A Decoupled Access Scheme with Reinforcement Learning Power Control for Cellular-Enabled UAVs

DOI:
10.1109/JIOT.2021.3078188

Document Version
Accepted author manuscript

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3078188

A Decoupled Access Scheme with Reinforcement Learning Power Control for Cellular-Enabled UAVs

Yao Shi, Mutasem Q. Hamdan, *Student Member, IEEE*, Emad Alsusa, *Senior Member, IEEE*, Khairi A. Hamdi, *Senior Member, IEEE*, and Mohammed W. Baidas, *Senior Member, IEEE*

***Abstract*—This paper proposes a downlink/uplink decoupled (DUDe) access scheme for cellular-enabled unmanned aerial vehicle (UAV) communication systems. To minimise interference, the proposed scheme separates the control and data links of UAVs, as well as the uplinks (ULs) and downlinks (DLs) of ground users (GUEs), onto different serving base-stations and operating frequencies. Since power availability is a major constraint in UAV communications, two power allocation schemes based on Q-learning (QL) and deep Q-learning (DQL) are proposed to optimize the communication energy-efficiency (EE) of this DUDe network. To quantify the improvements achieved, the proposed schemes are compared with the fractional power control (FPC) scheme used in 4G and 5G networks, as well as, a convex optimization based optimal power allocation scheme. The results demonstrate that the proposed DUDe scheme can achieve up to several times higher sum-rates and EE in the UL direction than its coupled counterparts. Moreover, it is shown that the EE performance of the QL and DQL power allocation schemes approach the optimal performance and surpass the conventional FPC scheme by $80\% - 100\%$ in the UHF band, and by $160\% - 170\%$ in the mmWave band.**

***Index Terms*—Cellular-enabled UAV communication, downlink and uplink decoupling, Q-learning, deep Q-learning, millimeter-wave communications**

I. Introduction

The worldwide market for unmanned aerial vehicles (UAVs) has rapidly expanded over the past decade, with UAV-based services becoming a driver of financial developments and opportunities for the telecom operators [1]. UAVs can be aerial users to assist with aerial mapping, disaster rescue, agricultural irrigation, etc. [2]–[4], or, flying base-stations (BSs) to boost throughput, coverage, and quality-of-service (QoS) of cellular networks [5]–[7]. Currently, most UAVs in the market rely on direct point-to-point communication with ground control stations (GCSs) or ground pilots, and transmit over the unlicensed spectrum, such as Wi-Fi [8]. Despite the fact that the unlicensed spectrum is free and can fulfill some applications (e.g. visual line-of-sight (LOS) aerial photography), it offers unstable data rates, and is vulnerable to interference. The rapidly growing number of UAVs and increasing communication requirements for UAV applications call for a more reliable and effective communication system. A promising solution is to utilize cellular networks to support UAV communications, leading to cellular-enabled UAV communications [9,10].

LTE-enabled UAV communications have been approved by 3GPP [11], with their feasibility demonstrated in [1,12]; however, the interference between UAVs and ground user equipments (GUEs)/ground base-stations (GBSs) is still a major concern. This is because most of the UAV-GUE/GBSs links are LOS, and, UAVs usually have omnidirectional antennas, which causes more interference to GUEs. Most of the current research solves this problem by power control, resource block (RB) allocation, and/or trajectory design [13]–[17]. Although some of the existing solutions are efficient and robust, in this paper, this problem is solved from a new perspective by eliminating interference via effective decoupled base-station (BS) association and spectral resource allocation.

In heterogeneous networks (HetNets), the transmit power of macro BSs (MBSs) is much higher than that of small BSs (SBSs), and thus, more UEs are connected to the MBSs, since the MBSs can provide higher biased reference signal received power (RSRP) in the downlink (DL) [18]. However, in the uplink (UL), the received signal strength depends on the UE-BS distance, and the MBS's edge UEs may suffer from poor QoS which deteriorates with higher operating frequencies. Under these circumstances, downlink/uplink decoupled (DUDe) access was initially proposed in [19] to allow UEs to connect to different BSs in the UL and DL, as opposed to 1G-4G networks where the UEs are associated with the same BS in both link directions. According to DUDe, the UEs can connect to the nearest BS in the UL rather than the same DL serving BS. The first DUDe access scheme—namely minimum path-loss (min-PL)—was proposed in [19], where the UEs are connected to the BS with the lowest path-loss in the UL. The UL performance improvement brought by DUDe is investigated in [20], where it is shown that DUDe can improve the load balancing, especially for ultra-dense networks. Inspired by such findings, and in comparison with existing works, this paper decouples the links from the perspective of serving BSs and the operating frequency bands in the downlink and uplink. By utilizing different frequency bands for the LOS and non-LOS (NLOS) links, and also for the data links and control links, the interference between UAVs and GUEs can be eliminated from the very beginning.

Energy-efficiency (EE) optimization is essential for UAV communications, as efficient energy utilization can prolong the operation of UAV applications for the network stake-

Yao Shi, Mutasem Q. Hamdan (corresponding author), Emad Alsusa and Khairi A. Hamdi are with the School of Electrical and Electronic Engineering, University of Manchester, Manchester M1 3WE, UK (e-mail: {yao.shi, mutasem.hamdan, e.alsusa, k.hamdi}@manchester.ac.uk).

Mohammed W. Baidas is with the Electrical Engineering Department, College of Engineering and Petroleum, Kuwait University, Kuwait (e-mail: m.baidas@ku.edu.kw).

This work was partially supported by the Kuwait Foundation for the Advancement of Sciences (KFAS) under project code PN17-15EE-02.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3078188

holders, ultimately improving users experience (e.g. in processing sensory data and flight times for the UAVs) [21]. Due to the rapid changes in the UAV wireless environment—such as air-to-ground channels, spatial and time variations of non-stationary signal behaviour, and detection of UAVs via UAVs-enabled protocols—conventional optimization methods with ideal assumptions (e.g. perfect CSI) may not work in practical and real-time applications. Hence, it is necessary to augment classical algorithms and solutions with artificial intelligence (AI) and machine learning (ML)-based techniques [22]. Moreover, when both UAVs and GUEs transmission resources are optimized, the network access schemes can be designed to overcome the classical methods' excessive overhead and delays, while incorporating ML techniques to achieve an acceptable/sub-optimal EE solution in rapidly changing wireless environments.

Reinforcement learning (RL) algorithms are among the most promising ML techniques to use in radio resource management (RRM) for UAV-enabled cellular communications [23]. This is due to the nature of RL, which is based on maximizing a reward function by exploring the action(s) domain—via trial-and-error interactions—to allow the learner to discover the best choices based on the received rewards [24]. In turn, RL has become a base for resource allocation in wireless networks, due to its simplicity and ability to provide reliable and efficient learning through interaction with the network. Q-learning (QL) is a model-free RL approach, which is based on finite states and actions to obtain acceptable/near-optimal solutions with low computational-complexity [25]. However, in QL, the sizes of the state and action spaces grow exponentially for each additional unknown network feature and/or parameter, leading to the curse of dimensionality, especially in the training phase. Alternatively, deep Q-learning (DQL) has been proposed, which utilizes a deep neural network (DNN), called deep Q-network (DQN) along with other techniques (e.g. replay memory) to perform a stable and efficient training, and reliably estimate the Q-function [26]. Particularly, DQL is based on quickly performing predictions using only a small number of simple operations to obtain an output, which greatly reduces execution time. Consequently, deep RL approaches have found numerous applications in cellular networks [27]–[29]. Add to this the 3GPP technical requirements for the enhancement of UAVs [30], which only proposes AI/ML to control the UAVs, but did not discuss how AI/ML can be used in the scheduling and resource allocation. This motivates us to investigate the potentials of AI/ML techniques in cellular-enabled UAVs by applying QL and DQL as resource allocation tools.

A. Related Works

Recently, a number of research works have proposed learning-based resource allocation for cellular-enabled UAV networks [28]. For instance, in [31], an interference management scheme is proposed with the aim of achieving a tradeoff between maximizing EE and minimizing wireless latency and interference to the ground network. Specifically, a DQL algorithm based on echo state network (ESN) cells is devised to allow each UAV to map each observation of the network state to an action, and hence learn its optimal path, transmit power and cell association. The proposed algorithm has been shown to minimize the interference to the GUEs and the transmission delay of the UAVs. A 3D energy-efficient and fair UAV scheduling scheme based on deep RL (DRL) is proposed in [32] to allow the UAVs to hover around and serve the users, and also recharge their batteries. The proposed algorithm has been shown to outperform existing scheduling algorithms in terms of coverage, energy-efficiency and fairness. In [33], a novel DRL-based control algorithm is devised for energy-efficient coverage and connectivity, which is demonstrated to outperform baseline schemes in terms of coverage, fairness and energy consumption. In [34], a DRL-based channel and power allocation scheme for UAV-enabled IoT systems is proposed. Particularly, the UAV-BS is able to schedule channels and allocate transmit power for uplink transmissions to maximize the minimum energy-efficiency among all the IoT nodes, yielding superior performance over state-of-the-art schemes. In [35], a DRL-based approach for distributed energy-efficient multi-UAV navigation has been proposed to ensure long-term communication coverage, while optimizing geographical fairness, UAV energy consumption and connectivity. A Q-learning algorithm has been designed in [36] for the optimal positioning of UAV small cells, with the aim of maximizing the network lifetime. In [37] and [38], the proposed QL- and DRL-based methods have been applied to UAV-BSs with energy constraints to achieve energy-efficiency and coverage fairness to the GUEs, while reducing the collision incidents and co-channel interference (CCI). The majority of researchers have considered UAVs as BSs, rather than UEs and to the best of our knowledge, none of the prior works in the literature have considered decoupled access in cellular-enabled UAVs with RL-based power control for energy-efficiency maximization, except our previous work in [39], where we solved the formulated problems using conventional convex optimization methods. However, in this paper, the formulated problems are solved using the proposed QL and DQL algorithms, which are model-free, and can handle none convex problems with stochastic transitions. Moreover, we show that the performance of our proposed RL-based algorithms approach that of the upper-bound solution obtained in [39].

B. Main Contributions

The main contributions of this paper can be summarized as follows:

- A DUDe access scheme is proposed for UAVs and GUEs, in which the serving BSs and operating frequency bands of UAV data links and control links, as well as GUE ULs and DLs are decoupled.
- A novel and simple QL algorithm is proposed for EE-maximizing power control, while alleviating the excessive computational delays of the classical fractional programming and successive convex approximation solutions. This algorithm has outperformed the benchmark schemes in terms of EE.

- A novel DQL algorithm is proposed to optimize the EE and overcome the large state-action matrix in the QL algorithm. Although the DQL performance is slightly worse than the QL, it outperforms the conventional fractional power control (FPC) scheme.
- The performance of the proposed DUDe QL and DQL power control schemes are compared with state-of-art alternatives in terms of EE, sum-rate and data rate per GUE/UAV. It is demonstrated that the proposed DUDe can achieve several times higher sum-rates and EE than the coupled benchmark counterparts. The QL (DQL) power control scheme improves EE by around 80% (100%) for the UHF band, and by around 160% (170%) for the mmWave band, in comparison to conventional FPC scheme.

C. Organization

The rest of this paper is organized as follows. Section II describes the system model, while Section III introduces the DUDe cellular-enabled UAV communication scheme. Section IV formulates the GUEs and UAVs EE maximization problems based on the DUDe access scheme, with constraints on the maximum transmit power and minimum data rate per GUE/UAV. Section V outlines the QL and DQL algorithms for EE maximization, while Section VI discusses their implementation. Section VII evaluates the performance of the proposed QL and DQL power control schemes, and compares them with several benchmarks. Finally, Section VIII draws the conclusions.

II. System Model

A. Network Model

An OFDMA HetNet consisting of MBSs, SBSs, UAVs and GUEs is considered, which are deployed uniformly with densities of λ_m, λ_s and λ_g, respectively. The horizontal positions and heights of the UAVs also follow uniform distribution with intensity λ_u. The full-buffer UE traffic model is assumed in this paper[1].

B. Propagation Model

The path-loss $L(d)$ is given by

$$L(d) = 20\log\left(\frac{4\pi d_0 f}{c}\right) + 10\phi\log\left(\frac{d}{d_0}\right) + \chi, \tag{1}$$

where d_0 refers to close-in reference distance, f denotes the operating frequency, c represents the speed of light, d is the GUE/UAV-BS distance, ϕ is the path-loss exponent, and χ is the log-normal shadowing. The blockage models for the mmWave links of GUEs and UAVs are different. For GUEs, the generalized blockage ball model in [40] is adopted, which is widely accepted in many studies [41,42]. Particularly, if the distance between a GUE/UAV and its serving BS is less than μ = 200 m, this link is assumed to be LOS with probability ω = 0.2; otherwise, this link is assumed to be NLOS. For UAV communications, the blockage model in [43] is utilized, where the LOS probability is given by

$$\Pr^A(LOS, \theta) = \frac{1}{1 + \exp(-b(\theta - a))}, \tag{2}$$

in which θ is the elevation angle of the UAV at the BS antenna, and a and b are S-curve parameters related to the environment.

C. Antenna Elements

In this work, it is assumed that the BSs, UAVs and GUEs support both mmWave and ultra-high frequency (UHF) bands. Specifically, each BS, UAV and GUE is assumed to have one UHF omnidirectional antenna, while uniform planar square arrays (UPA) with half-wavelength antenna element spacing are utilized for mmWave transmissions [44]–[46]. The mmWave BS antenna gain is modeled as

$$G_b(\Theta) = \begin{cases} G_M, & |\Theta| \leq \Theta_b/2, \\ G_m, & \text{otherwise}, \end{cases} \tag{3}$$

where Θ_b is the mainlobe beamwidth, G_M is the mainlobe gain, and G_m is the sidelobe gain. The mmWave GUE/UAV are assumed to be in perfect alignment with their serving BSs [46]. Table I summarizes the antenna parameters.

TABLE I
Antenna Parameters

Frequency Band	UHF	mmWave	mmWave
Number of antenna elements	1	4	16
Half-power beamwidth (degree) Θ_b	360	49.6	24.8
Main-lobe gain (dBi) G_M	0	6	12
Side-lobe gain (dBi) G_m	0	-0.8839	-1.1092

D. Transmission Rate

The greedy RB allocation algorithm in [47] is adopted, where each RB has bandwidth B. For each GUE/UAV, all the RBs that are not yet assigned are sorted according to their corresponding signal-to-interference-plus-noise (SINR) values, and those with high SINR are preferentially assigned to the GUE/UAV. The transmission rate between transmitter $i \in \mathcal{I} = \{1, 2, ..., I\}$ and its receiver $m \in \mathcal{M} = \{0, 1, 2, ..., M\}$ on RB $n \in \mathcal{N} = \{1, 2, ..., N\}$ is expressed as[2]

$$\mathbb{R}_{i,m,n} = B\log_2\left(1 + \frac{P_{i,n}G_{i,m}|h_{i,m,n}|^2}{\sum\limits_{j\in\mathcal{I}, j\neq i} P_{j,n}G_{j,m}|h_{j,m,n}|^2 + \sigma^2}\right), \tag{4}$$

where $G_{i,m}$ is the antenna gain. Moreover, $\sigma^2 = N_0 B$ is the variance of the AWGN, where N_0 is the noise spectral density. Also, $P_{i,n}$ is the transmit power of transmitter i on RB n, and $|h_{i,m,n}|^2$ is the channel gain between transmitter i and receiver m on RB n, which includes both fading and path-loss. The fading models of the mmWave and UHF communications are Nakagami$-m$ fading [48], and Rayleigh fading [49], respectively.

[1]It is noteworthy that UE is used to collectively refer to a GUE or a UAV.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3078188

TABLE II
TRANSMISSION PARAMETERS OF THE DIFFERENT SCHEMES

Scheme	DUDe				Coupled (UHF)		Coupled (mmWave)	
Direction	DL		UL		DL	UL	DL	UL
Frequency Band	mmWave	UHF	mmWave	UHF	UHF		mmWave	
Bandwidth	4.8 MHz	1.2 MHz	4.8 MHz	1.2 MHz	1.2 MHz		4.8 MHz	
UE	GUE	GUE	UAV	GUE	GUE	UAV+GUE	GUE	UAV+GUE
Cell Association	Biased RSRP		Min-PL		Biased RSRP		Biased RSRP	

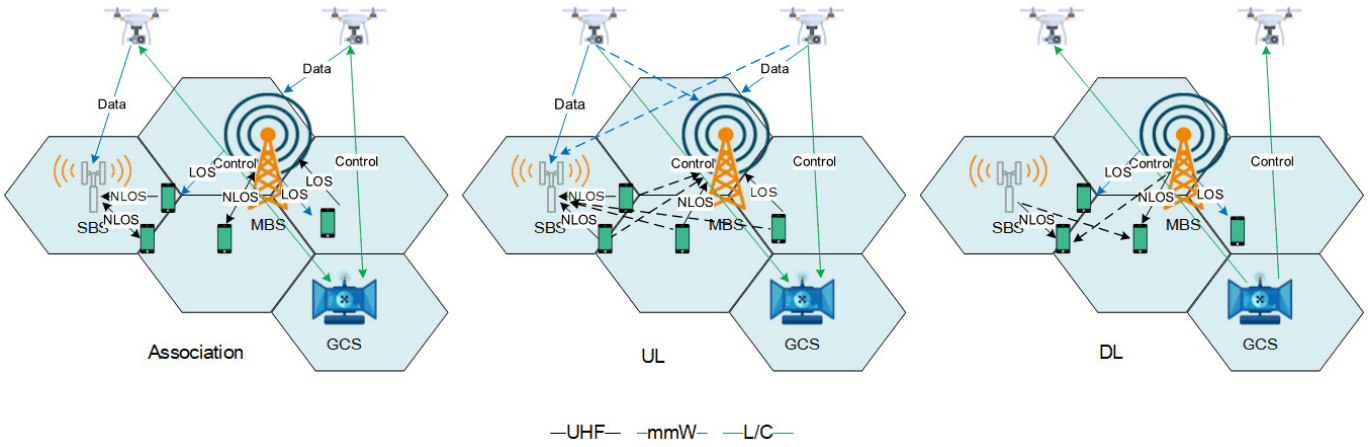

Fig. 1. Cell Association and Interference map of the proposed scheme.

III. DUDE ACCESS FOR UAVS AND GUES

The requirements for UAV data links and control links are quite different. Particularly, UAV data links require high data rates that can be of the order of hundreds of Mbps. On the contrary, UAV control and non-payload communication (CNPC) links are of low data rate [1], but require low-latency, and ultra-reliability, which is difficult to guarantee in cellular networks due to interference [50]. This motivates the splitting of the UAV data links from the control links. Likewise, the GUE ULs and DLs are imbalanced. Also, there is a significant gap between the the BS UL and DL coverage, and GUEs require much higher data rate in DL than in UL. Thus, transmitting over high-frequency bands in the DL and low-frequency bands in the UL is intuitive. Additionally, transmitting on a dedicated band helps eliminate interference, and guarantee the reliability of control links.

Taking the above two aspects into consideration, a DUDe access scheme for cellular-enabled UAV communications is proposed. Specifically, GCSs operating on the L/C bands are deployed to support CNPC links to avoid excessive switching of serving BSs during the flight, and provide wider coverage. In particular, approximately 17 MHz (960-977 MHz) in the L-band and 61 MHz (5.03-5.091 GHz) in the C-band are presently allocated for UAV CNPC links [51]. On the other hand, UAV data links are mainly UL, LOS-dominated, and require high data rates. In this case, mmWave communication is particularly suitable. Although some may question the practicability of mmWave-enabled UAV communication, the challenges and solutions have been well addressed in [52,53]. Furthermore, to prevent the interference between UAVs and GUEs in the UL, UHF bands are utilized for the GUE UL transmission, and the min-PL scheme is utilized for BS association, such that the UL coverage can be guaranteed and the GUE/UAV-BS distances are shortened. As for GUE DL transmission, UHF bands are utilized to provide umbrella coverage, while mmWave bands are applied to improve the data rate over LOS links. The biased RSRP scheme is applied for DL BS access. Moreover, time-division duplexing (TDD) and static consistent DL/UL configuration are considered in this paper, such that cross-link interference is avoided [54]. For both mmWave and UHF band communications, 50% subframes/symbols are used for UL transmission, and the remaining 50% subframes/symbols are used for DL transmission, while special subframes are neglected[3]. The whole band allocation and cell association strategy of the proposed DUDe scheme is given in Table II. Moreover, Fig. 1 illustrates the cell association and interference map of the proposed scheme, where the dotted lines refer to interference.

IV. ENERGY-EFFICIENCY MAXIMIZATION

As the battery capacities of UAVs and GUEs are limited, the aim of this paper is to optimize their EE. In order to solve this problem, the network interference is analyzed first. As shown in Fig. 1, in the DL, NLOS GUEs transmit over the UHF band. Thus, the GUEs in different cells cause inter-cell interference (ICI) to each other, leading to an interference-limited scenario. Also, LOS GUEs transmitting over the mmWave band suffer from interference; however, since mmWave signals are sensitive to blockage, and beamforming is applied, the interference

[2]As both UL and DL are considered in this paper, a transmitter can be a GUE/UAV or BS.

[3]In LTE TDD, if a subframe is configured for DL (or UL), all of the symbols within the subframe should be used as DL (or UL). However, in 5G NR, the symbols within a slot can be configured in various ways [55].

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3078188

level becomes negligible (i.e. a noise-limited scenario [56,57]). In the UL, all GUEs transmit over the UHF band, and suffer from ICI, while all UAVs transmit over the mmWave band, and the interference can be ignored. Due to the application of decoupling in terms of operating bands and associated BSs, the interference between UAVs and GUEs is greatly reduced, which is very different from the scenario in most existing research. The communication EE of UAVs and GUEs in the UL is optimized in this paper, and the DL EE can be obtained analogously.

A. Optimal GUEs EE Maximization

The data rate of GUE i to its serving BS m over RB n is expressed as

$$\mathbb{R}^G_{i,m,n}\left(\mathbf{P}\right) = B\log_2\left(1+\gamma^G_{i,n}\right), \tag{5}$$

in which $\gamma^G_{i,n} = \frac{P_{i,n}|h_{i,m,n}|^2}{\sum\limits_{j\in\mathcal{I}^G, j\neq i} P_{j,n}|h_{j,m,n}|^2+\sigma^2}$ is the received SINR, while $\mathcal{I}^G$ is the GUEs index set over the UHF band. Also, $\mathbf{P}$ is the power allocation matrix of all GUEs over all RBs. In turn, the data rate of GUE i over its allocated RBs is

$$\mathbb{R}^G_i\left(\mathbf{P}\right) = \sum_{n\in\mathcal{N}_{UHF}} \mathbb{R}^G_{i,m,n}\left(\mathbf{P}\right), \tag{6}$$

while its total transmit power consumption is

$$\mathbb{P}^G_i\left(\mathbf{P}\right) = \sum_{n\in\mathcal{N}_{UHF}} P_{i,n}, \tag{7}$$

where $\mathcal{N}_{UHF}$ is the index set of UHF RBs. In turn, the GUEs EE maximization (GUEs-EE-MAX) problem is formulated as

GUEs-EE-MAX: (8)

$$\max_{\mathbf{P}} \quad \mathbb{EE}^G\left(\mathbf{P}\right) = \frac{\sum_{i\in\mathcal{I}^G}\mathbb{R}^G_i\left(\mathbf{P}\right)}{\sum_{i\in\mathcal{I}^G}\mathbb{P}^G_i\left(\mathbf{P}\right)} \tag{8a}$$

$$\text{s.t.} \quad \mathbb{R}^G_i\left(\mathbf{P}\right) \geq R^G_{\min}, \forall i\in\mathcal{I}^G, \tag{8b}$$

$$\mathbb{P}^G_i\left(\mathbf{P}\right) \leq P^G_{\max}, \forall i\in\mathcal{I}^G, \tag{8c}$$

$$P_{i,n}\geq 0, \qquad \forall i\in\mathcal{I}^G, \forall n\in\mathcal{N}_{UHF}. \tag{8d}$$

where $P^G_{\max}$ and $R^G_{\min}$ are the maximum transmit power and minimum data rate of each GUE, respectively. In problem **GUEs-EE-MAX**, (8a) is the objective function, while Constraint (8b) guarantees that the total rate of each GUE is higher than the minimum data rate. Constraint (8c) ensures the transmit power of each GUE does not exceed the maximum transmit power, while Constraint (8d) enforces the non-negativity of the transmit power of each GUE.

Remark 1: The rate function of each GUE is non-convex because of the interference terms, and thus the objective function $\mathbb{EE}^G\left(\mathbf{P}\right)$ is non-convex either [58]. Also, the constraints set is non-convex due to the minimum rate constraint.

To solve problem **GUEs-EE-MAX**, an efficient power allocation scheme is proposed in our previous work [39], which utilizes a lower-bound rate approximation, convex variable substitution, Dinkelbach's inner-layer algorithm [59], and an iterative outer-layer algorithm to efficiently obtain the global optimal EE solution. The overall complexity of the proposed power allocation scheme is $\mathcal{O}\left(\left(\frac{1}{\epsilon^2}\log(|\mathcal{I}^G|)\right)^2\right)$, where ϵ is the error tolerance (i.e. stopping criterion).

B. Optimal UAVs EE Maximization

Different from GUE communications, UAV communications is noise-limited. The transmission rate between UAV i and its serving BS m over RB n is denoted by

$$\mathbb{R}^A_{i,m,n}\left(P_{i,n}\right) = B\log_2\left(1+\gamma^A_{i,n}\right), \tag{9}$$

in which $\gamma^A_{i,n} = \frac{P_{i,n}G_{i,m}|h_{i,m,n}|^2}{\sigma^2}$ is the received SINR, and $\mathcal{I}^A$ is the UAVs index set. The data rate of UAV i on all its allocated RBs is expressed as

$$\mathbb{R}^A_i\left(\mathbf{P}\right) = \sum_{n\in\mathcal{N}_{mmW}} \mathbb{R}^A_{i,m,n}\left(P_{i,n}\right), \tag{10}$$

where $\mathcal{N}_{mmW}$ is the index set of mmWave RBs.

Remark 2: The sum-rate function $\mathbb{R}^A_i\left(\mathbf{P}\right)$ can be verified to be concave in $\mathbf{P}$, since the data rate function $\mathbb{R}^A_{i,n}\left(P_{i,n}\right)$ is concave in $P_{i,n}$ [60].

The total transmit power consumption of UAV i is

$$\mathbb{P}^A_i\left(\mathbf{P}\right) = \sum_{n\in\mathcal{N}_{mmW}} P_{i,n}. \tag{11}$$

The UAVs EE maximization (UAVs-EE-MAX) problem is written as

UAVs-EE-MAX: (12)

$$\max_{\mathbf{P}} \quad \mathbb{EE}^A\left(\mathbf{P}\right) = \frac{\sum_{i\in\mathcal{I}^A}\mathbb{R}^A_i\left(\mathbf{P}\right)}{\sum_{i\in\mathcal{I}^A}\mathbb{P}^A_i\left(\mathbf{P}\right)} \tag{12a}$$

$$\text{s.t.} \quad \mathbb{R}^A_i\left(\mathbf{P}\right) \geq R^A_{\min}, \forall i\in\mathcal{I}^A, \tag{12b}$$

$$\mathbb{P}^A_i\left(\mathbf{P}\right) \leq P^A_{\max}, \forall i\in\mathcal{I}^A, \tag{12c}$$

$$P_{i,n}\geq 0, \qquad \forall i\in\mathcal{I}^G, \forall n\in\mathcal{N}_{mmW}. \tag{12d}$$

where $P^A_{\max}$ and $R^A_{\min}$ are the maximum transmit power and minimum data rate of each UAV, respectively.

Remark 3: The $\mathbb{EE}^A\left(\mathbf{P}\right)$ function is a ratio of a concave function $\sum_{i\in\mathcal{I}^A}\mathbb{R}^A_i\left(\mathbf{P}\right)$ to a linear function $\sum_{i\in\mathcal{I}^A}\mathbb{P}^A_i\left(\mathbf{P}\right)$ in $\mathbf{P}$.

Problem **UAVs-EE-MAX** is globally optimized in our previous work [39] via convex variable substitution and Dinkelbach's algorithm, with overall computational-complexity of $\mathcal{O}\left(\frac{1}{\epsilon^2}\log(|\mathcal{I}^A|)\right)$.

The **GUEs-EE-MAX** and **UAVa-EE-MAX** power allocation schemes are utilized in this paper as optimal benchmarks. Another benchmark is the FPC scheme adopted in 4G and 5G cellular networks [61,62], where the transmit power (in dBm) of a GUE/UAV is given by

$$P_{i,n} = \frac{1}{N'}\min\{P_{\max}, 10\log_{10}N' + wL + P_0\}, \tag{13}$$

where $P_{\max}$ is the GUE's/UAV's maximum transmit power, N' is the number of allocated RBs, $w \in \{0, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1\}$ is the compensation factor for path-loss L and P_0 is the target received power.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3078188

V. RL-Based Optimization of Energy-Efficiency

A. Q-Learning (QL)

In this work, the power control is centralized, and the QL agent is assumed to be located at the MBS, where the learning process is modeled as a Markov decision process (MDP). Now, let $\mathcal{S}$ be the set of possible transmit power states over the assigned RBs, and $\mathcal{A}(\mathbf{s})$ be the discrete set of actions in terms of the transmit powers over the assigned RBs in state $\mathbf{s}$. Assuming discrete time-steps t resembling the training rounds, the QL agent takes $\mathbf{a}^{(t)} \in \mathcal{A}\left(\mathbf{s}^{(t)}\right)$ based on some policy ϕ. Particularly, $\phi\left(\mathbf{s},\mathbf{a}\right)$ represents the probability of taking action vector $\mathbf{a}$ in state $\mathbf{s}$[4]. By applying $\mathbf{a}^{(t)} \in \mathcal{A}\left(\mathbf{s}^{(t)}\right)$ and transitioning from state $\mathbf{s}^{(t)}$ to $\mathbf{s}^{(t+1)}$, a reward $r^{(t+1)} \triangleq \left(\mathbf{s}^{(t)},\mathbf{a}^{(t)}\right)$ is given to characterize the benefit from taking action vector $\mathbf{a}^{(t)}$ in state $\mathbf{s}^{(t)}$. The well-known QL algorithm aims to find the optimal policy ϕ^* that maximizes an expected reward function. Thus, let the future cumulative discounted reward at time-step t be given by [63]

$$\mathcal{R}^{(t)} = \sum_{\tau=0}^{\infty} \delta^{\tau} r^{(t+\tau+1)}, \tag{14}$$

where $\delta \in [0,1)$ is the discount factor for future rewards. Also, let the Q-function associated with policy ϕ as the expected reward when $\mathbf{a}$ is taken in state $\mathbf{s}$, as

$$Q_{\phi}(\mathbf{s},\mathbf{a}) = \mathbb{E}\left[\mathcal{R}^{(t)} | \mathbf{s}^{(t)} = \mathbf{s}, \mathbf{a}^{(t)} = \mathbf{a}\right], \tag{15}$$

which satisfies the Bellman optimality equation as [64]

$$Q_{\phi}(\mathbf{s},\mathbf{a}) = \mathcal{R}(\mathbf{s},\mathbf{a}) + \delta \sum_{\mathbf{s}' \in \mathcal{S}} \mathcal{P}^{\mathbf{a}}_{\mathbf{s},\mathbf{s}'} \left(\sum_{\mathbf{a}' \in \mathcal{A}(\mathbf{s}')} \phi\left(\mathbf{s}',\mathbf{a}'\right) Q_{\phi}\left(\mathbf{s}',\mathbf{a}'\right) \right), \tag{16}$$

with $\mathcal{R}(\mathbf{s},\mathbf{a}) = \mathbb{E}\left[r^{(t+1)} | \mathbf{s}^{(t)} = \mathbf{s}, \mathbf{a}^{(t)} = \mathbf{a}\right]$ being the expected reward of $(\mathbf{s},\mathbf{a}) \in \mathcal{S} \times \mathcal{A}$. Moreover, $\mathcal{P}^{\mathbf{a}}_{\mathbf{s},\mathbf{s}'} = \Pr\left(\mathbf{s}^{(t+1)} = \mathbf{s}' | \mathbf{s}^{(t)} = \mathbf{s}, \mathbf{a}^{(t)} = \mathbf{a}\right)$ represents the transition probability from state $\mathbf{s}$ to state $\mathbf{s}'$ upon applying $\mathbf{a}$. In turn, the optimal Q-function associated with ϕ^* is obtained as

$$Q_{\phi^*}\left(\mathbf{s},\mathbf{a}\right) = \mathcal{R}(\mathbf{s},\mathbf{a}) + \delta \sum_{\mathbf{s}' \in \mathcal{S}} \mathcal{P}^{\mathbf{a}}_{\mathbf{s},\mathbf{s}'} \max_{\mathbf{a}'} Q_{\phi^*}\left(\mathbf{s}',\mathbf{a}'\right). \tag{17}$$

The QL agent assigned a Q matrix for each GUE/UAV in the network, denoted $\mathbf{Q}(\mathbf{s},\mathbf{a})$, which serves as a lookup table for each action-value combination. Moreover, the QL algorithm updates each entry in the $\mathbf{Q}$ matrix in each time-step t as

$$\begin{aligned} \mathbf{Q}\left(\mathbf{s}^{(t)},\mathbf{a}^{(t)}\right) \leftarrow (1-\eta)\mathbf{Q}\left(\mathbf{s}^{(t)},\mathbf{a}^{(t)}\right) \\ + \eta\left(r^{(t+1)} + \delta \max_{\mathbf{a}} \mathbf{Q}\left(\mathbf{s}^{(t+1)},\mathbf{a}\right)\right), \end{aligned} \tag{18}$$

where $0 < \eta \leq 1$ is the learning rate to control the speed of reaching a solution. To avoid being stuck at non-optimal policies and to deal with the exploitation versus exploration trade-off issue [65], the ε-greedy policy is used for each time-step t, which implies that the QL agent takes action $\mathbf{a}^*$ that maximizes the Q-function with probability $1 - \varepsilon + \frac{\varepsilon}{|\mathcal{A}(\mathbf{s})|}$ for exploitation, and a random action with probability $\varepsilon + \frac{\varepsilon}{|\mathcal{A}(\mathbf{s})|}$ for exploration [65].

In this work, the reward function $r(\mathbf{s},\mathbf{a})$ takes one of the predefined values $v_1 > v_2 > v_3$, described as

$$r(\mathbf{s},\mathbf{a}) = \begin{cases} v_1, & (\mathbb{EE} \geq \zeta_{\min}) \cap (\mathbb{R}_i \geq R_{\min}) \cap (\mathbb{P}_i \leq P_{\max}), \\ v_2, & (\mathbb{EE} \geq \zeta_{\min}) \cup (\mathbb{R}_i \geq R_{\min}) \cap (\mathbb{P}_i \leq P_{\max}), \\ v_3, & \text{otherwise}, \end{cases} \tag{19}$$

where $\mathbb{EE}$ is the energy-efficiency value, $\mathbb{R}_i$ is the data rate of GUE/UAV i, and $\mathbb{P}_i$ is its the total transmit power over all used RBs, with thresholds $\zeta_{\min}$, $R_{\min}$, and $P_{\max}$, respectively. Every time an the action vector $\mathbf{a}$ is selected, the MBS calculates the rewards, and measures how well the action vector contributes to the maximization of GUEs/UAVs energy-efficiency, while ensuring the minimum date rate and maximum transmit power constraints are satisfied. The QL algorithm is outlined in **Algorithm 1**, which summarizes the process of evaluating the Q values, and obtaining the GUEs/UAVs allocated RBs states, and transmit power action vectors.

Algorithm 1 Q-Learning

1: **Initialization:** $\mathbf{Q}(\mathbf{s},\mathbf{a})$ with zero values, δ, η, and ε.
2: **for** each time-step t **do**
3: For the current state $\mathbf{s}^{(t)}$, pick the action vector $\mathbf{a}^{(t)}$ using the ε-greedy policy, as

$$\mathbf{a}^{(t)} \leftarrow \begin{cases} \arg\max_{\mathbf{a}} \mathbf{Q}\left(\mathbf{s}^{(t+1)},\mathbf{a}\right), & \text{with prob. } 1-\varepsilon+\frac{\varepsilon}{|\mathcal{A}(\mathbf{s}^{(t)})|}, \\ \text{a random action vector, with prob. } \varepsilon + \frac{\varepsilon}{|\mathcal{A}(\mathbf{s}^{(t)})|}. \end{cases}$$

4: Perform action $\mathbf{a}^{(t)}$, obtain reward $r^{(t+1)} = r\left(\mathbf{s}^{(t)},\mathbf{a}^{(t)}\right)$ and observe the new state $\mathbf{s}^{(t+1)}$.
5: Update $\mathbf{Q}\left(\mathbf{s}^{(t)},\mathbf{a}^{(t)}\right)$ as

$$\begin{aligned} \mathbf{Q}\left(\mathbf{s}^{(t)},\mathbf{a}^{(t)}\right) \leftarrow (1-\eta)\mathbf{Q}\left(\mathbf{s}^{(t)},\mathbf{a}^{(t)}\right) \\ + \eta\left(r^{(t+1)} + \delta \max_{\mathbf{a}} \mathbf{Q}\left(\mathbf{s}^{(t+1)},\mathbf{a}\right)\right). \end{aligned}$$

6: Set $t \leftarrow t+1$ and current state $\mathbf{s}^{(t)} \leftarrow \mathbf{s}^{(t+1)}$.
7: **end for**
8: **Output:** State $\mathbf{s}$ and action $\mathbf{a}$ vectors.

The QL algorithm is guaranteed to converge when all actions are repeatedly sampled and the rewards are bounded [25,66]. More importantly, the QL algorithm has two serious issues: (1) the amount of memory need to store and update the $\mathbf{Q}(\mathbf{s},\mathbf{a})$ matrix grows exponentially as the number of states and actions increases, and (2) some states may rarely be visited, which excessively increases the time needed to explore all state-action combinations to obtain a good estimate of $\mathbf{Q}(\mathbf{s},\mathbf{a})$, which is impractical. The complexity for the QL algorithm can be discussed from three perspectives: the regret, and time and space complexity. By definition, the regret of exploration for the RL algorithm is the difference between the T-step cumulative reward obtained by an optimal policy and that

[4]It should be noted that both $\mathbf{s}$ and $\mathbf{a}$ are of dimension $1 \times N$, where N is the number of the reused RBs in the system.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3078188

by the RL algorithm [67]. For the QL algorithm, the upper confidence bound (UCB) regret is $\mathcal{O}\left(\sqrt{SATH^3}\right)$, where T, S, A, and H are the total number of steps, number of states, number of actions, and number of steps per episode, respectively [68]. Moreover, the time complexity are given by $\mathcal{O}(T)$, while the space complexity is $\mathcal{O}(SAH)$. In our work, $H = 1$ (i.e. one step per episode), and thus the space complexity is $\mathcal{O}(SA)$. In turn, our QL algorithm has linear time-complexity and polynomial space-complexity.

B. Deep Q-Learning (DQL)

As for DQL, and as mentioned earlier, a DNN called DQN is utilized to estimate the Q-function instead of the $\mathbf{Q}(\mathbf{s}, \mathbf{a})$ matrix in the QL algorithm. In this work, a multi-layer deep forward neural network is utilized to replace the classical state-action matrix and find the optimal policy. This is achieved by exploiting correlations in the space of the input raw data and identifying the important features that distinguish such input [69]. Moreover, an experience buffer mechanism is used to store the reciprocal experience and randomly pick a group of samples from the stored experience to train the DQL instead of the direct successive samples of the QL algorithm. Furthermore, a second neural network is added to provide the target Q-values. These values will be used to calculate the loss value for each action at DQL training round [28].

Now, let the DQN be denoted $\mathbf{Q}(\mathbf{s}, \mathbf{a}; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is a real-valued vector completely characterizing the function $\mathbf{Q}(\mathbf{s}, \mathbf{a}; \boldsymbol{\theta})$, such that $\mathbf{Q}(\mathbf{s}, \mathbf{a}; \boldsymbol{\theta}) \approx Q_{\phi^*}(\mathbf{s}, \mathbf{a})$. In turn, the search for the best Q-function translates to finding the best $\boldsymbol{\theta}$ of finite dimensions via training. In particular, the DQL agent gathers experiences and forms a data set $\mathcal{D}$ in the form of $\left(\mathbf{s}^{(t)}, \mathbf{a}^{(t)}, r^{(t+1)}, \mathbf{s}^{(t+1)}\right)$ by collecting experiences until-step t. To this end, two DQNs are defined, namely the target DQN with $\boldsymbol{\theta}^{(t)}_{\text{target}}$, and the train DQN with $\boldsymbol{\theta}^{(t)}_{\text{train}}$. Moreover, $\boldsymbol{\theta}^{(t)}_{\text{target}}$ is updated to become equivalent to $\boldsymbol{\theta}^{(t)}_{\text{train}}$ over a specific number of time-steps [63]. In each time-step t, the DQN is trained by minimizing a least squares loss function (i.e. a gradient-descent) based a random mini-batch from $\mathcal{D}$, which is expressed as [70]

$$\mathcal{L}\left(\boldsymbol{\theta}^{(t)}_{\text{train}}\right) = \mathbb{E}\left[y^{(t)} - Q\left(\mathbf{s}^{(t)}, \mathbf{a}^{(t)}; \boldsymbol{\theta}^{(t)}_{\text{train}}\right)\right]^2, \quad (20)$$

where $y^{(t)}$ is the target value function, given by

$$y^{(t)} = r\left(\mathbf{s}^{(t)}, \mathbf{a}^{(t)}\right) + \delta \max_{\mathbf{a}} Q\left(\mathbf{s}^{(t+1)}, \mathbf{a}; \boldsymbol{\theta}^{(t)}_{\text{target}}\right). \quad (21)$$

Due to the possible instability (or divergence) of the DQL, the aperiodic store experience is used to improve the learning stability of the DQL [71]. In addition, ε is updated using the decay rate υ as $\varepsilon = \varepsilon(1 - \upsilon)$, while slowly smoothing the target parameters in every training round with ξ, as

$$\begin{aligned} \boldsymbol{\theta}^{(t)}_{\text{train}} &= \xi \boldsymbol{\theta}^{(t-1)}_{\text{train}} + (1 - \xi)\boldsymbol{\theta}^{(t)}_{\text{train}} \\ \boldsymbol{\theta}^{(t)}_{\text{target}} &= \xi \boldsymbol{\theta}^{(t-1)}_{\text{target}} + (1 - \xi)\boldsymbol{\theta}^{(t)}_{\text{target}}, \end{aligned} \quad (22)$$

ultimately reducing the correlations between the target and estimated Q-values, and thus stabilizing the DQL algorithm.

For DQL, the reward function $r(\mathbf{s}, \mathbf{a})$ takes one of the predefined values and $v_1 > v_2$, as

$$r(\mathbf{s}, \mathbf{a}) = \begin{cases} v_1, & (\mathbb{EE} \geq \zeta_{\min}) \cap (\mathbb{R}_i \geq R_{\min}) \cap (\mathbb{P}_i \leq P_{\max}), \\ v_2, & \text{otherwise}, \end{cases} \quad (23)$$

which maintains the minimum capacity and maximum transmit power for each GUE/UAV, while maximizing the EE. The DQL algorithm is summarized in **Algorithm 2**, which is a model-free, online, off-policy reinforcement learning method that is guaranteed to converge efficiently [63,72][5]. The training procedure in **Algorithm 2** can be described with the help of Fig. 2. For RL, neural networks in most optimization algorithms assume that the samples are independent and identically distributed. However, this is no longer acceptable for samples that have been produced sequentially. Add to this, for efficient use of the existing hardware, it is essential to exploit sampled mini-batches from the stored experience buffer, rather than online experiences. As this experience buffer is a finite-sized cache, the oldest samples will be dropped when the buffer is full, and replaced by new ones that take into consideration the dynamic changes in the wireless environment. At each time-step, the train DQN and target DQN are updated by uniformly sampling a mini-batch from the buffer. Due to the off-policy nature of **Algorithm 2**, the experience buffer can be large, if learning across a set of uncorrelated transitions is required in some scenarios. Note that (20) may cause unstable performance in many environments, since the updated train DQN network is used in updating the target DQN in (21), which may result in Q-values divergence. A solution to this is proposed in [73], where the weights of the target network are updated by having them smoothed gradually with the learned train DQN network, as in (22). This ensures the slow change in the target DQN values, which improves the learning stability.

Since a multi-layer deep neural network is utilized in this work, Fig. 3 illustrates the operation of the proposed power control scheme using DQL. The state and action vectors each have $1 \times (N \times I)$ elements to describe each possible state and action, where N is the number of the reused RBs in each frequency band, and I is the total number of UEs in the UHF or mmWave band. Both $\mathbf{s}$ and $\mathbf{a}$ represent inputs to the DNN, while the output is the estimate of expected long-term reward based on a given status $\mathbf{s}$ of the DQL. The input layers for both $\mathbf{s}$ and $\mathbf{a}$ are followed by multiple deep layers; starting with fully connected layer, described by $\mathbf{y_1} = \mathbf{w_s} \cdot \mathbf{s} + \mathbf{b_s}$, where the input vector $\mathbf{s}$ is weighted by vector $\mathbf{w_s}$ and $\mathbf{b_s}$ is the bias vector. The next layer is the Rectified Linear Unit (ReLU) used to suppress any negative output value of the previous fully connected layers to zero, and the output is $\mathbf{y_2} = \max(\mathbf{y_1}, 0)$, then another fully connected layer is applied. In order to update $\mathbf{s}$ by adding the actions $\mathbf{a}$, the

[5]**Online** learning algorithms work with data as it is made available. Strictly online algorithms improve incrementally from each piece of new data as it arrives, then discard that data and not use it again. Also, **off-policy** algorithms work with two policies which are: (a) a policy being learned, called the target policy, and (b) a policy being followed, called the behaviour policy. Via an **online**, **off-policy** RL algorithm, the learning agent sets the task of behaving optimally in an environment. It may behave and gain observations from the behaviour policy, but learns a separate optimal target policy [24].

Algorithm 2 Deep Q-Learning

1: **Initialization:** Experience memory $\mathcal{D}$, δ, υ, ξ, ε, and $\varepsilon_{\min}$ with $\varepsilon > \varepsilon_{\min}$. Also, initialize training parameters $\boldsymbol{\theta}_{\text{train}}$, and target parameters as $\boldsymbol{\theta}_{\text{target}} = \boldsymbol{\theta}_{\text{train}}$.
2: **for** each time-step t **do**
3: For the current state $\mathbf{s}^{(t)}$, pick the action vector $\mathbf{a}^{(t)}$ using the ε-greedy policy, as

$$\mathbf{a}^{(t)} \leftarrow \begin{cases} \arg\max_{\mathbf{a}} \mathbf{Q}\left(\mathbf{s}^{(t+1)}, \mathbf{a}; \boldsymbol{\theta}_{\text{target}}^{(t)}\right), & \text{with prob. } 1-\varepsilon+\frac{\varepsilon}{|\mathcal{A}(\mathbf{s}^{(t)})|}, \\ \text{a random action vector}, & \text{with prob. } \varepsilon+\frac{\varepsilon}{|\mathcal{A}(\mathbf{s}^{(t)})|}. \end{cases}$$

4: Perform action $\mathbf{a}^{(t)}$, obtain reward $r^{(t+1)} = r\left(\mathbf{s}^{(t)}, \mathbf{a}^{(t)}\right)$ and observe the new state $\mathbf{s}^{(t+1)}$.
5: Store $\left(\mathbf{s}^{(t)}, \mathbf{a}^{(t)}, r^{(t+1)}, \mathbf{s}^{(t+1)}\right)$ in experiences memory $\mathcal{D}$.
6: Pick a random mini-batch of from $\mathcal{D}$.
7: Determine the target value function $y^{(t)}$ as

$$y^{(t)} = r\left(\mathbf{s}^{(t)}, \mathbf{a}^{(t)}\right) + \delta \max_{a} Q\left(\mathbf{s}^{(t+1)}, \mathbf{a}; \boldsymbol{\theta}_{\text{target}}^{(t)}\right).$$

8: Update parameters $\boldsymbol{\theta}_{\text{train}}^{(t)}$ by minimizing the loss function

$$\mathcal{L}\left(\boldsymbol{\theta}_{\text{train}}^{(t)}\right) = \mathbb{E}\left[y^{(t)} - Q\left(\mathbf{s}^{(t)}, \mathbf{a}^{(t)}; \boldsymbol{\theta}_{\text{train}}^{(t)}\right)\right]^2.$$

9: Update the target parameters $\boldsymbol{\theta}_{\text{train}}^{(t)}$ and $\boldsymbol{\theta}_{\text{target}}^{(t)}$ using ξ as

$$\boldsymbol{\theta}_{\text{train}}^{(t)} = \xi\boldsymbol{\theta}_{\text{train}}^{(t-1)} + (1-\xi)\boldsymbol{\theta}_{\text{train}}^{(t)}$$
$$\boldsymbol{\theta}_{\text{target}}^{(t)} = \xi\boldsymbol{\theta}_{\text{target}}^{(t-1)} + (1-\xi)\boldsymbol{\theta}_{\text{target}}^{(t)}.$$

10: Set $t \leftarrow t+1$ and current state $\mathbf{s}^{(t)} \leftarrow \mathbf{s}^{(t+1)}$.
11: **if** $\varepsilon > \varepsilon_{min}$ **then**
12: Update $\varepsilon = \varepsilon(1-\upsilon)$.
13: **end if**
14: **end for**
15: **Output:** State **s** and action **a** vectors.

Add layer has been used to obtain the output. To remove any negative power, a ReLU layer has been used. Lastly, a fully connected layer with a single output is used to provide the state-action function $Q(\mathbf{s}, \mathbf{a})$, as illustrated in Fig. 4, which presents a block diagram for proposed solution.

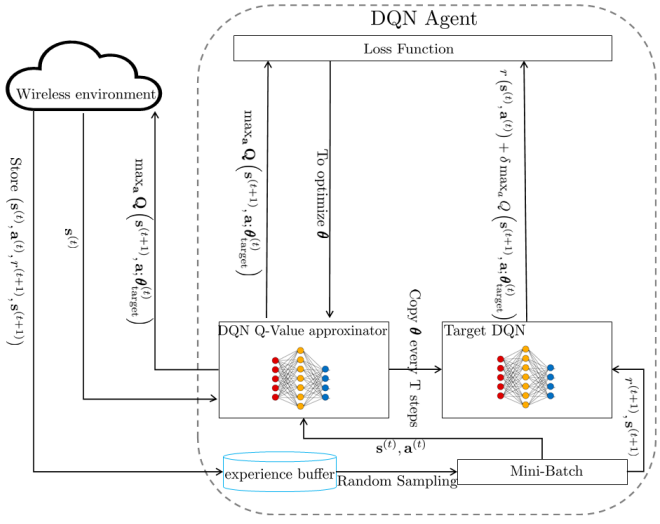

Fig. 2. Training block diagram of the DQN

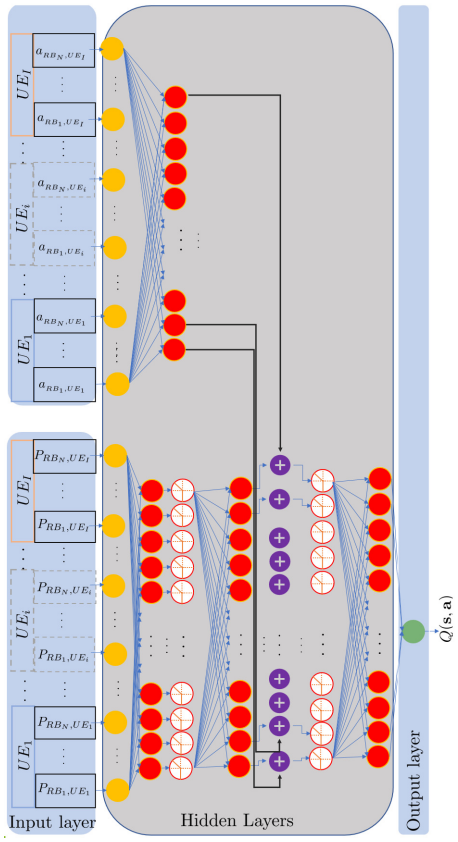

Fig. 3. Proposed power control scheme using DQL

VI. Implementation of QL and DQL for Optimizing DUDe Access Energy-Efficiency

This section discusses the implementation of the QL and DQL algorithms. It should be noted that since the UHF and mmWave UEs do not interfere with each other, the QL/DQL algorithm is executed for both the GUEs and UAVs over each band separately to obtain the transmit power values for GUEs and UAVs EE maximization.

Now, the state vector **s** contains the power value of the user RBs, say GUE/UAV i, starting with a low power value (e.g. $P_{i,n} = 2 \times 10^{-6}$ W) up to the maximum transmit power value $P_{\max} = 0.2$ W. For the QL algorithm, each action in the action vector **a** involves multiplying the RB power by one of three values in $\{0.1, 1, 10\}$ (as per the ε-greedy policy), which facilitates the exploration and exploitation to maximize the Q-function. For example, consider the case of two GUEs (or UAVs), each with two RBs; then, at the BS two independent QL agents will be created. Each of these GUEs (or UAVs) independently learns its own policy, and considers the other agent as a component in the wireless environment [74]. On the other hand, the DQL algorithm action vector **a** is a combination of adding or subtracting a step value c for each RB of each GUE/UAV. For instance, if $c = 5 \times 10^{-6}$ W, then for each element of the states vector **s**, the new observation is manipulated by adding or subtracting the c step randomly. As before, consider the case of two GUEs (or UAVs), each with two RBs. A possible action vector is $\mathbf{a} = [-1 \times c; 1 \times c; -1 \times c; 1 \times c]$. Upon applying an action

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3078188

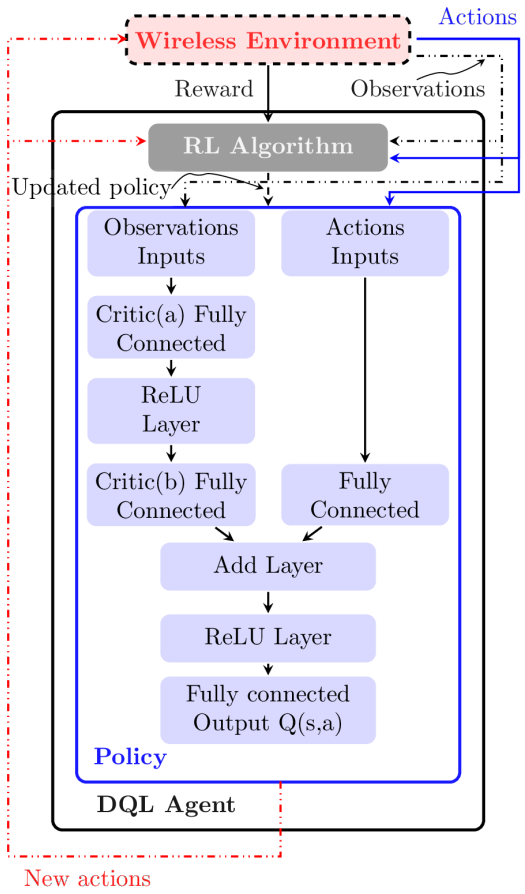

Fig. 4. Block diagram for proposed DQL solution

vector, the states vector is updated to values in the range $\left[10^{-6}, 0.2\right]$ W, which designates the transmit power values of either UHF or mmWave UEs. The input, as shown in Fig. 3, has the transmit power of each RB as a feature, while the BS has the information about the RB association scheme, received signal strength and interference. Then, the DQN learns the relative location of GUEs/UAVs, data rates, and EE. Also, the actions—as part of the inputs—are changing the power states to find the best possible state that achieves the near-optimal EE solution.

In this work, the final or exit state is not known apriori, since the optimal value for the EE is not known before executing the QL (or DQL) algorithm. However, convergence to a certain Q value or reaching the maximum number of iterations terminates the search process, and the final obtained Q value resembles the best EE value. In turn, the learning rate η, discount factor δ, and maximum number of iterations $T_{\max}$ govern the convergence speed and accuracy of the obtained EE solution. Particularly, a higher learning rate η allows better solution exploration, and a value of $\delta \rightarrow 1$ puts more emphasis on long-term higher rewards. Also, the higher $T_{\max}$ is, the better the exploration and exploitation, which guarantees the optimal states for GUE/UAV transmit powers. Hence, the values of η, δ and $T_{\max}$ pose a trade-off between accuracy of the obtained solution and speed of convergence. To highlight this, Table III summarizes the parameters for two scenarios, which will be considered in the performance evaluation in Section VII.

In a similar manner to the QL algorithm, DQL is evaluated based on two scenarios, as shown in Table IV.

TABLE III
QL PARAMETERS

Parameters	Scenario 1	Scenario 2
No. of States	No. of RBs = 3	
Possible States	$s \in \left[2 \times 10^{-6}, 0.2\right]$ W	
No. of Actions	3	
Possible Actions	$a \in \{0.1, 1, 10\}$ W	
Reward Function Values	$\upsilon_1 = 10$ $\upsilon_2 = 1$ $\upsilon_3 = -10$ $\zeta^G_{\min} = \zeta^A_{\min} = 4$ MBits/J $R^G_{\min} = 0.4$ MBits/s $R^A_{\min} = 4$ MBits/s $P^G_{\max} = P^A_{\max} = 0.2$ W	
Discount Factor δ	0.1	
ε-Greedy Parameter	0.1	
Learning Rate η	0.1	0.01
$T_{\max}$	30,000	50,000

TABLE IV
DQL PARAMETERS

Parameters	Scenario 1	Scenario 2
States Input Layer	Input Size: 5 Users × 3 RBs = 15 Neurons Output: 24 Neurons—Normalization: None	
Actions Input Layer	Input Size: 15 Neurons—Output: 50 Neurons Normalization: None	
States Critic(a) Fully Connected Layer	Input Size: 50 Neurons—Output: 50 Neurons Normalization: None	
States Critic(b) Fully Connected Layer	Input Size: 50 Neurons—Output: 50 Neurons Normalization: None	
Action Critic Fully Connected Layer	Input Size: 50 Neurons—Output: 50 Neurons Normalization: None	
ReLU Layers for Critic & Action Paths	$f(x) = \max(0, x)$	
Add layer	Adding neurons element wise	
Fully Connected Output $Q(\mathbf{s}, \mathbf{a})$	Input Size: 50 Neurons—Output: 1 Neuron Normalization: None	
Possible States	$s \in \left[10^{-6}, 0.2\right]$ W	
No. of Actions	3	
Possible Actions	$a \in \{+10^{-6}, -10^{-6}\}$ W	
Reward Function Values	$\upsilon_1 = 10$ $\upsilon_2 = -10$ $\zeta^G_{\min} = \zeta^A_{\min} = 4$ MBits/J $R^G_{\min} = 0.4$ MBits/s $R^A_{\min} = 4$ MBits/s $P^G_{\max} = P^A_{\max} = 0.2$ W	
ε-Greedy Parameter	0.1	
Decay Rate υ	0.005	
ε_{min}	0.01	
Smoothing Factor ξ	0.001	
Discount Factor δ	0.1	
Learning Rate η	0.1	0.01
$T_{\max}$	30,000	50,000

VII. PERFORMANCE EVALUATION

In this section, the performance of the coupled UHF and coupled mmWave with FPC are compared to the DUDe access scheme in terms of sum-rate, energy-efficiency, and data rate per GUE/UAV. Specifically, the performance of the QL and DQL power control schemes based on DUDe

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3078188

TABLE V
SIMULATION PARAMETERS

Parameters	GUE	UAV	MBS	SBS
Maximum Transmit Power	23 dBm	23 dBm	46 dBm	30 dBm
DL/UL Bias	N/A	N/A	0/0 dB	3/0 dB
Spatial Density	30 per km^2	30 per km^2	5 per km^2	20 per km^2
Operating Frequency	2 GHz & 28 GHz	28 GHz	2 GHz & 28 GHz	2 GHz & 28 GHz
Spatial Distribution	Uniform Distribution			
Altitude of UAVs	50-200 m [11,50]			
S-curve Parameters	$a = 9.6$, $b = 0.28$ [43,46]			
Blockage Ball Model Parameters	$\mu = 200$ m, $\omega = 0.2$ [40]			
Bandwidth	UHF: 1.2 MHz; mmWave: 4.8 MHz			
Subcarrier Spacing	UHF: 15 kHz; mmWave: 60 kHz [75]			
Power Control	FPC with $P_0 = -85$ dBm, and $\alpha = 0.8$ [11]			
Noise Spectral Density	-174 dBm/Hz			
Path-Loss Exponent	UHF: GUE-UAV 2, GUE-BS 3, $d_0 = 1$ m [49]; mmWave: LOS 2.55, NLOS 5.76, $d_0 = 5$ m [56,76]			
Lognormal Shadowing	UHF: $\mu = 0$, $\sigma = 4$ dB [49]; mmWave: LOS $\mu = 0$, $\sigma = 8.66$ dB, NLOS $\mu = 0$, $\sigma = 9.02$ dB [56]			
Nakagami-m Parameters	$m_L = 3$, $m_N = 2$ [46,48]			

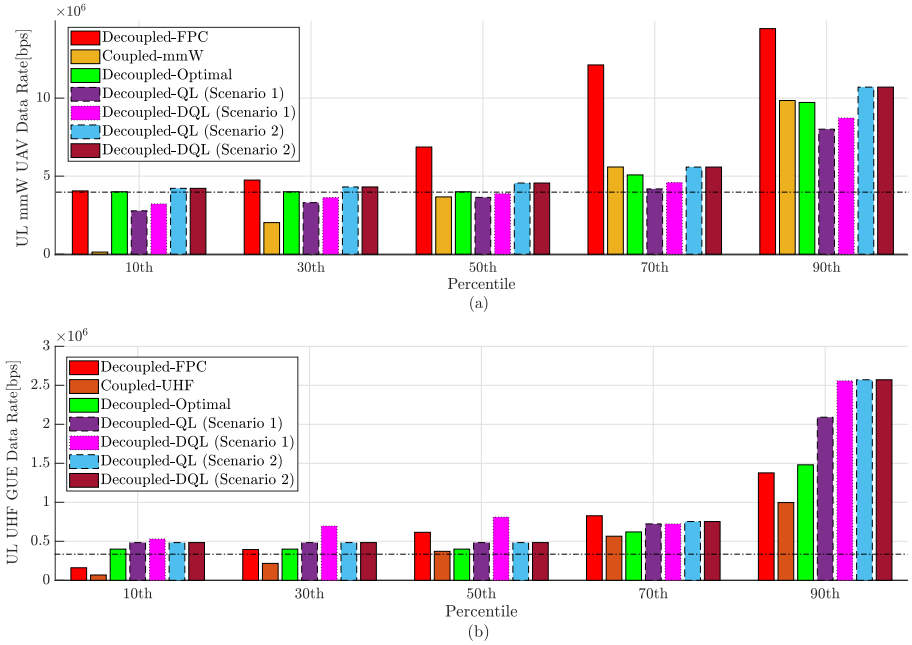

Fig. 5. 10^{th}, 30^{th}, 50^{th}, 70^{th}, and 90^{th} percentile data rate per user in the UL: (a) mmWave band and (b) UHF band - SBS to MBS ratio $= 4$

access are evaluated and compared with the optimal and FPC schemes, namely **Decoupled-Optimal** and **Decoupled-FPC**, respectively[6]. Since the UAV CNPC links require low data rate, and will not interfere with other links, they are not considered in the simulations, for simplicity. The frequency allocation and BS association parameters are as given in Table II, while Table V summarizes the simulated transmission parameters.

Fig. 5 illustrates the 10^{th}, 30^{th}, 50^{th}, 70^{th}, and 90^{th} percentile data rate per GUE and UAV in the UL, where the ratio of SBS to the MBS is 4. In Fig. 5a, the UAV data rates shows that due to EE optimization the 70^{th} and 90^{th} percentile data rates in **Decoupled-Optimal**, QL and DQL are lower than the decoupled FPC. Although the minimum rates are respectively 4×10^6 and 4×10^5 bps for the UHF GUEs and mmWave UAVs, some of the mmWave UAVs under the QL and DQL schemes are below the thresholds. This is because the EE thresholds (i.e. $\zeta^G_{\min}$ and $\zeta^A_{\min}$) appear as soft thresholds (as per (19) and (23)), which leads to a tradeoff between the data rate and EE. Additionally, Fig. 5a and Fig. 5b show that both QL and DQL algorithms improve their learning policies and assign power to GUEs/UAVs to increase their data rates when the number of training iterations is increased and their learning rates are decreased. This can be verified by comparing **Scenarios 1** and **2** for the **Decoupled-QL** and **Decoupled-DQL** schemes, and this is due to the fact that more states are visited in search for the best state. More importantly, this implies that improving the learning increases the rate of the UAVs/GUEs for EE-maximization in both the mmWave and UHF bands.

In Fig. 6, it can be seen that the GUEs sum-rate of the decoupled schemes are at least 60% higher than the **Coupled-UHF** scheme. This is because the decoupled schemes have

[6]The optimal EE-maximizing power control schemes are based on the solutions of problems **GUEs-EE-MAX** and **UAVs-EE-MAX**, as discussed in subsections IV-A, and IV-B, respectively.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3078188

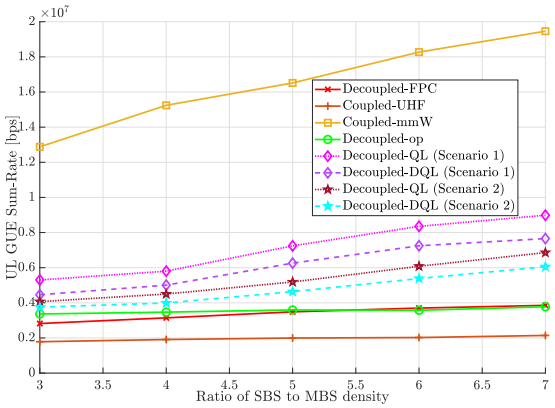

Fig. 6. UL GUE sum-rate vs. SBS to MBS ratio.

wider bandwidth for GUE UL communications (as shown in Table II), shorten the distances between the GUEs and BSs, and eliminate the interference between UAVs and GUEs. The sum-rate of the **Coupled-mmW** scheme is the highest, since it utilizes the mmWave band for GUE UL communication, while the other schemes utilize the UHF band, and the mmWave bandwidth is wider than the UHF bandwidth. The sum-rate of the **Decoupled-Optimal** scheme remains relatively constant with the increase in the SBS to MBS ratio, since it mainly aims to achieve the optimal energy-efficiency, as will shown in Figs. 8 and 9. In comparison to the **Decoupled-Optimal** scheme, both the **Decoupled-QL** and **Decoupled-DQL** schemes yield higher data rates at the expense of higher transmit power, which will translate to lower EE values. To see this, for both schemes, **Scenario 2** yields lower sum-rate than **Scenario 1**, as increasing the training iterations and reducing the learning rate lower the sum-rate to improve the EE by carefully selecting the transmit power.

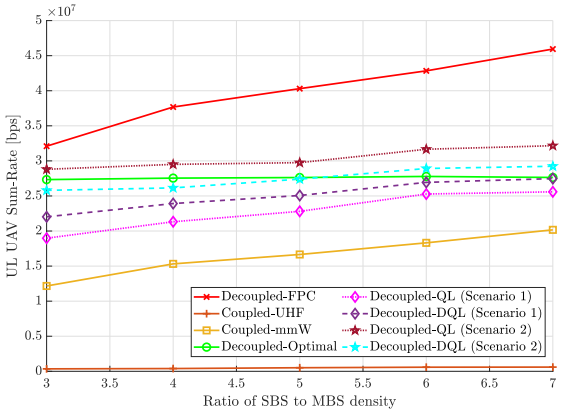

Fig. 7. UL UAV sum-rate vs. SBS to MBS ratio.

Similarly, in Fig. 7, the UAV sum-rates of the **Coupled-UHF** and **Coupled-mmW** schemes are much lower than the decoupled schemes, since the UAVs under those two schemes are allocated narrower bandwidth and suffer from higher path-loss. Besides, the UAVs under the **Coupled-UHF** scheme also suffer from ICI, while the UAVs under the other schemes are allocated the mmWave band, and thus, their ICI is minimal. In comparison to the **Decoupled-Optimal** scheme, both **Decoupled-DQL** and **Decoupled-QL** tend to explore if increasing the sum-rates achieves better EE for the UAV UL transmissions and this appears as a small increase in the sum-rate when the SBS to MBS ratio increases. Adding to this, **Scenario 2** improves the sum-rates for both the **Decoupled-DQL** and **Decoupled-QL** in comparison to **Scenario 1**.

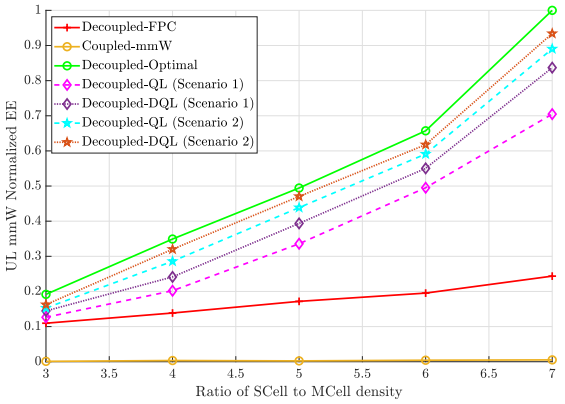

Fig. 8. UL mmWave UEs normalized EE vs. SBS to MBS ratio.

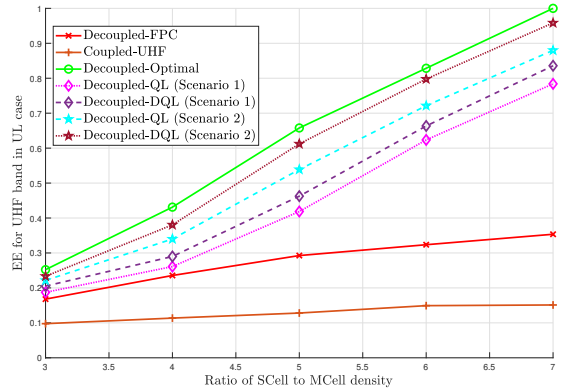

Fig. 9. UL UHF UEs normalized EE vs. SBS to MBS ratio.

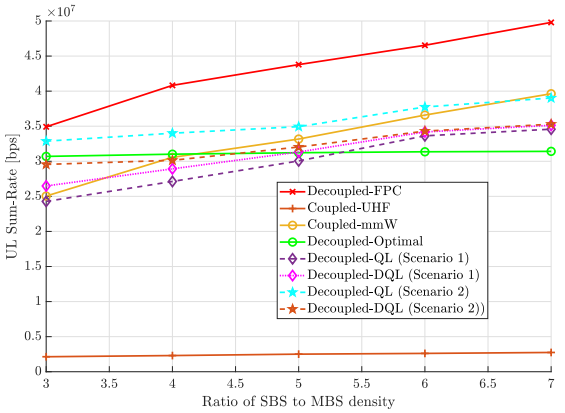

Fig. 10. Network UL sum-rate vs. SBS to MBS ratio.

As for EE, as shown in Fig. 8 and Fig. 9, the decoupled schemes can achieve up to several times higher EE than the

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3078188

coupled schemes, as they prevent the interference between UAVs and GUEs, reduce the interference among GUEs, and shorten the GUE/UAV-BS distances. Also, Figs. 8 and 9 demonstrate that the EE improvement for the **Decoupled-Optimal**, **Decoupled-QL** and **Decoupled-DQL** schemes as the SBS to MBS ratio increases. This is attributed to the decrease in the number of GUEs/UAVs associated with the same SBS or MBS, and the decrease in GUE/UAV-BS distances. In addition, the **Decoupled-QL** and **Decoupled-DQL** schemes yield an improvement in the EE as the training iterations increase and the learning rate decreases, which can be verified by comparing **Scenario 1** and **Scenario 2** for both schemes.

Fig. 10 illustrates the total UL sum-rate, where one can see that the sum-rate of the **Decoupled-QL** and **Decoupled-DQL** schemes improves as the ratio of SBS to MBS increases. Also, the sum-rate for the **Decoupled-QL (Scenario 2)** and **Decoupled-DQL (Scenario 2)** schemes have a minor improvement over the **Decoupled-QL (Scenario 1)** and **Decoupled-DQL (Scenario 1)** schemes, respectively. Lastly the QL and DQL algorithms are limited by the maximum number of iterations in search for the best trade-off between sum rate, minimum user rate, and the EE, which also control how long it takes to run the optimization process.

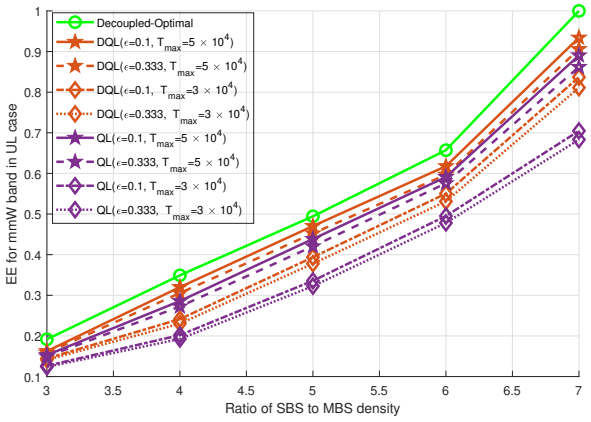

Fig. 11. Different (ε, T_{max}) combinations for EE mmW vs. SBS to MBS ratio.

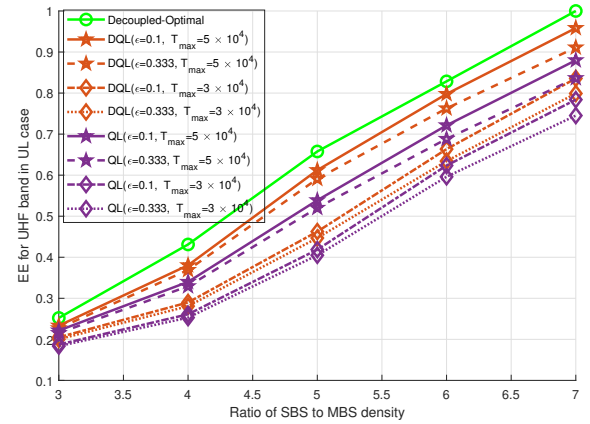

Fig. 12. Different (ε, T_{max}) combinations for EE UHF vs. SBS to MBS ratio.

Fig. 11 and 12 illustrate that by decreasing the value of ε and increasing the value of $T_{\max}$, the obtained EE approaches that of the optimal scheme. More importantly, the DQL algorithm outperforms its QL counterpart algorithm, as the DQL searches more states using actions with smaller transmit power steps. Also increasing $T_{\max}$ has higher impact on improving the solution than reducing the ε. However, this costs more time (i.e. more iterations) to approach the optimal EE value.

Figs. 13a and b reveal that the DQL and QL solutions converge as the iterations number increases. Although the DQL algorithm converges slower than the QL, it has better results than the QL when both are compared with the optimal values. Also, when the SBS to MBS ratio increases, the EE increases, as the distance between the BSs and GUEs/UAVs decreases, which requires lower transmit power. However, this adds more complexity (hence more iterations) to find the optimal solution.

VIII. Conclusion

In this paper, the merits of adopting DUDe in cellular-enabled UAV networks have been investigated. Specifically, the UAV data links and CNPC links, as well as GUE ULs and DLs have been decoupled in terms of serving BSs and operating frequencies. Moreover, two power control schemes based on QL and DQL have been proposed to improve the network EE. The proposed decoupled schemes with QL and DQL have been compared with the FPC scheme, and the optimal EE-maximizing benchmark power allocation scheme. The results revealed that the proposed DUDe access schemes can achieve several times higher sum-rates and EE than their coupled counterparts. Moreover, it is shown that although the RL methods can achieve optimal results, in practical scenarios with predominantly dynamic environments, and limited time to execute the optimization process, QL and DQL, with limited number of iterations, may only achieve a near-optimal EE performance. Nonetheless, the proposed QL (DQL) algorithm has been shown to achieve better EE performance than the baseline FPC scheme by around 80% (100%) for UHF band, and by around 160% (170%) for the mmWave band, in comparison to conventional FPC scheme. Also, the proposed DQL achieves better performance than the proposed QL for both **Scenario 1** and **Scenario 2**. This is because DQL has higher number of states and considers all GUEs/UAVs jointly as one agent, while the QL considers each GUE/UAV independently. Lastly, by decreasing the value of ε and increasing the value of $T_{\max}$, the obtained EE approaches that of the optimal scheme, and the DQL algorithm has been shown to outperform the QL algorithm.

References

[1] Y. Zeng, J. Lyu, and R. Zhang, "Cellular-connected UAV: Potential, challenges, and promising technologies," *IEEE Wirel. Commun.*, vol. 26, no. 1, pp. 120–127, 2018.

[2] S. Zhang, Y. Zeng, and R. Zhang, "Cellular-enabled UAV communication: A connectivity-constrained trajectory optimization perspective," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2580–2604, March 2019.

[3] G. Zhang, Q. Wu, M. Cui, and R. Zhang, "Securing UAV communications via joint trajectory and power control," *IEEE Trans. Wirel. Commun.*, vol. 18, no. 2, pp. 1376–1389, Feb 2019.

This article has been accepted for publication in IEEE Internet of Things Journal. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/JIOT.2021.3078188

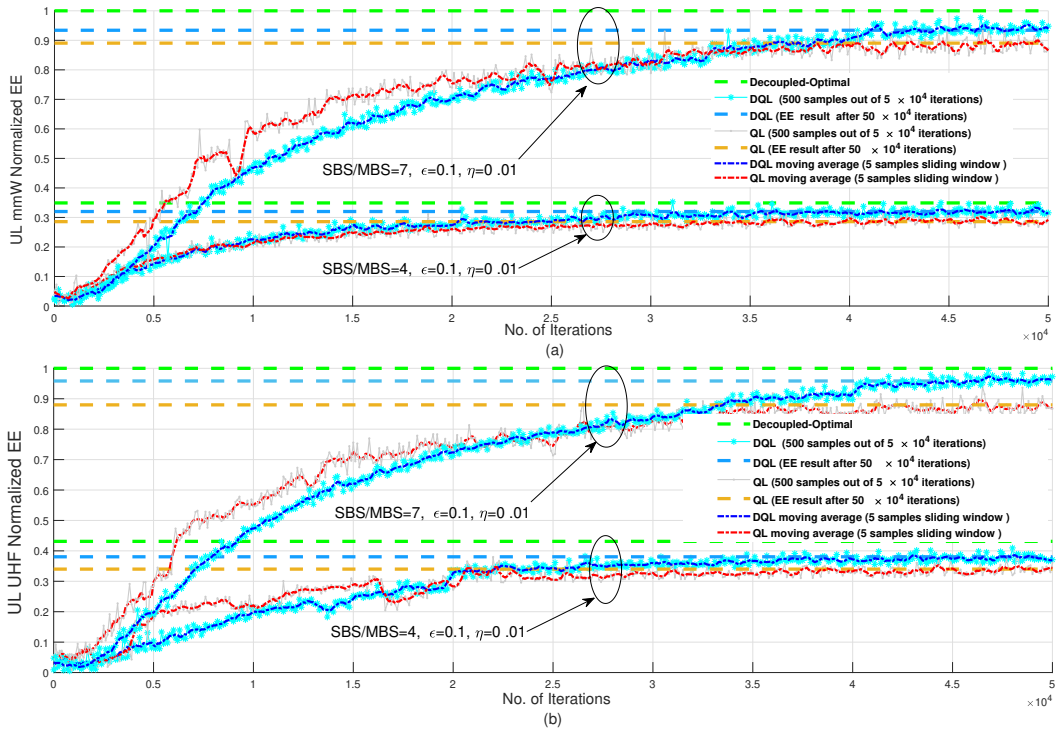

Fig. 13. EE vs No. of iterations

[4] J. Wang, C. Jiang, Z. Han, Y. Ren, R. G. Maunder, and L. Hanzo, "Taking drones to the next level: Cooperative distributed unmanned-aerial-vehicular networks for small and mini drones," *IEEE Veh. Technol. Mag.*, vol. 12, no. 3, pp. 73–82, 2017.

[5] H. Zhao, H. Wang, W. Wu, and J. Wei, "Deployment algorithms for UAV airborne networks toward on-demand coverage," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 9, pp. 2015–2031, 2018.

[6] H. Wang, H. Zhao, W. Wu, J. Xiong, D. Ma, and J. Wei, "Deployment algorithms of flying base stations: 5G and beyond with UAVs," *IEEE Internet Things J.*, vol. 6, no. 6, pp. 10 009–10 027, 2019.

[7] J. Wang, C. Jiang, Z. Wei, C. Pan, H. Zhang, and Y. Ren, "Joint UAV hovering altitude and power control for space-air-ground IoT networks," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1741–1753, 2019.

[8] S. Zhang, Y. Zeng, and R. Zhang, "Cellular-enabled UAV communication: A connectivity-constrained trajectory optimization perspective," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2580–2604, 2018.

[9] M. Mozaffari, A. T. Z. Kasgari, W. Saad, M. Bennis, and M. Debbah, "Beyond 5G with UAVs: Foundations of a 3D wireless cellular network," *IEEE Trans. Wirel. Commun.*, vol. 18, no. 1, pp. 357–372, Jan. 2019.

[10] M. Mozaffari, W. Saad, M. Bennis, Y.-H. Nam, and M. Debbah, "A tutorial on UAVs for wireless networks: Applications, challenges, and open problems," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2334–2360, Mar. 2019.

[11] 3GPP, "Study on Enhanced LTE Support for Aerial Vehicles," 3rd Generation Partnership Project (3GPP), Technical Report (TR) 36.777, 12 2017, version 1.0.0. [Online]. Available: https://ftp.3gpp.org//Specs/archive/36_series/36.777/36777-100.zip

[12] B. V. Der Bergh, A. Chiumento, and S. Pollin, "LTE in the sky: Trading off propagation benefits with interference costs for aerial nodes," *IEEE Commun. Mag.*, vol. 54, no. 5, pp. 44–50, May 2016.

[13] Y. Zeng and R. Zhang, "Energy-efficient UAV communication with trajectory optimization," *IEEE Trans. Wirel. Commun.*, vol. 16, no. 6, pp. 3747–3760, Jun. 2017.

[14] D. Yang, Q. Wu, Y. Zeng, and R. Zhang, "Energy tradeoff in ground-to-UAV communication via trajectory design," *IEEE Trans. Veh. Technol.*, vol. 67, no. 7, pp. 6721–6726, Jul. 2018.

[15] S. Zhang, Y. Zeng, and R. Zhang, "Cellular-enabled UAV communication: A connectivity-constrained trajectory optimization perspective," *IEEE Trans. Commun.*, vol. 67, no. 3, pp. 2580–2604, Mar. 2019.

[16] S. Yin, S. Zhao, Y. Zhao, and F. R. Yu, "Intelligent trajectory design in UAV-aided communications with reinforcement learning," *IEEE Trans. Veh. Technol.*, vol. 68, no. 8, pp. 8227–8231, Aug. 2019.

[17] J. Wang, Z. Na, and X. Liu, "Collaborative design of multi-UAV trajectory and resource scheduling for 6G-enabled Internet of Things," *IEEE Internet Things J.*, pp. 1–1, 2020.

[18] J. Sangiamwong, Y. Saito, N. Miki, T. Abe, S. Nagata, and Y. Okumura, "Investigation on cell selection methods associated with inter-cell interference coordination in heterogeneous networks for LTE-advanced downlink," in *Proc. of 17th European Wireless 2011 - Sustainable Wireless Technologies*, Apr. 2011, pp. 1–6.

[19] H. Elshaer, F. Boccardi, M. Dohler, and R. Irmer, "Downlink and uplink decoupling: A disruptive architectural design for 5G networks," in *IEEE Global Communications Conference (GLOBECOM)*, Dec. 2014, pp. 1798–1803.

[20] L. Zhang, W. Nie, G. Feng, F.-C. Zheng, and S. Qin, "Uplink performance improvement by decoupling uplink/downlink access in HetNets," *IEEE Trans. Veh. Technol*, vol. 66, no. 8, pp. 6862–6876, 2017.

[21] C. Zhang, W. Zhang, W. Wang, L. Yang, and W. Zhang, "Research challenges and opportunities of UAV millimeter-wave communications," *IEEE Wirel. Commun.*, vol. 26, no. 1, pp. 58–62, 2019.

[22] B. Li, Z. Fei, and Y. Zhang, "UAV communications for 5G and beyond: Recent advances and future trends," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 2241–2263, 2018.

[23] J. Cui, Y. Liu, and A. Nallanathan, "Multi-agent reinforcement learning-based resource allocation for UAV networks," *IEEE Trans. Wirel. Commun.*, vol. 19, no. 2, pp. 729–743, Feb. 2020.

[24] R. S. Sutton and A. G. Barto, *Reinforcement learning: An introduction*. MIT press, 2018.

[25] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, no. 3, pp. 279–292, May 1992.

[26] V. Mnih and et al, "Human-level control through reinforcement learning," *Nature*, vol. 518, no. 7540, pp. 529–533, 2015.

[27] C. Zhang, P. Patras, and H. Haddadi, "Deep learning in mobile and wireless networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 3, pp. 2224–2287, Mar. 2019.

[28] N. C. Luong, D. T. Hoang, S. Gong, D. Niyato, P. Wang, Y.-C. Liang, and D. I. Kim, "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3133–3174, May 2019.

[29] G. L. Santos, P. T. Endo, D. Sadok, and J. Klener, "When 5G meets deep learning: A systematic review," *Algorithms*, vol. 13, no. 208, pp. 1–34, Aug. 2020.

[30] 3GPP, "Enhancement for Unmanned Aerial Vehicles; Stage 1," 3rd Generation Partnership Project (3GPP), Technical Report (TR) 22.829, 09 2019, version 17.1.0. [Online]. Available: https://portal.3gpp.org/desktopmodules/Specifications/SpecificationDetails.aspx?specificationId=3557

[31] U. Challita, W. Saad, and C. Bettstetter, "Interference management for cellular-connected UAVs: A deep reinforcement learning approach," *IEEE Trans. Wirel. Commun.*, vol. 18, no. 4, pp. 2125–2140, Apr. 2019.

[32] H. Qi, Z. Hu, H. Huang, X. Wen, and Z. Lu, "Energy efficient 3-D UAV control for persistent communication service and fairness: A deep reinforcement learning approach," *IEEE Access*, vol. 8, pp. 53 172–53 184, Mar. 2020.

[33] C. H. Liu, Z. Chen, J. Tang, J. Xu, and X. Piao, "Energy-efficient UAV control for effective and fair communication coverage: A deep reinforcement learning approach," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 9, pp. 2059–2070, Sept. 2018.

[34] Y. Cao, L. Zhang, and Y. C. Liang, "Deep reinforcement learning for channel and power allocation in UAV-enabled IoT systems," *Proc. IEEE Global Communications Conference (GLOBECOM)*, pp. 1–6, 2019.

[35] C. H. Liu, X. Ma, X. Gao, and J. Tang, "Distributed energy-efficient multi-UAV navigation for long-term communication coverage by deep reinforcement learning," *IEEE Trans. Mobile Comput.*, vol. 19, no. 6, pp. 1274–1285, 2020.

[36] A. F. dos Reis, G. Brante, R. Parisotto, R. D. Souza, P. H. V. Klaine, J. P. Battistella, and M. A. Imran, "Energy efficiency analysis of drone small cells positioning based on reinforcement learning," *Internet Technol. Lett.*, vol. 3, no. 5, p. e166, 2020. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/itl2.166

[37] H. V. Abeywickrama, Y. He, E. Dutkiewicz, B. A. Jayawickrama, and M. Mueck, "A reinforcement learning approach for fair user coverage using UAV mounted base stations under energy constraints," *IEEE Open Journal of Vehicular Technology*, vol. 1, pp. 67–81, 2020.

[38] J. Qiu, J. Lyu, and L. Fu, "Placement optimization of aerial base stations with deep reinforcement learning," *IEEE International Conference on Communications (ICC)*, pp. 1–6, 2020.

[39] Y. Shi, E. Alsusa, and M. W. Baidas, "Energy-efficient decoupled access scheme for cellular-enabled UAV communication systems," *IEEE Syst. J.*, Jan. 2021, DOI: 10.1109/JSYST.2020.3046689.

[40] M. Shi, K. Yang, C. Xing, and R. Fan, "Decoupled heterogeneous networks with millimeter wave small cells," *IEEE Trans. Wirel. Commun.*, vol. 17, no. 9, pp. 5871–5884, Sept. 2018.

[41] H. Elshaer, M. N. Kulkarni, F. Boccardi, J. G. Andrews, and M. Dohler, "Downlink and uplink cell association with traditional macrocells and millimeter wave small cells," *IEEE Trans. Wirel. Commun.*, vol. 15, no. 9, pp. 6244–6258, Sept. 2016.

[42] S. Singh, M. N. Kulkarni, A. Ghosh, and J. G. Andrews, "Tractable model for rate in self-backhauled millimeter wave cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 10, pp. 2196–2211, 2015.

[43] A. Al-Hourani, S. Kandeepan, and S. Lardner, "Optimal LAP altitude for maximum coverage," *IEEE Wireless Commun. Lett.*, vol. 3, no. 6, pp. 569–572, Dec 2014.

[44] K. Venugopal, M. C. Valenti, and R. W. Heath, "Device-to-device millimeter wave communications: Interference, coverage, rate, and finite topologies," *IEEE Trans. Wirel. Commun.*, vol. 15, no. 9, pp. 6175–6188, Sep. 2016.

[45] R. Ma, W. Yang, Y. Zhang, and S. Wang, "Secure on-off transmission in UAV relay-assisted mmwave networks," *Applied Sciences*, vol. 9, no. 19, p. 4138, 2019.

[46] Y. Zhu, G. Zheng, and M. Fitch, "Secrecy rate analysis of UAV-enabled mmWave networks using Matérn hardcore point processes," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 7, pp. 1397–1409, July 2018.

[47] Y. Kim and J. Kim, "An efficient subcarrier allocation scheme for capacity enhancement in multiuser OFDM systems," in *IEEE Vehicular Technology Conference (VTC-Spring)*, 2008, pp. 1915–1919.

[48] T. Bai and R. W. Heath, "Coverage and rate analysis for millimeter-wave cellular networks," *IEEE Trans. Wirel. Commun.*, vol. 14, no. 2, pp. 1100–1114, Feb. 2015.

[49] J. Chakareski, S. Naqvi, N. Mastronarde, J. Xu, F. Afghah, and A. Razi, "An energy efficient framework for UAV-assisted millimeter wave 5G heterogeneous cellular networks," *IEEE Trans. Green Commun. Netw.*, vol. 3, no. 1, pp. 37–44, March 2019.

[50] A. Fotouhi, H. Qiang, M. Ding, M. Hassan, L. G. Giordano, A. Garcia-Rodriguez, and J. Yuan, "Survey on UAV cellular communications: Practical aspects, standardization advancements, regulation, and security challenges," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3417–3442, 2019.

[51] D. W. Matolak and R. Sun, "Unmanned aircraft systems: Air-ground channel characterization for future applications," *IEEE Veh. Technol. Mag.*, vol. 10, no. 2, pp. 79–85, 2015.

[52] C. Zhang, W. Zhang, W. Wang, L. Yang, and W. Zhang, "Research challenges and opportunities of UAV millimeter-wave communications," *IEEE Wirel. Commun.*, vol. 26, no. 1, pp. 58–62, Feb. 2019.

[53] L. Zhang, H. Zhao, S. Hou, Z. Zhao, H. Xu, X. Wu, Q. Wu, and R. Zhang, "A survey on 5G millimeter wave communications for UAV-assisted wireless networks," *IEEE Access*, vol. 7, pp. 117 460–117 504, 2019.

[54] D. W. Yun and W. C. Lee, "LTE-TDD interference analysis in spatial, time and frequency domain," in *Proc. of Ninth International Conference on Ubiquitous and Future Networks (ICUFN)*, July 2017, pp. 785–787.

[55] 3GPP, "Physical layer procedures for control," 3rd Generation Partnership Project (3GPP), Technical Report (TR) 138.213, 10 2019, version 15.7.0. [Online]. Available: https://www.etsi.org/deliver/etsi_ts/138200_138299/138213/15.07.00_60/ts_138213v150700p.pdf

[56] I. A. Hemadeh, K. Satyanarayana, M. El-Hajjar, and L. Hanzo, "Millimeter-wave communications: Physical channel models, design considerations, antenna constructions, and link-budget," *IEEE Commun. Surveys Tuts.*, vol. 20, no. 2, pp. 870–913, 2018.

[57] Z. Pi and F. Khan, "An introduction to millimeter-wave mobile broadband systems," *IEEE Commun. Mag.*, vol. 49, no. 6, pp. 101–107, 2011.

[58] A. Zappone, E. Jorswieck *et al.*, "Energy efficiency in wireless networks via fractional programming theory," *Found. Trends Commun. Inf. Theory*, vol. 11, no. 3-4, pp. 185–396, 2015.

[59] W. Dinkelbach, "On nonlinear fractional programming," *Manag. Sci.*, vol. 13, no. 7, pp. 492–498, Mar. 1967.

[60] S. Boyd and L. Vandenberghe, *Convex optimization*. Cambridge university press, 2004.

[61] 3GPP, "Physical layer procedures," https://www.etsi.org/, accessed Feb. 28, 2019.

[62] E. Tejaswi and B. Suresh, "Survey of power control schemes for LTE uplink," *Int. Journal Computer Sci. and Inform. Technol*, vol. 10, p. 2, 2013.

[63] Y. S. Nasir and D. Guo, "Multi-agent deep reinforcement learning for dynamic power allocation in wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2239–2250, Oct. 2019.

[64] R. E. Bellman, *Dynamic programming*. Princeton University Press, 1957.

[65] M. Kearns and S. Singh, "Near-optimal reinforcement learning in polynomial time," *Machine Learning*, vol. 49, pp. 209–232, Nov. 2002.

[66] S. S, T. Jaakkola, M. L. Littman, and C. Szepesvari, "Convergence results for single-step on-policy reinforcement-learning algorithms," *Machine Learning*, vol. 38, pp. 287–308, Mar. 2000.

[67] L. Li, *Sample Complexity Bounds of Exploration*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 175–204. [Online]. Available: https://doi.org/10.1007/978-3-642-27645-3_6

[68] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan, "Is Q-learning provably efficient?" *Proc. 32nd International Conference on Neural Information Processing Systems (NIPS)*, pp. 4868–4878, Dec. 2018.

[69] F. R. Yu and Y. He, *Deep Reinforcement Learning for Wireless Networks*. Springer, 2019.

[70] F.-L. Luo, *Machine Learning for Future Wireless Communications*. John Wiley & Sons, 2020.

[71] C. H. Liu, Z. Chen, J. Tang, J. Xu, and C. Piao, "Energy-efficient UAV control for effective and fair communication coverage: A deep reinforcement learning approach," *IEEE J. Sel. Areas Commun.*, vol. 36, no. 9, pp. 2059–2070, 2018.

[72] Y. LeCun, Y. Bengio, and G. Hinton, "Nature," *Deep Learning*, vol. 521, pp. 436—444, May 2015.

[73] T. P. Lillicrap, J. J. Hunt, A. Pritzel, N. Heess, T. Erez, Y. Tassa, D. Silver, and D. Wierstra, "Continuous control with deep reinforcement learning," *arXiv preprint arXiv:1509.02971*, 2015.

[74] M. Tan, "Multi-agent reinforcement learning: Independent vs. cooperative agents," in *Proc. of the Tenth International Conference on Machine Learning*, 1993, pp. 330–337.

[75] C. J., "Understanding the 5G NR physical layer," https://www.keysight.com/upload/cmc_upload/All/Understanding_the_5G_NR_Physical_Layer.pdf, accessed Feb. 28, 2019.

[76] Y. Azar, G. N. Wong, K. Wang, R. Mayzus, J. K. Schulz, H. Zhao, F. Gutierrez, D. Hwang, and T. S. Rappaport, "28 GHz propagation measurements for outdoor cellular communications using steerable beam antennas in New York city," in *Proc. of IEEE International Conference on Communications (ICC)*, Jun. 2013, pp. 5143–5147.