# WIP: Demand-Driven Power Allocation in Wireless Networks with Deep Q-Learning

A. Giannopoulos[1], S. Spantideas[1], N. Capsalis[1], P. Gkonis[2], P. Karkazis[3], L. Sarakis[2], P. Trakadas[2] and C. Capsalis[1]

[1]*National Technical University of Athens, 9, Iroon Polytechniou str., Athens, 15780, Greece.*
*E-mails: angianno@mail.ntua.gr, sspantideas@central.ntua.gr, ncapsalis@gmail.com, ccaps@central.ntua.gr*
[2]*General Department, National and Kapodistrian University of Athens, Sterea Ellada, 34400 Dirfies Messapies, Greece.*
*E-mails: pgkonis@uoa.gr, ptrakadas@uoa.gr, lsarakis@uoa.gr*
[3]*Department of Informatics and Computer Engineering, School of Engineering, University of West Attica, 12243 Athens, Greece. E-mail: p.karkazis@uniwa.gr*

***Abstract***—Power allocation is strongly related to the coverage and capacity of wireless networks, playing a critical role in the development of 5G networks. This paper proposes a Demand-Driven Power Allocation (DDPA) algorithm aiming to fulfill the requested throughput of individual users and accommodate their needs. DDPA is based on model-free Deep Reinforcement Learning (DRL) approaches and has the ability to proactively adjust the power levels of network transmitters. The performance of the developed algorithm is evaluated for a variety of simulation parameters and variable user demands. According to the presented results, the DDPA scheme exhibits a near-optimal performance for up to 50 users in the network area (i.e. satisfaction percentage exceeds 95%), with each one requesting 1 Mbps. Moreover, performance comparison between DDPA and two typical baseline methods reveals that the former results into enhanced total allocated throughput solutions (i.e. a performance increase by a factor of approximately 9% against baseline methods).

***Keywords***—Reinforcement learning, Power allocation, Deep Q Network, Resource allocation

## I. Introduction

Fifth-generation (5G) networks will act as a unified connectivity platform for future innovations, embracing the evolving cellular systems to incorporate diverse services, devices and deployments. It is expected that 5G networks will support billions of inter-connected devices, ensuring high-quality of experience [1]. Such networks are provisioned to present self-organizing capabilities, including self-configuration of radio resources and self-optimization of network parameters [2]. However, due to the large number of mobile users and/or densely-deployed network elements, overall interference defines an upper-bound in the network performance.

In such complex environments, the optimal resource allocation inevitably becomes a non-convex problem, enforcing the idea of finding sub-optimal solutions [3]. Traditionally, heuristic algorithms have been widely used to solve Radio Resource Management (RRM) problems, such as brute-force search, genetic algorithms, rule-based and branch-and-cut approaches [4]-[5]. The main drawback of such methods is the excessive computational cost, as well as the inability to generalize their solutions, thus becoming infeasible for large-scale cellular systems [3]. Reinforcement learning (RL) algorithms have yielded promising results in various RRM problems so far and have been recently adapted for network optimization [6].

In the framework of designing and implementing the interference mitigation techniques towards ensuring sufficient Quality of Service (QoS) to mobile users, intelligent optimization methods aim to decrease the operational costs associated with commissioning additional network elements. Among a variety of physical layer optimization algorithms, power allocation has attracted considerable scientific interest. To this end, several power configuration algorithms have been proposed to optimize the network capacity, eliminate inter-cell interferences and regulate the coverage area of the network cells [7]-[10]. Specifically, the authors in [11] proposed a cooperative Q-learning algorithm to control the power of dense cells, while ensuring fairness across users. Moreover, a different RL strategy was followed in [12], where a multi-cell power allocation scheme targeting at the maximization of the overall network capacity was presented. In addition, a joint user association and resource allocation algorithm was proposed in [13], in order to maximize the long-term overall network utility in heterogeneous networks.

In this paper, a *Demand-Driven Power Allocation* (DDPA) algorithm is formulated and implemented on realistic network configurations. A general-purpose power control scheme is described, aiming to ensure acceptable QoS to mobile users by adjusting the power levels of radio units (RUs). The algorithm is based on Deep Reinforcement Learning (DRL), taking advantage of its generalizability and applicability to large state-action spaces. The developed algorithm targets at the fulfillment of the maximum possible allocated throughput, driven by the user-specific requested throughput. The main contributions of this paper include: (i) The proposed DRL-based power allocation scheme relies on a model-free algorithm, (ii) the presented framework allows the generalizability of the DDPA algorithm to complex telecommunication environments, (iii) as opposed to several existing power allocation approaches focusing on the total network-wide throughput maximization, DDPA addresses the power control problem from a demand-driven perspective, (iv) a novel rewarding system is established in order to overcome the significant drawback of the total network-wide throughput maximization (i.e. unbalanced throughput allocation that leads to over-satisfaction of some users and poor QoS for others) and (v) the DDPA algorithm can take advantage of training-inference split. In specific, the pre-trained DQL agent can be effortlessly inferred throughout the online network operation.

## II. System Model and Problem Formulation

### A. Network and Interference Model

A network area accommodating $M$ RUs is considered, each one having available $F$ resource blocks (RBs) that may be

978-1-6654-2263-5/21/$31.00 ©2021 IEEE.

grouped in $N$ equally-divided channels with the same bandwidth $B$. Moreover, the $m^{\text{th}}$ RU transmits over each channel $n$ with a specified power level $P_{m,n} = p(l)$, which is selected from a set of available power levels $\{l = 1,2,\dots,L\}$. Finally, a maximum total power constraint $P_{max}^{m}$ is considered for each RU. Each user $u \in \{1,2,\dots,U\}$ located inside the network area may be associated with $\geq 1$ RBs of a particular RU $m$. This user requests a service corresponding to a throughput demand vector $D_u$ (in Mbps). The system is supervised by a centralized intelligent controller, which effectively adjusts the transmit power of the RUs' channels (denoted as $P_{m,n}$).

The wireless environment implies accumulated interference signals from operating neighboring RUs in the network area. The signal-to-interference-plus-noise ratio (SINR) of the $u^{\text{th}}$ user that is associated with the $n^{\text{th}}$ channel of the $m^{\text{th}}$ RU may be expressed as:

$$SINR_u^{m,n,f} = \frac{P_{m,n,f} \cdot G_{m,n,u}}{(\sum_{m' \neq m}^{M} P_{m',n,f} \cdot G_{m',n,u}) + N_0}, \quad (1)$$

where $P_{m,n,f}$ denotes the transmit power of the $m^{\text{th}}$ RU over RB $f$ of channel $n$, $G_{m,n,u}$ denotes the channel gain (log-normal shadowing and the corresponding path-loss [7], [14]) from the $m^{\text{th}}$ RU to the $u^{\text{th}}$ user over the $n^{\text{th}}$ channel and $N_0{}'$ stands for the noise power at the receiver level. The downlink capacity can be calculated from the Shannon formula as:

$$R_u^{m,n} = j_{u,n,m} \frac{N \cdot B_n}{F} \cdot log\left(1 + SINR_u^{m,n,f}\right), \quad (2)$$

where $j_{u,n,m}$ is number of RBs assigned to satisfy the demands of user $u$.

### B. Problem Formulation

The target of the DDPA modeling methodology is to adjust the power vectors of the $M$ RUs in order to fulfil the requested throughput of each user. The optimization problem $(P)$ may be defined as follows:

$$(P) \quad min \sum_{u=1}^{U} (D_u - R_u^{m,n}) \quad (3)$$

*s.t.:*

$$(C1) \quad \sum_{m=1}^{M} a_{u,m,n} \leq 1, \forall u, \quad (4)$$

$$(C2) \quad \sum_{n=1}^{N} P_{m,n} \leq P_{max}^{m}, \forall m \quad (5)$$

$$(C3) \quad \sum_{u=1}^{U} a_{u,m,n} j_{u,n,m} \leq \frac{F}{N}, \forall m, n \quad (6)$$

$$(C4) \quad R_u^{m,n} \leftarrow min\{D_u, R_u^{m,n}\} \quad (7)$$

Equations (3)–(7) constitute the optimization problem constraints. Specifically, $(C1)$ ensures that each user can only be associated to a single RU, $(C2)$ reflects the power budget limitation, $(C3)$ secures the channel capacity overflow and, finally, $(C4)$ ensures that the user data-rate is upper bounded by the requested throughput requirements.

### C. The Deep Q-Learning Framework

A DRL framework is described via (i) the state space of the environment observed by the agent, (ii) the action space that contains all possible actions (iii) the rewarding function to model the environment responses. *State space*: The environment state is efficiently described via a three-fold information, namely the user-specific (i) Channel Quality Indicator (CQI, [14]), (ii) the serving RU number and (iii) the allocated channel ID. Formally, the system state is $S_t = [(CQI_1, RU_1, CH_1), \dots, (CQI_U, RU_U, CH_U)]$, at time $t$. The user association is based on the maximum throughput criterion. *Action space*: At a given time $t$, the performed action is $A_t = [(P_{1,1}, \dots, P_{1,N}), \dots, (P_{M,1}, \dots, P_{M,N})]$. *Reward system*: The feedback of the telecommunication environment may be described via the rewarding function:

$$r_t = \begin{cases} I = \sum_{u=1}^{U} min\{D_u, R_u^{m,n}\}_t - \sum_{u=1}^{U} min\{D_u, R_u^{m,n}\}_{t-1}, & if\ I > 0 \\ \sum_{u=1}^{U} \{D_u\}_t, & if\ \{R_u^{m,n}\}_t \geq \{D_u\}_t,\ \forall u \\ 0, & otherwise \end{cases} \quad (8)$$

The main goal of the rewarding function is to uniformly increase the allocated sum-rate among the individual users. Specifically, the agent receives:

*Case 1*: a positive reward equal to the difference $I$ between the current and the previous sum-rate in case of sum-rate increase,

*Case 2*: a high-valued positive reward equal to the total requested throughput if the demands of all users are totally fulfilled,

*Case 3*: a zero value when the current action does not increase the sum-rate with respect to its previous value.

*Action selection policy*: The *ε-greedy* method was used for the action selection strategy.

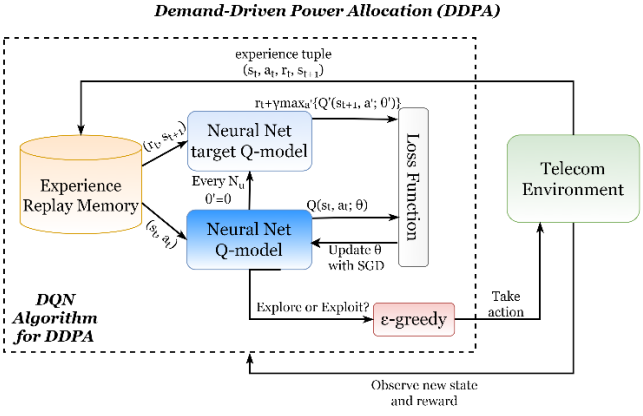


Fig. 1. The process of the DQL algorithm for DDPA.

The neural network that is employed to solve the optimization problem is depicted in Fig. 1, along with

progression of the DDPA algorithm. The number of neurons in the input layer is related to the state space, which in turn is positively correlated to the number of users. The output layer is based on the number of RUs and channels, i.e. the action space includes all the possible combinations of the power regulation for the channels of all RUs.

## III. Simulation Results

### *A. Simulation setup and DQN fine-tuning*

In this section, simulation results are provided regarding the performance of the proposed DDPA algorithm, which was implemented in Python 3.8. In this context, we consider a network area consisting of four RUs with a minimum inter-RU distance of 750m [14]. To validate the proposed RL scheme in extreme interference conditions, we also consider two channels for each RU ($B = 9MHz$, $f_c = 2GHz$, $F = 50$ RBs). Each channel operates at a specific power level selected by the set of available powers {6.4, 9.6, 12.8, 16, 19.2} W (with $P_{max} = 38\ W$). The antenna system of each RU had three equally-spaced sectors (HPBW = 70° and minimum gain = -35dB per sector [14]). The path loss part of the channel gain was computed according to the model specified in [14], with a noise power density of -174dBm/Hz. For simulation purposes, we indicatively consider three requested QoS levels, namely 0.1, 1 and 2.5 Mbps.

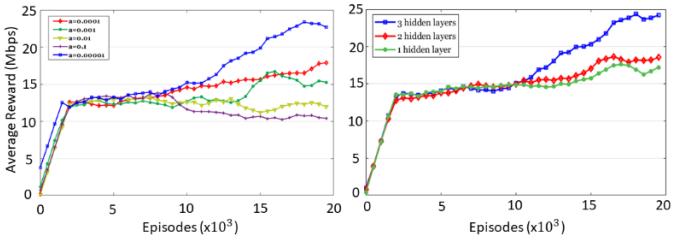


Fig. 2. Accumulated reward for different values of the learning rate $\boldsymbol{\alpha}$ (left) and number of hidden layers (right).

The performance of the DDPA algorithm is evaluated by monitoring the training process using various hyper-parameters in the neural network configuration. Fig. 2 depicts the accumulated reward for several values of the learning rate as a function of the training episodes (*left*), for different number of hidden layers in the *Q*- and *target Q*-networks (*right*).

Integrating the results of Figs. 2-3, the value of the learning rate is set at $10^{-4}$, whereas the neural networks are deployed using 3 fully-connected hidden layers, discount factor=0.99, update target frequency every 500 episodes, memory size=5000, batch size=64 and *Huber loss* as loss function. The number of neurons in 1-3 hidden layers was 4×, 3× and 2× number of actions, respectively.

### *B. Performance Evaluation*

The algorithm performance is verified by conducting Monte-Carlo simulations for different user positioning realizations within the network area. In each positioning scenario, the users are randomly placed within the network area. We define the performance of each simulation setting as the ratio between the total allocated throughput and the total requested throughput (i.e. the optimal solution) for $N_i = 10^3$ different user positioning scenarios. Formally, the performance metric for a simulation setting with $U$ users is given by:

$$P_U(\%) = \frac{\sum_{i=1}^{N_i}\sum_{u=1}^{U} R_u}{\sum_{i=1}^{N_i}\sum_{u=1}^{U} D_u} \times 100, \quad (9)$$

where $R_u$ is the allocated throughput to the user $u$ and $D_u$ is the corresponding throughput demand.

In the first part of the evaluation, we investigate the performance of the proposed scheme as (i) a function of the number of users and (ii) the requested service becomes more stringent. To that end, the DDPA algorithm is verified for 10, 20, 30, 40 and 50 users and for varying demand vector, namely all users request QoS (i) class 1, (ii) class 2 and (iii) class 3.

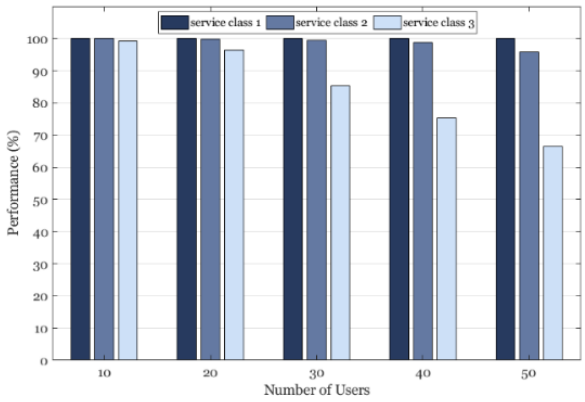


Fig. 3. DDPA performance in cases that all users request service 1, 2 and 3, respectively, as a function of the number of users

Evidently, the results shown in Fig. 4 indicate that the DDPA algorithm achieves 96-100% performance when the demand vectors are 0.1 and 1 Mbps, regardless of the number of users. As expected, the performance ratio gradually deteriorates in the case of the strictest demand vector (2.5 Mbps). Although the ratio remains above 85% for 10, 20 and 30 users, the results of Fig.4 imply the need for additional channel capacity in cases of increased total requested throughput (substantial number of users, very challenging demands, etc.).

In the second set of simulations, we consider five different realistic scenarios with increasing complexity, as the number of active users ranges from 10 to 50. Specifically, the configuration of the 5 scenarios included: **(i)** 10 users requesting 17.5 Mbps, **(ii)** 20 users requesting 30.5 Mbps, **(iii)** 30 users requesting 40.5 Mbps, **(iv)** 40 users requesting 46 Mbps and **(v)** 50 users requesting 63.5 Mbps. The resulting performance metrics for all scenarios, along with the total allocated throughput (in Mbps) are demonstrated in Fig. 5. For comparison purposes, Fig. 5 also depicts the performance metric resulted from two baseline methods, namely the Weighted Minimum Mean Square Error (WMMSE) algorithm and a fixed power allocation policy (Average Power), according to which each RU/channel transmits with the average/median power level as a reasonable trade-off between the achieved coverage and the potential interferences. WMMSE algorithm is implemented as suggested in [5] and [12], with the objective of maximizing the weighted sum-rate across users.

As readily observed in Fig. 6, the three methods exhibit near-optimal outcomes in the first scenario (10 users). However, as the simulation scenario becomes more challenging with respect to the number of users and their requested throughput, the benefits of the developed DDPA algorithm against the two baseline methods become more noticeable. Evidently, all methods show above 93% performance for the simple scenarios (1 & 2), whereas for scenarios 3, 4 and 5 a significantly improved performance is achieved for the DDPA case. Specifically, the DQN based algorithm allocates (39.54, 44.21, 54.9) Mbps out of the total requested throughput of (40.5, 46, 63.5) Mbps in scenarios 3, 4 and 5, respectively (Fig. 6). The average performance of DDPA across the five scenarios was equal to 95.6%, while the same for metric for WMMSE was 87.9, resulting to a performance increase of approximately 9%.

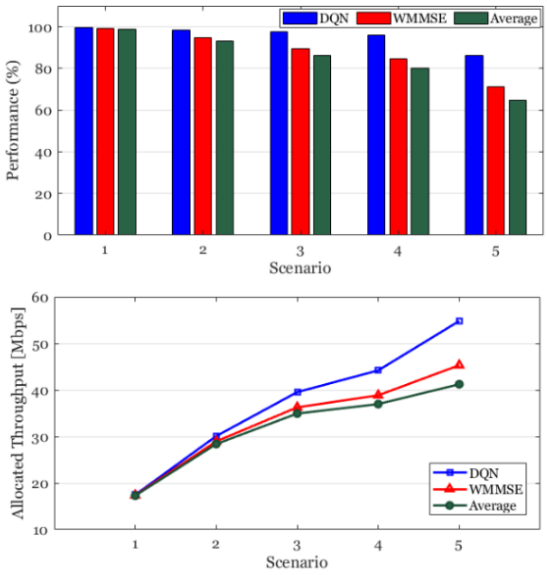


Fig. 4. Comparison among methods: DDPA performance against *WMMSE* and *Average* methods for five different simulation scenarios (up) and total allocated throughput for the three methods (down).

The results clearly indicate that the DDPA algorithm outperforms the two baseline methods in terms of total allocated throughput, especially for increasing number of users and/or more strict demands. Moreover, it should be noted that the pre-trained DQL models can be effortlessly inferred during the online network operation in order to provide accurate estimates of the most beneficial power configuration (real-time decision-making). To enhance the efficiency of these models, offline training can also be carried out without affecting the real-time network performance.

## IV. Conclusions

In the present study, a DRL approach for power adjustment is proposed, which efficiently adjusts the transmit power of RUs following a demand-driven strategy. In contrast to several algorithms attempting to maximize the total network throughput, DDPA uses an alternative state space and rewarding system definition to, finally, proactively allocate power levels to RUs. Evaluation results show enhanced performance of the algorithm, even in extreme demand scenarios, as compared to other baseline methods.

## V. 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/).

## VI. References

[1] P. Trakadas, et al., "Hybrid clouds for data-Intensive, 5G-Enabled IoT applications: an overview, key issues and relevant architecture", *Sensors*, 19 (16), pp. 3591, 2019.

[2] F. D. Calabrese, L. Wang, E. Ghadimi, G. Peters, L. Hanzo and P. Soldati, "Learning Radio Resource Management in RANs: Framework, Opportunities, and Challenges," in *IEEE Communications Magazine*, vol. 56, no. 9, pp. 138-145, Sept. 2018, doi: 10.1109/MCOM.2018.1701031.

[3] M. E. Morocho-Cayamcela, H. Lee and W. Lim, "Machine Learning for 5G/B5G Mobile and Wireless Communications: Potential, Limitations, and Future Directions," in *IEEE Access*, vol. 7, pp. 137184-137206, 2019, doi: 10.1109/ACCESS.2019.2942390.

[4] Q. Qi, A. Minturn and Y. Yang, "An efficient water-filling algorithm for power allocation in OFDM-based cognitive radio systems," *2012 International Conference on Systems and Informatics (ICSAI2012)*, Yantai, 2012, pp. 2069-2073, doi: 10.1109/ICSAI.2012.6223460.

[5] Q. Shi, M. Razaviyayn, Z. Luo and C. He, "An Iteratively Weighted MMSE Approach to Distributed Sum-Utility Maximization for a MIMO Interfering Broadcast Channel," in *IEEE Transactions on Signal Processing*, vol. 59, no. 9, pp. 4331-4340, Sept. 2011, doi: 10.1109/TSP.2011.2147784.

[6] C. Zhang, P. Patras and H. Haddadi, "Deep Learning in Mobile and Wireless Networking: A Survey," in *IEEE Communications Surveys & Tutorials*, vol. 21, no. 3, pp. 2224-2287, thirdquarter 2019, doi: 10.1109/COMST.2019.2904897.

[7] K. I. Ahmed, H. Tabassum and E. Hossain, "Deep Learning for Radio Resource Allocation in Multi-Cell Networks," *in IEEE Network*, vol. 33, no. 6, pp. 188-195, Nov.-Dec. 2019, doi: 10.1109/MNET.2019.1900029.

[8] N. Naderializadeh, J. Sydir, M. Simsek and H. Nikopour, "Resource management in wireless networks via multi-agent deep reinforcement learning," *in IEEE Transactions on Wireless Communications*, 2021.

[9] N. Zhao, Y. Liang, D. Niyato, Y. Pei, M. Wu and Y. Jiang, "Deep Reinforcement Learning for User Association and Resource Allocation in Heterogeneous Cellular Networks," *in IEEE Transactions on Wireless Communications*, vol. 18, no. 11, pp. 5141-5152, Nov. 2019, doi: 10.1109/TWC.2019.2933417.

[10] 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," *2017 IEEE International Conference on Communications (ICC)*, Paris, 2017, pp. 1-6, doi: 10.1109/ICC.2017.7997286.

[11] R. Amiri, H. Mehrpouyan, L. Fridman, R. K. Mallik, A. Nallanathan and D. Matolak, "A Machine Learning Approach for Power Allocation in HetNets Considering QoS," *2018 IEEE International Conference on Communications (ICC)*, Kansas City, MO, 2018, pp. 1-7, doi: 10.1109/ICC.2018.8422864.

[12] M. Zhang and M. Chen, "Power Allocation in Multi-cell System Using Distributed Deep Neural Network Algorithm," *2019 International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob)*, Barcelona, Spain, 2019, pp. 1-4, doi: 10.1109/WiMOB.2019.8923201.

[13] 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," *in IEEE Access*, vol. 7, pp. 177254-177265, 2019, doi: 10.1109/ACCESS.2019.2937438.

[14] ETSI TS 136 213 *LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures*, V14.2.0 (2017-04).