

# Dynamic power control for relay-aided transmission based on deep reinforcement learning

Qin Cai, Wang Chaowei(✉), Wang Weidong, Zhang Yinghai

1. School of Electronic Engineering, Beijing University of Posts and Telecommunications, Beijing 100876, China

2. Key Laboratory of Universal Wireless Communications, Beijing University of Posts and Telecommunication, Beijing 100876, China

## Abstract

Using relay in the wireless communication network is an efficient way to ensure the data transmission to the distant receiver. In this paper, a dynamic power control (DPC) approach is proposed for the amplify-and-forward (AF) relay-aided downlink transmission scenario based on deep reinforcement learning (DRL) to reduce the co-channel interference caused by spectrum sharing among different nodes. The relay works in a two-way half-duplex (HD) mode. Specifically, the power control of the relay is modeled as a Markov decision process (MDP) and the sum rate maximization of the network is formulated as a DRL problem. Simulation results indicate that the proposed method can significantly improve the system sum rate.

**Keywords** power control, deep reinforcement learning, relay, downlink

## 1 Introduction

Relay-aided wireless network attracted lots of attentions in the past because of its advantages in coverage extension and transmission enhancement [1]. Recently, the study of dual-hop transmission through relays gained new actuality in terms of cooperative wireless communication systems [2–5]. However, with the introduction of dual-hop transmission link, the co-channel interference caused by spectrum sharing in the network is more serious.

To solve the problem above, many methods such as interference alignment (IA) [3–5], cognitive radio

[6–7], and interference coordination [8], etc. were proposed to manage the interference in the relay-aided network. In Ref. [9], a hybrid relaying algorithm is proposed to opportunistically switch the transmission mode under different interference conditions. A joint relay and antenna selection method is investigated to optimize the end-to-end error performance in general full-duplex (FD) relay networks [10]. However, the transmit power of relay in Refs. [9–10] is set to the maximum value which results in excessive self-interference among different nodes and limiting the transmission rate of each link.

As an efficient way of interference coordination, power control was widely studied in wireless communication systems. Most of existing literatures, e.g. [11–13] studied power control in the scenario of cognitive network in which the transmit power of the secondary user was controlled to obtain the desired transmission performance. The energy-efficient of

cognitive radio spectrum sharing system was studied in Ref. [11] and a power control scheme was proposed to maximize the energy efficiency of the secondary user. In Refs. [12 – 13], the authors performed power control scheme on the secondary user to ensure the interference under a specified threshold. In recent years, power control was also investigated in relay-aided network [14 – 15]. In Ref. [14], a power control algorithm for two-way FD relay was proposed to reduce the power consumption of relay and ensure the data transmission of the network. However, the algorithm in Ref. [14] only adjusted the power of relay but did not consider the optimal system performance. In order to improve the efficiency of power control, a double-layer Stackelberg game theory model was proposed to jointly allocate the power consumption of both user terminals and relays [15]. In Refs. [16 – 17], the power control was performed on both source and relay nodes to satisfy the power constraints and achieve the global optimal system performance, respectively. However, all the algorithms above were studied to reduce the interference among dual-hop transmission links. For the scenario that direct link is coexistent with dual-hop link, the power control which can suppress co-channel interference are not extensively studied.

On the other hand, reinforcement learning (RL) attracted attentions in resource allocation and management [18], especially by combining with deep neural network (DNN) and further developed as DRL [19] to satisfy the high-dimensional observations. For example, DRL was used to implement the dynamic channel allocation of multi-beam satellite communication system in Ref. [20]. In Ref. [21], it was used to solve the power-efficient resource allocation problem to meet user demands. In Ref. [22], the problem of allocating the powers among all active nodes was mapped as a game-theory-based model and Q-learning algorithm was used to find the Nash equilibrium points. Especially in Ref. [23], a power control method based on DRL was developed in the non-cooperative two-user cognitive radio system. However, for the relay-aided network, the power control scheme based on DRL is not widely investigated

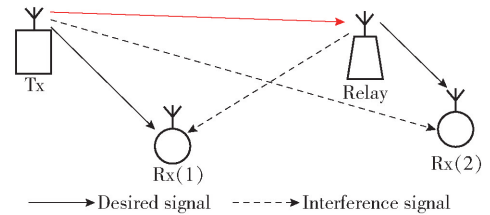
yet, especially for the scenario which mentioned above (i. e. coexistence of direct link and dual-hop link).

In this paper, the DPC approach for the downlink transmission of AF relay-aided network is studied from the DRL perspective. The relay works in a two-way HD mode such that it is able to collect the state from environment in addition to forwarding the signal from transmitter (Tx). The DRL algorithm is used to train the relay to obtain a power control policy to reduce the co-channel interference caused by spectrum sharing. Specifically, the total received power strength at each receiver (Rx) is used to construct the system state space. It should be noted that the data transmission occupies the different frequency from state collection. The experimental results show that relay can dynamically control its power to reach the target state through the proposed DRL-based power control scheme.

The rest of this paper is organized as follows. In Sect. 2, the system model of relay-aided network is described. In Sect. 3, the details of proposing the DPC approach are illustrated. Simulation results are presented and discussed in Sect. 4. Finally, Sect. 5 concludes the paper.

## 2 System model

The system model of relay-aided wireless communication network is illustrated in Fig. 1.



**Fig. 1** System model of relay-aided wireless communication network

As can be seen from Fig. 1, there are 1 Tx, 1 relay and 2 Rx in the network. Specifically, the  $k$ th Rx in the network is denoted as  $Rx(k)$ . The signal to  $Rx(1)$  can be directly received and to  $Rx(2)$  is forwarded through the relay. Both Tx and Rx are equipped with 1 antenna. Furthermore, all the nodes occupy the same

frequency band to carry on data transmissions, which results in the co-channel interference. Nevertheless, it should be noted that the feedback signal used in DRL training process (i. e. the strength of total received signal at each Rx) which takes a low data rate cost is transmitted on another spectrum.

According to Fig. 1, since each node in the network is equipped with only 1 antenna, the data transmissions from the Tx to relay and Rx(1) are assumed to perform in 2 time slots to avoid the interference between relay and Rx(1). Specifically, it is assumed that the data transmission to relay (i. e. the red solid arrow in Fig. 1) is performed in the first time slot, which can be expressed as:

$$y_{t_1}(r) = h_{tr}x_{t_1}(2) + n_{t_1}(r) \quad (1)$$

In Eq. (1),  $h_{tr}$  is the channel gain between Tx and relay,  $x_{t_1}(2)$  and  $n_{t_1}(r)$  denote the desired signal to Rx(2) and the additive white Gaussian noise in time slot  $t_1$ , respectively.

In the second time slot, both Tx and relay send data simultaneously to their destinations (i. e. the black solid arrow in Fig. 1). Hence, the co-channel interference occurs (i. e. the black dotted arrow in Fig. 1), and the signal received at each Rx can be written as Eqs. (2) and (3):

$$y_{t_2}(1) = h_1x_{t_2}(1) + h_{r1}y_{t_1}(r) + n_{t_2}(1) = \underbrace{h_1x_{t_2}(1)}_{\text{Desired}} + \underbrace{h_{r1}h_{tr}x_{t_1}(2)}_{\text{Interference}} + \bar{n}_{t_2}(1) \quad (2)$$

In Eq. (2)  $h_1$  and  $h_{r1}$  are channel gains from Tx and relay to Rx(1), respectively.  $\bar{n}_{t_2}(1) = h_{r1}n_{t_1}(r) + n_{t_2}(1)$  stands for the equivalent noise coefficient at Rx(1) in time slot  $t_2$ .

$$y_{t_2}(2) = \underbrace{h_2x_{t_2}(2)}_{\text{Desired}} + \underbrace{h_2x_{t_2}(1)}_{\text{Interference}} + n_{t_2}(2) \quad (3)$$

In Eq. (3),  $h_2$  and  $h_{r2}$  are channel gains from the Tx and relay to Rx(2), respectively, and  $n_{t_2}(2)$  denotes the additive white Gaussian noise at Tx(2) in time slot  $t_2$ .

As can be seen from Eqs. (2) and (3), the received signal at each Rx contains both desired and interference signals. And the system performance is directly affected by the transmit nodes (i. e. the power of Tx and relay). And the objective of the work in this paper is to help the relay learn an appropriate policy to

adjust its transmit power, such that the optimal system performance can be obtained after rounds of learning. Assume that the transmit power of Tx and relay are denoted by  $p_1$  and  $p_{r2}$ . Hence, for both the two Rxs, the signal to interference and noise ratio (SINR) is defined as:

$$\Phi_1 = \frac{g_1 |h_1|^2 p_1}{g_{r1} |h_{r1}h_{tr}|^2 p_{r2} + \bar{N}_1} \quad (4)$$

and

$$\Phi_2 = \frac{g_2 |h_2|^2 p_1}{g_2 |h_2|^2 p_1 + N_2} \quad (5)$$

In Eqs. (4) and (5),  $\bar{N}_1$  and  $N_2$  represent the noise power at Rx (1) and Rx (2), respectively.  $g_k$  represents the path loss between Tx and Rx( $k$ ) in free space as well as  $g_{rk}$ , as mentioned in Ref. [23], which are given as:

$$\left. \begin{aligned} g_k &= \left( \frac{\lambda}{4\pi d_k} \right)^2 \\ g_{rk} &= \left( \frac{\lambda}{4\pi d_{rk}} \right)^2 \end{aligned} \right\} \quad (6)$$

It should be noted that the sum rate of the network is used to measure the system performance, which can be obtained based on Eqs. (4) and (5) and expressed as:

$$R_{\text{sum}} = \sum_{k=1}^2 \text{lb}(1 + \Phi_k) \quad (7)$$

To meet the performance requirement, it is assumed that the sum rate has to satisfy a minimum  $R_{\text{sum}}$  requirement (i. e. no less than a given sum rate threshold  $\varphi$ ) for data transmission, i. e.

$$\left. \begin{aligned} R_{\text{sum}} &\geq \varphi \\ \text{s. t.} \quad & \\ &p_1 \leq p_{\text{max}} \\ &p_{r2} \leq p_{\text{max}} \end{aligned} \right\} \quad (8)$$

According to Eqs. (4) and (5), when the transmit power of Tx increases, the rate of Rx(1) increases with the power. While the interference caused by Tx also increases, which decreases the data rate of Rx(2). And it has further influence on the sum rate performance of the network. The same problem occurred when the power of relay is changed. Hence, it is necessary to propose an appropriate scheme to ensure the sum rate of the network.

In this paper, the DRL method is utilized to control

the transmit power of relay. Firstly, in order to meet the sum rate requirement, the Tx is supposed to adaptively adjust the transmit power based on its own scheme. Specifically, the transmit power of Tx includes two states: unchanged and changed. The transition probability matrix is set to be:

$$\mathbf{P}_{\text{Tx}} = (0.6, 0.4) \quad (9)$$

Eq. (9) means that the transmit power of Tx keeps unchanging with probability of 0.6 during the training process, and changing with probability of 0.4. For the rule of changing, the transmit power of Tx adjusted according to the classical power control algorithm [23], i. e.

$$p_1(z+1) = D\left(\frac{\eta_1 p_1(z)}{\Phi_1(z)}\right) \quad (10)$$

In Eq. (10),  $\Phi_1(z)$  and  $p_1(z)$  denote the measured SINR and transmit power of Rx(1) at the  $z$ th time frame, here it is assumed that the transmit power of Tx is adjusted according to a time framed basis.  $\eta_1$  is a given fixed SINR coefficient to constrain the range of changing.  $D(\cdot)$  denotes the discretization operation which is able to map the continuous-valued levels into a set of discrete values [23].

The relay in the network plays as an agent in DRL, which is responsible to select an action and interact with the wireless environment. The configuration is as similar as the secondary user in Ref. [23] and the scheduler in Ref. [24]. In the relay-aided network, we assume that the two-way relay is connected to each Rx through low rate backhaul links which occupies different frequency band from the data transmission link. The system state is constructed by the total received power of each Rx (details will be given in the Sect. 3). After obtaining the total received power of received signal, each Rx sends it to the relay through the wireless backhaul link.

Based on the description above, the relay is able to adjust its transmit power dynamically. As similar as in Ref. [23], the transmit power of relay is assume to be adopted from a finite set, i. e.

$$\begin{aligned} P &\triangleq \{p_1^1, p_1^2, \dots, p_1^M\} \\ \text{s. t. } &\left. \begin{aligned} p_1^1 &\leq p_1^2 \leq \dots \leq p_1^M \end{aligned} \right\} \end{aligned} \quad (11)$$

In Eq. (11),  $p_1^1, p_1^2, \dots, p_1^M$  denote the  $M$  elements

of the finite set. In order to satisfy the sum rate requirement i. e. Eq. (8), the relay need to be trained to properly select an item from the set in each training step such that the maximum system sum rate can be obtained after rounds of training steps.

### 3 DPC approach based on DRL

In this section, the procedure of the proposed power control approach is delivered. Firstly, the principle of DRL algorithm is briefly introduced. Then, the problem of power control is mapped as a MDP process. Finally, the introduction about how to implement the proposed DPC approach is given.

#### 3.1 Introduction of DRL

As we all know, reinforcement learning is an important component of machine learning, and the most remarkable architecture of which is that DRL is able to learn from the external environment. Generally, the procedure of training the agent in reinforcement learning is usually regarded as a MDP. It aims at finding an optimal state-action policy and converting the environment into the optimal state by learning the feature of environment i. e. finding the optimal action iteratively to obtain the maximum rewards. According to the property of MDP, for a given policy  $a = \pi(s) \in A$ ,  $s \in S$  (where  $a$  stands for the action in action space  $A$  and  $s$  denotes the state in state space  $S$ ), the state value function can be written as defined in Ref. [25].

$$V^\pi(s) = R(s, a) + \gamma \sum_{s' \in S} \Pr(s'|s, a) V^\pi(s') \quad (12)$$

In Eq. (12),  $R(s, a)$  is the expected value when performing action  $a$  at state  $s$ .  $\gamma \in (0, 1)$  is the discount factor which is used to discount the future rewards.  $\Pr(s'|s, a)$  is the transition probability from  $s$  to  $s'$  when performing action  $a$ . The objective of training is to obtain the optimal policy which satisfies:  $\pi^* = \arg \max_{\pi: s \rightarrow a} [V^\pi(s)]$ ;  $\forall s \in S, a \in A$  (13)

As a widely used method, Q-learning is often used to solve the problem above. Specifically, the Q-function (i. e. the state-action function) based on a discount factor  $\varepsilon$  is given as:

$$Q^\pi(s, a) = R(s, a) + \varepsilon \sum_{s' \in S} \Pr(s'|s, a) V^\pi(s') \quad (14)$$

Based on that, the discounted cumulative reward can be obtained after performing action  $a$  at state  $s$ . By selecting the action with maximum Q-value in each iteration, the agent learns how to dynamically perform an action based on an optimal policy after a number of iterations and Eq. (13) can be expressed as:

$$\pi^* = \arg \max_{\pi: s \rightarrow a} [Q^\pi(s, a)]; \quad \forall s \in S, a \in A \quad (15)$$

Through that, the optimal policy can be obtained. It should be noted that the Q-function is recursively achieved based on an available transfer sample  $(s, a, r, s')$ , where  $r$  denotes the instantaneous reward,  $\tau$  denotes the training time slice, and update with a learning rate  $\alpha$  i. e. Eq. (16) and converge to an optimal action-value function  $Q^{\pi^*}(s, a)$  at last.

$$Q_{\tau+1}(s, a) = Q_\tau(s, a) + \alpha(r + \gamma \max_{\pi: s' \rightarrow a} Q_\tau(s', a') - Q_\tau(s, a)) \quad (16)$$

RL is generally used to deal with the discrete state-limited problem. However, there are always some cases that the state space expanding to a huge status in reality. Then, the RL method will be insufficient on dealing with the challenges in such a complex environment. Hence, the DRL method combining the DNN and RL is introduced to solve this problem [24]. Deep Q-network (DQN) is a typical example of DRL, where the Q-learning process is approximated through DNN i. e.  $Q(s, a) \approx Q(s, a; \theta)$ ,  $\theta$  stands for the weight parameters of DNN. It need to be trained to reach the target value i. e. Eq. (17) by minimizing the loss function in each iteration i. e. Eq. (18), where  $\mathbb{E}[\cdot]$  denotes the mean value operation.

$$Q_{\text{target}} = r + \gamma \max_{a'} Q_\tau(s', a'; \bar{\theta}) \quad (17)$$

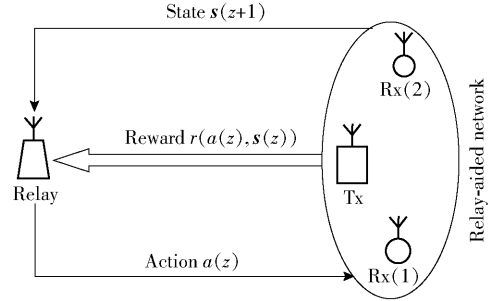
$$\theta; \min L(\theta) = \mathbb{E}[Q_{\text{target}} - Q(s, a; \theta)] \quad (18)$$

Note that the weight parameters of target Q-network updates in period (e. g. every  $N_u$  steps) to ensure the convergence of loss function. It is intended to use DQN in this paper to optimize the sum rate performance of the relay-aided network, and details of formulation will be given in the next subsection.

### 3.2 MDP model of power control

In this section, the power control process of relay is mapped to a MDP model as shown in Fig. 2, which reflects interactions between the relay and environment

(i. e. relay-aided network). According to Fig. 2, the next state in Markov chain model is decided by the action implemented in current state. The main objective of MDP is to learn an optimal policy, which is able to help the agent (i. e. the relay in Fig. 2) to perform the optimal action and obtain the maximum discounted cumulative reward.



**Fig. 2** MDP model of power control in relay-aided network

The essential components of MDP contains the selected action, the state space, reward function and transition probabilities. Specifically, the state space, action space and reward of the MDP-modeled power control system is defined.

#### 3.2.1 System state

Action is generated based on the current system state i. e. the total received power of signal collected at each Rx. The system state vector in the  $z$ th iteration is constructed by each Rx's state, which can be written as:

$$s(z) = [P_1^{\text{sum}}(z), P_2^{\text{sum}}(z)]^H \quad (19)$$

#### 3.2.2 Action index

According to the system state, the relay needs to take an action to adjust the transmit power. Specifically, the system action can be expressed as:

$$a(z) = p_1(z+1) \quad (20)$$

In Eq. (20),  $a(z)$  represents the power control of the relay i. e. changing its transmit power based on the current state  $s(z)$ . The rule of changing and transition probability are given in Eqs. (9) and (10).

#### 3.2.3 System reward function

The system optimization objective is directly designated by the reward function. In this paper, it is

expected to obtain the optimal global system performance by maximizing the sum rate performance of the whole network. According to Eq. (7), the reward function can be defined as

$$r = \begin{cases} \mu^+; & R_{\text{sum}} \geq \varphi \\ \mu^-; & \text{else} \end{cases} \quad (21)$$

In Eq. (21)  $\mu^+$  represents a positive value when the sum rate is content with the requirement, while  $\mu^-$  is a negative reward value when the sum rate requirement is not satisfied. Since the influence of path loss between TxS and RxS is considered, the values of data sent to the Q-network are distributed in a relatively small range (about  $10^{-8} \sim 10^{-5}$ ). For simplicity, the parameters  $\mu^+$  and  $\mu^-$  in the simulation of this paper are set to be  $\mu^+ = 1$  and  $\mu^- = 0$ , respectively. It should be noted that other values are also applicable as long as the gap between  $\mu^+$  and  $\mu^-$  is big enough to support the learning process.

### 3.3 Implementation of the proposed scheme

Since the state space, action space and reward function of the relay-aided system was defined, the implementation of the proposed DPC approach will be elaborated in this subsection. The core of the proposed approach is the Q-network, whose architecture directly determines the mapping from input  $s(z)$  to the output  $Q(s(z), a(z); \theta)$ . And the value of  $Q(s(z), a(z); \theta)$  directly reflects the accumulated rewards when taking an action  $a(z)$ . Specifically, the DNN is adopted to approximate the non-linear function. The construction of DNN is same as in Ref. [23], which contains three fully-connected hidden layers. The number of neurons in each layer are set as 256, 256 and 512, respectively. The rectified linear units (ReLU) is adopted in the first two hidden layers and the tanh function is adopted in the third hidden layer are set as the activation functions in DNN.

Furthermore, in order to improve the stability of the proposed scheme, the experience replay is used to train the Q-network. The replay memory  $B$  is used to store the initial generated tuple  $(s(z), a(z), r(z), s(z+1))$  i. e. experience data, and the capacity of which is set as  $N_B$ . When the number of stored experience tuple

is larger than  $N_D, N_M$  (i. e. the size of minibatch) experience data are randomly sampled from  $B$  in each iteration to start training the Q-network.

Besides, it should be noted that for the action selection in this paper, the  $\varepsilon$ -greedy policy is utilized to balance the exploitation and exploration. During the iteration process, the exploration rate  $\varepsilon$  linearly decreases from an initial value  $\varepsilon_s$  to a final value  $\varepsilon_e$  with the number of iterations  $z$ . i. e.

$$\varepsilon_e = \varepsilon_s \left(1 - \frac{z}{Z}\right) \quad (22)$$

The detailed process of the proposed DPC scheme is illustrated below i. e. the procedure of the proposed DPC scheme.

#### Algorithm 1 Implementation of DPC scheme

1. Initialization
2. Initialize relay-aided network parameters i. e. channel coefficients, distance between each two nodes, path-loss model and transmit power, etc.
3. Initialize the capacity of replay memory  $B$  with  $N_B$ , the parameters of Q-network with random  $\theta_0$ , the parameters of target Q-network with  $\theta = \theta_0$
4. Implementation
5. For  $z = 1, 2, \dots, Z$  do
6.     Calculate the beginning state  $s$  according to Eq. (19)
7.     Generate a random probability  $p$
8.     If  $p \geq \varepsilon$ ,
9.         Choose  $a(z) = \arg \max_{a \in A} Q(s(z), a(z); \theta)$
10.     Else
11.         Randomly choose an action
12.         Obtain the reward  $r(z)$  and next state  $s(z+1)$
13.         Store the data tuple  $(s(z), a(z), r(z), s(z+1))$  in replay memory
14.     If  $z \geq N_D$
15.         Randomly sample a minibatch of data tuples from  $B$
16.         Minimizing the loss function based on Eq. (18)
17.         Update the parameters of Q-network
18.         Reset the target Q network every  $N_u$  steps
19.     End if
20.     End if
21. End for

## 4 Simulation analysis and discussions

In this section, the performance of the proposed

DPC approach in relay-aided network is simulated. Firstly, the simulation platform with software environment and the parameters of both considered scenario and DRL are provided. Then, the training process under different network settings and algorithm configuration are simulated and analyzed. Finally, the sum rate performance under the proposed DPC scheme is examined.

#### 4.1 Simulation configurations

In the simulation, the CPU-based server with version Intel i7-3770, 8 GB memory is adopted. The software environment of DRL-based algorithm is TensorFlow 1.11.0 with Python 3.5 in Anaconda 5.1.0 since it deploys lots of machine learning models and provides Python APIs. The structure of relay-aided network presented in Sect. 2 is taken as the simulation scenario. Both the network parameters such as the antenna configurations and the restriction of transmit powers, etc. are adopted as same as Ref. [23]. For example, the channel gains from the Tx/relay to the Rxs are assumed to be  $h_k/h_{rk} = 1, k = 1, 2$ . Similarly, let assume that the transmit power (in W) of both Tx and relay is selected from a predefined finite set  $P \triangleq \{0.1, 0.2, \dots, 1\}$ . The fixed SINR coefficient  $\eta_1$  that constrains the range of changing in Eq. (10) is set to  $\eta_1 = 1.2$  and the sum rate threshold is determined when the Tx and relay send data with the maximum transmit power. The power of noise is adjusted according to the transmit power of Tx and relay to satisfy the signal-to-noise ratio (SNR) requirement. It should be noted that the nodes deployment in the network are set based on the coverage of Femto BS to approach the real application scenario as far as possible. The detailed parameters of both simulation scenario and DRL are shown in Table 1.

#### 4.2 Numerical results

As shown in Fig. 3, the values of loss function are compared to show the convergence performance under different learning rates. It should be noted that the numerical results are averaged every 100 training steps to constrict amplitude of the training hopping. It can be

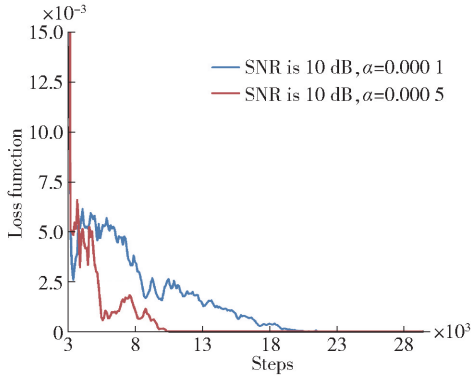
observed from Fig. 3 that the loss function is very high at the beginning of training process. With the increase of training steps, the loss function decreases until it converges to some stable values around zero. Besides, the convergence performance is affected by the learning rate of Adam Optimizer. Specifically, it can be seen from Fig. 3 that the loss function curve at learning rate of 0.0005 converges faster than that of 0.0001. However, a larger learning rate leads to a higher shaking amplitude even over flitting, which decreases the learning accuracy. Hence, in the simulation, the learning rate of 0.0001 is adopted to ensure the accuracy of results in the training process.

**Table 1** The simulation parameters

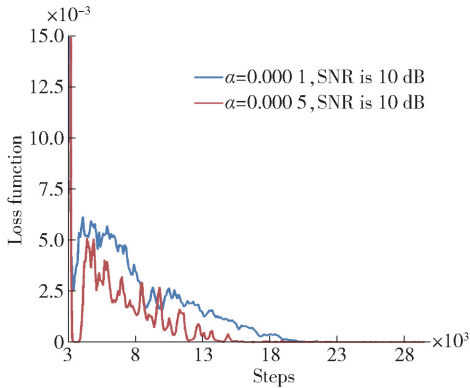
Parameters of simulation	Value
Distance between Tx and Rx(1) $d_1/\text{m}$	5
Distance between Tx and Rx(2) $d_2/\text{m}$	10
Distance between relay and Rx(2) $d_{r2}/\text{m}$	2
Distance between relay and Rx(1) $d_{r1}/\text{m}$	8
Maximum transmit power $P_{\max}/\text{W}$	1
Bandwidth $B/\text{MHz}$	10
Carrier frequency $f/\text{GHz}$	2
Number of antennas at each node	1
Antenna pattern/dB	0
Replay memory capacity $N_B$	10 000
Experience tuple size $N_D$	3 000
Minibatch size $N_m$	16
Activation function of hidden layers	ReLU
Activation function of output layers	tanh
The training optimizer	Adam
The learning rate $\alpha$	0.0001
The value of positive reward $\mu^+$	1
The value of negative reward $\mu^-$	0
The value of discount factor $\gamma$	0.9
Target network update frequency $N_u$	100
The initial value of exploration rate $\varepsilon_s$	0.99
The end value of exploration rate $\varepsilon_e$	0.001

In Fig. 4, the effects of different SNR values on the convergence of each gradient step in the DRL-based algorithm is plotted. The value of SNR directly determines the data inputted to the DNN (the training Q-network). From Fig. 4, it can be seen that the curve converges faster at SNR is 20 dB and the fluctuation is stronger compared to the case of SNR is 10 dB. That is because with the increase of SNR, the range of input data space is spanned and the influence caused by noise is weakened. It should be noted that the system

state in this paper is defined as the total receive power of each user. That means the state space is enhanced and the data structure is prominent due to the increase of SNR. Fig. 4 illustrates that the proposed algorithm performs better on suppressing interference.

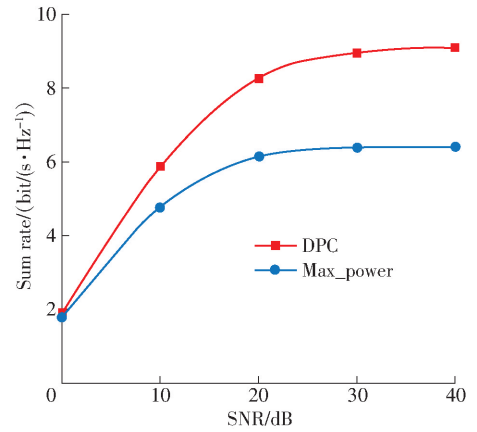


**Fig. 3** The loss function vs. the number of training steps under different learning rates



**Fig. 4** The loss function vs. the number of training steps under different SNR

In Fig. 5, the sum rate performance of the proposed DPC scheme is plotted. It can be obviously seen that the curves of both max\_power and the proposed DPC schemes are firstly increase with SNR and then almost remain unchanged, since both the two transmission links occupy the same frequency and the interference becomes stronger with the increase of transmit power. Furthermore, by comparing the tendency of two curves, it can be seen that with the increase of SNR, the gap between them becomes larger. The reason is that by implementing the DPC scheme, the power of relay is dynamically adjusted based on the network environment to suppress co-channel interference and maximize the sum rate performance.



**Fig. 5** Comparison of sum rate performance vs. SNR of different schemes

### 4.3 Algorithm analysis and discussions

In this section, all the calculations of sum rate are based on Shannon's theorem e. g. Eqs. (4) and (5) which is a widely used general mathematical model in wireless communication system. And the DPC algorithm only uses the data to train the relay to obtain the optimal power control policy but not involved in the calculation of simulation results. Furthermore, the feasibility of simulation parameters analyzed in Sect. 4.1 supports the simulation results as well. Hence, based on the analysis above, the numerical results are authentic.

The two-receiver model is used to derive and verify the feasibility of DRL as a tool to solve power control problems in relay-aided network. In theory, by expanding the dimension of the state space in the DPC algorithm and reasonably setting the threshold of reward function, the existing scheme can be extended to multiple Rx scenarios. However, it should be noted that this paper studies the multiplexing of the same frequency band by multiple Rx's. The introduction of new nodes will multiply the total number of interference signals in the network and then affect the rate of each Rx, which can not guarantee the improvement of the total transmission rate gain of the whole system. Besides, the training process will be complicated with the increase of nodes.

The application of DRL aims at training the relay to adapt to the changes of the external environment so as

to dynamically control the transmission power according to the network environment. There is no doubt that the training process is relatively complex, but it should be noted that the practical application is the algorithm after training (that is to say, Q-network in DQN uses the parameters after training). Through the training process, the agent (i. e. the relay) is able to obtain the optimal power control strategy, and then adjust its transmit power to the optimal state as soon as possible. The application of the DPC after training is relatively simple, which is of great significance to the improvement of sum rate in the relay-aided network with the same frequency multiplexing.

## 5 Conclusions

In this paper, the dynamic power control approach based on DRL (i. e. DPC) is studied to suppress the interference and maximum the sum rate performance in relay-aided network. Specifically, the power control problem in the relay-aided network is modeled as a MDP. The relay plays as the agent to train the DQN to obtain the optimal control policy by adjusting its power. Simulation results show that the DPC method is effective to suppress interference and can significantly improve the sum rate performance of the networks. In future, how to extend the DPC method to multi-cell multiple input multiple output (MIMO) heterogeneous network and jointly optimize the power control and channel allocation will be consider to improve the quality of service performance.

## Acknowledgements

This work was supported by the National Key R&D Program of China (2017YFC0804404), and the Beijing Talents Foundation (2017000020124G067).

## References

- Hua Y B, Bliss D W, Gazor S, et al. Guest editorial: theories and methods for advanced wireless relays: Issue I. *IEEE Journal on Selected Areas in Communications*, 2012, 30(8): 1297 – 1303
- Hasna M O, Alouini M S. Performance analysis of two-hop relayed transmission over Rayleigh fading channels. *Proceedings of the IEEE 56th Vehicular Technology Conference*, Sept, 2002, Vancouver, Canada. Piscataway, NJ, USA: IEEE, 2002: 1992 – 1996
- Wang R, Yuan X J. MIMO multiway relaying with pairwise data exchange: a degrees of freedom perspective. *IEEE Transactions on Signal Process*, 2014, 62(20): 5294 – 5307
- Li X H, Sun Y, Zhao N, et al. A novel interference alignment scheme with a full-duplex MIMO relay. *IEEE Communications Letters*, 2015, 19(10): 1798 – 1801
- Wang R, Yuan X J, Tao M X. Degrees of freedom of MIMO multiway relay channel with clustered pairwise exchange. *IEEE Journal on Selected Areas in Communications*, 2015, 33(2): 337 – 351
- Haykin S. Cognitive radio: brain-empowered wireless communications. *IEEE Journal on Selected Areas in Communications*, 2005, 23(2): 201 – 220
- Mitola J, Maguire G Q. Cognitive radios: making software radio more personal. *IEEE Personal Communications*, 1999, 6(4): 13 – 18
- Jin J, Gao X C, Li X W, et al. Achievable degrees of freedom for the two-cell two-hop MIMO interference channel with half-duplex relays. *IEEE Access*, 2017, 5: 1376 – 1381
- Riihonen T, Werner S, Wichman R. Hybrid full-duplex/half-duplex relaying with transmit power adaptation. *IEEE Transactions on Wireless Communications*, 2011, 10(9): 3074 – 3085
- Yang K, Cui H Y, Song L Y, et al. Joint relay and antenna selection for full-duplex AF relay networks. *Proceedings of the 2014 IEEE International Conference on Communications (ICC'14)*, Jun 10 – 14, 2014, Sydney, Australia. Piscataway, NJ, USA: IEEE, 2014: 4454 – 4459
- Ozcan G, Gursoy M C. Energy-efficient power adaptation for cognitive radio systems under imperfect channel sensing. *Proceedings of the 2014 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS'14)*, Apr 27 – May 2, 2014, Toronto, Canada. Piscataway, NJ, USA: IEEE, 2014: 706 – 711
- Mitliagkas I, Sidiropoulos N D, Swami A. Joint power and admission control for Ad-hoc and cognitive underlay networks: Convex approximation and distributed implementation. *IEEE Transactions on Wireless Communications*, 2011, 10(12): 4110 – 4121
- Tadrous J, Sultan A, Nafie M. Admission and power control for spectrum sharing cognitive radio networks. *IEEE Transactions on Wireless Communications*, 2011, 10(6): 1945 – 1955
- Li Y, Li N, Peng M G, et al. Relay power control for two-way full-duplex amplify-and-forward relay networks. *IEEE Signal Processing Letters*, 2016, 23(2): 292 – 296
- Guo Y, Jiang F, Hu J K. Distributed power control with double-layer Stackelberg game and utility learning in cooperative relay networks. *Proceedings of the IEEE 10th Conference on Industrial Electronics and Applications (ICIEA'15)*, Jun 15 – 17, 2015, Auckland, New Zealand. Piscataway, NJ, USA: IEEE, 2015: 306 – 311
- Yu B, Yang L Q, Cheng X, et al. Transmit power optimization for full duplex decode-and-forward relaying. *Proceedings of the 2013 IEEE Global Communications Conference (GLOBECOM'13)*, Dec 9 – 13, 2013, Atlanta, GA, USA. Piscataway, NJ, USA: IEEE, 2013: 3347 – 3352

- signature scheme in quantum cryptosystem. *International Journal of Theoretical Physics*, 2014, 53(1): 28–38
10. Wang C, Liu, J W, Shang T. Enhanced arbitrated quantum signature scheme using Bell states. *Chinese Physics B*, 2014, 23(6): 060309/8p
  11. Yu C H, Guo G D, Lin S. Arbitrated quantum signature scheme based on reusable key. *Science China: Physics Mechanics & Astronomy*, 2014, 57(11): 2079–2085
  12. Li F G, Shi J H. An arbitrated quantum signature protocol based on the chained CNOT operations encryption. *Quantum Information Processing*, 2015, 14(6): 2171–2181
  13. Zhang L, Sun H W, Zhang K J, et al. An improved arbitrated quantum signature protocol based on the key-controlled chained CNOT encryption. *Quantum Information Processing*, 2017, 16: Article 70/15p
  14. Zhang M L, Liu Y H, Nie M, et al. Arbitrated quantum signature of quantum messages with a semi-honest arbitrator. *International Journal of Theoretical Physics*, 2018, 57(5): 1310–1318
  15. Guo Y, Feng Y, Huang D, et al. Arbitrated quantum signature scheme with continuous-variable coherent states. *International Journal of Theoretical Physics*, 2016, 55(4): 2290–2302
  16. Feng Y Y, Shi R H, Guo Y. Arbitrated quantum signature scheme with continuous-variable squeezed vacuum states. *Chinese Physics B*, 2018, 27(2): 020302/10p
  17. Gao F, Qin S J, Guo F Z, et al. Cryptanalysis of the arbitrated quantum signature protocols. *Physical Review A*, 2011, 84(2): 022344/7p
  18. Choi J W, Chang K Y, Hong D. Security problem on arbitrated quantum signature schemes. *Physical Review A*, 2011, 84(6): 062330/4p
  19. Zhang K J, Zhang W W, Li D. Improving the security of arbitrated quantum signature against the forgery attack. *Quantum Information Processing*, 2013, 12(8): 2655–2669
  20. Zhang K J, Li D, Su Q. Security of the arbitrated quantum signature protocols revisited. *Physica Scripta*, 2014, 89(1): 015102/9p
  21. Zhang K J, Qin S J, Sun Y, et al. Reexamination of arbitrated quantum signature: the impossible and the possible. *Quantum Information Processing*, 2013, 12(9): 3127–3141
  22. Zhang L, Sun H W, Zhang K J, et al. The security problems in some novel arbitrated quantum signature protocols. *International Journal of Theoretical Physics*, 2017, 56(8): 2433–2444
  23. Kim T, Choi J W, Jho N S, et al. Quantum messages with signatures forgeable in arbitrated quantum signature schemes. *Physica Scripta*, 2015, 90(2): 025101/7p
  24. Kim T, Lee H S, Lee S. Forgeable quantum messages in arbitrated quantum signature schemes. *Physica Scripta*, 2017, 16: Article 268/9p
  25. Xu G L, Zou X F. Security analysis of an arbitrated quantum signature scheme with Bell states. *International Journal of Theoretical Physics*, 2016, 55(9): 4142–4156
  26. Boykin P O, Roychowdhury V. Optimal encryption of quantum bits. *Physical Review A*, 2003, 67: 042317/6p

(Editor: Wang Xuying)

## From p. 43

17. Ugurlu U, Wichman R, Riihonen T, et al. Power control and beamformer design for the optimization of full-duplex MIMO relays in a dual-hop MISO link. *Proceedings of the 9th International Conference on Cognitive Radio Oriented Wireless Networks and Communications (CROWNCOM'14)*, Jun 2–4, 2014, Oulu, Finland. Piscataway, NJ, USA: IEEE, 2014: 545–549
18. Gao L, Duan L J, Huang J W. Two-sided matching based cooperative spectrum sharing. *IEEE Transactions on Mobile Computing*, 2017, 16(2): 538–551
19. Galindo-Serrano A, Giupponi L. Distributed Q-learning for aggregated interference control in cognitive radio networks. *IEEE Transactions on Vehicular Technology*, 2010, 59(4): 1823–1834
20. Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning. *Nature*, 2015, 518: 529–533
21. Han G A, Xiao L, Poor H V. Two-dimensional anti-jamming communication based on deep reinforcement learning. *Proceedings of the 2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'17)*, Mar 5–9, 2017, New Orleans, LA, USA. Piscataway, NJ, USA: IEEE, 2017: 2087–2091
22. Shams F, Bacci G, Luise M. Energy-efficient power control for multiple-relay cooperative networks using Q-learning. *IEEE Transactions on Wireless Communications*, 2015, 14(3): 1567–1580
23. Li X J, Fang J, Cheng W, et al. Intelligent power control for spectrum sharing in cognitive radios: a deep reinforcement learning approach. *IEEE Access*, 2018, 6: 25463–25473
24. He Y, Zhang Z, Yu F R, et al. Deep-reinforcement-learning-based optimization for cache-enabled opportunistic interference alignment wireless networks. *IEEE Transactions on Vehicular Technology*, 2017, 66(11): 10433–10445
25. Rappaport T S. *Wireless communications: principles and practice*. Englewood Cliffs, NJ, USA: Prentice-Hall, 2002

(Editor: Wang Xuying)