Deep Reinforcement Learning for Channel Selection and Power Allocation in D2D Communications

Jun Zhou*
School of Electronics and Information Engineering
Shenzhen University
Shenzhen, China
* Corresponding author: 1900432044@email.szu.edu.cn

Abstract—Device-to-device (D2D) communication is regarded as a key technical component of the fifth-generation (5G). D2D communication usually reuses spectrum resources with cellular users (CUs). To mitigate interference to cellular links and improve spectrum efficiency, this paper investigates a sum-rate maximization problem in the underlay of D2D communication. Particularly, a joint channel selection and power allocation framework based on multi-agent deep reinforcement learning is proposed, named Double Deep Q-Network (DDQN). It can adeptly select the channel and allocate power in a dynamic environment. The proposed scheme only requires local information and some outdated nonlocal information, which reduces signaling overheads significantly. Simulation results show that the proposed scheme can improve the D2D sum rate and ensure quality-of-service (QoS) of CUs compared with other benchmarks.

1. INTRODUCTION

D2D communication is a promising technology, it allows two proximal users to communicate with each other directly. D2D communication has many advantages. First, D2D communication can improve the spectrum and energy efficiency of the network. Second, D2D communication can bring less communication delay to users [1]. Third, D2D communication can offload heavy mobile traffic from the base station (BS) at a low cost.

Despite the many potential benefits, the interference management of the D2D network is a challenge. Power control and channel selection are effective methods of interference mitigation, attracting the attention of many scholars. There are many channel selection and power allocation methods in the existing literature, which can be divided into the traditional scheme and machine learning-based. In traditional schemes, the authors in [2] propose an algorithm based on game theory, which is used for joint channel selection and power control in D2D networks. Weighted MMSE (WMMSE) is used for power allocation [3]. In [4], the authors develop a fractional programming algorithm (FP) for resource allocation. These algorithms require massive information and instantaneous signaling exchanges, and the computational complexity of these algorithms is high.

Machine learning is a new resource allocation tool in wireless communication [5], especially reinforcement learning [6]. The authors in [7] use centralized Q-learning and distributed Q-learning for D2D power control. In [8], a two-step resource allocation scheme is proposed to control the power of D2D, the first step uses graph theory to select the channel, and the second step uses Q-learning to control D2D power. The authors in [9] propose a novel Q-learning-based cognitive internet of things (IoT) transmission scheduling mechanism, which aims to maximize the system throughput. Authors in [10] develop a convergent fully autonomous Q-learning algorithm in heterogeneous cellular networks.
Although the above work performs well in wireless resource allocation, they all assume a static channel model.

In this paper, a distributed deep reinforcement learning is proposed, named Double Deep Q-Network (DDQN) [11], to optimize the transmission power and selection channel of the D2D pair. The proposed scheme only requires local information and neighbor users' historical information, this is easier to implement in practice.

The main contributions of this paper are as follows:
1) The proposed algorithm is model-free, which is based on Double Deep Q-Network and can adapt to changes in the dynamic wireless environment.
2) The reward function of the proposed scheme is novel and unique. The proposed algorithm does not rely on existing training data. Each agent continuously interacts with the environment, which can generate data for training. The base station uses these data to train the network centrally.

2. System Model and Problem Formulation

2.1. System Model
As shown in Fig. 1, This paper considers a single-cell communication scenario, where D2D users and cellular users coexist. A set of M CUs denoted as $M = \{1, 2, \ldots, M\}$, and a set of N active D2D pairs, denoted as $N = \{1, 2, \ldots, N\}$. Assuming the BS is located in the center of the cell, and D2D users and cellular users are randomly distributed. Each user is equipped with one antenna. The network provides K orthogonal channels, denoted as $K = \{1, 2, \ldots, K\}$, a fully-loaded network is considered, the orthogonal channel is equal to the number of CUs, $M=K$. For brevity, cellular user $k$ access channel $k$.

At slot $t$, the channel gain from transmitter $x$ to receiver $y$ on the channel $k$, which can be written as,

$$g_{x,y,k}^{(t)} = |h_{x,y,k}^{(t)}|^2 \beta_{x,y}^k$$

where $h_{x,y,k}^{(t)}$ is the small-scale Rayleigh fading component, $\beta_{x,y}^k$ is the large-scale fading component, which includes path loss and log-normal shadowing. This paper adopts Jake’s model to describe Rayleigh fading. Accordingly, the small-scale Rayleigh fading for each channel can be modeled as a first-order complex Gauss-Markov process

$$h^{(t)} = e h^{(t-1)} + \epsilon$$

where $\epsilon \in [0,1)$ is the correlation factor between the two consecutive time slots. $h^{(0)} \sim CN(0,1)$ and $\epsilon \sim CN(0, \sqrt{1-\epsilon^2})$.
\( \chi_{n,k}^{(t)} \) is the indicator of channel allocation, which is given as follows:

\[
\chi_{n,k}^{(t)} = \begin{cases} 
1, & \text{if D2D pair } n \text{ accesses channel } k; \\
0, & \text{otherwise.}
\end{cases}
\]  

(3)

The SINR for cellular user \( k \) on the channel \( k \) is given as follows:

\[
\gamma_{c,k}^{(t)} = \frac{p_c g_{k,bs}^{(t)}}{\sum_{n} \chi_{n,k}^{(t)} p_{n,k} g_{n,bs,k}^{(t)} + \sigma^2}
\]  

(4)

The transmit power of all cellular users is the same, denoted by \( p_c \). \( p_{n,k}^{(t)} \) is the transmission power of the \( n^{th} \) D2D pair on channel \( k \). \( g_{k,bs}^{(t)} \) is the channel gain from the CU \( k \) to the BS. \( g_{n,bs,k}^{(t)} \) is the channel gain from the transmitter of D2D pair \( n \) to BS over the channel \( k \). \( \sigma^2 \) is the Gaussian white noise power.

The received SINR of D2D pair \( n \) on the channel \( k \) is given as follows:

\[
\gamma_{d,n,k}^{(t)} = \frac{p_{n,k}^{(t)} g_{n,n,k}^{(t)}}{p_c g_{k,n}^{(t)} + \sum_{j} \chi_{j,k}^{(t)} p_{j,n}^{(t)} g_{j,n,k}^{(t)} + \sigma^2}
\]  

(5)

where \( g_{n,n,k}^{(t)} \) is the channel gain from D2D transmitter \( n \) to D2D receiver \( n \) over the channel \( k \). \( g_{k,n}^{(t)} \) is channel gain from the CU \( k \) to BS on the channel \( k \). \( g_{j,n,k}^{(t)} \) is the channel gain from D2D transmitter \( j \) to D2D receiver \( n \) on the channel \( k \).

At time slot \( t \), the rate obtained by D2D pair \( n \) on channel \( n \), which can be expressed as,

\[
U_{n,k}^{(t)} = \log_2(1 + \gamma_{d,n,k}^{(t)})
\]  

(6)

2.2. Problem Formulation

In this paper, to maximize the D2D sum-rate and ensure the communication quality of cellular users, the problem is described as follows:

\[
\begin{aligned}
& \max_{p_{n,k}} \sum_{k=1}^{K} \sum_{n=1}^{N} U_{n,k}^{(t)} \\
& \text{s.t.} \quad \gamma_{c,k}^{(t)} > \gamma_{\min} \forall k \in K \\
& \quad \chi_{n,k}^{(t)} \in \{0, 1\}, \forall n \in N, \forall k \in K \\
& \quad \sum_{k=1}^{K} \chi_{n,k}^{(t)} \leq 1, \forall n \in N \\
& \quad \sum_{n} \chi_{n,k}^{(t)} \leq 1, \forall n \in N \\
& \quad p_{n}^{(t)} \leq p_{\max}, \forall n \in N
\end{aligned}
\]  

(7)

where \( \gamma_{\min} \) represents the requirements of the minimum SINR for CUs. \( \chi_{n,k}^{(t)} \) represents the channel allocation indicator for D2D pairs \( n \). \( \sum_{k=1}^{K} \chi_{n,k}^{(t)} \) ensures that at most one channel is allocated to one D2D pair. \( p_{n}^{(t)} \) represents the maximum transmit power of D2D pairs.

3. Reinforcement Learning Based Power Control Algorithm

This section addresses the channel selection and power allocation problem in underlaying cellular networks. This paper first models a multi-agent environment, then a distributed reinforcement learning algorithm is proposed, it only uses local information to train the network, greatly reducing communication overhead.

3.1. Overview of Reinforcement Learning

At time step \( t \), the agent takes action \( a_t \) according to the currently observed state \( s_t \). The agent gets a reward \( r^{t+1} \) and the environment transits to a new state \( s_{t+1} \). The agent continuously interacts with the environment, and the system reaches a dynamic balance. In the RL, the goal is to maximize the total discount cumulative reward, as given by,
\[ R^t = \sum_{k=0}^{\infty} \mu^{k} r^{t+k+1} \quad (8) \]

where \( \mu \in [0,1] \) is the discount factor.

Q-learning is the most classic reinforcement learning algorithm, it first creates a Q-table, the agent continuously interacts with the environment, which can update the Q-value. Under policy \( \pi \), the Q-value of the action taken \( a \) in state \( s \)

\[ Q^\pi(s,a) = E[R^t | s^t = s, a^t = a] \quad (9) \]

In order to find the optimal Q-value \( Q^*(s,a) = \max_\pi Q^\pi(s,a) \), according to the Bellman optimality equation, the optimal Q-value can be updated continuously. The update equation of the action-value function \( q(s^t,a^t) \), which is given as follows:

\[ q(s^t,a^t) = q(s^t,a^t) + \alpha [r^{t+1} + \mu \max_a q(s^{t+1},a^{t+1}) - q(s^t,a^t)] \quad (10) \]

3.2 Multiagent Reinforcement Learning Algorithm

In a multi-agent system, the state space, \( S \), the action space, \( A \), and the reward function \( R \), each D2D pair acts as an agent, each part of reinforcement learning is defined as follows:

**State space:** To optimize the objective function (7), and enable the agent to better perceive the environment, the agent not only considers local observations but also considers the historical information of neighbor users, which can make the agent selects the appropriate channel and transmit power. In this paper, each agent considers three neighbors, including the Interfering Neighbors and the Interfered Neighbors. Local information can be obtained in time, assuming that obtaining non-local information requires a slot delay. The state obtained by this agent consists of four parts.

This first part is the local observation information: To perceive changes in the environment, local information considers the measurement of the previous one-time slots, the previous transmit power of agent \( n \), \( p_{n,k}^{(t-1)} \), the previous channel selected by the agent \( n \), \( k_{n}^{(t-1)} \), the previous spectrum efficiency achieved by agent \( n \), \( U_{n,k}^{(t-1)} \), the instant channel information of the agent \( n \), \( g_{n,n,k}^{(t)} \), total interference-plus-noise power at agent \( n \), \( p_c g_{k,n}^{(t)} + \sum_{j \neq n} \chi_{j,n}^{(t-1)} p_{j,n}^{(t-1)} g_{j,n,k}^{(t)} + \sigma^2 p_c \).

This second part considers the state of the base station: To ensure the communication quality of CUs, this state allows the agent to pay attention to the interference to the BS, \( p_n^{(t-1)} g_{n,b}^{(t-1)} \).

The third part is the state of Interfering Neighbors: Interfering neighbor users include cellular users or other D2D pair transmitters, which can reduce the spectrum efficiency of agent \( n \). The previous transmit power of interfering neighbor user \( j \), \( p_{j,k}^{(t-1)} \), the previous channel selected by the interfering neighbor user \( j \), \( k_{j}^{(t-1)} \), the previous spectrum efficiency achieved by interfering neighbor user \( j \), \( U_{j,k}^{(t-1)} \), the interference from interfering neighbor user \( j \) to agent \( n \), \( \chi_{j,n}^{(t-1)} p_{j,n}^{(t-1)} g_{j,n,k}^{(t-1)} \), the direct channel gain of interfering neighbor user \( j \), \( g_{j,n,k}^{(t-1)} \).

The fourth part is the state of the Interfered Neighbors: The interfered neighbor is the receiver of other D2D pairs. Due to the interference of the agent \( n \), the rate value of the user of the interfered neighbor will decrease. The previous transmit power of interfered neighbor user \( i \), \( p_{i,k}^{(t-1)} \), the previous channel selected by the interfered neighbor user \( i \), \( k_{i}^{(t-1)} \), the previous spectrum efficiency achieved by interfered neighbor user \( i \), \( U_{i,k}^{(t-1)} \), the interference from agent \( n \) to its interfered neighbor user \( i \), \( \chi_{i,n}^{(t-1)} p_{n,k}^{(t-1)} g_{n,i,k}^{(t-1)} \), the direct channel gain of interfered neighbor user \( i \), \( g_{i,n,k}^{(t-1)} \).

**Action space:** At time slot \( t \), each agent observes the current state, according to the output of DDQN, it can determine its transmission power and channel selection. The maximum transmits power of the D2D pair is \( p_m \), which is discretized into L levels. The total number of channels is represented by \( K \), therefore, the dimension of the behavior space is \( K*L \). The action space \( A \) of the agent as,
\[ A = \{(k, p) | \forall k \in K, p \in \{0, \frac{P_m}{L}, \frac{2P_m}{L}, \ldots, p_m\}\} \tag{11} \]

**Reward function:** The learning process of reinforcement learning is driven by the reward function. When interacting with the environment, each agent maximizes its reward by taking appropriate actions. To optimize the goal (7), and ensure the QoS for CUs, the reward function consists of three parts, the first part is the rate of the agent \( n \) itself.

\[
U^{(t)}_n = \log_2(1 + \frac{p^{(t)}_{n,k}g_{n,n,k}}{p_c g_{k,n} + \sum_{j \in M_n} x^{(t)}_{j,k} p^{(t)}_{j,k} g_{j,n,k} + \sigma^2})
\tag{12}
\]

The second part is the reduced value of the spectrum efficiency of the interfered neighbor. \( M^R_n \) represents the set of interfered neighbors. \( U^{(t)}_{i \setminus n} \) spectrum efficiency obtained by neighbor \( i \) without considering the interference of agent \( n \), which can be written as:

\[
U^{(t)}_{i \setminus n} = \log_2(1 + \frac{p^{(t)}_{i,k} g_{i,i,k}}{p_c g_{k,i} + \sum_{j \in M_i} x^{(t)}_{j,k} p^{(t)}_{j,k} g_{j,i,k} + \sigma^2})
\tag{13}
\]

Due to the interference of agent \( n \), the total spectral efficiency reduction value of the interfered neighbor. which can be written as:

\[
\bar{U}^{(t)}_n = \sum_{i \in M^R_n} (U^{(t)}_{i \setminus n} - U^{(t)}_i)
\tag{14}
\]

The third part is the penalty for interference to base station, it may degrade the QoS of CUs. If the minimum SINR of the cellular user is not satisfied, and agent \( n \) will be punished according to the ratio of its interference to the total interference, which is given as follows:

\[
l^{(t)}_n = \frac{p^{(t)}_{n,k} g_{k,s}}{\sum_j p^{(t)}_{j,k} g_{j,s}} \beta
\tag{15}
\]

where \( \beta \) is an adjustable coefficient, which is given as,

\[
\beta = \begin{cases} 
0, & \gamma_c > \gamma_{\text{min}}; \\
C, & \text{otherwise}.
\end{cases}
\tag{16}
\]

If the minimum SINR of the cellular user can be satisfied, \( \beta \) is equal to 0, otherwise \( \beta \) is equal to \( C \), C is a non-negative variable, and its size can affect the magnitude of interference from D2D users to CUs.

The reward function of agent \( n \in N \) taking action \( a^{(t)}_n \) is given as,

\[
r^{(t)}_n = U^{(t)}_n - \bar{U}^{(t)}_n - l^{(t)}_n
\tag{17}
\]

As shown in Figure 2, the training method adopts centralized training and distributed execution. In the process of distributed execution. At slot \( t \), each pair of D2D transmitters is based on the current state \( s_t \), which includes the user’s local observations and the historical information of neighbor users. All agents synchronize and take actions \( a_t \) at the same time, and transition to the next state \( s_{(t+1)} \), and the environment will feedback one reward \( r_t \). When the communication is in a good state, each pair of D2D will upload historical information to the base station at time slot \( t \), including \( (s_t, a_t, r_t, s_{(t+1)}) \). The process of centralized training is implemented at the base station. All information from all agents is used to train a network, which can effectively reduce the problem of insufficient memory and computing resources caused by training. At the same time, using similar parameter sharing methods, all agents can share the same set of parameters, which can also make the algorithm converge faster. In the training process of the network, the base station will broadcast the trained weights to the D2D transmitter every time \( T \), until the network reaches the state of convergence.
Figure 2. Description of the proposed DRL algorithm.

In the problem of channel and power allocation, all D2D links update their strategies independently, but this is a multi-agent environment. The state transition of the agent is not only affected by its behavior, but also by the joint behavior of other agents. Its environment is unstable for any agent, and it is difficult to converge in a multi-agent environment.

To solve the problem of environmental instability, a deep reinforcement learning method using neighbor user information is proposed. Each agent can not only obtain timely local observations, and also obtain the historical information of neighbor users. In this way, each agent can overcome the unstable environment caused by multi-agents, which can better perceive the dynamic environment.

At time step \( t \), each agent \( n \in N \) takes actions \( a_n^t \) according to the current state \( s_n^t \), and get timely rewards \( r_n^t \). The agent will transition from the current state \( s_n^t \) to the new state \( s_n^{t+1} \). Then, the experience samples are stored in the experience replay buffer. In the training phase, each batch of data is extracted from the experience replay set to train the network, the loss function is given as follows:

\[
\text{Loss}(\theta) = \frac{1}{B} \sum_{t=1}^{B} (y_n - Q(s,a) | \theta) \tag{18}
\]

where \( Q(s,a) \) is the output value of train_DQN, and \( y_n \) is the target Q-value, which is given as follows:

\[
y_n = r_n^{t+1} + \mu q(s_n^{t+1}, a_n^{t+1} | \theta) \tag{19}
\]

where \( q(s_n^{t+1}, a_n^{t+1} | \theta) \) is the estimated value of the target Q-network, and the behavior \( a_n^{t+1} \) is to select the behavior with the largest Q-value in the train_DQN, which can solve the problem of behavior overestimation and better choose behavior according to the current state.

4. SIMULATION RESULTS

This paper considers a single-cell scenario with a radius of 500m, the base station is located in the center of the cell, and users are randomly distributed. The parameters that were used for the simulation are listed as follows. The bandwidth of the subband is 180 KHz. The transmit power of CUs is 23 dBm, and the max transmit power of D2D users is 23 dBm. The number of cellular users and channels is 10. The number of D2D pairs is \( N \in [5,25] \). The maximum communication distance between the D2D transmitter (TX) and receiver (RX) is 50 m. The noise power spectral density is -174dBm/Hz. The minimum SINR for CUs is 8dB. The path loss model for cellular links is \( 128.1 + 37.6 \log(d_{km}) \). The pathloss model for D2D links is \( 148 + 40 \log(d_{km}) \). The Shadowing standard deviation is 10dB for cellular mode links, and 12 dB for D2D mode links. Small-scale fading is Rayleigh fading. If the received SINR is greater than 30 dB, it is capped at 30 dB.

A five-layer feed-forward neural network (FNN) is chosen, and the neuron numbers of three hidden layers are 100 and 50, 50, respectively. The first two layers use “Relu” as the activation function, and the
third layer uses “Tanh” as the activation function. The “Tanh” activation function can accelerate the convergence of the network. The learning rate is 0.001. The memory size R is set to be 2000, and the minibatch size is 256. Adam optimizer is adopted to minimize the loss function Eq. (18). The learning rate is 0.001.

The Q-learning algorithm, Random power allocation, and Open-loop power allocation schemes are treated as benchmarks to evaluate the proposed algorithm.

![Graph](image)

**Figure 3.** With different discount values, the recorded average rate during the training period.

In this subsection, the performance of different discount factors $\mu$ is studied, discount indicates the impact of future rewards on the current. The number of D2D is set to 10. The $\mu \in \{0, 0.3, 0.5, 0.7, 0.9\}$, and the average rate over the training period is shown in Fig. 3. As the $\mu$ value increases, the average rate obtained by D2D users decreases. It shows that $\mu = 0$, the algorithm can obtain the very highest average rate. The proposed algorithm can converge at about 4000 slots. In the following experimental simulation, $\mu = 0$.

Fig. 4 compares the QoS Satisfying Degree of different algorithms. The number of D2D is set to 10. The proposed scheme can achieve a higher QoS Satisfaction Degree than the other four algorithms and guarantee the communication quality of CUs. The proposed scheme can ensure the QoS of cellular users than other benchmarks. Here is a further explanation that the algorithm converges in 4000 slots. The QoS Satisfaction Degree is given as follows:

$$QoS\ Satisfaction\ Degree = \frac{N_1}{N_2}$$  \hspace{1cm} (20)

where $N_1$ is the number of cellular users that meet the minimum SINR, and $N_2$ is the total number of CUs.
Figure 4. Compare different algorithms with QoS Degree Satisfaction.

Figure 5. D2D sum rate for different number of D2D links

Fig. 5 shows the variation of the D2D sum rate with the number of D2D pairs. The result implies that as the number of D2D pairs increases, the D2D sum rate of the network increases. The result implies that our proposed algorithm can achieve a higher D2D sum rate than Q-learning, Random, and Open-loop.

5. CONCLUSION
This paper has studied the channel selection and power allocation in the underlay of D2D communication, and formulated the D2D sum-rate maximization problem as a decentralized multi-agent deep reinforcement learning problem. To ease implementation and to improve stability, the training method adopts centralized training and distributed execution. The proposed scheme only requires local information and some outdated nonlocal information. Simulation results show that the proposed scheme can improve the D2D sum rate and ensure QoS of CUs compared with other benchmarks.

REFERENCES
[1] M. Shafi et al., “5G: A Tutorial Overview of Standards, Trials, Challenges, Deployment, and Practice,” in IEEE Journal on Selected Areas in Communications, vol. 35, no. 6, pp. 1201-1221, June 2017.
[2] X. Wang, Z. Qian, Y. Cong and X. Wang, “Joint Channel and Power Allocation Based on Stackelberg for D2D Communications in Cellular Networks,” 2020 IEEE Globecom Workshops GC Wkshps, 2020, pp.1-6.

[3] Q. Shi, M. Razaviyayn, Z. Luo and C. He, “An Iteratively Weighted MMSE Approach to Distributed Sum-Utility Maximization for a MIMO Interfering Broadcast Channel,” in IEEE Transactions on Signal Processing, vol. 59, no. 9, pp. 4331-4340, Sept. 2011.

[4] K. Shen and W. Yu, “Fractional Programming for Communication Systems—Part I: Power Control and Beamforming,” in IEEE Transactions on Signal Processing,” vol. 66, no. 10, pp. 2616-2630, 15 May15, 2018.

[5] W. Lee, M. Kim and D. Cho, “Deep Learning Based Transmit Power Control in Underlaid Device-to-Device Communication,” in IEEE Systems Journal, vol. 13, no. 3, pp. 2551-2554, Sept. 2019.

[6] R. S. Sutton, A. G. Barto, et al., “Introduction to reinforcement learning.” vol. 135. MIT press Cambridge, 1998.

[7] S. Nie, Z. Fan, M. Zhao, X. Gu and L. Zhang, “Q-learning based power control algorithm for D2D communication,” 2016 IEEE 27thAnnual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), Valencia, Spain, 2016, pp. 1-6.

[8] S. Maghsudi and S. Sta´nczak, “Hybrid Centralized–Distributed Resource Allocation for Device-to-Device Communication Underlaying Cellular Networks,” in IEEE Transactions on Vehicular Technology, vol. 65, no. 4, pp. 2481-2495, April 2016.

[9] J. Zhu, Y. Song, D. Jiang and H. Song, “A New Deep-Q-Learning-Based Transmission Scheduling Mechanism for the Cognitive Internet of Things,” in IEEE Internet of Things Journal, vol. 5, no. 4, pp. 2375-2385, Aug. 2018.

[10] A. Asheralieva and Y. Miyanaga, “An Autonomous Learning-Based Algorithm for Joint Channel and Power Level Selection by D2D Pairs in Heterogeneous Cellular Networks,” in IEEE Transactions on Communications, vol. 64, no. 9, pp. 3996-4012, Sept. 2016.

[11] J. Dai, J. Liu, Y. Shi, S. Zhang, and J. Ma, “Analytical modeling of resource allocation in D2D overlying multihop multichannel uplink cellular networks,” IEEE Trans. Veh. Technol., vol. 66, no. 8, pp. 6633–6644, Aug. 2017.