Joint Energy-efficient and Throughput-sufficient Transmissions in 5G Cells with Deep Q-Learning

Sotirios T. Spantideas
General Department
National and Kapodistrian University of Athens
Athens, Greece
sospanti@uoa.gr

Anastasios E. Giannopoulos
General Department
National and Kapodistrian University of Athens
Athens, Greece
angianno@uoa.gr

Nikolaos C. Kapsalis
Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
ncapsalis@gmail.com

Alexandros Kalafatelis
General Department
National and Kapodistrian University of Athens
Athens, Greece
kalafatelis.alexander@gmail.com

Christos N. Capsalis
Electrical and Computer Engineering
National Technical University of Athens
Athens, Greece
ccaps@central.ntua.gr

Panagiotis Trakadas
General Department
National and Kapodistrian University of Athens
Athens, Greece
ptrakadas@uoa.gr

***Abstract*— As a consequence of the 5G network densification and heterogeneity, there is a competitive relationship between the sufficient satisfaction of the cell users and the power-efficiency of 5G transmissions. This paper proposes a Deep Q-Learning (DQL) based power configuration algorithm by jointly optimizing the energy-efficiency (EE) and throughput-adequacy (JET) of 5G cells. The algorithm exploits the user demands to effectively learn-and-improve the user fulfillment rate, while ensuring cost-efficient power adjustment. To evaluate the potency of the developed methodology, several validation setups were conducted comparing the outcomes of the JET-DQL with those derived from conventional power control schemes, namely a Water-filling (WF) algorithm, a weighted minimum mean squared error (WMMSE) method, a heuristic solution and three fixed power allocation policies. JET-DQL algorithm exhibits a remarkable trade-off between the allocated throughput (ensuring high user satisfaction rates and average behavior in total allocated throughput relative to baselines), while resulting into low-valued (almost minimum) power configurations. In particular, even for strict demand scenarios, JET-DQL outperforms the other baselines with respect to EE showing a gain of 2.9-4.5 relative to others, although it does not provide the optimal sum-rate utility and minimum power levels.**

Keywords—5G network, Reinforcement learning, Energy efficiency, Deep Q-learning, Power allocation, Radio resource management

I. INTRODUCTION

Driven by the rapid evolution of wireless communication systems, spanning from novel schemes on the radio channel, such as the concept of Non-Orthogonal Multiple Access (NOMA) [1], to the introduction of software-defined, virtualized network services, the fifth-generation (5G) networks have gained great interest to enable previously-unseen capabilities to support services in several vertical domains [2]. According to 5G expectations [3], there are three main types of services, namely the massive machine-type communication (mMTC), enhanced mobile broadband (eMBB) and ultra-reliable low-latency communication (URLLC) services. All 5G types of communication will exhibit strict QoS with network connectivity requirements even for cell-edge users and under severe interference [3].

Thus, in 5G networks, billions of wireless devices are provisioned to be interconnected, communicating in a fast, heterogeneous and reliable manner. The considerable increase in the spatial density of the network architecture raises, in turn, significant challenges in the radio resource management (RRM) entity. In general, network densification results in lower probability of finding uncovered areas and/or users, but, contradictorily, increases the probability of severe interference. Moreover, the dense wireless environment also implies the simultaneous operation of multiple transmitters (e.g. macro and small-cell radio units, IoT devices) that are inevitably forced to use prohibitive power levels in order to surmount the experienced interference in their established links and provide adequate QoS to end users. On the other hand, demanding energy efficiency (EE) requirements of the wireless networks necessitate for deployment of novel power regulation schemes that incorporate autonomous, high-speed and smart solutions. Considering the significance of improving the EE by a factor of 2000 compared to existing network configurations [4], the power adjustment of both macro- and small-cell radio units has to be effectively re-addressed not only to guarantee interference mitigation, but also to provide considerably increased EE levels.

To this end, several suboptimal methods have been proposed to solve the EE maximization problem in wireless networks, which is typically an NP-hard problem [5]. The EE optimization problem has conflicting objectives, since it entails both the minimization of power levels of the transmitters and, at the same time, the maximization of the experienced throughput and interference mitigation, parameters that are usually positively correlated to the power increase. Traditional approaches to the non-convex EE maximization problem include interference mitigation methodologies and exploitation of the orthogonality between transmissions [4], [6]. More advanced techniques that involve fractional programming and sequential convex optimization or heuristic algorithms target at finding suboptimal solutions [7], [8]. However, the former methods suffer from poor resource utilization (e.g. spectral efficiency), while the latter are unable to provide effective solutions to large-scale wireless networks due to the complexity of the telecommunication environment and their poor convergence time.

The tremendous progress in both computing power and artificial intelligence (AI) algorithms have placed the solution of non-convex optimization problems more closely to

automated processes, rather than ruled-based, brute-force approaches. With reinforcement learning (RL), a specific branch of AI, it is possible to find the optimal strategy according to which an agent will achieve an objective after interacting with the environment [9]. As opposed to deep learning (DL), deep-RL (DRL) methods do not require training data to learn from, instead they can capture knowledge following a trial-and-error approach, while also exhibiting powerful generalization capabilities, since they can be easily implemented on large and complex environments [10]. Additionally, DRL pre-trained agents can be directly deployed in the network and provide near-real time predictions. Several studies have recently used RL methods in resource allocation problems, showing that usually RL outperforms the previously known, rule-based search algorithms [11]. Specifically, the authors in [12] propose an RL-based online learning power allocation framework in order to maximize the energy efficiency in multi-tier 5G heterogeneous networks. Moreover, a branch-and-cut method combined with DL was proposed in [4] to provide global EE-driven power control, whereas a power allocation RL strategy has been proposed in [13] to mitigate the network power consumption in cloud radio access networks (RANs), while maintaining the user demands. Other studies [14], [16] proposed similar RL frameworks aiming at maximizing the total network throughput by adjusting the power of the wireless transmitters.

In this paper, an urban heterogeneous coverage area is considered and the problem of configuring the power of transmitters is formulated towards ensuring near-optimal system-level EE. To this end, we propose a Deep Q-Learning (DQL) framework which, given the association scheme and the demand vector of diverse services, attempts to jointly maximize the EE, without disrespecting the requested throughput (JET-DQL). The main contributions of this work are: (i) The JET-DQL algorithm is tested in a heterogeneous 5G-compliant wireless network including both primary and secondary transmitters and can be easily adjusted to diverse network configurations (number of cells/transmitters, 5G numerology, etc.), (ii) while the conventional approaches to solve the EE non-convex optimization problem are usually computationally intractable, pre-trained JET-DQL model can be effortlessly inferred for EE-targeted power control, (iii) the EE and the throughput adequacy are jointly optimized and (iv) as indicated by the results, the developed method achieves a remarkable balance between the conflicting objectives of total allocated throughput maximization and sum-power consumption minimization.

II. Network Model and Problem Formulation

A. Network and Interference Model

An urban 5G network area is considered to be heterogeneously deployed, consisting of a macro-cell (UMa) and K overlapped micro-cells (UMi). Each cell k is covered by the respective transmitter k (Tx_k, $\forall k = 1,2,\dots,K+1$, where $k = 1$ corresponds to the UMa transmitter). According to the selected operational 5G band and physical resource block (PRB) segmentation (5G numerology), each cell has N available PRBs for physical-layer transmissions. A complete frequency reuse scheme across cells is also assumed. The system is controlled by a centralized cognitive controller, which targets at jointly accommodating the user requested services and ensuring low-valued power configuration. Each user ($u = 1,2,\dots,U$) requested a throughput-specific service s defined by the service level agreement (SLA) profile of the available services ($s = 1,2,\dots,S$). Each service corresponds to a particular throughput demand in order to ensure adequate QoS. Therefore, a demand vector D, with respective elements d_i ($i = 1,2,\dots,U$), is adapted to notify the requested service class of user u, expressed in terms of throughput (Mbps). Each single user u may be associated with a PRB n of a particular cell k, thus defining the association matrix A ($A_{k,n,u} = 1$ when the association occurs, or 0 otherwise). The power level of cell k over PRB n is denoted as $P_{n,k}$. Moreover, a sum-power constraint is established for each cell due to power budget limitations (separate thresholds for UMa and UMi cells), i.e. $\sum_{n=1}^{N} P_{1,n} \leq P_{max}^{UMa}$ and $\sum_{n=1}^{N} P_{k,n} \leq P_{max}^{UMi}, \forall k \geq 2$. To account for signaling/sleeping mode operations, a PRB-specific minimum power level is also defined, i.e. $P_{k,n} \geq P_{min},, \forall k \geq 1, n \geq 1$. Noteworthy, we assume that (i) each user is connected to a single PRB, (ii) the cell capacity is upper-bounded by the number of available PRBs and (iii) each cell can cover multiple users (at most N users).

In heterogeneous environments, each user receives not only the signal from the associated UMa or UMi Tx, but also the accumulated interference signals from other operating stations. Inter-cell interference is taken into account via calculating the signal-to-interference-plus-noise (SINR) ratio γ. Given a particular association link between cell k and user u over PRB n, parameter γ is given by:

$$\gamma_{k,n,u} = \frac{P_{k,n} \cdot L_{k,n,u}}{(\sum_{k' \neq k} P_{k',n} \cdot L_{k',n,u}) + n_0}, \quad (1)$$

where $L_{k,n,u}$ denotes the channel losses that characterize the link between Tx k and user u over PRB n, and n_0 stands for the received noise power. Notably, the channel losses reflect the shadowing and path losses, while they are also positively correlated to the Tx-user distance [17]. The cell heterogeneity implies that different channel models should be considered, separately for UMa and UMi transmitters. Specifically, to reflect 5G-compliant conditions, UMa model and UMi model were assumed for the channel loss estimation of macro- and micro-links, respectively, as suggested in [17]. The achievable data rate of a particular link is characterized according to the respective γ status. The computation of the reachable transmission rate R of the link between Tx k and user u over PRB n is based on the Shannon formula, as follows:

$$R_{k,n,u} = W_n \cdot \log_2(1 + \gamma_{k,n,u}), \quad (2)$$

where W_n is the bandwidth of PRB n, depending on the selected 5G numerology and operating band.

B. Problem Formulation

In this section, we address the problem of adjusting the transmitting power levels across all cells and PRBs of the network in order to maximize the ratio of the total network throughput and total power budget (i.e. EE). The system-level EE is, thus, expressed by:

$$EE_{system} = \frac{\sum_{u=1}^{U} R_u}{\sum_{k=1}^{K+1} \sum_{n=1}^{N} P_{k,n}}, \quad (3)$$

where R_u is the experienced throughput of user u given the respective associated cell and PRB. To avoid over-satisfaction of the users' QoS, the numerator of the EE in Eq. 3 is modified so as each user-specific EE term to be upper-bounded by the

requested throughput. The *EE optimization problem (P)* is then formulated as follows:

$$(P) \quad \max_{\boldsymbol{P}} \left\{ EE_{system} = \frac{\sum_{u=1}^{U} min\{d_u, R_u\}}{\sum_{k=1}^{K+1} \sum_{n=1}^{N} P_{k,n}} \right\} \tag{4}$$

s.t.:

$$(C1) \quad \sum_{n=1}^{N} P_{1,f} \leq P_{max}^{UMa} \tag{5}$$

$$(C2) \quad \sum_{n=1}^{N} P_{k,n} \leq P_{max}^{UMi}, \forall k = 2, \dots, K+1 \tag{6}$$

$$(C3) \quad P_{k,n} \geq P_{min}, \forall k = 1, \dots, K+1; n = 1, \dots, N \tag{7}$$

$$(C4) \quad \sum_{k=1}^{K+1} \sum_{n=1}^{N} A_{k,n,u} \leq 1, \forall u = 1, \dots, U \tag{8}$$

Optimization problem *(P)* is solved by finding a power configuration vector $\boldsymbol{P}$ such that the system-level E is maximized. Moreover, constraints (C1)-(C4) guarantee the sum-power limitations of UMa (C1) and UMi (C2) transmissions, the minimum power level of each PRB (C3) and the link allocation scheme according to which every association link contains at most one single-user (C4).

C. Deep Q-Learning Principles

In principle, a DRL agent observes the current state $s \in S$ of the environment, takes an action $a \in A$ (action space) and receives a reward r, reflecting the impact of the performed action. Q-learning allows the agent to predict the quality of being in state s_t and performing the action a_t, based on the Bellman equation [9]:

$$Q_t(s_t, a_t) = (1-\alpha) \cdot Q_{t-1}(s_t, a_t) + \alpha \cdot (r(s_t, a_t) + \gamma \cdot \max_{a'} \{Q(s_{t+1}, a')\}) \tag{9}$$

The Bellman equation defines the update rule of the Q-table at time t and implies that the new Q-value depends on both the previous Q-value for a given state-action pair (first term), while the second term represents the immediate reward ($r(s_t, a_t)$) and the optimal future discounted reward ($\gamma \cdot \max_{a'} \{Q(s_{t+1}, a')\}$). Additionally, the learning rate $\alpha \in [0,1]$ and the discount factor $\gamma \in [0,1]$ are used to balance between old/new Q-values and discount the future rewards, respectively. The agent gradually gathers experience about the beneficial actions and finally gains sufficient knowledge about the environment, meaning that the temporal difference (TD) between the learned value $r(s_t, a_t) + \gamma \cdot \max_{a'} \{Q(s_{t+1}, a')\}$ and the old value $Q_{t-1}(s_t, a_t)$ is minimized. The TD function may be expressed as [9]:

$$TD_t(s_t, a_t) = \left(r_t(s_t, a_t) + \gamma \cdot \max_{a'} \{Q(s_{t+1}, a')\}\right) - Q_{t-1}(s_t, a_t) \approx 0 \tag{10}$$

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 main idea of DQL relies on the usage of two identical neural networks: (i) the *Q*-network which is used to estimate the current best action and (ii) the *target Q*–network which is used to predict the next action that will return the maximum long-term reward ($\max_{a'} \{Q(s_{t+1}, a')\}$). The outputs of the *Q*- and target *Q*-networks are considered as the features and the labels, respectively, of the deep learning part of DQL. During the training phase, TD minimization is achieved by appropriately adjusting the weights of the *Q*-network neurons, such that the loss function (difference between the predicted and actual values) is significantly reduced. Finally, the trained *Q*-network acts as a consultant of the agent, guiding its action selection policy throughout the inference (online) process.

D. JET-DQL Framework

To obtain a (sub)optimal solution for the optimization problem (P), a DQL agent is set to interact with a heterogeneous wireless environment. Fig. 1 outlines the training and inference pseudocode of the JET-DQL algorithm. The JET-DQL design parameter are defined as follows:

***State space*:** It is a function that describes the telecom environment, transforming the action taken in the previous step into a reward and a new set of actions. In the proposed algorithm, the state space includes the association information of each user u and a throughput tolerance indicator tol_u, discretized in 5 tolerance levels. Specifically, the quantized values of tolerance are given by:

$$tol_u = \begin{cases} 0, & R_u/d_u < 0 \\ 1, & 1 \leq R_u/d_u < 1.2 \\ 2, & 1.2 \leq R_u/d_u < 1.5 \\ 3, & 1.5 \leq R_u/d_u < 2 \\ 4, & R_u/d_u \geq 2 \end{cases} \tag{11}$$

where the categorical values of tol_u reflect the user's satisfaction status, ranging from under- to over-satisfaction levels (range 0-4). Therefore, the temporal sequence of the state space is $S = \{S_1, \dots, S_t, \dots, S_T\}$, where at a given time t, $S_t = [(Tx_1, PRB_1, tol_1), \dots, (Tx_U, PRB_U, tol_U)]$ determines the system state. This means that the controller, before taking an action, knows a triplet for each user u, namely the associated Tx and PRB identifiers, as well as a demand-related tolerance level. Noteworthy, user association is applied according to the maximum throughput criterion.

***Action space*:** The controller performs a sequence of actions $\{A_1, \dots, A_t, \dots, A_T\}$ during an episode, i.e. a complete series of agent-environment interactions, beginning from the initial state and terminating in the final state. At a given step t, the agent performs a power 'increment', 'decrement' or 'null' move on a selected PRB of each Tx. Formally, the action taken at time t is denoted as $A_t = [(n_1, a_1), \dots, (n_{K+1}, a_{K+1})]$ and the power change on the n-th PRB of the k-th Tx is expressed by $a_{n_k} \in \{P_{step}, -P_{step}, 0\}$, where the power step P_{step} is constant. After selecting an action, the power update rule at Tx k is given by $P_{k,n_k}(t+1) = P_{k,n_k}(t) + a_{k_n}(t)$.

***Rewarding system*:** The action taken by the agent results into a different system state, thus leading to association configuration and tolerance levels. The reward returned at time t is defined as follows:

$$r_t(S_t, A_t) = \begin{cases} 100 \dfrac{EE_t - EE_{t-1}}{EE_{t-1}}, & if\ EE_t > EE_{t-1} \\ 0, & otherwise \end{cases} \tag{12}$$

where EE_t is system-level EE at time t. Note that, a non-zero reward quantifies the percentage increment in EE resulted by

taking the current action. Intuitively, this rewarding system guides the agent to gradually prefer those sequence of actions that constantly improve the network EE.

Algorithm 1. JET-DQL Algorithm

Training phase:

Initialize $\varepsilon = 1$, learning rate and discount factor
Initialize Q- and $target\ Q$-nets (random weights θ)
Initialize replay memory and mini-batch size
Initialize the power levels of all transmitters.
for $episode = 1, \dots, T$ **do**
  Place the users randomly in the network area
  Associate users (max throughput criterion)
  for $t = 1, \dots, T'$ **do**
    With probability ε select a random A_t, otherwise select A_t with the highest Q-value.
    Take action A_t and observe r_t, S_{t+1}
    Save tuple (S_t, A_t, r_t, S_{t+1}) in replay memory
    Select randomly a mini-batch of experience tuples from replay memory
    Set target values:

$$y_j = \begin{cases} r_j, if\ terminal\ state \\ r_j + \gamma\ max_{a'}\ Q(s+1, a'; \theta), otherwise \end{cases}$$

    Do gradient descent on: $(y_j - Q(s_t, a_j; \theta))^2$
  end for
  Decay ε
  Every X episodes set the weights of $target\ Q$-net equal to Q-net
end for

Inference phase:

Load the pre-trained Q-model
for $validation\_episode = 1, \dots, M$ **do**
  Observe the state S_t
  Feed S_t to the Q-net and select the action with the highest Q-value
  Observe new state and reward
  while $EE\ increment$ (or positive reward) is observed:
    repeat steps 3-4
  end while
  Calculate the accumulative EE increment
  Store the final Throughput, Power vectors.
end for

Fig. 1. JET-DQL algorithm pseudocode for training and inference phases.

III. Simulation Results

Simulations of the JET-DQL algorithm were implemented in Python 3.8 using Tensorflow 2.4. The training phase of the developed algorithm took ~1 hour running on a personal PC with a CPU i7-8700 at 3.2 GHz and a RAM of 8 GB (no GPU usage). Initially, the impact of the algorithm hyper-parameters is assessed and their values are stabilized depending on the performance of the algorithm. Subsequently, the proposed JET-DQL algorithm is compared against multiple baseline power allocation methods to elucidate its effectiveness. Table I summarizes the setup parameters both for the configuration of the telecommunication wireless network and the architectural design of the DQL network. Evidently, considering that each user may occupy a single PRB, each UMi cell reaches its full capacity (in terms of spectrum utilization) when the number of associated users equals to the number of available PRBs (6 PRBs for the selected 5G numerology 4). The above network realization ensures that the UMa transmitter unavoidably causes interferences across all established UMi links. Finally, we assume three different types of requested SLAs, with an SLA-throughput mapping of SLA-{1, 2, 3} = {1, 2.5, 5} Mbps.

TABLE I. Simulation Setup Parameters

Network Parameters		**JET-DQL Architecture**	
Parameter	**Value**	**Parameter**	**Value**
Central Frequency	6 GHz	Update target frequency	100
Number of PRBs	6	Memory size	5000
Number of users per UMi	6	Mini-batch size	64
5G numerology	4	Loss function	Huber loss
PRB Bandwidth	2.88 MHz	Number of hidden layers	3
Number of UMi cells	4	Activation function of input and hidden layers	Rectified Linear (ReLu)
UMa/UMi Power Constraint $P_{max}^{UMa}/P_{max}^{UMi}$	80/25 W	Activation function of output layer	Linear
Minimum Power Constraint P_{min}	0.1 W	Monte-Carlo simulations	1000
UMa/UMi radius	500/100 m	Optimizer	Adam
Noise Power Density	-174 dBm/Hz	ε decay	Linear

A. Impact of hyper-parameters on JET-DQL training

The stochastic behavior of the DQN algorithms implies cautious fine-tuning among the crucial learning parameters, namely the number of episodes (directly affecting the exploration duration), the learning rate (a, balances the contribution of the new and old Q-value in the Bellman formula), the discount factor (df, defines the significance of future rewards), as well as the power granularity (P_{step}).

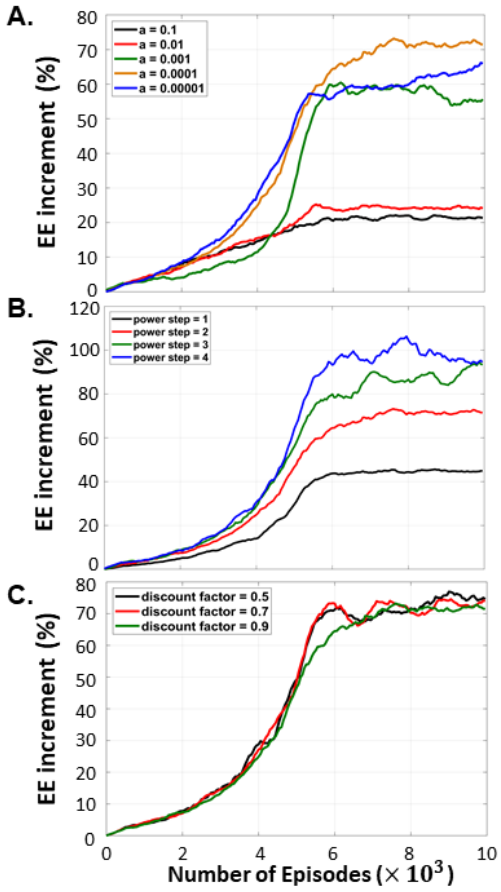

Fig. 2. Learning curves of the JET-DQL training phase for different values of learning rate (A), power step (B) and discount factor (C).

In the first part of the training, we examined the impact of the number of training episodes on the convergence rewards, observing that more than 10000 episodes result into similar reward values. Next, we set constant $df = 0.9$ and $P_{step} = 2$ and experimented with different learning rate values $a \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$ (see Fig. 2A), noticing the best total reward (73%) for $\alpha = 10^{-4}$. The impact of the df was then investigated (see Fig. 2C), showing a systematic independence between the reward convergence (~70%) and df (0.5, 0.7 and 0.9). Moreover, power granularity impact is shown in Fig. 2B, where the learning curve is depicted for different values of P_{step} (~92% for both 3 and 4 Watts).

Finally, to confirm the absence of significant convergence alterations in cases that diverse hyper-parameter combinations are used, we also conducted training simulations experimenting with all the possible dyads of $a \in \{10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}\}$ and $P_{step} \in \{1, 2, 3, 4\}$ Watts. Fig. 3 depicts the reward converge value (% accumulative EE increment) of the above training setups. Again, it was verified that the optimal converge values are obtained for $\alpha = 10^{-4}$, independently of using P_{step} of 3 (92.3% total reward) or 4 Watts (92.4% total reward). For the rest of the simulations, we used the pre-trained JET-DQL model parameterized with the optimal hyper-parameters.

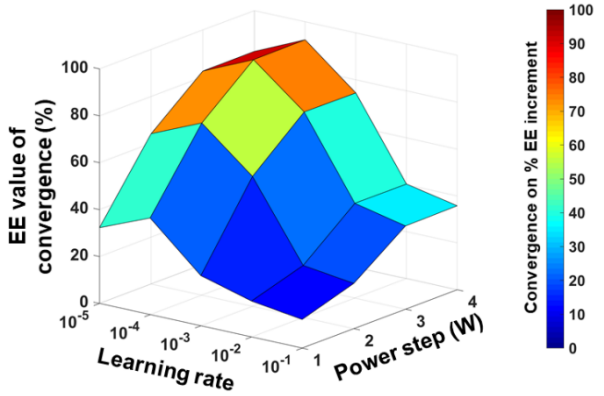

Fig. 3. A 3-D representation of the JET-DQL convergence value (EE increment) as function of both learning rate and power step.

B. Comparison with Baseline Methods

To evaluate the performance of the developed methodology, we compare the JET-DQL outcomes against five well-established baseline power allocation methods. Specifically, two global network-wide throughput optimization algorithms, namely the Water-Filling (WF) and the Weighted Minimum Mean Square Error (WMMSE) methods, were implemented according to [18] and [19], respectively. Moreover, two additional fixed power control policies were considered for comparison purposes: the minimum power allocation scheme (MIN), according to which all UMi cells transmit with the minimum power level as the optimal power consumption solution, as well as the average power allocation scheme (AVG), which offers a reasonable trade-off between the consumed power and the achievable throughput. Finally, a heuristic Particle Swarm Optimization (PSO) solution was also implemented, targeting at the system-level EE maximization with identical power constraints.

All methods were contrasted in terms of the three following metrics: (i) total allocated throughput, (ii) the total power consumption and (iii) the system-level EE. The above metrics were calculated as the average values across 1000 different validation simulations (Monte-Carlo simulations), where in each simulation the users are randomly placed within the network area and the initial power levels of all transmitters are also randomly initialized. To account for various demanding situations, 4 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 or SLA-2 services and *Scenario 4:* random requests of SLA-1, SLA-2 or SLA-3 services.

As depicted in Fig. 4, WF and WMMSE algorithms are the best methods with regards the system throughput maximization (Fig. 4A). This is directly attributed to their objective in searching for power configuration that results into increased experienced throughput for the individual links with good channel conditions. However, these approaches have the drawback of over-satisfying already fulfilled users as an attempt to significantly improve the total network-wide throughput. In this context, both methods provide poor power consumption (Fig. 4B) and EE (Fig. 4C) performance, primarily due to the enhanced allocated power levels at PRBs with high-SINR.

As expected, MIN method showed constantly the best performance in total power consumption as it uses the most cost-efficient power levels. On the contrary, minimum power allocation results into poor system throughput, mainly because the received signals are not sufficient to accommodate the SLA requirements. Overall, MIN results show also poor EE performance, since the high reduction in the denominator of the EE is counterbalanced by extremely low nominator values (i.e. total throughput). Moreover, AVG method exhibits median performance among all evaluation metrics, since it systematically guarantees a reasonable ratio between the experienced throughput and the allocated power, showing also acceptable EE solutions. Regarding the PSO performance, we observed adequate values across all evaluation metrics. Although PSO outperforms the other four baseline methods in the majority of validation scenarios, it fails to yield better performance than JET-DQL. Notably, unlike the other baseline methods (MIN, AVG, WF, WMMSE and JET-DQL inference phase) that do not have considerable response time to find the power allocation solution (all in the order of ms), PSO requires a run time in the order of min (~6 minutes in our simulations) to obtain a sub-optimal solution. However, it provides ~29% enhanced EE solutions in relation to WF, MIN and WMMSE, as well as noticeably lower power consumption. Integrating the above observations, we noticed remarkably beneficial outcomes of the JET-DQL approach, primarily exhibiting an EE gain in the range of 2.9-4.5 relative to the other baselines. It is also worth noting that the proposed algorithm provides median solution (similar to that of AVG) in terms of the total allocated throughput (~36% below WF and WMMSE), while showing near-optimal performance in power consumption (follows the MIN solution).

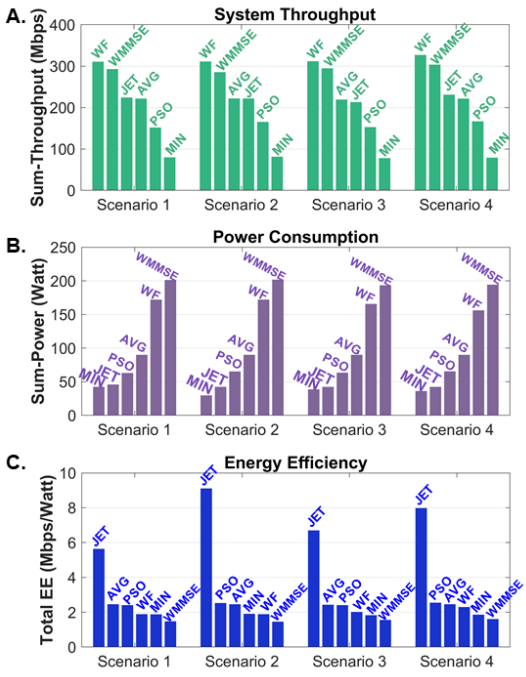

Fig. 4. Performance comparisons between JET-DQL algorithm and the five baseline methods in terms of (A) System Throughput, (B) Power Consumption and (C) Energy-efficiency.

To sum up, results confirmed the effectiveness of both WF and WMMSE in optimizing the total network-wide throughput (around 36% enhancement relative to JET), while neglecting the cost-efficiency of the allocated power vectors (increased power consumption by a factor of ~4.5 compared with JET). The achieved equilibrium between the total allocated throughput and the wasted sum-power offered by JET-DQL algorithm may highlight its main objective: maximization of the EE at the cost of degrading the total system throughput.

IV. Conclusion

In the present work, a multi-channel 5G-compliant system is considered, including heterogeneous cells, while the power levels of the transmitters are supervised by a centralized DQL agent. Towards the stabilization of the DQL hyper-parameters, we initially present the impact of the latter on the algorithm performance (percentage of EE increment) employing numerous training simulations. To evaluate the proposed scheme, several gradually demanding (in terms of SLA requirements) validation scenarios were considered. Although WF and WMMSE methods were the optimal techniques for total network-wide throughput maximization, JET-DQL showed dominant performance regarding the EE optimization compared to all baseline methods, illustrating an average solution for the sum-rate utility and near-optimal power consumption abilities. Overall, results confirm that JET-DQL can achieve a reasonable trade-off between power savings and system throughput degradation.

Acknowledgment

This work has been partially supported by the Affordable5G project, funded by the European Commission under Grant Agreement H2020-ICT-2020-1, number 957317 through the Horizon 2020 and 5G-PPP programs (www.affordable5g.eu/).

References

[1] Nomikos, N., et al., (2019). Flex-NOMA: exploiting buffer-aided relay selection for massive connectivity in the 5G uplink. *IEEE Access*, *7*, 88743-88755.

[2] Trakadas, P., et al., (2019). Hybrid clouds for data-intensive, 5G-enabled IoT applications: An overview, key issues and relevant architecture. *Sensors*, *19*(16), 3591.

[3] Andrews, J. G., Buzzi, S., Choi, W., Hanly, S. V., Lozano, A., Soong, A. C., & Zhang, J. C. (2014). What will 5G be?. *IEEE Journal on selected areas in communications*, *32*(6), 1065-1082.

[4] Matthiesen, B., Zappone, A., Besser, K. L., Jorswieck, E. A., & Debbah, M. (2020). A globally optimal energy-efficient power control framework and its efficient implementation in wireless interference networks. *IEEE Transactions on Signal Processing*, *68*, 3887-3902.

[5] Luo, Z. Q., & Zhang, S. (2008). Dynamic spectrum management: Complexity and duality. *IEEE journal of selected topics in signal processing*, *2*(1), 57-73.

[6] Xu, Q., Li, X., Ji, H., & Du, X. (2014). Energy-efficient resource allocation for heterogeneous services in OFDMA downlink networks: Systematic perspective. *IEEE Transactions on Vehicular Technology*, *63*(5), 2071-2082.

[7] Zappone, A., Sanguinetti, L., Bacci, G., Jorswieck, E., & Debbah, M. (2015). Energy-efficient power control: A look at 5G wireless technologies. *IEEE Transactions on Signal Processing*, *64*(7), 1668-1683.

[8] Adedoyin, M., & Falowo, O. (2016, September). An energy-efficient radio resource allocation algorithm for heterogeneous wireless networks. In *2016 IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)* (pp. 1-6). IEEE.

[9] Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press

[10] Morocho-Cayamcela, M. E., Lee, H., & Lim, W. (2019). Machine learning for 5G/B5G mobile and wireless communications: Potential, limitations, and future directions. *IEEE Access*, *7*, 137184-137206.

[11] Zhang, C., Patras, P., & Haddadi, H. (2019). Deep learning in mobile and wireless networking: A survey. *IEEE Communications Surveys & Tutorials*, *21*(3), 2224-2287.

[12] AlQerm, Ismail, and Basem Shihada. "Energy-efficient power allocation in multitier 5G networks using enhanced online learning." *IEEE Transactions on Vehicular Technology* 66.12 (2017): 11086-11097.

[13] Xu, Z., Wang, Y., Tang, J., Wang, J., & Gursoy, M. C. (2017, May). A deep reinforcement learning based framework for power-efficient resource allocation in cloud RANs. In *2017 IEEE International Conference on Communications (ICC)* (pp. 1-6). IEEE.

[14] Zhao, G., Li, Y., Xu, C., Han, Z., Xing, Y., & Yu, S. (2019). Joint Power Control and Channel Allocation for Interference Mitigation Based on Reinforcement Learning. *IEEE Access*, *7*, 177254-177265.

[15] Giannopoulos, A., Spantideas, S., Tsinos, C., & Trakadas. P. (2021). Power Control in 5G Heterogeneous Cells considering User Demands using Deep Reinforcement Learning. In *17th International Conference on Artificial Intelligence Applications and Innovations,* in press.

[16] Zhang, Y., Kang, C., Ma, T., Teng, Y., & Guo, D. (2018, August). Power allocation in multi-cell networks using deep reinforcement learning. In *2018 IEEE 88th Vehicular Technology Conference (VTC-Fall)* (pp. 1-6). IEEE.

[17] 3GPP, Study on channel model for frequencies from 0.5 to 100 GHz. *Technical report (TR) 38.901, 3rd Generation Partnership Project (3GPP)*, 201

[18] Qi, Q., Minturn, A., & Yang, Y. (2012, May). An efficient water-filling algorithm for power allocation in OFDM-based cognitive radio systems. In *2012 International Conference on Systems and Informatics (ICSAI2012)* (pp. 2069-2073). IEEE.

[19] Shi, Q., Razaviyayn, M., Luo, Z. Q., & He, C. (2011). An iteratively weighted MMSE approach to distributed sum-utility maximization for a MIMO interfering broadcast channel. *IEEE Transactions on Signal Processing*, *59*(9), 4331-4340.