Hindawi
Wireless Communications and Mobile Computing
Volume 2021, Article ID 6946115, 10 pages
https://doi.org/10.1155/2021/6946115

Research Article

Reinforcement Learning for Interference Coordination Stackelberg Games in Heterogeneous Cellular Networks

Chen Sun, Shiyi Wu, and Bo Zhang

School of Software, Nanchang Hangkong University, 330063 Nanchang, China

Correspondence should be addressed to Chen Sun; sunchen@nchu.edu.cn

Received 3 October 2021; Accepted 5 November 2021; Published 14 December 2021

Academic Editor: Issa Elfergani

In future heterogeneous cellular networks with small cells, such as D2D and relay, interference coordination between macro cells and small cells should be addressed through effective resource allocation and power control. The two-step Stackelberg game is a widely used and feasible model for resource allocation and power control problem formulation. Both in the follower games for small cells and in the leader games for the macro cell, the cost parameters are a critical variable for the performance of Stackelberg game. Previous studies have failed to adequately address the optimization of cost parameters. This paper presents a reinforcement learning approach for effectively training cost parameters for better system performance. Furthermore, a two-stage pretraining plus ε-greedy algorithm is proposed to accelerate the convergence of reinforcement learning. The simulation results can demonstrate that compared with the three beachmarking algorithms, the proposed algorithm can enhance average throughput of all users and cellular users by up to 7% and 9.7%, respectively.

1. Introduction

In the existing and future cellular networks, i.e., 5G and beyond, small cells, such as device-to-device (D2D) and relay, are promised to constitute the heterogeneous networks to amplify system capacity and/or expand coverage [1]. D2D technology improves spectrum efficiency and reduces the power consumption of User Equipment (UE) by allowing two terminals to communicate with each other directly without Base Station (BS). Using in-band underlying D2D pairs can reuse the same spectrum with cellular UEs (CUEs) at the same time, which means more resources and also increased interference between CUEs and D2D pairs. Additionally, relay technology is applied to relay UEs (RUEs) that are far away from the BS. The RUEs can connect to the relay node (RN) by reusing the spectrum resources with the CUEs, and finally, the RN sends the information to the BS.

In recent literature on interference coordination in heterogeneous cellular networks (e.g., [2–15]), several approaches, i.e., heuristic algorithms, convex optimization, intelligent optimization, reinforcement learning, and game theory, are discussed. The authors in [2–4] proposed heuristic algorithms. In [2], by limiting the minimum Signal-to-Interference-plus-Noise-Ratio (SINR), each BS controls the distance between links reusing the same frequency by calculating the minimum restricted area where frequency reuse is not allowed. The authors in [3] divided the available spectrum into the inner region frequency and the outer region frequency. By limiting the reusable region, a cell sectorization method was proposed to solve the resource allocation and power control problems. In [4], user association of D2D communication is formulated based on maximizing received power. Then, a sequential max search-based algorithm was developed to solve resource allocation problem. However, heuristic algorithms with low complexity can hardly reach the optimal solution.

In addition, some researchers proposed convex optimization algorithms to deal with interference coordination problems. An optimization problem of network throughput for D2D underlaying cellular networks was formulated in [5] and solved by a convex function while ensuring the quality-of-service (QoS) constraints. The authors in literature [6] studied the power control problem in the D2D heterogeneous cellular networks based on partial frequency reuse

and proposed a dynamic power control scheme based on the basis of partial power control. With the goal of maximizing system capacity, the objective function of power control is established, the nonconvex function is transformed into a convex function, and the improved Lagrangian dual decomposition method is introduced to reduce the algorithm complexity. The interference coordination algorithms based on convex optimization can approximate optimal solutions by establishing optimization models and solving them via convex optimization algorithms. However, the idealized optimization model and the NP-hard problem in the solving process make these algorithms impracticable.

Some researchers also tried to use intelligent optimization algorithms to solve the NP-hard problem in optimization. Authors in [7] proposed a joint resource allocation and user matching scheme based on genetic algorithm to minimize interference and maximize spectrum efficiency, which used a limited number of resource blocks to serve a large number of users. In [8], a simple particle swarm optimization algorithm for resource allocation was proposed to improve the system capacity performance. The simulation shows that with 10 particles, the proposed scheme can obtain suboptimum performance with quick convergence.

The intelligent optimization algorithms still model the optimization target as an ideal mathematical model, which cannot perfectly describe the real scenarios. Nevertheless, Reinforcement learning (RL) has attracted more attention to solve the interference coordination in a new way. RL is a machine learning paradigm in which agents measure the quality of their actions through reward in an episode and determine the actions under their states to maximize long-term returns. For example, a scheme using Q learning was proposed in [9] to allow a small cell to learn the appropriate transmitting power for less interference between the small cell and BS. The authors in [10] investigated multiagent Q learning for D2D user (DUE) selecting frequency resources in multilayer D2D heterogeneous networks. The issues to be addressed in above two methods are shorter learning process and avoiding local optimal solutions.

Game theory is also expected to solve the interference coordination problem in heterogeneous cellular networks. The authors in [11] formulated the cochannel interference coordination problem between the D2D link, the micro cell link, and the macro cell link into a potential game problem. The players' strategies are updated iteratively by message passing to achieve the Nash equilibrium. The relay selection problem was modeled as a noncooperative game in [12], which was used to improve spectral and energy efficiency. The authors in [13] utilized the Nash bargaining model to deal with the frequency reuse problem between D2D and the macro cell. A bargaining factor is introduced for performance optimization, which is solved by a maximum weight maximum stream algorithm and the Lagrange Multiplier method. A cooperative game (CG) theory-based resource allocation in cluster-based D2D communication network was discussed in [14]. In CG-D2D, a utility function for allocating the resources between the D2D pairs and the cluster was proposed. Stackelberg game with pricing was introduced in [15]. BS evaluates the QoS of the CUEs and decides a price to be paid by the D2D pairs for reusing the resource of a CUE. The purpose is to allocate channel and power levels to D2D pairs and to optimize their transmission rates.

In a Stackelberg game-based interference coordination model of heterogeneous networks, the cost parameter is critical for leaders and followers, because it can affect the results of the two-stage game. Therefore, many studies proposed several cost parameter setting methods. Authors in [16] presented an artificial adjustment method of cost parameter to guarantee both sides of the reaction function the same order of magnitude, and the value of the reaction function is within a reasonable range. The cost parameters can also be determined by the channel state of the follower [17] or by the channel state of the leader [18]. In [19], an iterative strategy was proposed to improve the cost parameter, updating the global cost parameter in a fixed number of steps according to the result of the game. The authors in [20] reckon that the cost parameters need to be set for each D2D link in each channel. In the current researches on the interference coordination algorithms based on Stackelberg game, most of the cost parameters are fixed or self-iteratively updated. At present, there are few explorations on the advanced setting of cost parameters, and thus, the effectiveness of the Stackelberg game is difficult to be guaranteed.

Centralized RL-based interference coordination methods, such as in [9], require accurate channel state information from small cells, which brings heavy burden to networks. Meanwhile, the RL-based distributed interference coordination methods, such as in [10], are executed by each UE, which can consume excessive computing resources. Stackelberg game-based interference coordination methods [15–20] allow distributed follower games in each UE of small cells; however, important cost parameters for follower games are not effectively optimized so far. Therefore, this paper focuses on applying reinforcement learning to the Stackelberg game model to address interference coordination problem in D2D and relay heterogeneous cellular networks. The main contributions of this paper can be summarized as follows.

(1) An interference coordination architecture containing a reinforcement learning model and a Stackelberg game model is first introduced to model the interference coordination problem in D2D and relay heterogeneous cellular networks. This architecture allows distributed interference coordination in D2D pairs and RUEs based on local channel state information and centralized reinforcement learning in BS to improve the performance of interference coordination

(2) A reinforcement learning model is proposed to optimize cost parameters for the Stackelberg game model. The proposed Q-learning model defines the current resource reuse situation with CUEs as the state space, the cost parameters as the action space, and the utility changes of all links as the reward

(3) A two-stage pretraining plus ε-greedy algorithm is proposed for a better update of the Q table. In the pretraining stage, the agent randomly picks the

actions for dozens of episodes, and in the second stage, the agent changes between random action choice and best action choice according to the probability ε. This algorithm is aimed at faster convergence and better optimization of the Q table at the same time

(4) The proposed interference coordination algorithm using two-stage Q learning for the Stackelberg game model, called RL game for short, proves its effectiveness in network throughput by comparing with benchmark algorithms in simulation. Better settings of probability ε and pretraining episode are also found through simulation results

The rest of this paper is organized as follows. Section 2 describes the system model and problem formulation of a heterogeneous cell. Section 3 explains the Stackelberg game for interference coordination. A reinforcement learning algorithm for cost parameters in Stackelberg game is proposed in Section 4. In Section 5, the performance evaluation results and their discussions are analyzed. Section 6 concludes the paper.

2. System Model

We consider a single heterogeneous cell which is shown in Figure 1 [21]. A number of CUE and DUEs are randomly distributed within the coverage area of the BS. In the 3GPP standards, resource reuse between underlying D2D links and CUE uplinks is considered in priority [21]. In the uplink, a set of CUE $M = \{1, \cdots, M\}$ and RUE $Q = \{1, \cdots, Q\}$ in a macro cell communicates with the BS and RN, respectively. There exists a set of N DUE pairs that comprise a D2D transmitter (DTx) and a D2D receiver (DRx). A D2D pair is able to communicate each other by reusing a unit of CUE uplink resource. In-band RUE-RN links and some CUE-BS links also use the same RBs. The RN-BS link shares the RBs with the CUE-BS links orthogonally in the backhaul subframes to avoid self-interference in the RN. Thus, the CUE-BS link may suffer from the interference from the RUE and the DTx. Similarly, the CUEs may also be the interference sources to DRx of the D2D links and RNs of the RUE-RN links.

Therefore, the SINR of a CUE on a Physical Resource Blocks (PRBs) in the uplink is defined as follows:

$$\mathrm{SINR}_{m,k} = \frac{\alpha_{m,k} P_{m,k} \mathrm{PL}_m}{N_{0,k} + \gamma_{n,k} P_{n,k} \mathrm{PL}_n + \beta_{q,k} P_{q,k} \mathrm{PL}_q}, \quad (1)$$

where $P_{m,k}$ denotes the transmitting power of CUE m using PRB k, $P_{n,k}$ represents the DTx transmitting power of the D2D pair n using PRB k, and $P_{q,k}$ indicates the transmitting power of RUE q using PRB k. Moreover, PL_m represents path loss of the link between CUE m and BS. PL_n denotes path loss of the link between DTx n and BS. PL_q indicates path loss of the link between RUE q and BS. In addition, $N_{0,k}$ means Gaussian white noise; $\alpha_{m,k}$, $\beta_{q,k}$, and $\gamma_{n,k}$ are binary variables, where 0 shows that PRB k is not used and 1 means that PRB k is used.

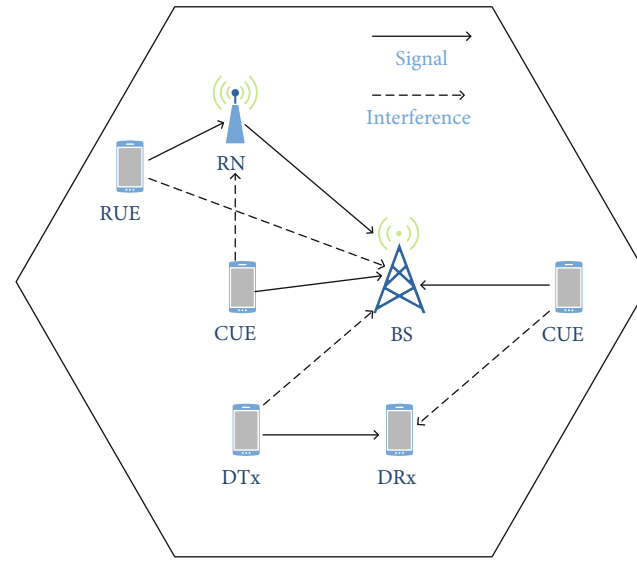

Figure 1: Single heterogeneous cell model.

When D2D performs uplink communication, the SINR of a certain DRx on a certain PRB can be expressed as follows:

$$\mathrm{SINR}_{n,k} = \frac{\gamma_{n,k} P_{n,k} \mathrm{PL}_n}{N_{0,k} + \alpha_{m,k} P_{m,k} \mathrm{PL}_{m,n} + \beta_{q,k} P_{q,k} \mathrm{PL}_{q,n}}, \quad (2)$$

where $\mathrm{PL}_{m,n}$ represents the path loss of the link between CUE m and DUE n and $\mathrm{PL}_{q,n}$ denotes the path loss between DRx n and RUE q.

Similarly, when the RUE performs uplink communication on the access link, the SINR of a certain RUE on a certain PRB can be written as follows:

$$\mathrm{SINR}_{q,k} = \frac{\beta_{q,k} P_{q,k} \mathrm{PL}_q}{N_{0,k} + \alpha_{m,k} P_{m,k} \mathrm{PL}_{m,q} + \gamma_{n,k} P_{n,k} \mathrm{PL}_{n,q}}, \quad (3)$$

where $\mathrm{PL}_{m,q}$ represents the path loss of the link between CUE m and RUE q and $\mathrm{PL}_{n,q}$ denotes the path loss between DTx n and RUE q.

In summary, the total data transmission rate of the cellular communication system on PRB k with a bandwidth B can be expressed as follows:

$$R_k = B \sum_{M,N,Q} \begin{cases} \log_2(1 + \mathrm{SINR}_{m,k}), \\ +\log_2(1 + \mathrm{SINR}_{n,k}) + \log_2\left(1 + \mathrm{SINR}_{q,k}\right). \end{cases} \quad (4)$$

The purpose of interference coordination in this paper is to maximize the system throughput for all links on each PRB, and thus, its objective function is defined as follows:

$$\max R_K = \max R_K\left(\alpha_{m,k}, \beta_{q,k}, \gamma_{n,k}, P_{q,k}, P_{n,k}\right), \quad (5)$$

$$\text{Subject to}: \begin{cases} \sum_{M} \alpha_{m,k}, \forall k \in K, \\ \sum_{Q} \beta_{q,k} + \sum_{N} \gamma_{n,k} \le 1, \forall k \in K, \\ P_{q,\min} \le P_{q,k} \le P_{q,\max}, \\ P_{n,\min} \le P_{n,k} \le P_{n,\max}, \end{cases} \tag{6}$$

where $P_{q,\min}$ and $P_{q,\max}$ represent the minimum and maximum transmitting power, respectively, allowed by the RUE q; $P_{n,\min}$ and $P_{n,\max}$ represent the minimum and maximum transmitting power, respectively, allowed by the DTx n.

3. Stackelberg Game-Based Interference Coordination

In order to execute distributed interference coordination, this study uses the Stackelberg two-step game model to allow DUEs and RUEs to control their transmission power and determine the resource allocation for both small cell UEs and CUEs. DUEs and RUEs consider follower games with the information from the leaders, and CUEs compete each other in a leader game in BS. Our goal is to minimize interference on macro cells while ensuring the basic performance of small cells. Therefore, using Stackelberg's two-stage game model to model interference coordination is in line with realistic needs.

3.1. Leader Utility Function. In the leader game, the utility function consists of CUE m, DTx n, and RUE q in PRB k, which can be expressed as follows:

$$U_{m,n,q,k} = B \log_2 \left(1 + \frac{P_{m,k}\mathrm{PL}_m}{N_{0,k} + P_{n,k}\mathrm{PL}_n}\right) + \lambda_m \left(\gamma_{n,k} P_{n,k}\mathrm{PL}_n + \beta_{q,k} P_{q,k}\mathrm{PL}_q\right), \tag{7}$$

where λ_m denotes the cost parameter provided by each CUE m for any other underlay links. Similarly, reusing parameters $\gamma_{n,k}$ and $\beta_{q,k}$ cannot be 1 at the same time. In other words, the DTx n and RUE q cannot reuse the PRB of the same CUE simultaneously.

3.2. Follower Utility Function. In the follower game, the payment utility functions of DTx n and RUE q in PRB k can be expressed as follows:

$$V_{m,n,k} = B \log_2 \left(1 + \frac{P_{n,k}\mathrm{PL}_n}{N_{0,k} + P_{m,k}\mathrm{PL}_m}\right) - \lambda_m (P_{n,k}\mathrm{PL}_n), \tag{8}$$

$$V_{m,q,k} = B \log_2 \left(1 + \frac{P_{q,k}\mathrm{PL}_q}{N_{0,k} + P_{m,k}\mathrm{PL}_m}\right) - \lambda_m \left(P_{q,k}\mathrm{PL}_q\right). \tag{9}$$

In the model proposed in this paper, neither resource reuse between DTx and RUE nor resource reuse between different D2D pairs is considered.

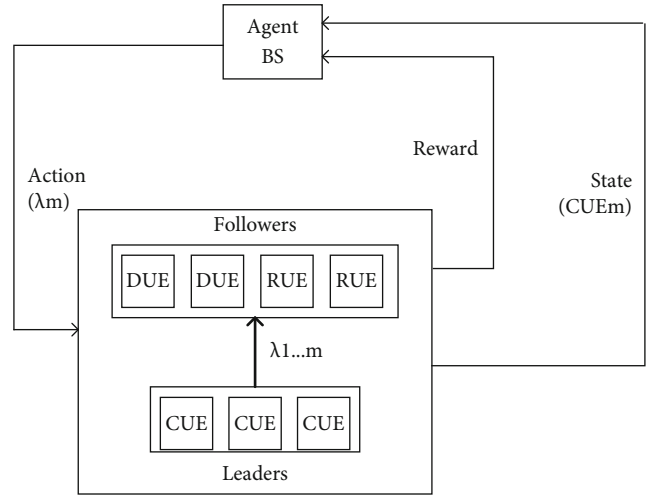

Figure 2: Interference coordination architecture model.

3.3. Power Control in Small Cells. In small cells, the transmitting power is decided by letting the partial derivative of the follower utility function equal 0. Taking the transmitting power of DTx n on PRB k as an example, find the partial derivative function of $P_{n,k}$ in the above Equation (8), and set it to 0. Thus, the DTx transmitting power $P_{n,k}^{m^*}$ for maximizing the function (8) can be derived for different cost parameter λ_m.

$$\frac{\partial V_{m,n,k}}{\partial P_{n,k}} = \frac{B}{\ln 2} * \frac{\mathrm{PL}_n}{N_{0,k} + P_{m,k}\mathrm{PL}_m + P_{n,k}\mathrm{PL}_n} - \lambda_m \mathrm{PL}_n = 0, \tag{10}$$

$$P_{n,k}^{m^*} = \frac{B}{\lambda_m \mathrm{PL}_n \ln 2} - \frac{N_{0,k} + P_{m,k}\mathrm{PL}_m}{\mathrm{PL}_n}. \tag{11}$$

The partial derivative function of $P_{q,k}$ in the above Equation (9) can be obtained alike, and set it to 0. The transmitter power of RUE for maximizing the function (9) can be calculated:

$$P_{q,k}^{m^*} = \frac{B}{\lambda_m \mathrm{PL}_q \ln 2} - \frac{N_{0,k} + P_{m,k}\mathrm{PL}_m}{\mathrm{PL}_q}. \tag{12}$$

After obtaining $P_{n,k}^{m^*}$ and $P_{q,k}^{m^*}$, they need to be limited to the maximum and minimum transmitting power.

$$P_{n,k}^{m^*} = \max\left(\min\left(P_{n,k}^{m^*}, P_{n,\max}\right), P_{n,\min}\right), \tag{13}$$

$$P_{q,k}^{m^*} = \max\left(\min\left(P_{q,k}^{m^*}, P_{q,\max}\right), P_{q,\min}\right). \tag{14}$$

3.4. Resource Allocation in the Macro Cell BS. The macro cell BS takes the $P_{n,k}^{m^*}$ and $P_{q,k}^{m^*}$ provided by each DTx and RUE in Equation (7), and the matrix size of the utility function $M * (N + Q)$ can be obtained when different D2D pairs, RUEs, and CUEs reuse the same PRB. In order to achieve the optimization of interference coordination in formula (5), the

Initializes S=S0 and Q(s,a)=0.
Sets the values α =0.1
Loop % start an update episode t
If t<50 % In the pre-training stage
The agent selects an action randomly from the action set;
Update Q value according to
$Q_t(s_t, a_t) = (1 - \alpha)Q_{t-1}(s_t, a_t) +$
$\alpha[r_t(s_t, a_t)]$
Else
Generate a random number num;
If num<ε %'exploration' is selected
The agent selects an action which can get largest Q value from action set;
Else% 'exploitation' is selected
Agent select an action randomly from the action set;
Update Q value according to
$Q_t(s_t, a_t) = (1 - \alpha)Q_{t-1}(s_t, a_t) +$
$\alpha[r_t(s_t, a_t)]$
Update state
End if
End if
Until t is terminal
End Loop

Algorithm 1: Two-stage Q-learning algorithm.

objective function of resource allocation can be rewritten as follows:

$$\sum_{M} \alpha_{m,k} \left(\sum_{N} \gamma_{n,k} U_{m,n,q,k} + \sum_{Q} \beta_{q,k} U_{m,n,q,k} \right). \quad (15)$$

To meet the optimization requirements in formula (15), a resource allocation algorithm based on the Hungarian algorithm is proposed. The specific steps of the algorithm are given below:

Step 1. Traverse all columns in the $U_{m,n,q,k}$ matrix and find the maximum value of $U_{m,n,q,k}$ in $N + Q$ columns and all the corresponding rows m.

Step 2. Judge whether m in different columns are different. If so, jump to Step 4.

Step 3. Find the corresponding columns n or q for all nonrepeated rows m, and remove them from the $U_{m,n,q,k}$ matrix. Jump to Step 1 (note: since $N + Q$ should be less than M, the matrix must not be empty).

Step 4. Output all (m, n) and (m, q) correspondences, and use the round-robin algorithm to fairly allocate all resources to CUE m, D2D pair n, and RUE q according to the selected (m, n) and (m, q).

4. Reinforcement Learning of Cost Parameter in Stackelberg Game Model

The cost parameter is the key factor in the Stackelberg game model, because it determines the transmitting power of DTx and RUE and then affects the resource reuse between D2D/RUE and CUE. However, it is difficult to set a suitable parameter for each CUE m. The appropriate cost parameter of CUE m in the follower games should keep the transmitting power of D2D/RUE within a reasonable range to realize power control and should improve overall system performance in the leader game. Hence, this section proposes a Q-learning method of cost parameter, which is aimed at determining an appropriate cost parameter for each CUE m. An interference coordination architecture combining the reinforcement learning model and Stackelberg game model is proposed as shown in Figure 2.

4.1. Reinforcement Learning Model. The BS performs a learning process for all D2D pairs and RUEs in each slot t to update the triplet variables, which are state s, action a, and reward r. Three basic elements necessary for the reinforcement learning model are defined as follows.

State. It is the current situation of resource reuse with CUEs, and if a D2D/RUE is currently reusing the resource of CUE m, the state denotes m.

Action. A set of cost parameters is defined as the action space. Note that the value range of λ_m is assumed from 138 dB to 197 dB with the value interval of 1 dB in this study.

Reward. The reward function reflects the learning goal, expressed as the total throughput of D2D/RUE and CUE m on PRB k with cochannel interference minus the

Table 1: Simulation parameters.

Parameters	Values
Cell radius	500 m
Number of users (CUE+RUE)	30
Distance between RN and BS	$(0.35 \sim 0.5) * 500$ m
Number of PRB	50
Path loss model of cellular link	3GPP TR 36.814 V9 A.2.1.1.2
Path loss model of relay link	Heterogeneous system simulation baseline parameters [21]
Path loss model of D2D link	3GPP TR 36.843 V12 A.2.1.2 channel models [22]
UE transmitting power	-40 to 24 dBm
D2D link interference threshold	-105 dBm
Noise power spectral density	-174 dBm/Hz

throughput of CUE m without interference caused by resource reuse. It can be described as follows:

$$r(s,a) = \begin{cases} R_{m,k} + R_{n,k} - I_{m,k}, & \text{if } \gamma_{n,k} = 1, \beta_{q,k} = 0, \\ R_{m,k} + R_{q,k} - I_{m,k}, & \text{if } \gamma_{n,k} = 0, \beta_{q,k} = 1, \\ 0, & \text{otherwise}, \end{cases} \quad (16)$$

where $R_{m,k}$ denotes the throughput of CUE m using the PRB k, $R_{n,k}$ represents the throughput of the D2D pair n reusing the PRB k, and $R_{q,k}$ denotes the throughput of RUE q when it reuses the PRB k. Finally, $I_{m,k}$ indicates the throughput of CUE m without cochannel interference from DUE or RUE.

4.2. A Two-Stage Q-Learning Algorithm. In a Q-learning method, Q values indicate the expected rewards in all states and actions, which are saved and updated in a Q table. Based on link information and throughput feedback from D2D pairs and RUEs, updating the Q values in an episode t is carried out by BS, which can be expressed as follows.

$$Q_t(s_t, a_t) = (1-\alpha)Q_{t-1}(s_t, a_t) + \alpha[r_t(s_t, a_t) + \gamma \max Q_{t+1}(s_{t+1}, a_{t+1})], \quad (17)$$

where α is the learning rate that represents the update rate of the Q values and 0.1 in this study; γ is the discount rate, which denotes the impact of the final reward on the intermediate state, and needs to be 0 in this study.

To update the Q values, an execution strategy based on the Q table is required. As a feasible strategy, an ε-greedy algorithm [22] in an update episode t chooses an action at random with a probability ε, called "exploration," otherwise chooses an action with the highest Q values according to the current state, called "exploitation." The smaller exploration probability ε not only means more update episodes t to the convergence, but also better optimization.

For faster convergence and better optimization, a two-stage "pretraining plus ε-greedy" algorithm is proposed, which divides the update episodes into two stages, the pretraining stage and the ε-greedy stage. In the first pretraining stage containing several episodes, the agent always chooses a random action regardless of the state. In the following ε-greedy stage, the traditional ε-greedy algorithm with large ε is carried out based on the pretrained Q table, which is expected to accelerate the convergence and achieve better optimization.

The proposed two-stage Q-learning algorithm is summarized in Algorithm 1.

5. Simulation

In this section, a single-sector system-level simulation to compare the performance of the proposed algorithm and three benchmarks is described. In this simulation, UE of different communication modes are distributed in a sector randomly, including 30 CUEs or RUEs and several DUEs. The simulation parameters are listed in Table 1.

Three benchmark algorithms are considered for comparison. The first is the round-robin-based resource allocation algorithm, which is abbreviated "RR." In the RR algorithm, RUEs and D2D pairs reuse CUE link resources randomly, regardless of their channel information and transmitting power. The second is labeled "greedy." In the greedy optimization algorithm without power control, the sum of the throughput of the CUE and RUE or D2D pair on each PRB is optimized without considering the interference between them. Moreover, the maximum transmission power of DTx and RUEs is assumed. The last is the interference coordination algorithm based on the Stackelberg game with fixed cost parameters, which is abbreviated "FC game." In the FC game, the CUEs will select two fixed cost parameters which are on the 33rd and 67th percentiles of the cost parameter value range.

The RL game algorithm proposed in this paper will compare performance indicators with the above three benchmark algorithms, such as DUE average throughput, CUE average throughput, and average throughput of all users.

Figures 3–5 depict the cumulative distribution functions (CDF) of the CUE, DUE, and RUE throughput, respectively, with different interference coordination algorithms. Compared to the RR and greedy algorithms, the RL game algorithm proposed in this study reaches greater CUE performance, which means more CUEs have throughput of over 1100 Kbps as shown in Figure 3. It can be observed that the DUE performance using the proposed RL game

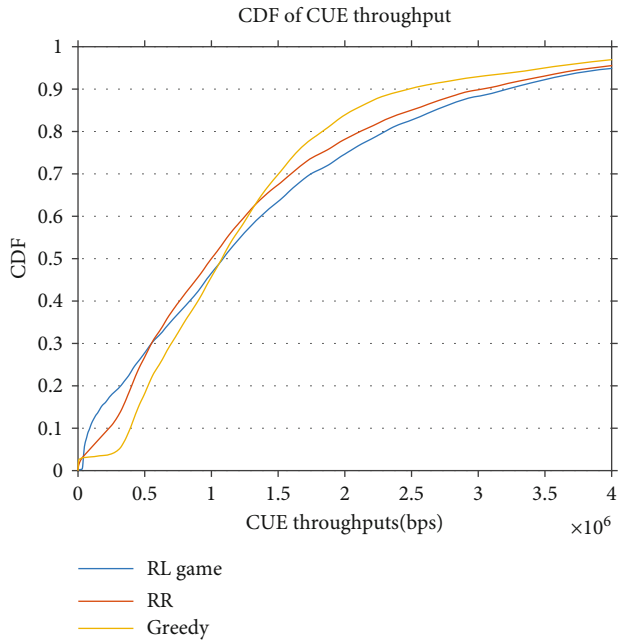

Figure 3: CDF of CUE throughputs.

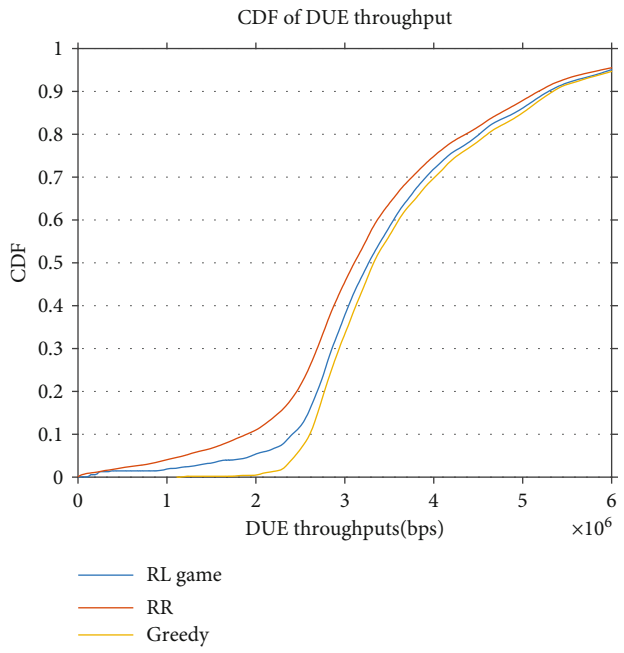

Figure 4: CDF of DUE throughputs.

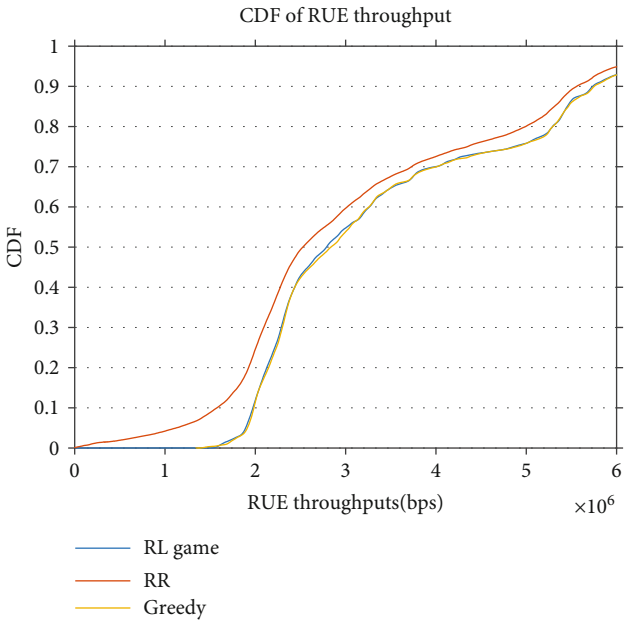

Figure 5: CDF of RUE throughputs.

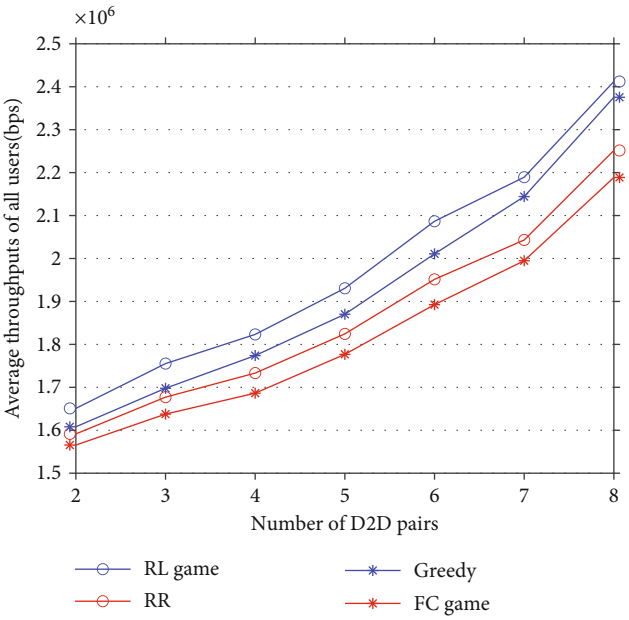

Figure 6: Average throughput of all users with different number of D2D pairs.

algorithm is intermediate between that using the greedy algorithm and the RR algorithm, while the RUE performance using the RL game algorithm and the greedy algorithm are almost the same and both significantly better than that using the RR algorithm.

Figures 6 shows that the average throughput of all users grows up along with the increasing number of D2D pairs. Since more D2D pairs reuse more RBs, the average throughput of all users using the proposed RL game algorithm increases from 1.65 to 2.41 Mbps with the numbers of D2D pairs from 2 to 8. Figure 6 also demonstrates the top rank of the proposed RL game algorithm in the average throughput of all users with the benchmark algorithms.

From Figure 7, it can be noted that the average throughput of D2D pairs also has strong positive correlation to the number of D2D pairs. When the D2D pairs increase from 2 to 8, the DUE throughput increases significantly from 3.22 to 4.2 Mbps using the RL game algorithm. We can speculate that higher diversity of D2D pairs is exploited with more D2D pairs reusing the resources. However, the greedy algorithm has better performance because the D2D pairs will not control their power to reduce their interference to CUEs.

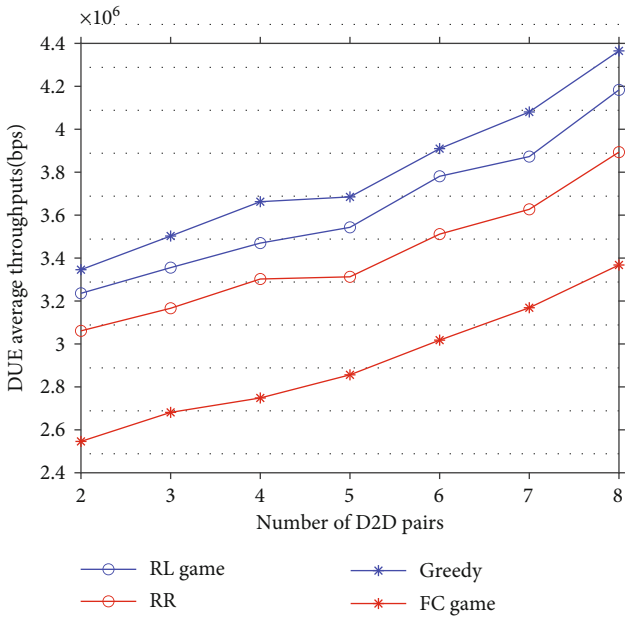

Figure 7: DUE throughput with different number of D2D pairs.

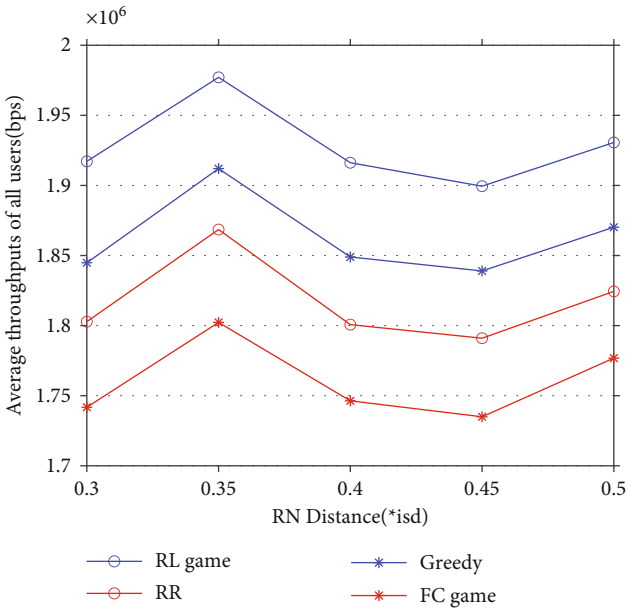

Figure 8: Average throughput of all users in different RN distances from BS.

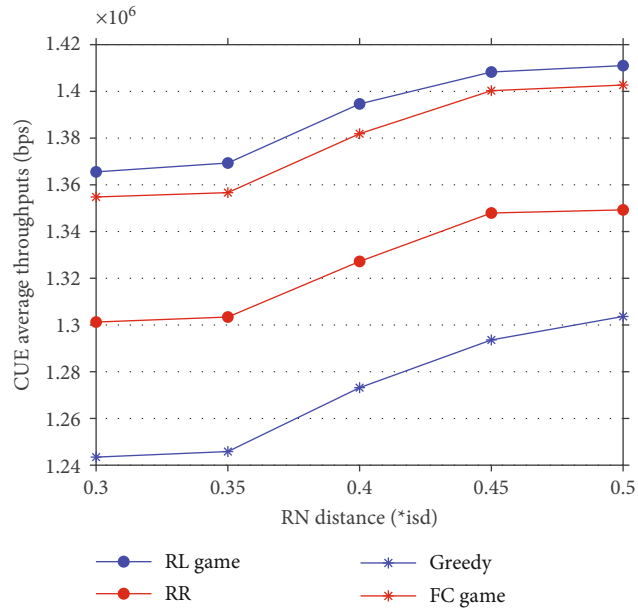

Figure 9: CUE throughput in different RN distances from BS.

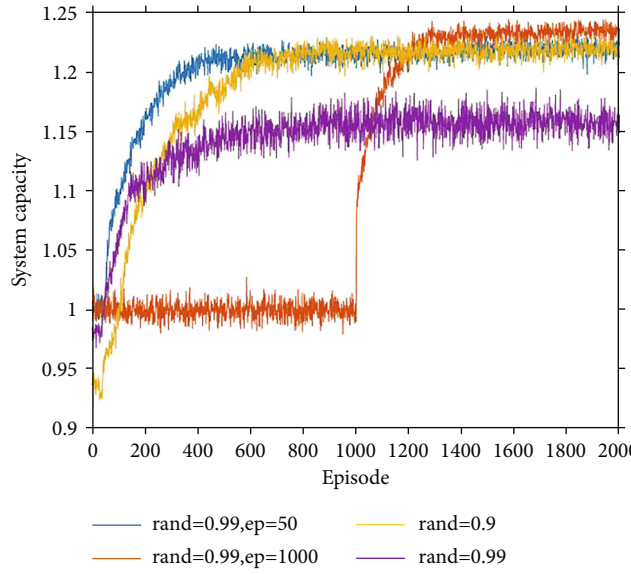

Figure 10: Convergence with different exploration probabilities ε and pretraining episodes.

Figure 8 shows that when the distance between BS and RN increases, the average throughput of all users varies greatly using all algorithms. Note that "isd" is the abbreviation of inter site distance. As is seen in Figure 9, when RN moves away from BS, the CUE throughputs using different algorithms increase. The reason is that the CUE with poor signal quality in the cell edge area can be improved as the RN approaches the edge. However, when the RN is close to the BS, larger interference between CUEs and RUEs will result in a decreasing average throughput of all users from $0.3 * \text{isd}$ to $0.35 * \text{isd}$ of the RN distance from the BS.

Figure 10 shows the convergence of the normalized system capacity in an episode with different random exploration probabilities ε and pretraining episodes. The system capacity with random action is selected as normalization. It can be observed that after about 600 episodes, the system capacity with 99% random exploration probability and 1% exploitation probability reaches the convergence, which is slower than that with 90% random exploration probability and 10% exploitation probability. After the convergence, the system capacity with 99% random exploration probability is larger than that with 90% random exploration

probability. This implies that a larger random exploration probability can obtain more optimization solutions with slower convergence. Pretraining the agents with random exploration actions is expected to accelerate the convergence, which can be validated by the system capacity with 99% random actions after 50 and 1000 pretraining episodes. With 1000 pretraining episodes, the system capacity reaches convergence after about 200 episodes, while about 350 episodes are taken using 50 pretraining episodes. However, the system capacity with 50 pretraining episodes can converge to the maximum in the least episodes between these methods, which is suggested in the proposed algorithm.

6. Conclusions

This paper investigates an interference coordination architecture in D2D and relay heterogeneous cellular networks that combines reinforcement learning and the Stackelberg game. A reinforcement learning model for cost parameters in Stackelberg games is proposed along with a two-stage Q -learning algorithm. The simulation results prove the network throughput advantages using the proposed algorithm and the benchmark algorithms. The better episode number in the pretraining stage and the better exploration probability ε are also investigated through the simulation.

Data Availability

The simulation data used to support the findings of this study are available from the corresponding author upon request.

Conflicts of Interest

The authors declare that they have no conflicts of interest.

References

[1] 3GPP, "Requirements for further advancements for E-UTRA (LTE-Advanced)," *TR36.913*, vol. 16.0.0, 2020.

[2] K. Wen, Y. Chen, and Y. Hu, "A resource allocation method for D2D and small cellular users in HetNet," in *2017 3rd IEEE International Conference on Computer and Communications (ICCC)*, Chengdu, China, 2017IEEE.

[3] D. D. Ningombam and S. Shin, "Radio resource allocation and power control scheme to mitigate interference in device-to-device communications underlaying LTE-A uplink cellular networks," in *2017 International Conference on Information and Communication Technology Convergence (ICTC)*, Jeju, Korea (South), 2017.

[4] A. Algedir and H. H. Refai, "A user association and energy efficiency analysis of D2D communication under HetNets," in *2018 14th International Wireless Communications & Mobile Computing Conference (IWCMC)*, Limassol, Cyprus, 2018.

[5] H. H. Esmat, M. M. Elmesalawy, and I. I. Ibrahim, "Uplink resource allocation and power control for D2D communications underlaying multi-cell mobile networks," *AEU-International Journal of Electronics and Communications*, vol. 93, pp. 163–171, 2018.

[6] F. Jiang, B. C. Wang, C. Y. Sun, Y. Liu, and X. Wang, "Resource allocation and dynamic power control for D2D communication underlaying uplink multi-cell networks," *Wireless Networks*, vol. 24, no. 2, pp. 549–563, 2018.

[7] H. Takshi, G. Dogan, and H. Arslan, "Joint optimization of device-to-device resource and power allocation based on genetic algorithm," *IEEE Access*, vol. 6, pp. 21173–21183, 2018.

[8] Y.-F. Huang, T.-H. Tan, B.-A. Chen, S.-H. Liu, and Y.-F. Chen, "Performance of resource allocation in device-to-device communication systems based on particle swarm optimization," in *2017 IEEE International Conference on Systems, Man and Cybernetics (SMC)*, Banff, AB, Canada, 2017IEEE.

[9] L. Xiao, H. Zhang, Y. Xiao et al., "Reinforcement learning based downlink interference control for ultra-dense small cells," *IEEE Transactions on Wireless Communications*, vol. 19, no. 1, pp. 423–434, 2019.

[10] K. Zia, N. Javed, M. N. Sial, S. Ahmed, A. A. Pirzada, and F. Pervez, "A distributed multi-agent RL based autonomous spectrum allocation scheme in D2D enabled multi-tier HetNets," *IEEE Access*, vol. 7, pp. 6733–6745, 2019.

[11] D. D. Penda, A. Abrardo, M. Moretti, and M. Johansson, "Distributed channel allocation for D2D-enabled 5G networks using potential games," *IEEE Access*, vol. 7, pp. 11195–11208, 2019.

[12] S. Selmi and R. Bouallegue, "Joint spectral and energy efficient multi-hop D2D communication underlay 5G networks," in *2020 International Conference on Software, Telecommunications and Computer Networks (SoftCOM)*, Split, Croatia, 2020.

[13] G. Wang, T. Liu, and C. Zhao, "Joint channel and power allocation based on generalized Nash bargaining solution in device-to-device communication," *Access*, vol. 7, pp. 172571–172583, 2019.

[14] S. Ghosh and D. De, "CG-D2D: cooperative game theory based resource optimization for D2D communication in 5G wireless network," in *2020 Fifth International Conference on Research in Computational Intelligence and Communication Networks (ICRCICN)*, Bangalore, India, 2020.

[15] S. Dominic and L. Jacob, "Distributed resource allocation for D2D communications underlaying cellular networks in time-varying environment," *IEEE Communications Letters*, vol. 22, no. 2, pp. 388–391, 2018.

[16] C. Xia, S. Xu, and K. S. Kwak, "Resource allocation for device-to-device communication in LTE-A network: a Stackelberg game approach," in *2014 IEEE 80th Vehicular Technology Conference (VTC2014-Fall)*, Vancouver, BC, Canada, 2014.

[17] C. Sun, M. Peng, Y. Sun, Y. Li, and J. Jiang, "Distributed power control for device-to-device network using Stackelberg game," in *2014 IEEE Wireless Communications and Networking Conference (WCNC)*, Istanbul, 2014.

[18] S. Lindner, R. Elsner, P. N. Tran, and A. Timm-Giel, "A two-game algorithm for device-to-device resource allocation with frequency reuse," in *2019 IEEE 90th Vehicular Technology Conference (VTC2019-Fall)*, Honolulu, HI, USA, 2019.

[19] R. Yin, G. Yu, C. Zhong, and Z. Zhang, "Distributed resource allocation for D2D communication underlaying cellular networks," in *2013 IEEE International Conference on Communications Workshops (ICC)*, Budapest, Hungary, 2013.

[20] Y. Yuan, T. Yang, H. Feng, and B. Hu, "An iterative matching-Stackelberg game model for channel-power allocation in D2D

underlaid cellular networks," *IEEE Transactions on Wireless Communications*, vol. 17, no. 11, pp. 7456–7471, 2018.

[21] 3GPP, "Further advancements for E-UTRA physical layer aspects," *TR 36.814*, vol. 9.2.0, 2017.

[22] R. S. Sutton and A. G. Barto, "Reinforcement learning: an introduction," *IEEE Transactions on Neural Networks*, vol. 9, no. 5, pp. 1054–1054, 1998.