Power Optimization in Device-to-Device Communications: A Deep Reinforcement Learning Approach With Dynamic Reward

Zelin Ji, Adnan K. Kiani, Zhijin Qin, *Member, IEEE*, and Rizwan Ahmad

***Abstract*—Device-to-Device (D2D) communication can be used to improve system capacity and energy efficiency (EE) in cellular networks. One of the critical challenges in D2D communications is to extend network lifetime by efficient and effective resource management. Deep reinforcement learning (RL) provides a promising solution for resource management in wireless communication systems. This letter aims to maximise the EE while satisfying the system throughput constraints as well as the quality of service (QoS) requirements of D2D pairs and cellular users in an underlay D2D communication network. To achieve this, a deep RL based dynamic power optimization algorithm with dynamic rewards is proposed. Moreover, a novel algorithm with two parallel deep Q networks (DQNs) is designed to maximize the EE of the considered network. The proposed deep RL based power optimization method with dynamic rewards achieves higher EE while satisfying the system throughput requirements.**

***Index Terms*—Device-to-device communication, deep reinforcement learning, energy efficiency, power control.**

I. Introduction

As ONE of the enabling technologies of the fifth generation (5G) and beyond communication systems, device-to-device (D2D) communication can benefit the system with lower communication latency and improved throughput. Moreover, due to the physical proximity and the potential reuse gain, D2D communication can improve the energy efficiency (EE) of wireless communication systems, while at the same time offloading traffic in cellular networks to avoid congestion [1]. However, due to the absence of base stations (BS), interference management becomes challenging in D2D networks [2]. Particularly, D2D users usually experience high interference from other D2D pairs and cellular users, which could degrade the overall system throughput and EE.

Transmit power can be optimized to limit the interference to an acceptable level. Some work has been reported in literature to optimize the transmit power to limit the interference to an acceptable level. In [3], the joint optimization of data rate and EE has been considered to obtain the optimal transmit power to maximize the EE in the orthogonal frequency-division multiple access (OFDMA) based D2D communications. In [4], the joint mode selection and spectrum sharing have been modeled as a coalition formation game. However, the main challenge is that the configuration of D2D users, i.e., position, transmit power, channel gain, are usually dynamic, which means the aforementioned optimization approaches may be unrealistic for the fast-changing systems. In such scenarios, reinforcement learning (RL) becomes a promising tool to perform efficient resource allocation and interference management.

RL is a machine learning method to solve decision-making problems, which has been employed in wireless communication systems, i.e., for End-to-End (E2E) communications, resource allocation, and power control [5]–[7]. It has been demonstrated that RL enabled E2E systems are able to circumvent the problem of missing gradients from channels when optimizing the transmitter [5]. However, the space of action and state grows significantly as the number of D2D and cellular users becomes large, which makes the legacy RL not applicable for large-scale networks.

In [6], [7], deep RL is employed to improve the QoS and minimize interference for D2D and V2V communications. Leveraging neural network (NN) to evaluate the approximate Q-value, deep Q network (DQN) is more applicable for solving the complicated problems with lower complexity. However, these algorithms suffer from lack of tradeoff between throughput and EE, which could lead to reduced lifetime of D2D users. In the proposed algorithm, a dynamic reward enabled parallel DQN algorithm is considered to achieve the tradeoff between obtaining the acceptable system throughput and maximising the EE.

The contributions are summarized as follows:

1) For the considered D2D network, a EE maximization problem is formulated with constraints of system throughput as well as the QoS requirements of D2D and cellular users.
2) To improve EE while satisfying the required QoS, we propose a deep RL enabled power optimization algorithm with two parallel DQNs. Each DQN represents a metric, i.e., data rate and EE, and the BS chooses from the two DQNs according to current state to achieve a better tradeoff between system throughput and EE.
3) The proposed algorithm efficiently converges to near optimal EE under different QoS constraints, providing flexibility for the system to set up throughput constraints in different communication scenarios.

Manuscript received October 4, 2020; accepted October 31, 2020. Date of publication November 4, 2020; date of current version March 9, 2021. The associate editor coordinating the review of this article and approving it for publication was K. Adachi. *(Corresponding author: Zhijin Qin.)*

Zelin Ji and Zhijin Qin are with the Department of EECS, Queen Mary University of London, E1 4NS London, U.K. (e-mail: z.ji@qmul.ac.uk; z.qin@qmul.ac.uk).

Adnan K. Kiani is with the Essex Pathways Department, University of Essex, CO4 3SQ Colchester, U.K. (e-mail: a.kiani@essex.ac.uk).

Rizwan Ahmad is with the Department of Electrical Engineering, National University of Sciences and Technology, Islamabad 44000, Pakistan (e-mail: rizwan.ahmad@seecs.edu.pk).

Digital Object Identifier 10.1109/LWC.2020.3035898

II. System Model

We consider a scenario consisting of a base station (BS) and multiple cellular users and D2D pairs. The sets of cellular users, D2D transmitters, and D2D receivers are denoted as $\mathcal{CU} = \{CU_1, \ldots, CU_N\}$, $\mathcal{TU} = \{TU_1, \ldots, TU_M\}$ and $\mathcal{RU} = \{RU_1, \ldots, RU_M\}$ respectively, where *M* and *N* denote the number of D2D pairs and cellular users. In this letter, we assume only uplink cellular channels are shared with D2D pairs since uplink resources are less intensively used and interference at the BS is more manageable. Flat-fading channel is considered. Chahennel gain $g^D_{m,m}$ between t *m*th D2D transmitter TU_m and its receiver RU_m is given by

$$g^D_{m,m} = \beta l^{-d}_{m,m} |h|^2, \tag{1}$$

where β is a constant value, $l_{m,m}$ is the distance between TU_m and RU_m, *d* is the path loss parameter, and *h* is the complex Gaussian random variables with zero mean. The interference experienced by a D2D receiver could be from a number of signals, including signals from other D2D transmitters and signals transmitted between cellular users and the BS, which shares the same resource block with it. Then, the signal-to-interference-plus-noise ratio (SINR) γ^D_m of RU_m is given by

$$\gamma^D_m = \frac{p^D_m g^D_{m,m}}{\sigma^2 + \sum_{i=1, i\neq m}^{M} k^D_{i,m} p^D_i g^D_{i,m} + \sum_{n=1}^{N} k^C_{n,m} p^C_n g^C_{n,m}}, \tag{2}$$

where σ^2 denotes the Gaussian noise variance, p^D_m represents the transmit power of the *m*th transmitter DU_m, p^C_n is the uplink transmit power of the *n*th cellular user CU_n, $g^C_{n,m}$ denotes the channel gain from CU_n to DU_m. $k^D_{i,m}$ denotes the resource sharing coefficient for D2D pairs, with $k^D_{i,m} = 1$ indicating the *i*th D2D user and the *m*th D2D user share the same resource block and $k^D_{i,m} = 0$ otherwise. Similarly, $k^C_{n,m}$ denotes the resource sharing coefficient between the *m*th D2D user and the CU_n. The resource sharing coefficient can be optimized by the method in [8] with low complexity. The interference experienced by the BS for *n*th cellular user is caused by the signals from D2D transmitters that reuse the *n*th resource block. Therefore, the SINR γ^C_n received at the BS for the CU_n can be defined as

$$\gamma^C_n = \frac{p^C_n g^C_{n,bs}}{\sigma^2 + \sum_{i=1}^{M} k^C_{n,i} p^D_i g^C_{i,bs}}, \tag{3}$$

where $g^C_{n,bs}$, $g^C_{i,bs}$ represent the channel gain between the BS and the CU_n and DU_i, respectively. Then, the sum throughput *T* achieved by D2D pairs and cellular networks can be calculated as

$$T = \sum_{m=1}^{M} B_m \log_2(1 + \gamma^D_m) + \sum_{n=1}^{N} B_n \log_2(1 + \gamma^C_n), \tag{4}$$

where B_m and B_n represent the bandwidth of TU_m to RU_m and CU_n to the BS, respectively. Our objective is to maximize the EE with throughput constraints to guarantee QoS of the system and each users, the problem can then be formulated as

$$(\text{P0})\{\boldsymbol{p^D}, \boldsymbol{p^C}\} \quad \frac{T}{\sum_{m=1}^{M} p^D_m + \sum_{n=1}^{N} p^C_n} \tag{5a}$$

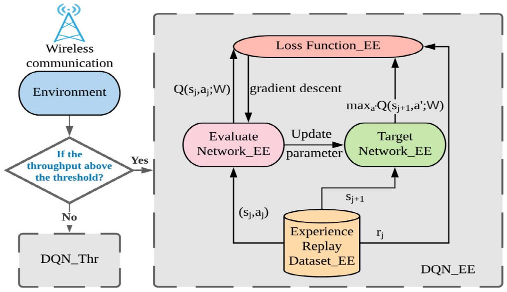


Fig. 1. The proposed parallel DQNs enabled power allocation in D2D networks.

$$\text{subject to} \quad C_1{:}\, T > \Pi \tag{5b}$$

$$C_2{:}\, \gamma^D_m > \gamma^{D*}, \forall m \in \{1, \ldots, M\} \tag{5c}$$

$$C_3{:}\, \gamma^C_n > \gamma^{C*}, \forall n \in \{1, \ldots, N\} \tag{5d}$$

$$C_4{:}\, p^D_m \leq p_{max}, \forall m \in \{1, \ldots, M\} \tag{5e}$$

$$C_5{:}\, p^C_n \leq p_{max}, \forall n \in \{1, \ldots, N\} \tag{5f}$$

$$C_6{:} \sum_{n=1}^{N} k^C_{n,m} \leq 1, \forall m \in \{1, \ldots, M\}, \tag{5g}$$

where $\boldsymbol{p^D} = [p^D_1, \ldots, p^D_M]$ and $\boldsymbol{p^C} = [p^C_1, \ldots, p^C_N]$ represents the transmit power of D2D transmitters and cellular users respectively, Π represents the threshold for system throughput that needs to be achieved. Moreover, γ^{D*} and γ^{C*} represent the SINR requirements for D2D receivers and the BS, respectively. C_4 and C_5 constrain the transmit power of cellular users and D2D transmitters, respectively. C_6 assumes that each D2D pair only shares the resource block with one cellular user, which makes (P0) an NP-hard problem, therefore, the exhaustive search algorithm could be adopted to find the optimal solution but with high complexity.

III. Dynamic Power Optimization Algorithm

To solve the above problem, we propose a deep RL based power optimization algorithm with low complexity to achieve the near-optimal solution. In the following, we first introduce the deep RL components shown in Fig. 1, which is followed by the description of proposed algorithm.

A. Reinforcement Learning Components

Agent: We propose a centralized power optimization approach and the BS is employed as the agent. The components of deep RL include $(s_t; a_t; r_t; s_{t+1})$, which means the BS in state s_t chooses an action a_t according to certain policy at step *t*. Then it will receive reward r_t and fall into state s_{t+1} at the next training step.

State: The state $\mathcal{S}$ includes the SINR information at D2D receivers and the BS, and current transmission power of D2D transmitters and cellular users. Assuming that the SINR of D2D receivers are broadcast and the BS acquires the information in real time. Formally, state $\mathcal{S}$ can be defined as $\mathcal{S} = \{\gamma^D_1, \ldots, \gamma^D_M; \gamma^C_1, \ldots, \gamma^C_N; \boldsymbol{p^D}; \boldsymbol{p^C}\}$.

Action and Policy: The action space $\mathcal{A}$ represents the power adjustment operation. At each iteration *t*, the action $\boldsymbol{a}_t \in \mathcal{A}$

can be defined as $\boldsymbol{a}_t = \{\Delta p_1^D, \ldots, \Delta p_M^D; \Delta p_1^C, \ldots, \Delta p_N^C\}$. The variable quantity $\Delta p_m^D, \Delta p_n^C \in \{-\delta, 0, \delta\}$, $\forall m \in \{1, \ldots, M\}, n \in \{1, \ldots, N\}$. In order to strike a trade-off between exploration and exploitation, the agent chooses actions according to ϵ-greedy algorithm [9].

Reward: Once the transmission finishes, agent receives the corresponding reward from the environment. Reward represents the optimization goals of the RL task. In order to achieve a balance between throughput and EE, two different rewards are designed to encourage throughput and EE. The rewards are switched based on different system states, generating two different Q-values for one state-action pair.

In Fig. 1, if the system throughput is above the threshold and all the other constraints are satisfied, the learning process tends to encourage higher EE to reduce the energy consumption of the system. Therefore, the reward function is defined to maximise the EE of the whole system. The EE of the D2D communications system is the ratio of the sum rate to the power consumption. The reward function is defined to maximise the EE, which can be defined as

$$r(\gamma, p) = \frac{T^D + T^C}{\sum_{m=1}^{M} p_m^D + \sum_{n=1}^{N} p_n^C}. \tag{6}$$

However, if the system throughput cannot satisfy the constraint C_1 while all the other constraints are satisfied, the optimization goal becomes increasing the throughput. Then the reward function composed of the system throughput is defined as

$$r(\gamma) = T^D + T^C. \tag{7}$$

Failure to fulfill constraints C_2, C_3, C_4, C_5 or C_6 means the QoS requirements of D2D pairs and cellular users cannot be satisfied. This action results in penalty and the cost is determined by the interference caused by the action. In such states, the reward function is defined as

$$\begin{aligned} r_{\text{fail}} = -\Big(\sum\nolimits_{m}^{M}\Big(&\sum_{i=1, i\neq m}^{M} k_{i,m}^D p_i^D g_{i,m}^D + \sum_{n=1}^{N} k_{n,m}^C p_n^C g_{n,m}^C\Big) \\ &+ \sum\nolimits_{n}^{N} \sum\nolimits_{i=1}^{M} k_{n,i}^C p_i^D g_{i,bs}^C\Big). \end{aligned} \tag{8}$$

In summary, the overall reward corresponding to EE, r_{EE} and corresponding to throughput, r_{Thr} can be expressed as

$$r_{EE} = \begin{cases} r(\gamma, p), \text{ if (5c), (5d), (5e), (5f) and (5g) are satisfied;} \\ r_{\text{fail}}, \quad \text{else,} \end{cases} \tag{9}$$

and

$$r_{Thr} = \begin{cases} r(\gamma), \text{ if (5c), (5d), (5e), (5f) and (5g) are satisfied;} \\ r_{\text{fail}}, \text{ else}. \end{cases} \tag{10}$$

According to Bellman equation [9], the optimal strategy is to select the action that maximizes

$$Q^*(s_t, a_t) = \mathbb{E}_{s_{t+1}}[r_t + \Gamma \max_{a' \in A} Q^*(s_{t+1}, a') | s_t, a_t], \tag{11}$$

where Γ represents the discount rate, $Q^*(s, a)$ is the desired value function such that $Q(s_t, a_t) \to Q^*(s_t, a_t)$ as $t \to \infty$. Since the iteration is discrete, it is impractical to directly acquire the accurate value. The NNs are applied to be function approximator to estimate the action-value function, i.e., $Q(s_t, a_t; \boldsymbol{W}) \approx Q^*(s, a)$.

Algorithm 1 Parallel DQN Training Algorithm

```
Input: Environment simulator, Q network, minibatch size;
Initialize: action-value function Q with random weights $\boldsymbol{W}$ and $\boldsymbol{W}^*$, replay memory
    $\mathcal{D}_{EE}$ and $\mathcal{D}_{Thr}$;
for each training episode do
    for each training step do
        Choose DQN according to current throughput threshold;
        Choose action $a_t$ from action space $\mathcal{A}$ according to $\epsilon$-greedy algorithm;
        Execute $a_t$, calculate dynamic reward $r_t$ by (9),(10) and observe $s_{t+1}$;
        Store transition $(s_t, a_t, r_t, s_{t+1})$ in $\mathcal{D}$;
        Replay memory:
         Sample random minibatch of transitions
         $(s_j, a_j, r_j, s_{j+1})$ in $\mathcal{D}$;
         Calculate $q_{target}$;)
         Perform a gradient descent step on
         $(q_{target} - Q(s_j, a_j; \boldsymbol{W}))^2$;
    end for
    Update the target network $\boldsymbol{W}^* = \boldsymbol{W}$
end for
Return: Trained DQNs and state-action values.
```

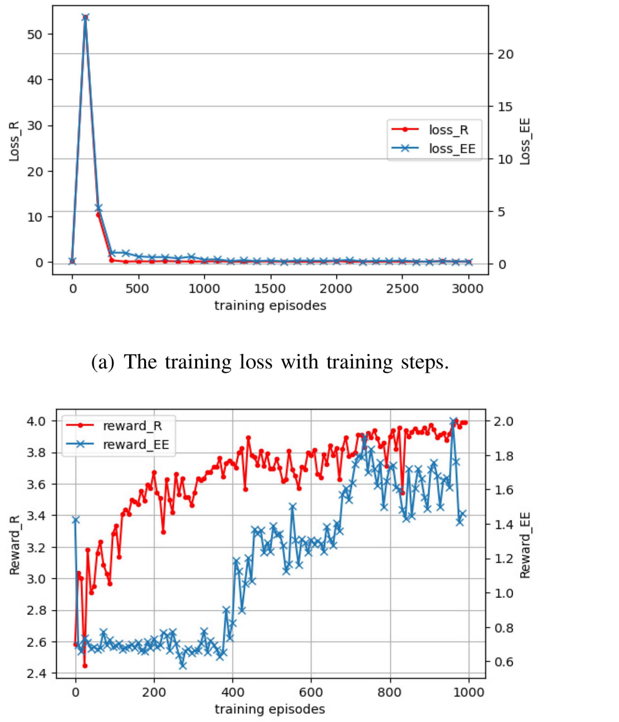
B. DQN Training Process

Acting as the agent, the BS takes state $\mathcal{S}$ as the input and outputs the evaluated Q-value for the state-action pairs. As mentioned above, two parallel Q-value tables are generated to represent EE and system throughput, therefore, two parallel DQNs, i.e., *DQN_Thr* and *DQN_EE*, are invoked to evaluate the two Q-value tables. The BS chooses between *DQN_Thr* and *DQN_EE* according to current threshold and trains NNs as shown in Fig. 1. To collect the training data set, the replay memories $\mathcal{D}_{EE}$ and $\mathcal{D}_{Thr}$ are leveraged. If the throughput is higher than the threshold, *DQN_EE* is selected and the transition $\boldsymbol{e}_t^{EE} = (s_t, a_t, r_t, s_{t+1})$ would be stored in $\mathcal{D}^{EE}$, where r_t is calculated by (9). During training process, a minibatch $\boldsymbol{e}_j^{EE} = (s_j, a_j, r_j, s_{j+1})$ is sampled from training data set $\mathcal{D}^{EE}$.

Each DQN consists of two NNs, called evaluation network and target network, respectively. The evaluation network takes state s_j and action a_j as the input and outputs the evaluated Q-value, while the target network takes the next state s_{j+1} as the input and generates the target Q-value for the *j*th minibatch. To minimize the difference between the estimated Q-value and the target Q-value, we define the loss function

$$\text{Loss}(\boldsymbol{W}) = \mathbb{E}[(q_{target} - Q(s_j, a_j; \boldsymbol{W}))^2], \tag{12}$$

where $q_{target} = r_j + \Gamma \max_{a'} Q(s_{j+1}, a'; \boldsymbol{W}^*)$ represents the target Q-value. Additionally, $\boldsymbol{W}^*$ and $\boldsymbol{W}$ denotes the weights of the target and evaluation network, respectively. The weights W are optimized by the gradient descent method [9]. *DQN_Thr* can be trained in a similar approach with *DQN_EE* by the reward in (10). The details of the proposed algorithm are shown in Algorithm 1. In each training step, only one DQN is chosen and trained, so the computational complexity of the proposed algorithm for each training step is same as the legacy DQN algorithm.



(a) The training loss with training steps.

(b) The training reward with training episodes.

Fig. 2. Performance of the training loss and reward of the parallel DQNs, $M = 10$, $N = 30$.

IV. Numerical Results

In this section, performance of the proposed dynamic power allocation and reward conversion method based on parallel DQN algorithm is presented. In the simulation, D2D pairs and cellular users are uniformly distributed over a $1000m \times 1000m$ area. The maximum transmit power p_{max} and minimum transmit power p_{min} are 30dBm and 0dBm, respectively. The power variable quantity $\delta = 1$dBm. The minimum SINR requirements for D2D receiver γ^{D*} and for cellular uplink γ^{C*} are -20dB and -26dB, respectively. Gaussian noise variance σ^2 is -116dBm. Path loss parameter d is set to 3. The system throughput threshold is set to 3.6Mbps. We set the exploration $\epsilon = 1 - 0.8 \times t/T$, where t represents current episode and and $T = 1000$ is maximum training episodes.

Fig. 2 shows the training performance of the parallel DQNs. In Fig. 2(a), the training losses of both DQNs decrease quickly with the increasing number of training episodes. After around 500 training steps, the losses of both DQNs remain unchanged at very low level. Fig. 2(b) shows the average reward per episode. At the beginning 400 training episodes, the reward for EE remains stable, while reward for throughput increases fast until reaching the predefined threshold. Then the reward for EE is increased, achieving more than two folds of the initial EE value, and the reward for throughput keeps meeting the threshold.

Fig. 3 shows the EE comparison of proposed algorithm with other benchmarks. The optimal curve is acquired by the exhaustive search. We can see that when the number of D2D users is relatively small, EE of all algorithms drops as the number of D2D users increases. Both the proposed algorithm and the legacy RL can achieve the near-optimal EE when number of D2D users is less than 10. As the number of D2D users becomes more than 15, the legacy RL cannot converge within the training episodes because of the high complexity, and the gap between the optimal result and legacy RL becomes bigger, while the proposed algorithm achieves the near-optimal EE.

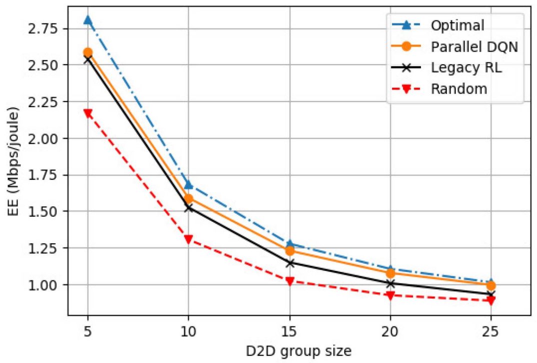


Fig. 3. System EE for different D2D group size.

V. Conclusion

This letter proposes a deep RL enabled dynamic power optimization algorithm with parallel DQNs. The novelty of this letter lies in designing an improved reward function and DQN structure where the tradeoff is obtained between achieving acceptable system throughput and maximising the EE. The results show that the proposed algorithm achieves a good balance and much higher EE is obtained compared to the traditional power optimization methods while providing guaranteed QoS.

References

[1] D. Feng, L. Lu, Y. Yuan-Wu, G. Y. Li, S. Li, and G. Feng, "Device-to-device communications in cellular networks," *IEEE Commun. Mag.*, vol. 52, no. 4, pp. 49–55, Apr. 2014.

[2] Y. Kai, J. Wang, H. Zhu, and J. Wang, "Resource allocation and performance analysis of cellular-assisted OFDMA device-to-device communications," *IEEE Trans. Wireless Commun.*, vol. 18, no. 1, pp. 416–431, Jan. 2019.

[3] M. R. Mili, P. Tehrani, and M. Bennis, "Energy-efficient power allocation in OFDMA D2D communication by multiobjective optimization," *IEEE Wireless. Commun. Lett.*, vol. 5, no. 6, pp. 668–671, Dec. 2016.

[4] H. Chen, D. Wu, and Y. Cai, "Coalition formation game for green resource management in D2D communications," *IEEE Commun. Lett.*, vol. 18, no. 8, pp. 1395–1398, Aug. 2014.

[5] Z. Qin, H. Ye, G. Y. Li, and B. F. Juang, "Deep learning in physical layer communications," *IEEE Wireless Commun.*, vol. 26, no. 2, pp. 93–99, Apr. 2019.

[6] H. Ye, G. Y. Li, and B. F. Juang, "Deep reinforcement learning based resource allocation for V2V communications," *IEEE Trans. Veh. Technol.*, vol. 68, no. 4, pp. 3163–3173, Apr. 2019.

[7] L. Liang, H. Ye, G. Yu, and G. Y. Li, "Deep-learning-based wireless resource allocation with application to vehicular networks," *Proc. IEEE*, vol. 108, no. 2, pp. 341–356, Feb. 2020.

[8] G. Yang, Y. Liao, Y.-C. Liang, and O. Tirkkonen, "Reconfigurable intelligent surface empowered underlaying device-to-device communication," Jun. 2020. [Online]. Available: arXiv:2006.02103.

[9] V. Mnih *et al.*, "Playing atari with deep reinforcement learning," 2013. [Online]. Available: arXiv:1312.5602.