Received 25 September 2022; revised 25 October 2022; accepted 25 October 2022. Date of publication 11 November 2022; date of current version 30 November 2022.

Digital Object Identifier 10.1109/OJCOMS.2022.3219014

Deep Q-Learning-Based Resource Allocation in NOMA Visible Light Communications

AHMED AL HAMMADI[1] (Member, IEEE), LINA BARIAH[2,3] (Senior Member, IEEE), SAMI MUHAIDAT[3,4] (Senior Member, IEEE), MAHMOUD AL-QUTAYRI[1] (Senior Member, IEEE), PASCHALIS C. SOFOTASIOS[3,5] (Senior Member, IEEE), AND MEROUANE DEBBAH[2,6] (Fellow, IEEE)

[1]Department of Electrical Engineering and Computer Science, Khalifa University, Abu Dhabi, UAE

[2]Technology Innovation Institute, Abu Dhabi, UAE

[3]Center for Cyber-Physical Systems, Department of Electrical Engineering and Computer Science, Khalifa University, Abu Dhabi, UAE

[4]Department of Systems and Computer Engineering, Carleton University, Ottawa, ON K1S 5B6, Canada

[5]Department of Electrical Engineering, Tampere University, 33014 Tampere, Finland

[6]Centrale Supelec, University Paris-Saclay, 91192 Gif-sur-Yvette, France

CORRESPONDING AUTHOR: L. BARIAH (e-mail: lina.bariah@ieee.org)

ABSTRACT Visible light communication (VLC) has been introduced as a key enabler for high-data rate wireless services in future wireless communication networks. In addition to this, it was also demonstrated recently that non-orthogonal multiple access (NOMA) can further improve the spectral efficiency of multi-user VLC systems. In this context and owing to the significantly promising potential of artificial intelligence in wireless communications, the present contribution proposes a deep Q-learning (DQL) framework that aims to optimize the performance of an indoor NOMA-VLC downlink network. In particular, we formulate a joint power allocation and LED transmission angle tuning optimization problem, in order to maximize the average sum rate and the average energy efficiency. The obtained results demonstrate that our algorithm offers a noticeable performance enhancement into the NOMA-VLC systems in terms of average sum rate and average energy efficiency, while maintaining the minimum convergence time, particularly for higher number of users. Furthermore, considering a realistic downlink VLC network setup, the simulation results have shown that our algorithm outperforms the genetic algorithm (GA) and the differential evolution (DE) algorithm in terms of average sum rate, and offers considerably less run-time complexity.

INDEX TERMS Deep reinforcement learning, multiple access, resource allocation, sum-rate, visible light communications.

I. INTRODUCTION

THE RAPIDLY emerging services and technologies have paved the way for shaping the vision of future sixth-generation (6G) wireless networks, imposing new challenging constraints relating to system reliability, latency, rate, and energy efficiency. Such constraints are a consequence of the massive increase in the number of connected data-hungry, delay-sensitive wireless devices [1]. Motivated by this, visible light communication (VLC) was identified, among others, as a key enabling technology that is capable of meeting the ever-growing demand for efficient high-rate wireless data services. One of the most attractive advantages of VLC is the abundant visible light spectrum, which is in the order of hundreds of Terahertz. Another distinct characteristic of VLC networks is that they exhibit inherent security, and are immune to electromagnetic interference [2]. Due to spectrum availability, VLC can be adopted in various applications such as healthcare, vehicle-to-vehicle communications, and Internet-of-Things (IoT). The large bandwidth in VLC makes it an attractive technique for realizing efficient and high data-rate IoT connectivity. In an indoor environment, short communication [3] is realized through the use of LEDs.

The short-range communication with LEDs can be accomplished by modulating the intensity of the LED, a process known as intensity modulation (IM). At the receiver side, a photodetector (PD) is used to perform direct detection (DD) by converting the received light intensity fluctuations into an electrical current for data demodulation [2]. Yet, although VLC systems offer a large amount of available bandwidth, existing off-the-shelf LEDs have a restricted bandwidth, limiting the number of served users by an LED. Therefore, to reap the full potential of VLC networks, spectrally-efficient multiple access schemes are required for improved connectivity and the overall quality of service. To that end, one of the emerging multiple access techniques that is capable of improving the spectral efficiency is non-orthogonal multiple access (NOMA) [4]. Through power domain superposition coding at the transmitter and successive interference cancellation (SIC) at the receiver, all users in NOMA systems can utilize the entire modulation bandwidth of the system simultaneously. Thus, NOMA offers higher connectivity and spectrum efficiency (SE) in IoT networks, as compared to the orthogonal frequency division multiple access (OFDMA) scheme. Moreover, it has been shown that NOMA perform considerably better in high signal-to-noise ratio (SNR) scenarios [5], rendering it a prominent candidate for VLC systems that enjoy high SNRs, which are attributed to the relatively short distances between the transmitter and the receiver. The performance of NOMA-enabled VLC systems has been extensively studied [6], [7], [8], [9]. The main challenge of applying NOMA in VLC systems is the non-negative real-valued requirement imposed on VLC signals, rendering current power allocation schemes in RF-NOMA systems inapplicable to VLC scenarios. Accordingly, in our paper, we revisit these schemes and develop a deep Q-learning (DQL) approach in order to obtain optimal resource allocation while considering the VLC channel characteristics. The authors in [6] proposed a gain ratio power allocation (GRPA) method and suggested NOMA as a possible candidate for high-speed VLC systems. The studies in [7] and [8] reported more advanced power allocation methods for NOMA-VLC, at the expense of increased computational complexity. Likewise, to improve the error rate performance of uplink NOMA-VLC systems, the authors in [9] proposed a phase pre-distortion approach.

It is recalled that the typically high energy consumption of connected devices in wireless networks constitutes a fundamental challenge in designing future 6G wireless networks, which are envisioned to enable a wide range of essential but energy-consuming applications [10], [11]. Therefore, it is critical to improve the energy efficiency of future wireless communication systems while maintaining or increasing the desired quality of service (QoS). In this context, it is worthy to mention that the exploitation of superposition modulation in NOMA enables energy-efficient wireless transmission [12], [13]. Thus, in order to ensure a desired quality-of-service (QoS) levels for all superimposed users, several research efforts have been devoted in the efficient design of power allocation mechanisms. To that end, power allocation problems were studied in [14], [15], whereas the joint power allocation and sub-channel assignment problems were investigated in [16], [17], [18], [19].

Joint optimization problems in NOMA have received a considerable attention from the research community. For example, Zhao et al. [20] proposed a joint UAV trajectory and NOMA precoding optimization framework, with the aim to improve the system throughput. In another work, Peng et al. [21] considered a hybrid precoding and power allocation scheme in order to maximize the energy efficiency of mmWave-enabled NOMA UAV networks. Nevertheless, most of the reported contributions in joint resource allocation problems for NOMA-enabled networks are non-deterministic polynomial-time hard (NP-hard) [22], especially when users are mobile. Therefore, it is challenging to obtain an optimal solution due to the high amount of uncertainty and the high computational complexity. As a result, sub-optimal solutions were subsequently proposed in [23], [24], [25]. Heuristic optimization techniques like the genetic algorithm (GA) [26] and the differential evolution (DE) algorithm [27] can solve these NP-hard problems. However, these techniques often fall into a local optimum solution. Hence, using heuristic techniques may limit the performance of NOMA in different scenarios for future wireless networks. Therefore, it is of paramount importance to employ an efficient method for obtaining an optimal power allocation mechanism for VLC networks with uniformly distributed users. With this motivation, in the present contribution we utilize an algorithm based on deep reinforcement learning (DRL), in which an agent in the network continuously learns from the environment and adapts the network parameters accordingly. The proposed algorithm aims to improve the average sum rate of a VLC network, with uniformly distributed users. This also provides an answer to the following question: is it practically feasible to jointly optimize the power allocation and the LED transmission angle of an indoor VLC-NOMA network?

A. RELATED WORKS

Recently, Q-learning sparked an unprecedented interest by researchers and engineers in various fields. Q-learning is a subset of reinforcement learning that relies on Q tables to store the optimal sequence of actions, which maximizes the future reward. In the context of optimizing communication networks, several studies have adopted Q-learning to enhance the performance of wireless networks from different perspectives [25], [28], [29], [30], [31], [32], [33], [34], [35]. In [28], the authors proposed a fast RL-based power allocation scheme to improve the spectral efficiency of a multiple-input multiple-output (MIMO) NOMA system in the presence of a smart jammer interference. The study in [30] used Q-learning to develop a framework for enabling mobile edge computing with NOMA. By incorporating deep learning into RL, DRL addresses a challenge associated with Q-learning in terms of Q table storage and look-up. Based on this, Yang et al. [31] used a deep Q-network (DQN) to

model a multi-user NOMA offloading problem, whereas the authors in [32] designed a power allocation in cache-assisted NOMA systems using DRL. Likewise, Zhang et al. [33] proposed a dynamic power allocation mechanism based on the actor-critic RL, whilst DRL was used in [35] to arrive at sub-optimal power allocation solutions for an uplink multi-carrier NOMA system. Finally, He et al. [25] solved a joint power allocation and channel assignment problem in a two-user NOMA system using a DRL framework.

B. MOTIVATION

The aim of this work is to jointly optimize power allocation and LED transmission angle tuning, with uniformly distributed users in an indoor VLC network. In such a setup, the problem is NP-hard and cannot be tackled using conventional optimization methods. The most significant advantage offered by DQL is its ability to solve complex joint optimization problems in wireless communication, which cannot be solved by conventional mathematical tools [36]. The effectiveness of DQL was demonstrated in several works in the literature. For instance, the authors in [37] optimized an IRS-NOMA system by using DQL to predict and optimally tune the IRS phase shift matrices. In [38], DQL is used to allocate optimal channels to a cluster of users in order to maximize energy efficiency. The DQL algorithm allows the agent to learn about the communication environment and develop new knowledge that can lead to an optimal solution, with imperfect channel state information acquisition. Therefore, in the current contribution, we leverage the DQL algorithm in order to solve this complex problem. The motivation underlying the utilization of DQL in our optimization framework is two-fold:

- The proposed solution can cope with the channel uncertainty and does not require perfect knowledge of channel state information to maximize the average sum rate.
- The solution avoids an exhaustive search method to reach the optimal solution, which searches for all the power allocation coefficients, with all possible LED transmission angles, thus rendering it to an impractical solution.

C. CONTRIBUTIONS

To the best of the authors' knowledge, none of the previous studies proposed a DQL algorithm to maximize the average sum rate of uniformly distributed users in a NOMA VLC indoor network. In this work, we propose an efficient DQL-based algorithm that maximizes the average sum rate by jointly optimizing the power allocation and the transmission angle of the LEDs. The main contributions of this paper can be summarized as follows:

- We formulate a joint optimal power allocation and LED transmission angle tuning problem in the downlink of the considered NOMA-VLC network. The optimization problem aims to maximize the average sum rate of uniformly distributed users, under total power and the individual LED transmission angle constraints.

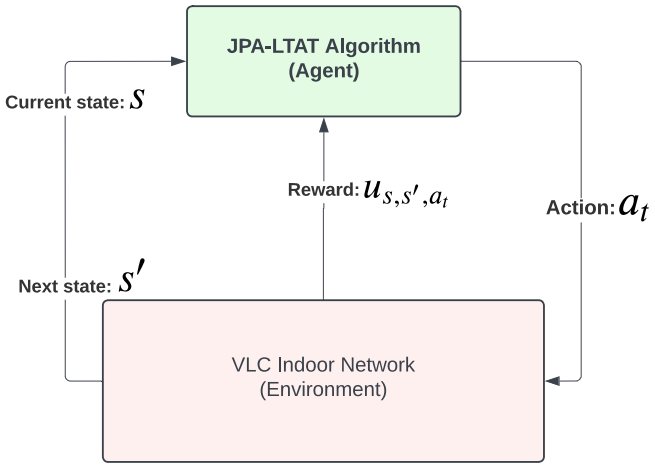


FIGURE 1. The interplay between the agent and the environment in reinforcement learning.

- We propose a joint power allocation and LED transmission angle tuning algorithm to solve the aforementioned non-convex optimization problem by introducing the DQL concept. In particular, we define a reward function that maximizes the sum rate while adhering to the constraints of power and LED transmission angles.
- We conduct a theoretical complexity analysis of the proposed deep Q-learning framework and draw valuable insights on the efficiency of the proposed scheme.
- We validate the superiority of the proposed algorithm over the fixed power allocation policy and the exhaustive search method. The simulation results indicate that after a few iterations, the proposed scheme converges and performs better under varying transmit SNR, cell radius, and VLC Access Point (AP) height.
- The offered results provide useful insights on the achievable performance of the proposed technique, which has a particularly practical importance.

II. INTRODUCTION TO DEEP REINFORCEMENT LEARNING

In this section, we introduce the concept of DRL, which is a special case of reinforcement learning. First, it is recalled that reinforcement learning is a sub-field of machine learning, where an agent interacts with the environment to perform the best series of actions that will maximize the expected future reward in an interactive environment. This interplay between the agent and the environment is depicted in Fig. 1.

In general, RL can be classified as single-agent or multi-agent based on the number of agents. In the case of single agent RL. If the agent can observe the environment's full state information, the sequential decision-making problem can be modeled using the Markov decision process (MDP) framework. On the other hand, multi-agent reinforcement learning is typically modeled as a Markov or random game (a generalized method of traditional parroted game) when two or more agents have complete environment observation,

and make decisions accordingly. Without loss of generality, our underlying framework assumes a single agent for a single VLC access point. In DRL, the best sequence of actions for an agent will be predicted based on a deep neural network. Therefore, the deep neural network in DRL acts as a universal function approximator.

The fundamental elements for RL are:

- *Observations:* These are the continuous measurements of the environment's properties. They are represented in vector $\mathbf{p}$ with $O \in \mathbb{R}_p$, where p denotes the number of the observed properties.
- *States:* The state $s_t \in S$ denotes the discretized observation at time step t.
- *Actions:* An action $a_t \in A$ is one of the valid decisions that the agent can take at time step t.
- *Policy:* A policy denoted by $\pi(.)$ is the mapping between the actions to be taken by the agent at any given state of the environment.
- *Rewards:* The value u_{s,s',a_t} is the reward obtained after an agents takes a particular action a_t in a given state s at time t, which leads to state s'.
- *State-action value:* Denoted by $Q_\pi(s, a)$, and defined as the expected discounted reward when the agent starts at state s and selects action a according to policy π.

At a given time step t, when an agent performs an action a_t, the agent's environment changes from the current state s_t to the following state s'. As a result of this transition, the agent receives an immediate reward u' that represents the outcome of performing action a_t while in state s_t. At time t', this system generates an experience tuple $e' = (s_t, a_t, u', s')$, which is stored in buffer $\mathcal{D}$. Based on this, the main goal of the agent is to maximize the long-term cumulative discounted reward, which is defined as

$$U_t = \sum_{i=0}^{\infty} \gamma^i u_{t+i}, \tag{1}$$

with discount factor $\gamma \in [0, 1]$. To accomplish this, an optimal policy π^* that maps the best actions to states is required. In other words, the optimal policy will act as a guide, informing the agent which actions should be taken at a any given state, in order to maximize the long-term cumulative reward. It is noted that the Q-value function [39] is a function that represents the expected cumulative reward U_t of starting at state s_t, performing action a_t, and following a certain policy π. This function is critical in solving RL problems, and is given by

$$\begin{aligned} Q_\pi(s_t, a_t) &= \mathbb{E}[U_t | s_t, a_t] = \mathbb{E}\left[\sum_{i=0}^{\infty} \gamma^i u_{t+i} | s_t, a_t\right] \\ &= \mathbb{E}\left[u_t + \gamma Q_\pi\left(s', a'\right) | s_t, a_t\right], \end{aligned} \tag{2}$$

where $\mathbb{E}[\,\cdot\,]$ denotes statistical expectation. The optimal π^* that maximizes (1) for all states and actions, also maximizes (2). Consequentially, the optimal Q-value function that follows π^* is obtained using

$$Q_{\pi^*}(s_t, a_t) = \mathbb{E}\left[u_t + \gamma \max_{a'} Q_{\pi^*}\left(s', a'\right) | s_t, a_t\right]. \tag{3}$$

The definition in (3) is known as the Bellman equation [40]. The purpose of the equation is to divide the value function into two components: the immediate reward u_t and the long-term cumulative discounted reward U_t. Rather than summing up over multiple time steps, the definition (3) simplifies the computation of the Q-value function by decomposing it into simpler, recursive sub-problems and determining their optimal solutions. Nevertheless, the Bellman equation in (3) is nonlinear, and hence, there are no closed-form solutions to it. As a result, numerous iterative methods have been proposed (e.g., Q-learning), each of which has been shown to converge to the optimal Q value function [39]. However, these methods become impractical in multi-user systems with a large state or action space, as the size of the Q-value table (e.g., all possible values of (2) for all possible states and actions) is extremely large. The solution to this problem is to estimate the Q value using function approximations, e.g., deep neural networks, which is the core idea of the underlying deep Q-network.

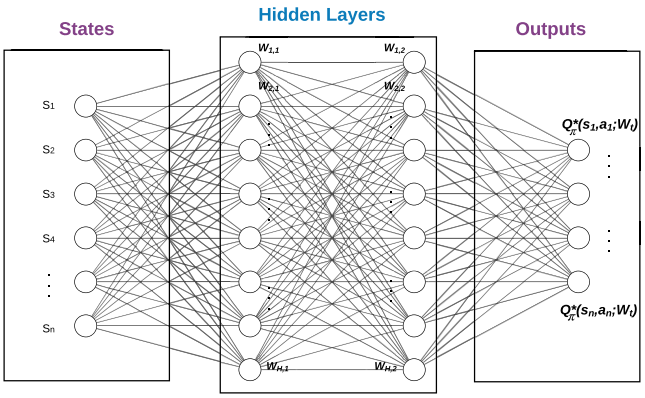


FIGURE 2. The structure of the DQN used with two hidden layers.

The DQN design, shown in Fig. 2, consists of the following three main components:

- The input layer represents the states of the environment.
- The hidden layer acts as a function approximator. In this component, the Rectified Linear Unit (ReLU) activation function is used to compute the hidden layer values. The ReLU function is defined as

$$y_o = \begin{cases} y_i & \text{for} \quad y_i \geq 0 \\ 0 & \text{for} \quad y_i < 0, \end{cases} \tag{4}$$

  where y_o is the output from the activation function while y_i is its input. The main advantage of employing ReLU as an activation function is its computational efficiency, since it does not compute exponentials and divisions [41]. Additionally, ReLU introduces more sparsity in the hidden units, as when $y_i < 0$, the output values become zero [42]. Therefore, the computational

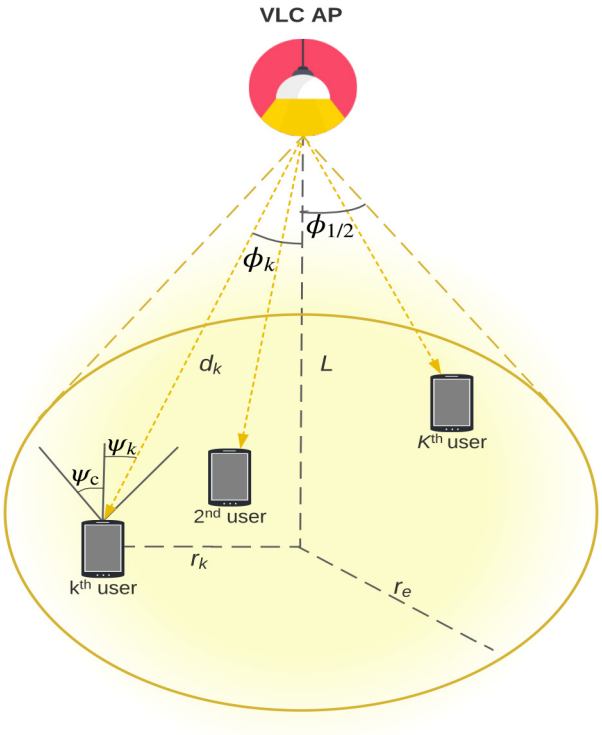


FIGURE 3. System Model.

efficiency and the increased sparsity can lead to faster convergence.

- The output layers represent the predicted state-action value function, $Q^*_\pi(s, a; W_t)$.

III. SYSTEM AND CHANNEL MODELS

Fig. 3 shows the underlying system model of the considered study. We consider a NOMA-VLC indoor network, which consists of a VLC AP installed on a ceiling at height L. The VLC AP serves K users, uniformly distributed over a polar coordinate plane of r radius. Without loss of generality, in this work, we will focus on the downlink communication. Although VLC channels consist of a line-of-sight (LOS) and non-LOS (NLOS) components, this study considers the direct LOS component, due to the fact that the NLOS component has much less energy.

A. VLC CHANNEL MODEL

The signal transmitted by the VLC AP can be expressed as

$$x_i = \sum_{i=1}^{K} \alpha_i \sqrt{P_e} s_i + I_{DC}, \tag{5}$$

where P_e denotes the total electrical transmit power, I_{DC} represents the LED DC bias, which is essential for intensity modulation-based optical baseband transmission, s_i represents the modulated symbol of the i^{th} out of K links, and α_i is the power allocation coefficient for the corresponding link. It is assumed that the transmitted signal for each user follows a uniform distribution with zero mean and unit variance. Based on this and a given total power constraint, the following constraint should hold,

$$\sum_{i=1}^{K} \alpha_i = 1. \tag{6}$$

Furthermore, the optical transmit power of the LED can be expressed as

$$P_{opt} = \eta \mathrm{E}[x] = \eta I_{DC}, \tag{7}$$

where η denote the LED efficiency, which, without loss of generality, is assumed to be normalized to unity. Based on this, the received signal at the k^{th} user can be expressed as

$$y_k = \sqrt{P_e} h_k \left(\sum_{i=1}^{k-1} \alpha_i s_i + \alpha_k s_k + \sum_{i=k+1}^{K} \alpha_i s_i \right) + z_k, \tag{8}$$

where the channel gain h_k is given by [43]

$$h_k = \frac{(m+1) A R_p}{2\pi d_k^2} \cos^m(\phi_k) T(\psi_k) g(\psi_k) \cos(\psi_k), \tag{9}$$

with z_k is the additive white Gaussian noise with zero mean, and variance σ_k^2, A denoting the area of the PD, R_p representing the responsivity of the PD, and d_k is the Euclidean distance between the VLC AP and the k^{th} user. Also, $T(\psi_k)$ and $g(\psi_k)$ denote the optical filter gain and the optical concentrator, respectively. It is also noted that (9) indicates that the channel gain h_k is inversely proportional to the distance of the k^{th} user. As shown in Fig. 3, the light emitted from the LED follows a Lambertian radiation pattern with an order.

$$m = -\frac{1}{\log_2(\cos(\phi_{1/2}))}, \tag{10}$$

where $\phi_{1/2}$ is the transmission angle of the VLC AP, ψ_c denotes the receiver's field of view (FOV), whereas ψ_k and ϕ_k denote the angle of incidence and the angle of irradiance, respectively.

It is recalled that in power-domain NOMA systems, users with stronger channel conditions are allocated lower signal power, whereas users with severe channel conditions are allocated more power, which implies that $\alpha_1 \geq \cdots \geq \alpha_k \geq \cdots \geq \alpha_{K-1} \geq \alpha_K$. Without loss of generality, we assume that the users in the considered setup are sorted in ascending order according to their channels, namely,

$$|h_K| \geq |h_{K-1}| \geq \cdots \geq |h_k| \geq \cdots \geq |h_1|. \tag{11}$$

In order to perform reliable signal detection, the k^{th} user performs SIC in order to cancel the incurred interference experienced from signals with higher power levels. Also, the signals of the users that are allocated with lower power coefficients are treated as noise.

B. IMPERFECT CSI MODEL

Unlike the majority of previous related contributions, which assumed perfect CSI knowledge for the underlying VLC system model, this work makes the practical assumption of imperfect CSI knowledge. CSI is typically obtained at receivers via pilot symbols. The channel coefficients are transmitted to the transmitter via an RF or infrared (IR) uplink, where channel uncertainty increases as uplink and downlink channel noise increases. Additionally, channel uncertainty is increased due to quantization errors introduced by the imperfect digital-to-analog, analog-to-digital conversion processes, which ultimately degrades system performance. It is worth noting that the current analysis uses the same noisy CSI model as [44], which takes into account the resultant CSI error regardless of the source of the error, i.e., location uncertainty, orientation uncertainty, and LED half-angle uncertainty.

The channel coefficient for the VLC link can be modeled by using the minimum mean squared error (MMSE) estimation method, yielding [44]

$$h_k = \hat{h}_k + e_k, \tag{12}$$

where $\hat{h}_k \sim \mathcal{N}(0, 1-\sigma_e^2)$ is the estimated channel gain and e_k denotes the estimated error in the channel which follows a Gaussian distribution with mean $= h_k$ and variance $= \sigma_e^2$. It is worth noting that the random variables $\hat{h}_k$ and e_k are uncorrelated.

C. VLC CHANNEL MODEL OF UNIFORMLY DISTRIBUTED USERS

Without loss of generality, we assume that the users are uniformly distributed within the attocell. This assumption is widely considered as a baseline in several contributions in the literature, e.g., [45], [46], [47], and it can be readily generalized into Poisson or normal distributions. Following this assumption, the relationship between the angle of incidence, the angle of irradiance, the Euclidian distance of the k^{th} user, the height L, and the radical distance r_k is given by

$$\cos(\phi_k) = \cos(\psi_k) = \frac{L}{d_k}, \tag{13}$$

where

$$d_k = \sqrt{r_k^2 + L^2}. \tag{14}$$

Substituting (13) and (14) in (9), the DC gain of the LOS component can be expressed as

$$h_k = \frac{\Xi(m+1)L^{m+1}}{\left(r_e^2 + L^2\right)^{\frac{m+3}{2}}}, \tag{15}$$

where

$$\Xi = \frac{AR_pU(\psi_k)g(\psi_k)}{2\pi} \tag{16}$$

is a constant. Furthermore, given that users are uniformly distributed, the following probability density function (PDF) is used $f_{r_k}(r) = 2r/r_e$. Therefore, the PDF of the corresponding channel gain is given by

$$f_{h_k}(t) = \frac{2\left(\Xi(m+1)L^{m+1}\right)^{\frac{2}{m+3}}}{{r_e}^2(p+3)t^{\frac{2}{m+3}+1}}, \quad t \in [\lambda_{\min}, \lambda_{\max}] \tag{17}$$

where

$$\lambda_{\min} = \frac{\Xi^2(m+1)^2L^{2m+2}}{\left({r_e}^2 + L^2\right)^{m+3}} \tag{18}$$

and

$$\lambda_{\max} = \frac{\Xi^2(m+1)^2L^{2m+2}}{L^{2(m+3)}}. \tag{19}$$

Based on this and in order to obtain the corresponding cumulative distribution function (CDF), we integrate (17) over the range $[\lambda_{\min}, \lambda_{\max}]$, yielding

$$F_{h_k^2}(t) = 1 + \frac{L^2}{r_e^2} - \frac{\left(\Xi(m+1)L^{m+1}\right)^{\frac{2}{m+3}}}{r_e^2 t^{\frac{1}{m+3}}}. \tag{20}$$

With the aid of order statistics [48], the PDF of the ordered channel gain of the k^{th} user denoted by $f'_{h_k}(t)$, can be obtained as

$$f'_{h_k^2}(t) = \frac{K! f_{h_k^2}(t)}{(k-1)!(K-k)!} F_{h_k^2}(t)^{k-1}\left[1 - F_{h_k^2}(t)\right]^{K-k}, \tag{21}$$

which after some algebraic manipulations can be equivalently expressed as follows:

$$f'_{h_k^2}(t) = \frac{\Omega}{m+3}\frac{K! t^{-\frac{1}{m+3}-1}}{(k-1)!(K-k)!} \times \left(\frac{\Omega}{t^{\frac{1}{m+3}}} - \frac{L^2}{{r_e}^2}\right)^{K-k}\left(1 - \frac{\Omega}{t^{\frac{1}{m+3}}} + \frac{L^2}{r_e^2}\right)^{k-1} \tag{22}$$

where the constant $\Omega = \frac{1}{r_e^2}(C(m+1)L^{m+1})^{\frac{2}{m+3}}$.

Note that the PDF in (22) has a convenient mathematical form as it consists of only elementary functions, which renders it tractable both analytically and computationally.

D. AVERAGE SUM RATE OF NOMA-VLC WITH UNIFORMLY DISTRIBUTED USERS

Following [43], the average sum-rate of NOMA VLC under imperfect CSI can be expressed as [1]

$$R_k^{\text{VLC}} = \begin{cases} \log_2\left(1 + \frac{\hat{h}_k^2\alpha_k^2}{\sum_{i=k+1}^{K}\hat{h}_k^2\alpha_i^2 + \frac{1}{\rho} + \sigma_e^2}\right), & k = 1, \ldots, K-1 \\ \log_2\left(1 + \rho\hat{h}_k^2\alpha_k^2 + \sigma_e^2\right), & k = K \end{cases} \tag{23}$$

1. It should be noted that Shannon's capacity equations is valid for VLC systems, if the transmitted signal is frequency-upshifted. Wherein, the real-valued baseband transmission signal model for VLC can be converted into a complex-valued baseband channel by applying a frequency-upshift to an intermediate frequency (IF), which has a slightly higher center frequency than half the bandwidth of the transmitted signal before applying the bias current.

where $\rho = P_e/\sigma_k^2$ denotes average transmit SNR. It is worth noting that (23) is derived under the assumption of perfect SIC process. In addition, the decoding order is assumed to be fixed and known to the receivers.

For a K number of uniformly distributed users and an arbitrary power allocation strategy, the average sum rate of NOMA-VLC is given at the bottom of the page [49]. The average sum rate in (24) is expressed in bits per second (bits/s), whereas the average energy efficiency metric is expressed in terms of bit per joule, and can be calculated using

$$\xi = \frac{\hat{R}_{\text{VLC}}^{\text{NOMA}}}{Q_{\text{VLC}}}, \tag{25}$$

where Q_{VLC} represents fixed power consumption of the VLC AP, expressed in watts.

IV. PROBLEM FORMULATION

The main objective of this work is to perform a joint power allocation and LED transmission angle $\phi_{1/2}$ tuning optimization, with the aim to maximize the average sum rate of uniformly distributed users. Accordingly, the joint optimization problem is formulated as

$$\max_{\substack{\alpha_k \\ \phi_{1/2}}} \quad \hat{R}_{\text{VLC}}^{\text{NOMA}} \tag{P1}$$

$$s.t. \quad P_e \leq P_{max}, \tag{P1.a}$$

$$\sum_{k=1}^{K} \alpha_k = 1, \forall k \in K, \tag{P1.b}$$

$$30^\circ \leq \phi_{1/2} \leq 70^\circ. \tag{P1.c}$$

where the first constraint (P1.a) refers to the maximum allowed transmission power, the second constraint (P1.b) is set to ensure that the total transmit power of the superimposed signal equals to P_e. The final constraint (P1.c) aims to ensure that the selected LED transmission angles fall within a practical range. Due to the high computational complexity and the varying nature of the channels, it is challenging to obtain a global optimum solution to (P1).

To solve the above optimization problem, two approaches can be considered. The first approach is the simplest in terms of implementation, in which the use of a fixed power allocation policy, and a fixed LED transmission angle is considered. However, such an approach results in a sub-optimal solution. The other approach is the exhaustive search, which can lead to an optimal solution; however, this comes at the expense of increased complexity. In the following sections, we introduce DRL as an alternative approach to solve the underlying optimization problem. In the next section, we will demonstrate the proposed DRL-based optimization framework for joint power allocation and LED transmission tuning.

V. JOINT POWER ALLOCATION AND LED TRANSMISSION ANGLE TUNING (JPA-LTAT): DRL-BASED FRAMEWORK

In what follows, we propose a DRL-based framework to solve the optimization problem (P1). First, we will present how the DQN is trained with an appropriate policy selection criterion. Then we introduce an algorithm that relies on the DRL framework to achieve optimal performance.

A. TRAINING PHASE

The DQN is trained and updated to approximate the action-value function of $Q_{\pi^*}(s, a)$. It is recalled that the experience tuple is defined as $e_t = (s_t, a_t, u_t, s')$. The agent saves its experiences in a buffer $\mathcal{D} = \{e_1 \;\; e_2 \;\; \ldots \;\; e_t\}$ that is used to train the DQN using the gradient descent algorithm [7].

While it is ideal for DQN training to use all data in each iteration, this is prohibitively expensive when the training set is large. A more efficient method is to evaluate the gradients in each iteration using a random subset of the replay buffer $\mathcal{D}$, referred to as mini-batch. Accordingly, the loss function is defined as follows

$$\mathcal{L}(\boldsymbol{W}) = \sum_{e \in D} \left(\underbrace{u + \gamma \max_{a'} Q_{\pi^*}\left(s', a', \hat{\boldsymbol{W}}\right)}_{\text{target}} - Q_{\pi^*}(s, a, \boldsymbol{W}) \right)^2, \tag{26}$$

where (26) denotes the DQN's loss function for a random mini-batch $\mathcal{D}$ at time slot t and $\hat{\boldsymbol{W}}$ denotes the quasi-static target parameters that are updated every t time slots. Finally, the optimal weights are obtained using

$$\boldsymbol{W}^* = \arg\min_{\boldsymbol{W}} \mathcal{L}(\boldsymbol{W}). \tag{27}$$

In order to minimize the loss function defined in (26), the weights of the DQN are updated at every time step t using a stochastic gradient descent (SGD) algorithm on a mini-batch sampled from the replay buffer $\mathcal{D}$. To this effect, the SGD algorithm will update the weights $\boldsymbol{W}$ in an iterative process with a learning rate of $\mu > 0$ as follows [50]

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \mu \bigtriangledown \mathcal{L}_t(\boldsymbol{W}_t). \tag{28}$$

$$\hat{R}_{\text{VLC}}^{\text{NOMA}} = \frac{\Xi K}{\ln(2)(m+3)} \left\{ \sum_{l=0}^{K-1} \frac{(K-1)!(-\Xi)^l}{l!\,(K-1-l)!(v_1+1)} \left(\frac{L^2}{r_{\text{e}}^2}+1\right)^{K-1-l} [\Omega(\lambda_{\max}, v_1, b_1) - \Omega(\lambda_{\min}, v_1, b_1)] \right.$$
$$\left. + \sum_{k=1}^{K-1} \sum_{p=0}^{k-1} \sum_{q=0}^{K-k} \frac{(K)!(\Xi)^{p+q}(-1)^{p+K-k-q}}{p!\,(k-1-p)q!(K-k-q)!} \left(\frac{L^2}{r_{\text{e}}^2}+1\right)^{k-1-p} \left(\frac{L^2}{r_{\text{e}}^2}\right)^{K-k-q} [\Omega(\lambda_{\max}, v_2, b_2) - \Omega(\lambda_{\min}, v_2, b_2)] \right\}. \tag{24}$$

B. POLICY SELECTION

Generally speaking, Q-Learning is considered as an off-policy algorithm, which means without actually following any greedy policy, it estimates the reward for future actions and adds a value to the new state [51]. Based on this, we consider a near-greedy action selection policy. The near-greedy policy has two modes:

1) *Exploration:* The agent tries different actions at every time step t to discover an effective action a_t.
2) *Exploitation:* The agent chooses an action at time step t that maximizes the state-action value function $Q_\pi(s, a; \boldsymbol{W}_t)$ based on the previous experience.

In the near-greedy policy, the agent has an exploration rate of ϵ and an exploitation rate of 1 - ϵ, where $0 < \epsilon < 1$, and ϵ is a hyper-parameter that controls the trade-off between exploitation rate and exploration rate of the agent. For every time step t, the agent performs a specific action a_t at a given current state s_t. Accordingly, the agent receives a positive or negative reward $u_{s,s',a}[t]$ and moves into a target state $s' := s_{t+1}$.

The period of time in which the agent interacts with the environment is called an episode, where each episode has a total duration time of T time steps. The convergence of an episode is governed by the target objective being fulfilled. Also, the dimension of the input layer is set equal to the number of the states in $\boldsymbol{S}$, the dimension of the output layer is equal to the number of possible actions $\boldsymbol{A}$. For the hidden layer, we choose a smaller depth, as it has a considerable impact on the computational complexity. Therefore, we opted for a depth that offers a reasonable balance between performance and computational complexity.

C. PROPOSED ALGORITHM

In this subsection, we propose the joint power allocation and LED transmission angle tuning (JPA-LTAT) algorithm; an optimization framework based on DRL. The JPA-LTAT algorithm optimizes the average sum rate of the VLC system, assuming that the CSI of each user is unknown. At each time step t, the algorithm calculates the average sum rate of NOMA users in the considered VLC network, which is given in (24). In what follows, we provide some details on the action space, state space, and the reward function.

1) STATE SPACE

All possible states form the state space, denoted as $\boldsymbol{\mathcal{S}}$, which are characterized by power allocation coefficients of each user in the VLC network.

In this paper, the state space $\boldsymbol{\mathcal{S}}$ contains the power allocation coefficients of each user in the VLC network and the LED transmission angle of the VLC AP. Accordingly, the resultant state space $\boldsymbol{\mathcal{S}} = \{\alpha_1 \ \alpha_2 \ \ldots \ \alpha_K \ \phi_{1/2}\}$. For instance, assuming an initial equal power allocation for 4 users, the initial state space for $K = 4$ users and $M = 1$ VLC access point of 45° LED transmission angle is $\boldsymbol{\mathcal{S}} = \begin{Bmatrix} \alpha_1 = 0.25 & \alpha_2 = 0.25 & \alpha_3 = 0.25 & \alpha_4 = 0.25 \\ \phi_{0.5} = 45^\circ \end{Bmatrix}$

TABLE 1. VLC network parameters.

Parameter	Value
Vertical separation between the LED and PDs, L	3 m
Cell radius, r_e	4 m
LED fixed transmission angle	45°
Total signal power, P_e	0.25 W
PD FOV, Ψ_{fov}	60°
PD responsivity R_p	0.4 A/W
PD detection area, A	1 cm^2
Reflective index, n	1.5
Optical filter gain, T	1
Signal bandwidth, B	20 MHz
Noise PSD, N_0	20^{-21} A^2 /Hz

2) ACTION SPACE

All the actions can be taken by the agent from the action space, denoted as $\boldsymbol{\mathcal{A}}$. The possible actions in the action space $\boldsymbol{\mathcal{A}}$ are:

- Increase / Decrease power allocation factor of user k by a step size of Λ_k, where Λ_k is a fixed value to be added to (or subtracted from) each α_k where $k \in K$, while maintaining a unity sum such that $\sum_{k=1}^{K} a_k = 1, \forall k \in K$.
- Increase / Decrease the LED transmission angle of the VLC AP by step size ι_m, where ι_m is a fixed value to be added to (or subtracted from) the value of the LED transmission angle of the m^{th} VLC AP, such that the LED transmission angle is $30^\circ \leq \phi_{1/2} \geq 70^\circ$.

The total number of actions in the action space $\boldsymbol{\mathcal{A}}$ are calculated using $|A| = 2M + 2K$.

3) REWARD FUNCTION

The reward function plays an essential role in the RL algorithm. We use the average sum rate of the VLC-NOMA system, which is calculated using (24), to represent the immediate reward u_t returned after choosing action a_t in state s_t.

Having described the State Space, Action Space, and the Reward Function. In the following, we describe in detail the operational steps of the JPA-LTAT algorithm. Algorithm 1 further summarizes the JPA-LTAT algorithm.

1) The VLC network environment is initialized according to Table 1. The DRL hyper-parameters are initialized as in Table 2. The policy network weights $\boldsymbol{W}_t$ are randomly initialized.
2) The power allocation coefficients are reset to their initial values at the start of each episode to improve the learning experience. Similarly, the LED transmission angle is also reset to the initial value of 45°.

Algorithm 1: JPA-LTAT Algorithm

Input: The average sum rate of the VLC network.
Output: The optimal power allocation coefficients of each user, and the optimal LED transmission angle.

Initialize time, actions, states, and replay buffer $\mathcal{D}$
while *No convergence or Not aborted* **do**
  while $t < T$ **do**
    t:= t+1
    Observe current state s_t
    $\epsilon = \max(\epsilon.\text{d}, \epsilon_{\min})$
    Sample $\tau \sim$ Uniform(0,1)
    if $\tau \leq \epsilon$ **then**
      Select a random action a_t
    else
      Select an action based on a_t=arg max $Q_\pi(s, a; \boldsymbol{W}_t)$
    if $30° \leq \phi_{1/2} \geq 70°$ **then**
      Abort episode.
    Compute the average sum rate based on (24).
    Store experience $e_t = (s_t, a_t, u_{s,s'}, a_t, s')$ in $\mathcal{D}$.
    Minibatch sample from $\mathcal{D}$, $e_j = (s_j, a_j, u_j, s_{j+1})$.
    Set $y_j := u_j + \gamma \max_{a'} Q_{\pi^*}(s_{j+1}, a'; \boldsymbol{W}_t)$.
    Obtain the optimal weights $\boldsymbol{W}^\star$ by performing SGD on $((y_j - Q_{\pi^*}(s_j, a_j, \boldsymbol{W}_t))^2$
    Update $\boldsymbol{W}_t := \boldsymbol{W}^\star$ in the DQN.
    Record the Loss $\mathcal{L}_t$.
    Update $s_t := s'$.

TABLE 2. Deep Q-learning hyper-parameters.

Parameter	Value
Discount factor, γ	0.995
Power allocation step size, Λ	0.01
LED transmission angle step size, ι	5°
Initial exploration rate, ϵ	1.000
Number of states	8
Deep Q-network width	24
Exploration decay rate, κ	0.9995
Minimum exploration rate, $\epsilon_{\min}$	0.1
Number of actions	$2 + (2K)$
DQN depth	2

3) JPA-LTAT uses the ϵ-greedy algorithm to select an action from the action space for a given state in our time-sequential decision process.
4) To allow the exploration of the action space, τ is randomly sampled from a uniform distribution.
   a) If the sampled value is less than or equal to the value of ϵ, then the agent takes a random action.
   b) Otherwise, the agent will select an action based on the learned policy a_t=arg max $Q_\pi(s, a; \boldsymbol{W}_t)$, which aims to maximize the cumulative future reward.
5) In order to maximize the Q-value, which is constructed from the policy network outputs, the agent observes the next state and performs the following set of possible actions:
   a) Increase or decrease the power allocation factor α_k by step size Λ_k, $\forall k \in K$ for each user in the VLC network.
   b) Increase or decrease the LED transmission angle of the VLC AP $\phi_{1/2}$, by step size ι.
6) Following (24), compute the average sum rate for the new set of power allocation factors and the newly modified LED transmission angle and store it as a reward u_t.
7) If the agent tries to exceed the constraint of the LED transmission angle, outside the specified range $30° \leq \phi_{1/2} \geq 70°$, abort the episode.
8) Following that, s_t, s', a_t, and u_t are stored in the replay memory buffer $\mathcal{D}$, which has a capacity of $\mathcal{M}$.
9) Using the gradient descent algorithm with a learning rate μ, a mini-batch is sampled from the buffer and is used to train the policy network to minimize the loss function, which is given by (26).
10) The resulted loss $\mathcal{L}(\boldsymbol{W})$ at time step t is recorded and the next state s' is updated as current state s_t.

D. COMPLEXITY ANALYSIS OF THE PROPOSED ALGORITHM

It is crucial to quantify computational complexity of the proposed algorithm. However, since deep learning algorithms are dependent on hyperparameters, applying analytical methodologies to guarantee the convergence of the proposed DQL-based method is difficult. This is a common challenge in the literature for analytically proving optimality and convergence [52], [53], [54], [55]. Therefore, instead of convergence, we are presenting the following theorem that shows the amount of work per iteration in **Algorithm 1**.

Theorem 1: For an indoor NOMA-VLC system with K users and M access points, the computational complexity of the proposed **Algorithm 1** is given by:

$$\mathcal{O}\Big(\Big(2M^2 + 2K^2 + 4MK\Big) \times \mathcal{H} + \mathcal{C}_1(K)\Big), \tag{29}$$

Proof: First, the DQL agent observes the state of the system, executes the most valuable action, and calculates the reward based on (24). Assuming that the computational complexity of calculating the reward is

$$\mathcal{C}_1(K), \tag{30}$$

where $\mathcal{C}_1$ is directly proportional with K. Second, it is known that the size of the state space and the size of the action space have a significant role in the complexity of the deep Q-learning algorithm. Following [56], the computational complexity of the Q-learning algorithm with the greedy policy is estimated to be $\mathcal{O}(\mathcal{S} \times \mathcal{A} \times \mathcal{H})$ each iteration, where $\mathcal{S}$ is the number of states, $\mathcal{A}$ is the number of actions, and $\mathcal{H}$ is the number of steps per episode. It is recalled that

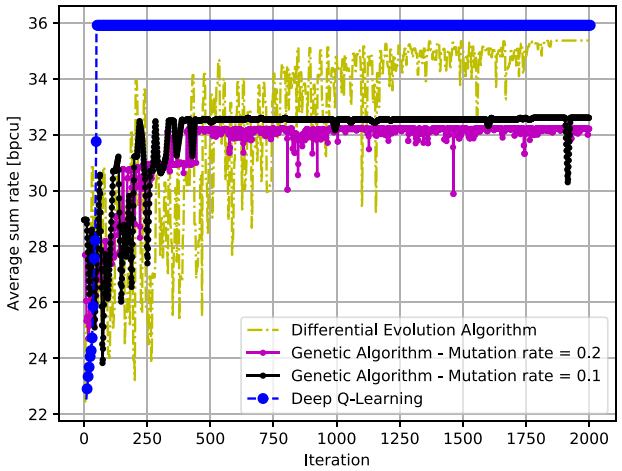


FIGURE 4. Convergence comparison between DQL, GA, and DE for $K = 4$ users.

the size of the state space is $K + M$, and the size of the action space is $2K + 2M$. Therefore, the amount of work per iteration is

$$\mathcal{O}\Big(\Big(2M^2 + 2K^2 + 4MK\Big) \times \mathcal{H}\Big). \tag{31}$$

Based on this and by incorporating (30) into (31), equation (29) is deduced, which completes the proof. ■

E. FIXED POWER ALLOCATION

FPA is considered as one of the simplest power allocation schemes. In this scheme, the allocated power among the users is predefined and fixed according to the following,

$$P_k = \alpha P_{k-1}, \tag{32}$$

where α is the fixed power allocation factor. It is worth noting that FPA yields a complexity of $\mathcal{O}(1)$; however, it does not yield optimal or near-optimal performance.

VI. ACHIEVED RESULTS AND DISCUSSION

This section discusses and analyzes the performance of the proposed DQL-based algorithm, which maximizes the average sum rate of the NOMA-VLC indoor network. Without loss of generality, we assume K users, uniformly distributed in an indoor environment, with a room size of 4×4 meters and a height of 3 meters. The room has a single VLC AP in the ceiling, with a fixed power consumption of 4 Watt and 1 Watt/Amps conversion efficiency. The rest of the simulation parameters are summarized in Table 1. The DQL Algorithm was realized and trained on a PC equipped with Nvidia GPU 2080Ti and an 18-core 2.6GHz processor. Note that we have developed our framework using Python and TensorFlow library [57]. The Deep Q-Learning hyper-parameters are shown in Table 2.

Fig. 4 shows the convergence performance comparison between the proposed DQL-based algorithm, the GA, and the DE algorithm. The settings for the GA is as follows: the number of bits per variable is 8, the population size is 20, crossover rate is 0.9, and we chose two typical mutation rates of 0.1 and 0.2. It worth mentioning that each algorithm has a different execution time per iteration, which is shown in Table 3. To begin with, the proposed algorithm converges after 48 iterations with a maximum average sum rate of 35.9 bpcu. Notably, the convergence rate is faster than the two baseline schemes. For example, the GA with mutation rate = 0.1, takes approximately 478 iterations for convergence. On the other hand, the GA with mutation rate = 0.2 converges after 481 iterations, which is similar with the case of mutation rate = 0.1. The DE algorithm converges after 1787 iterations, which is the highest amongst all the techniques. The rapid convergence of the proposed algorithm is partly attributed to the fact that the DQL algorithm can leverage the GPU cores in order to parallelize the operations. s for the average sum rate performance, the proposed algorithm achieves a maximum average sum rate of 35.9 bpcu, which outperforms both baselines. The DE algorithm achieves a maximum average sum rate of 35.5 bpcu, which outperforms the GA with both mutation rates. The GA with the lower mutation rate achieves 32.5 bpcu, which is slightly better than GA with higher mutation rate that achieves an average sum rate of 32.1 bpcu.

TABLE 3. Time per iteration for each algorithm.

Algorithm	Time (seconds)
Deep Q-learning (JPA-LTAT) Algorithm	1.62
Differential Evolution Algorithm	1.68
Genetic Algorithm	1.89

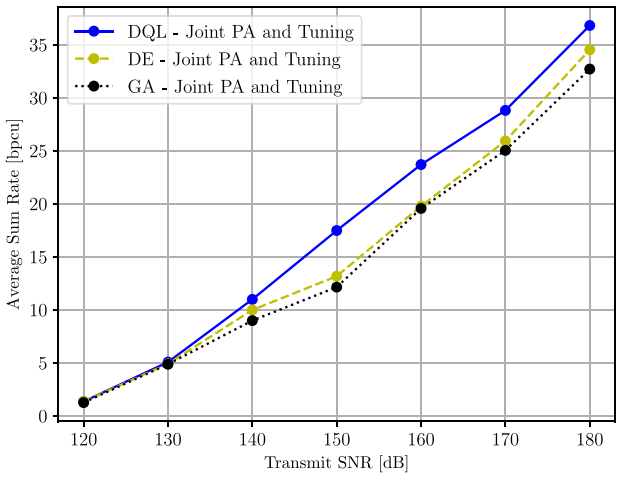


FIGURE 5. Average sum rate vs. transmit SNR for $K = 4$ for NOMA, using DQL, GA, and DE using NOMA.

Fig. 5 shows the average sum rate vs the transmit SNR, where we compare the proposed DQL-based algorithm, the GA with mutation rate of 0.1, and the DE algorithm for $K = 4$ users. It can be shown that the proposed algorithm outperforms both baselines (GA and DE) in the medium to high SNR values. When the SNR $\leq$ 130 dB, all algorithms achieve nearly the same average sum rate. The divergence

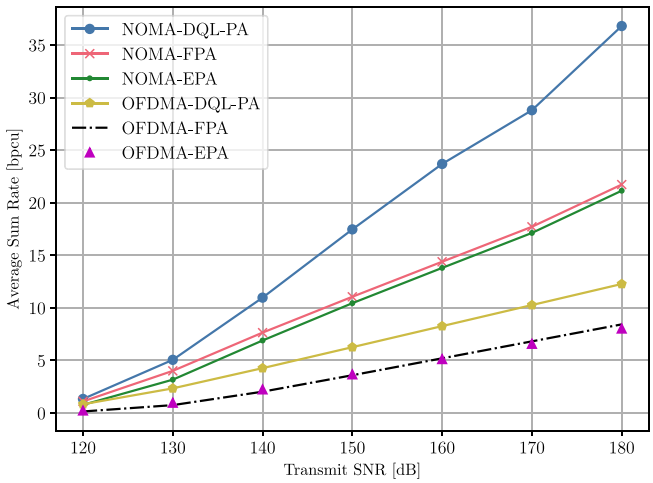


FIGURE 6. Average sum rate vs. transmit SNR for $K = 4$, using DQL-PA, EPA, and FPA for NOMA and OFDMA.

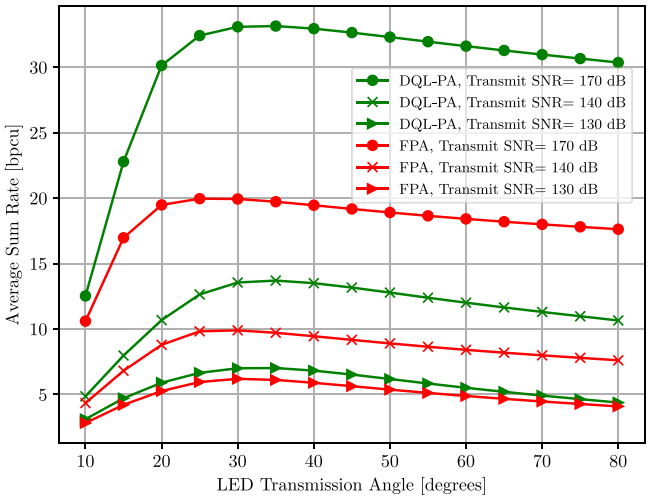


FIGURE 7. Average sum rate vs. LED transmission angle $\phi_{1/2}$ for $K = 4$, using DQL-PA and FPA.

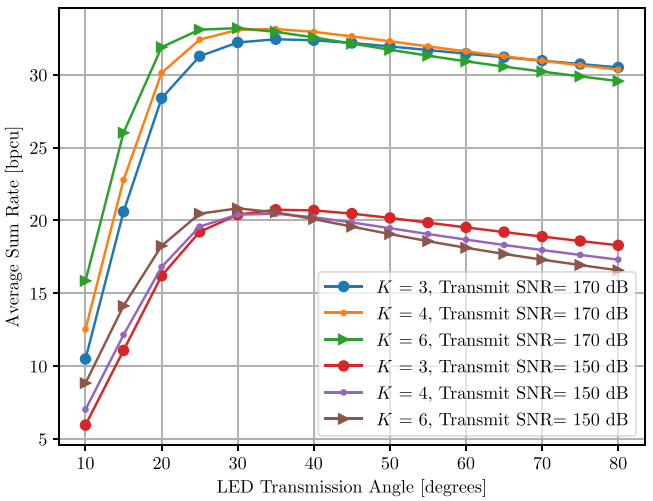


FIGURE 8. Average sum rate vs. LED transmission angle $\phi_{1/2}$ using DQL-PA, with different K and transmit SNR values.

between the curves begin when the SNR = 140 dB, where the proposed algorithm outperforms both the GA and DE baselines. As the SNR approaches 150 dB, the proposed algorithm yields an average sum rate of 17.8 bpcu, which is around 33% more than the average sum rate achieved by DE, and 47% more than average sum rate of the GA. Finally, it can be deduced that the DE outperforms the GA in the medium to high SNR range. However, the difference between the DE and GA fluctuates as the SNR increases.

Fig. 6 depicts the average sum rate as a function of SNR for equal power allocation (EPA), FPA, and DQL-based power allocation for both NOMA and OFDMA as a benchmark solution, with $K = 4$ users. In the case of NOMA, it can be shown that our algorithm outperforms both techniques in the entire SNR range. Moreover, it can be further observed that as the SNR increases, the performance gap between our algorithm and the other two methods, FPA and EPA, becomes more substantial. For instance, at SNR = 150 dB, NOMA-DQL-PA yields an average sum rate of 17.1 bpcu, compared to 10.2 bpcu and 11 bpcu achieved by NOMA-EPA and NOMA-FPA, respectively. For the case of SNR = 180 dB, NOMA-FPA and NOMA-EPA techniques yield an average sum rate of 21 bpcu and 22 bpcu, respectively, whereas DQL-PA achieves an average sum rate of 36 bpcu, which is approximately 71% higher than the NOMA-EPA and NOMA-FPA techniques. For the OFDMA counterpart, it can be seen that even with OFDMA technique, the proposed algorithm offers a noticeable enhancement over FPA and EPA.For instance, when the SNR = 180 dB, the proposed algorithm achieves 12 bpcu, compared to 8 bpcu in FPA and EPA. Finally, it can be seen that NOMA-based techniques outperform OFDMA-based techniques in terms of the average sum rate. This is expected since in NOMA, each user utilizes the entire bandwidth, whereas OFDMA divides the bandwidth between the 4 users.

In Fig. 7, we compare FPA and DQL-PA algorithms in terms of average sum rate as a function of LED transmission angle $\phi_{1/2}$, with $K = 4$. It can be shown that our algorithm outperforms FPA over the entire LED transmission angle range. More specifically, the performance gap between the two techniques increases as the transmit SNR increases. Furthermore, the LED transmission angle's impact on the performance follows a similar pattern in both techniques. Therefore, it becomes evident by Fig. 8 that there is an optimal LED transmission angle, which is both unique and significant.

In Fig. 9, the average sum rate is shown versus the LED transmission angle $\phi_{1/2}$, for a different number of users K, using the DQL-PA algorithm. Similar to Fig. 6, we observe that the number of users K plays a vital role in defining the optimal LED transmission angle. More specifically, for the case of SNR = 170 dB, the optimal transmission angle for $K = 3$ is 35°, whereas, for $K = 6$, the optimal transmission angle is around 30°. Interestingly, the optimal angle tends to decrease as the number of users gets higher. This

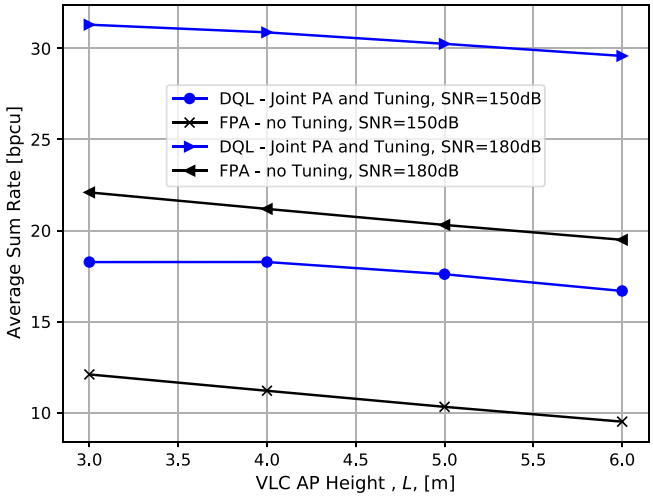


FIGURE 9. Average sum rate vs.VLC AP height *L*, using DQL-PA with tuning, and FPA without tuning.

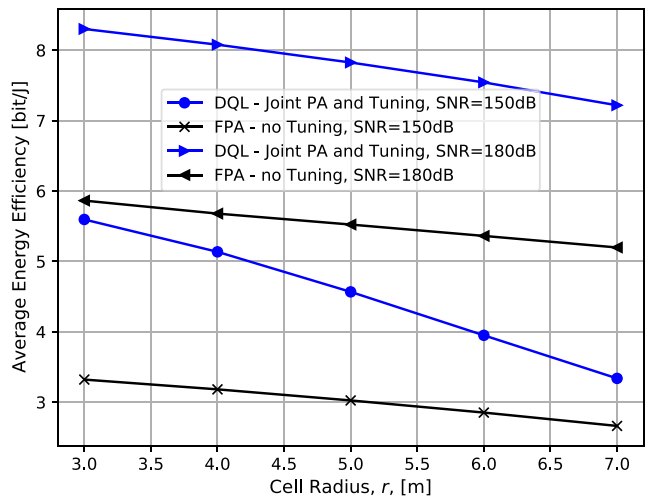


FIGURE 10. Average energy efficiency vs. cell radius *r*, using DQL-PA with tuning, and FPA without tuning.

phenomenon is analogous to water-filling power allocation techniques [58], in which the strong users are allocated more power, and conversely, weak users are allocated less power. Also, the fact that there is a unique optimal LED transmission angle for each K necessitates the need for jointly optimizing the power allocation and LED transmission angle using the DQL technique.

Fig. 9 shows the average sum rate as a function of the VLC AP vertical length L, using DQL-PA with tuning, and FPA with fixed LED transmission angle, with five users. In this scenario, the impact of the channel symmetry dilemma in VLC is investigated. As the vertical distance becomes large, the channel symmetry becomes worse. At SNR = 180 dB, our DQL-PA with tuning outperforms the FPA approach with no tuning by 65% to 70%. Even at the worst channel symmetry conditions for DQL-PA with tuning, the average sum rate is 29.5 bpcu, which is still higher than the best-case scenario for the FPA with no tuning, which is 19.2 bpcu. This shows that our proposed framework outperforms the other benchmark method of FPA with no tuning, even with varying channel symmetry.

Finally, Fig. 10 demonstrates the average energy efficiency as a function of the cell radius r, using DQL-PA with tuning and FPA with a fixed LED transmission angle. This is an important metric since it can quantify how much energy we expect to save from the use of our approach compared to the conventional scheme. It is shown that DQL-PA with tuning outperforms FPA with no tuning, even after varying the distances between the users from 3 to 7 meters. For instance, the average energy efficiency of DQL-PA with tuning at $r = 7$ and SNR = 180 dB is 7.28 b/J, compared to 5.24 b/J in the case of FPA with no tuning. Moreover, DQL-PA with tuning in the case of $r = 7$ meters outperforms the FPA with no tuning in the case of $r = 3$ meters by 21%.

VII. CONCLUSION

In this work, we proposed an algorithm to maximize the average sum rate and average energy efficiency in an indoor NOMA-VLC network. We leveraged the DRL algorithm to train an agent, in order to obtain an optimal power allocation policy for the users. Jointly with the power allocation, the agent can select the optimal LED transmission angle at the VLC AP. To this effect, the obtained results demonstrated that our algorithm outperforms the GA and the DE in terms of average sum rate, and offers considerably less run-time complexity. It was also shown that the joint optimization of the power allocation and the LED transmission angle is more effective as the number of users increases compared to the sole optimal power allocation approach.

REFERENCES

[1] L. Bariah et al., "A prospective look: Key enabling technologies, applications and open research topics in 6G networks," *IEEE Access*, vol. 8, pp. 174792–174820, 2020.
[2] H. Chang et al., "A 100-Gb/s multiple-input multiple-output visible laser light communication system," *J. Lightw. Technol.*, vol. 32, no. 24, pp. 4121–4127, Dec. 15, 2014.
[3] A. Memedi and F. Dressler, "Vehicular visible light communications: A survey," *IEEE Commun. Surveys Tuts.*, vol. 23, no. 1, pp. 161–181, 1st Quart., 2021.
[4] L. Dai, B. Wang, Y. Yuan, S. Han, I. Chih-Lin, and Z. Wang, "Non-orthogonal multiple access for 5G: Solutions, challenges, opportunities, and future research trends," *IEEE Commun. Mag.*, vol. 53, no. 9, pp. 74–81, Sep. 2015.
[5] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5G systems with randomly deployed users," *IEEE Signal Process. Lett.*, vol. 21, no. 12, pp. 1501–1505, Dec. 2014.
[6] H. Marshoud, V. M. Kapinas, G. K. Karagiannidis, and S. Muhaidat, "Non-orthogonal multiple access for visible light communications," *IEEE Photon. Technol. Lett.*, vol. 28, no. 1, pp. 51–54, Jan. 1, 2016.
[7] X. Zhang, Q. Gao, C. Gong, and Z. Xu, "User grouping and power allocation for NOMA visible light communication multi-cell networks," *IEEE Commun. Lett.*, vol. 21, no. 4, pp. 777–780, Apr. 2017.
[8] L. Yin, W. O. Popoola, X. Wu, and H. Haas, "Performance evaluation of non-orthogonal multiple access in visible light communication," *IEEE Trans. Commun.*, vol. 64, no. 12, pp. 5162–5175, Dec. 2016.
[9] X. Guan, Q. Yang, Y. Hong, and C. C.-K. Chan, "Non-orthogonal multiple access with phase pre-distortion in visible light communication," *Opt. Exp.*, vol. 24, no. 22, pp. 25816–25823, Oct. 2016.

[10] H. Zhang, F. Fang, J. Cheng, K. Long, W. Wang, and V. C. M. Leung, "Energy-efficient resource allocation in NOMA heterogeneous networks," *IEEE Wireless Commun.*, vol. 25, no. 2, pp. 48–53, Apr. 2018.

[11] J. Tang et al., "Energy efficiency optimization for NOMA with SWIPT," *IEEE J. Sel. Topics Signal Process.*, vol. 13, no. 3, pp. 452–466, Jun. 2019.

[12] A. E. Mostafa, Y. Zhou, and V. W. S. Wong, "Connection density maximization of narrowband IoT systems with NOMA," *IEEE Trans. Wireless Commun.*, vol. 18, no. 10, pp. 4708–4722, Oct. 2019.

[13] A. Shahini and N. Ansari, "NOMA aided narrowband IoT for machine type communications with user clustering," *IEEE Internet Things J.*, vol. 6, no. 4, pp. 7183–7191, Aug. 2019.

[14] Y. Zhang, H.-M. Wang, T.-X. Zheng, and Q. Yang, "Energy-efficient transmission design in non-orthogonal multiple access," *IEEE Trans. Veh. Technol.*, vol. 66, no. 3, pp. 2852–2857, Mar. 2017.

[15] T. A. Zewde and M. C. Gursoy, "NOMA-based energy-efficient wireless powered communications," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 3, pp. 679–692, Sep. 2018.

[16] J. Shi, W. Yu, Q. Ni, W. Liang, Z. Li, and P. Xiao, "Energy efficient resource allocation in hybrid non-orthogonal multiple access systems," *IEEE Trans. Commun.*, vol. 67, no. 5, pp. 3496–3511, May 2019.

[17] H. Zhang et al., "Energy efficient dynamic resource optimization in NOMA system," *IEEE Trans. Wireless Commun.*, vol. 17, no. 9, pp. 5671–5683, Sep. 2018.

[18] F. Fang, J. Cheng, and Z. Ding, "Joint energy efficient subchannel and power optimization for a downlink NOMA heterogeneous network," *IEEE Trans. Veh. Technol.*, vol. 68, no. 2, pp. 1351–1364, Feb. 2019.

[19] Z. Song, Q. Ni, and X. Sun, "Spectrum and energy efficient resource allocation with QoS requirements for hybrid MC-NOMA 5G systems," *IEEE Access*, vol. 6, pp. 37055–37069, 2018.

[20] N. Zhao et al., "Joint trajectory and precoding optimization for UAV-assisted NOMA networks," *IEEE Trans. Commun.*, vol. 67, no. 5, pp. 3723–3735, May 2019.

[21] X. Pang, J. Tang, N. Zhao, X. Zhang, and Y. Qian, "Energy-efficient design for mmWave-enabled NOMA-UAV networks," *Sci. China Inf. Sci.*, vol. 64, no. 4, pp. 1–14, Apr. 2021.

[22] Y.-F. Liu and Y.-H. Dai, "On the complexity of joint subcarrier and power allocation for multi-user OFDMA systems," *IEEE Trans. Signal Process.*, vol. 62, no. 3, pp. 583–596, Feb. 2014.

[23] Z. Xiong, Y. Zhang, D. Niyato, R. Deng, P. Wang, and L.-C. Wang, "Deep reinforcement learning for mobile 5G and beyond: Fundamentals, applications, and challenges," *IEEE Veh. Technol. Mag.*, vol. 14, no. 2, pp. 44–52, Jun. 2019.

[24] H. Ye, L. Liang, G. Y. Li, J. Kim, L. Lu, and M. Wu, "Machine learning for vehicular networks: Recent advances and application examples," *IEEE Veh. Technol. Mag.*, vol. 13, no. 2, pp. 94–101, Jun. 2018.

[25] C. He, Y. Hu, Y. Chen, and B. Zeng, "Joint power allocation and channel assignment for NOMA with deep reinforcement learning," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2200–2210, Oct. 2019.

[26] G. Wang, Y. Shao, L.-K. Chen, and J. Zhao, "Subcarrier and power allocation in OFDM-NOMA VLC systems," *IEEE Photon. Technol. Lett.*, vol. 33, no. 4, pp. 189–192, Feb. 15, 2021.

[27] Z. Dong, T. Shang, Q. Li, and T. Tang, "Differential evolution-based optimal power allocation scheme for NOMA-VLC systems," *Opt. Exp.*, vol. 28, no. 15, pp. 21627–21640, 2020.

[28] L. Xiao, Y. Li, C. Dai, H. Dai, and H. V. Poor, "Reinforcement learning-based NOMA power allocation in the presence of smart jamming," *IEEE Trans. Veh. Technol.*, vol. 67, no. 4, pp. 3377–3389, Apr. 2018.

[29] N. C. Luong et al., "Applications of deep reinforcement learning in communications and networking: A survey," *IEEE Commun. Surveys Tuts.*, vol. 21, no. 4, pp. 3133–3174, 4th Quart., 2019, doi: 10.1109/COMST.2019.2916583.

[30] Z. Yang, Y. Liu, Y. Chen, and N. Al-Dhahir, "Cache-aided NOMA mobile edge computing: A reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6899–6915, Oct. 2020, doi: 10.1109/TWC.2020.3006922.

[31] P. Yang, L. Li, W. Liang, H. Zhang, and Z. Ding, "Latency optimization for multi-user NOMA-MEC offloading using reinforcement learning," in *Proc. 28th Wireless Opt. Commun. Conf. (WOCC)*, May 2019, pp. 1–5, doi: 10.1109/WOCC.2019.8770605.

[32] K. N. Doan, M. Vaezi, W. Shin, H. V. Poor, H. Shin, and T. Q. S. Quek, "Power allocation in cache-aided NOMA systems: Optimization and deep reinforcement learning approaches," *IEEE Trans. Commun.*, vol. 68, no. 1, pp. 630–644, Jan. 2020.

[33] S. Zhang et al., "A dynamic power allocation scheme in power-domain NOMA using actor-critic reinforcement learning," in *Proc. IEEE/CIC Int. Conf. Commun. China (ICCC)*, Aug. 2018, pp. 719–723.

[34] Y. Zhang, X. Wang, and Y. Xu, "Energy-efficient resource allocation in uplink NOMA systems with deep reinforcement learning," in *Proc. 11th Int. Conf. Wireless Commun. Signal Process. (WCSP)*, Oct. 2019, pp. 1–6.

[35] H. T. H. Giang, T. N. K. Hoan, P. D. Thanh, and I. Koo, "Hybrid NOMA/OMA-based dynamic power allocation scheme using deep reinforcement learning in 5G networks," *Appl. Sci.*, vol. 10, no. 12, p. 4236, Jun. 2020.

[36] V. Andiappan and V. Ponnusamy, "Deep learning enhanced NOMA system: A survey on future scope and challenges," *Wireless Pers. Commun.*, vol. 123, no. 1, pp. 839–877, Mar. 2022.

[37] M. Shehab, B. S. Ciftler, T. Khattab, M. M. Abdallah, and D. Trinchero, "Deep reinforcement learning powered IRS-assisted downlink NOMA," *IEEE Open J. Commun. Soc.*, vol. 3, pp. 729–739, 2022.

[38] T. Manglayev, R. C. Kizilirmak, Y. H. Kho, N. A. W. A. Hamid, and Y. Tian, "AI based power allocation for NOMA," *Wireless Pers. Commun.*, vol. 124, pp. 3253–3261, Jan. 2022.

[39] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

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

[41] J.-B. Wang et al., "A machine learning framework for resource allocation assisted by cloud computing," *IEEE Netw.*, vol. 32, no. 2, pp. 144–151, Mar./Apr. 2018.

[42] X. Glorot, A. Bordes, and Y. Bengio, "Deep sparse rectifier neural networks," in *Proc. 14th Int. Conf. Artif. Intell. Stat.*, 2011, pp. 315–323.

[43] H. Marshoud, D. Dawoud, V. M. Kapinas, G. K. Karagiannidis, S. Muhaidat, and B. Sharif, "MU-MIMO precoding for VLC with imperfect CSI," in *Proc. 4th Int. Workshop Opt. Wireless Commun. (IWOW)*, Sep. 2015, pp. 93–97.

[44] H. Marshoud, P. C. Sofotasios, S. Muhaidat, G. K. Karagiannidis, and B. S. Sharif, "On the performance of visible light communication systems with non-orthogonal multiple access," *IEEE Trans. Wireless Commun.*, vol. 16, no. 10, pp. 6350–6364, Oct. 2017.

[45] A. Khazali, D. Tarchi, M. G. Shayesteh, H. Kalbkhani, and A. Bozorgchenani, "Energy efficient uplink transmission in cooperative mmWave NOMA networks with wireless power transfer," *IEEE Trans. Veh. Technol.*, vol. 71, no. 1, pp. 391–405, Jan. 2022.

[46] P. S. Bouzinis, P. D. Diamantoulakis, and G. K. Karagiannidis, "Incentive-based delay minimization for 6G-enabled wireless federated learning," *Front. Commun. Netw.*, vol. 3, p. 7, Mar. 2022.

[47] A. Fahim and Y. Gadallah, "An optimized LTE-based technique for drone base station dynamic 3D placement and resource allocation in delay-sensitive M2M networks," *IEEE Trans. Mobile Comput.*, early access, Jun. 15, 2021, doi: 10.1109/TMC.2021.3089329.

[48] H. A. David and H. N. Nagaraja, *Order Statistics*. Hoboken, NJ, USA: Wiley, 2004.

[49] A. Al Hammadi, P. C. Sofotasios, S. Muhaidat, M. Al-Qutayri, and H. Elgala, "Non-orthogonal multiple access for hybrid VLC-RF networks with imperfect channel state information," *IEEE Trans. Veh. Technol.*, vol. 70, no. 1, pp. 398–411, Jan. 2021.

[50] L.-J. Lin, *Reinforcement Learning for Robots Using Neural Networks*. Pittsburgh, PA, USA: Carnegie Mellon Univ., 1992.

[51] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 2018.

[52] U. Challita, L. Dong, and W. Saad, "Proactive resource management for LTE in unlicensed spectrum: A deep learning perspective," *IEEE Trans. Wireless Commun.*, vol. 17, no. 7, pp. 4674–4689, Jul. 2018.

[53] Y. Wei, F. R. Yu, M. Song, and Z. Han, "User scheduling and resource allocation in HetNets with hybrid energy supply: An actor-critic reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 17, no. 1, pp. 680–692, Jan. 2018.

[54] Y. Sun, M. Peng, and S. Mao, "Deep reinforcement learning-based mode selection and resource management for green fog radio access networks," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1960–1971, Apr. 2019.

[55] Z. Dai, Y. Zhang, W. Zhang, X. Luo, and Z. He, "A multi-agent collaborative environment learning method for UAV deployment and resource allocation," *IEEE Trans. Signal Inf. Process. Netw.*, vol. 8, pp. 120–130, Feb. 2022, doi: 10.1109/TSIPN.2022.3150911.
[56] C. Jin, Z. Allen-Zhu, S. Bubeck, and M. I. Jordan, "Is Q-learning provably efficient?" in *Proc. Int. Conf. Adv. Neural Inf. Process. Syst.*, vol. 31, 2018, pp. 4868–4878.
[57] M. Abadi et al., "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Symp. Oper. Syst. Des. Implement. (OSDI)*, Nov. 2016, pp. 265–283.
[58] C. Xing, Y. Jing, S. Wang, S. Ma, and H. V. Poor, "New viewpoint and algorithms for water-filling solutions in wireless communications," *IEEE Trans. Signal Process.*, vol. 68, pp. 1618–1634, Feb. 2020, doi: 10.1109/TSP.2020.2973488.

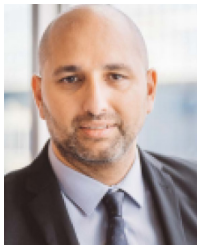
AHMED AL HAMMADI (Member, IEEE) received the B.Sc. and M.Sc. degrees from the Electrical Engineering and Computer Science Department, Khalifa University, Abu Dhabi, UAE, in 2011 and 2015, respectively, where he is currently pursuing the Ph.D. degree in electrical and computer engineering. His research interests include visible light communications, mmWave massive MIMO, machine learning, and optimization techniques for next-generation wireless networks.

LINA BARIAH (Senior Member, IEEE) received the M.Sc. and Ph.D. degrees in communications engineering from Khalifa University, Abu Dhabi, UAE, in 2015 and 2018, respectively. She was a Visiting Researcher with the Department of Systems and Computer Engineering, Carleton University, Ottawa, ON, Canada, in 2019, and an Affiliate Research Fellow with the James Watt School of Engineering, University of Glasgow, U.K. She is currently a Senior Researcher with Technology Innovation Institute, a Visiting Research Scientist with Khalifa University, and an Affiliate Researcher with University at Albany, USA. She serves as the Session Chair and an Active Reviewer for numerous IEEE conferences and journals. She is currently an Associate Editor for the IEEE COMMUNICATION LETTERS, an Associate Editor for the IEEE Open Journal of the Communications Society, and an Area Editor for *Physical Communication* (Elsevier). She is a Guest Editor in *IEEE Network Magazine*, IEEE OPEN JOURNAL OF VEHICULAR TECHNOLOGY, and *IEEE Communication Magazine*. She was a member of the technical program committee of a number of IEEE conferences, such as ICC and Globecom. She is currently organizing/chairing a number of workshops. She is a Senior Member of the IEEE Communications Society, IEEE Vehicular Technology Society, and IEEE Women in Engineering.

SAMI MUHAIDAT (Senior Member, IEEE) received the Ph.D. degree in electrical and computer engineering from the University of Waterloo, Waterloo, ON, Canada, in 2006. From 2007 to 2008, he was a Postdoctoral Fellow with the Department of Electrical and Computer Engineering, University of Toronto, Toronto, ON, Canada. From 2008 to 2012, he was an Assistant Professor with the School of Engineering Science, Simon Fraser University, Burnaby, BC, Canada. He is currently a Professor with Khalifa University, Abu Dhabi, UAE, and a Visiting Professor with the Department of Electrical and Computer Engineering, Carelton University, Ottawa, ON, Canada. He is also a Visiting Reader with the Faculty of Engineering, University of Surrey, Guildford, U.K. He was a recipient of several scholarships during his undergraduate and graduate studies and the winner of the 2006 Postdoctoral Fellowship Competition. He was a Senior Editor of the IEEE COMMUNICATIONS LETTERS, and an Associate Editor of IEEE TRANSACTIONS ON COMMUNICATIONS, IEEE COMMUNICATIONS LETTERS, and IEEE TRANSACTIONS ON VEHICULAR TECHNOLOGY. He is currently an Area Editor of the IEEE TRANSACTIONS ON COMMUNICATIONS.

MAHMOUD AL-QUTAYRI (Senior Member, IEEE) received the B.Eng. degree in electrical and electronic engineering from Concordia University, Canada, in 1984, the M.Sc. degree in electrical and electronic engineering from the University of Manchester, U.K., in 1987, and the Ph.D. degree in electrical and electronic engineering from the University of Bath, U.K., in 1992. He is currently a Full Professor with the Department of Electrical and Computer Engineering and the Associate Dean for Graduate Studies with the College of Engineering, Khalifa University, UAE. Prior to joining Khalifa University, he worked with De Montfort University, U.K., and University of Bath, U.K. He has authored/coauthored numerous technical papers in peer-reviewed journals and international conferences. He also coauthored a book titled *Digital Phase Lock Loops: Architectures and Applications* and edited a book titled *Smart Home Systems*. This is in addition to a number of book chapters and patents. His current research interests include wireless sensor networks, embedded systems design, in-memory computing, mixed-signal integrated circuits design and test, and hardware security.

PASCHALIS C. SOFOTASIOS (Senior Member, IEEE) was born in Volos, Greece, in 1978. He received the M.Eng. degree from Newcastle University, U.K., in 2004, the M.Sc. degree from the University of Surrey, U.K., in 2006, and the Ph.D. degree from the University of Leeds, U.K., in 2011. He was with the University of Leeds; the University of California at Los Angeles, CA, USA; Tampere University of Technology, Finland; the Aristotle University of Thessaloniki, Greece; and the Khalifa University of Science and Technology, UAE, where he is currently an Associate Professor with the Department of Electrical Engineering and Computer Science, Khalifa University. His research interests include digital and optical wireless communications and special functions and statistics. He received the Scholarship from UK-EPSRC for his M.Sc. studies and from UK-EPSRC and Pace plc for his Ph.D. studies. He received the Exemplary Reviewer Award from the IEEE COMMUNICATIONS LETTERS in 2012, the Best Paper Award from ICUFN 2013, and the IEEE TRANSACTIONS ON COMMUNICATIONS in 2015 and 2016. He is a Regular Reviewer of several international journals and a member of the Technical Program Committee of numerous IEEE conferences. He is currently the Editor of the IEEE COMMUNICATIONS LETTERS.



MÉROUANE DEBBAH (Fellow, IEEE) received the M.Sc. and Ph.D. degrees from the École Normale Supérieure Paris-Saclay, Cachan, France, in 1999 and 2002, respectively. He was with Motorola Labs, Saclay, France, from 1999 to 2002, and the Vienna Research Center for Telecommunications, Vienna, Austria, until 2003. From 2003 to 2007, he was an Assistant Professor with the Mobile Communications Department, Institut Eurecom, Sophia Antipolis, France. In 2007, he was appointed as a Full Professor with Centrale Supélec, Gif-sur-Yvette, France. From 2007 to 2014, he was the Director of the Alcatel-Lucent Chair on Flexible Radio. From 2014 to 2021, he was the Vice-President of the Huawei France Research Center, Boulogne-Billancourt, France, and jointly the Director of the Mathematical and Algorithmic Sciences Laboratory and the Lagrange Mathematical and Computing Research Center, Paris, France. Since 2021, he has been the Chief Research Officer with the Technology Innovation Institute, Abu Dhabi, UAE. His research interests lie in fundamental mathematics, algorithms, statistics, information, and communication sciences research. He was a recipient of the ERC Grant MORE (Advanced Mathematical Tools for Complex Network Engineering) from 2012 to 2017. He received more than 20 best paper awards, including the Mario Boella Award in 2005, the IEEE Glavieux Prize Award in 2011, the Qualcomm Innovation Prize Award in 2012, the 2019 IEEE Radio Communications Committee Technical Recognition Award, and the 2020 SEE Blondel Medal.