Dynamic Power Allocation in D2D Communications Using Deep Reinforcement Learning

Jun Zhou

1 School of Electronics and Information Engineering, Shenzhen University, Shenzhen, Guangdong,518000, China
*Corresponding author’s e-mail: 17369423604@163.com

Abstract. Due to the development of communication technology, mobile devices continue to increase. As one of the critical technologies of 5G, D2D communication is an up-and-coming technology. In this paper, multiple D2D pairs usually multiplex the same channel, which will cause serious channel interference. To solve this problem, a distributed deep reinforcement learning framework is proposed, which can well adapt to the power allocation in a dynamic environment. The simulation results show that compared with other benchmark methods, the proposed scheme can improve the overall D2D rate.

1. Introduction
In the future, wireless network data transmission has experienced explosive growth, the density of users is increasing, and the interference between users is also increasing, how to improve the capacity of the system, which is very important. In this paper, the problem of maximizing system capacity is investigated.

Different from data-driven methods, many articles have studied the issue of power control. In traditional schemes, Game theory is used for the performance of device-to-device (D2D) communication links through carrier sense multiple [1]. Weighted MMSE (WMMSE) is used as a linear beamformer to maximize system throughput [2]. In [3], a fractional programming algorithm (FP) is used to optimize discrete variables. The above algorithm requires a lot of timely channel state information (CSI).

Machine learning is a new method of resource allocation in the field of wireless communication [4], especially reinforcement learning. The authors in [5] propose centralized Q-learning and distributed Q-learning methods, and shows the advantages of distributed Q-learning. In [6], authors propose a two-stage resource allocation strategy, which guarantee the quality-of-service (QoS) of D2D users. The author proposes a Q-learning-based cognitive radio network transmission scheduling mechanism, the proposed algorithm can transmit data packets better, with lower power consumption and packet loss rate [7]. The authors propose a fully autonomous multi-agent Q-learning algorithm for heterogeneous cellular networks with multiple base stations. The algorithm can show faster convergence after a few iterations [8].

In this paper, a distributed deep reinforcement learning method, called Deep-Q Network (DQN), is proposed, which can be well adapted for power control in dynamic environments [9]. The main contributions are summarized as follows:

1) The proposed method uses centralized training and distributed execution methods, which fully utilizes the computing power of the base station.
2) The proposed algorithm relies on the data generated by the environment to train the network, and is well adapted to power control in a dynamic environment.

2. System Model and Problem Formulation

2.1 System Model

As shown in Fig. 1, in this paper, considers a full-power multiplexed full-synchronization dedicated D2D network. The network contains N pairs of D2D links, it is represented as \( N = \{1, 2, \ldots, N\} \). The base station covers the entire cell and is located in its center, the D2D links are evenly distributed within the coverage of the base station. Each user's transmitter and receiver are equipped with an antenna.

![Fig.1 System Model](image)

At slot \( t \), \( g_{x,y}^{(t)} \) represents the channel gain in time from the D2D transmitter \( x \) to the receiver \( y \), which can be written as,

\[
g_{x,y}^{(t)} = \| h_{x,y}^{(t)} \|^2 \beta_{x,y} \tag{1}
\]

where \( h_{x,y}^{(t)} \) is the small-scale fading, small scale fading is described using Rayleigh fading, \( \beta_{x,y} \) is the large-scale fraction containing path loss and shadow fading, which includes path loss and shadow fading. According to Jake’s model, Rayleigh fading is modeled as the first sequence of complex Gaussian Markov processes:

\[
h^{(t)} = \varepsilon h^{(t-1)} + \sqrt{1 - \varepsilon^2} e^{(0)} \tag{2}
\]

The correlation \( \varepsilon = J_0(2\pi f_d) f_d \) is maximum Doppler frequency, \( J_0 \) is the zeroth-order Bessel function of the first kind. \( h(0) \) and \( e(0) \) follow a complex normal distribution.

The received signal-to-interference-plus-noise ratio (SINR) of D2D link \( n \), which is written as follows:

\[
\gamma^{(t)}_{d,n} = \frac{p^{(t)}_{n,k} g_{n,n}^{(t)}}{\sum_{j\neq n}^{N} p^{(t)}_{j,n} g_{j,n}^{(t)} + \sigma^2} \tag{3}
\]

At the slot \( t \), \( g_{n,n,k}^{(t)} \) represents the channel gain of the D2D link \( n \) in time. \( g_{j,n}^{(t)} \) represents the channel gain from the transmitter of D2D link \( k \) to the receiver end of D2D link \( n \), \( \sigma^2 \) represents Gaussian white noise.

The rate obtained by D2D link \( n \), which is expressed as follows:

\[
U^{(t)}_n = \log_2(1 + \gamma^{(t)}_{d,n}) \tag{4}
\]
2.2 Problem Formulation

This paper investigates the D2D total rate maximization problem, users can select a reasonable transmission power to minimize the interference between users. The problem can be planned as follows:

\[
\max_p \sum_{n=1}^{N} U_n^{(t)} \\
\text{s.t. } p_n \leq p_{\text{max}}, \forall n \in N
\]  

(5) restricts the maximum transmission power of D2D pair not to exceed \( p_{\text{max}} \).

3 Deep Reinforcement Learning Based Power Control

3.1 Overview of reinforcement learning

At time slot \( t \), each agent takes action \( a^t \) according to the current state \( s^t \), The environment will feedback to the agent a reward \( r^t \), and agent from the current state \( s^t \) to the next state \( s^{t+1} \). The agent achieves dynamic balance by constantly interacting with the environment. In the reinforcement learning, the goal of reinforcement learning is to maximize the cumulative discount reward, which is written as follows,

\[
R^{(t)} = \sum_{k=0}^{\infty} \mu^k r^{t+k+1}
\]  

(6)

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

The most classic reinforcement learning algorithm is Q-learning, it maintains a Q-table, and the agent keeps interacting with the environment to keep up with the new Q-table. In the current state, the Q-value is expressed as follows:

\[
Q^\ast(s, a) = E[R^t \mid s^t = s, a^t = a]
\]  

(7)

In order to find the optimal Q value, the update of the Q function obeys Bellman's equation, which is given as follows:

\[
g(s', a') = g(s', a') + \alpha \gamma r^{t+1} + \mu \max_a q(s^{t+1}, a^{t+1}) - g(s', a')
\]  

(8)

3.2 Deep Multiagent Reinforcement Learning

Reinforcement learning contains several basic elements, the action space \( A \), the state space \( S \), and the reward function \( R \), basic elements represented by a tuple \(< S, A, R >\), which is defined as follows:

**State space:** To maximize the objective function (6), and enable the agent to better adapted to the environment, consider measuring the local state of the agent in two time slots, and the interfering neighbor users and the interfered neighbor users consider the information of the previous one time slot. In this article, the state of each agent is composed of three different state information, local observation information, information about interfering neighbor users, and information about interfered neighbor users.

Local Information: The user's local information takes into account the information of the first two time slots, which allows the agent to better perceive changes in the environment. the transmit power of agent \( n \), \( p_n^{(t-1)} \), the spectrum efficiency achieved by agent \( n \), \( U_n^{(t-1)} \), the instant channel gain of the agent \( n \), \( g_n^{(1)} \), the channel gain of the agent \( n \), \( g_n^{(2)} \), the interference received by agent \( n \), \( \sum_{j \neq n} p_{j,n}^{(t-2)} g_{j,n}^{(t-1)} \), and \( \sum_{j \neq n} p_{j,n}^{(t-2)} g_{j,n}^{(t-2)} \).
Interfering Neighbors State: Interfering with the spectrum efficiency of neighbor agent j, $U_j^{(r-1)}$, interference to agent n, $p_f^{(r-1)}g_{j,n}^{(r-1)}$, channel gain of interfering neighbor user j, $g_{j,j}^{(r-1)}$, interfering with the channel gain from neighbor user j to agent n, $g_{j,n}^{(r-1)}$.

Interfered Neighbors State: Spectrum efficiency obtained by interfered neighbor i, $U_i^{(r-1)}$, interference of agent n to interfered neighbor i, $p_f^{(r-1)}g_{j,n}^{(r-1)}$, the direct channel gain of interfered neighbor user i, $g_{i,i}^{(r-1)}$, channel gain from agent n to interfered neighbor i, $g_{i,n}^{(r-1)}$.

Action space: The agent needs to choose the appropriate power, which can reduce the interference between users, $p_{\text{max}}$ is maximum transmit power of D2D link, which is discretized into L levels, the action space $A$ is given as follows:

$$A = \{ p \mid p \in \{ 0, \frac{2p_{\text{max}}}{L}, \ldots, p_{\text{max}} \} \}$$  \hspace{1cm} (9)

Reward function: Reinforcement learning is driven by a reward function, which is related to the rate of acquisition of each agent, which can better optimize the objective function (5), the reward function of the agent consists of two parts. The first part is the rate obtained by agent n.

$$U_n^{(i)} = \log_2(1 + \frac{p_n^{(i)}g_{n,n}^{(i)}}{\sum_{j,n\neq n} p_f^{(i)}g_{j,n}^{(i)} + \delta^2})$$ \hspace{1cm} (10)

When the interfered neighbor is interfered, the rate obtained will be reduced, and the second part is the reduced rate. When the interference of agent n is not considered, the rate obtained by agent i is denoted by $U_{i\setminus n}^{(i)}$, which can be given as:

$$U_{i\setminus n}^{(i)} = \log_2(1 + \frac{p_i^{(i)}g_{i,i}^{(i)}}{\sum_{j,n\neq n} p_f^{(i)}g_{j,j}^{(i)} + \delta^2})$$ \hspace{1cm} (11)

Total spectral efficiency reduction value of the interfered neighbor, which is given as follows:

$$\overline{U}_n^{(i)} = \sum_{i=\mathcal{M}_n} (U_{i\setminus n}^{(i)} - U_i^{(i)})$$ \hspace{1cm} (12)

The set representing the disturbed neighbor is denoted by $\mathcal{M}_n$.

Finally, the reward function obtained by agent n, which is given as:

$$r_n^{(i)} = U_n^{(i)} - \overline{U}_n^{(i)}$$ \hspace{1cm} (13)
In order to make full use of the computing resources of the base station, the neural network adopts centralized training, distributed execution at each D2D transmitter, and centralized training to simplify implementation and improve stability. Each agent \( n \) shares the same set of DQN parameters. As shown in Figure 2, in the training phase, the centralized network trainer fetches small batches of data from the experience replay buffer, and these are minimized by optimizing the equation (14). The network training is carried out at the base station, which can make full use of the powerful computing power of the base station, and use the information of all agents to train the same network, which allows the agents to learn faster and accelerates the convergence of the network. After a while, the base station sends the trained network weights to each D2D pair through the backhaul link, and then the D2D transmitter loads the trained neural network parameters, and the D2D user’s transmitter selects the appropriate one according to the current state, behavior.

There are two reasons why reinforcement learning can maintain intelligence. First, the way of experience replay makes good use of past experience, and secondly, the correlation of the data is broken by the method of target network. In the time slot \( t \), the agent notices the current state \( s \), takes actions \( a \), the agent transitions to the next state, and obtains a reward, and then the experience samples are stored in the experience replay buffer. After collecting enough experience samples, one B samples are taken from the experience replay buffer to calculate the loss function, which is written as follows:

\[
\text{Loss}(\theta) = \frac{1}{B} \sum_{i=1}^{B} (y_n - Q(s, a) | \theta) \tag{14}
\]

\( y_n \) is the output Q-value of the target DQN, which is given as follows:

\[
y_n = r^{t+1} + \mu Q(s^{t+1}, a^{t+1} | \theta') \tag{15}
\]

4 Simulation Results

Consider a single-cell uplink communication scenario, users are randomly distributed within the coverage of the base station. The maximum distance from the D2D transmitter to the receiver is 60m, the maximum transmission power of the user is 5.42781W, and the number of D2D links is \( N \in [5, 17] \). Due to the actual digital signal processing, the user’s maximum SINR is 30dB, and Gaussian white noise is -114dBm/Hz. The pathloss model for D2D links is \( 128.1 + 37.6\log(d_{km}) \) The Shadowing standard deviation is 10dB.

This paper considers a five-layer neural network, one input layer, one output layer, and three hidden layers. The number of neurons in the hidden layer is 100, 50, and 50. In order to ensure the convergence of the network, the hidden layer uses "Relu" as the activation function, and the size of the experience replay buffer set to 2000, the minibatch size is set to 128, Adam optimizer is used to minimize the loss.
function, which can speed up network convergence and find the optimal solution quickly, the learning rate is set to 0.001.

The paper uses four benchmark algorithms to show the superiority of the proposed algorithm. The Q-learning algorithm, Random power allocation, Fractional Programming (Fp), power allocation, and P-max power allocation schemes.

In Fig 3, the performance of the same learning rate and different batch size algorithms is compared. It can be seen from the figure that as the training time increases, the algorithm learning under different batch size conditions reaches a stable state, the learning performance of the algorithm is not sensitive under the conditions of different batch sizes.

![Fig.3 Algorithm learning performance with same learning rate and different batch sizes.](image)

In Fig 4, when the D2D link in the system increases, the interference between users becomes more and more serious, and the average rate obtained by the D2D link will decrease. Random power allocation and maximum power allocation are relatively simple to allocate power, and the interference between links is the largest. When the number of links increases, the average rate of the links drops the fastest. It can be seen from the figure that the proposed algorithm is compared with the other four algorithms. With the increase of D2D links, users can obtain a higher average rate and can better select the appropriate transmit power in a dynamic environment.

![Fig.4 The performance of the average rate obtained by the DQN algorithm different links](image)

5 Conclusion

This article studies the problem of power allocation in a dynamic environment. The power allocation problem is formulated as a D2D total rate maximization problem. In this article, a distributed power control method is proposed. Each agent can take appropriate actions according to the current state, and can adapt to power control problems in a dynamic environment. In the future, the power control problem will be applied to the scenario of millimeter-wave communication, and the joint optimization
of channel and power will be considered, the current environment will be closer to the natural environment.

References
[1] Lyu, J., Chew, H.Y., Wong, W.C., (2016) A Stackelberg Game Model for Overlay D2D Transmission With Heterogeneous Rate Requirements. in IEEE Transactions on Vehicular Technology., 65: 8461-8475.
[2] Shi, Q., Razaviyayn, M., Luo, Z.Q, He, C., (2011) An Iteratively Weighted MMSE Approach to Distributed Sum-Utility Maximization for a MIMO Interfering Broadcast Channel. in IEEE Transactions on Signal Processing., 59: 4331-434.
[3] Shen, K., Yu, W., (2018) Fractional Programming for Communication Systems—Part II: Uplink Scheduling via Matching, in IEEE Transactions on Signal Processing., 66: 2631-2644.
[4] Lee, W., Kim, M., Cho, D., (2019) Deep Learning Based Transmit Power Control in Underlaid Device-to-Device Communication. in IEEE Systems Journal., 13: 2551-2554.
[5] Nie, S., Fan, Z., Zhao, M., Gu, X., Zhang, L. (2016) Q-learning based power control algorithm for D2D communication. In 2016 IEEE 27th Annual International Symposium on Personal. Valencia., pp: 1-6.
[6] Maghsudi, S., Stańczak, S., (2016) Hybrid Centralized–Distributed Resource Allocation for Device-to-Device Communication Underlaying Cellular Networks. IEEE Transactions on Vehicular Technology., 65: 2481-2495.
[7] Zhu, J., Song, Y., Jiang, D., Song, H., (2018) A New Deep-Q-Learning-Based Transmission Scheduling Mechanism for the Cognitive Internet of Things. in IEEE Internet of Things Journal., 5: 2375-2385.
[8] Asheralieva, A., Miyanaga, Y., (2016) An Autonomous Learning-Based Algorithm for Joint Channel and Power Level Selection by D2D Pairs in Heterogeneous Cellular Networks. in IEEE Transactions on Communications., 64: 3996-4012.
[9] Nasir, Y.S., Guo, D., (2019) Multi-Agent Deep Reinforcement Learning for Dynamic Power Allocation in Wireless Networks. in IEEE Journal on Selected Areas in Communications., 37: 2239-2250.