Energy Efficiency Optimization for SWIPT-Based D2D-Underlaid Cellular Networks Using Multiagent Deep Reinforcement Learning

Sengly Muy, Dara Ron, and Jung-Ryun Lee, *Senior Member, IEEE*

***Abstract*—In this article, we study the optimization of energy efficiency in wireless device-to-device (D2D-underlaid cellular networks where multiple D2D pairs adopt simultaneous wireless information and power transfer functionality. We formulate the optimization problem, which is a NP-hard combinatorial problem with nonlinear constraints. First, we use optimization-based-iterative techniques such as exhaustive search (ES) and gradient search (GS) with barrier, which are generally used to obtain the global optimum and local optimum of the nonconvex optimization problem, respectively. Considering that these techniques require a centralized unit to share information with each other, we propose multiagent deep reinforcement learning to solve this optimization problem in a distributed manner, which provides optimal decision making together with efficient deep network training under inequality constraints including transmit power, power splitting ratio, and minimum requirement data rate for D2D and cellular users. In this proposed method, we consider the virtual environment in which each agent can train their model according to shared information, and then, we deploy the trained model into the actual environment where each agent can only know their channel gain, interference power, and required minimum throughput. Simulation results show that the proposed algorithm can afford a near-global-optimum solution with much lower computation complexity than ES and outperforms the GS.**

***Index Terms*—Device-to-device (D2D) communication, energy dissipation, energy efficiency, multiagent deep reinforcement learning (DRL), power control, simultaneous wireless information and power transfer (SWIPT).**

I. Introduction

THE rapidly growing number of mobile devices along with the plethora of multimedia applications such as mobile gaming, high-definition (HD) movies, and video conferencing causes an increase in demand for high data rate and quality of service (QoS) [1]. The device-to-device (D2D)-underlaid cellular network has been expected as a key technology in 5 G cellular networks, because it can use the benefits of the communicating devices in proximity to improve the network performances, such as spectral efficiency, energy efficiency, and transmission delay. However, the performance of D2D communication-underlaid cellular networks is degraded by interference between D2D and cellular users that share the same frequency band [2]. Therefore, a provisioning solution for mitigating the interference is crucial. In addition, how to improve energy efficiency is also essential because D2D and cellular users typically use handheld equipment with an energy-limited battery [3].

Simultaneous wireless information and power transfer (SWIPT) technology is a promising approach to enhance the energy efficiency in D2D communication networks because it can convert the interference signals to electricity. In general, the strong interference increases amount of energy harvesting, but decreases the system throughput. Therefore, it is essential to optimize the tradeoff between the system throughput and energy harvesting. Energy efficiency is basically defined as the system throughput to energy dissipation ratio. Hence, the tradeoff between the system throughput and energy harvesting can be optimized via energy efficiency optimization.

Previous studies focusing on SWIPT were proposed in [4]–[7]. Zhang and Ho [4] investigated two practical receiver designs, namely time switching and power splitting, for the case of colocated receivers. For the case of separated receivers, the authors proposed an optimal transmission strategy to achieve different tradeoffs for maximal information rate versus energy transfer, which are characterized by the boundary of a so-called rate energy (R-E) region. The dynamic power splitting (DPS) and two types of practical receiver architectures (namely, separated versus integrated information and energy receivers) were proposed in [5]. The R-E performances for the two proposed receivers are further characterized by the R-E region. With receiver circuit power consumption taken into account, it is shown that the ON-OFF power splitting scheme is optimal for both receivers. Liu *et al.* [6] obtained the optimal transmit beam forming and power allocation solution by applying the technique of semidefinite relaxation to maximize the secrecy rate and the weighted sum-energy transferred to energy receivers (ERs) for the information receiver (IR). A novel EH-balancing technique for robust beamformers design in the multiuser multiple-input–single-output (MISO) broadcast system with SWIPT was

Manuscript received December 9, 2020; revised May 18, 2021; accepted July 16, 2021. Date of publication August 9, 2021; date of current version June 13, 2022. This work was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC support program (IITP-2021-2018-0-01799) supervised by the IITP (Institute for Information and communications Technology Planning and Evaluation, in part by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Trade, Industry and Energy (MOTIE) of the Republic of Korea under Grant 20214000000280, and in part by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) under Grant NRF-2020R1A2C1010929. *(Corresponding author: Jung-Ryun Lee.)*

Sengly Muy and Dara Ron are with the School of Electronics and Electrical Engineering, Chung-Ang University, Seoul 156-756, Republic of Korea (e-mail: muysengly@cau.ac.kr; drron@cau.ac.kr).

Jung-Ryun Lee is with the School of Electronics and Electrical Engineering, Chung-Ang University, Seoul 156-756, Republic of Korea, and also with the Department of Intelligent Energy and Industry, Chung-Ang University, Seoul 156-756, Republic of Korea.

Digital Object Identifier 10.1109/JSYST.2021.3098860

proposed in [7]. That article aimed to maximize the worst energy receiver harvested power under both the signal-to-interference-and-noise ratio (SINR) constraints at information receivers and total transmission power constraints at the transmitter. Gao *et al.* [8] studied the joint time resource assignment and pricing in a back-scatter-assisted RF-powered network with the consideration that the secondary gateway knows the statistical information about the harvested power of the secondary transmitter. For this purpose, the authors developed the contract model and designed an optimal contract for the secondary gateway to maximize its profits. To improve the energy efficiency, An *et al.* [9] proposed an iterative successive convex approximation to obtain the numerical solution of the joint optimization problem of the base station (BS) association and beamforming in coordinated multicell multiuser downlink systems.

On the other hand, with the increasing diversity and complexity of mobile network architectures, it is infeasible to monitor and manage the multitude of network elements. Machine learning (ML) has been successfully applied in many areas because it is capable of solving complex problem with large amounts of data, detecting anomalies, predicting future scenarios, and generally discovering the patterns that a human can miss [10]; therefore, the use of ML in future mobile networks has attracted tremendous research attention.

Reinforcement learning (RL) is one of ML algorithms, which enables the agent to learn by interacting with its environment in order to maximize the cumulative reward [11]. In RL, the agent observes the current state and take an action from the action space according to the optimal decision-making policy. Then, the agent transits to a new state and receives a reward obtained from its environment. With the key features of RL, various studies focusing on RL dealing with complex and control problems in wireless communication. In [12], the authors proposed two RL-based algorithms to address a problem formulation that jointly considers vehicle mobility and resource constraint with aiming to meet the offloading requirement of traffic vehicles. The authors of [13] formulated an efficient power allocation policy as an RL problem to optimize the transmission power to prolong battery life and maximize throughput. Zhao *et al.* [14] proposed a joint power control and channel allocation based on the RL algorithm combined with statistical channel state information (CSI) to reduce the interference adaptively.

Deep neural network (DNN) is another branch of ML algorithms, which learns multiple layers of nonlinear representations for given prediction tasks. The DNN inherently fuses the process of feature extraction with classification into learning by using the fuzzy support vector machine and enables the decision making. It is very powerful method for solving real-world problems such as automated image classification, natural language processing, human action recognition, or physics [15]. The DNN has gained more and more attention in wireless communication researchers. Lee *et al.* [16] proposed the centralized DNN to optimize the transmit power allocation in wireless-powered networks. Lee *et al.* [17] applied the convolutional neural network to optimize the transmit power allocation with the purpose of maximizing the energy efficiency.

Despite RL successes in recent years, these results suffer from the lack of scalability and disable to manage the high-dimensional problem. Deep reinforcement learning (DRL) can efficiently overcome these problems by combining both RL and DNN. DRL uses the DNN as a powerful function approximator together with the use of RL, which provides an autonomous decision-making mechanism for the learning agent. With the key advantage of DRL, the authors of [18] applied DRL to solve the joint optimization problem of the subchannel assignment and power allocation for multiusers in the nonorthogonal multiple access (NOMA) systems.

Q-learning is a specific type of RL and its deep reinforcement learning analog is deep Q network (DQN). Q-learning is a temporal difference learning method, where it has a Q-table to look for best action possible in the current state based on the Q-value function. At each state, the agent tries an action and receives the immediate reward, which is used to determine the long-term discounted reward of each state–action pairs called action-value (or Q-table). After that, the agent stores the action-value in its Q-table [19]. By trying all actions in all states repeatedly, the agent knows which action is the best for selection by judging its Q-value stored in Q-table. Deep Q-learning (DQL) is a learning method that combines the Q-leaning and DNN. In DQL, the DQN is basically constructed based on the DNN in which its input and output are the states, Q-values, respectively. This method uses the DQN to approximate the action-values together with the use of RL, which selects an action according to the optimal policy.

In this article, we aim to optimize the energy efficiency in the context of the SWIPT-based D2D-underlaid cellular networks, where a D2D receiver can simultaneously harvest energy and decode information. It is noted that, to the best of our knowledge, there is no study that applied the multiagent DRL to optimize the energy efficiency for SWIPT-based D2D-underlaid cellular networks. The main contributions of this article can be summarized as follows. First, we propose a multiagent DRL algorithm to address the energy efficiency maximization problem for SWIPT-based D2D-underlaid cellular networks. Specifically, we design the states, actions, and reward of multiagent DRL to facilitate the addressal of the nonconvex problem, and the proposed multiagent DRL algorithm enables each agent to optimize its transmit power allocation and splitting ratio in a way to minimize the loss function. Second, we compare the time-complexity of the proposed multiagent DRL and benchmark schemes. Finally, we evaluate the performances of the proposed DRL, GS, and ES schemes, which show that the proposed DRL scheme achieves a near-global optimum solution with very low computational complexity.

The rest of this article is organized as follows. Section II describes the system model and formulates the optimization problem for a SWIPT-based D2D communication-underlaid cellular networks in detail, and Section III explains the iterative approaches to optimize this problem, including GS and ES methods. Section IV explicates the proposed multiagent DRL to optimize the problem and analyzes computational complexity, and Section V shows the comparable results between GS, ES, and the proposed algorithm. Section VI concludes this article.

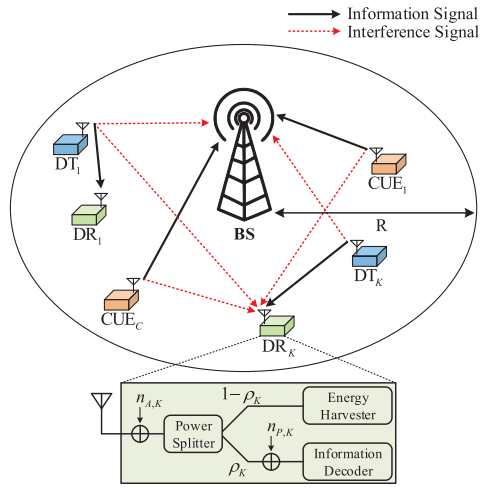

Fig. 1. System model of D2D-underlaid cellular networks.

II. System Model

We consider an SWIPT-based D2D-underlaid cellular networks wherein multiple cellular user equipments (CUEs) and multiple D2D pairs are randomly deployed over the cell coverage of a BS, as show in Fig. 1. Each D2D transceiver pair consists of one D2D transmitter and one D2D receiver. Each CUE transmits data to the BS and shares its subchannel with multiple D2D pairs. Let $\mathcal{K} = \{1, 2, \ldots, k, \ldots K\}$ and $\mathcal{C} = \{1, 2, \ldots, c, \ldots C\}$ represent the set of CUEs and D2D pairs, respectively. To perform the SWIPT functionality, we assume that all D2D receivers are adopted with power splitting policy, which splits the received power into two parts for information decoding and energy harvesting [20].

A. Cellular Communication Model

In this model, we consider uplink communication, which enables the CUE to transmit data to the BS. The signal received at the BS from the cth CUE is given by

$$y_{c,b} = \sqrt{p_c} g_{c,b} x_{c,b} + \sum_{k=1}^{K} \sqrt{p_k} g_{k,b} x_{k,b} + n_{A,b} \tag{1}$$

where p_c and p_k are the transmit powers of the cth CUE and the kth D2D transmitter, respectively. Furthermore, $x_{c,b}$ and $x_{k,b}$ are the message signals that are transmitted from the cth CUE and the kth D2D transmitter to the BS, respectively. $n_{A,b} \sim \mathcal{CN}(0, \sigma^2_{A,b})$ is the additive white Gaussian noise (AWGN). $g_{c,b}$ and $g_{k,b}$ are the channel gains of the cth CUE and the BS link, and the kth D2D transmitter and the BS link, respectively. Each channel coefficient suffers from small-scale-fading and path loss. According to (1), the SINR received at the BS from the cth CUE is given by

$$\Gamma_{c,b} = \frac{p_c \left| g_{c,b} \right|^2}{\sum_{k=1}^{K} p_k \left| g_{k,b} \right|^2 + \sigma^2_{A,b}} \tag{2}$$

where $|g_{c,b}|^2 = |\tilde{g}_{c,b}|^2 (d_{c,b}^{-\eta})$ and $|g_{c,b}|^2 = |\tilde{g}_{k,b}|^2 (d_{k,b}^{-\eta})$. $|\tilde{g}_{c,b}|^2$ and $|\tilde{g}_{k,b}|^2$ follow the independent Rician distribution with mean μ because the line-of-sight propagation is appropriate for energy harvesting [21]. The $d_{c,b}$ and $d_{k,b}$ denote the distances from the cth CUE to BS, and from the kth D2D transmitter to BS, respectively, and η is the path-loss exponent.

B. Direct D2D Communication Model

In this model, the D2D transmitters directly communicate with their receivers using the cellular channel. Therefore, the D2D receiver should be interfered by the following two scenarios:

1) the D2D receivers may receive signal from CUE transmitters;
2) D2D receivers may receive signals from the D2D transmitters of other D2D pairs.

Let $\rho_k \in [0, 1]$ be the power splitting ratio. The kth D2D receiver is capable of decoding information and harvesting energy from the received signal with ratio of ρ_k and $1 - \rho_k$, respectively. Therefore, the information decoded by the kth D2D receiver can be calculated as

$$y_k^{\text{ID}} = \sqrt{\rho_k} \left(\sqrt{p_k} g_{k,k} x_{k,k} + \sum_{k' \neq k}^{K} \sqrt{p_{k'}} g_{k',k} x_{k',k} + \sum_{c=1}^{C} \sqrt{p_c} g_{c,k} x_{c,k} + n_{A,k} \right) + n_{P,k} \tag{3}$$

and the energy harvesting is given by

$$y_k^{\text{EH}} = \sqrt{1 - \rho_k} \left(\sqrt{p_k} g_{k,k} x_{k,k} + \sum_{k' \neq k}^{K} \sqrt{p_{k'}} g_{k',k} x_{k',k} + \sum_{c=1}^{C} \sqrt{p_c} g_{c,k} x_{c,k} + n_{A,k} \right) \tag{4}$$

where $p_k, p_{k'}$, and p_c are the transmission powers of the kth D2D transmitter, the k'th D2D transmitter of other D2D pairs, and the cth CUE, respectively. $x_{k,k}, x_{k',k}$, and $x_{c,k}$ are the message signals that are transmitted from the kth D2D transmitter, the k'th D2D transmitter of other D2D pairs, and the cth CUE to the kth D2D receiver, respectively. $n_{A,k} \sim \mathcal{CN}(0, \sigma^2_{A,k})$ and $n_{P,k} \sim \mathcal{CN}(0, \sigma^2_{P,k})$ are the AWGN and the additive noise received at the kth D2D receiver, respectively. $g_{k,k}$, $g_{k',k}$, and $g_{c,k}$ are the channel coefficients of the kth D2D transmitter and the kth D2D receiver link, the k'th D2D transmitter and the kth D2D receiver link, and the cth CUE and the kth D2D receiver link, respectively. According to the information decoding in (3), the SINR of the kth D2D link is given by

$$\Gamma_{k,k} = \frac{\rho_k p_k \left| g_{k,k} \right|^2}{\rho_k \left[\sum_{k' \neq k}^{K} p_{k'} \left| g_{k',k} \right|^2 + \sum_{c=1}^{C} p_c \left| g_{c,k} \right|^2 + \sigma^2_{A,k} \right] + \sigma^2_{P,k}} \tag{5}$$

where $|g_{k,k}|^2 = |\tilde{g}_{k,k}|^2 (d_{k,k}^{-\eta})$, $|g_{k',k}|^2 = |\tilde{g}_{k',k}|^2 (d_{k',k}^{-\eta})$, and $|g_{c,k}|^2 = |\tilde{g}_{c,k}|^2 (d_{c,k}^{-\eta})$. Here, $|\tilde{g}_{k,k}|^2$, $|\tilde{g}_{k',k}|^2$, and $|\tilde{g}_{c,k}|^2$ follow the independent Rician distribution with mean μ. $d_{k,k}$, $d_{k',k}$, and $d_{c,k}$ are the distance from the kth D2D transmitter, the

k'th D2D transmitter, and the cth CUE to the kth D2D receiver, respectively.

C. Total Throughput and Energy Dissipation

From (2) and (5), the throughputs of the cth CUE and kth D2D pair are given by

$$R_c = B_c \log_2 (1 + \Gamma_{c,b}) \tag{6}$$

$$R_k = B \log_2 (1 + \Gamma_{k,k}) \tag{7}$$

where B_c and B are the bandwidths allocated to the cth CUE and kth D2D pair, respectively. It is noted that the aggregate bandwidth allocated to multiple CUEs are reused by each D2D pair; therefore, the bandwidth of kth D2D pair is given by $B = \sum_{c=1}^{C} B_c$. From (4), the energy harvested by the kth D2D receiver can be calculated as

$$\mathrm{EH}_k = (1 - \rho_k)\, \xi \left(\sum_{c=1}^{C} p_c \left| g_{c,k} \right|^2 + \sum_{l=1}^{K} p_l \left| g_{l,k} \right|^2 \right) \tag{8}$$

where ξ is the energy conversion efficiency. Therefore, the total throughput and energy harvesting can be formulated as

$$R_{\mathrm{sum}}(\vec{p}, \vec{\rho}) = \sum_{c=1}^{C} R_c + \sum_{k=1}^{K} R_k \tag{9}$$

$$\mathrm{EH}_{\mathrm{sum}}(\vec{p}, \vec{\rho}) = \sum_{k=1}^{K} \mathrm{EH}_k \tag{10}$$

where $\vec{p} = (p_1, p_2, \ldots, p_K)$ and $\vec{\rho} = (\rho_1, \rho_2, \ldots, \rho_K)$ are the transmit power vector of D2D transmitters and the power splitting vector of D2D receivers, respectively. According to (10), we can formulate the total energy dissipation of the network as

$$\mathrm{ED}_{\mathrm{sum}}(\vec{p}, \vec{\rho}) = \sum_{k=1}^{K} (p_k + P_C - \mathrm{EH}_k(\vec{p}, \vec{\rho})) + \sum_{c=1}^{C} p_c \tag{11}$$

where P_C is the constant circuit power consumption at the D2D pair.

D. Problem Formulation

Energy efficiency is defined as the data rate per unit energy in [bits/hz/joule], which is a measure of how efficiently energy can be used for transferring information. Therefore, the energy efficiency of a network is defined as $\mathrm{EE}(\vec{p}, \vec{\rho}) = \frac{R_{\mathrm{sum}}(\vec{p}, \vec{\rho})}{\mathrm{ED}_{\mathrm{sum}}(\vec{p}, \vec{\rho})}$. Thus, we can formulate the optimization problem to determine the optimal transmit power and power splitting ratio $(\vec{p}^*, \vec{\rho}^*)$ that maximize the energy efficiency under the constraint of maximum transmit power $P_{\max}$, and minimum data rate requirement for CUEs $R_{\min}^{\mathrm{CUE}}$ and D2D users $R_{\min}^{\mathrm{D2D}}$, which is given by

$$\max_{\vec{p}, \vec{\rho}} \mathrm{EE}(\vec{p}, \vec{\rho}) \tag{12}$$

$$\begin{aligned} \text{s.t. } & C_1 : 0 \le p_k \le P_{\max}, k \in \{1, 2, \ldots, K\} \\ & C_2 : 0 \le \rho_k \le 1, k \in \{1, 2, \ldots, K\} \\ & C_3 : R_c \ge R_{\min}^{\mathrm{CUE}} \\ & C_4 : R_k \ge R_{\min}^{\mathrm{D2D}}. \end{aligned}$$

The objective function of (12) is a nonconvex problem with nonlinear constraints; therefore, it is impossible to determine the close-form solution for the optimal transmit power and power splitting ratio. However, the local and global optimum solution can be obtained numerically using the gradient search (GS) and exhaustive search (ES), respectively. To obtain the global optimum solution using ES, the transmit power and power splitting ratio are quantized with equally spaced values, and all combinations of quantized values are evaluated to find the optimal one. Unfortunately, ES is infeasible for large numbers of D2D pairs and CUEs owing to the resulting large time complexity. DRL, which combines both RL and DNN, would be a suitable method to address the problem with high-dimensional input space because DRL can provide an autonomous decision-making mechanism for the agent together with the use of the DNN as a powerful function approximator.

$$\begin{aligned} P(\vec{p}, \vec{\rho}) = \frac{1}{t} \Bigg\{ \sum_{k=1}^{K} \big[& \ln (P_{\max} - p_k) + \ln (p_k) + \ln (1 - \rho_k) \\ & + \ln (\rho_k) + \ln \left(R_k - R_{\min}^{\mathrm{D2D}} \right) \big] + \sum_{c=1}^{C} \ln \left(R_c - R_{\min}^{\mathrm{CUE}} \right) \Bigg\}. \end{aligned} \tag{13}$$

III. Optimization-Based Iteration Method

In this section, we introduce ES and GS with barrier to find numerically the global and local optimal values of $\vec{p}$ and $\vec{\rho}$, respectively.

A. Exhaustive Search (ES)

The ES algorithm is a general global optimization technique, which checks all possible enumerating candidates that satisfies the problem's statement. In our problem, the control parameters, $\vec{p}$ and $\vec{\rho}$, are quantized with S equally spaced length, and all possible solutions are examined to determine the maximum value of the objective function while satisfying all constraints. The ES can reach the global solution when S reach infinity with an exponentially increasing time complexity, which is given by $O(S^{2K})$, where K is the number of D2D pairs.

B. Gradient Search (GS) With Barrier

The GS, also known as the steepest ascent/descent method, is the simplest and most fundamental optimization method for unconstrained optimization. Because our objective function is a constrained optimization problem, we need to convert this problem to an unconstrained optimization problem before applying the GS algorithm. To transform our objective function from a constrained to unconstrained optimization problem, we use the penalty technique in which the constraint is added to the objective function [22]. In our problem, we employ the logarithmic barrier function with the penalty parameter $t > 0$. By adding constraints such as C_1, C_2, C_3, and C_4, to the objective function, we can construct a new penalty function as expressed in (13). Then, we can convert the original problem to a smooth approximation function, which has eventually convergence to

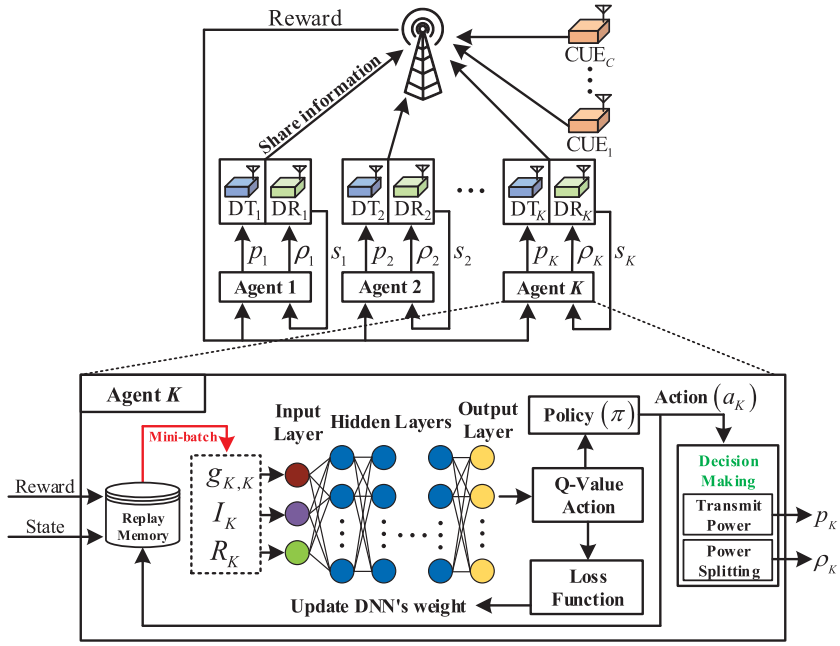

Fig. 2. Proposed multiagent DRL-based power control.

Algorithm 1: GS algorithm.

1: Initialize p_i, ρ_i randomly in feasible region
2: **set** β_p, β_ρ, and δ
3: **repeat** :
4: $\quad \vec{p}_{\text{new}} = \vec{p}_{\text{old}} + \beta_p \times \frac{\partial \mathcal{B}}{\partial \vec{p}}$
5: $\quad \vec{\rho}_{\text{new}} = \vec{\rho}_{\text{old}} + \beta_\rho \times \frac{\partial \mathcal{B}}{\partial \vec{\rho}}$
6: **until** $|\mathcal{B}_{\text{new}}(\vec{p}_{\text{new}}, \vec{\rho}_{\text{new}}) - \mathcal{B}(\vec{p}_{\text{old}}, \vec{\rho}_{\text{old}})| < \delta$

the solution of the original problem, which is given by

$$\mathcal{B}(\vec{p}, \vec{\rho}) = EE(\vec{p}, \vec{\rho}) + P(\vec{p}, \vec{\rho}). \tag{14}$$

To obtain the suboptimal solution for (14), we use the simple GS technique, which is represented in Algorithm 1. It is noted that the step sizes β_p and β_ρ, which are used to literately update $\vec{p}$ and $\vec{\rho}$, respectively, should be different with each other because the available range of $\vec{p}$ is quite different from $\vec{\rho}$. We use the tolerance error δ to break the loop of the GS algorithm, then the time complexity of GS has the upper bound of $O(\delta^{-2})$ [23], [24].

IV. Proposed Multiagent Deep Reinforcement Learning Algorithm

A. Markov Decision Process (MDP)

In RL, MDP enables the agent to transition from the current state to the next state according to the policy. From the definition in [25], an MDP is defined as a tuple $\langle \mathcal{S}_k, \mathcal{A}_k, \mathcal{R}_k, \mathcal{P}_k \rangle$, where $\mathcal{S}_k$ is a finite set of environment states of the kth agent, $\mathcal{A}_k$ is a finite set of agent actions, and $\mathcal{R}_k$ is the reward function. $\mathcal{P}_k$ is the transition probability of moving from the current state $s_k^t \in \mathcal{S}_k$ to the next state $s_k^{t+1} \in \mathcal{S}_k$ according to the policy, which is denoted by $\mathrm{P}^\pi(s_k^{t+1}|s_k^t)$. $\mathrm{P}(s_k^{t+1}|s_k^t, a_k^t)$ is the transition probability from the current state s_k^t to the next state s_k^{t+1} given the action $a_k^t \in \mathcal{A}_k$, and $\pi(a_k^t|s_k^t)$ is a mapping from the current state s_k^t to the action a_k^t, called the policy. Therefore, $\mathrm{P}^\pi(s_k^{t+1}|s_k^t)$ is defined as the transition probability $\mathrm{P}(s_k^{t+1}|s_k^t, a_k^t)$ weighted by the policy $\pi(a_k^t|s_k^t)$. The goal of an MDP aims to find an optimal policy π^* to maximize the reward function $\mathcal{R}_k$.

B. Multiagent DRL

In our article, we propose a design of multiagent DRL-based power control, where each agent can control the transmit power of the D2D transmitter and power splitting ratio of the D2D receiver, as shown in Fig. 2. Each agent is equipped with a classical Q-learning algorithm, and it learns without cooperating with the other agents. We assume that all devices (D2D pairs and CUEs) share their necessary information, such as data rate, the transmit power, and the amount of harvested energy, with the BS. With the necessary information received from all devices, the BS calculates the future reward. After that, the BS gives feedback on the reward to all agents. The agent is required to store the reward in the replay memory, and then, it randomly selects the reward, current states, actions, and the next state from the replay memory to train the neural networks. Then, each agent separates the overall data in the replay memory, including the rewards, the current states, the actions, and the next states, into multiple minibatch samples and uses them as the input data of the DNN to train the neural networks in a way to minimize the loss function. The output of the DNN is the approximate Q-value. The proposed scheme provides an autonomous decision-making mechanism to select the optimal transmit power and power splitting ratio according to the maximum Q-value.

In the proposed multiagent DRL algorithm, we define the agents, states, actions, and reward function as follows.

1) Agent: The system consists of K agents corresponding to the number of D2D pairs.
2) State: The state of the kth D2D pair is defined by

$$s_k^t = \{g_{k,k}, I_k, R_k\} \tag{15}$$

where $g_{k,k}$, I_k, and R_k are the channel gain of the kth D2D pair, the interference of the kth D2D pair received from other transmitters, and the throughput of the kth D2D pair, respectively. It is noted that $I_k = \sum_{k' \neq k}^{K} p_{k'} |g_{k',k}|^2 + \sum_{c=1}^{C} p_c |g_{c,k}|^2$ is the interference from another D2D transmitter and CUE transmitter.
3) Action: In our multiagent DRL scheme, an agent corresponds to a D2D pair and determines the transmit power and the power splitting ratio to optimize the energy efficiency. Here, it is assumed that the information about the transmit power and the power splitting ratio is included in the packet sent from the D2D transmitter to the D2D receiver, and the amount of the harvested energy is measured by the D2D receiver using the power splitting ratio received from the D2D transmitter.
Therefore, we define the action as the set of the transmit power and power splitting ratio, which is given by

$$a_k^t = \{p_k, \rho_k\} \tag{16}$$

where $p_k \in \{0, \frac{P_{\max}}{N-1}, \frac{2P_{\max}}{N-1}, \ldots, P_{\max}\}$ and $\rho_k \in \{0, \frac{1}{M-1}, \frac{2}{M-1}, \ldots, 1\}$ are defined as N and M discrete levels, respectively.
4) Reward: Once an action is taken, the environment feedbacks a reward to the agent that can score how good the designed goal is achieved. In our system, the data rate of CUEs and D2D pairs should be never less than the minimum throughput requirement to guarantee the network connectivity. Therefore, we define the reward function as follows:

$$r_k^t = \begin{cases} \text{EE}\left(\vec{p}, \vec{\rho}\right), & C_3 \text{ and } C_4 \\ -100, & \text{otherwise.} \end{cases} \tag{17}$$

From (17), the reward function is equal to the energy efficiency $\text{EE}(\vec{p}, \vec{\rho})$ when its condition satisfies the inequality constraints C_3 and C_4 in (12). Otherwise, the reward is set to -100. It is noted that the small negative value of the reward will quite affect the DRL's performance.

The multiagent Q-learning algorithm finds optimal Q-value $Q^*(s_k, a_k)$ in a recursive way after receiving a transition information $\langle s_k, a_k, s'_k, \pi_k \rangle$, where $s_k^t = s_k \in \mathcal{S}_k$ and $s_k^{t+1} = s'_k \in \mathcal{S}_k$ are the environment states observed by the agent k at time t and $t+1$, respectively; $a_k^t = a_k \in \mathcal{A}_k$ and $a_k^{t+1} = a'_k \in \mathcal{A}_k$ are the kth agent's action at time slot t, $t+1$, respectively. The Q-value can be updated as follows:

$$\begin{aligned} Q_k\left(s_k, a_k\right) &\leftarrow\!\rightarrow Q_k\left(s_k, a_k\right) \\ &+ \alpha[r_k + \gamma \max_{a \in \mathcal{A}_k} Q_k\left(s'_k, a\right) - Q_k\left(s_k, a_k\right)] \end{aligned} \tag{18}$$

where α is a learning rate and γ is a discount factor. It is noted that when the state and action spaces are too large, the application of Q-learning becomes impractical because of two significant problems: the lookup table storage should be sufficiently large, and the training process converges slowly. To tackle this issue, the DNN has been proposed to estimate the Q-value function.

In our article, we design the fully connected DNN with zero bias. The DNN is separated into three types of layers, including the input layer, the hidden layer, and the output layer. For our model, the input layer consists of three neurons that correspond to three different types of input data, including channel gain, interference, and throughput. We train the neural networks with all samples in minibatch and compute the average of the DNN output. In our scheme, the DNN is used to approximate the Q-value; therefore, the output layer of the DNN represents the Q-value. The output layer of the DNN consists of $N \times M$ neurons, where N and M are the discrete levels of the transmit power and power splitting ratio, respectively. There are L hidden layers in which the number of neurons of each hidden layer is equal to the number of neurons of the output layer. Furthermore, we use the ReLU activation function to activate every layer of the DNN.

We denote the estimated Q-value function of the kth agent as $Q_k(s, a; \boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the weight of the DNN. The neural network's weights are updated using the backpropagation of error gradient, which is defined as the partial derivative loss function with respect to weights. The loss function is given by

$$L_k = E\left[\left(y_k - Q_k\left(s_k, a_k; \boldsymbol{\theta}\right)\right)^2\right] \tag{19}$$

where y_k is the target value estimated by

$$y_k = r_k + \gamma \max_{a \in \mathcal{A}_k} Q_k\left(s'_k, a; \boldsymbol{\theta}_{\text{target}}\right). \tag{20}$$

The model's parameters (weights) are updated via training the DNN by picking a random sample minibatch $\hat{D}$ from the replay memory D that are used as the input data. By using the chain rule, the error gradient can be calculated as

$$\begin{aligned} &\nabla_\theta L_k \\ &\approx \frac{1}{\left|\hat{D}\right|} \sum_{\left\{s_k, a_k, r_k^t, s'_k\right\} \in \hat{D}} 2\left(y_k - Q_k(s_k, a_k; \boldsymbol{\theta})\right) \nabla_\theta Q_k(s_k, a_k; \boldsymbol{\theta}). \end{aligned} \tag{21}$$

At each iteration, each agent updates its own policy according to the Q-value estimation. Then, the agent selects an action from the action space using the ϵ-greedy policy, which is given by

$$a'_k = \begin{cases} \underset{a \in \mathcal{A}_k}{\text{argmax}} Q_k\left(s_k, a, \boldsymbol{\theta}\right), & \text{with probability } 1 - \epsilon \\ \text{random action}, & \text{with probability } \epsilon. \end{cases} \tag{22}$$

The proposed algorithm for energy efficiency optimization in the context of SWIPT-based D2D-underlaid cellular networks is summarized in Algorithm 2.

C. Time Complexity Analysis

In this section, we analyze the time computation complexity of multiagent DRL. All K agents of DRL are computed in parallel. Therefore, the time complexity of multiagent is same as that of a single agent. In the DNN, the number of neurons in the input and output layer is equal to the dimension of the state n_{state}, and the discrete size of action space $n_{\text{action}} = N \times M$, respectively. In our proposed scheme, the DNN is designed as a fully connected network with zero bias and consists of L hidden layers. Let

Algorithm 2: Multi-Agent Deep Q-Learning.

Initialization:
for all $k \in K$ **do**
  Randomly initialize the Q-network $Q_k(s_k, a_k; \boldsymbol{\theta})$
  Randomly initialize the target Q-network $\hat{Q}_k(s_k, a_k; \boldsymbol{\theta}_{\text{target}})$
  Initialize the replay memory D with capacity C
end for
while not convergence **do**
  forall k^{th} D2D pair; $k \in K$ **do**
    for Iteration **do**
      Select action a'_k using ϵ-greedy policy via $Q_k(s_k, a_k, \boldsymbol{\theta})$
      Observe the new state s'_k by measuring channel gain, interference level, and throughput
      Observe the reward r_k^t
      Store transition (s_k, a_k, r_k^t, s'_k) in D
      If the replay memory D is full **then**
        Sample a mini-batch of K transitions from memory D
        Train DNN using mini batch gradient descent to minimize the loss function (19)
      end if
      Update DNN weight $\boldsymbol{\theta}_{\text{target}}$
      Update the state $s_k = s'_k$
  endfor
end while

TABLE I
COMPUTATIONAL COMPLEXITY COMPARISON

Parameter	Value
Cell radius R	1 Km
Number of CUEs C	2 CUEs
Number of D2D pairs K	$\{3, 4, ..., 7\}$ pairs
CUE Tx power P_c	23 dBm
D2D maximum Tx power P_k	$\{11, 14, ..., 32\}$ dBm
Circuit power consumption P_C	28 dBm
Noise power σ^2, σ_A^2	-70 dBm, -100 dBm
Energy conversion factor ξ	50%
Rician fading gain	5 dBm
Path-loss exponent m	3.6
Total bandwidth B	1 MHz
$R_{\min}^{CUE}$ and $R_{\min}^{D2D}$	1 Mbps

V_l be the number of neurons of the lth hidden layer. The time complexity for computing the lth hidden-to-$(l+1)$th hidden layer and the Lth hidden layer-to-output of the DNN is given by $O(V_{l-1}V_l + V_lV_{l+1})$ and $O(V_{L-1}V_L + V_L n_{\text{action}})$, respectively, where $1 \leq l \leq L-1$, [26]. In this article, the number of neurons in the hidden layer is equal to the number of actions n_{action}; therefore, the time complexity in a feed-forward network can be calculate as $O(L \times n_{\text{action}}^2)$. We assume that the proposed algorithm converges after T epochs with I iterations. Therefore, the total number of sample in the simulation process is $T \times I$, and thus, the time complexity of the apprenticeship learning-based algorithm is $O(T \times I \times L \times n_{\text{action}}^2)$. Table I summarizes the comparison of the computational complexity with the running time of the ES, GS, and DRL by giving the number of D2D pairs $K = 3$, the quantization parameter in ES $S = 10^3$, tolerance error of GS $\delta = 10^{-5}$, the number of hidden layers $L = 10$, and the number of action $n_{\text{action}} = N \times M = 100$.

TABLE II
SIMULATION PARAMETERS FOR UNDERLAY D2D

Parameter	Value
Learning rate α	0.01
Number of hidden layer L	10 layers
N and M	10 levels
ϵ-greedy ϵ	0.1
Discount factor γ	0.99
Replay memory size D	1000
Mini batch size	25
Optimizer	SGD
Activation function	ReLu

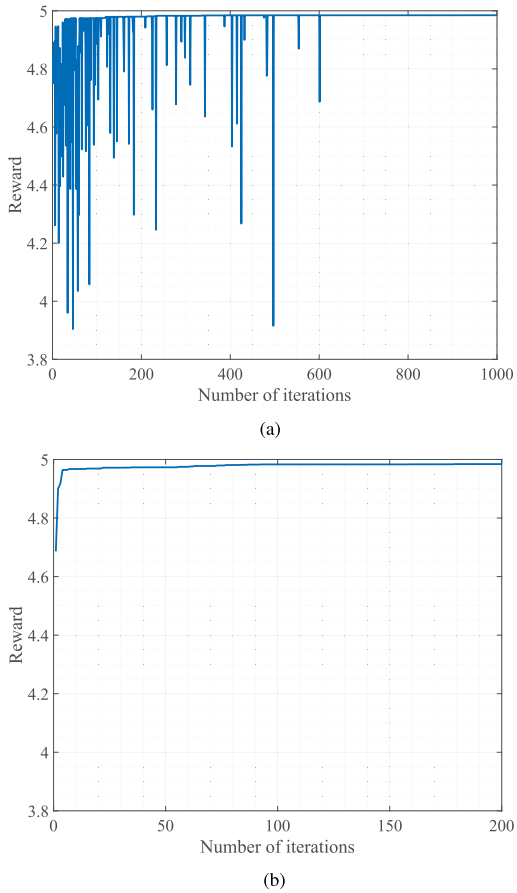

Fig. 3. Training and testing convergence of an agent.

V. SIMULATION RESULTS AND DISCUSSION

Table II summarizes the simulation parameters that are used to evaluate the performance of the proposed DRL algorithm in the context of SWIPT-based D2D-underlaid uplink cellular networks. The CUEs and D2D pairs are randomly deployed over the cell coverage, and the distance between D2D transmitter and receiver is set following the normal distribution with mean of 10 m. The performances of ES, GS, and multiagent DRL are evaluated using MATLAB 2020b, which was installed in a PC with AMD Ryzen 5 3600 6-core processors CPU, NVIDIA

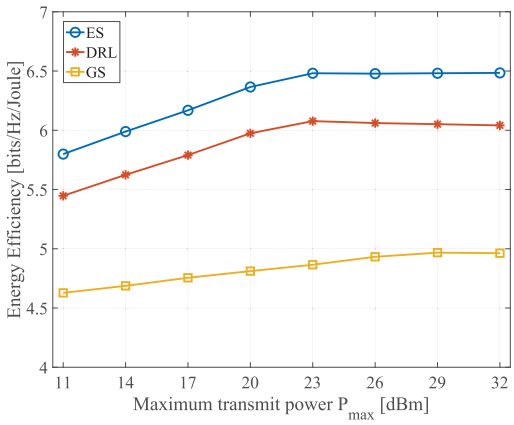

Fig. 4. Energy efficiency versus $P_{\max}$ when $K = 3$.

TABLE III
SIMULATION PARAMETERS FOR PROPOSED ALGORITHM

Algorithm	Computation complexity	The running time
ES	$O\left(S^{2K}\right)$	8.5 days
GS	$O\left(\delta^{-2}\right)$	0.025056 seconds
DRL	$O\left(T \times I \times L \times n_{\text{action}}^2\right)$	0.053323 seconds

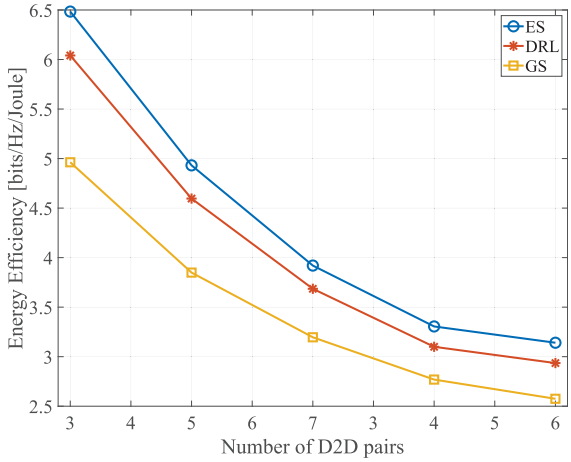

Fig. 5. Energy efficiency versus K when $P_{\max} = 32$ dBm.

GeForce RTX 2080 Ti GPU, and 16 GB RAM. To evaluate the performance of ES, the numbers of equispaced divisions of p and ρ are set to 1000. For the GS, we set the penalty parameter $t = 100$, the learning rate $\beta_p = P_{\max}/10^3$ and $\beta_\rho = 1/10^3$, and the tolerance error δ is set to 10^{-5}. In the proposed multiagent DRL, we use the fully connected neural network for training the action-value function. This network consists of an input layer with three neurons for state space, ten hidden layers, and an output layer with $M \times N$ neurons for action space. The parameters related to the multiagent DRL model are summarized in Table III. In Figs. 4 and 5, we use Monte Carlo simulation runs for the results of ES, DRL, and GS with 10^3, 10^4, and 10^4 different channel gains, respectively.

Fig. 3(a) shows the example of training convergence in terms of the reward (or energy efficiency) and the number of iterations with $(K = 3)$ and $P_{\max} = 11$ dBm for a fixed channel gain. The result shows that the proposed multiagent DRL converges after 600 rounds of iteration because, at this iteration, the optimal solution of the neural network's weights and policy have been obtained. Here, it is noticed that, with the optimal policy, the agent can select the optimal action, which results in receiving the maximum reward. Fig. 3(b) shows the testing convergence in terms of the reward and number of iteration. In the testing phase, each agent uses the trained model to calculate the transmit power and splitting ratio and obtains the reward value. After that, the agent uses the transmit power and splitting ratio calculated at iteration t to update the agent state at iteration $t + 1$, which is defined as the function of the channel gain, interference, and data rate. In this way, the DRL state is updated at every iterations. As a result, the proposed algorithm converges after 100 rounds of iterations, with very low time complexity.

Fig. 4 shows the energy efficiency as a function of the maximum transmit power $(P_{\max})$ with three D2D pairs $(K = 3)$. The result show that the energy efficiency increases as the maximum transmit power increases. However, it converges when the maximum transmit power is greater than or equal to 23 dBm. This result indicates that the use of extra transmit power beyond 23 dBm causes a loss in energy efficiency. In the case of the GS, suboptimal solutions of $\vec{p}$ and $\vec{\rho}$ are found and the duality gap causes a degradation in the energy efficiency up to 25% compared to the ES. However, DRL achieves a near-optimal energy efficiency with less than 10% loss, which shows that the DRL outperforms the GS with 15% increase in energy efficiency. We can see that the gap between ES and DRL increases when the maximum transmits power increases because we use the fixed quantization level of the transmit power in the action space.

Fig. 5 shows the energy efficiency versus the number of D2D pairs (K) with $P_{\max} = 32$ dBm. A larger K results in degrading the energy efficiency due to increased interference. However, DRL achieves the near-global optimum with higher performance compared to that of GS.

VI. CONCLUSION

This article studied the power management of a wireless D2D under cellular networks for maximizing energy efficiency where multiple D2D pairs adopt the SWIPT functionality with a power splitting policy. To solve this problem, we first use optimization-based-iterative techniques such as ES and GS with barrier to find the global optimum and suboptimum, respectively. We propose a multiagent DRL that can self-organize the transmit power and power splitting ratio to get the system's optimum energy efficiency. This proposed method considers that each agent can only know its channel gain, interference power, and required minimum throughput. Through the simulation results, we show that multiagent DRL can achieve near-global optimal with low computation complexity compared to the optimization-based iterative algorithm techniques.

In our article, we applied the multiagent DRL to optimize the transmit power and power splitting ratio with considering the perfect channel coefficient estimation. However, it is difficult to obtain a perfect channel coefficient in practice due to the channel estimation error, feedback delay, quantization error, and channel variation caused by the fast fading [9]. Therefore, we will build

a plan to study the application of federated learning on energy efficiency maximization for SWIPT-based D2D communication underlaid cellular networks considering the imperfect channel coefficient estimation, as a future work.

REFERENCES

[1] F. Jameel, Z. Hamid, F. Jabeen, S. Zeadally, and M. A. Javed, "A survey of device-to-device communications: Research issues and challenges," *IEEE Commun. Surv. Tuts.*, vol. 20, no. 3, pp. 2133–2168, Jul.–Sep. 2018.

[2] P. Sun, K. G. Shin, H. Zhang, and L. He, "Transmit power control for D2D underlaid cellular networks based on statistical features," *IEEE Trans. Veh. Technol.*, vol. 66, no. 5, pp. 4110–4119, May 2017.

[3] F. Wang, C. Xu, L. Song, and Z. Han, "Energy-efficient resource allocation for device-to-device underlay communication," *IEEE Trans. Wireless Commun.*, vol. 14, no. 4, pp. 2082–2092, Apr. 2015.

[4] R. Zhang and C. K. Ho, "MIMO broadcasting for simultaneous wireless information and power transfer," *IEEE Trans. Wireless Commun.*, vol. 12, no. 5, pp. 1989–2001, May 2013.

[5] X. Zhou, R. Zhang, and C. K. Ho, "Wireless information and power transfer: Architecture design and rate-energy trade off," *IEEE Trans. Commun.*, vol. 61, no. 11, pp. 4757–4767, Nov. 2013.

[6] L. Liu, R. Zhang, and K. C. Chua, "Secrecy wireless information and power transfer with MISO beamforming," *IEEE Trans. Signal Process.*, vol. 62, no. 7, pp. 1850–1863, Apr. 2014.

[7] H. Zhang, K. Song, Y. Huang, and L. Yang, "Energy harvesting balancing technique for robust beamforming in multiuser MISO SWIPT system," in *Proc. Int. Conf. Wireless Commun. Signal Process.*, Oct. 2013, pp. 1–5.

[8] X. Gao, D. Niyato, P. Wang, K. Yang, and J. An, "Contract design for time resource assignment and pricing in backscatter-assisted RF powered networks," *IEEE Wireless Commun. Lett.*, vol. 9, no. 1, pp. 42–46, Jan. 2020.

[9] J. An, Y. Zhang, X. Gao, and K. Yang, "Energy-efficient base station association and beamforming for multi-cell multiuser systems," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2841–2854, Apr. 2020.

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

[11] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 1998.

[12] Y. Wang, K. Wang, H. Huang, T. Miyazaki, and S. Guo, "Traffic and computation co-offloading with reinforcement learning in fog computing for industrial applications," *IEEE Trans. Ind. Inform.*, vol. 15, no. 2, pp. 976–986, Feb. 2019.

[13] A. Masadeh, Z. Wang, and A. E. Kamal, "Reinforcement learning exploration algorithms for energy harvesting communications systems," in *Proc. IEEE Int. Conf. Commun.*, 2018, pp. 1–6.

[14] G. Zhao, Y. Li, C. Xu, Z. Han, Y. Xing, and S. Yu, "Joint power control and channel allocation for interference mitigation based on reinforcement learning," *IEEE Access*, vol. 7, pp. 177254–177265, 2019.

[15] W. Samek, A. Binder, G. Montavon, S. Lapuschkin, and K.-R. Muller, "Evaluating the visualization of what a deep neural network has learned," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 18, no. 11, pp. 2660–2673, Nov. 2017.

[16] K. Lee, J. Lee, and H. Choi, "Learning-based joint optimization of transmit power and harvesting time in wireless-powered networks with co-channel interference," *IEEE Trans. Veh. Technol.*, vol. 69, no. 3, pp. 3500–3504, Mar. 2020.

[17] W. Lee, M. Kim, and D. Cho, "Deep power control: Transmit power control scheme based on convolutional neural network," *IEEE Commun. Lett.*, vol. 22, no. 6, pp. 1276–1279, Jun. 2018.

[18] X. Wang, Y. Zhang, R. Shen, Y. Xu, and F.-C. Zheng, "DRL-based energy-efficient resource allocation frameworks for uplink NOMA systems," *IEEE Internet Things J.*, vol. 7, no. 8, pp. 7279–7294, Aug. 2020.

[19] C. Watkins and P. Dayan, "Q-learning," *Mach. Learn.*, vol. 8, pp. 279–292, 1992.

[20] K. Xu, Z. Shen, Y. Wang, X. Xia, and D. Zhang, "Hybrid time-switching and power splitting SWIPT for full-duplex massive MIMO systems: A beam-domain approach," *IEEE Trans. Veh. Technol.*, vol. 67, no. 8, pp. 7257–7274, Aug. 2018.

[21] B. Chung, K. Lee, and D. Cho, "Proportional fair energy-efficient resource allocation in energy-harvesting-based wireless networks," *IEEE Syst. J.*, vol. 12, no. 3, pp. 2106–2116, Sep. 2018.

[22] W. Sun and Y.-X. Yuan, *Optim. Theory and Methods*. New York, Ny, USA: Springer, 2006.

[23] C. Cartis, N. I. M. Gould, and P. H. L. Toint, "On the complexity of steepest descent, newton's and regularized newton's methods for nonconvex unconstrained optimization problems," *SIAM J. Optim.*, vol. 20, no. 6, pp. 2833–2852, Oct. 2010.

[24] R. Sedgewick and K. Wayne, *Computer Science: An Interdisciplinary Approach*. Boston, MA, USA: Addison-Wesley, 2016.

[25] C. H. Yong and R. Miikkulainen, "Cooperative coevolution of multi-agent systems," Dept. of Comput. Sci., Univ. Texas at Austin, Austin, TX, USA, Tech. Rep. AI0 1–287, 2001.

[26] Y. Du, F. Zhang, and L. Xue, "A kind of joint routing and resource allocation scheme based on prioritized memories-deep Q. network for cognitive radio Ad Hoc networks," *Sensors*, vol. 18, no. 7, p. 2119, 2018.

Sengly Muy received the B.S. degree in electrical engineering from the Institute of Technology of Cambodia, Phnom Penh, Cambodia, in 2018. He is currently working toward the M.S. and Ph.D. integrated program with the School of Intelligent Energy and Industry, Chung-Ang University, Seoul, Republic of Korea.

His current research interests include performance optimization in wireless networks, artificial intelligence, and machine learning.

Dara Ron received the B.S. degree in electrical engineering from the Institute of Technology of Cambodia, Phnom Penh, Cambodia, in 2017. He is currently working toward the M.S. and Ph.D. integrated program with the School of Intelligent Energy and Industry, Chung-Ang University, Seoul, Republic of Korea.

His current research interests include bioinspired algorithms, LoRaWAN protocol, and artificial-intelligence-based wireless networks.

Jung-Ryun Lee (Senior Member, IEEE) received the B.S. and M.S. degrees in mathematics from the Seoul National University, Seoul, South Korea, in 1995 and 1997, respectively, and the Ph.D. degree in electrical and electronics engineering from the Korea Advanced Institute of Science and Technology, Daejeon, South Korea, in 2006.

From 1997 to 2005, he was a Chief Research Engineer with LG Electronics, Korea. From 2006 to 2007, he was a Full-Time Lecturer in electronic engineering with the University of Incheon. Since 2008, he has been a Professor with the School of Electrical and Electronics Engineering, Chung-Ang University, Seoul, South Korea. His research interests include low energy networks and algorithms, bioinspired autonomous networks, and artificial-intelligence-based networks.

Dr. Lee is a Member of the Institute of Electronics, Information and Communication Engineers, Korean Institute of Information Scientists and Engineers, and Korean Institute of Communications and Information Sciences.