter was proved the optimal EE solution. To this end, we reconsider the topology of Section V.B and we extract the throughput satisfaction (i.e. percentage ratio between the allocated and the requested throughput) and the total power consumption for varying values of η. Simulations involved random SLAs (1, 5 or 10 Mbps) among users and random user types (pedestrian, low- or high-speed vehicles). This is illustrated in Fig. 10, where the total allocated throughput and the total consumed power are illustrated as a function of η.

As readily observed from Fig. 10, low values of η correspond to high power consumptions and high system throughput satisfaction, whereas high values of η correspond to power-efficient system operation at the expense of throughput degradation. A reasonable trade-off between the allocated throughput and the consumed power is observed for $\eta = 30\%$, which means that a user satisfaction rate of ∼97% can beneficially influence (10-Watt reduction relative to the solution of $\eta = 0$) the power consumptions

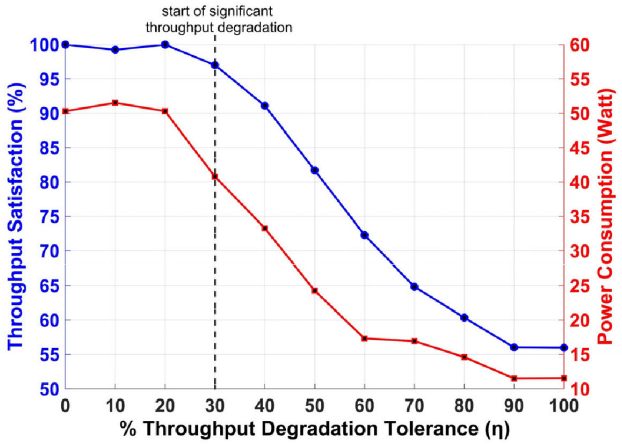


FIGURE 10. Throughput satisfaction rate (blue) and consumed power (red) derived by the MA-DQN scheme for varying values of throughput degradation tolerance η. The individual points of throughput satisfaction and power consumption curves correspond to the mean value across 1000 validation scenarios.

(and thus the system EE). Note also that, both throughput and power curves are quasi-stable for $\eta < 30\%$, whereas the significant throughput degradation and power reduction are observed for $\eta > 30\%$. In addition, the throughput satisfaction rate is lower-bounded by 55%, even when η allows for >70% throughput degradation because the power levels of all cells have reached their minimum values.

It should be noted that parameter η can have significant impact on the EE optimization and its selection highly depends on how user- or operator-centric the system actually is. To that end, a user-specific priority label ('High' for 100% satisfaction, 'Medium' for >70% satisfaction or 'Low' for >40% satisfaction) could be established in the EE optimization framework to indicate whether a specific user (or link) can actually experience throughput degradation towards EE enhancements. This priority assignment highly depends on the operator's policy, energy consumption thresholds, environmental aspects and user requirements. However, the presented framework considers equal fairness across users and does not distinguish between user priorities, thus the user priority labels are ignored.

G. DEPLOYMENT OPTIONS WITHOUT COLLECTED DATA

In this section, we present three conceptual deployment options for implementing the proposed framework without (or with limited) need for historically collected data. Given the serious confidential and privacy issues of mobile operators to store and share the network traffic data, it is worth noting the deployment options of the resource management algorithms to prove their applicability in real conditions. Note that, once a DRL agent is pre-trained (i.e. the neural network weights are adjusted to provide a good action given a state), it can be directly deployed in a real network physical or virtual infrastructure for purposes of providing suggestions or predictions.

In the proposed framework, the data required by the agent (both for training and inference phase) at each time slot are determined by the state space, which is basically what the agent needs to observe before taking an action. The defined state space includes a two-fold information, namely the user-specific associated PRB and the SS. In a realistic operation of the proposed algorithm, the online information reported by the network to the agent is:

(i) the associated link of each user (i.e. PRB ID)
(ii) the demand of each user (i.e. the SLA of the request)
(iii) the experienced throughput of the user.

1) OPTION 1 – TRAINING WITH SIMULATIVE MEASUREMENTS

In this option, the training of the algorithm is conducted offline by using simulations, as exactly done in this paper. Information (i) and (ii) can be easily adopted to reflect realistic values, although the agent is trained on unreal conditions. The first serious concern in using the simulated channel models (urban models UMa and UMi) is their potential variation from real channel conditions. This means that, since our agents are trained in the simulated environment, the experienced throughput values (information (iii)) that is visible by the agent during the training may be different from the throughput of a real environment, mainly due to the inaccuracies of the channel models. Based on this limitation, we used the SS as part of the state space in order to transform the continuous (and possibly untrue) values of throughput into a discrete variable (of 5 levels). This trick allows us to train our agent in a simulated environment aiming to visit as much as possible pairs of state variables (PRB ID and SS) per user. In this sense, the training of the agent does not need collected training data, since it is trained in a simulative DRL environment, while also it is insensitive to channel model imperfections, since the values of throughput are translated into discrete satisfaction levels.

2) OPTION 2 – TRAINING WITH GAN-GENERATED MEASUREMENTS

Another option relies on an ongoing field of ML and wireless networks, which basically produces ''fake'' network measurements that are difficult to be distinguished from real network measurements. This data generation is based on the usage of Generative Adversarial Networks (GANs) and has been recently proposed as *experienced DRL* [34]. Given a limited training dataset, the goal of a GAN is to generate new data with the same statistics as the training set [35]. Note that, the key idea of this approach is to enrich an existing-but-limited dataset with synthetic data, thus its application requires a preliminary time-limited data collection from the real network. Then, the dataset can be extended with synthetic data showing superficially authentic properties. Such an approach has been shown that, by gaining experience, DRL can become more reliable to extreme events and faster to recover and converge [34].

3) OPTION 3 – BACKGROUND TRAINING DURING NETWORK OPERATION

The final option is the most straight-forward approach for training a DRL agent, which is literally to adapt it on the real network. By allowing the agent to interact with the real environment (i.e. it observes the actual throughput, the actual RSSI, etc.), the errors produced by the channel estimations are eliminated, offering the opportunity to the agent to be trained accurately on the exact same environment that will be used for inference. The main drawback of this approach is that it is time-consuming and inflexible, mainly due to the long period required for the training (hyperparameter fine-tuning) and the exploration phase.

A promising way to combine the benefits of options 1 and 3 is to train an agent in a simulation environment (option 1) and, before deploying it to the real network, to test and validate its performance for possible parameter adaptation based on a short-range real inferences.

VI. CONCLUSION

In the present work, a DQN-based framework is proposed, appropriate for improving the system-level EE of two-tier 5G heterogeneous cells with multi-channel transmissions. To efficiently solve the EE problem and contrast different approaches, three DQN-based methods (centralized, multi-agent and transfer learning) are established and properly modified to ensure QoS adequacy and power-efficient adjustment in each transmission link. Following extensive simulations on 5G-compliant channel models, we showed that DQN-based assistance on EE optimization can achieve at least 2 Mbps/Watt, even in strict demanding scenarios and densely-deployed areas. We observed that multi-agent scheme (achieves up to 14 Mbps/Watt) constantly outperforms the other schemes, mainly due to the coordinated behavior among agents with simultaneously low DQN dimensionality. Centralized solutions (performance of $<$5 Mbps/Watt) have to be cautiously selected for solving EE problems, since they provoke not only extremely high overhead to collect the network-wide information, but also require massive computation resources to be trained on all possible network states. Transfer learning principles have to be seriously considered to reduce computationally-intensive re-training and simplify the optimization strategies. This is especially applied when the transferred models are inherited across micro-cells with quasi-identical interference profile and the inter-cell distances are very low ($\sim$2.5 Mbps/Watt) or very high ($\sim$13 Mbps/Watt).

REFERENCES

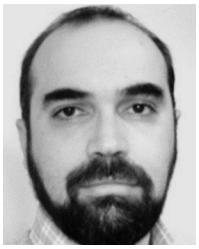
[1] H. Kim, *Design and Optimization for 5G Wireless Communications*. Hoboken, NJ, USA: Wiley, 2020.
[2] P. Trakadas, P. Karkazis, H. C. Leligou, T. Zahariadis, F. Vicens, A. Zurita, P. Alemany, T. Soenen, C. Parada, J. Bonnet, E. Fotopoulou, A. Zafeiropoulos, E. Kapassa, M. Touloupou, and D. Kyriazis, ''Comparison of management and orchestration solutions for the 5G era,'' *J. Sens. Actuator Netw.*, vol. 9, no. 1, p. 4, Jan. 2020.

[3] A. Ahmad, M. H. Rehmani, H. Tembine, O. A. Mohammed, and A. Jamalipour, "IEEE access special section editorial: Optimization for emerging wireless networks: IoT, 5G, and smart grid communication networks," *IEEE Access*, vol. 5, pp. 2096–2100, 2017.
[4] *Cisco Visual Networking Index: Forecast and Methodology, 2016–2021; White Paper; Cisco Systems*, Cisco Vis. Netw. Index, San Jose, CA, USA, 2017.
[5] A. Gupta and E. R. K. Jha, "A survey of 5G network: Architecture and emerging technologies," *IEEE Access*, vol. 3, pp. 1206–1232, Jul. 2015.
[6] S. Buzzi, C.-L. I, T. E. Klein, H. V. Poor, C. Yang, and A. Zappone, "A survey of energy-efficient techniques for 5G networks and challenges ahead," *IEEE J. Sel. Areas Commun.*, vol. 34, no. 4, pp. 697–709, Apr. 2016.
[7] H. Tullberg, P. Popovski, Z. Li, M. A. Uusitalo, A. Hoglund, O. Bulakci, M. Fallgren, and J. F. Monserrat, "The METIS 5G system concept: Meeting the 5G requirements," *IEEE Commun. Mag.*, vol. 54, no. 12, pp. 132–139, Dec. 2016.
[8] N. Nomikos, E. T. Michailidis, P. Trakadas, D. Vouyioukas, T. Zahariadis, and I. Krikidis, "Flex-NOMA: Exploiting buffer-aided relay selection for massive connectivity in the 5G uplink," *IEEE Access*, vol. 7, pp. 88743–88755, 2019.
[9] L. Xiao, Y. Li, C. Dai, H. Dai, and H. V. Poor, "Reinforcement learning-based NOMA power allocation in the presence of smart jamming," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3377–3389, Apr. 2018.
[10] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, no. 3, pp. 379–423, Jul./Oct. 1948.
[11] D. Tse and P. Viswanath, *Fundamentals of Wireless Communication*. Cambridge, U.K.: Cambridge Univ. Press, 2005.
[12] Z. E. Ankarali, B. Peköz, and H. Arslan, "Flexible radio access beyond 5G: A future projection on waveform, numerology, and frame design principles," *IEEE Access*, vol. 5, no. 1, pp. 18295–18309, 2017.
[13] A. Zappone, L. Sanguinetti, G. Bacci, E. Jorswieck, and M. Debbah, "Energy-efficient power control: A look at 5G wireless technologies," *IEEE Trans. Signal Process.*, vol. 64, no. 7, pp. 1668–1683, Apr. 2016.
[14] J. Tang, D. K. C. So, E. Alsusa, K. A. Hamdi, and A. Shojaeifard, "Resource allocation for energy efficiency optimization in heterogeneous networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 10, pp. 2104–2117, Oct. 2015.
[15] J. Kaur, M. A. Khan, M. Iftikhar, M. Imran, and Q. E. Ul Haq, "Machine learning techniques for 5G and beyond," *IEEE Access*, vol. 9, pp. 23472–23488, 2021.
[16] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.
[17] *Study on Channel Model for Frequencies From 0.5 to 100 GHz*, document TR 38.901, 3GPP, 2017.
[18] A. Salh, N. S. M. Shah, L. Audah, Q. Abdullah, W. A. Jabbar, and M. Mohamad, "Energy-efficient power allocation and joint user association in multiuser-downlink massive MIMO system," *IEEE Access*, vol. 8, pp. 1314–1326, 2020.
[19] M. E. Morocho-Cayamcela, H. Lee, and W. Lim, "Machine learning for 5G/B5G mobile and wireless communications: Potential, limitations, and future directions," *IEEE Access*, vol. 7, pp. 137184–137206, 2019.
[20] A. Zappone, M. Di Renzo, M. Debbah, T. T. Lam, and X. Qian, "Model-aided wireless artificial intelligence: Embedding expert knowledge in deep neural networks for wireless system optimization," *IEEE Veh. Technol. Mag.*, vol. 14, no. 3, pp. 60–69, Jul. 2019.
[21] Y. Zhang, C. Kang, T. Ma, Y. Teng, and D. Guo, "Power allocation in multi-cell networks using deep reinforcement learning," in *Proc. IEEE 88th Veh. Technol. Conf. (VTC-Fall)*, Chicago, IL, USA, Aug. 2018, pp. 1–6.
[22] A. Giannopoulos, S. Spantideas, N. Capsalis, P. Gkonis, P. Karkazis, L. Sarakis, P. Trakadas, and C. Capsalis, "WIP: Demand-driven power allocation in wireless networks with deep Q-learning," in *Proc. IEEE 22nd Int. Symp. World Wireless, Mobile Multimedia Netw. (WoWMoM)*, Pisa, Italy, Jun. 2021, pp. 248–251.
[23] A. Giannopoulos, S. Spantideas, C. Tsinos, and P. Trakadas, "Power control in 5G heterogeneous cells considering user demands using deep reinforcement learning," in *Proc. IFIP Int. Conf. Artif. Intell. Appl. Innov.*, Jun. 2021, pp. 95–105.
[24] O. Naparstek and K. Cohen, "Deep multi-user reinforcement learning for dynamic spectrum access in multichannel wireless networks," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Singapore, Dec. 2017, pp. 1–7.
[25] F. B. Mismar, B. L. Evans, and A. Alkhateeb, "Deep reinforcement learning for 5G networks: Joint beamforming, power control, and interference coordination," *IEEE Trans. Commun.*, vol. 68, no. 3, pp. 1581–1592, Mar. 2020.
[26] Z. Xu, Y. Wang, J. Tang, J. Wang, and M. C. Gursoy, "A deep reinforcement learning based framework for power-efficient resource allocation in cloud RANs," in *Proc. IEEE Int. Conf. Commun. (ICC)*, Paris, France, May 2017, pp. 1–6.
[27] G. Zhao, Y. Li, C. Xu, Z. Han, Y. Xing, and S. Yu, "Joint power control and channel allocation for interference mitigation based on reinforcement learning," *IEEE Access*, vol. 7, pp. 177254–177265, 2019.
[28] N. Zhao, Y.-C. Liang, D. Niyato, Y. Pei, M. Wu, and Y. Jiang, "Deep reinforcement learning for user association and resource allocation in heterogeneous cellular networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 11, pp. 5141–5152, Nov. 2019.
[29] I. Alqerm and B. Shihada, "Energy-efficient power allocation in multitier 5G networks using enhanced online learning," *IEEE Trans. Veh. Technol.*, vol. 66, no. 12, pp. 11086–11097, Dec. 2017.
[30] H. Ding, F. Zhao, J. Tian, D. Li, and H. Zhang, "A deep reinforcement learning for user association and power control in heterogeneous networks," *Ad Hoc Netw.*, vol. 102, May 2020, Art. no. 102069.
[31] M. Miozzo, L. Giupponi, M. Rossi, and P. Dini, "Switch-on/off policies for energy harvesting small cells through distributed Q-learning," in *Proc. IEEE Wireless Commun. Netw. Conf. Workshops (WCNCW)*, San Francisco, CA, USA, Mar. 2017, pp. 1–6.
[32] *ECC Report 296: National Synchronisation Regulatory Framework Options in 3400–3800 MHz: A Toolbox for Coexistence of MFCNs in Synchronised Unsynchronised and Semi-Synchronised Operation in 3400–3800 MHz*, Electron. Commun. Committee, Eur. Conf. Postal Telecommun. Admin., Copenhagen, Denmark, 2019.
[33] *NR; Physical Channels and Modulation*, document TS38.211, Version 15.1.0, 3GPP, 2018.
[34] A. T. Z. Kasgari, W. Saad, M. Mozaffari, and H. V. Poor, "Experienced deep reinforcement learning with generative adversarial networks (GANs) for model-free ultra reliable low latency communication," *IEEE Trans. Commun.*, vol. 69, no. 2, pp. 884–899, Feb. 2021.
[35] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial networks," *Commun. ACM*, vol. 63, no. 11, pp. 139–144, Oct. 2020.

ANASTASIOS GIANNOPOULOS received the Diploma degree in electrical and computer engineering from the National Technical University of Athens and the M.Eng. degree from the National Technical University of Athens, in 2018, where he is currently pursuing the Ph.D. degree. He is included in the team composition of five European projects, among those an ongoing European 5G project, and has participated in one 5GPPP white paper from the technical board. He has authored articles in international scientific journals and conferences. His research interests include advanced optimization techniques for wireless systems, ML-assisted resource allocation, cognitive radios, and multi-dimensional data analysis.

SOTIRIOS SPANTIDEAS received the Diploma degree in electrical and computer engineering from the Polytechnic School, University of Patras, in 2010, with the thesis "Design of Meander Printed Antennas in Portable Terminal Devices," the M.Sc. degree in electrophysics from the KTH Royal Institute of Technology, Stockholm, in 2013, and the D.Eng. degree from NTUA with a doctoral dissertation "Development of Methods for obtaining DC and low frequency AC magnetic cleanliness in space missions." He is currently a Postdoctoral Research Associate with the Wireless and Long Distance Communications Laboratory, NTUA. His research interests include electromagnetic compatibility, machine learning for wireless networks, magnetic cleanliness for space missions, and optimization algorithms for computational electromagnetics.

NIKOLAOS KAPSALIS received the B.Sc. degree in computer science from the Department of Informatics, Athens University of Economics and Business, in 2008, and the M.B.A. degree in engineering economic systems and the Ph.D. degree from the School of Electrical and Computer Engineering, NTUA, in 2011 and 2015, respectively. He has authored articles in international scientific journals and conferences and coauthored a published book's chapter. His research interests include advanced signal processing, radio resource management, and optimization methods for wireless networks.



PANAGIOTIS KARKAZIS received the Ph.D. degree from the Technical University of Crete, in 2014. He has extensive experience as a Software Engineer for embedded systems in the telecommunications industry (research and development). He is currently an Assistant Professor with the Department of Informatics and Computer Engineering, University of West Attica. Moreover, he has participated as the Principal Researcher in many national and European research projects on the fields of the embedded systems, the IoT, cloud computing, and SDN (Google Scholar: https://scholar.google.gr/citations?user=d4vpm4EAAAAJ&hl=el).

PANAGIOTIS TRAKADAS received the Diploma and M.E. degrees in electrical and computer engineering and the Ph.D. degree from the National Technical University of Athens. He has been with the Technological Educational Institute of Athens as an Assistant Professor and the Technological Educational Institute of Chalkida and the National Technical University of Athens as a Senior Researcher. He is currently an Associate Professor with the Department of Ports Management and Shipping, National and Kapodistrian University of Athens. He has been actively involved in many research projects. He has published more than 130 papers in magazines, journals, books, and conferences (Google Scholar: https://scholar.google.com/citations?user=pP-GyYEAAAAJ&hl=el). His main interests include 5G technologies, such as NFV and NOMA, wireless sensor networks, signal processing, and routing protocols. He is a TPC member of conferences and a reviewer of several journals.