



Resource Allocation Based on Deep Reinforcement Learning for Wideband Cognitive Radio Networks

Fuhui Zhou⁽¹⁾, Yuhang Wu⁽¹⁾, and Qihui Wu⁽¹⁾

(1) Nanjing University of Aeronautics and Astronautics, Nanjing, China

Abstract

A wideband cognitive radio network relying on energy harvesting is studied under a practically non-linear energy harvesting model. In order to efficiently exploit the harvesting energy and enhance the performance of the secondary users (SUs), an intelligent resource allocation scheme based on deep reinforcement learning is proposed. The scheme intelligently selects the operation mode and the transmit power of SUs as well as allocates the sub-channels to maximize the defined reward function. Simulation results demonstrate the efficiency of our proposed resource allocation scheme.

1 Introduction

Cognitive radio (CR) is deemed a promising technology in the next-generation wireless networks. It can effectively improve the spectral efficiency by allowing secondary users (SUs) to coexist with the primary users (PUs) at the same spectrum band [1]. In the underlay mode, SUs can access the PU spectrum band as long as the interference caused to the PUs is kept below a certain threshold. Since the transmit power of the SUs cannot be too large in order to protect the PUs' quality of services (QoS), the achievable capacity of SUs is quite limited. In order to enhance the performance of SUs, wideband CR networks that operate in multiple sub-carrier bands have been proposed [3]-[5].

Besides the spectral efficiency issue, the contradiction between the intensive energy consumption requirement for supporting diverse computing and communication services and the limited energy storage capacity of mobile devices is increasingly obvious. Towards that end, wireless power transfer (WPT) techniques were proposed to tackle energy shortage issues on the mobile devices [6]. WPT enables mobile devices to exploit the radio frequency (RF) signals for harvesting energy. It can provide users with stable and controllable energy. There have been two research streams based on the energy harvesting (EH) model, namely, the linear EH model [3] and the non-linear EH model [7]. In [3], the authors proposed a robust beamforming design for the security of the secondary network under a linear EH model. The authors in [7] proposed a resource allocation scheme for a multiple-input multiple-output wireless powered communication network based on a non-linear EH model. It

was shown that the performance of users achieved under the non-linear EH model can be better than that obtained under the linear EH model.

Most existing works address dynamic resource allocation problems by using optimization-based approaches [3]-[5]. In [5], we studied robust resource allocation problem in CRNs with simultaneous wireless information and power transfer (SWIPT) and proposed an algorithm based on successive convex approximation (SCA) for achieving a good tradeoff between the transmission power and the harvested energy. Besides optimization-based approaches, the authors in [8] introduced a method based on reinforcement learning (RL) to solve the resource allocation problem for cloud-based cache-enabled small cell networks. Moreover, the authors in [9] investigated dynamic resource allocation of multiple UAVs enabled communication networks and developed a multi-agent RL framework to maximize the long-term rewards.

However, it is impractical for RL to optimize policies for high-dimensional Markov decision process (MDP) problems as their action-value tables are infinite. Instead deep learning was introduced to replace the action-value table in RL and the new method was called deep reinforcement learning (DRL), also known as deep Q-network (DQN) [10]-[11]. In [10], the authors considered a dynamic multichannel access problem and obtained the optimal access policy by DQN without knowing the system dynamics. In [11], a DRL-based approach was developed for the SUs to learn an efficient power control policy and to access the spectrum band of the PUs in the CRN. It was shown that the performance of SUs can be improved by using the proposed DRL-based power allocation scheme.

Unlike these works, our work considers a wideband CRN where a secondary transmitter (ST) has a rechargeable battery and an energy harvesting circuit that can harvest energy for transmitting information to the SUs. How to balance the ST's energy harvesting and information transmission is also of crucial importance since there exists a non-trivial performance tradeoff between them. In order to tackle this issue, an intelligent resource allocation scheme that decides operation mode, channel selection, and power allocation is proposed based on DRL. Simulation results elucidate the tradeoff between energy harvesting and information transmission.

2 System Model

In the considered downlink CRN, the primary network co-exists with the secondary network under the spectrum sharing paradigm. The primary network consists of one primary transmitter (PT) and one PU while the secondary network consists of one ST and K SUs. All the devices are equipped with a single antenna. The ST has a rechargeable battery and an energy harvesting circuit that can harvest energy for transmitting information to SUs. Let c_i denote the battery storage at SUs at the i th time frame. From a practical circuit architecture design perspective, SUs cannot simultaneously harvest energy and transmit information. Thus, the ST has to make a binary decision regarding whether to transmit information or to harvest energy at each time frame. Let $x_i \in \{0, 1\}$ be the indicating variable at the i th time frame, where $x_i = 0$ denotes that the ST harvests energy and $x_i = 1$ denotes that the ST transmits information to a selected SU at the i th time frame.

Let $\mathbf{h}_k = \{h_{k,m} | k \in \mathcal{K}, m \in \mathcal{M}\}$ denote the channel gain state vector of the links from the ST to the k th SU, where $\mathcal{K} = \{1, \dots, K\}$ and $\mathcal{M} = \{1, \dots, M\}$. K denotes the number of SUs and M denotes the number of sub-carrier channels. e , g and q respectively denote the channel gains of the links from the PT to the ST, from the ST to the PU, and from the PT to the SUs; the values of h_m , e , g , q are independently chosen from the corresponding finite state space $\mathcal{H}$, $\mathcal{E}$, $\mathcal{G}$ and $\mathcal{Q}$ at the beginning of each time frame. Similar to [12], all the channel state transitions across the time frames are modeled as a finite-state discrete-time Markov chain.

The ST dynamically selects one sub-channel to transmit information to a selected SU during the communication process. Let $y_i \in \{-1\} \cup \{m | m \in \mathcal{M}\}$ denote the selected channel. Specially, $y_i = -1$ indicates that the ST harvests energy and does not transmit information while $y_i = m$ indicates that the ST selects the m th sub-channel to transmit information. Let $P_{k,i}^s \leq c_i$ denote the transmit power. Thus, the transmit rate of the k th SU is given as

$$R_i = \frac{B}{M} \log_2 \left(1 + \frac{|h_{k,m,i}|^2 P_{k,i}^s}{|q_i|^2 P_p + \sum_{j \neq k} |h_{j,m,i}|^2 P_{j,i}^s + \sigma^2} \right), \quad (1)$$

where B is the total bandwidth available at the primary user; P_p is the transmit power of the PT; σ^2 is the variance of noise at the SU. In order to guarantee the QoS of the SUs, one has $R_i \geq R_{\min}$, where $R_{\min}$ represents the minimum required transmit rate of the SUs.

In the energy harvesting mode, the ST harvests energy from the radio frequency signals transmitted by the PT and restores the harvested energy in the battery in the ST. For the ideally linear energy harvesting (EH) model, the harvested power of the ST is modeled as $\Phi_i^{\text{linear}} = \mu P_p$, where $0 \leq \mu \leq 1$ denotes the energy conversion efficiency of the ST. Since the practical energy harvesting circuit results in a

non-linear end-to-end energy conversion, a practically non-linear parametric energy harvesting model is applied in this paper. According to [3] and [7], the harvested power at the ST denoted by Φ_i is given by

$$\Phi_i = \frac{\Psi_i - E_{\max} \Omega}{1 - \Omega}, \quad \Omega = \frac{1}{1 + \exp(ab)}, \quad (2a)$$

$$\Psi_i = \frac{E}{1 + \exp(-a(e_i P_p - b))}, \quad (2b)$$

where $E_{\max}$ represents the maximum harvested power at the ST when the EH circuit is saturated; Parameters a and b are related to the specifications of the EH circuit. Moreover, the battery storage is limited to its maximal capacity $C_{\max}$, thus one has $c_i + \Phi_i \leq C_{\max}$. In the case that $c_i + \Phi_i > C_{\max}$, the battery capacity at the next time frame is still capped at $C_{\max}$.

3 DQN-Based Resource Allocation Scheme

3.1 Problem Formulation

In order to efficiently exploit the harvested energy and enhance the performance of SUs, an intelligent resource allocation scheme is proposed for SUs to select the operation mode, the transmit power, and the channel to be used. Let $s_i = (c_i, \mathbf{h}_{k,i}, e_i, g_i, q_i)$ denote the network state of the ST at each time frame i . With observation s_i at the beginning of the i th time frame, the ST strategically takes a joint action $\mathbf{a}_i = (x_i, y_i) \in \{(0, -1)\} \cup \{(1, m) | \forall m\}$ following a stationary policy π . $\{s_i | i \in N_+\}$ is a controlled Markov chain with the state transition probability given as

$$\Pr\{s_{i+1} | s_i, \pi(s_i)\} = \left(\prod_{m \in \mathcal{M}} \Pr\{h_{k,m,i+1} | h_{k,m,i}\} \right) \times \Pr\{e_{i+1} | e_i\} \Pr\{g_{i+1} | g_i\} \Pr\{q_{i+1} | q_i\}. \quad (3)$$

If the ST decides to transmit information, i.e. $x_i = 1$, in order to improve the spectral efficiency of CRNs, the transmit power of the ST is designed to maximize the average transmit rate of the SUs subject to the average interference power constraint and the average transmit power constraint, which are respectively given as

$$E\{g_i P_{k,i}^s\} \leq P_{in}, \quad (4a)$$

$$E\{P_{k,i}^s\} \leq P_{th}, P_{k,i}^s \geq 0, \quad (4b)$$

where $E\{\cdot\}$ denotes the expectation operator; P_{in} denotes the maximum average interference power that can be tolerable for the PU; P_{th} denotes the maximum average transmit power of the SUs. According to our previous work [?], a deep neural network (DNN) which takes channel gains $\{h_{m,i}, g_i, q_i\}$ as inputs is applied to obtain an optimal transmit power $P_{k,i}^s$ at each time frame. Thus, the battery storage in the next time frame can be given as

$$c_{i+1} = \begin{cases} c_i - P_{k,i}^s, & \text{if } x_i = 1, \\ c_i + \Phi_i, & \text{if } x_i = 0. \end{cases} \quad (5)$$

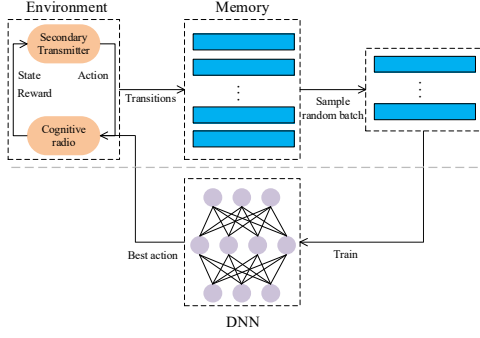


Figure 1. The structure of our proposed DQN algorithm.

After executing the chosen action at the current state, the ST receives a corresponding reward $r_i = r(s_i, a_i)$, defined as

$$r_i = \begin{cases} R_i, & \text{if } x_i = 1, c_i \geq P_{k,i}^s \text{ and } R_i \geq R_{\min}, \\ f_0, & \text{if } x_i = 0 \text{ and } c_i < P_{k,i}^s, \\ 0, & \text{otherwise,} \end{cases} \quad (6)$$

where f_0 is a constant. Then, the ST moves into the next state and all the channel states are changed at the beginning of the new time frame.

The core of MDPs for making a decision is to learn a policy π that specifies the action $\pi(s)$ in the state s . Specifically, the goal of the ST is to learn a policy π for selecting its action a_i on the basis of the current state s_i in order to maximize the expected discounted cumulative reward. The cumulative reward is given as

$$V(s, \pi) = E_\pi \left[\sum_{i \in N_+} \gamma^{i-1} r_i \right], \quad (7)$$

where γ is the discount factor. Thus, the optimal policy π^* can be obtained by maximizing $V(s, \pi)$ at any given network state. In this case, the problem can be formulated as

$$P_1 : \max_{\pi} V(s, \pi). \quad (8)$$

The optimal policy π^* can be given as $\pi^* = \arg \max V(s, \pi)$.

3.2 DQN-Based Scheme

In our formulated problem, the state space s is infinite since the battery storage c is continuous. Thus, the Q-learning method used in [8] is impossible to converge in our problem. In order to tackle this challenge caused by the infinite state space, the deep Q-network proposed in [14] is applied. Instead of generating a finite Q-value table, a deep neural network is applied to approximate the Q-value function in DQN. When the Q-value function is learned, the optimal policy can be obtained by selecting the action with the highest Q-value at each state.

The structure of our proposed DQN algorithm is illustrated in Fig. 1. Let θ denote the set of weights of the Q-network.

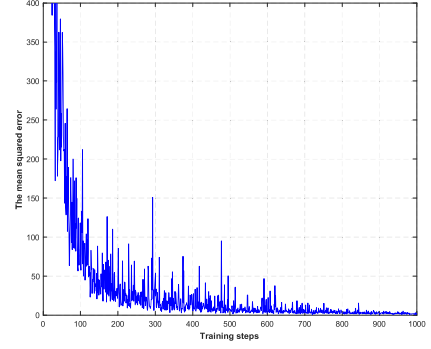


Figure 2. The loss function value versus the DQN training steps.

The Q-value can be expressed as $Q(s, a; \theta)$. The training data used to train the DQN are generated as follows. Given state s_i at time frame i , the ST randomly selects an action with probability ε ; Otherwise, it selects an action a_i which has the largest Q-value. After taking the action a_i , the ST receives a reward r_i and observes a new state s_{i+1} . This transition $d_i = \{s_i, a_i, r_i, s_{i+1}\}$ is stored in the replay memory D . The training of the DQN begins when D collects a sufficient number of transitions.

In the training stage, according to the experience replay technique [13], [14], the ST randomly samples a minibatch of transitions from D to update the weights θ in the direction of minimizing the loss function, which can be defined as

$$L(\theta) = E \left[\sum_{d_i \in \mathcal{O}} (Q_e - Q(s_i, a_i; \theta))^2 \right], \quad (9)$$

where $\mathcal{O}$ is the set of sampled transitions batch from D . Q_e is a estimated Q-value with the current weight θ_0 , given as

$$Q_e = r_i + \gamma \max_a Q(s_{i+1}, a; \theta_0). \quad (10)$$

4 Simulation Results

In this section, simulation results are presented to demonstrate the system performance of our proposed DQN-based resource allocation scheme. The simulation parameters are given as follows. For the DQN, the replay memory has a capacity of $U = 1000$, the size of minibatch is 32 and the learning rate is $\alpha = 0.001$. Throughout the simulation, it is assumed that the number of channels is $M = 8$. Similar to [12], the channel gains $h_m \in \mathbf{h}$ and e are chosen from a finite set $\mathcal{H} = \mathcal{E} = \{0.05, 0.1, \dots, 2.0\}$, the channel gains g and q are chosen from a finite set $\mathcal{G} = \mathcal{Q} = \{0.01, 0.12, 0.23, \dots, 1.0\}$. In addition, $B = 10^6$ Hz, $\varepsilon = 0.1$, $\gamma = 0.9$, $p_p = 0.05$ W, $\sigma = 10^{-8}$, $R_{\min} = 2$ bits/s, $\mu = 0.5$, $E_{\max} = 0.24$ W, $a = 150$, $b = 0.014$, $P_{in} = 50$ mW and $P_{th} = 100$ mW.

Fig. 2 shows the loss function value of the Q-network in the training process. For a better presentation, we take samples

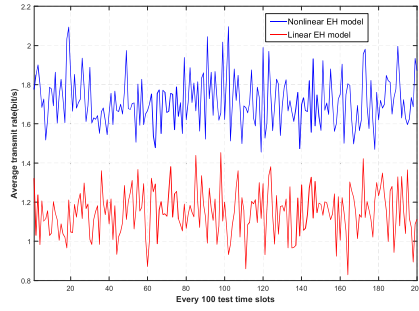


Figure 3. The comparison of the average transmit rate based on the non-linear EH model versus the linear EH model.

every 200 points from the first 200 thousand training steps. It can be seen that the loss value decreases and approaches 0 with the increase of the training steps in the training process. It demonstrates that our proposed algorithm is convergent. Based on the convergence of the loss value, the results shown in Figs. 3 are obtained from the same well trained DQN configuration.

Fig. 3 shows the comparison of the average transmit rate obtained by our proposed DQN resource allocation schemes based on the non-linear EH model with that achieved by using the linear EH model. For a better comparison, the average transmit rate is defined as the average value of every 100 transmit rates among the total 20 thousand test time slots. It can be seen from Fig. 3 that the performance achieved under the non-linear EH model given by eq. (3) is better than that obtained by using the model given by eq. (2). The reason is that the non-linear energy harvesting model considers the nonlinear nature of the practical EH circuit and its model more accurate than the linear one. The same results have been shown in [7].

5 Conclusions

An intelligent resource allocation scheme based on DQN was proposed in a wideband cognitive radio network under the practical non-linear energy harvesting model. The operation mode and the transmit power of the SUs and the sub-carrier channel were jointly optimized to maximize the achievable reward. It was shown that our proposed resource allocation scheme based on deep reinforcement learning under the non-linear EH model outperforms that achieved under the linear EH model.

References

- [1] C. Huang, *et al.*, “Machine Learning-Enabled LOS/NLOS Identification for MIMO Systems in Dynamic Environments,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 3643-3657, June 2020,.
- [2] M. Yang *et al.*, “Machine-Learning-Based Scenario Identification Using Channel Characteristics in Intelli-

gent Vehicular Communications,” *IEEE Trans. on Intell. Transportation Syst.*, to be published, 2021.

- [3] E. Ahmed, *et al.*, “Channel Assignment Algorithms in Cognitive Radio Networks: Taxonomy, Open Issues, and Challenges,” *IEEE Commun. Surveys Tuts.*, vol. 18, pp. 795-823, Oct. 2016.
- [4] Y. Wang, *et al.*, “Resource Allocation and Trajectory Design in UAV-Assisted Jamming Wideband Cognitive Radio Networks,” *IEEE Trans. Cog. Commun. Net.*, to be published, 2021.
- [5] X. Zhang, *et al.*, “Robust Resource Allocation for MISO Cognitive Radio Networks Under Two Practical Non-linear Energy Harvesting Models,” *IEEE Commun. Lett.*, vol. 22, no. 9 pp. 1874-1877, Sept. 2018.
- [6] X. Lu, *et al.*, “Wireless Networks with RF Energy Harvesting: A Contemporary Survey,” *IEEE Commun. Surveys Tuts.*, vol. 17, pp. 757-789, 2nd Quart., 2015.
- [7] E. Boshkovska, *et al.*, “Robust Resource Allocation for MIMO Wireless Powered Communication Networks Based on a Non-Linear EH Model,” *IEEE Trans. Commun.*, vol. 65, no. 5, pp. 1984-1999, Feb. 2017.
- [8] Y. S. Nasir and D. Guo, “Multi-Agent Deep Reinforcement Learning for Dynamic Power Allocation in Wireless Networks,” *IEEE J. Sel. Areas Commun.*, vol. 37, no. 10, pp. 2239-2250, Oct. 2019
- [9] J. Cui, Y. Liu and A. Nallanathan, “Multi-Agent Reinforcement Learning Based Resource Allocation for UAV Networks,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 2, pp. 729-743, Feb. 2020,.
- [10] S. Wang, *et al.*, “Deep Reinforcement Learning for Dynamic Multichannel Access in Wireless Networks,” *IEEE Trans. Cog. Commun. Net.*, vol. 4, no. 6, pp. 257-265, June 2018.
- [11] Y. Xu, *et al.*, “The Application of Deep Reinforcement Learning to Distributed Spectrum Access in Dynamic Heterogeneous Environments With Partial Observations,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 7, pp. 4494-4506, July 2020.
- [12] F. Tang, *et al.*, “Deep Reinforcement Learning for Dynamic Uplink/Downlink Resource Allocation in High Mobility 5G HetNet,” *IEEE J. Sel. Areas Commun.*, vol. 38, no. 12, pp. 2773-2782, Dec. 2020.
- [13] H. Zhang, *et al.*, “Power Control Based on Deep Reinforcement Learning for Spectrum Sharing,” *IEEE Trans. Wireless Commun.*, vol. 19, no. 6, pp. 4209-4219, June 2020.
- [14] H. H. Chang, *et al.*, “Distributive Dynamic Spectrum Access Through Deep Reinforcement Learning: A Reservoir Computing-Based Approach,” *IEEE Internet Things J.*, vol. 6, no. 2, pp. 1938-1948, Apr. 2019.