IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received May 18, 2022, accepted June 6, 2022, date of publication June 13, 2022, date of current version June 22, 2022.

Digital Object Identifier 10.1109/ACCESS.2022.3182686

Balancing Fairness and Energy Efficiency in SWIPT-Based D2D Networks: Deep Reinforcement Learning Based Approach

EUN-JEONG HAN[1], MUY SENGLY[1], AND JUNG-RYUN LEE[1,2], (Senior Member, IEEE)
[1]School of Intelligent Energy and Industry, Chung-Ang University, Seoul 06974, South Korea
[2]School of Electrical and Electronics Engineering, Chung-Ang University, Seoul 06974, South Korea

Corresponding author: Jung-Ryun Lee (jrlee@cau.ac.kr)

This work was supported in part by the Ministry of Science and ICT (MSIT), South Korea, through the ITRC Support Program supervised by the Institute for Information & Communications Technology Planning & Evaluation (IITP) under Grant IITP-2022-2018-0-01799; in part by the Korea Institute of Energy Technology Evaluation and Planning (KETEP) and the Ministry of Trade, Industry & 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.

ABSTRACT In this study, we propose a method to balance between user fairness and energy efficiency of users in the context of simultaneous wireless information and power transfer (SWIPT)-based device-to-device (D2D) networks. For this purpose, we build an optimization model which determines the subchannel allocation, transmit power level, and power splitting ratio of D2D users, with the purpose of maximizing the objective function which presents the trade-off level between the harvested energy and the average logarithmic data rate of users. To solve this problem, we employ deep reinforcement learning (DRL) which combines deep neural network (DNN) with reinforcement learning (RL). Despite the use of DRL, the dimension of the action space in our work is still very high because it should include subchannel allocation indicator, power splitting ratio, and transmit power of all D2D users. We therefore apply an interior point method to the output of the DNN in DRL to avoid the excess convergent time of DRL. Through the simulations, we compare the performance of our proposed algorithm to that of the conventional iterative algorithms; exhaustive search (ES) and gradient search (GS). Results show that the objective function value remains stable regardless of the change in the maximum transmission power. In addition, it is verified that varying the power splitting ratio has little effect on the system performance, which justifies using a constant power splitting ratio in SWIPT-based D2D networks. Furthermore, it is verified that the proposed DRL achieves near-global-optimal solution compared with conventional algorithms, with lower computational complexity.

INDEX TERMS Energy efficiency, packet scheduling, D2D network, joint optimization, deep reinforcement learning.

I. INTRODUCTION

Because the data rate of future generation networks is expected to be 100 to 1000 times faster than that of the current generation networks, the use of high frequency bands becomes inevitable [1]. Therefore, path-loss and cell coverage are expected to be increased and reduced, respectively, resulting in the deterioration of the quality of service. Thus, D2D communication, which can communicate directly between proximity users without going through a base station, is gaining popularity as a solution for dealing with big data while maintaining high spectral efficiency and energy efficiency. However, because most D2D user devices are powered by batteries and the capacity of a battery is limited and insufficient to satisfy the high energy demands in wireless networks, increasing the energy efficiency of D2D user devices in D2D communication is a challenging task [2], [3]. In this context, SWIPT is considered a solution to the power shortage problem of D2D communications, as it can extend the lifetime of wireless nodes by transmitting

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

data while harvesting energy from radio-frequency signals [4] [5]. However, the emphasis on energy efficiency across the network may cause degradation of fairness among individual D2D devices. Therefore, improving energy efficiency while guaranteeing user fairness is an important issue from the perspective of D2D users.

Many studies have been conducted on improving energy efficiency to solve the energy shortage in a D2D network. The authors of [6] investigated efficient energy management using Q-learning while considering the trade-off in energy efficiency and delay. In [7], the author considered the energy-efficient resource allocation problem to reduce dependence on the battery with the EH time slot allocation, resource block, and power allocation. The author of [8] applied the scheduling method to select D2D links that satisfy both the signal-to-interference-plus-noise (SINR) ratio and the transmit power constraints.

In addition, there are some researches of SWIPT-enabled D2D networks to solve the problem of limited energy of D2D devices. In [9], the authors considered power allocation in SWIPT-based D2D networks to improve the energy efficiency. The author of [10] focused on wireless powered communication network to improve the bandwidth utilization and reduce energy loss in SWIPT-based D2D networks.

On the other hand, as mobile network architecture has become more diverse and complex, various machine learning algorithms have been used to control and adapt proper network parameters. Deep learning, which uses multiple layers to progressively extract high-level features from raw input, has been successfully applied to complex mobile network environments because it can extract highly correlated features without being explicitly programmed. Reinforcement learning enables environmental automatic exploration and self-decision, and it has received considerable attention because of its ability to solve dynamic resource allocation problems [11]. Deep reinforcement learning (DRL) is a recently developed technology that combines deep learning and reinforcement learning. While reinforcement learning is only applicable to data with low-dimensional features due to its exponentially increasing computational complexity, DRL can train the agent to learn actions using deep learning to approximate high-dimensional raw data [12], [13]. Because DRL does not require prior knowledge of the environment to obtain optimal performance, it possesses characteristics of autonomous exploration and optimal decision-making functionalities, which justifies embedding deep learning into wireless networks.

Many researchers have applied machine learning algorithms to resource allocation problems in various network systems. In [14], the authors proposed a multi-agent DRL to optimize joint energy-efficient subchannel assignment and power control in a massive access management problem. The authors of [15] proposed a resource allocation mechanism based on DRL that minimizes interference to vehicle-to-infrastructure communications. In [16], the authors developed a DRL-based algorithm to maximize the weighted-sum rate of a D2D network by formulating a joint channel selection and power control optimization problem. In [17], the deep Q-network (DQN) algorithm was used to solve the problems of caching deployment based D2D and mobile edge computing caching system. The authors of [18] applied DRL to solve the joint optimization of sub-carrier assignment and power allocation in D2D networks. In [19], a machine learning was applied to deploy the energy resource appropriately.

In this paper, we study energy efficiency optimization for individual D2D user pairs from a fairness perspective in a SWIPT-based D2D network using DRL. For this purpose, we build a model for joint optimization of energy efficiency and proportional fair scheduling in the context of a SWIPT-based D2D network. We propose DRL for the joint optimization model to determine an attractive trade-off between harvested energy and proportional fairness of D2D users. The main contributions of this paper are summarized as follows.

1) We build a joint optimization model of energy harvesting and proportional fair scheduling in D2D communication subject to the transmit power control of D2D users and subchannel allocation, and design a DRL to solve the joint optimization model.
2) Because the output space of the DNN in the proposed DRL is composed of the transmit power and subchannel allocation of D2D users, the computational time for convergence grows exponentially as the numbers of D2D users increases, quantization levels of transmit power becomes high, and the number of system subchannel increases. To obtain faster convergence of the proposed DRL algorithm, we apply the interior point method to the outputs of the proposed DRL method, which transforms the original objective function with inequality constraints into an equivalent optimization problem with equality constraints using the barrier function.
3) We compare the results of the proposed DRL algorithm with optimization-based iterative methods, such as exhaustive search (ES) and gradient search (GS) with the barrier function. The results reveal that the proposed DRL algorithm achieves a nearly-global-optimal solution as compared to other iteration-based optimization methods while reducing the time-complexity.

The remainder of this paper is organized as follows. Section II describes our system model and states the optimization model. In Section III, we explain the proposed DRL model and interior point method. Section IV evaluates the performance of the proposed algorithm with that of the comparison algorithms. Finally, Section V presents the conclusions of this article.

II. SYSTEM MODEL AND PROBLEM FORMULATION

We consider a SWIPT-based D2D network consisting of $\mathcal{K} \in \{1, 2, \ldots, K\}$ D2D pairs randomly deployed under a single base station with the coverage of R. Each D2D receiver (D2D-Rx) is equipped with an energy harvesting sensor that

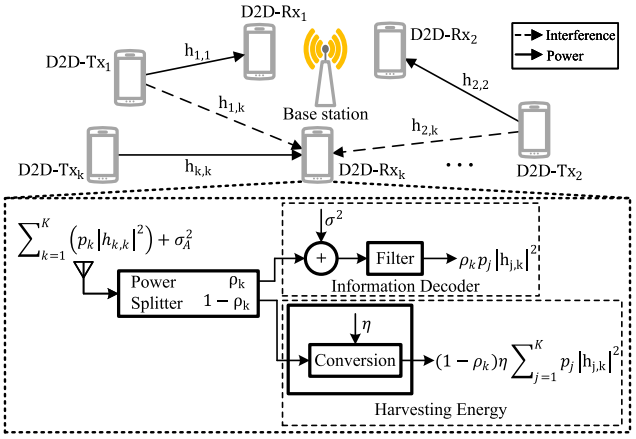


FIGURE 1. System model.

can divide the received power into two, such as information decoding and energy harvesting, as illustrated in Fig.1. There are N subchannels for scheduling D2D users, and p_k^t and ρ_k^t are the transmission power of the k^{th} D2D transmitter (D2D-Tx) at time t and the power splitting ratio of the D2D pair $k \in \mathcal{K}$ at time t, respectively. The channel between D2D-Tx k and D2D-Rx k at time t is denoted by $h_{k,k}^t$, and the channel gain is given by $\left|h_{k,k}^t\right|^2$, which is an independent and identically distributed Rician random variable with mean $\mu_{k,k}$. The interference between j^{th} D2D-Tx and k^{th} D2D-Rx at time t is given by

$$I_{j,k}^t = \rho_k^t \sum_{j=1, j\neq k}^{K} p_j^t \left|h_{j,k}^t\right|^2. \tag{1}$$

Therefore, the SINR received at the k^{th} D2D-Rx from k^{th} D2D-Tx at time t is expressed by

$$\Gamma_{k,k}^t = \frac{\rho_k^t p_k^t \left|h_{k,k}^t\right|^2}{\sigma^2 + \rho_k^t \sigma_A^2 + I_{j,k}^t}, \tag{2}$$

where the noise is $\sigma^2 + \rho_k^t \sigma_A^2 + I_{j,k}^t$, σ_A^2 and σ^2 are the antenna noise and base-band noise at the wireless-power receiver, respectively. Then, the received data rate of the k^{th} D2D-Rx at time t is given by

$$R_k^t = \sum_{n=1}^{N} s_{n,k}^t \log_2\left(1 + \Gamma_{k,k}^t\right), \tag{3}$$

where $s_{n,k}^t$ is an indicator of the subchannel allocation at time t. If k^{th} D2D-Rx is allocated to subchannel n at time t, $s_{n,k}^t = 1$, otherwise $s_{n,k}^t = 0$.

The previously aggregated received data rate of the k^{th} D2D-Rx during time window T is given by

$$AR_k^{t_c} = \begin{cases} \sum_{t=1}^{t_c-1} R_k^t, & t_c < T \\ \sum_{t=t_c-T}^{t_c-1} R_k^t, & t_c \geq T \end{cases}, \tag{4}$$

where t_c is the current time slot. The average of the previously received data rate of the k^{th} D2D-Rx during time window T becomes

$$\overline{R}_k^{t_c} = \begin{cases} \frac{1}{t_c} \sum_{t=1}^{t_c-1} R_k^t, & t_c < T \\ \frac{1}{T} \sum_{t=t_c-T}^{t_c-1} R_k^t, & t_c \geq T \end{cases}. \tag{5}$$

Then, the sum of the previous logarithmic data rate of the receivers during time window T is given by

$$\begin{aligned} R_{PF}^{t_c} &= \sum_{k=1}^{K} \log_2 AR_k^{t_c} \\ &= \begin{cases} \sum_{k=1}^{K} \log_2 \sum_{t=0}^{t_c-1} R_k^t, & t_c < T \\ \sum_{k=1}^{K} \log_2 \sum_{t=t_c-T}^{t_c-1} R_k^t, & t_c \geq T \end{cases}. \end{aligned} \tag{6}$$

We use the sum of logarithmic form to ensure the fairness of the user's data rate because the marginal gain is larger in a low-rate region than that in a high-rate region in the logarithmic function.

We define the total energy dissipation which describes the energy usage including the consumed energy for data communication and harvested energy from ambient energy sources as defined in [20] and [21]. Therefore, the total energy dissipation in EH-based wireless networks (EHWNs) during time window T is given by

$$ED^{t_c} = \begin{cases} \sum_{t=1}^{t_c-1} \sum_{k=1}^{K} \left[P_c + p_k^t\right] - E, & t_c < T \\ \sum_{t=t_c-T}^{t_c-1} \sum_{k=1}^{K} \left[P_c + p_k^t\right] - E, & t_c \geq T \end{cases}, \tag{7}$$

where P_c is the power circuit consumption, and E is the total harvested energy from all D2D-Rxs during time window T, which is given by

$$E = \sum_{k=1}^{K} \sum_{t=t_c-T}^{t_c-1} \sum_{i=1}^{K} (1 - \rho_k^t) \eta p_i^t \left|h_{i,k}^t\right|^2 \tag{8}$$

where η denotes the conversion rate of energy harvesting.

From the definitions of $R_{PF}^{t_c}$ and ED^{t_c}, we define proportional fair scheduling with energy efficiency in SWIPT D2D pairs as

$$PS^{t_c} = \frac{R_{PF}^{t_c}}{ED^{t_c}}. \tag{9}$$

In this work, we aim to determine a transmit power $\vec{p}$, power splitting ratio $\vec{\rho}$, and subchannel allocation indicator $\vec{s}$ that maximize the proportional fair energy efficiency at the current time. Therefore, the optimization problem can be formulated as

$$\begin{aligned} \max_{\vec{p},s,\vec{\rho}} \; & PS^{t_c}(\vec{p}, s, \vec{\rho}) \\ \text{s.t.} \; & C1: 0 \leq p_k^t \leq p_{\max}, \\ & \text{for } 1 \leq k \leq K, \text{and } t_c - T \leq t \leq t_c - 1 \\ & C2: s_{n,k}^t \in \{0, 1\}, \text{ for } 1 \leq n \leq N \\ & C3: \overline{R}_k^{t_c} \geq R_{\min} \\ & C4: 0 \leq \rho_k^t \leq 1. \end{aligned} \tag{10}$$

Here, constraint $C1$ denotes the boundary conditions for the transmit power, which states that the transmit power cannot be less than 0 and limited by the maximum transmit power $p_{\max}$. The constraint $C2$ is a binary criterion for subchannel allocation, and $C3$ denotes the minimum received rate for each D2D pair to guarantee the quality of services (QoS). The constraint $C4$ is a boundary of the power splitting ratio.

III. PROPOSED DEEP REINFORCEMENT LEARNING WITH INTERIOR POINT METHOD

A. PROPOSED DEEP REINFORCEMENT LEARNING

A Markov decision process (MDP) is a mathematical framework to describe an environment in reinforcement learning. The MDP is defined as a tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{R}, \mathcal{P} \rangle$, where $\mathcal{S}$ denotes a finite set of states, $\mathcal{A}$ represents a finite set of actions, $\mathcal{R}$ indicates the reward function, and $\mathcal{P}$ is the transition probability from current state $S[t_c] \in \mathcal{S}$ at time t_c to the next state $S[t_c+1] \in \mathcal{S}$ at time t_c+1. $P^{\pi}(S[t_c+1] \mid S[t_c], A[t_c])$ is the transition probability from current state $S[t_c]$ to the next state $S[t_c+1]$ given the action $A[t_c] \in \mathcal{A}$, and $\pi(A[t_c] \mid S[t_c])$ is a mapping from the current state $S[t_c]$to the action $A[t_c]$, called the policy. Therefore, $P^{\pi}(S[t_c+1] \mid S[t_c])$ is defined as the transition probability $P^{\pi}(S[t_c+1] \mid S[t_c], A[t_c])$ weighted by the policy $\pi(A[t_c] \mid S[t_c])$. The goal of the MDP is to find the policy π^* that maximizes the reward function R.

In our study, we propose a DRL-based power control algorithm where the agent of DRL is installed at the base station (BS). The network design for the proposed DRL algorithm is as follows. The agent in the BS manages the current channel allocation $\boldsymbol{s}^{t_c}$, transmit power $\overrightarrow{p}^{\,t_c}$ of the D2D transmitters, and power splitting ratio $\overrightarrow{\rho}^{\,t_c}$ of the D2D receivers. We assume that all D2D pairs share all of the information they need with the BS, including channel gain and data rate of each D2D pair. With the information received from all D2D pairs, the agent in the BS calculates the reward and chooses an action for the next transmit power level $\overrightarrow{p}^{\,t_c}$, next power splitting level $\overrightarrow{\rho}^{\,t_c}$, and next channel allocation $\boldsymbol{s}^{t_c}$ for all D2D devices, based on ϵ-greedy policy π. After that, the agent shares the action with all D2D pairs, and stores the state, action, reward, and next state in the replay memory. Then, the agent separates the overall data of the replay memory into multiple mini-batch samples. Mini-batch samples are used as the input data of DNN to train the DNN in a way to minimize the loss function.

In training phase, the output of DNN becomes the initial value of interior point method. Based on the (sub-)optimal solution set $\left(p^{subopt}, \rho^{subopt}, s^{subopt}\right)$ obtained from interior point method, the next action is chosen using ϵ-greedy policy. It is noted that, whenever the DNN is trained by the samples, the (sub-)optimal solution obtained from interior point method goes back to DNN, and acts as the target value in the loss function of DNN. This procedure is provided to accelerate the convergence of DRL, and characterizes the proposed algorithm in this paper. It is noticed that this procedure is applied only for training phase, but not for testing phase; that is, the output of DNN is directly delivered to the action space without using interior point method for testing. Fig.2 illustrates the architecture of the proposed DRL algorithm.

DNN in the proposed DRL algorithm is trained using the network training function in MATLAB Toolbox, 'trainscg', which updates weight and bias values based on the scaled conjugate gradient method. The detailed definitions of the state space, the action space, and the reward function in the proposed DRL algorithm are as follows.

- State space: The state space of the proposed DRL is define by

$$S[t_c] = \left\{ \mathbf{H}^{t_c}, \overline{\mathbf{R}}^{t_c} \right\}, \tag{11}$$

  where $\mathbf{H}^{t_c} = \left(\left| h_{j,k}^{t_c} \right|^2, \left| h_{j,k}^{t_c-1} \right|^2, \ldots, \left| h_{j,k}^{t_c-T} \right|^2 \right)$ is the set of channel gain from previous time window T include current time slot, and $\overline{\mathbf{R}}^{t_c} = \left(\overline{R}_1^{t_c}, \overline{R}_2^{t_c}, \ldots, \overline{R}_K^{t_c} \right)$ is the set of average of the previously received data rate during previous time window T.
- Action space: An agent corresponds to each D2D pair, which interacts with others to adjust the transmission power, power splitting ratio, and subchannel allocation indicator. Therefore, the action of each D2D pair at current time slot t_c is defined as

$$A[t_c] = \left\{ \overrightarrow{p}^{\,t_c}, \overrightarrow{\rho}^{\,t_c}, s^{t_c} \right\}, \tag{12}$$

  where $\overrightarrow{p}^{\,t_c}$, $\overrightarrow{\rho}^{\,t_c}$, and s^{t_c} are transmit power, power splitting ratio, and channel allocation indicator for all D2D pairs at current time slot t_c, respectively. Note that $\overrightarrow{p}^{\,t_c} \in \left\{ 0, \frac{p_{\max}}{L}, \frac{2p_{\max}}{L}, \ldots, p_{\max} \right\}$ for $(L+1)$ quantization levels, $\overrightarrow{\rho}^{\,t_c} \in \left\{ 0, \frac{1}{M}, \frac{2}{M}, \ldots, 1 \right\}$ for $(M+1)$ quantization levels, and $s^{t_c} \in \{0, 1\}$.
- Reward: In our problem, we aim to maximize the objective function which maximize the proportional fair with energy efficiency. Therefore, we define the reward for the proposed DRL model at the current time slot t_c as follows:

$$R[t_c] = \mathrm{PS}^{t_c}(\vec{p}, \boldsymbol{s}, \vec{\rho}) = \frac{\mathrm{R_{PF}}(\vec{p}, \boldsymbol{s}, \vec{\rho})}{\mathrm{ED}(\vec{p}, \boldsymbol{s}, \vec{\rho})}, \tag{13}$$

  when constraint C3 is satisfied; otherwise, the reward is negative.

In RL, Q-value identifies an optimal action-selection policy for any given finite Markov decision process, with given infinite exploration time and a partly-random policy. In our work, the Q-value is updated by

$$\begin{aligned} &Q^{t_c}(S[t_c], A[t_c]) \\ &\quad = (1-\alpha)\, Q^{t_c}(S[t_c], A[t_c]) \\ &\qquad + \alpha \left[R[t_c] + \gamma \max_{a \in \mathcal{A}_k} Q^{t_c+1}(S[t_{c+1}], A) \right], \end{aligned} \tag{14}$$

where $(\alpha > 0)$ and $(0 \le \gamma \le 1)$ are the learning rate and the discount factor, respectively.

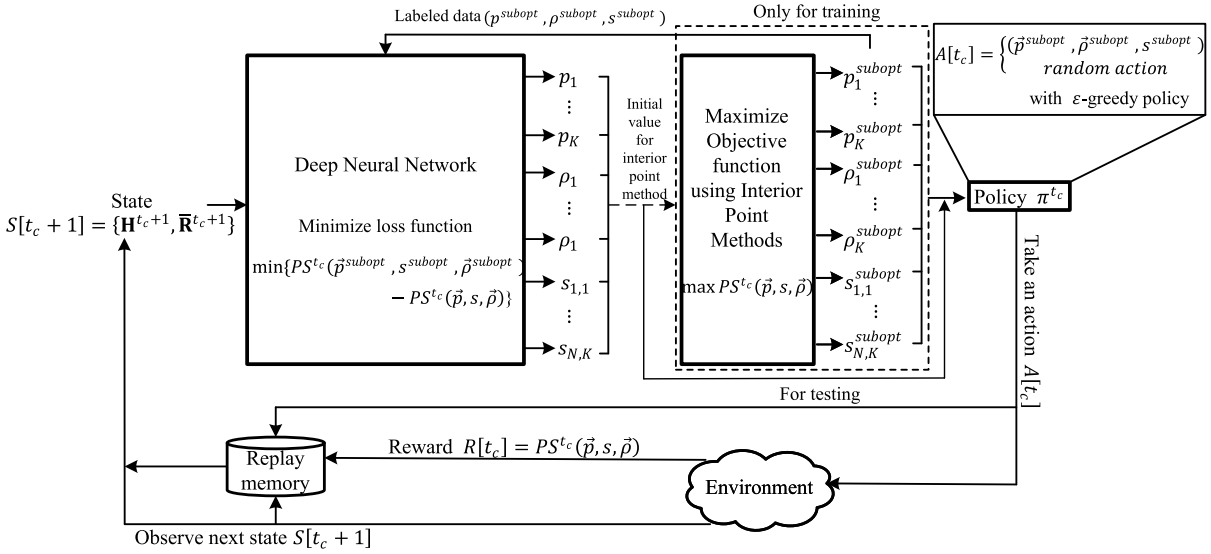


FIGURE 2. Proposed DRL model.

It is noted that we employ DRL rather than RL in our work, because the agent must consider the large input (state) space matrix at each time slot t, which includes the channel gain, received data rate, transmit power level, subchannel allocation indicator, and power splitting ratio. In addition, it is also noted that the output space of DNN (used for action space) is also very large because it should include huge number of actions which increases in proportional to the number of D2D pairs K, the number of subchannel N, quantization levels of transmit power L, and power splitting ratio M, and it may result in high convergence time.

To reduce the convergence time of the proposed DRL algorithm, we employ the interior point method in our work. The interior point method in the proposed DRL algorithm obtains (local-)optimal solutions for the problem of maximizing our objective function (9) using the output of DNN as input parameters. As explained earlier, the solution of the interior point method goes back to the DNN as the target value of the loss function of DNN. The details of the interior point method in this work is described in the following subsection.

B. INTERIOR POINT METHOD

The interior point method is an optimization method that can determine the local or global optimum of nonlinear objective functions and constraints. The original optimization problem with nonlinear constraints can be transformed into equality constraints using interior point method by adding an inequality with a slack variable. In this work, we build the optimization model using the reward function, which is given by

$$\begin{aligned}
&\max_{\vec{p},s,\vec{\rho}} PS^{t_c}(\vec{p}, s, \vec{\rho}) = \min_{\vec{p},s,\vec{\rho}} \left[-PS^{t_c}(\vec{p}, s, \vec{\rho})\right] \\
&\text{s.t.} \quad s_i \le 0, i = 1, 2, 3 \\
&\qquad C1' : p_k - p_{max} - s_1 = 0 \\
&\qquad C2' : s_{n,k} - s_{max} - s_2 = 0 \\
&\qquad C3' : \rho_k - \rho_{max} - s_3 = 0 \\
&\qquad C4' : \overline{R}_k^{t_c} \ge R_{\min},
\end{aligned} \tag{15}$$

where s_i is a slack variable. The barrier function replaces the inequality constraints by adding a penalizing term in the objective function. With the barrier function, the inequality constrained optimization problem becomes an equality constrained problem, and the objective function can be optimized more easily, which is given by,

$$\min_{\vec{p},s,\vec{\rho}} \left[-PS^{t_c}(\vec{p}, s, \vec{\rho}) + \Psi(\vec{p}, s, \vec{\rho})\right] \tag{16}$$

under the constraints $C1' - C4'$. The log barrier function is given by

$$\begin{aligned}
\Psi(\vec{p}, s, \vec{\rho}) &= -\mu \sum_{i=1}^{3} \log s_i, \\
s_1 &= \sum_{k=1}^{K} (\log\ (p_{max}\ -\ p_k\ ) + \log\ p_k) \\
s_2 &= \sum_{k=1}^{K} \sum_{n=1}^{N} (\log\ (\ s_{max}\ -\ s_k\ )\quad +\quad \log\ s_k\ ) \\
s_3 &= \sum_{k=1}^{K} (\log\ (\rho_{max} - \rho_k) + \log \rho_k)
\end{aligned} \tag{17}$$

where μ is a parameter of the log barrier function, and s_i represents the barriers for the inequality constraints of the original problem $C1$, $C2$, and $C3$, respectively. The solution is only searched in a feasible interior space. Then the Newton-Raphson method is employed to solve the KKT solu-

TABLE 1. Simulation environments.

Parameters	Values
Number of subchannels	N = 3
Number of D2D pairs	K = 6
Energy conversion efficiency	η_k = 0.5 [22]
Distance between D2D pairs	10-20 m
Energy consumption of circuit	P_c = 20 dBm
Rician factor	5 dB
Base-band noise power spectrum	σ^2=-70 dBm
AWGN power spectrum	σ_A^2 = -100 dBm
Maximum transmit power	1-23 dBm
Window size	T = 5
Path-loss exponent	3.6

tion for the problem with equality constraints, given by

$$\nabla(-PS^{t_c}(\vec{p}, s, \vec{\rho}) + \sum_{j=1}^{3} \lambda_j \nabla c_j - z_i) = 0, \tag{18}$$

for the stationarity condition, where c_j denotes the equality constraints, and λ_j represents an unknown variable for the Lagrange multiplier method. The search directions d_k^s, d_k^λ, and d_k^z are obtained using the Newton-Raphson method in the interior point method, which is given by

$$\begin{bmatrix} \nabla_{ss}^2 \left(-PS^{t_c} + c_k^T \lambda_k - z_k\right) & \nabla c_k & -I \\ \nabla c_k^T & 0 & 0 \\ Z_k & 0 & S_k \end{bmatrix} \begin{bmatrix} d_k^s \\ d_k^\lambda \\ d_k^z \end{bmatrix} = - \begin{pmatrix} \nabla f + \nabla c_k - z_k \\ c_k \\ S_k Z_k e - \mu e \end{pmatrix} \tag{19}$$

where $z_i = \mu / s_i$, $e = (1, \ldots, 1)^T$, $S = diag(s)$, and $Z = diag(z)$. The iterative search for the interior point method is performed until it converges with tolerance while satisfying the KKT conditions.

IV. PERFORMANCE EVALUATION

For simulation setup, we assume 3 subchannels, and 6 D2D pairs to verify the effect of channel scheduling. The energy conversion efficiency η_k is set to 0.5 [22]. The base-band noise power spectrum and additional white Gaussian noise power spectrum are set to $\sigma^2 = -70$ dBm, and $\sigma_A^2 = -100$ dBm, respectively [23]. The channel gain between nodes i and j is defined as $|h_{i,j}|^2 = g_{i,j}/d_{i,j}^m$, where $d_{i,j}^m$ is the physical distance between two nodes. The averages of direct link and interference link following the normal distribution are 10 m and 20 m, respectively. The path-loss exponent m is equal to 3.6 [24]. $g_{i,j}$ is the Rician small scale fading gain with a 5 dB K-factor. The constant energy consumption of the circuit is $P_c = 20$ dBm. The detailed simulation environments are summarized in Table 1. In addition, we assume the D2D communication with no delay and pairing failure to focus on the balancing performance between energy efficiency and fair scheduling in SWIPT-based D2D communications.

To evaluate the performance of the proposed DRL algorithm, we compare the proposed scheme with two algorithms as follows [24].

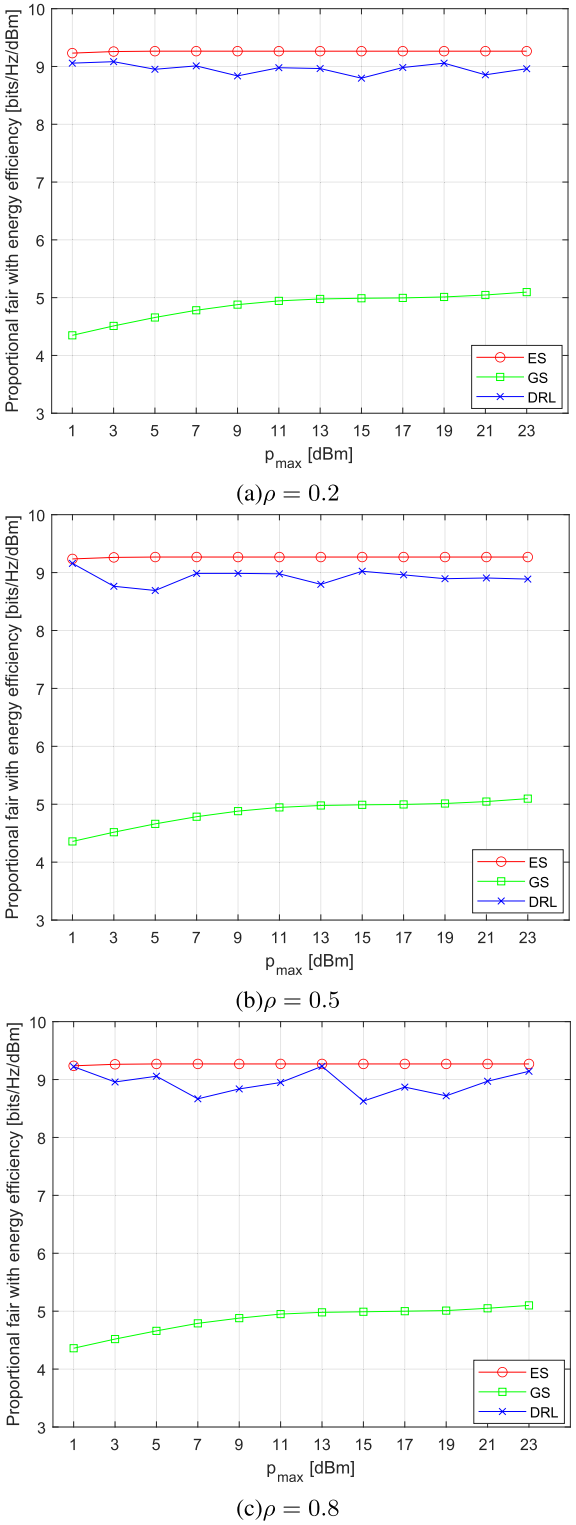
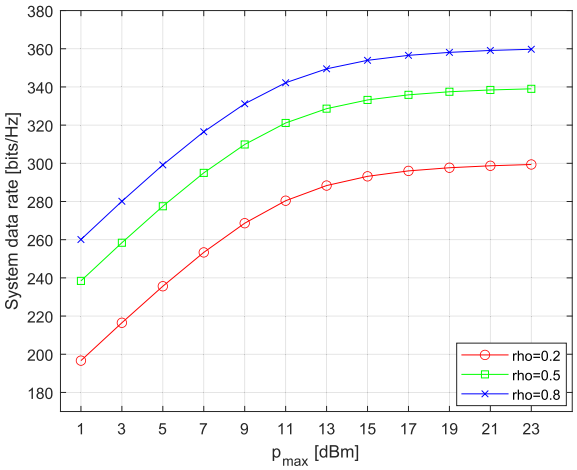
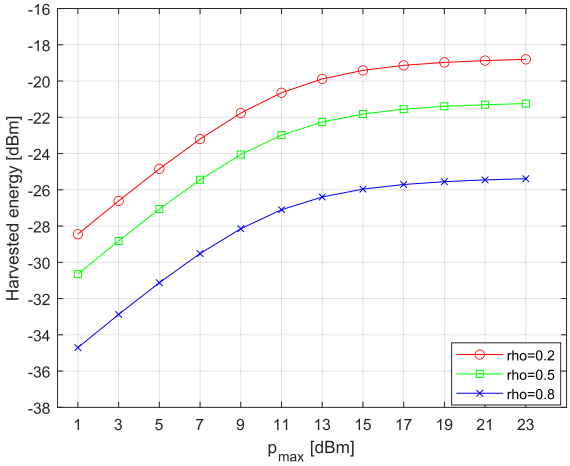


FIGURE 3. Proportional fair with energy efficiency for different power splitting ratio.

1) Exhaustive search (ES) algorithm: The kinds of optimization method used to find the global optimal solution, which enumerates and compares all possible candidates for the optimization problem. In our work,



(a)System data rate vs p_{max}



(b)Harvested energy vs p_{max}

FIGURE 4. System data rate and harvested energy.

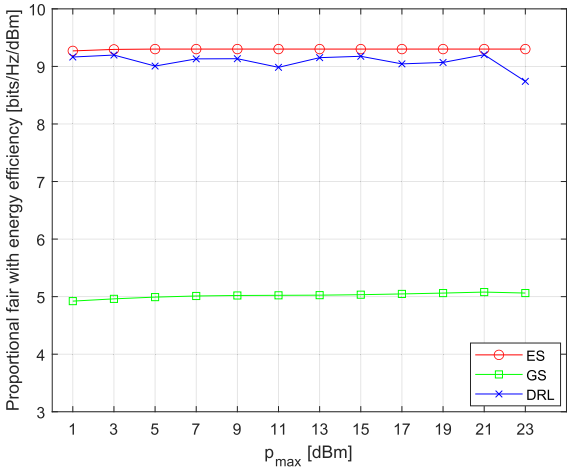


FIGURE 5. Proportional fair with energy efficiency with changeable power splitting ratio.

the control parameters $(\vec{p}, \boldsymbol{s}, \vec{\rho})$ are quantized, and all possible combinations of the quantized control parameters are examined to determine the maximum value of the objective function while satisfying the constraints.

TABLE 2. Computational complexity comparison.

Algorithm	Computation complexity	The running time
ES	$O((LM)^{NK})$	2.5 days
GS	$O(\epsilon^{-3})$	12.3 minutes
DRL	$O(HK^4N^2)$	4.2 minutes

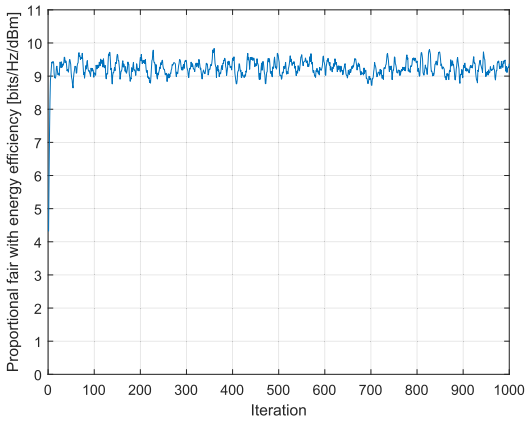


FIGURE 6. Convergence of Proposed DRL when $p_{max} = 11$ dBm.

2) Gradient search (GS) algorithm: The method for determining the local optimum that searches for the solution according to the direction of derivative.

Fig. 3. presents the convergence results of the proportional fairness with energy efficiency for the ES, DNN, and GS vs. p_{max} with $K = 6$ and $N = 3$ when the power splitting ratio ρ is 0.2, 0.5, and 0.8, respectively. We perform the simulation under simple conditions with $N = 3$ because ES has extremely high computational complexity of $O((LM)^{NK})$. ES can get the global optimal solution with extremely high time complexity of $O((LM)^{NK})$, which exponentially increases as N or K increase. Therefore, even though the ES algorithm can find a global optimal solution, it has been used for comparison purpose only, as shown in [25], [26]. The results reveal that the objective function value increases as p_{max} increases, which explains that the degree of deterioration in the sum of the logarithmic data is slightly smaller than or almost equal to the degree of improvement in the energy efficiency as p_{max} increases. The results indicate that the objective function value of GS is significantly lower than that of ES. However, the difference between the proposed DRL and ES is less than 5 % even if p_{max} increases. It is noticed that the proposed DRL algorithm achieves near-global optimal solutions while reducing the time complexity compared to ES as shown in Table 2, where H is the number of hidden layers and ϵ is set to 10^{-5}.

Fig.4 shows system data rate and harvested energy when power splitting ratio ρ is fixed as 0.2, 0.5, and 0.8. Although the values of proportional fairness with energy efficiency did not change significantly when the power splitting ratio changes, the system data rate decreases and the harvested

energy increases as the power splitting ratio increases. It also shows that the system data rate and harvested energy converge as p_{max} increases.

On the other hand, Fig.5 is obtained when the ρ can vary every time slot according to the policy of DRL. This figure indicates that the proportional fairness with energy efficiency is slightly larger than but almost same to the result of using the fixed splitting ratio. Overall, we can verify that the objective function value increases as p_{max} increases.

Fig.6 shows the convergence of the proposed DRL algorithm when $p_{max} = 11$ dBm, $\rho = 0.5$, $N = 3$, and $K = 6$.

V. CONCLUSION

In this paper, we suggested the optimization model to balance the system performance between fair scheduling and energy efficiency in SWIPT-based D2D networks. To solve this problem, we proposed the DRL model, which determines the best transmit power, power splitting ratio, and subchannel allocation indicator to maximize the objective function, defined as the sum of the previous logarithmic data rate over the total energy dissipation in the system model. To reduce the convergence time of the proposed model, we converted the original optimization problem with inequality constraints into an equality constraint using a barrier function. And we used the interior point method for the output space of the DNN in the proposed DRL model. Simulation results said that the system logarithmic data rate and harvested energy increase at nearly the same rate as the maximum power increases, whereas the objective function index remains almost fixed. Furthermore, it was verified that varying the power splitting ratio has little effect on the system performance, which justifies the use of a constant power splitting ratio in SWIPT-based D2D networks. In addition, we can see that, throughout the simulation runs, the performance of the proposed DRL model outperforms GS and achieves the near-global-optimal solution with lower time complexity, which shows the benefit of using DRL.

Furthermore, we plan to apply another kinds of optimization methods such as deep deterministic policy gradient (DDPG) method, twin delayed DDPG (TD3) method, and soft actor critic (SAC) method, to joint optimization of energy efficiency and scheduling in SWIPT-based D2D network.

REFERENCES

[1] S. Zhang, J. Liu, H. Guo, M. Qi, and N. Kato, "Envisioning device-to-device communications in 6G," *IEEE Netw.*, vol. 34, no. 3, pp. 86–91, Jun. 2020.

[2] X. Lin, J. Wu, A. K. Bashir, J. Li, W. Yang, and J. Piran, "Blockchain-based incentive energy-knowledge trading in IoT: Joint power transfer and AI design," *IEEE Internet Things J.*, early access, Sep. 15, 2020, doi: 10.1109/JIOT.2020.3024246.

[3] J. Li, Z. Zhou, J. Wu, J. Li, S. Mumtaz, X. Lin, H. Gacanin, and S. Alotaibi, "Decentralized on-demand energy supply for blockchain in Internet of Things: A microgrids approach," *IEEE Trans. Computat. Social Syst.*, vol. 6, no. 6, pp. 1395–1406, Dec. 2019.

[4] R. Zhang, R. G. Maunder, and L. Hanzo, "Wireless information and power transfer: From scientific hypothesis to engineering practice," *IEEE Commun. Mag.*, vol. 53, no. 8, pp. 99–105, Aug. 2015.

[5] J. Huang, C.-C. Xing, and C. Wang, "Simultaneous wireless information and power transfer: Technologies, applications, and research challenges," *IEEE Commun. Mag.*, vol. 55, no. 11, pp. 26–32, Nov. 2017.

[6] Y. Luo, M. Zeng, and H. Jiang, "Learning to tradeoff between energy efficiency and delay in energy harvesting-powered D2D communication: A distributed experience-sharing algorithm," *IEEE Internet Things J.*, vol. 6, no. 3, pp. 5585–5594, Jun. 2019.

[7] Z. Kuang, G. Liu, G. Li, and X. Deng, "Energy efficient resource allocation algorithm in energy harvesting-based D2D heterogeneous networks," *IEEE Internet Things J.*, vol. 6, no. 1, pp. 557–567, Feb. 2019.

[8] L. Zhang, M. Xiao, G. Wu, and S. Li, "Efficient scheduling and power allocation for D2D-assisted wireless caching networks," *IEEE Trans. Commun.*, vol. 64, no. 6, pp. 2438–2452, Jun. 2016.

[9] J. Huang, C.-C. Xing, and M. Guizani, "Power allocation for D2D communications with SWIPT," *IEEE Trans. Wireless Commun.*, vol. 19, no. 4, pp. 2308–2320, Apr. 2020.

[10] I. Budhiraja, N. Kumar, S. Tyagi, S. Tanwar, and M. Guizani, "SWIPT-enabled D2D communication underlaying NOMA-based cellular networks in imperfect CSI," *IEEE Trans. Veh. Technol.*, vol. 70, no. 1, pp. 692–699, Jan. 2021.

[11] K. Arulkumaran, M. P. Deisenroth, M. Brundage, and A. A. Bharath, "Deep reinforcement learning: A brief survey," *IEEE Signal Process. Mag.*, vol. 34, no. 6, pp. 26–38, Nov. 2017.

[12] M. Bennis, M. Debbah, and H. V. Poor, "Ultrareliable and low-latency wireless communication: Tail, risk, and scale," *Proc. IEEE*, vol. 106, no. 10, pp. 1834–1853, Oct. 2018.

[13] S. Rajendran, W. Meert, D. Giustiniano, V. Lenders, and S. Pollin, "Deep learning models for wireless signal classification with distributed low-cost spectrum sensors," *IEEE Trans. Cogn. Commun. Netw.*, vol. 4, no. 3, pp. 433–445, Sep. 2018.

[14] H. Yang, Z. Xiong, J. Zhao, D. Niyato, C. Yuen, and R. Deng, "Deep reinforcement learning based massive access management for ultra-reliable low-latency communications," *IEEE Trans. Wireless Commun.*, vol. 20, no. 5, pp. 2977–2990, May 2021.

[15] H. Ye and G. Y. Li, "Deep reinforcement learning for resource allocation in V2V communications," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–6.

[16] J. Tan, Y.-C. Liang, L. Zhang, and G. Feng, "Deep reinforcement learning for joint channel selection and power control in D2D networks," *IEEE Trans. Wireless Commun.*, vol. 20, no. 2, pp. 1363–1378, Feb. 2021.

[17] R. Zhang, F. R. Yu, J. Liu, T. Huang, and Y. Liu, "Deep reinforcement learning (DRL)-based device-to-device (D2D) caching with blockchain and mobile edge computing," *IEEE Trans. Wireless Commun.*, vol. 19, no. 10, pp. 6469–6485, Oct. 2020.

[18] X. Zhang, Z. Lin, B. Ding, B. Gu, and Y. Han, "Deep multi-agent reinforcement learning for resource allocation in D2D communication underlaying cellular networks," in *Proc. 21st Asia–Pacific Netw. Oper. Manage. Symp. (APNOMS)*, Sep. 2020, pp. 55–60.

[19] K. Wang, J. Wu, X. Zheng, A. Jolfaei, J. Li, and D. Yu, "Leveraging energy function virtualization with game theory for fault-tolerant smart grid," *IEEE Trans. Ind. Informat.*, vol. 17, no. 1, pp. 678–687, Jan. 2021.

[20] D. Altinel and G. K. Kurt, "Modeling of hybrid energy harvesting communication systems," *IEEE Trans. Green Commun. Netw.*, vol. 3, no. 2, pp. 523–534, Jun. 2019.

[21] J. Anees, H.-C. Zhang, B. G. Lougou, S. Baig, Y. G. Dessie, and Y. Li, "Harvested energy scavenging and transfer capabilities in opportunistic ring routing," *IEEE Access*, vol. 9, pp. 75801–75825, 2021.

[22] V. D. Nguyen, T. Q. Duong, H. D. Tuan, O.-S. Shin, and H. V. Poor, "Spectral and energy efficiencies in full-duplex wireless information and power transfer," *IEEE Trans. Commun.*, vol. 65, no. 5, pp. 2220–2233, Feb. 2017.

[23] L. Liu, R. Zhang, and K.-C. Chua, "Wireless information transfer with opportunistic energy harvesting," *IEEE Trans. Wireless Commun.*, vol. 12, no. 1, pp. 288–300, Jan. 2013.

[24] M. Sengly, K. Lee, and J.-R. Lee, "Joint optimization of spectral efficiency and energy harvesting in D2D networks using deep neural network," *IEEE Trans. Veh. Technol.*, vol. 70, no. 8, pp. 8361–8366, Aug. 2021.

[25] S. Zhang, B. Di, L. Song, and Y. Li, "Sub-channel and power allocation for non-orthogonal multiple access relay networks with amplify-and-forward protocol," *IEEE Trans. Wireless Commun.*, vol. 16, no. 4, pp. 2249–2261, Apr. 2017.

[26] Y. Zhao, Z. Li, B. Hao, and J. Shi, "Sensor selection for TDOA-based localization in wireless sensor networks with non-line-of-sight condition," *IEEE Trans. Veh. Technol.*, vol. 68, no. 10, pp. 9935–9950, Oct. 2019.

EUN-JEONG HAN received the B.S. degree from the School of Electrical and Electronics Engineering, College of ICT Engineering, Chung-Ang University, Seoul, South Korea, in 2021. She is currently pursuing the M.S. degree with the School of Intelligent Energy and Industry, Chung-Ang University. Her current research interests include optimization problem using machine learning in wireless networks, federated learning, and artificial intelligence.

MUY SENGLY received the B.S. degree from the Institute of Technology of Cambodia (ITC), Phnom Penh, Cambodia, in 2018. He is currently pursuing the integrated M.S. and Ph.D. degree with the School of Intelligent Energy and Industry, Chung-Ang University, Republic of Korea. His current research interests include performance optimization in wireless networks, artificial intelligence, and machine learning.

JUNG-RYUN LEE (Senior Member, IEEE) received the B.S. and M.S. degrees in mathematics from Seoul National University, in 1995 and 1997, respectively, and the Ph.D. degree in electrical and electronics engineering from the Korea Advanced Institute of Science and Technology (KAIST), in 2006. From 1997 to 2005, he was a Chief Research Engineer at LG Electronics, South Korea. From 2006 to 2007, he was a full-time Lecturer in electronic engineering at the University of Incheon. Since 2008, he has been a Professor with the School of Electrical and Electronics Engineering, Chung-Ang University, South Korea. His research interests include energy-efficient networks and algorithms, bio-inspired autonomous networks, and artificial intelligence-based networking. He is a Regular Member of IEICE, KIISE, and KICS. He received the Excellent Paper Award at ICUFN 2012, the Best Paper Award at ICN 2014, the Best Paper Award at QSHINE 2016, and the Excellent Paper Award at ICTC 2018.