Received December 4, 2020, accepted December 17, 2020, date of publication December 22, 2020, date of current version January 7, 2021.

Digital Object Identifier 10.1109/ACCESS.2020.3046499

Real-Time Energy Harvesting Aided Scheduling in UAV-Assisted D2D Networks Relying on Deep Reinforcement Learning

KHOI KHAC NGUYEN[1], NGO ANH VIEN[2], LONG D. NGUYEN[3], (Member, IEEE), MINH-TUAN LE[4], LAJOS HANZO[5], (Fellow, IEEE), AND TRUNG Q. DUONG[1], (Senior Member, IEEE)

[1]School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, Belfast BT7 1NN, U.K.
[2]Bosch Center for Artificial Intelligence, Renningen, Germany
[3]Duy Tan University, Da Nang 550000, Vietnam
[4]MobiFone Research and Development Center, MobiFone Corporation, Hanoi, Vietnam
[5]School of Electronics and Computer Science, University of Southampton, Southampton SO17 1BJ, U.K.

Corresponding author: Trung Q. Duong (trung.q.duong@qub.ac.uk)

This work was supported in part by the U.K. Royal Academy of Engineering (RAEng) through the RAEng Research Chair and Senior Research Fellowship Scheme under Grant RCSRF2021\11\41 and the RAEng Research Fellowship Scheme under Grant RF1415\14\22. It was also supported in part by the Newton Fund Institutional Link through the Fly-by Flood Monitoring Project under Grant ID 428328486 and in part by Researcher Links grant, ID 527612186, under the Newton Fund partnership, the grants are funded by the UK Department for Business, Energy and Industrial Strategy and delivered by the British Council, for further information, please visit www.newtonfund.ac.uk. The work of Lajos Hanzo was supported by the Engineering and Physical Sciences Research Council of the Royal Society's Global Challenges Research Fund Grant and the European Research Council's Advanced Fellow Grant QuantCom under Project EP/N004558/1, Project EP/P034284/1, Project EP/P034284/1, and Project EP/P003990/1 (COALESCE).

ABSTRACT Unmanned aerial vehicle (UAV)-assisted device-to-device (D2D) communications can be deployed flexibly thanks to UAVs' agility. By exploiting the direct D2D interaction supported by UAVs, both the user experience and network performance can be substantially enhanced at public events. However, the continuous moving of D2D users, limited energy and flying time of UAVs are impediments to their applications in real-time. To tackle this issue, we propose a novel model based on deep reinforcement learning in order to find the optimal solution for the energy-harvesting time scheduling in UAV-assisted D2D communications. To make the system model more realistic, we assume that the UAV flies around a central point, the D2D users move continuously with random walk model and the channel state information encountered during each time slot is randomly time-variant. Our numerical results demonstrate that the proposed schemes outperform the existing solutions. The associated energy efficiency game can be solved in less than one millisecond by an off-the-shelf processor using trained neural networks. Hence our deep reinforcement learning techniques are capable of solving real-time resource allocation problems in UAV-assisted wireless networks.

INDEX TERMS Energy harvesting, time scheduling, resource allocation, UAV-assisted D2D communications, and deep reinforcement learning.

I. INTRODUCTION

Unmanned aerial vehicles (UAVs) have various wireless applications ranging from public safety, environmental monitoring, and enhanced network connectivity as a benefit of their nimble mobility features. UAVs are capable of assisting wireless networks in providing ubiquitous coverage, robust handovers, and flawless real-time multi-media communications. However, the performance of UAV-assisted networks is limited by the UAVs' energy-storage and the resultant flying time. Recent research has tackled some of the challenges in UAV-supported wireless communications [1]–[9]. Yet, most techniques rely on unrealistic simplifications and focus predominantly on data transmission. Hence it is crucial to find solutions to the associated problems in realistic dynamic environments, as detailed below.

Device-to-device (D2D) communications solutions have been designed for diverse applications, such as smart city

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

operation [10] and video streaming [11], by exploiting direct D2D connections. UAV-supported D2D communications is eminently suitable for providing emergency notifications or simple text messages, when deployed in inaccessible disaster zones or remote areas. But again, they potentially suffer from limited UAV flying time as well as energy constraints, and often have strict computational deadlines in support of D2D communications. These stringent requirements call for powerful solutions in support of real-time resource allocation for enhancing the network performance while satisfying all the constraints. Several techniques have been proposed for solving the associated resource management problems, some of which achieve excellent performance, but they cannot satisfy the stringent time constraints of real-life applications.

Machine learning-based solutions are regarded as powerful techniques for tackling the challenges in UAV-assisted wireless communications. Very recently, deep reinforcement learning (DRL)-based methods, which rely on a combination of reinforcement learning and neural networks, have demonstrated impressive results in resource allocation [12]–[15]. Upon interacting directly with the environment to learn by trial-and-error, the approaches based on DRL algorithms have demonstrated self-organising capability to adapt to a dynamic environment that exhibits rapidly fluctuating channel state information. Furthermore, by using neural networks for training, flexible and prompt decisions may be made according to the environment's state. Inspired by the success of DRL algorithms in solving the resource allocation problems tackled in [12]–[15], we also rely on the broad family of DRL algorithm-based techniques to deal with the energy harvesting scheduling problem of UAV-assisted D2D communications. The UAV is considered to be an agent, who interacts with the environment in order to find the optimal policy. After being trained, the agent embarks on the most appropriate action in each time step by deciding to opt for either energy harvesting or for information transmission, approximately maximising the network performance. We will demonstrate that our proposed models outperform the benchmarks in terms of the processing time imposed, requiring less than one millisecond to decide upon the most beneficial action for the next time step.

A. STATE-OF-THE-ART AND CHALLENGES

UAVs have been adopted for various applications such as geographic surveys, security control, agriculture, and goods delivery. For example, as seen in [16], Amazon provided a service that allows customers to opt for UAV-delivery and receive packages within 30 minutes. In [17], a scouting UAV suitable for counter-terrorism was developed in order to discover weapons and hide-out locations, or to conduct battle damage assessment. Moreover, UAVs are becoming practical resources for rescue teams and medical emergencies, as in payload delivery missions [18]. UAVs have also been used to enhance wireless network performance [3], [5]–[7], [19], [20] or in disaster relief networks [3], [19], public safety communications [21] and sensor data tracking systems [19], [22]. For example, the authors of [7] proposed a cooperative interference cancellation scheme for multi-beam UAV communication, with the objective of maximising the uplink sum-rate, while suppressing the interference at the ground base stations (BSs). But again, UAVs also suffer from some stringent constraints owing to limitation to their power, flying time and reliability. Thus, it is necessary to optimise their flight trajectory and resource allocation in order to enhance the network performance.

Sophisticated techniques have been developed to deal with resource allocation problems in UAV-assisted wireless networks. In [19], the authors proposed a system model using UAVs in the aftermath of natural disasters. A real-time optimisation method having a low complexity was proposed to design the optimal path for gathering data in wireless sensor networks. In [22], a UAV was used to collect data from sensor nodes in Internet-of-Things (IoT)'s networks. By jointly optimising the sensor nodes's wake-up schedule and the UAV trajectory, the authors minimised the energy consumption, while reducing the data collection time. In [3], the authors provided network coverage to an inaccessible disaster-stricken area with the support of UAVs. The K-means clustering method combined with a sophisticated power allocation algorithm was proposed for the real-time support of users sending information about their positions and conditions to families and rescue teams during and after a disaster. However, there is still a paucity of solutions for realistic real-time scenarios.

To elaborate a little further, the associated resource allocation problems are complex owing to the dynamic positions of UAVs. In this context, the authors of [8] used the classic Lagrangian relaxation method to incorporate their constraints into the objective function and conceived resource optimisation for harvest-and-transmit protocols of UAV-assisted D2D communications. Explicitly, the D2D transmitters harvested energy from the UAV and then used the harvested energy for information transmission to the D2D receivers. In [6], a novel solution based on a logarithmic inequality was introduced for jointly optimising the power allocation and energy harvesting in UAV-aided D2D communications.

Both deep learning and reinforcement learning have become popular for mitigating the violently fluctuating channel quality effects of realistic wireless networks. In [3], the authors proposed an unsupervised learning algorithm, namely the K-means algorithm, for clustering the remote users located in a disaster area into small groups and then a UAV was deployed to serve them separately. Indeed, numerous DRL algorithms have been proposed in the literature [12]–[15], [23] for finding the optimal policy in the face of the near-instantaneously fluctuating propagation environment. The agents observe the environment's state and take actions; then, trial-and-error based learning is employed until the performance saturates. In realistic scenarios it would take excessive time to find the optimal solution, when relying on mathematical models. By contrast, DRL based model-free schemes are potentially capable of finding the solutions more promptly by neural networks following a training session.

In [15], the authors proposed models based on both deep Q-learning, as well as on double deep Q-learning and dueling deep Q-learning algorithms to solve a multi-agent aided power allocation problem in D2D communications. However, the action space in those algorithms has to be discrete, and human intervention may also be required. In [13], the authors improved the model by introducing a distributed deep deterministic policy gradient algorithm to solve the power control problem of D2D-based vehicle-to-vehicle communications. Inspired by the encouraging results of the above applications, we conceive bespoke DRL algorithms for optimising the energy harvesting time scheduling of UAV-enabled D2D communications.

B. CONTRIBUTIONS

Our main contributions are as follows:

- We conceive a system model capable of reflecting the dynamic position of UAVs and the unknown channel state information (CSI).
- We then formulate our energy-harvesting time scheduling problem for UAV-assisted D2D communications, where the UAV is considered to be an agent in the game. The agent will observe the environment's state and then takes the action of approximately choosing the specific time span τ that maximises the energy efficiency (EE) of the network. In each time step of the DRL algorithm, the UAV chooses the most appropriate action to be taken according to the change of the environment and the CSI.
- We propose a novel deep deterministic policy gradient (DDPG) algorithm for solving the energy efficiency optimisation game for the UAV-supported D2D scenario considered. Our method outperforms conventional optimisation techniques in terms of its EE vs. complexity, hence resulting in a reduced processing time.
- We further improve our model with the aid of an efficient sampling technique by appropriately adapting the proximal policy optimisation (PPO) algorithms of [24]. More explicitly, we rely on the clipping surrogate objective and Kullback-Leibler (KL) divergence penalty [24] for formulating the objective function (OF) of the PPO algorithm. These techniques improve the speed of convergence, and are robust in terms of adapting to the changes of the environment.

In the remainder of our paper, Section II formulates the system model of UAV-enabled D2D communications. The background of DRL algorithms is introduced in Section III. In Section IV, we solve our energy-harvesting time scheduling problem by using the DDPG algorithm for the continuous action space of D2D communications supported by a UAV. To improve training and sampling, in Section V, we propose a model based on the PPO algorithm to constrain the size of policy updating at each iteration. Our numerical results are presented in Section VI for characterising DRL algorithm-based UAV-assisted wireless networks. Finally, we conclude the paper and propose some future research directions in Section VII.

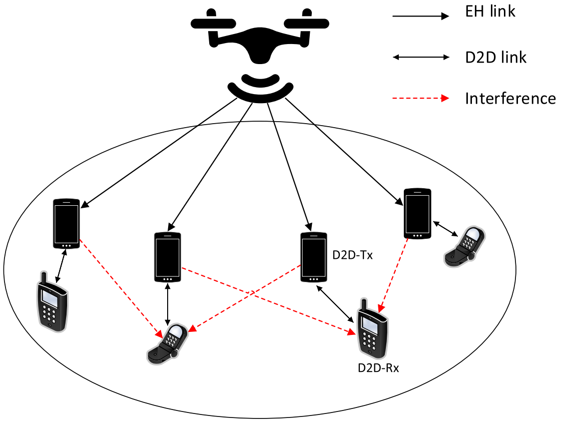


FIGURE 1. System model of D2D communications supported by a UAV.

II. SYSTEM MODEL AND PROBLEM FORMULATION

Our system includes N D2D pairs and one UAV equipped with a single-antenna, as seen in Fig. 1. Each D2D pair consists of a single-antenna D2D transmitter (D2D-Tx) and a single antenna D2D receiver (D2D-Rx). The N D2D pairs are randomly distributed within the UAV's coverage area. In each time step, the D2D pairs are moving continuously following the random walk model with the velocity v. The UAV is moving randomly in a zone around a point in the centre of coverage are due to the effect of wind. We adopt the harvest-then-transmit protocol of [25]. The energy harvesting and information transmission take place in their dedicated phases. In the first phase the D2D-Tx harvests energy from the UAV during time span $\tau\mathbb{T}$ with $0 \leq \tau \leq 1$. Then in the remaining phase spanning $(1-\tau)\mathbb{T}$ the information transmission between D2D-Tx and D2D-Rx takes place. For convenience, we assume that the block time is normalised to 1, $\mathbb{T} = 1$.

We assume that the central point in the 3D place of the UAV is (x_0, y_0, H_0), where H_0 is the altitude of the UAV's antenna. In a real-life application, the UAV is affected by environmental factors, such as wind and rain, which is the reason for its random movement around the position $(x_{UAV}, y_{UAV}, H_{UAV})$. The locations of the ith D2D-Tx and the jth D2D-Rx are $(x_i^{\mathrm{Tx}}, y_i^{\mathrm{Tx}})$ and $(x_i^{\mathrm{Rx}}, y_i^{\mathrm{Rx}})$ with $i, j = 1, \ldots, N$, respectively. The distance between the UAV and the ith D2D-Tx is given by

$$D_i = \sqrt{d_i^2 + H_{UAV}^2}, \tag{1}$$

where $d_i = \sqrt{(x_{UAV} - x_i^{\mathrm{Tx}})^2 + (y_{UAV} - y_i^{\mathrm{Tx}})^2}$ is the Euclidean distance between the UAV and the ith D2D-Tx.

The air-to-ground channel between the UAV and the D2D-Tx is subjected to blockage effects of buildings. The probability of having a line-of-sight (LoS) connection between the UAV and the ith D2D user is given by [26]

$$P_i^{LoS} = \frac{1}{1 + a\exp\Big[-b\Big(\Theta_i - a\Big)\Big]}, \tag{2}$$

where a and b are constants that depend on the environment. The elevation angle θ is defined as

$$\Theta_i = \frac{180}{\pi}\sin^{-1}\left(\frac{H_{UAV}}{D_i}\right). \tag{3}$$

The probability of having a non-line-of-sight (NLoS) link is $P_i^{NLoS} = 1 - P_i^{LoS}$. Thus, the channel gain between the UAV and the ith D2D user characterised by the LoS and NLoS is given by

$$h_i = P_i^{LoS} \times d_i^{-\alpha_E} + P_i^{NLoS} \times \omega d_i^{-\alpha_E}, \tag{4}$$

where ω and α_E are the NLOS connection factor and the path-loss exponent between the UAV and the user link, respectively.

Furthermore, we define P_0 and g_i as the maximum total transmit power of the UAV and the link's power gain between the UAV and the ith D2D-Tx, respectively. The energy harvested at the ith D2D-Tx over the time span τ is given by

$$\mathcal{E} = \tau\eta P_0 g_i, \tag{5}$$

where $0 < \eta < 1$ is the energy harvesting efficiency. We denote the information transmission power usage between D2D-Tx and D2D-Rx of the ith D2D pair by p_i. Thus, the total power usage available for information transmission in the entire network is given by

$$\Phi = \sum_i^N (1-\tau)p_i, \quad i \in N. \tag{6}$$

We define the channel power gain at the reference distance by β_{D2D}, the small scale fading channel power gain (an exponentially distributed random variable) by f_i^2, and the path-loss exponent by α. The channel's power gain between the ith D2D-Tx and the jth D2D-Rx is defined as

$$h_{ij} = \beta_{D2D} f_i^2 d_{ij}^{-\alpha}, \tag{7}$$

where $d_{ij} = \sqrt{(x_i^{\text{Tx}} - x_j^{\text{Rx}})^2 + (y_i^{\text{Tx}} - y_j^{\text{Rx}})^2}$ is the Euclidean distance between ith D2D-Tx and jth D2D-Rx.

The signal-to-interference-plus-noise (SINR) ratio at the ith D2D user's receiver is:

$$\gamma_i = \frac{p_i h_{ii}}{\sum_{j\in N}^{j\neq i} p_j h_{ji} + \sigma^2}, \tag{8}$$

where h_{ji} is the channel gain between the jth D2D-Tx and the ith D2D-Rx, while σ^2 is the AWGN power.

The information throughput at the ith D2D pair is given by

$$R_i(\tau, p_i) = (1-\tau)W\log_2(1+\gamma_i), \tag{9}$$

where W is the bandwidth. The D2D link's communication constraint is formulated as:

$$R_i(\tau, p_i) \geq r_{min}, \quad \forall i \in N, \tag{10}$$

where the threshold r_{min} represents the quality-of-service (QoS) constraint. The power total consumption during the energy harvesting phase between the UAV and the ith D2D user as well as of the information transmission phase between the ith D2D-Tx and the ith D2D-Rx is formulated as

$$\rho(\tau, p) = \sum_{i=1}^N (1-\tau)p_i + \tau\eta P_0 + P_{cir}, \tag{11}$$

where we have $p = [p_i]|_{i=1}^N$ and P_{cir} is the total circuit power dissipation at the UAV and the D2D users.

In harvesting scheduling optimisation, we assume that each D2D-Tx uses the maximum energy harvested from the UAV for transmitting information, yielding

$$(1-\tau)p_i = \tau\eta P_0 g_i. \tag{12}$$

Our objective is to maximise the EE defined as

$$\chi = \frac{\sum_{i=1}^N R_i(\tau, p_i)}{\rho(\tau, p)}, \tag{13}$$

Thus, we formulate the EE optimisation problem as

$$\begin{aligned} \max_{\tau} \; & \frac{\sum_{i=1}^N (1-\tau)W\log_2(1+\gamma_i)}{\sum_{i=1}^N (1-\tau)p_i + \tau\eta P_0 + P_{cir}} \\ s.t. \; & 0 < \tau < 1 \\ & R_i(\tau, p_i) \geq r_{min}, \quad \forall i \in N, \end{aligned} \tag{14}$$

We proceed by setting up our EE game as a Markov decision process (MDP) [27], defined by the five tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}, \mathcal{R}, \zeta \rangle$, where $\mathcal{S}$ is the state space, $\mathcal{A}$ is the action space, $\mathcal{P} : \mathcal{S} \times \mathcal{A} \times \mathcal{S} \rightarrow \mathbb{R}$ is the state transition probability, $\mathcal{R} : \mathcal{S} \rightarrow \mathbb{R}$ is the reward function, and $\zeta \in (0, 1)$ is the discount factor. We define the policy mapping as a distribution $\pi : \mathcal{S} \times \mathcal{A} \rightarrow [0, 1]$. Then, the game is formulated as follows:

- *Agent*: The UAV is an agent. The agent observes the state and takes an action to interact with the environment for finding the optimal policy.
- *State space*: The state space is defined as a cooperative state of all nodes in the network as

$$\mathcal{S} = \{\mathcal{I}_1, \ldots, \mathcal{I}_i, \ldots \mathcal{I}_N\}, \tag{15}$$

where $\mathcal{I}_i$ indicates whether the ith D2D pair satisfies the SINR constraints:

$$\mathcal{I}_i = \begin{cases} 1, & \text{for} \quad R_i(\tau, p_i) \geq r_{min} \\ 0, & \text{for} \quad \text{otherwise.} \end{cases} \tag{16}$$

- *Action space*: The agent at state s selects an action a from the legitimate action space to obtain the reward r,

$$\mathcal{A} = \{\tau\}, 0 < \tau < 1. \tag{17}$$

- *Reward function*: At each time step t, the agent will take action a following the policy π to maximise the reward r of the network. The reward function is a joint function of all D2D pairs formulated as

$$\mathcal{R} = \frac{\sum_{i=1}^N (1-\tau)W\log_2(1+\gamma_i)}{\sum_{i=1}^N (1-\tau)p_i + \tau\eta P_0 + P_{cir}}. \tag{18}$$

III. PRELIMINARIES

The DRL algorithms rely either on a value function-based model or on a policy gradient-based model. In this section, we briefly present the concepts and mathematical formulation of both the value function and the policy gradient based models.

A. VALUE FUNCTION

The value function at state s is the expected reward, while following the policy π

$$V^{\pi}(s) = \mathbb{E}\bigg[\sum_{t\geq 0} \zeta r^t | s_0 = s, \pi\bigg], \tag{19}$$

where the expectation $\mathbb{E}[.]$ denotes the empirical average over a batch of samples.

The action-value function is the expected reward obtained after taking action a at state s under the policy π, which is expressed as

$$Q^{\pi}(s, a) = \mathbb{E}\bigg[\sum_{t\geq 0} \zeta r^t | s_0 = s, a_0 = a, \pi\bigg]. \tag{20}$$

The optimal performance obtained at state s when taking action a is associated with the maximum expected reward defined as

$$Q^*(s, a) = Q^{\pi^*}(s, a) = \max_{\pi} \mathbb{E}\bigg[\sum_{t\geq 0} \zeta r^t | s_0 = s, a_0 = a, \pi\bigg], \tag{21}$$

where π^* is the optimal policy.

We can reach the optimal performance by finding the optimal policy π^* that follow the Bellman equation [28]

$$Q^*(s, a) = \mathbb{E}\Big[r + \zeta \max_{a'} Q^*(s', a') | s, a\Big]. \tag{22}$$

Recently, the deep Q-learning algorithm of [29] has gained substantial attention, since it is eminently suitable for estimating the action-value function. To estimate the action-value function, we use a function approximator $Q(s, a; \theta) \approx Q^*(s, a)$ where θ is the parameter of the neural network. Our objective in deep Q-learning is that of minimising the loss $L_i(\theta_i)$ at each iteration i as follows:

$$L_i(\theta_i) = \mathbb{E}\Big[\Big(y_i - Q(s, a; \theta_i)^2\Big)\Big], \tag{23}$$

where $\hat{Q}(s', a'; \theta_{targ})$ is the target network with parameter θ_{targ} and the target value y_i for iteration i is defined as

$$y_i = \mathbb{E}\Big[r + \zeta \max_{a'} \hat{Q}(s', a'; \theta_{targ}) | s, a\Big]. \tag{24}$$

The gradient update is written as

$$\nabla_{\theta_i} L_i(\theta_i) = \mathbb{E}\Big[r + \zeta \max_{a'} \hat{Q}(s', a'; \theta_{targ}) - Q(s, a; \theta_i) \nabla_{\theta_i} Q(s, a; \theta_i)\Big]. \tag{25}$$

B. POLICY SEARCH

As for the policy search-based method, we can directly search for an optimal policy π^* for the agents to reach the best performance in terms of maximising the reward value of

$$J(\theta) = \mathbb{E}\bigg[\sum_{t\geq 0} \zeta r^t | \pi_0\bigg]. \tag{26}$$

The optimal policy parameters can be formulated as

$$\phi^* = \arg\max J(\phi). \tag{27}$$

Mathematically, the average value can be written as

$$J(\phi) = \int^{\kappa} R(\kappa) p(\kappa, \phi) d\kappa, \tag{28}$$

where κ is represented by the trajectory $\{s^0, a^0, s^1, a^1, \ldots, s^{T-1}, a^T, s^T\}$, while $p(\kappa; \phi)$ is a trajectory distribution given by

$$p(\kappa; \phi) = p(s^0) \prod_{t=0}^{T-1} p(s^{t+1} | s^t, a^t) \pi(a^t | s^t; \phi), \tag{29}$$

where ϕ denotes the parameters of the policy π. Upon differentiating the expected reward, we have [30]

$$\begin{aligned} \nabla_{\phi} J(\phi) &= \int^{\kappa} R(\kappa) \nabla_{\phi} p(\kappa; \phi) d\kappa \\ &= \int^{\kappa} R(\kappa) \nabla_{\phi} \log p(\kappa; \phi) p(\kappa; \phi) d\kappa \\ &= \mathbb{E}\Big[R(\kappa) \nabla_{\phi} \log p(\kappa; \phi)\Big]. \end{aligned} \tag{30}$$

Thus, we can estimate the gradient of the reward function by:

$$\nabla_{\phi} J(\phi) \approx \sum_{t=0}^{T-1} R(\kappa) \nabla_{\phi} \log \pi(a^t | s^t; \phi). \tag{31}$$

The parameter ϕ corresponding to the policy π can be updated by using the stochastic gradient descent algorithm as

$$\phi \leftarrow \phi + \alpha \nabla_{\phi} J(\phi), \tag{32}$$

where $\alpha \in [0, 1]$ is the learning rate.

Several algorithms have been developed in the literature based on policy search, such as natural policy gradient methods [31] and vanilla policy gradient methods [32].

IV. ENERGY HARVESTING TIME SCHEDULING IN UAV-POWERED D2D COMMUNICATIONS: A DEEP DETERMINISTIC POLICY GRADIENT APPROACH

In this section, we propose a deep deterministic policy gradient algorithm (DDPG) [33] for energy harvesting time scheduling in UAV-powered D2D communications. The DDPG algorithm is a hybrid model of the value function and policy search methods. By exploiting the benefits of both models, the DDPG algorithm improves the convergence speed of the optimisation to be suitable even for large-scale action spaces.

The DDPG algorithm consists of two fundamental elements: the actor function and critic function. The actor function $\mu(s;\theta_\mu)$ maps the states to a specific action according to the current policy, while the critic function $Q(s,a)$ is learned as in Q-learning for qualifying the action taken. The pair of techniques that we advocate in the DRL algorithm are as follows:

- *Experience replay buffer*: We use a replay memory pool $\mathcal{D}$ for storing the transitions (s^t, a^t, r^t, s^{t+1}), which are inferred from the environment under the exploration policy. A mini-batch K of samples stored in the replay buffer $\mathcal{D}$ will be randomly taken for training the actor and critic network. Additionally, the buffer $\mathcal{D}$ is set to a finite size. Thus, the oldest transitions are discarded for updating samples space, hence the buffer is always up-to-date.
- *Target network*: One of the challenges during the training step is the unstable nature of the network, if we use a shifting set of Q values for calculating the target value. To overcome this challenge, we use the target network for estimating the target values. Here particularly, in the DDPG algorithm, we use the target actor network and the target critic network, $\mu'(s;\theta_{\mu'})$ and $Q'(s,a;\theta_{q'})$, respectively.

We create a mini-batch of K transitions (s^k, a^k, r^k, s^{k+1}) from the buffer $\mathcal{D}$ by random sampling for training. The critic network parameters are updated for minimising the loss function

$$L = \frac{1}{K}\sum_{k}^{K}\left(y^k - Q(s^k, a^k;\theta_q)\right)^2, \tag{33}$$

where we have

$$y^k = r^k(s^k, a^k) + \zeta Q'(s^{k+1}, a^{k+1};\theta_{q'})|_{a^{k+1}=\mu'(s^{k+1};\theta_{\mu'})}. \tag{34}$$

The actor policy is updated using the sampled policy gradient as follows:

$$\nabla_{\theta_\mu} J \approx \frac{1}{K}\sum_{k}^{K}\nabla_{a^k} Q(s^k, a^k;\theta_q)|_{a^k=\mu(s^k)}\nabla_{\theta_\mu}\mu(s^k;\theta_\mu). \tag{35}$$

The parameters θ_q and $\theta_{\mu'}$ of the target actor network and the target critic network are then updated by using soft target updates associated with $\varkappa \ll 1$

$$\theta_{q'} \leftarrow \varkappa\theta_q + (1-\varkappa)\theta_{q'}, \tag{36}$$

$$\theta_{\mu'} \leftarrow \varkappa\theta_\mu + (1-\varkappa)\theta_{\mu'}. \tag{37}$$

It makes the target values be constrained to change significantly more slowly which allows the Q function approach to supervised learning more closely. However, the price is that this may slow down the training due to the delayed value estimators propagation in the target networks μ' and Q'. In continuous action space, we have to find a good exploration policy for attaining better convergence. Thus, we add a noise process of $\mathcal{N}(0,1)$ associated with a small constant ψ to our actor policy, which is formulated as [33]

$$\mu'(s^t) = \mu(s^t;\theta_\mu^t) + \psi\mathcal{N}(0,1). \tag{38}$$

Algorithm 1 Deep Deterministic Policy Gradient Algorithm for Energy Harvesting Time Scheduling in UAV-Assisted D2D Communications

```
Initialise the critic network $Q(s,a;\theta_q)$ and the actor network $\mu(s;\theta_\mu)$ with random parameter $\theta_q$ and $\theta_\mu$, respectively
Initialise the target critic networks $Q'$ and the target actor network $\mu'$ with parameter $\theta_{q'} \leftarrow \theta_q, \theta_{\mu'} \leftarrow \theta_\mu$, respectively
Initialise the replay memory pool $\mathcal{D}$
for episode $= 1, \ldots, M$ do
    Initialise a random process $\mathcal{N}$ for the action exploration
    Receive initial observation state $s^0$
    for iteration $= 1, \ldots, T$ do
        Obtain the action $a^t$ at state $s^t$ according to the current policy and the exploration noise
        Measure the achieved SINR according to (8)
        Update the reward $r^t$ according to (18)
        Observe the new state $s^{t+1}$
        Store transition $(s^t, a^t, r^t, s^{t+1})$ into the replay buffer $\mathcal{D}$
        Sample randomly a mini-batch of $K$ transitions $(s^k, a^k, r^k, s^{k+1})$ from $\mathcal{D}$
        Update the critic by minimising the loss as in (33)
        Update the actor policy using the sampled policy gradient as in (35)
        Update the target networks as in (36) and (37)
        Update the state $s_i^t = s_i^{t+1}$
    end for
end for
```

The details of our DDPG algorithm-based technique of solving the energy harvesting time scheduling in UAV-assisted D2D communications are described in Algorithm 1 where M and T are the number of maximum episode and time step per episode, respectively.

V. EFFICIENT LEARNING WITH PROXIMAL POLICY OPTIMISATION ALGORITHMS TO SOLVE THE ENERGY HARVESTING TIME SCHEDULING PROBLEM IN D2D COMMUNICATIONS ASSISTED BY UAV

In this section, we propose a novel model based on the PPO algorithm relying on an efficient sampling technique for solving the energy harvesting time scheduling game in UAV-assisted D2D communications. The PPO algorithm allows the policy to carry out the most significant possible improvement step using the current data without overestimation that might degrade the performance.

A. CLIPPING SURROGATE METHOD

Let p_θ^t denote the probability ratio $p_\theta^t = \frac{\pi(s,a;\theta)}{\pi(s,a;\theta_{old})}$. Then our main objective is to maximise $\mathcal{L}$ as follows:

$$\begin{aligned}\mathcal{L}(s,a;\theta) &= \mathbb{E}\left[\frac{\pi(s,a;\theta)}{\pi(s,a;\theta_{old})}A^\pi(s,a)\right]\\ &= \mathbb{E}\left[p_\theta^t A^\pi(s,a)\right],\end{aligned} \tag{39}$$

where $A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$ is an estimator of the advantage function defined in [34]. Again, we create a mini-batch K and then use the classic stochastic policy gradient descent to train our neural networks. The policy parameter is updated via

$$\theta^{k+1} = \arg\max \mathbb{E}\left[\mathcal{L}(s,a;\theta^k)\right]. \tag{40}$$

Without improving any constraints, the maximisation of $\mathcal{L}(s,a;\theta)$ may lead to an incentive for the policy to move the probability p_θ^t away from 1. Thus, we opt for a suitable clipping technique for modifying the objective of (40) to [24]

$$\begin{aligned}\mathcal{L}^{CLIP}(s,a;\theta) = \mathbb{E}\Big[&\min(p_\theta^t A^\pi(s,a),\\ &clip(p_\theta^t, 1-\epsilon, 1+\epsilon)A^\pi(s,a))\Big],\end{aligned} \tag{41}$$

where ϵ is a small hyper-parameter. We use the function $clip(p_\theta^t, 1-\epsilon, 1+\epsilon)$ for limiting the probability ratio to avoid the excessive modification of p_θ^t outside the interval $[1-\epsilon, 1+\epsilon]$. In this paper, we use an estimate of the advantage function $A^\pi(s,a)$ formulated as [32]

$$A^\pi(s,a) = r^t + \zeta V^\pi(s^{t+1}) - V^\pi(s^t). \tag{42}$$

B. KULLBACK-LEIBLER DIVERGENCE PENALTY

Instead of using clipping surrogate objective as in Section V-A, we can also use the KL divergence penalty [24] based technique, where the parameters are updated by optimising the KL penalty objective [24]

$$\begin{aligned}\mathcal{L}^{KL}(s,a;\theta) = \mathbb{E}\Big[&\frac{\pi(a|s;\phi)}{\pi(a|s;\phi_{old})}A(s,a)\\ &-\varphi KL[\pi(.|s;\phi_{old}), \pi(.|s;\phi)]\Big].\end{aligned} \tag{43}$$

Then, we compute $d = \mathbb{E}\Big[KL[\pi(.|s;\phi_{old}), \pi(.|s;\phi)]\Big]$ based on the target value d_{targ} of KL divergence [24] :

- if $d < d_{targ}/1.5$, $\varphi \leftarrow \varphi/2$,
- if $d < d_{targ} \times 1.5$, $\varphi \leftarrow \varphi \times 2$.

The parameter φ is promptly updated in the next episode. The details of the PPO based algorithm of solving the energy harvesting time scheduling in UAV-assisted D2D communications are presented in Algorithm 2.

Algorithm 2 Our Proposed Method Based on the PPO Algorithm for Energy Harvesting Time Scheduling in UAV-Assisted D2D Communications

```
Initialise the policy parameter $\theta_\pi$
Initialise the penalty method parameters
for episode $= 1, \ldots, M$ do
    Receive initial observation state $s^0$
    for iteration $= 1, \ldots, T$ do
        Obtain the action $a^t$ at state $s^t$ according to the current policy
        Measure the achieved SINR according to (8)
        Update the reward $r^t$ according to (18)
        Observe the new state $s^{t+1}$
        Update the state $s_i^t = s_i^{t+1}$
        Collect a set of partial trajectories with $K$ transitions
        Estimate the advantage function according to (42)
    end for
    Update the policy parameters using stochastic gradient descent with mini-batch $K$
        $$\theta^{k+1} = \arg\max \frac{1}{K}\sum^{K}\mathcal{L}(s,a;\theta^k) \tag{44}$$
end for
```

TABLE 1. Simulation parameters.

Parameters	Value
Bandwidth (W)	1 MHz
UAV transmission power	5 W
The central point of UAV	$(x_0, y_0, H_0) = (0, 0, 200)$
Max moving distance of users per time step	1m
Max distance of UAV from central point	1m
Path-loss parameter	$\alpha_h = 3$
D2D-Tx and D2D-Rx max distance	50 m
Environment parameter	$a = 11.95, b = 0.136$
Channel power gain	$\beta = -30$ dB
EH efficiency	$\eta = 0.5$
The NLOS connection factor	$\omega = 20$ dB
Non-transmit power of UAV and D2D	4 W
Clipping parameter	$\epsilon = 0.2$
Discounting factor	$\zeta = 0.9$
Max number of D2D pairs	30
Initial batch size	$K = 32$
Number of units per layer	100

VI. SIMULATION RESULTS

In this section, we illustrate the efficiency of our DRL algorithms over the conventional approaches. All the algorithms are implemented using Tensorflow 1.13.1 [35] and the Adam optimisation algorithm [36] for training the neural networks. The algorithm in [6] is implemented using Python and CVXPY library [37] for convex optimisation. All the other simulation parameters are provided in Table 1.

A. PERFORMANCE COMPARISION

Fig. 2 characterises our DRL algorithms when the number of D2D pairs is set to $N = 30$, in comparison to the optimal

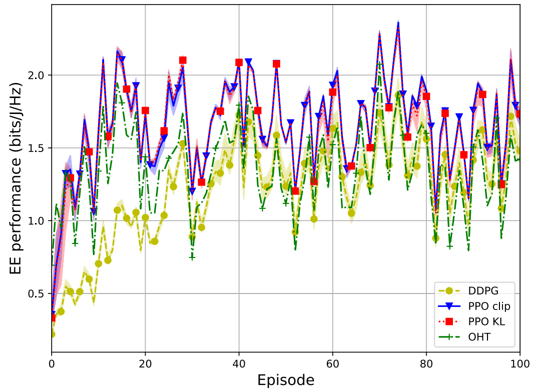


FIGURE 2. The EE of optimal harvesting time using our DDPG, PPO algorithm and the OHT optimisation [6] when the number of D2D pairs $N = 30$.

harvesting time optimisation (OHT) solution of [6]. We use two hidden layers for the DDPG algorithm associated with 100 nodes per layer, while we use a single hidden layer associated with 100 nodes per layer for the PPO algorithm. This is because the DDPG algorithm has an off-policy nature, while the PPO algorithm is of on-policy nature. The position of the UAV is changing over time, hence the channel state information also fluctuates dynamically in every time step. Both algorithms approach the optimal performance after about 50 episodes. The results based on our DDPG and PPO algorithm combination are better than the ones using the OHT optimisation. Moreover, as can be observed in Fig. 2, the methods based on both the PPO algorithm using the above mentioned clip surrogate and KL divergence penalty achieve both similar EE and convergence speed. The convergence speed of the scheme using the PPO algorithm is substantially faster than that of the ones using the DDPG algorithm.

In Fig. 3, we present the performance of our DRL algorithm-based methods and of the OHT optimisation-based method [6], while considering different numbers of D2D pairs. We average the performance over 200 episodes. The average performance of the PPO algorithm-based method is higher than that of the DDPG algorithms. Furthermore, the EE of our proposed solutions is better than that of OHT optimisation regardless of the numbers of D2D pairs within the UAV's coverage area.

The performance for different values of connection constraints of the DDPG and PPO algorithms is presented in Fig. 4. The EE of methods based on the DDPG and PPO algorithm are similar, when r_{min} is in the range of 0.2 to 1.0. The results suggest that the PPO algorithm is indeed flexible and robust in the scenarios considered.

B. PARAMETER ANALYSIS

In Fig. 5, we present the EE of our method based on the PPO algorithm for different batch sizes, K. Upon increasing the batch size K, the convergence speed is reduced. This is because when we take a smaller batch size of samples to train the networks in the PPO algorithm, the policy parameters are updated more frequently. Thus, we can approach the optimal performance faster. However, if we take the training time into consideration, the smaller batch size requires more time for training the neural networks in order collect enough samples. In this study, we opted for the batch size of $K = 32$, for the implementation of the DDPG and PPO algorithms.

In Fig. 6, we consider the difference in EE between the method using the DDPG algorithm in conjunction with

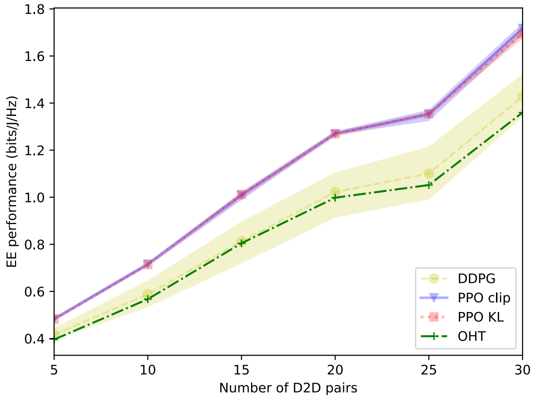


FIGURE 3. The performance results of optimal harvesting time scheduling with different numbers of D2D pairs N.

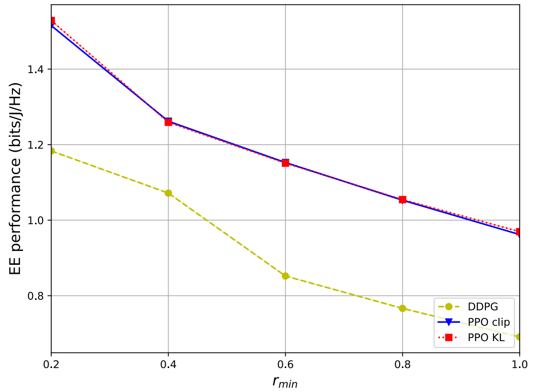


FIGURE 4. The performance results of optimal harvesting time scheduling with different QoS constraints.

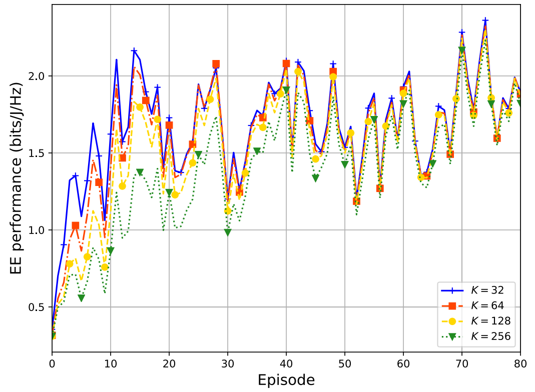


FIGURE 5. The performance of optimal harvesting time using our PPO algorithm relying on the clipping objective technique while considering different value of batch size K.

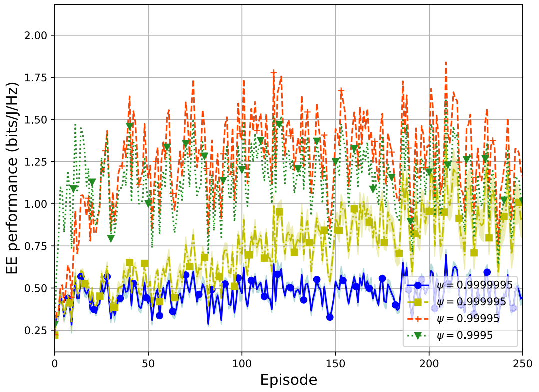


FIGURE 6. The performance of optimal harvesting time using our DDPG algorithm with different value of exploration parameter ψ.

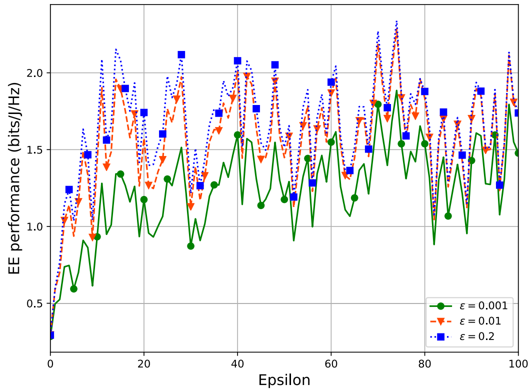
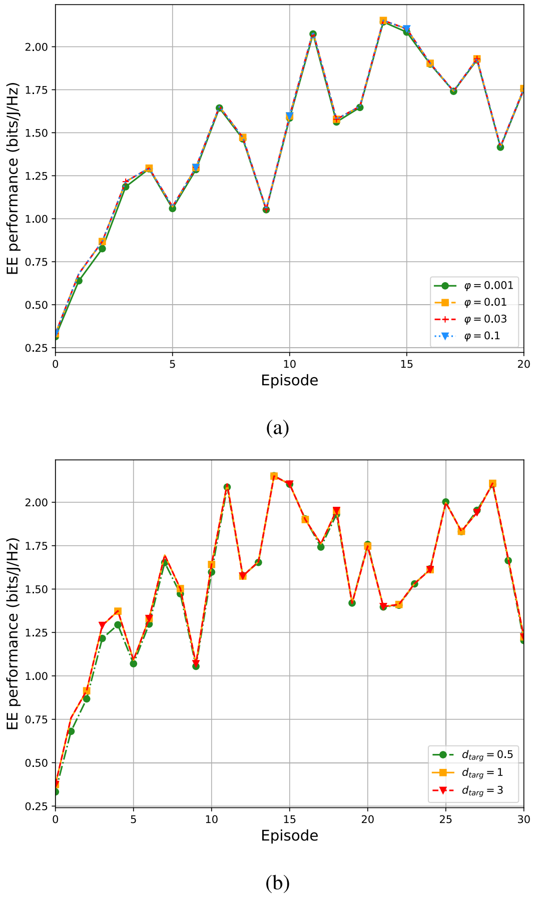


FIGURE 7. The performance of optimal harvesting time by using our PPO algorithm relying on the clipping objective technique with different value of ϵ.

various exploration parameters, ψ. Choosing the appropriate value of exploration is one of the challenges in designing the DDPG algorithm. If we choose the value ψ to be too small in Fig. 6, our algorithm will be stuck at a local optimum because the DRL algorithm is a trial-and-error based method. Hence the agents cannot reach the optimal policy, if we do not allow the agent to try all the possible circumstances. By contrast, if we choose an excessive exploration parameter, ψ, the convergence speed will be affected, because the agents may bounce around the optimal value to explore more hitherto unexplored information. This reduces the convergence speed. As a compromise, we opted for the exploration ratio of $\psi = 0.99995$ for the DDPG algorithm.

Fig. 7 presents the performance of UAV-assisted D2D communications upon using the PPO algorithm relying on the clipping surrogate method, while considering different clipping thresholds, ϵ. The results show that we can achieve the best performance with a threshold of $\epsilon = 0.2$. Meanwhile, Fig. 8 illustrates the EE of the PPO algorithm using the KL penalty divergence in our optimal harvesting time scheduling problem of UAV-assisted D2D communications. In Fig. 8a, the initial value of φ is not critical in the PPO algorithm using the KL penalty method, because φ quickly adjusts. Thus, the results are similar for different values of φ. As a further result, Fig. 8b shows the performance when we employ the method based on the PPO algorithm using the KL method with different values of d_{targ}. The results suggest that we should choose the value of d_{targ} to be moderate for rapid convergence.



(a)

(b)

FIGURE 8. The performance of optimal harvesting time by using our PPO algorithm relying on the KL divergence penalty technique.

C. COMPUTATIONAL COMPLEXITY

We compare the computational complexity of our DDPG algorithm and PPO algorithm in the training phase for the energy harvesting time scheduling problem of D2D communications supported by a UAV. With the DDPG algorithm, the computational complexity is $O(MTK(Nn_{l1} + n_{l1}n_{l2} + \ldots))$ with n_{li} is the number of nodes at layer i in a neural network. On the one hand, the complexity is $O(MT + MK(Nn_{l1} + n_{l1}n_{l2} + \ldots))$ with the PPO algorithm. Furthermore, we compare the time processing of our neural networks in the testing phase and the OHT algorithm [6]. After training, simple calculations are required to predict the proper action in each time-step. The computational complexity of trained networks is $O(Nn_{l1} + n_{l1}n_{l2} + \ldots)$. Specifically, our proposed DDPG algorithm solved the problem in only 0.229ms, 0.259ms, and 0.255ms with the number of D2D pairs at 5, 15, and 30, respectively. Meanwhile, the OHT optimisation takes 54.1ms, 115.5ms, and 170.3ms. Thus, our DDPG algorithm

TABLE 2. The processing time while considering varied number of D2D pairs.

	DDPG	PPO clip	PPO KL	OHT [6]
$N = 5$	0.229ms	0.259ms	0.255ms	54.1ms
$N = 15$	0.246ms	0.277ms	0.281ms	115.5ms
$N = 30$	0.260ms	0.276ms	0.286ms	170.3ms

and PPO algorithm outperforms mathematical models in terms of robustness, EE, and complexity.

VII. CONCLUSION

In this paper, we presented the efficiency of our proposed DRL algorithms to schedule the energy harvesting time of UAV-assisted D2D communications. Our proposed techniques outperform benchmarks in terms of EE and complexity. By utilising the advantages of deep learning, the energy harvesting time scheduling game can be solved almost instantly. The results suggest the DRL algorithm can be a potential technique for real-time applications under limitation of the energy storage and the flying time-constrained UAVs. In the future, we will solve more complicated problems by jointly optimising power allocation, trajectory planning, and multiple UAV scenarios.

ACKNOWLEDGMENT

This article will be presented in part at the 7th International Conference on Industrial Networks and Intelligent Systems (INISCOM 2021), Hanoi, Vietnam, April 22 and 23, 2021.

REFERENCES

[1] R. Duan, J. Wang, C. Jiang, Y. Ren, and L. Hanzo, "The transmit-energy vs computation-delay trade-off in gateway-selection for heterogenous cloud aided multi-UAV systems," *IEEE Trans. Commun.*, vol. 67, no. 4, pp. 3026–3039, Apr. 2019.

[2] R. Rajashekar, M. D. Renzo, K. V. S. Hari, and L. Hanzo, "A beamforming-aided full-diversity scheme for low-altitude Air-to-Ground communication systems operating with limited feedback," *IEEE Trans. Commun.*, vol. 66, no. 12, pp. 6602–6613, Dec. 2018.

[3] L. D. Nguyen, K. K. Nguyen, A. Kortun, and T. Q. Duong, "Real-time deployment and resource allocation for distributed UAV systems in disaster relief," in *Proc. IEEE 20th Int. Workshop Signal Process. Adv. Wireless Commun. (SPAWC)*, Cannes, France, Jul. 2019, pp. 1–5.

[4] L. Xie, J. Xu, and R. Zhang, "Throughput maximization for UAV-enabled wireless powered communication networks," *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1690–1703, Apr. 2019.

[5] Q. Wu, Y. Zeng, and R. Zhang, "Joint trajectory and communication design for multi-UAV enabled wireless networks," *IEEE Trans. Wireless Commun.*, vol. 17, no. 3, pp. 2109–2121, Mar. 2018.

[6] M.-N. Nguyen, L. D. Nguyen, T. Q. Duong, and H. D. Tuan, "Real-time optimal resource allocation for embedded UAV communication systems," *IEEE Wireless Commun. Lett.*, vol. 8, no. 1, pp. 225–228, Feb. 2019.

[7] L. Liu, S. Zhang, and R. Zhang, "Multi-beam UAV communication in cellular uplink: Cooperative interference cancellation and sum-rate maximization," *IEEE Trans. Wireless Commun.*, vol. 18, no. 10, pp. 4679–4691, Oct. 2019.

[8] H. Wang, J. Wang, G. Ding, L. Wang, T. A. Tsiftsis, and P. K. Sharma, "Resource allocation for energy harvesting-powered D2D communication underlaying UAV-assisted networks," *IEEE Trans. Green Commun. Netw.*, vol. 2, no. 1, pp. 14–24, Mar. 2018.

[9] T. Q. Duong, L. D. Nguyen, H. D. Tuan, and L. Hanzo, "Learning-aided realtime performance optimisation of cognitive UAV-assisted disaster communication," in *Proc. IEEE Global Commun. Conf. (GLOBECOM)*, Waikoloa, HI, USA, Dec. 2019, pp. 1–6.

[10] C. Kai, H. Li, L. Xu, Y. Li, and T. Jiang, "Energy-efficient device-to-device communications for green smart cities," *IEEE Trans. Ind. Informat.*, vol. 14, no. 4, pp. 1542–1551, Apr. 2018.

[11] N.-S. Vo, T. Q. Duong, H. D. Tuan, and A. Kortun, "Optimal video streaming in dense 5G networks with D2D communications," *IEEE Access*, vol. 6, pp. 209–223, Oct. 2018.

[12] L. Huang, S. Bi, and Y.-J.-A. Zhang, "Deep reinforcement learning for online computation offloading in wireless powered mobile-edge computing networks," *IEEE Trans. Mobile Comput.*, vol. 19, no. 11, pp. 2581–2593, Nov. 2020.

[13] K. K. Nguyen, T. Q. Duong, N. A. Vien, N.-A. Le-Khac, and L. D. Nguyen, "Distributed deep deterministic policy gradient for power allocation control in D2D-based V2 V communications," *IEEE Access*, vol. 7, pp. 164533–164543, Nov. 2019.

[14] Y. Yu, T. Wang, and S. C. Liew, "Deep-reinforcement learning multiple access for heterogeneous wireless networks," *IEEE J. Sel. Areas Commun.*, vol. 37, no. 6, pp. 1277–1290, Jun. 2019.

[15] K. K. Nguyen, T. Q. Duong, N. A. Vien, N.-A. Le-Khac, and M.-N. Nguyen, "Non-cooperative energy efficient power allocation game in D2D communication: A multi-agent deep reinforcement learning approach," *IEEE Access*, vol. 7, pp. 100480–100490, Jul. 2019.

[16] *Amazon Prime Air*. Accessed: Dec. 4, 2020. [Online]. Available: https://www.amazon.com/Amazon-Prime-Air/b?ie=UTF8&node=8037720011

[17] A. Vacca, F. Cucch, and H. Onishi, "Drones: Military weapons, surveillance or mapping tools for environmental monitoring? The need for legal framework is required," *Transp. Res. Procedia*, vol. 25, pp. 51–62, 2017.

[18] C. A. Thiels, J. M. Aho, S. P. Zietlow, and D. H. Jenkins, "Use of unmanned aerial vehicles for medical product transport," *Air Med. J.*, vol. 34, no. 2, pp. 104–108, Mar. 2015.

[19] T. Q. Duong, L. D. Nguyen, and L. K. Nguyen, "Practical optimisation of path planning and completion time of data collection for UAV-enabled disaster communications," in *Proc. 15th Int. Wireless Commun. Mobile Comput. Conf. (IWCMC)*, Tangier, Morocco, Jun. 2019, pp. 372–377.

[20] L. D. Nguyen, A. Kortun, and T. Q. Duong, "An introduction of real-time embedded optimisation programming for uav systems under disaster communication," *EAI Endorsed Trans. Ind. Netw. Intell. Syst.*, vol. 5, no. 17, pp. 1–8, Dec. 2018.

[21] S. Shakoor, Z. Kaleem, M. I. Baig, O. Chughtai, T. Q. Duong, and L. D. Nguyen, "Role of UAVs in public safety communications: Energy efficiency perspective," *IEEE Access*, vol. 7, pp. 140665–140679, Sep. 2019.

[22] C. Zhan, Y. Zeng, and R. Zhang, "Energy-efficient data collection in UAV enabled wireless sensor network," *IEEE Wireless Commun. Lett.*, vol. 7, no. 3, pp. 328–331, Jun. 2018.

[23] H. Ye, G. Y. Li, and B.-H.-F. Juang, "Deep reinforcement learning based resource allocation for V2 V communications," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3163–3173, Apr. 2019.

[24] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, "Proximal policy optimization algorithms," *CoRR*, vol. abs/1707.06347, 2017.

[25] H. Ju and R. Zhang, "Throughput maximization in wireless powered communication networks," *IEEE Trans. Wireless Commun.*, vol. 13, no. 1, pp. 418–428, Jan. 2014.

[26] 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.

[27] M. L. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Hoboken, NJ, USA: Wiley, 1994.

[28] D. P. Bertsekas, *Dynamic Programming and Optimal Control*, vol. 1, no. 2. Belmont, MA, USA: Athena Scientific, 1995.

[29] V. Mnih, K. Kavukcuoglu, D. Silver, A. Graves, I. Antonoglou, D. Wierstra, and M. Riedmiller, "Playing Atari with deep reinforcement learning," *CoRR*, vol. abs/1312.5602, 2013.

[30] R. S. Sutton, D. McAllester, S. Singh, and Y. Mansour, "Policy gradient methods for reinforcement learning with function approximation," in *Proc. Adv. Neural Inf. Process. Syst.*, 2000, pp. 1057–1063.

[31] S. Kakade, "A natural policy gradient," in *Proc. Adv. Neural Inf. Process. Syst.*, 2002, pp. 1531–1538.

[32] V. Mnih, A. P. Badia, M. Mirza, A. Graves, T. Lillicrap, T. Harley, D. Silver, and K. Kavukcuoglu, "Asynchronous methods for deep reinforcement learning," *CoRR*, vol. abs/1602.01783, 2016.

[33] 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," in *Proc. 4th Int. Conf. Learn. Represent. (ICLR)*, 2016, pp. 1–14.

[34] J. Schulman, P. Moritz, S. Levine, M. I. Jordan, and P. Abbeel, "High-dimensional continuous control using generalized advantage estimation," in *Proc. 4th Int. Conf. Learn. Represent. (ICLR)*, 2016, pp. 1–14.

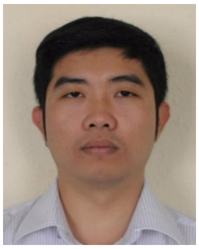
[35] M. Abadi *et al.*, "TensorFlow: A system for large-scale machine learning," in *Proc. 12th USENIX Conf. OSDI*, Nov. 2016, pp. 265–283.

[36] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2014, *arXiv:1412.6980*. [Online]. Available: http://arxiv.org/abs/1412.6980

[37] S. Diamond and S. Boyd, "CVXPY: A Python-embedded modeling language for convex optimization," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 2909–2913, Jan. 2016.

KHOI KHAC NGUYEN was born in Bac Ninh, Vietnam. He received the B.S. degree in information and communication technology from the Hanoi University of Science and Technology (HUST), Vietnam, in 2018. He is currently pursuing the Ph.D. degree with the School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, U.K. His research interests include machine learning and deep reinforcement learning for real-time optimization in wireless networks and massive Internet of Things (IoTs).

NGO ANH VIEN received the B.S. degree in computer engineering from the Hanoi University of Science and Technology, Vietnam, in 2005, and the Ph.D. degree in computer engineering from Kyung Hee University, South Korea, in 2009. He was a Postdoctoral Researcher with the National University of Singapore, from 2009 to 2011, the Ravensburg-Weingarten University of Applied Sciences, Germany, from 2011 to 2013, and the Machine Learning and Robotics Laboratory, University of Stuttgart, from 2013 to 2017. He was a Lecturer (also known as Assistant Professor) with Queen's University Belfast, U.K., from 2017 to 2020. He is currently a Researcher with the Bosch Center for Artificial Intelligence (BCAI), Renningen, Germany. His research interests include machine learning and robotics.



LONG D. NGUYEN (Member, IEEE) was born in Dong Nai, Vietnam. He received the B.S. degree in electrical and electronics engineering and the M.S. degree in telecommunication engineering from the Ho Chi Minh City University of Technology (HCMUT), Vietnam, in 2013 and 2015, respectively, and the Ph.D. degree in electronics and electrical engineering from Queen's Univeristy Belfast (QUB), U.K., in 2018. He was a Research Fellow with Queen's University Belfast, U.K., for a part of Newton project from 2018 to 2019. He is currently with Dong Nai University, Vietnam, as an Assistant Professor, and Duy Tan University as an Adjunct Assistant Professor. His research interests include convex optimization techniques for resource management in wireless communications, energy-efficiency approaches (heterogeneous networks, relay networks, cell-free networks, and massive MIMO), and real-time embedded optimization for wireless networks and the Internet of Things (IoTs).

MINH-TUAN LE was born in Thanh Hóa, Vietnam, in 1976. He received the B.E. degree in electronic engineering from the Hanoi University of Science and Technology, Vietnam, in 1999, and the M.S. and Ph.D. degrees in electrical engineering from Information and Communication University (Department of Electrical and Engineering of Korean Advanced Institute of Science and Technology (KAIST), Daejon, South Korea), in 2003 and 2007, respectively. From 1999 to 2001 and from 2007 to 2008, he worked as a Lecturer with the Posts and Telecommunication Institute of Technology (PTIT), Vietnam. From November 2012 to 2015, he worked at the Hanoi Department of Science and Technology, Vietnam. He is currently working at the MobiFone Reasearch and Development Center, MobiFone Corporation, Vietnam. His research interests include space–time coding, space–time processing, and MIMO systems. He was a recipient of the 2012 ATC Best Paper Award from the Radio Electronics Association of Vietnam (REV) and the IEEE Communications Society.

LAJOS HANZO (Fellow, IEEE) received the master's and Ph.D. degrees from the Technical University (TU) of Budapest, in 1976 and 1983, respectively, the D.Sc. degree from the University of Southampton, in 2004, and the Honorary Doctorates from the TU of Budapest and from The University of Edinburgh, in 2015 and 2009, respectively. He has published more than 1900 contributions at IEEE Xplore and 19 Wiley-IEEE Press books and has helped the fast-track career of 123 Ph.D. students. More than 40 of them are professors at various stages of their careers in academia, and many of them are leading scientists in the wireless industry. He is currently a Foreign Member of the Hungarian Academy of Sciences and the former Editor-in-Chief of IEEE Press. He is also a Fellow of the Royal Academy of Engineering (REng.), IET, and EURASIP. He has served several terms as a Governor for IEEE ComSoc and of VTS.

TRUNG Q. DUONG (Senior Member, IEEE) received the Ph.D. degree in telecommunications systems from the Blekinge Institute of Technology (BTH), Sweden, in 2012. He was a Lecturer (an Assistant Professor), from 2013 to 2017, and a Reader (an Associate Professor), from 2018 to July 2020, with Queen's University Belfast, U.K., where he has been a Full Professor, since August 2020. He is currently a Research Chair of the Royal Academy of Engineering and a Professor with Queen's University Belfast. His current research interests include the Internet of Things (IoT), wireless communications, molecular communications, and signal processing. He received the Best Paper Award at the IEEE Vehicular Technology Conference (VTC-Spring), in 2013, the IEEE International Conference on Communications (ICC), in 2014, the IEEE Global Communications Conference (GLOBECOM), in 2016, the IEEE Digital Signal Processing Conference (DSP), in 2017, and the prestigious Newton Prize in 2017. He was a recipient of prestigious Royal Academy of Engineering Research Fellowship for the term 2016–2021. He also serves as an Editor for the IEEE Transactions on Wireless Communications and IEEE Transactions on Communications and a Lead Senior Editor for IEEE Communications Letters.