Performance of Resource Allocation for D2D Communications in Q-Learning Based Heterogeneous Networks

Yung-Fa HUANG, *Member, IEEE*, Tan-Hsu TAN, *Member, IEEE*, Yu-Ling LI, and Shao-Chieh HUANG

***Abstract*--This paper investigates energy efficiency issues of device-to-device (D2D) communications in heterogeneous networks. To minimize the total transmit power, an approach based on Q-learning together with adaptive ε-greedy is proposed to optimize the connection of user equipment (UE) with base station (BS) or Access point (AP). The proposed adaptive *ε-greedy* can conduct the adequate exploration and exploitation operations for effective optimization. Simulation results indicate that in the single-cell scenario, the proposed adaptive *ε-greedy* can obtain performance close to the best solution.**

I. NTRODUCTION

Device-to-device (D2D) communication technology [1] is regarded as an important technical standard for next-generation mobile communication. D2D communication can use the Wi-Fi Direct technology recently developed by the Alliance for Wi-Fi [2] which allows a user device to relaying a signal of a base station (BS) as an access point to other user devices as an access point (AP) to extend the service range of the base station.

In mobile communication systems, in order to increase the spectrum usage, the cell size of the honeycomb gradually shrinks. The traditional concept of the cellular network is heading toward the heterogeneity of integrating the wireless LAN with the cell layers of different sizes Heterogeneous networks (HetNets) [3]. Reference [4] explores the application of D2D communication in heterogeneous networks with the user device acting as a relay, forwards the signal from the base station to another user, and uses the Q-learning algorithm [5] to select the best of the relay device to connect, can reduce the total system transmission power.

In this study, the D2D communication system is used as the basic architecture to explore the effectiveness of HetNets in which the cellular network and the wireless LAN technology coexist. This study propose the adaptive ε-greedy to balance Q Exploration and Exploitation effectiveness of Q-learning to bring performance closer to optimal solution

II. SYSTEM MODELS OF Q-LEARNING BASED SOURCE ALLOCATION

In this paper, we assume that in the OFDMA communication network in which each UE can be converted into a relay AP and provide wireless links to other UEs. The channel model includes the path loss and shadowing fading. The transmission power includes the power consumed by the BSs for transmission to the APs or UEs and the APs for transmission to the UEs. The total system transmission power can be expressed as

$$P_{TOT} = \sum_{j=1}^{J} P_{RB,j} \cdot M_{req,j} + \sum_{k=1}^{K} P_A \cdot \Theta_k , \tag{1}$$

where the first term is the total transmission power of the BSs, which can be obtained by multiplying the RBs provided by the BSs by the power of transmitting each RB. The second is the total transmission power of all the APs obtained by multiplied by the transmission power P_A of the AP to the bandwidth of the UEs required. Therefore, the optimization objective function can be expressed as

$$\min_{b_{k,n},c_{n,j}} P_{TOT} = \min_{b_{k,n},c_{n,j}} \left(\begin{array}{c} \sum_{j=1}^{J} P_{RB,j} \cdot M_{req,j}(b_{k,n}, c_{n,j}) \\ + \sum_{k=1}^{K} P_A \cdot \Theta_k(b_{k,n}) \end{array} \right) . \tag{2}$$

When an AP $A_k (k = 1,\ldots,K)$ and a BS $S_j (j = 1,\ldots,J)$ are used by the UE u_n, the UE stores its usage experience in $Q_{AP,n}(k)$ and $Q_{BS,n}(j)$, respectively. The higher its value, the better the connection quality. In the future, when the AP or BS is connected by the UE, it is updated $Q_{AP,n}(k)$ or $Q_{BS,n}(j)$, respectively. This study is a two-hop connection system, so Q-learning can be expressed as

$$Q_{AP,n}(k) \leftarrow (1-\alpha) \cdot Q_{AP,n}(k) + \alpha \cdot W_{AP,n}(k) \tag{3}$$

$$Q_{BS,n}(j) \leftarrow (1-\alpha) \cdot Q_{BS,n}(j) + \alpha \cdot W_{BS,n}(j) \tag{4}$$

where the direction of the arrow represents the pre-post-update value, $\alpha \in (0,1)$ is the learning rate; and $W_{AP,n}(k)$ 和 $W_{BS,n}(j)$ are the rewards of the AP and the BS, respectively.

When the UE links through the AP, its reward $W_{AP,n}(k)$ can be expressed as

$$W_{AP,n}(k) = \begin{cases} 0, & \text{if } \hat{r}_n < R_n \\ 1 - \dfrac{P_A \cdot \theta_{k,n} + P_{k,n,j}}{P_A + P_{\max,j}}, & \text{otherwise} \end{cases} \tag{5}$$

where $P_A \cdot \theta_{k,n}$ is the transmission power required for the AP to serve the UE; $P_{k,n,j}$ is the transmission power for the BS to transmit the service to the UE via the AP. The transmission power of the BS to the UE can be known from $P_{RB,j} \cdot M_{k,n,j}$.

When the UE is directly connected to the BS, its reward $W_{BS,n}(j)$ can be expressed as

$$W_{BS,n}(j) = \begin{cases} 0, & \text{if } \hat{r}_n < R_n \\ 1 - \dfrac{P_{n,j}}{P_{\max,j}}, & \text{otherwise} \end{cases} \tag{6}$$

where $P_{n,j}$, the transmission power of the serving UE for the BS can be obtained from $P_{RB,j} \cdot M_{n,j}$, where $M_{n,j}$ is the

MOST 106-2221-E-324-020

number of RBs that the BS needs to provide to transmit the service to the UE and may be known from $R_n / r_{U,n,j}$.

At the beginning, the BS or AP was not previously used by the UE, whose initials $Q_{AP,n}(k)$ and $Q_{BS,n}(j)$ can be computationally obtainable through similar payoffs (5) and (6). The initial connection of the UE to the AP can be expressed as

$$Q_{AP,n,initial}(k)=\begin{cases}0, & \text{if } (\theta_{k,n}>1) \text{ or } (M_{k,n,j}>M)\\ 1-\dfrac{P_A\cdot\theta_{k,n}+P_{k,n,j}}{P_A+P_{\max,j}}, & \text{otherwise}\end{cases} \tag{7}$$

where the first condition is that the channel state of the D2D link is not allowed to be reached (for example, $\theta_{k,n}>1$) or the BS cannot provide available RBs to the UE.

The initials $Q_{BS,n}(j)$ of the connection of UE directly to BS can be expressed as

$$Q_{BS,n,initial}(j)=\begin{cases}0, & \text{if } (M_{n,j}>M)\\ 1-\dfrac{P_{n,j}}{P_{\max,j}}, & \text{otherwise}\end{cases} \tag{8}$$

In ε-*greedy* , the agent will randomly select the action at a fixed probability ($0\le\varepsilon\le1$) in each time step, and the chances $1-\varepsilon$ are that the action corresponding to the maximum Q will be executed. The selection method $\pi(n)$ is

$$\pi(n)=\begin{cases}\text{ramdom action } A_k \text{ or } S_j, \text{if } (\xi<\varepsilon)\\ \arg\max_{k\in A_k,\ j\in S_j}\left(Q_{AP,n}(k),Q_{BS,n}(j)\right), \text{otherwise}\end{cases} \tag{9}$$

wherein each T_s generated uniform random variable with $0\le\xi\le1$. If ξ less than ε, the UE randomly selects an AP or a BS link; otherwise, the UE selects an AP or a BS link with a current maximum Q value. In addition, the adaptive ε-*greedy* parameters $\varepsilon=\varepsilon_0/\sqrt{T_s}$ that can be used allow the UE to explore early with a higher probability and gradually decrease with time to increase the probability of developing highly reward moves.

III. SIMULATION RESULTS

In this study, suppose a base station is located in a grid of 400 meters in length and width and has coordinates (0,0) respectively. Six APs are uniformly distributed in in the grid of 200 m long and wide. Twelve UEs (green triangles) are uniformly distributed in the grid of 400 m long and wide. The simulation parameters are shown in Table 1. The effect of the parameters ε in ε-*greedy* and α of Q-learning on the transmission power performance is shown in Fig. 1, with high values searching for better APs or BS links in short time steps, so Q-learning (α=0.9 and ε=0.9) performs best with 90% probability exploration.

The effect of adaptive ε-*greedy* and α parameters on the consumed power performance of Q-learning is shown in Fig. 2. After $T_s\ge50$, the adaptive ε-*greedy* with α=0.9 and $\varepsilon_0=0.9$ can reduce transmission power to 13 Watts.

TABLE 1 SIMULATION PARAMETERS

Number of RB in BS (M, RBs)	25
Bandwidth of each RB in BS (kHz)	180
Total bandwidth of AP (B_A, MHz)	20
Number of BS (J)	1
Number of AP (K)	6
Number of UE (N)	12
BS transmission power per RB ($P_{RB,j}$, dBm)	29
AP transmission power (P_A, dBm)	20
Throughput threshold of AP and UE (R_A, Rn,Mb/s)	5
Power spectral density of AWGN (N_o , dBm/Hz)	-164

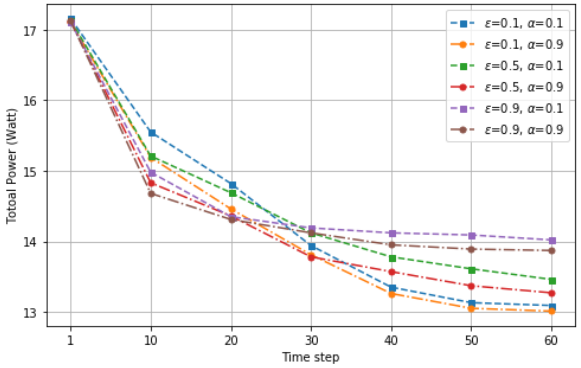

Fig. 1 The performance comparisons of parameters ε in ε-*greedy* and α of Q-learning on the transmission power.

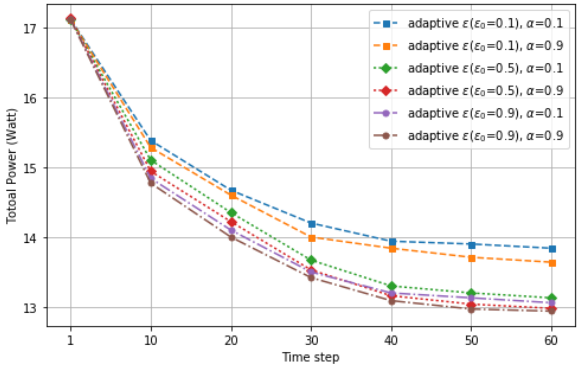

Fig. 2 Transmission power comparisons for adaptive ε-*greedy* with different ε_0 and α.

REFERENCE

[1] L. Wei, R. Wu, Y. Qian, and G. Wu, "Enable device-to-device communications underlaying cellular networks: challenges and research aspects," *IEEE Commun. Mag.*, vol. 52, no. 6, pp. 90-96, Jun. 2014.

[2] D. Camps-Mur, A. Garcia-Saavedra, and P. Serrano, "Device-to-device communications with Wi-Fi Direct: Overview and experimentation," *IEEE Wireless Commun.*, vol. 20, no. 3, pp. 96-104, Jun. 2013.

[3] A. Damnjanovic, J. Montojo, Y. Wei, T. Ji, T. Luo, M. Vajapeyam, T. Yoo, O. Song, and D. Malladi, "A survey on 3GPP heterogeneous networks," *IEEE Wireless Commun.*, vol. 18, no.3, pp. 10-21, Jun. 2011.

[4] J. Pérez-Romero, J. Sánchez-González, R. Agustí, B. Lorenzo and S. Glisic, "Power-efficient resource allocation in a heterogeneous network with cellular and D2D capabilities," *IEEE Trans. Vehi. Techno.*, vol. 65, no. 11, pp. 9272-9286, Nov. 2016.

[5] R. S. Sutton and A. G. Barto, Reinforcement Learning: An Introduction. Cambridge, MA, USA: MIT Press, 1998.