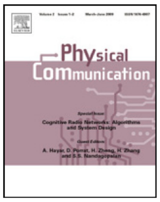
Physical Communication 45 (2021) 101262

Contents lists available at ScienceDirect

Physical Communication

journal homepage: www.elsevier.com/locate/phycom



Full length article

Joint resource allocation and power control for D2D communication with deep reinforcement learning in MCC☆

Dan Wang[a], Hao Qin[a], Bin Song[a,1,*], Ke Xu[a], Xiaojiang Du[b,2], Mohsen Guizani[c,2]

[a] State Key Laboratory of Integrated Services Networks, Xidian University, 710071, China
[b] Department of Computer and Information Sciences, Temple University, Philadelphia PA 19122, USA
[c] Department of Engineering, Qatar University, Qatar

ARTICLE INFO

Article history:
Received 31 May 2019
Received in revised form 6 November 2020
Accepted 23 December 2020
Available online 26 December 2020

Keywords:
D2D
DRL
Resource allocation
Power control

ABSTRACT

Mission-critical communication (MCC) is one of the main goals in 5G, which can leverage multiple device-to-device (D2D) connections to enhance reliability for mission-critical communication. In MCC, D2D users can reuses the non-orthogonal wireless resources of cellular users without a base station (BS). Meanwhile, the D2D users will generate co-channel interference to cellular users and hence affect their quality-of-service (QoS). To comprehensively improve the user experience, we proposed a novel approach, which embraces resource allocation and power control along with Deep Reinforcement Learning (DRL). In this paper, multiple procedures are carefully designed to assist in developing our proposal. As a starter, a scenario with multiple D2D pairs and cellular users in a cell will be modeled; followed by the analysis of issues pertaining to resource allocation and power control as well as the formulation of our optimization goal; and finally, a DRL method based on spectrum allocation strategy will be created, which can ensure D2D users to obtain the sufficient resource for their QoS improvement. With the resource data provided, which D2D users capture by interacting with surroundings, the DRL method can help the D2D users autonomously selecting an available channel and power to maximize system capacity and spectrum efficiency while minimizing interference to cellular users. Experimental results show that our learning method performs well to improve resource allocation and power control significantly.

1. Introduction

With the development of the fifth-generation (5G) cellular networks, mission-critical communication (MCC) [1] and Gigabit mobile connectivity will be supporting diverse modern applications and services [2]. As one of the main goals in the 5G network, MCC imposes enormous challenges to fulfill its vital performance. At present, the demand for enhancing MCC technology becomes more and more urgent, especially for further improving both the network communication capacity and spectrum efficiency. As one of the most critical techniques employed in MCC networks, device-to-device (D2D) communication allows mobile devices to perform direct peer-to-peer transmissions by reusing the licensed spectrum allocated to cellular services, and to enhance reliability for mission-critical communication by leveraging multiple D2D connections [3]. Recently, D2D technology has attracted the attention of academia and industry because of its ability to provide mobile services with large capacity, high speed, and guaranteed quality-of-service (QoS) [4]. In an emergency where network infrastructure is absent, the D2D network is critical for the (Mission-critical) MC site to exchange information and provide MCC networks the MC voice and data services. D2D communications can be in-band or out-band [5], which reuse the non-orthogonal wireless resources of the cellular user by selecting a communication mode and performing data transmission without a base station (BS) in D2D networks. D2D users inevitably generate co-channel interference to cellular users when they multiplex channel resources. As the interference intensifies, the communication of cellular users may be interrupted [6]. Generally, there are three types of interference in the underlying cellular and D2D communication system, i.e., the D2D-to-cellular interference, the cellular-to-D2D interference, and the D2D-to-D2D

☆ This work has been supported by the National Natural Science Foundation of China (Nos. 61772387, 62071354), the National Natural Science Foundation of Shaanxi Province, China (Grant Nos. 2019ZDLGY03-03), Graduate Student Innovation Fund (10221150004), and also supported by the ISN State Key Laboratory.

* Corresponding author.
E-mail addresses: wangdanxdty@gmail.com (D. Wang), hqin@mail.xidian.edu.cn (H. Qin), bsong@mail.xidian.edu.cn (B. Song), ke_xu_39@163.com (K. Xu), dxj@ieee.org (X. Du), mguizani@ieee.org (M. Guizani).

[1] Senior Member, IEEE.

[2] Fellow, IEEE.

https://doi.org/10.1016/j.phycom.2020.101262

interference [7]. At present, many studies have been conducted to address these problems. They have proposed tremendous effective solutions to the channel interference of D2D users, amongst which maximizing resource utilization and improving the system capacity gain the most attention of research on D2D communication. However, in most studies, game-theoretical approaches are adopted, which is unsuitable for complicated communication scenarios due to their high computational complexity [8]. Recently, reinforcement learning (RL) is a prevalent and effective algorithm to solve wireless communication problems, especially for interference management, resource allocation, and power control [9,10]. RL is a learning method with decision-making ability, which mainly includes agent, state, action, and policy. During the learning process, an agent can make the decision, interact with the environment, and then automatically explore its strategy to get the optimal policy. However, as the state and action spaces become more massive in a complex communication network, it will be more difficult or even impossible to find the optimal policy [11]. Deep reinforcement learning (DRL), a combination of RL and deep learning, has been developed to overcome the above shortcomings [12,13]. This paper aims to propose a single-objective optimization approach, namely, sophisticated joint resource allocation and power control mechanism with DRL. The method can mitigate interference and enhance the spectrum efficiency as well.

In this paper, we investigate the resource allocation and power control problems in which the D2D pairs utilize the uplink resources of cellular users. We consider a 5G network scenario that involves multiple cells with multiple cellular users, D2D pairs, and a BS. Our goal is to maximize the total system capacity while guaranteeing the QoS of cellular users in different MC services. The main contributions of this work as follows:

(1) The algorithms of the D2D communications interference problem have been researched and compared in detail to further reveal the issues and dig out the potential recommendation.

(2) The system model has been developed to meet the optimization goal by creating a DRL algorithm to jointly improve resource allocation and power control for D2D users which consume different services in cellular systems.

(3) The problem is decomposed into two separate subproblems to clarify the breaking point, and this procedure helps with targeting the objectives of our solution and, subsequently, the in-depth design of our objective optimization.

(4) Furthermore, this method achieved the goal of comprehensively improving the QoS of the system, such as optimizing system capacity and simultaneously reducing interference.

The remainder of this paper is organized as follows. Section 2 briefly introduces the related works of resource allocation and power control in D2D communications. Section 3 describes the proposed system model and optimization goal. Also, Section 4 provides the detailed design of the method, combining joint resource allocation and power control with DRL. Section 5 proves our achievement by showing the performance evaluation and analysis of the proposed algorithm after running a series of experiments, and Section 6 finally concludes the paper.

2. Related works

Recently, there are three main aspects in the management of D2D interference in D2D communications, namely mode selection, resource allocation, and power selection. Increasingly new methods have been proposed to reduce communication interference in D2D communication. In traditional communication research, an interference avoidance mechanism has been introduced in the hybrid cellular and the D2D system, mitigating the interference from the cellular transmission to D2D communication by users' mode selection [14]. Moreover, the method of joint mode selection and resource allocation scheme has been studied to improve users' throughput, extending the battery lifetime of user equipment by facilitating the reuse of spectrum resources between D2D and cellular links [15].

In addition to traditional communication methods, game theory and RL methods have become a popular method for solving wireless communication interference management problems. In [16], the author proposed joint scheduling and resource allocation algorithms and adopt the Stackelberg game to improve D2D communication performance, where cellular TIE and D2D TIE are grouped into leader–follower pair. The author developed a coalitional game with transferable utility. Each user intended to maximize its efficiency and had the incentive to cooperate with other users to form a strengthened user group, thus increasing the opportunity to win its preferred spectrum resources [17].

The RL method has been used instead of the traditional way and game theory to achieve resource allocation, mode selection, and power control. The author proposed two power control methods with RL for D2D users, namely team-Q learning and distributed-Q learning, to achieve power control in D2D communication [18]. They regard D2D communication as a multi-agent system, and power control is achieved by maximizing system capacity while maintaining the requirement of QoS from cellular users. The authors in [19] presented power control for D2D communication, which used multi-agent reinforcement learning (MARL) to maximize system throughput by adjusting the transmitted power of each D2D user. In [20], the authors proposed a joint mode selection and power adaptation approach using a conjecture based multi-agent Q-learning algorithm.

Although RL has some advantages in solving some problems in communication networks, it still has limitations. Specifically, when confronting the complicated network system and the large state–action space, the RL shows poor performance, and its convergence speed may suffer.

Therefore, the DRL approach is to address emerging problems in communications and networking [21]. In [22], the author aims at maximizing the sum rate of a D2D network under the assumption of realistic time-varying channels and D2D interference. They proposed to use a centralized DRL transmission scheme for D2D communications, in which transmission decisions are made by one agent expertized in the D2D network. At present, few articles are adopting the DRL method for resource allocation in D2D communication. There are often large action spaces and state space in joint power control and resource allocation issues. Hence, Q-learning can no longer meet the task requirements. Deep Q-Learning (DQL) can provide effective solutions for these problems [23]. In our work, a joint channel and power allocation algorithm with DQL has been investigated, which can be used to solve the problem with high dimensional state space and complexly discrete action space. In the next section, the system model of D2D communication underlying cellular networks will be introduced in detail.

3. System model

In this work, a small cellular communication system includes two basic types of communication modes, a direct D2D communication mode, and a traditional cellular communication mode.

We consider that M cellular users and N D2D pairs are deployed in each cell. Each D2D pair consists of a D2D transmitter ($\mathrm{DT_n}$) and a D2D receiver (DR_n) where D2D pairs reuse the same spectrum resource as cellular users. In addition, we assume that (1) in D2D communication, cellular users utilize the uplink (UL) resources of a small cell, while D2D pairs reuse the uplink resources non-orthogonally, (2) a cellular user and D2D pairs share a same resource block and each resource block is allocated to one cellular user and shared with multiple D2D pairs. As illustrated in Fig. 1, the D2D pairs reuse the UL resource in the central cell. Therefore, there are two kinds of interference, and the one is D2D-to-cellular interference, the other is cellular-to-D2D interference [23]. In the multi-cell model, in addition to the above interference, D2D communication interferes with the communication of neighboring cells when neighboring cell users use the identical spectrum. Cellular users and D2D users are subject to interference from nearby cell users.

We assume that the channel bandwidth is B, which is divided into K physical resource blocks (PRBs). Each channel bandwidth is defined as $b_i = \frac{B}{K}, i \in \{1, 2, \ldots, K\}$. In the multi-cellular scenario, our goal is to learn an effective joint channel allocation and power control strategy for each D2D transmitter based on different MCC services. We consider that D2D users have L service types, denoted by $S_l \in \{S_1, S_2, \ldots, S_L\}$, and each service has different requirements for channel transmission rates. We consider that a D2D pair can reuse multiple channel resources to ensure successful transmission of packets while meeting the QoS requirements of the entire communication system with minimal power consumption. Additionally, the signal to interference plus noise ratio (SINR) of the cellular user can be expressed as $\gamma_i^{C_n}$. For successful transmission, the SINR is above γ^*:

$$\gamma_i^{C_n} > \gamma^*, \ \forall \mathrm{i} \in N \tag{1}$$

where γ^* is a threshold in different service types. The SINR of the nth D2D link on the ith channel is defined as:

$$\gamma_i^{D_n} = \frac{g_i^{D_n} P_i^{D_n}}{\sigma^2 + P_i^{C_m} \cdot g_i^{C_m} + \sum_{x=1,x\neq n}^{N} P_i^{D_x} \cdot g_i^{D_x} + N_a} \tag{2}$$

where $g_i^{D_n}$ denotes the channel gain between DT_n and DR_n, $P_i^{D_n}$ is the transmission power of the nth D2D link on the ith channel. On the ith channel, $P_i^{C_m}$ is the transmission power of the mth cellular user, and $P_i^{D_x}$ is the transmission power of the xth D2D link. The $g_i^{C_m}$ and $g_i^{D_x}$ are the link gain of mth cellular user and xth D2D link on the ith channel. Here, σ^2 is the power of the Additive White Gaussian Noise (AWGN). N_a is the interference coming from neighboring cells, a represents the average noise of other cells. It denotes as:

$$N = G \cdot \sum P_z \cdot d_z^{-2} \quad \forall \mathrm{z} \in \{1, 2, \ldots, Z\} \tag{3}$$

where z is the number of neighboring cells, d is the distance between two cells, and G is the link gain. Then, the SINR of the mth cellular user on the ith channel is given by:

$$\gamma_i^{C_m} = \frac{P_i^{C_m} \cdot g_i^{C_m}}{\sigma^2 + \sum_{y=1}^{N} \sum_{j=1}^{K} P_j^{D_y} \cdot g_j^{D_y} + N_a} \tag{4}$$

where $P_i^{C_m}$ is the transmission power of the mth cellular user, $g_i^{C_m}$ is the link gain of the mth cellular user. $P_j^{D_y}$ is the transmission power of the yth D2D link that reuse the jth channel. Hence, the system capacity of a cell is given by:

$$\mathrm{C} = \mathrm{B} \sum_{i=1}^{K} [\sum_{n \in N} \log 2 \left(1 + \gamma_i^{D_n}\right) + \log 2 \left(1 + \gamma_i^{C_m}\right)] \tag{5}$$

In this paper, we assume that D2D users can reuse multiple channels under different services. To guarantee the communication quality of cellular users, we consider the joint channel selection and power control method, which is a single-object optimal problem. We focus on maximizing the total networking capacity of the cellular system and meeting the QoS requirements.

4. Resource allocation and power control method with deep reinforcement learning

The goal of this paper is that the D2D transmitter learns an efficient joint channel selection and power control policy after interacting with the environment. Generally, the increase of the D2D users' transmission power and the more interference for the cellular users as a result of the increasingly D2D users reuse channel. In our model, each D2D pair can adaptively learn multi-channel selection and power control strategies, maximizing system capacity and meeting service demands. The above problem is a decision-making problem that can be solved by adopting RL methods. In a communication scenario, the D2D user can select more than one channel, which complicates the decision-making problem with large state and action spaces. Therefore, we adopt a DRL method to solve the issues, which can significantly improve the learning speed, especially the problems with large state and action spaces [24]. In this section, we first introduce the framework of the DRL algorithm with multiple users. Based on that, we then design the resource allocation and power control algorithm in D2D communication.

4.1. Deep reinforcement learning

We model the D2D interference problem as a Markov Decision Processes (MDP), which is a decision-making problem. Generally, MDP is defined as a tuple (S_t, A_t, P, R), in which (S_t, A_t, P, R) is a set of states, a set of actions, state transition probabilities, and the reward function, respectively [25,26]. The target of MDP is to find a optimal policy, then solving the RL decision-making problem (that is, to maximize the reward function). In MDP, the agent first senses the environment state $s_t \in S_t$, and take a random action $a_t \in A_t$ by interacting with the environment. Then, the agent generates a new state s_{t+1} and gives an immediate reward $r_t \in R$. In the procedure, the policy is defined as $\pi\left(a|s\right) = \mathrm{p}[\, a_t |\, s_t]$, which is a mapping from state to action. Further data are generated through interaction between the agent and the environment to optimize the strategy. After many iterations, the agent learns an optimal policy π^*. Generally, the future reward is denoted as:

$$G_t = r_{t+1} + \mathit{\Gamma} r_{t+2} + \cdots = \sum_{k=0}^{\infty} \mathit{\Gamma}^k r_{t+k+1}, (\mathrm{r} \in R) \tag{6}$$

where $\mathit{\Gamma} \in [0, 1]$ is a discount factor. When the agent adopts the policy π, the action-value function is defined as:

$$\mathrm{Q}\left(\mathrm{s}_t, \mathrm{a}_t\right) = \mathrm{E}[\sum_{k=0}^{\infty} \mathit{\Gamma}^k r_{t+k+1} \Big| s_t, a_t, \pi] \tag{7}$$

The Q-value is evaluated as follows,

$$\mathrm{Q}\left(\mathrm{s}_t, \mathrm{a}_t\right) = \mathrm{Q}\left(\mathrm{s}_t, \mathrm{a}_t\right) + \alpha[r_t + \mathit{\Gamma} \max_a \mathrm{Q}\left(\mathrm{s}_{t+1}, a_{t+1}\right) - \mathrm{Q}\left(\mathrm{s}_t, \mathrm{a}_t\right)] \tag{8}$$

where $\alpha \in [0, 1]$ is the learning rate, $r_t + \mathit{\Gamma} \max_a \mathrm{Q}\left(\mathrm{s}_{t+1}, a_{t+1}\right)$ is the expected value. The process is repeated until the agent obtains the optimal policy π^*. The optimal Q-value $Q^*\left(\mathrm{s}_t, \mathrm{a}_t\right)$ can be defined as:

$$Q^*\left(\mathrm{s}_t, \mathrm{a}_t\right) = \mathrm{Q}^{\pi^*}\left(\mathrm{s}_t, \mathrm{a}_t\right) \tag{9}$$

The value function is:

$$V^*\left(\mathrm{s}\right) = \max_a Q^*\left(\mathrm{s}_t, \mathrm{a}_t\right) \tag{10}$$

Generally, the $Q\left(s_t, a_t\right)$ is estimated by a linear function approximator. However, a non-linear function approximator is used to estimate the action-value function in the DRL, such as a neural

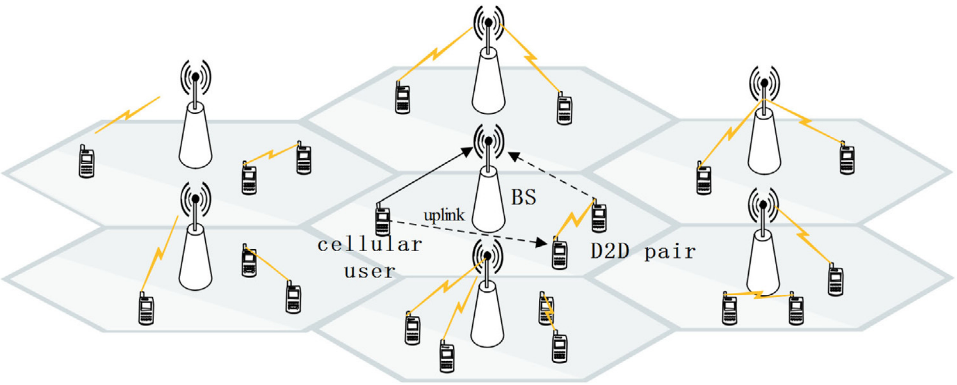


Fig. 1. A scenario where a D2D link is located in the small cells (uplink and downlink).

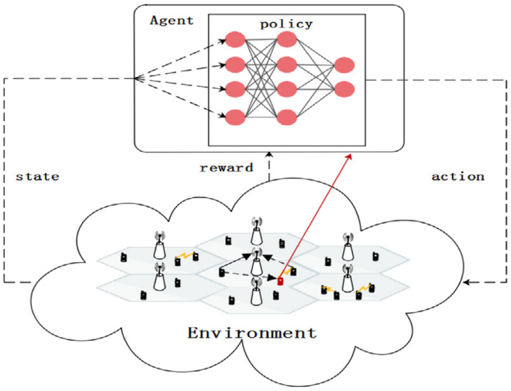


Fig. 2. Deep reinforcement learning for D2D communication in multiple small cells.

network (that is Q-network) [27]. In the DRL, the Q-value is defined as:

$$Q(s_t, a_t, \theta) \approx Q^*(s_t, a_t) \tag{11}$$

where θ is the weight of the network, and it also represents the value function. A Q-network can be trained to update θ through the gradient descent. The loss function $L_i(\theta_i)$ is denoted as:

$$L_i(\theta_i) = [r_j + \Gamma \max_{a_{j+1}} Q\left(s_{j+1}, a_{j+1}; \theta_{i-1}\right) - Q\left(s_j, a_j; \theta_i\right)]^2 \tag{12}$$

Hence, the update of the value function is given as follows:

$$\theta_{i+1} = \theta_i + \alpha \left[r_j + \Gamma \max_{a_{j+1}} Q\left(s_{j+1}, a_{j+1}; \theta_{i-1}\right) - Q\left(s_j, a_j; \theta_i\right) \right] \times \nabla Q\left(s_j, a_j; \theta_i\right) \tag{13}$$

4.2. Resource allocation and power control method

The DRL framework of D2D communications is illustrated in Fig. 2. In the scenario, we assume that an agent is a D2D transmitter in each D2D pair. In a cell, there are many D2D users. The scenario is the multi-agent system. The environment is multiple cellular users and D2D users. During the interaction period between agents and the environment, the D2D transmitter takes action, including select channel and power level. Then we introduce the state, action space, reward function and update the rule of channel allocation and power control problem.

4.2.1. State and action space of D2D users

We assume that the local SINR information of the D2D link is available, and the information of the cellular user is also available. We consider learning on one RB and dividing it into K PRBs. D2D transmitter can reuse multiple (physical resource block) PRB and control its power level to achieve optimal capacity under different services. Hence, each agent has the same learning goal. In this work, the environment is straightforward, including cellular users and D2D links.

Agent: In our proposed model, we design D2D transmitters as agents. Each agent is an individual with the abilities of learning and decision-making. A D2D transmitter is expressed as DT_x, $1 \leq x \leq N$, where N denotes the number of the D2D link.

States: At the time t, the state is determined by the channel and power level. We define the state space, including the channel state of uses, the state of power level and the number of the D2D pairs. It is defined as a three-dimensional matrix, as follows:

$$S(t) = \begin{bmatrix} s_{11}(t) & \cdots & s_{1K}(t) \\ \vdots & \ddots & \vdots \\ s_{N1}(t) & \cdots & s_{NK}(t) \end{bmatrix} = \begin{bmatrix} [c_{11}(t), p_{11}^l(t)] & \dots & [c_{1K}(t), p_{1K}^l(t)] \\ \vdots & \ddots & \vdots \\ [c_{N1}(t), p_{N1}^l(t)] & \dots & [c_{NK}(t), p_{NK}^l(t)] \end{bmatrix} \tag{14}$$

where $S(t)$ denotes state space, $c_{NK}(t)$ denotes the channel state and $p_{Nl}(t)$ denotes power level. We define $c_{NK}(t)$ as:

$$\begin{cases} c_{ij} = 1, \textit{the ith D2D user reuse jth channel} \\ c_{ij} = 0, \textit{otherwise} \end{cases}$$
$$(\forall i\epsilon\{1, 2, \ldots, N\}, j\epsilon\{1, 2, \ldots, K\}) \tag{15}$$

In addition, $p_{ij}^l(t)$ $(\forall i\epsilon\{1, 2, \ldots, N\}, j\epsilon\{1, 2, \ldots, K\}, l\epsilon 1, 2, \ldots, L)$ denotes the ith D2D user selects the jth channel and the lth power level. The transmission power is divided into L levels. If there are K PRB resource blocks, the dimension of action space is $L * K$. Hence, the state is complicated in our learning process. At the time t, the agent sends a communication request, and when cellular users, agents, and other D2D users share a same channel, there is interference between them. The huge state space makes it difficult to learn information with Q-learning, so we utilize the deep Q-learning to learn the high-dimensional inputs. The state space is inputted to the deep Q-network. We adopt a convolutional neural network (CNN) to learn features of the three-dimensional matrix.

Action: At the time t, the action is defined as:

$$A(t) = \{A_1(t), A_2(t)\} \tag{16}$$

where $A_1(t)$ represents to select channel, and $A_2(t)$ represents to select power level. More specifically, in our learning model, the action is defined as follows:

$$\begin{cases} A_1\left(\mathrm{t}\right) = \left(a_1^k, a_2^k, \ldots, a_n^k\right), \forall k \in \{1, 2, \ldots, K\}, n \in \{1, 2, \ldots, N\} \\ A_2\left(\mathrm{t}\right) = \left(p_1^k, p_2^k, \ldots, p_l^k\right), \forall k \in \{1, 2, \ldots, K\}, l \in \{1, 2, \ldots, L\} \end{cases} \tag{17}$$

where k is the kth PRB, and n is the number of D2D pairs, a_n^k represents that nth D2D transmitter select kth PRB. In addition, l is the power level, and p_l^k represents the power level of the agent in$k$th channel.

Algorithm 1

Algorithm 1: Joint resource allocation and power control method

```
begin
  Initialization:
    For $t = 0$, $t = (t_1, \ldots, t_N)$
        Randomly create a state matrix: $S(t)$
        Create an action matrix: $A(t) = 0$
      Initialization D2D system model parameter
        Initialization D2D user reuse m channel of one cellular user
        Set different services of D2D users, $S$
        (Min. Bandwidth Requirements)
      D2D user randomly select a first channel and power level
    End for
  Processing:
  Loop:
   For $i$ in $N$, do
  (1) Selected channel $C$ and power $P$, A(t)=[0,1,…]
  (2) Calculate:
        $\gamma_i^{C_m}$ of the ith channel of the cellular user
        $\gamma_i^{D_n}$ of the ith D2D pair
        system capacity $C$
  (3) Check SINR to guarantee QoS of users according to Constraints
  (4) Run Algorithm 2, learning channel and power selection policy
  (5) If the D2D transmission restarts in this time slot
        End if
   End for
     Set $t = t + 1$
     Create a new potential state matrix: $S(t)$
  End loop
end
```

Reward function: Generally, the reward function is the learning goal. Our learning goal is to optimize the total system capacity, so we define the system capacity as the reward function. Therefore, we assume that the reward function is related to system capacity and constraints. The reward function is expressed as:

$$r_t = \begin{cases} C, \mathit{if\ the\ constraints\ are\ satisfied} \\ -\zeta C, \mathit{otherwise}. \end{cases} \tag{18}$$

where C denotes system capacity. In the model, we propose an approach based on different service types S of D2D users to guarantee their communication requests and meet the QoS demands of cellular users. Here, we propose the QoS demands metric for the MCC scenario, where the QoS metric, namely total system capacity. The target of our approach is maximizing the total system capacity and meanwhile maintaining the level of the QoS for D2D users. We define the constraints as follows:

$$\begin{cases} \gamma_i^{C_n} \geq \tau_0, \\ 0 \leq P_i^{D_n} \leq P_{max}, \\ \mathrm{C}_c \geq \mathrm{C}_{c,s}, \\ \mathrm{C}_{D,i} \geq \mathrm{C}_{D,s}, \end{cases} \tag{19}$$

where $\gamma_i^{C_n}$ is the SINR of the cellular user, τ_0 is the threshold of SINR. To ensure the communication quality of the cellular link, we consider the impact on the cellular user SINR when D2D user reuses the spectrum resource. When the SINR is greater than a threshold τ_0, the maximum power at this time is set to the transmit power of the D2D user. The $P_i^{D_n}$ is the transmit power of the D2D transmitter and P_{max} is the maximum D2D transmit power. The C_c represent the bandwidth requirements of the different type of services. The bandwidth requirements are different when the service arrives each time. Therefore, when the agent performs channel selection and power control, it should meet the service requirements.

Our method not only ensures the normal communication of cellular users but also maximizes the multiplexing of channel resources and optimize system capacity. When the above conditions are met, the reward is C, otherwise, a penalty is given. The r_t is an immediate reward. Long-term reward is the sum of all immediate rewards, and it denotes as:

$$G_t = \sum_{k=0,t=0}^{\infty} \Gamma^k r_t, (r_t \in R) \tag{20}$$

4.2.2. Updating algorithm

We have described the system model in the previous section, in which the D2D transmitter acts as an agent. The agent interacts with the environment, then takes action to the environment. In the learning process, the agent continuously updates the policy according to the rules of the DRL algorithm until the optimal strategy is learned. Our approach combines channel selection and power selection, in which the agent has two different actions to achieve the goal. Scenario with multiple D2D users and channels lead to large state space. In our approach, when the target network is updated, two actions are output. The dimensions of each action are different.

We adopt a deep Q-learning network (DQN) to learn resource allocation and power control policy. In the algorithm, we use a CNN instead of the Q-table to derive an approximate Q-value. Our training network is shown in Fig. 3. It is a five-layer network where the last layer is divided into two sub-layer, one for channel selection and the other for power selection. The output is two Q-values.

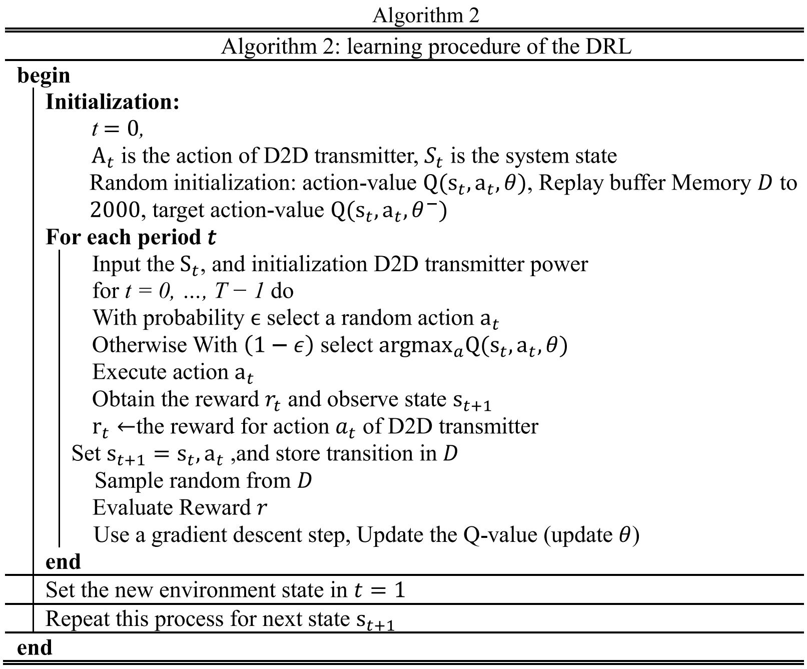
The update rules are shown in Algorithm 1 and Algorithm 2. Algorithm 1 shows the procedure of resource allocation and power control. Algorithm 2 mainly illustrates the update step of DQN. The method of selecting the channel and power by the D2D transmitter can ensure the QoS of cellular users and reduce interference.

During training, the DQN uses CNN to approximate the Q-value function of the selected channel and power. Regardless of channel selection or power level selection, actions can be learned by exploring the strategies under constraints. In this paper, we choose to use the Boltzmann distribution to balance exploitation and exploration. It is denoted by:

$$\mathrm{P} = \frac{e^{Q_{i,k,j}^t(a)/\tau}}{\sum_i e^{Q_{i,k,j}^t(a)/\tau}} \tag{21}$$

where $Q_{i,k,j}^t(a)$ is the Q-value for action when D2D transmitter select channel k and power j at time t. The τ is a temperature parameter, which controls the fluctuation of this Boltzmann distribution.

Here, we use the data of experience replay to train the neural network. Specifically, the state, action, reward and the next state are stored in the memory database each time. Then the data is sampled by uniform sampling, and the neural network is trained



Algorithm 2

```
Algorithm 2: learning procedure of the DRL
begin
  Initialization:
      $t = 0$,
      $A_t$ is the action of D2D transmitter, $S_t$ is the system state
      Random initialization: action-value $Q(s_t, a_t, \theta)$, Replay buffer Memory $D$ to
      2000, target action-value $Q(s_t, a_t, \theta^-)$
  For each period $t$
      Input the $S_t$, and initialization D2D transmitter power
      for $t = 0, \ldots, T-1$ do
      With probability $\epsilon$ select a random action $a_t$
      Otherwise With $(1-\epsilon)$ select $\mathrm{argmax}_a Q(s_t, a_t, \theta)$
      Execute action $a_t$
      Obtain the reward $r_t$ and observe state $s_{t+1}$
      $r_t$ ←the reward for action $a_t$ of D2D transmitter
    Set $s_{t+1} = s_t, a_t$ ,and store transition in $D$
      Sample random from $D$
      Evaluate Reward $r$
      Use a gradient descent step, Update the Q-value (update $\theta$)
  end
  Set the new environment state in $t = 1$
  Repeat this process for next state $s_{t+1}$
end
```

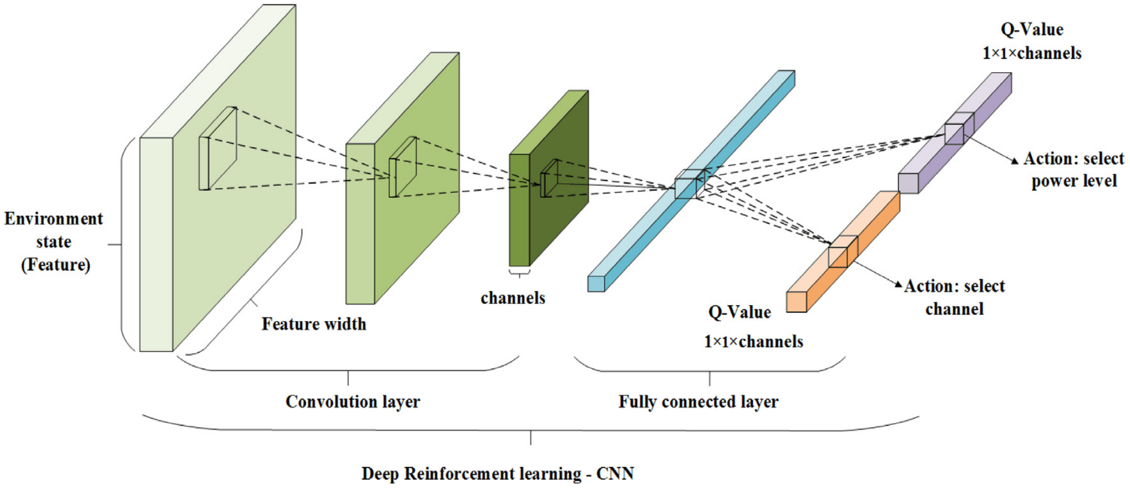


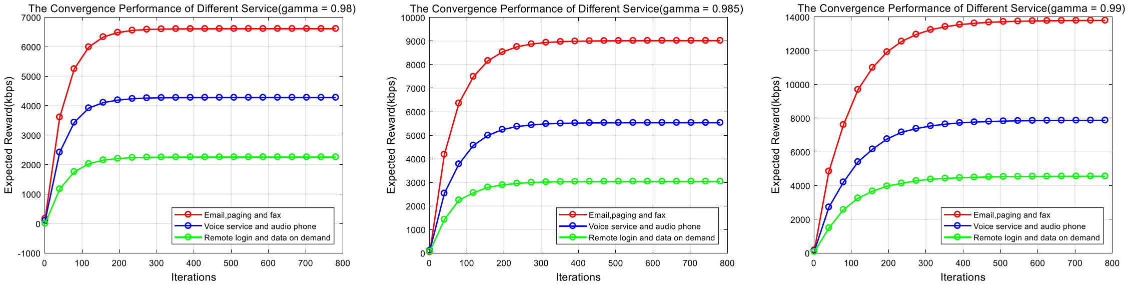
Fig. 3. Deep reinforcement learning – Convolution neural network.

by using the sampled data as shown in the algorithm of Algorithm 2. In addition, we consider the Q-learning algorithm as the baseline in this paper, where the number of states and the number of actions in practice make the solution space become very large. In this optimization problem, the number of state space is $N \cdot 2K$, and the number of action space is $K \cdot L$. Hence, the complexity of the Q-learning algorithm to search the optimal solution is $O\left((N \cdot 2K) \cdot (K \cdot L)\right) = O\left(2K^2NL\right)$. In our paper, we adopt the DQN algorithm to solve the optimization problem, which can obtain feasible lower-complexity solutions. This is because that DQN combines the Q-learning algorithm and neural network to process optimization problem. DQN can solve a large state space problem. It uses a neural network to approximate the Q-table and does not traverse the Q-table completely in each search. DQN is to update the calculation of network Q-value by training the quantitative data minibatch to make actions. Therefore, the computational complexity and computational time of DQN are less than $O\left(2K^2NL\right)$.

5. Simulation and analysis

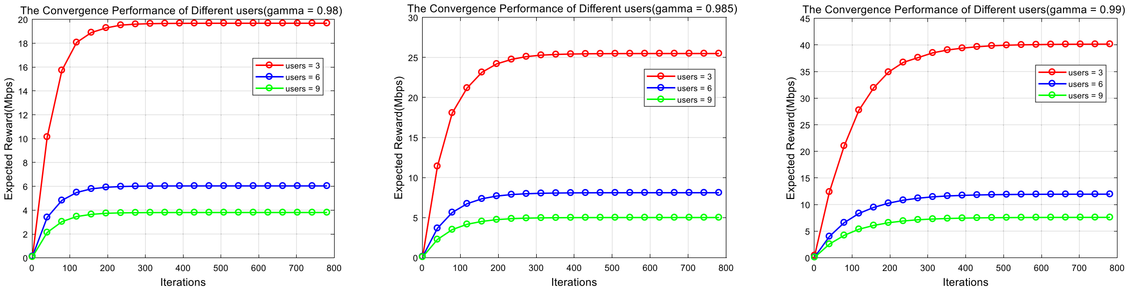
In this section, we present experiments to evaluate our proposed joint channel selection and power control method. Our experiments are based on an Ubuntu operating system (CPU Intel core i7-4790 3.6 GHz; memory 16 GB, GPU NVIDIA Quadro K2200, which contains 640 CUDA computing core units and 4 GB graphics memory).

In our experiment, the deep neural network is shown in Fig. 3. There are five layers, including three convolution layer and two fully connected layers. The final layer has two output values, and one represents the Q-value of channel selection; the other represents the Q-value of power selection. The main simulation parameters are presented in Table 1. The following results analyzed the convergence performance of different services and the convergence performance of various users under different discount factors, as well as the cumulative distribution function (CDF).



(a) gamma = 0.98 (b) gamma = 0.985 (c) gamma = 0.99

Fig. 4. The convergence performance of different service under different discount factors.



(a) (b) (c)

Fig. 5. The convergence performance of different users.

Table 1
The parameter of simulation.

Parameter	Value
Cell radius	500 m
D2D communication distance	50 m
D2D transmit power	[0–23 dB]
Resource block bandwidth	180 kHz
P_{max}	24 dB
Noise power/RB	−116 dB
path loss model between BS and users	15.3 + 37.6log(d(km))(dB)
path loss model between BS and users	28 + 40log10(d(km))(dB)
Macro BS antenna gain	17 dBi
User antenna gain	4 dBi
Learning rate	0.2
Discount factor	0.98, 0.985, 0.99
Exploration rate	Dynamic
Email, paging and fax	5 Kbps
Voice service and audio phone	30 Kbps
Remote login and data on demand	64 Kbps

Fig. 4 presents the convergence performance for three types of cellular users' services under different discount factors. Three various services, including Email, paging and fax, Voice service and audio phone, and Remote login and data on demand [28]. For the above three services, the minim bandwidth requirement of cellular users is 5 kbps, 30 kbps, and 60 kbps, respectively. Cellular users have different resource requirements for each service. The number of D2D users is 6, which reuse channels of one cellular user. We assume that the power level is [0, 4.8, 9.6, 14.4, 19.2, 24] (dB). In three different services, the agent learns the expected reward. The discount factors are gamma = 0.98, gamma = 0.985, gamma = 0.99. Fig. 4(a) shows that the number of iterations increases, the capacity can be gradually improved to a stable value. When the service is the email, paging, and fax, the convergence value is larger than the others. This is because the demand for the cellular user is smaller, so there are more reusable channel resources. As a result, the service of email, paging, and fax has a better convergence performance than others. Similarly, the same trend is seen in Fig. 4(b) and Fig. 4(c).

In addition, it can be seen that gamma = 0.98, gamma = 0.985, gamma = 0.99, the expected reward is increasing growth in a type of service in Fig. 4(a), (b), (c). This is because when gamma is set to be relatively large, transmitters will spend much more time identifying and reinforcing good actions. Hence, the value of the discount factor has an impact on our agent learning, where the larger discount factor results in a larger expected reward under the same services. With a larger discount factor, the system capacity stands a better chance to reach optimal convergence efficiency. The experiment proves that cellular communications and D2D communications can coexist, and RBs can be shared for their respective data transmissions. The proposed joint resources allocation and power selection method can maximize system capacity. During the learning process, the agent continuously updates the strategy to learn how to allocate resources and select power. In Fig. 4, the initially expected reward is low. This is because the agent was exploring the optimal strategy, and then the curve gradually rises and tends to stabilize. The optimal policy can be obtained faster through learning. The figure shows that DQN has a good convergence in the joint resource allocation and power selection, and the convergence time is short.

As shown in Fig. 5, we compare the expected reward of the different users under three discount factors. Fig. 5(a) depicts the expected rewards when the numbers of users are 3, 6, and 9 under gamma = 0.68. We can see that the expected reward is the maximum value under three users. The system's optimally expected reward drops as the number of users increase, which indicates the system performs better when there are fewer users. This is because of the interference by D2D links as a result of the number of D2D users. When the agent learns the strategy, more users will have more action and state space. Hence, the expected reward of a few users is higher than those with the many users in D2D communication. Furthermore, we can yet find the value of gamma has a major impact on the convergence performance in Fig. 5(a), (b), (c). However, the convergence speed in Fig. 5(c) is slower than in Fig. 5(b). This is because the agent learning process consisting of more iterations and larger gamma provides the agent with more efficient long-term observation to obtain better

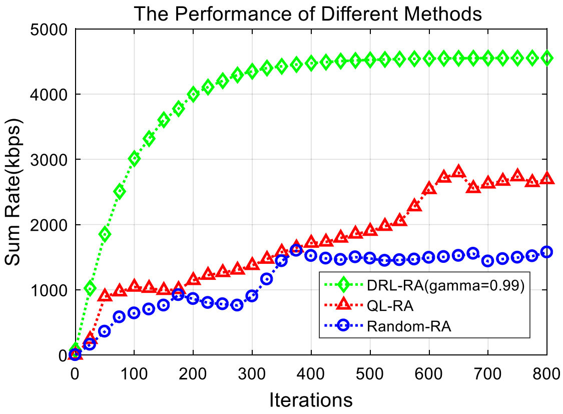


Fig. 6. The performance of different methods.

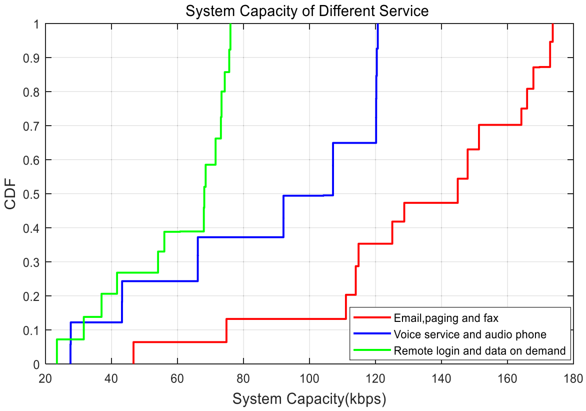


Fig. 7. The system capacity of different services.

learning efficiency. From the simulation results, each agent can learn how to satisfy the cellular communication constraint while minimizing D2D communications interference and maximizing the total system capacity.

As illustrated in Fig. 6, our proposed algorithm outperforms other existing D2D resource allocation algorithms under identical scenarios. As can be observed from the figure, we plot the system sum-rate with different methods, including random resource allocation (Random-RA), Q-learning resource allocation (QL-RA), and the proposed DRL resource allocation (DRL-RA). Fig. 6 shows that when the iteration increases, the rate performance of users is improved, and our proposed method is much better than other methods. Here, we set the gamma =0.99, and the service type is "Remote login and data on-demand". Compared to traditional resource allocation methods, cellular UEs achieve reasonable rate performance when interference is properly managed. D2D communication is more efficient.

Fig. 7 is showed that the maximization of system capacity as the CDF value. The values are plotted for all the iterations in the different services. It is illustrated in Fig. 7 that among the three services, the system performance is better when the service demand is smaller such as email, paging, and fax. The exploitation of our method in channel allocation and power control is efficient, increasing the system capacity.

6. Conclusion

This paper proposes a joint resource allocation and power control method with DRL in a sophisticated D2D communication. In our proposed learning method, all D2D pairs learn the strategies (channel selection and power control) by interacting with the environment. Furthermore, as the number of multiplexed channels increases, the performance of the algorithm does not decrease. Because of the different types of MCC service requirements, D2D users can select a number of channels to transmit services as soon as possible without affecting the normal communication of cellular users. Experimental results show that the learning process converges under different discount factors and users' number settings. The advantage of the proposed resource allocation and power selection method is to maximize the total system capacity according to different MCC services in the D2D network.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] Minh-Phung Bui, Nguyen-Son Vo, Sang Quang Nguyen, Quang-Nhat Tran, Social-Aware Caching and Resource Sharing Maximized Video Delivery Capacity in 5G Ultra-Dense Networks, ACM/Springer Mobile Networks & Applications, 2019, pp. 1–13.

[2] S. Mukherjee, C. Beard, A framework for ultra-reliable low latency mission-critical communication, in: 2017 Wireless Telecommunications Symposium (WTS), Chicago, IL, 2017, pp. 1–5.

[3] D. Wang, D. Chen, B. Song, N. Guizani, X. Yu, X. Du, From IoT to 5G I-IoT: The next generation IoT-based intelligent algorithms and 5G technologies, IEEE Commun. Mag. 56 (10) (2018) 114–120.

[4] L. Song, D. Niyato, Z. Han, E. Hossain, Game-theoretic resource allocation methods for device-to-device communication, IEEE Wirel. Commun. 21 (3) (2014) 136–144.

[5] R. Li, et al., Intelligent 5G: When cellular networks meet artificial intelligence, IEEE Wirel. Commun. 24 (5) (2017) 175–183.

[6] R.I. Ansari, C. Chrysostomou, S.A. Hassan, M. Guizani, S. Mumtaz, J. Rodriguez, J.J.P.C. Rodrigues, 5g d2d networks: Techniques, challenges, and future prospects, IEEE Systems J. PP (99) (2018) 1–15.

[7] G. Fodor, E. Dahlman, G. Mildh, S. Parkvall, N. Reider, G. Mikls, Z. Turnyi, Design aspects of network assisted device-to-device communications, IEEE Commun. Mag. 50 (3) (2012) 170–177.

[8] L. Song, D. Niyato, Z. Han, E. Hossain, Game-theoretic resource allocation methods for device-to-device communication, IEEE Wirel. Commun. 21 (3) (2014) 136–144.

[9] Y. Li, D. Jin, J. Yuan, Z. Han, Coalitional games for resource allocation in the device-to-device uplink underlaying cellular networks, IEEE Trans. Wirel. Commun. 13 (2014) 3965–3977.

[10] X. Du, M. Zhang, K. Nygard, S. Guizani, H.H. Chen, Self-healing sensor networks with distributed decision making, Int. J. Sens. Netw. 2 (5/6) (2007) 289–298.

[11] C.F. Silva, J.M.B. Silva Jr, T.F. Maciel, Radio resource management for device-to-device communications in long term evolution networks, in: Resource Allocation and MIMO for 4G and Beyond, Springer, 2014, pp. 105–156.

[12] N.C. Luong, D.T. Hoang, S. Gong, D. Niyato, P. Wang, Y.C. Liang, et al., Applications of deep reinforcement learning in communications and networking: a survey, 2018.

[13] Volodymyr Mnih, et al., Playing atari with deep reinforcement learning, Comput. Sci. (2013).

[14] T. Peng, Q. Lu, H. Wang, S. Xu, W. Wang, Interference avoidance mechanisms in the hybrid cellular and device-to-device systems, in: 2009 IEEE 20th International Symposium on Personal, Indoor and Mobile Radio Communications, Tokyo, 2009, pp. 617–621.

[15] M. Belleschi, G. Fodor, A. Abrardo, Performance analysis of a distributed resource allocation scheme for D2D communications, in: 2011 IEEE GLOBECOM Workshops (GC Wkshps), Houston, TX, 2011, pp. 358–362.

[16] X. Wang, F. Wang, Z. Song, Q. Zhao, Joint scheduling and resource allocation for device-to-device underlay communication, in: IEEE Wireless Communications and Networking Conference (WCNC), 2013.

[17] R. Zhang, L. Song, Z. Han, X. Cheng, B. Jiao, Distributed resource allocation for device-to-device communications underlaying cellular networks, in: IEEE International Conference on Communications (ICC), Budapest, 2013, pp. 1889–1893.

[18] S. Nie, Z. Fan, M. Zhao, X. Gu, L. Zhang, Q-learning based power control algorithm for D2D communication, in: IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC), Valencia, 2016, pp. 1-6.
[19] M. Zhao, Y. Wei, M. Song, G. Da, Power control for D2D communication using multi-agent reinforcement learning, in: IEEE/CIC International Conference on Communications in China (ICCC), Beijing, China, 2018, pp. 563–567.
[20] Y. Qiu, Z. Ji, Y. Zhu, G. Meng, G. Xie, Joint mode selection and power adaptation for D2D communication with reinforcement learning, in: 2018 15th International Symposium on Wireless Communication Systems (ISWCS), Lisbon, 2018, pp. 1–6.
[21] A. Moussaid, W. Jaafar, W. Ajib, H. Elbiaze, Deep reinforcement learning-based data transmission for D2D communications, in: 2018 14th International Conference on Wireless and Mobile Computing, Networking and Communications (WiMob), Limassol, 2018, pp. 1–7.
[22] S. Liu, X. Hu, W. Wang, Deep reinforcement learning based dynamic channel allocation algorithm in multibeam satellite systems, IEEE Access 6 (2018) 15 733–15 742.
[23] S. Wang, H. Liu, P.H. Gomes, B. Krishnamachari, Deep reinforcement learning for dynamic multichannel access in wireless networks, IEEE Trans. Cogn. Commun. Netw.
[24] Pieter Abbeel, John Schulman, Deep reinforcement learning through policy optimization, 2016, Tutorial at NIPS 2016.
[25] S. Gu, E. Holly, T. Lillicrap, S. Levine, Deep reinforcement learning for robotic manipulation with asynchronous off-policy updates, in: IEEE International Conference on Robotics and Automation (ICRA), 2017, pp. 3389–3396.
[26] Yue Zhang, Bin Song, Su Gao, et al., Monopolistic models for resource allocation: A probabilistic reinforcement learning approach, IEEE Access 6 (2018) 49721–49731.
[27] O. Naparstek, K. Cohen, Deep multi-user reinforcement learning for dynamic spectrum access in multichannel wireless networks, 2017, arXiv:1704.02613.
[28] M. Sanabani, et al., QoS provisioning for adaptive multimedia services in wireless/mobile cellular networks, in: International Rf & Microwave Conference IEEE, 2006.

Dan Wang received the B.Sc. degree on Electronic and Information Engineering from Northwest Normal University, Lanzhou, China, 2015. She is currently working towards her Ph.D. degree from Xidian University, Xi'an, China. Her research interests include machine learning, deep reinforcement learning, multi-agent reinforcement learning, game theory, Internet of Things, and Big data.

Hao Qin received the B.S., M.S., and Ph.D. degrees in communication and information systems from Xidian University, Xi'an, China, in 1996, 1999, and 2004, respectively. In 2004, he joined the School of Telecommunications Engineering, Xidian University, where he is currently an Associate Professor of communications and information systems. His research interests include wireless communications and satellite communications.

Bin Song received his B.S., M.S., and Ph.D. in communication and information systems from Xidian University, Xi'an, China in 1996, 1999, and 2002, respectively. He is currently a professor at the Xidian University, Xi'an, China. He has authored over 60 journal papers or conference papers and 30 patents. His research interests are in distributed video coding, compressed sensing based video coding, content-based image recognition and machine learning, deep reinforcement learning, Internet of Things, big data.

Ke Xu received the B.Sc. degree on Information Engineering and is currently working towards her Master degree in Xidian University, Xi'an, China. Her research interests include machine learning, deep reinforcement learning, wireless communication, satellite communication and Internet of Things.

Xiaojiang (James) Du is a tenured professor in the Department of Computer and Information Sciences at Temple University, Philadelphia, USA. Dr. Du received his B.S. and M.S. degree in electrical engineering from Tsinghua University, Beijing, China in 1996 and 1998, respectively. He received his M.S. and Ph.D. degree in electrical engineering from the University of Maryland College Park in 2002 and 2003, respectively. His research interests are security, wireless networks, and systems. He has authored over 300 journal and conference papers in these areas, as well as a book published by Springer. Dr. Du has been awarded more than $6 million US dollars research grants from the US National Science Foundation (NSF), Army Research Office, Air Force Research Lab, NASA, Qatar, the State of Pennsylvania, and Amazon. He won the best paper award at IEEE GLOBECOM 2014 and the best poster runner-up award at the ACM MobiHoc 2014. He serves on the editorial boards of three international journals. Dr. Du is a Senior Member of IEEE and a Life Member of ACM.

Mohsen Guizani (S'85–M'89–SM'99–F'09) received the B.S. (with distinction) and M.S. degrees in electrical engineering, the M.S. and Ph.D. degrees in computer engineering from Syracuse University, Syracuse, NY, USA, in 1984, 1986, 1987, and 1990, respectively. He is currently a Professor at the CSE Department in Qatar University, Qatar. Previously, he served in different academic and administrative positions at the University of Idaho, Western Michigan University, University of West Florida, University of Missouri-Kansas City, University of Colorado-Boulder, and Syracuse University. His research interests include wireless communications and mobile computing, computer networks, mobile cloud computing, security, and smart grid. He is currently the Editor-in-Chief of the IEEE Network Magazine, serves on the editorial boards of several international technical journals and the Founder and Editor-in-Chief of Wireless Communications and Mobile Computing journal (Wiley). He is the author of nine books and more than 500 publications in refereed journals and conferences. He guest edited a number of special issues in IEEE journals and magazines. He also served as a member, Chair, and General Chair of a number of international conferences. Throughout his career, he received three teaching awards and four research awards. He also received the 2017 IEEE Communications Society WTC Recognition Award as well as the 2018 AdHoc Technical Committee Recognition Award for his contribution to outstanding research in wireless communications and Ad-Hoc Sensor networks. He was the Chair of the IEEE Communications Society Wireless Technical Committee and the Chair of the TAOS Technical Committee. He served as the IEEE Computer Society Distinguished Speaker and is currently the IEEE ComSoc Distinguished Lecturer. He is a Fellow of IEEE and a Senior Member of ACM.