Spectrum Management in Dynamic Spectrum Access: A Deep Reinforcement Learning Approach

Hao Song

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Electrical Engineering

Lingjia Liu, Chair
Harpreet S. Dhillon
Yaling Yang

October 24, 2019
Blacksburg, Virginia

Keywords: Dynamic spectrum access, spectrum management, reinforcement learning, deep Q-network, echo state networks.

Spectrum Management in Dynamic Spectrum Access: A Deep Reinforcement Learning Approach

Hao Song

(ABSTRACT)

Dynamic spectrum access (DSA) is a promising technology to mitigate spectrum shortage and improve spectrum utilization. However, DSA users have to face two fundamental issues, interference coordination between DSA users and protections to primary users (PUs). These two issues are very challenging, since generally there is no powerful infrastructure in DSA networks to support centralized control. As a result, DSA users have to perform spectrum managements, including spectrum access and power allocations, independently without accurate channel state information. In this thesis, a novel spectrum management approach is proposed, in which Q-learning, a type of reinforcement learning, is utilized to enable DSA users to carry out effective spectrum managements individually and intelligently. For more efficient processes, powerful neural networks (NNs) are employed to implement Q-learning processes, so-called deep Q-network (DQN). Furthermore, I also investigate the optimal way to construct DQN considering both the performance of wireless communications and the difficulty of NN training. Finally, extensive simulation studies are conducted to demonstrate the effectiveness of the proposed spectrum management approach.

Spectrum Management in Dynamic Spectrum Access: A Deep Reinforcement Learning Approach

Hao Song

(GENERAL AUDIENCE ABSTRACT)

Generally, in dynamic spectrum access (DSA) networks, co-operations and centralized control are unavailable and DSA users have to carry out wireless transmissions individually. DSA users have to know other users' behaviors by sensing and analyzing wireless environments, so that DSA users can adjust their parameters properly and carry out effective wireless transmissions. In this thesis, machine learning and deep learning technologies are leveraged in DSA network to enable appropriate and intelligent spectrum managements, including both spectrum access and power allocations. Accordingly, a novel spectrum management framework utilizing deep reinforcement learning is proposed, in which deep reinforcement learning is employed to accurately learn wireless environments and generate optimal spectrum management strategies to adapt to the variations of wireless environments. Due to the model-free nature of reinforcement learning, DSA users only need to directly interact with environments to obtain optimal strategies rather than relying on accurate channel estimations. In this thesis, Q-learning, a type of reinforcement learning, is adopted to design the spectrum management framework. For more efficient and accurate learning, powerful neural networks (NN) is employed to combine Q-learning and deep learning, also referred to as deep Q-network (DQN). The selection of NNs is crucial for the performance of DQN, since different types of NNs possess various properties and are applicable for different application scenarios. Therefore, in this thesis, the optimal way to construct DQN is also analyzed and studied. Finally, the extensive simulation studies demonstrate that the proposed spectrum management framework could enable users to perform proper spectrum managements and achieve better performance.

Contents

List of Figures vi

List of Tables vii

1 Introduction 1

2 Spectrum Management Using Deep Reinforcement Learning in Distributed Dynamic Spectrum Access 5

2.1 Introduction 5

2.2 System Model 6

2.3 System Design 8

2.3.1 System procedure of interference information feedback 9

2.3.2 Interference information feedback method 10

2.4 Reinforcement learning 13

2.5 Reinforcement Learning Based Spectrum Management 15

2.5.1 Spectrum management with Q-learning 15

2.5.2 Definition of reward 18

2.5.3 Process of Q-learning based spectrum management 19

2.6 Deep Q-network Based Spectrum Management 20

2.6.1 Process of deep Q-network 21
2.6.2 Selection of neural networks 21
2.7 Simulation Results and Analysis 23
2.7.1 Simulation setup 23
2.7.2 Performance with the total interference broadcast 24
2.7.3 Performance with the specific interference feedback 27

3 Summary **30**

Bibliography **31**

List of Figures

2.1 DSA networks. 7

2.2 Received signals of a DSA user. 8

2.3 System procedure of interference information feedback. 10

2.4 Broadcast the total interference to all DSA users. 11

2.5 Feed the specific interference back to each DSA user. 12

2.6 Q-learning based spectrum managements. 20

2.7 An itaration of deep Q-network. 22

2.8 Echo state network. 23

2.9 Total reward with the total interference broadcast. 25

2.10 Total data rate with the total interference broadcast. 26

2.11 Total interference with the total interference broadcast. 27

2.12 Total reward with the specific interference feedback. 28

2.13 Total data rate with the specific interference feedback. 29

2.14 Total interference with the specific interference feedback. 29

List of Tables

2.1 Q-table in a DSA user 17

2.2 Simulation parameters. 24

Chapter 1

Introduction

According to the study and forecast released by CISCO [1], mobile data traffic is experiencing tremendous growth, which will increase sevenfold between 2016 and 2021 with a compound annual growth rate (CAGR) of 46%. This explosive growth makes spectrum resources extremely scarce and costly, since all the mobile operators seek for spectrum extension to meet mobile data traffic demand. However, practical measurements indicate that precious spectrum resources are being under-utilized. Even in some crowded urban areas, like New York and Chicago, the measured spectrum occupancy is below 40% [2]. These measurement results and statistics spur the the Federal Communication Commission (FCC) to consider that spectrum access is a more significant problem than the scarcity of spectrum. Moreover, the legacy static spectrum allocation policy is reviewed, which would limit potential spectrum users to obtain spectrum access opportunities [3]. As a result, the concept of Dynamic Spectrum Access (DSA) is come up, aiming at alleviating the problem of spectrum shortage and enhancing network capacity. Under DSA, spectrum will be intensively used, and shared among different classes of users dynamically and flexibly.

Many frequency bands have been opened up for unlicensed use, such as industrial, scientific and medical (ISM) bands, and unlicensed national information infrastructure (UNII) bands. LTE systems have been encouraged to extend their system bandwidth by accessing 5.8 GHz ISM bands, such as licensed-assisted access (LAA) and LTE-unlicensed (LTE-U) systems [4]. The corresponding enabling technologies have been widely studied, such as resource

allocations and co-existence between LTE-U and Wi-Fi [5], [6]. However, crowded Wi-Fi users and congested wireless environments have made detrimental interference occur on these bands, which are unable to accommodate more users. To cope with that, the FCC searches for more available spectrum to satisfy the demand of DSA users by expanding unlicensed bands to ultra-wideband millimeter-wave (mmWave) bands. The recent mmWave allocation policy issued by the FCC shows that 14 GHz of contiguous mmWave bands (57-71 GHz) have been opened up as unlicensed bands [7]. Unfortunately, complicated signal processing techniques and hardware are required to support effective mmWave transmissions, bringing in severe overhead for wireless communication systems. Therefore, to provide more DSA opportunities for users on superior low frequency bands, the FCC has decided to further exploit under-utilized licensed bands. For example, in 2015, an auction was held by the FCC for the secondary access on advanced wireless services (AWS-3) bands. In addition, 3.5 GHz bands from 3550 MHz to 3700 MHz will also be exploited as DSA bands [8].

Although opened licensed bands could provide the potential for spectrum extension, some technical challenges are also aroused. The first challenge that has to be addressed is the protection to primary users (PUs) or incumbent users, which hold the priority on spectrum usage. DSA users, generally as secondary users, should suppress their transmit power to protect PUs from detrimental interference. For example, on AWS-3 bands, the federal meteorological-satellite (MetSat) systems exist, which should be treated as PUs for DSA users [8]. Another big issue is that the interference coordination between DSA users is very challenging due to complicated interference environments, especially with no centralized control and no cooperation for DSA users, so-called distributed DSA networks [9]. In such a distributed DSA network, Channel State Information (CSI) information may be unavailable for DSA users, the acquisition of which requires DSA users to insert user-specific pilots, like the reference signals (RSs) in LTE systems, in their transmitted radio frames, and detect

the pilots from other users sharing the same channel with them, causing enormous overhead. Moreover, CSI estimations is unrealistic for distributed DSA networks, which cannot provide centralized measurement configurations for DSA users, since the effective detection of user-specific pilots need the synchronization among DSA users and the orthogonality among different user's pilots. As a result, in a distributed DSA network, the interference issue cannot be addressed by the traditional methods, like interference coordination [10] and interference cancellation [11], which depend upon cooperation between users or accurate channel state information (CSI) estimations for other users.

As powerful tools, applying machine learning and deep learning technologies in the wireless communication field has been widely studied to improve system performance or efficiency, such as beamforming managements [12], [13] and resource allocations [14]. However, these proposed methods are based on supervised learning, which require training data. Training data could be acquired by measurements or generated by the particular model of application scenarios. However, practical measurements are very costly, since tremendous data need to be collected and processed. Moreover, in a dynamic system like DSA networks, the model is normally unknown, as the information regarding network layout and channel states is unavailable. In this thesis, a novel framework of spectrum managements leveraging deep reinforcement learning is proposed in distributed DSA networks, enabling DSA users to learn dynamic wireless environments accurately and carry out spectrum managements appropriately through directly interacting with wireless environments, without requiring any cooperation among users and training data.

The remainder of this thesis is organized as follows. In Chapter 2, based on deep reinforcement learning, the proposed spectrum management approach in distributed DSA is elaborated, including system model, the system design of interference information feedback, the framework of the proposed spectrum management methods, as well as Simulation studies

and analysis. Then, the thesis is summarized in Chapter 3.

Notably, in the thesis, the research work mainly reproduces my published conference paper [28], and the corresponding journal paper [29] that is an extension of [28].

Chapter 2

Spectrum Management Using Deep Reinforcement Learning in Distributed Dynamic Spectrum Access

2.1 Introduction

Q-learning, a type of reinforcement learning, is utilized, the model free nature of which could enable DSA users to carry out appropriate spectrum managements, including spectrum access and power allocations, just by interacting with environments without depending on any training data [15]. Nonetheless, Q-learning cannot handle large exploration space. When the number of states and actions becomes large, it is hard for Q-learning to converge [16]. For fast convergence, neural networks (NNs) are utilized to perform Q-learning processes, including approximating the expected cumulative reward and exploring optimal state-action pairs, so-called deep Q-network (DQN), which is a type of deep reinforcement learning [17]. To the best of my knowledge, there is still no work done on using Q-learning and DQN in DSA networks to enable spectrum managements, including both spectrum access and power allocations. The key contributions of this thesis are summarized as follows:

1) A framework of spectrum managements is proposed based on deep Q-network, enabling DSA users to perform proper spectrum managements individually and intelligently without relying on accurate channel estimations and centralized control. In the proposed framework, the current spectrum management strategies, including spectrum access and power allocations, is defined as states, while the adjustment for spectrum managements is defined as actions which is conducted based on the reward obtained through interacting with environments directly.

2) I provide a comprehensive investigation of the proper way to constitute deep Q-network. The potential types of neural networks that are suitable to be applied in distributed DSA networks are discussed. Through simulations and comparison, the optimal selection of neural networks is found, which can bring in excellent performance in terms of both achievable data rate, PU protections and convergence behaviors.

2.2 System Model

As shown in Fig. 2.1, a DSA network consisting of multiple DSA users and primary users is considered, which is constructed in the distributed fashion without powerful infrastructures and centralized control support. Without loss of generality, assume that each DSA user is comprised of a transmitter (TX) and a receiver (RX), namely a DSA user pair. Each DSA user shares wireless channels with other DSA users and PUs, and opportunistically accesses wireless channels. For simplification, a reasonable assumption is made that each PU only uses one wireless channel and PUs occupy different channels to avoid making interference to each other.

The main notations are presented as follows. $\mathbf{N} = \{n | \, n = 1, 2, \cdots, N\}^T$ stands for the set of DSA users. Under the assumption that each PU only occupies one unique channel, let

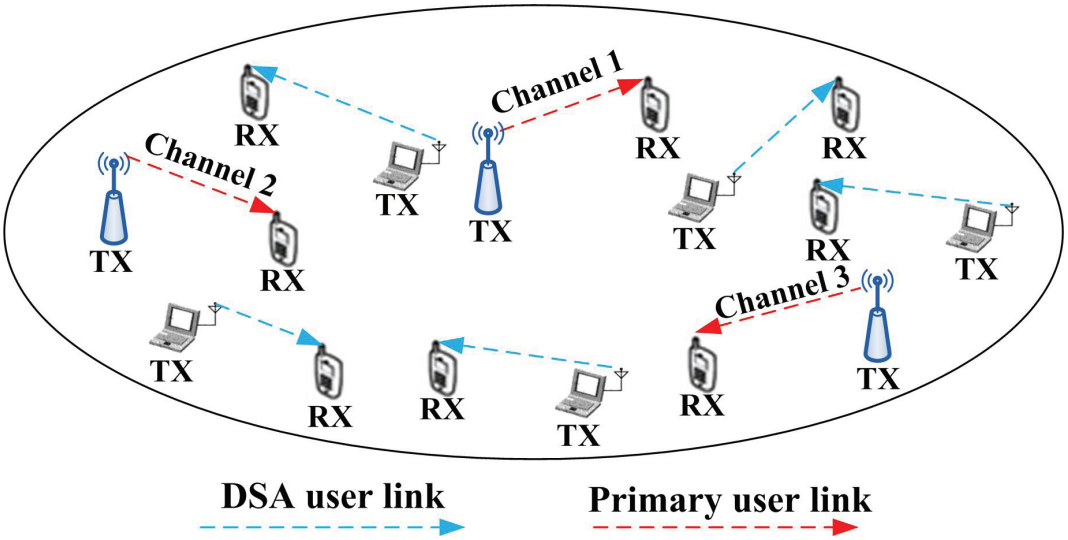

Figure 2.1: DSA networks.

$\mathbf{M} = \{m|\, m = 1, 2, \cdots, M\}^T$ be the set of both PUs and wireless channels. Additionally, m-th channel is the dedicated channel of m-th DSA user. $\Omega_n = \{m|\, m = 1, 2, \cdots, M_n\}^T$ and $\Phi_m = \{n|\, n = 1, 2, \cdots, N_m\}^T$ represent the set of the channels allocated to DSA user n and the set of the users accessing channel m, respectively.

Due to lack of centralized control in DSA networks, DSA users may suffer from the interference from both other DSA users and PUs. As shown in Fig. 2.2, the received signals of DSA user n on channel m is given by

$$y_n^m = x_n^m \cdot h_{nn}^m + x_m^m \cdot h_{mn}^m + \sum_{j \in \Phi_m, j \neq n} x_j^m \cdot h_{jn}^m + z_n^m, \tag{2.1}$$

where x_n^m denotes the desired signals sent by DSA user n on channel m. x_m^m and x_j^m stand for interference signals caused by DSA user j and PU m, respectively. Accordingly, h_{nn}^m, h_{mn}^m, and h_{jn}^m represent the channel gains of the links from the transmitter to receiver of DSA user n, from PU m to DSA user n, and from DSA user j to DSA user n, respectively. z_n^m is the received additive white Gaussian Noise (AWGN).

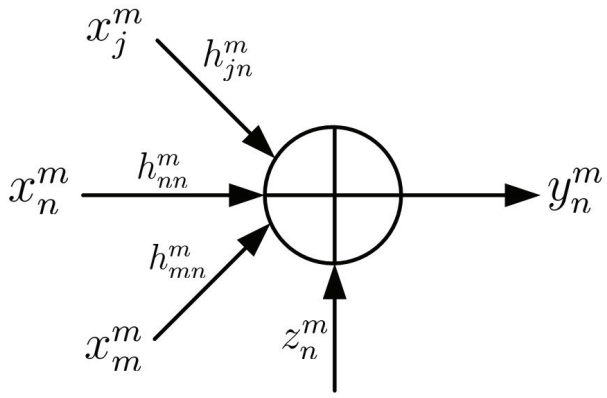

Figure 2.2: Received signals of a DSA user.

Accordingly, the signal to interference plus noise ratio (SINR) can be expressed by

$$r_n^m = \frac{p_n^m \cdot |h_{nn}^m|^2}{\underbrace{p_m^m \cdot |h_{mn}^m|^2}_{\substack{Interference\ from \\ PU\ m}} + \underbrace{\sum_{j \in \Phi_m, j \neq n} p_j^m \cdot \left|h_{jn}^m\right|^2}_{\substack{Interference\ from \\ other\ DSA\ users}} + \underbrace{B \cdot N_0}_{noise}}, \tag{2.2}$$

where p_n^m, p_m^m, and p_j^m denote transmit power of n, m, and j on channel m, respectively. B and N_0 are channel bandwidth and noise spectral density, respectively. The corresponding achievable data rate can be calculated by $B \cdot \log_2(1 + r_n^m)$.

2.3 System Design

Generally, no powerful infrastructure, like base stations (BSs) or control centers, is deployed in DSA networks to provide centralized control, so that DSA users have to carry out their spectrum managements individually. In such a network, a DSA user can only obtain very limited channel state information, namely that of the link between its own transmitter and

receiver by channel detection, while the channel state information regarding other DSA users and PUs is unavailable. As a result, it is difficult for DSA users to perform spectrum managements through resource allocation algorithms, which require accurate and sufficient channel state information. Thus, to protect PUs from harmful interference, PUs should at least provide the basic feedback of received interference to DSA users, enabling DSA users to adjusting their transmission parameters properly. However, DSA users and PUs may be operated by different mobile systems, and only limited information exchange can be achieved. Therefore, two possible interference information feedback methods of PUs are analyzed and corresponding system procedures are designed for viability.

2.3.1 System procedure of interference information feedback

The preliminary condition of effective information exchange between DSA users and PUs is the synchronization in time and frequency domain. In other words, DSA users need to know the frequency-time resource blocks that carry interference feedback information. Therefore, the system procedure of interference information feedback is designed. It is worth to note that since DSA users and PUs may be controlled by different wireless communication systems, the message exchange between them should be as few as possible to make the design interference information feedback processes less complicated and easy to realize.

Fig. 2.3 describes the process of interference information feedback. To let DSA users be aware of configurations regarding DSA, PUs should add the corresponding information in their system information (SI) and broadcast it periodically. SI is a proper carrier for DSA configurations, since SI is used to carry common control information that are fundamental and indispensable for all users to conduct wireless transmissions, and generally delivered upon fixed wireless channels [18]. Thus, in my design, when a DSA user attempts to access

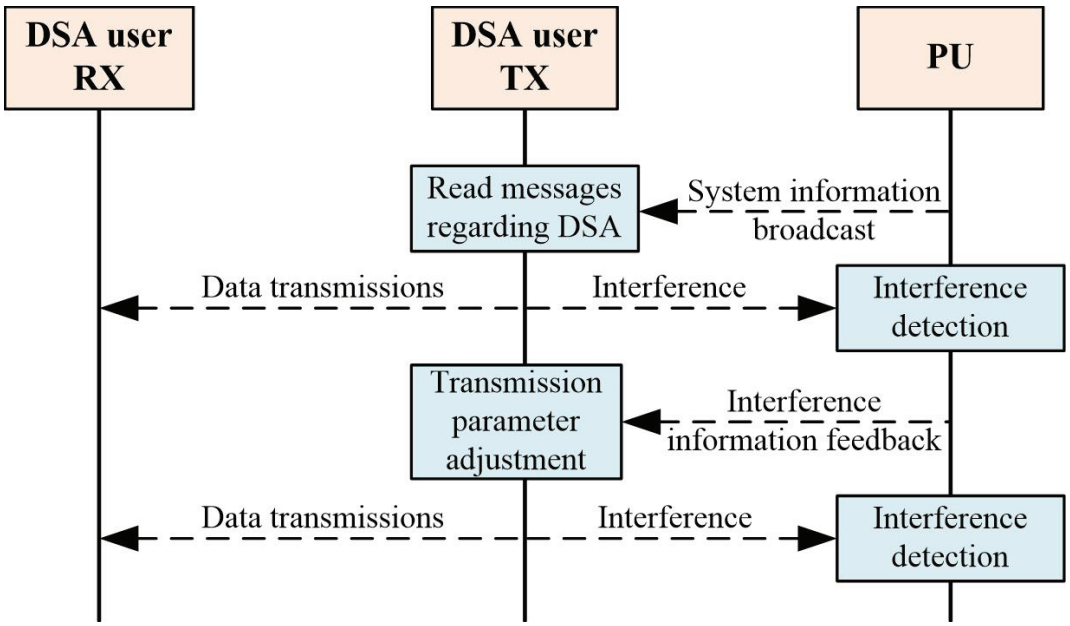

Figure 2.3: System procedure of interference information feedback.

a frequency band, it needs to receive the SI from the corresponding PUs to read DSA configurations. DSA configurations should provide the information related to transmit power constraints and the dedicated time-frequency resources to transmit interference feedback information. According to the received DSA configurations, the DSA user carries out data transmissions. Then, PUs measure received interference caused by DSA users and feed the corresponding interference information back to DSA users through the dedicated time-frequency resources indicated in DSA configurations. Based on the interference information feedback, DSA users adjust their DSA parameters to improve their own performance and guarantee PU protections.

2.3.2 Interference information feedback method

There is no doubt that DSA users would be able to make the more appropriate decision on DSA parameter adjustments if they can obtain more precise interference information

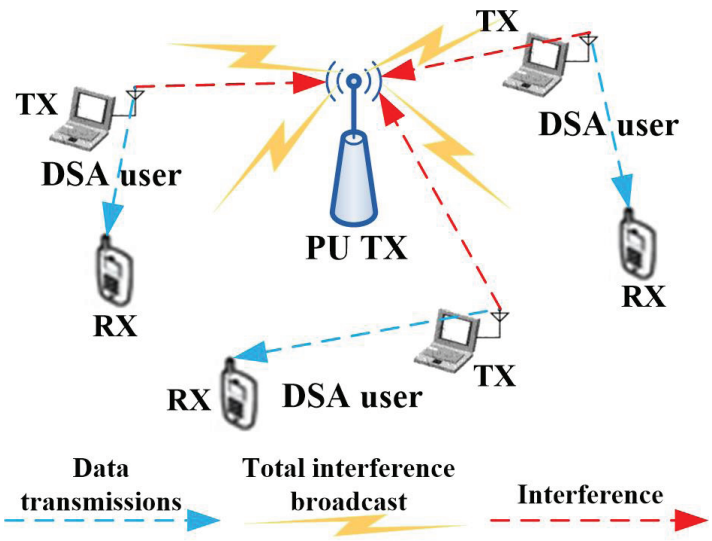

Figure 2.4: Broadcast the total interference to all DSA users.

feedback. However, the accuracy of interference information feedback is dominated by the way that PUs measure interference from DSA users. Here, based on the designed system procedure of interference information feedback, I discuss two possible methods of PUs performing interference measurements and the corresponding interference information that can be attained by DSA users.

1) Method 1: In general cases, a PU is only able to measure the total received interference, which can be realized by sensing blank time-frequency slots embedded in their occupied channels. Then, the PU broadcast the measurement results to DSA users. The method is presented in Fig. 2.4. Obviously, with this method, the overhead of interference information feedback is relatively small, while interference information that DSA users can get is really rare, only the total interference level that PUs are suffering.

2) Method 2: As shown in Fig. 2.5, PUs could identify and detect the interference caused by each individual DSA users, and feed the specific interference level back to each DSA user. Unfortunately, PU can only receive the mixed interference signals of all DSA users sharing

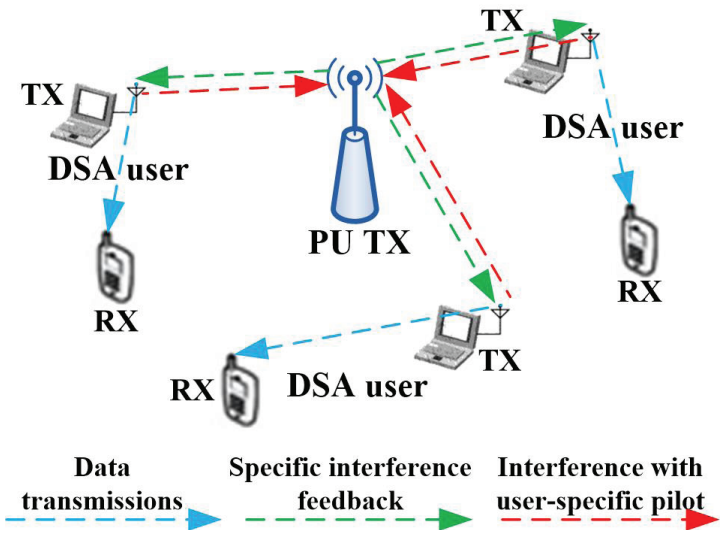

Figure 2.5: Feed the specific interference back to each DSA user.

the same channels. To distinguish the interference signals from different DSA users, each DSA user needs to be configured with user-specific pilots, by detecting which PUs can acquire the specific interference caused by different DSA users [19]. However, in DSA networks, there is no powerful infrastructure, like BSs, to conduct centralized measurement configurations for DSA users and PUs. Therefore, a low-complexity and efficient user-specific pilot assignment method is proposed, which is described as follows. To avoid pilot contamination, the user-specific pilots of different DSA users should be transmitted on different time-frequency resource blocks. A PU includes the information of unused user-specific pilots and the corresponding time-frequency resource blocks used to send different user-specific pilots in its SIs and broadcast to all DSA users. If a DSA user attempt to access the channels occupied by this PU, it needs to receive and read the PU's SI first. Then, the DSA user randomly selects a user-specific pilot and sends the chosen user-specific pilot on the corresponding time-frequency resource blocks. The PU needs to keep monitoring the time-frequency resource blocks used for user-specific pilot transmissions. If the PU notices that a user-specific pilot is transmitted on the corresponding time-frequency resource blocks, the PU should remove

the user-specific pilot from its SIs, and measure the user-specific pilot to obtain interference information. By this way, interference measurements for each particular DSA user could be achieved without relying on centralized measurement configurations supported by powerful infrastructures.

Although this method is able to provide more precise interference information feedback to DSA users, considerable overhead would also be aroused. Compared to the method 1, more time and frequency resources are consumed to perform user-specific pilot assignments, transmissions and measurements.

2.4 Reinforcement learning

Reinforcement learning is a promising machine learning paradigm, which has attracted more and more attentions in both academia and industry. With reinforcement learning, agents are able to learn which actions should be taken to yield the maximum reward without relying on labels, the acknowledged correct actions provided by authoritative external supervisor. Instead, agents need to try a variety of actions to accumulate the knowledge of rewards. Moreover, each action should be tried many times to obtain the reward knowledge associated with different states. By exploiting accumulated reward knowledge, agents will take the actions that are expected to bring in maximum rewards [15].

Q-learning is a type of reinforcement learning, which is widely used in various applications due to its model-free nature. The model-free nature makes agents learn optimal action policy directly through interacting with environments rather than investigating environment models, such as transition probability [20]. Since Q-learning uses iterative approach to update Q-value of each state and each action, a big challenge that has to be addressed is the tradeoff between exploitation and exploration. For exploitation, it is better for agents

to select actions that has been tried and found to be optimal for the current state, aiming at gaining high reward. The optimal policy π^* could be employed to guarantee exploitation, which could be expressed by

$$A_t^* = \arg \max_{A_t} Q^{\pi^*}(S_t, A_t) \tag{2.3}$$

where S_t and A_t^* represent an initial state and the corresponding action selected following the optimal policy π^*, respectively.

However, to obtain higher reward in the future, agents need to try actions that have not been experienced to accumulate reward knowledge, so-called exploration. Moreover, the exploration in Q-learning would be more important and meaningful in DSA networks with dynamic environments. This is because Q-value should be updated to adapt to the variations of wireless environments. In this thesis, the ε-greedy method is applied to take into account both exploitation and exploration, where $\varepsilon \in [0, 1]$ is the probability that agents randomly select actions regardless of Q-value [21]. The corresponding policy used for action selections is shown as

$$A_t = \begin{cases} \arg\max_{A_t} Q^{\pi^*}(S_t, A_t), \text{ with the probability of } 1-\varepsilon, \\ \text{Randomly select actions, with the probability of } \varepsilon. \end{cases} \tag{2.4}$$

Accordingly, an online Q-value update method is adopted, which is defined by

$$\begin{aligned} Q(S_t, A_t) \leftarrow\ & Q(S_t, A_t) \\ & + \alpha \cdot \left[R_{t+1} + \gamma \cdot \max_{A_{t+1}} Q(S_{t+1}, A_{t+1}) - Q(S_t, A_t) \right] \end{aligned} \tag{2.5}$$

where R_{t+1}, $\alpha \in (0, 1)$, and $\gamma \in [0, 1]$ denote the obtained reward, the learning rate, and the

discounted rate, respectively. It is noticeable that γ could be deemed as a factor to adjust the weights of immediate rewards and future rewards. If it is considered that the future reward is more important than the immediate reward, γ should be set to a relatively large value.

2.5 Reinforcement Learning Based Spectrum Management

To accomplish better performance under the condition of no centralized control and channel estimations, reinforcement learning (RL) will be employed, enabling DSA users to perform spectrum management individually and intelligently.

2.5.1 Spectrum management with Q-learning

Here, a Q-learning based spectrum management scheme is proposed. The distributed DSA will be formulated as a Q-learning problem, in which the essential components of Q-learning, including agents, states, and actions, are defined. The details of using Q-learning in spectrum managements are elaborated as follows.

1) Each DSA user will be regarded as a agent that carries out Q-learning processes independently, including action selections based on policy and Q-table updates.

2) The state in the Q-table of DSA user n is defined as a transmit power vector expressed by $S = \left(p_1, p_2, \cdots, p_{|\Omega_n|}\right)^T$, where p_i, $i = 1, 2, \cdots, M$, denotes the transmit power on i^{th} channel. To limit the size of the Q-table, the transmit power should be discretized properly. For example, if the total transmit power constraint of a user is 300 mW, its transmit power

on one channel could be discretized into 4 levels, namely 0 mW, 100 mW, 200 mW, and 300 mW.

3) The action in the Q-table is defined as a vector, indicating the change of transmit power of each channel. The vector is notated by $A = \left(a_1, a_2, \cdots, a_{|\Omega_n|}\right)^T$, where a_i, $i = 1, 2, \cdots, |\Omega_n|$, stands for the transmit power change of i^{th} channel. Considering the size of the Q-table, the number of possible actions should be restricted, therefore I only consider three sorts of transmit power changes, including increasing transmit power to next higher level, decreasing transmit power to next lower level, and no change, represented by In, De, and Un, respectively.

Based on the aforementioned configuration, a design example of a DSA user's Q-table is given as shown in Table 2.1 under the condition that 2 channels are used and the total transmit power constraint is 300 mW with 4 transmit power levels. For fairness, each DSA user accesses at least one channel, and accordingly the transmit power of at least one channel is non-zero. It should be noticed that when a user is in some specific states, some actions should avoid being chosen. For example, at state 7 (100 mW, 200 mW), taking actions 5, 6, and 9 will make the power consumption of the user exceed the total transmit power constraint. As for state 1 (100 mW, 0 mW), actions 2, 4, 8 cannot be chosen, since they will make the transmit power of the second channel become a negative value (-100mW). Therefore, for feasibility, some operating mechanisms need to be designed to tackle these issues, which are described as follows.

1) The corresponding Q-values of the inappropriate actions should be set to a very small value, represented by LR (low reward) in Table 2.1, to reduce the chance of selecting these actions.

2) If taking an action will make the transmit power vector unable to match any state in

Table 2.1: Q-table in a DSA user

	$A1:(Un,Un)$	$A2:(Un,De)$	$A3:(De,Un)$	$A4:(De,De)$	$A5:(Un,In)$	$A6:(In,Un)$	$A7:(De,In)$	$A8:(In,De)$	$A9:(In,In)$
$S1:(100mW,0mW)$	$Q(S1,A1)$	LR	LR	LR	$Q(S1,A5)$	$Q(S1,A6)$	$Q(S1,A7)$	LR	$Q(S1,A9)$
$S2:(0mW,100mW)$	$Q(S2,A1)$	LR	LR	LR	$Q(S3,A5)$	$Q(S2,A6)$	LR	$Q(S3,A8)$	$Q(S2,A9)$
$S3:(100mW,100mW)$	$Q(S3,A1)$	$Q(S3,A2)$	$Q(S3,A3)$	LR	$Q(S3,A5)$	$Q(S3,A6)$	$Q(S3,A7)$	$Q(S3,A8)$	$Q(S3,A9)$
$S4:(200mW,0mW)$	$Q(S4,A1)$	LR	$Q(S4,S3)$	LR	$Q(S4,A5)$	$Q(S4,A6)$	$Q(S4,A7)$	LR	LR
$S5:(0mW,200mW)$	$Q(S5,A1)$	$Q(S5,A2)$	LR	LR	$Q(S5,A5)$	$Q(S5,A6)$	LR	$Q(S6,A8)$	LR
$S6:(200mW,100mW)$	$Q(S6,A1)$	$Q(S6,A2)$	$Q(S6,A3)$	$Q(S6,A4)$	LR	LR	$Q(S6,A7)$	$Q(S6,A8)$	LR
$S7:(100mW,200mW)$	$Q(S7,A1)$	$Q(S7,A2)$	$Q(S7,A3)$	$Q(S7,A4)$	LR	LR	$Q(S7,A7)$	$Q(S7,A8)$	LR
$S8:(300mW,0mW)$	$Q(S8,A1)$	LR	$Q(S8,A3)$	LR	LR	LR	$Q(S8,A7)$	LR	LR
$S9:(0mW,300mW)$	$Q(S9,A1)$	$Q(S9,A2)$	LR	LR	LR	LR	LR	$Q(S9,A8)$	LR

the Q-table, the initial state is adopted as the next transited state. For example, when the initial state is state 6 (200 mW, 100 mW) and action 6 is taken, the transmit power vector will turn to (300 mW, 100 mW), which is unavailable in the Q-table. As a result, the next state will still be state 6.

Obviously, the proposed spectrum managements based on Q-learning can provide both spectrum access strategies and power allocation strategies for DSA users. Spectrum access is performed by no access to wireless channels, whose transmit power in the state represented by the transmit power vector is equal to 0. For wireless channels with non-zero elements in the state, the element values directly indicate power allocations.

2.5.2 Definition of reward

The definition of reward directly determines the performance of spectrum managements, which should consider both data rate enhancement and PU protections. In Chapter 3, two potential interference information feedback methods are discussed for future distributed DSA networks networks, based on which the reward used in Q-learning will be defined.

If the method 1 as shown in Fig. 2.4 is applied, PUs are merely able to broadcast total received interference to all the DSA users, the reward of DSA user n is defined as

$$\begin{aligned} R_n = & \sum_{m\in\Omega_n} \log_2\left(1+\frac{|h_{nn}^m|^2\cdot p_n^m}{|h_{mn}^m|^2\cdot p_m^m+\sum\limits_{j\in\Phi_m, j\neq n}\left|h_{jn}^m\right|^2\cdot p_j^m+B\cdot N_0}\right) \\ & -\kappa\cdot\sum_{m\in\Omega_n} e^{\frac{I^m}{\hat{I}^m}} \end{aligned} \tag{2.6}$$

where I^m and $\hat{I}^m$ are the total interference suffered by the PU m and the reference interference level, respectively. In equation (2.6), the first item and the second item are spectral

efficiency and the penalty regarding the interference caused to PUs, respectively. $\hat{I}^m$ could be deemed as a threshold. Once the interference received by PU m exceeds the threshold, the reward will exponentially decrease with the growth of I^m. κ is a weight to adjust the impact of the penalty on the reward function. Apparently, for a DSA user, the only external information needed to calculate the defined reward is the interference feedback from PUs, while its spectral efficiency can be monitored by itself.

Under the method 2 as shown in Fig. 2.5, A DSA user can obtain more detailed interference information from PUs, namely the specific interference caused by it. Accordingly, the reward of DSA user n is defined as

$$\begin{aligned} R_n = & \sum_{m \in \Omega_n} \log_2 \left(1 + \frac{|h_{nn}^m|^2 \cdot p_n^m}{|h_{mn}^m|^2 \cdot p_m^m + \sum\limits_{j \in \Phi_m, j \neq n} |h_{jn}^m|^2 \cdot p_j^m + B \cdot N_0} \right) \\ & - \kappa \cdot \sum_{m \in \Omega_n} e^{\frac{I_n^m}{\hat{I}_n^m}} \end{aligned} \tag{2.7}$$

where I_n^m is the interference received by PU m, which is caused by DSA user n. $\hat{I}_n^m$ denotes the reference interference level of DSA user n on channel m.

2.5.3 Process of Q-learning based spectrum management

Based on the aforementioned Q-learning configurations, the process of the corresponding spectrum managements is designed as follows. As shown in Fig. 2.6, a DSA user selects an action (transmit power changes) according to its current state (its transmit power vector) and Q-table, as well as the applied policy expressed by equation (2.4). Based on the chosen action, the DSA user adjusts its transmit power and updates its transmit power vector, based on which wireless transmissions are performed in wireless environments. Then, the updated transmit power vector will be used as the next state, and a reward is calculated based on

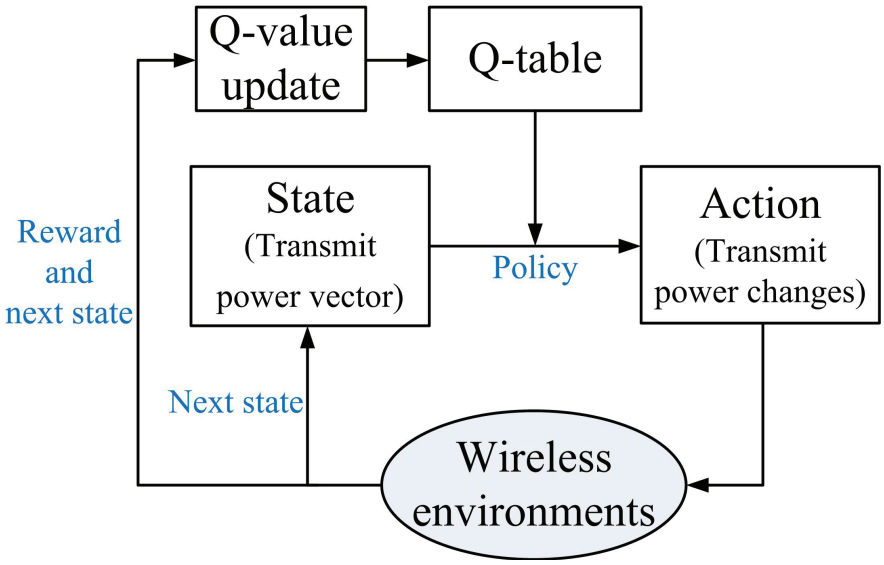

Figure 2.6: Q-learning based spectrum managements.

the performance of its wireless transmissions and the interference information fed back from PUs according to equation (2.6) or equation (2.7). Finally, the next state substitutes the initial one to be the current state of the DSA user, and the Q-value related to the initial state and the selected action in the Q-table is updated according to equation (2.5).

2.6 Deep Q-network Based Spectrum Management

From Table 2.1, it is easy to see that the size of Q-table will exponentially increase with the growth of the number of channels and the number of transmit power levels. The large size of Q-table makes Q-learning very hard or even impossible to converge [16]. Thus, neural networks (NNs) will be utilized to address this issue and support efficient Q-learning processes, so-called deep Q-network.

2.6.1 Process of deep Q-network

An iteration of DQN is depicted in Fig. 2.7. There are two NNs in DQN. The one, named evaluated NN, is used to generate Q-values of the initial state S_t for each action, $Q(S_t, A1), Q(S_t, A2), \cdots, Q(S_t, AL)$, where L is the total number of the actions defined in Q-table. Then, an action A_t is selected based on the adopted policy. After taking action A_t in environments, the corresponding reward and the next state S_{t+1} could be obtained. Another neural network, called target NN, will be utilized to update Q-value $Q(S_t, A_t)$. The next state S_{t+1} is input to the target NN to get the Q-values related to S_{t+1}, $Q(S_{t+1}, A1), Q(S_{t+1}, A2), \cdots, Q(S_{t+1}, AL)$. Based on the Q-values generated by TNN and the obtained reward, the $Q(S_t, A_t)$ is updated according to equation (2.5), which will be used as the target value to train the evaluated NN by the backpropagation method. After multiple iterations, the current ENN will be used as the new target NN to substitute the old one [17], [22].

2.6.2 Selection of neural networks

The selection of NNs is crucial for the performance of DQN, which should be based on the feature of applications. Feed-forward neural networks (FFNNs) are widely used in diverse applications because of its characteristics of simple structure and being easy to train [23]. However, in distributed DSA networks, recurrent neural networks (RNNs) may be a better choice to capture the dynamic of wireless environments. This is because the activation update in RNNs needs to take into account not only current input data, but also the previous activations of recurrent neurons and output neurons. These feedback connections make RNN capable of learning temporal correlations in dynamic systems [23], [24]. For example, a typical application of RNN is the natural language processing, since understanding a sentence

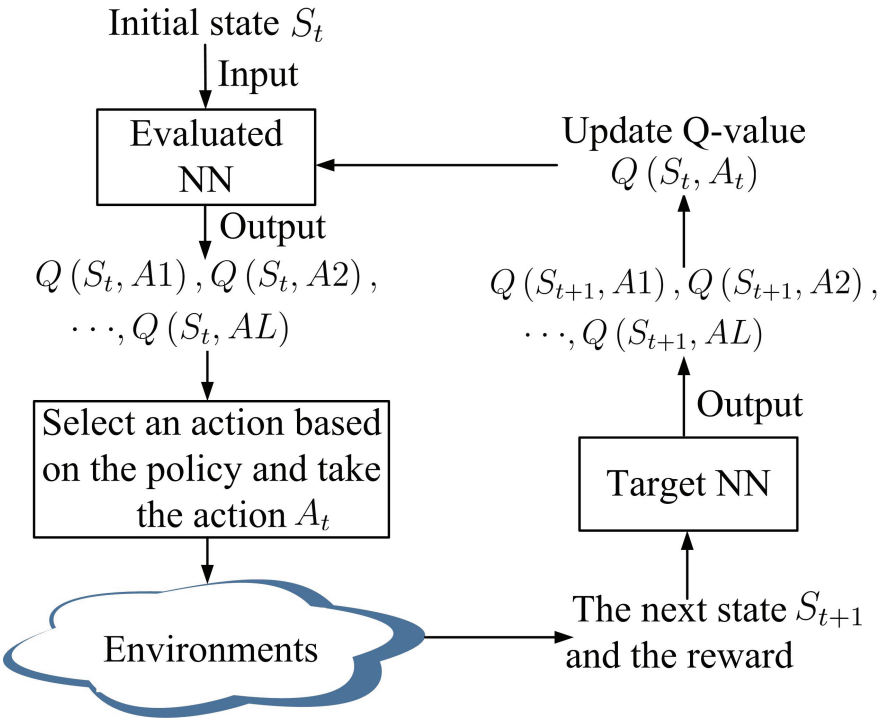

Figure 2.7: An itaration of deep Q-network.

normally needs to consider previous sentences, namely temporal correlations [25]. Similar to the natural language processing, temporal correlations also exist in the variations of wireless environments, since most of wireless devices adjust their transmission parameters following a fixed protocol, like the carrier sense multiple access with collision avoidance (CSMA/CA) in Wi-Fi systems. Unfortunately, compared to FFNN, the training of the RNN has been proven to be very difficult [16]. This barrier makes the application of RNN in distributed DSA networks more challenging. Thus, a special type of RNN, echo state networks (ESNs), will be utilized to construct DQN in this thesis. As shown in Fig. 2.8, ESN can be deemed as a simplified RNN, in which only the weights of the output layer will be trained, while other weights in the input layer and the reservoir layer are generated randomly and fixed in the training process [26]. By this way, the difficulty of training RNN could be significantly alleviated.

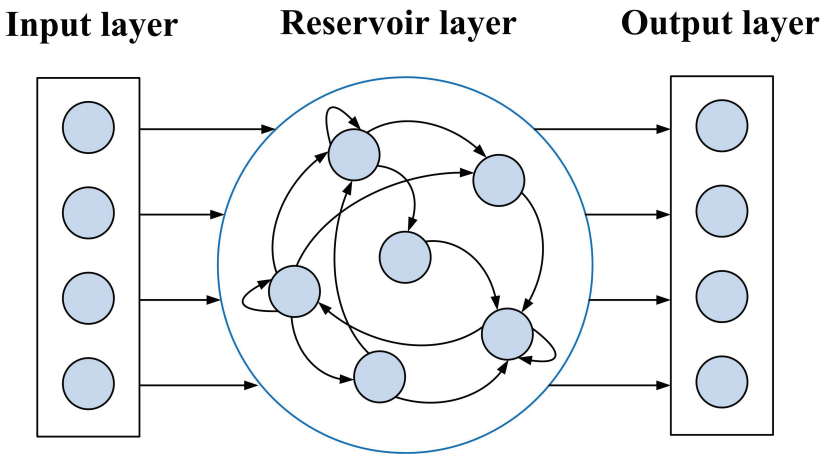

Figure 2.8: Echo state network.

2.7 Simulation Results and Analysis

By means of simulations, the performance of my proposed spectrum management scheme under two different interference information feedback methods will be investigated by extensive simulation studies. Additionally, the optimal way to constitute DQN is studied through simulations, which will take into account both system performance and the convergence of used machine learning or deep learning methods.

2.7.1 Simulation setup

A distributed DSA network with $M = 2$ wireless channels and $N = 4$ DSA users is considered, where PUs and DSA users are randomly distributed in a 150m $\times$ 150m square area. For fairness, it is assumed that each DSA user is at least access one channel and the transmit power constraint for a DSA user is 300mW. The WINNER II channel model and Rician channel model are employed to calculate channel gains [27]. According to the aforementioned analysis, the ε in the ε-greedy method is a critical parameter, which dominates

Table 2.2: Simulation parameters.

Parameters	**Values**
Transmit power of PUs	400mW
Transmit power constraint of DSA users	300mW
Channel bandwidth B	2MHz
Noise spectral density N_0	-174dBm/Hz
Center frequency	5GHz
Path-loss model (WINNER II)	$41 + 22.7 \cdot \log_{10}(d[m])$ $+20 \cdot \log_{10}(f_c[GHz]/5)$
K-factor	8
Penalty weight κ	0.3

the tradeoff between exploration and exploitation of Q-learning or DQN. In the simulation, the total number of training is 8000, in which 4000 times training is used for exploration with a relatively large ε, 0.5, facilitating DSA users to sufficiently explore all the possible spectrum management strategies. Then, ε will be adjusted to be 0 let DSA users select their spectrum management strategies with optimal rewards. The detailed simulation parameters are listed in Table 2.2. For comparison, Q-learning, FFNN based DQN, and ESN based DQN will be used to simulate the proposed spectrum management scheme, respectively. All the simulations are conducted by the Python and the Tensorflow is utilized to execute the training of neural networks. For the Q-value update in equation (2.5), the learning rate α and the discounted rate γ are set to be 0.01 and 0.9, respectively.

2.7.2 Performance with the total interference broadcast

First, the performance of the proposed spectrum management scheme is investigated under the condition that only the total interference broadcast can be provided by PUs as shown in Fig. 2.4. Accordingly, the reward of a DSA user is given by equation (2.6), and the reference interference level $\hat{I}^m$ is set to $8 * 10^{-6}$mW. Fig. 2.9 presents the total reward, which is the summation of all DSA users's, versus training steps. Obviously, ESN based DQN has the

better performance on the reward, as owning to the temporal correlation nature of ESN, DSA users can learn dynamic wireless environments better and make the more appropriate decision on spectrum managements. Besides, it can be seen that after the exploration stage the total reward of ESN based DQN becomes stable, indicating the excellent convergence behaviors of ESN based DQN.

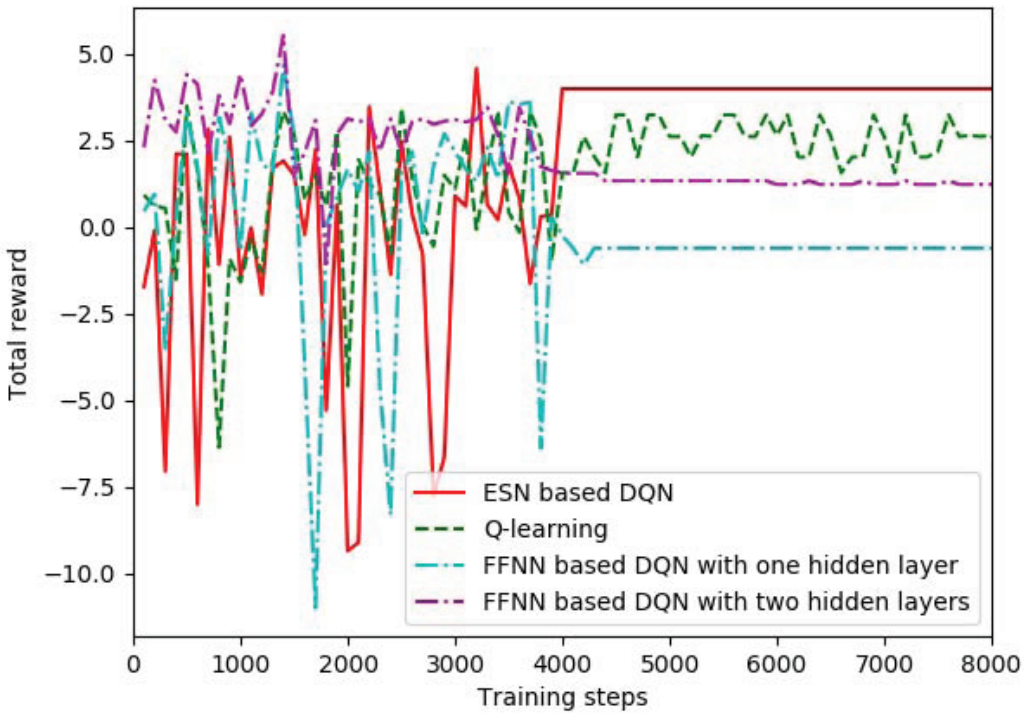

Figure 2.9: Total reward with the total interference broadcast.

Fig. 2.10 illustrates the total data rate of all DSA users with the unit of Mbits/s. Due to no centralized control, each DSA user attends to acquire more benefits in the competition with others. As a result, when a DSA user is experiencing the low reward, it may take the action of raising transmit power to boost its data rate. However, then the DSA user may encounter more severe interference from other DSA users, causing low data rate. This is because high

transmit power of a DSA user will incur more serious interference to other DSA users, which may also use the same method of rising their transmit power to preserve communication quality. Hence, the spectrum management scheme should enable DSA users to reach a balance on transmit power rather than unboundedly raising transmit power against serious interference. It can be observed from Fig. 2.10 that ESN based DQN is able to let DSA users reach a balance fast. In addition, spectrum managements with ESN based DQN could bring in higher data rate, indicating that the excellent balance is achieved among DSA users.

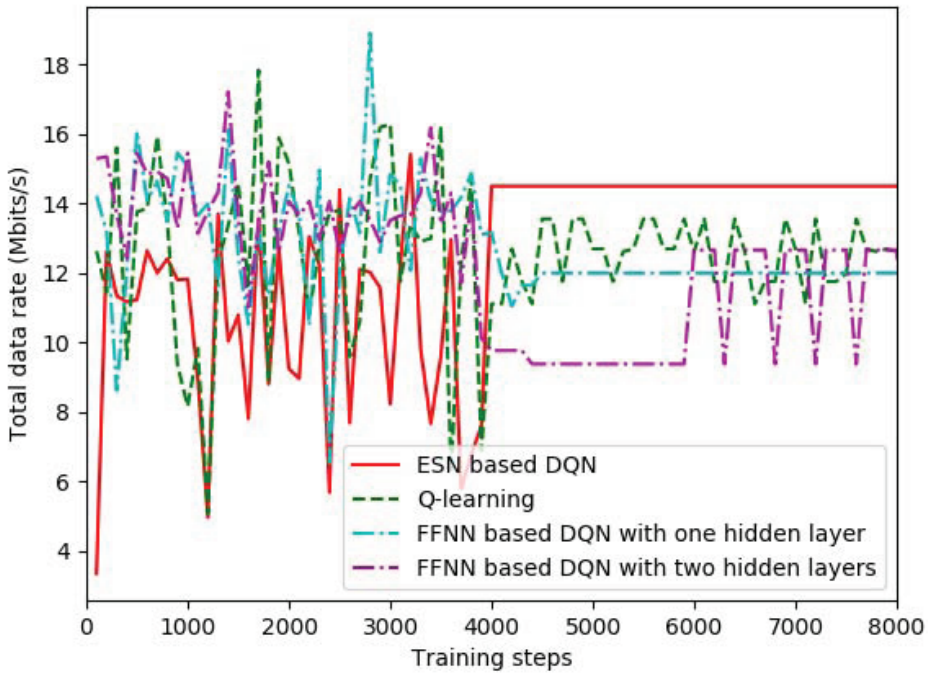

Figure 2.10: Total data rate with the total interference broadcast.

Fig. 2.11 shows the total interference to PUs caused by DSA users. According to equation (2.6), the interference is regarded as a penalty in the defined reward utility. DSA users are encouraged to lower their transmit power. Apparently, with ESN based DQN, DSA users are capable of effectively suppressing the interference to PUs in a relatively low level. The

reason is that the powerful ESN could enable DSA users to learn the interference tolerable level of PUs through interacting with environments and the received reward, so that more proper spectrum managements are performed to protect PUs from detrimental interference.

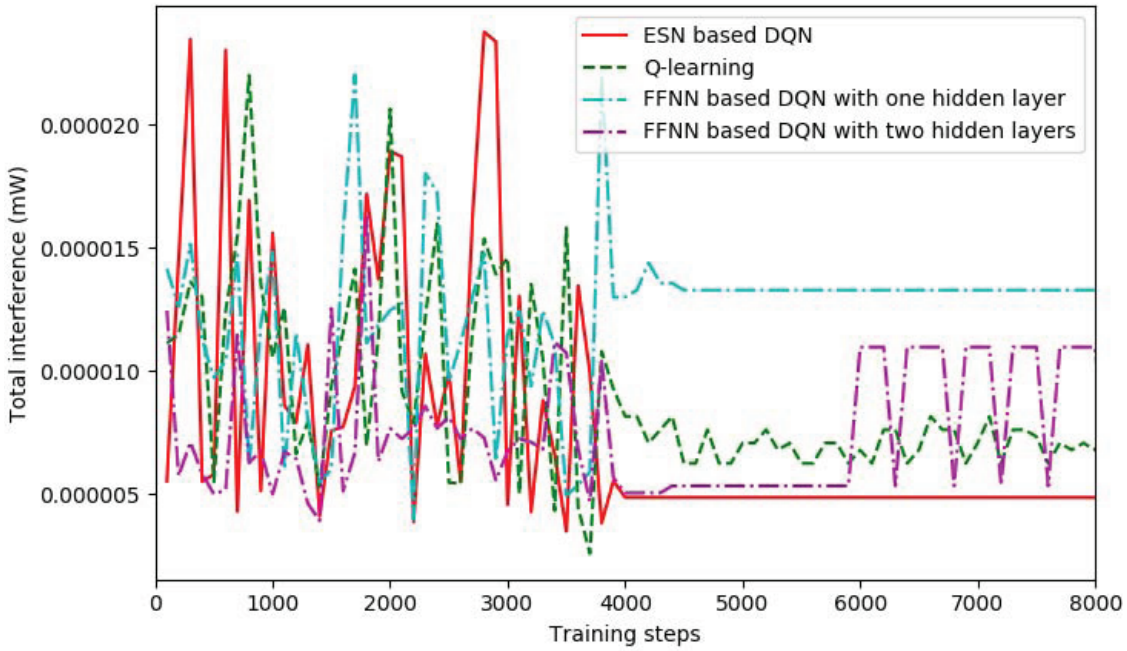

Figure 2.11: Total interference with the total interference broadcast.

2.7.3 Performance with the specific interference feedback

I also study the performance of the proposed spectrum management when DSA users can get more accurate interference feedback from PUs as shown in Fig. 2.5. In this case, the reward is calculated according to equation (2.7) and the reference interference level $\hat{I}_n^m$ is set to $2 * 10^{-6}$mW. Fig. 2.12, Fig. 2.13, and Fig. 2.14 show the simulation results of total reward, the total data rate, and the total interference, respectively. It is easy to observe that ESN based DQN can converge immediately once stepping into the exploitation

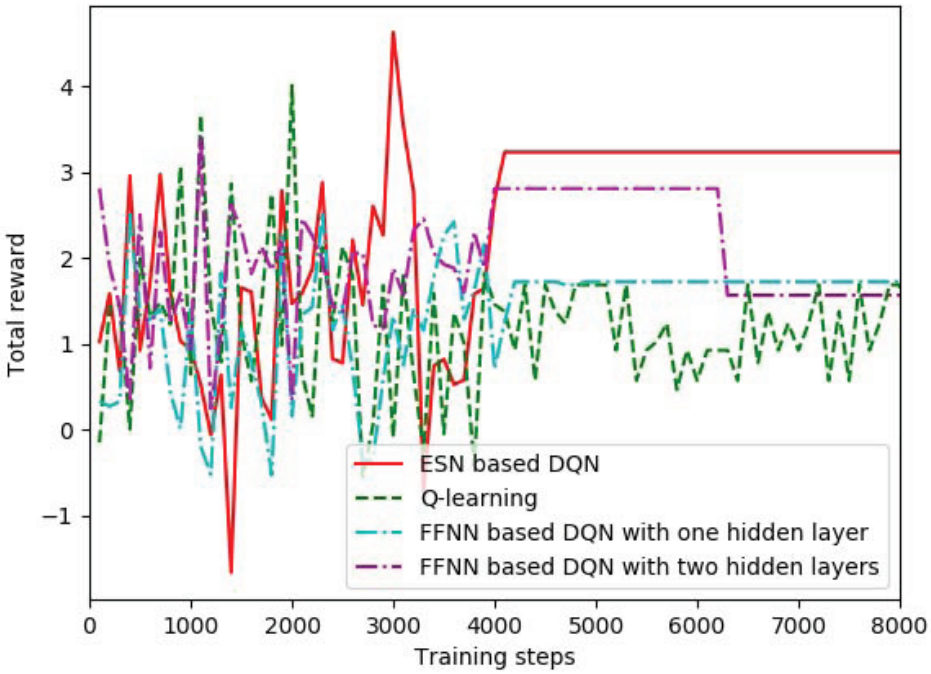

Figure 2.12: Total reward with the specific interference feedback.

stage, since ESN promote the environment learning ability of users and a excellent balance between users can rapidly be achieved. Additionally, ESN based DQN possesses the higher reward and the lower total interference than other methods. It is noted that in Fig. 2.13 the total data rate of ESN based DQN is lower than that of FFNN based DQN with two hidden layers. This phenomena manifests that ESN based DQN can make better use of interference information feedback from PUs when the feedback is more specific and detailed. For the reward enhancement, ESN based DQN mitigates the interference to PUs by reducing transmit power and sacrificing the data rate.

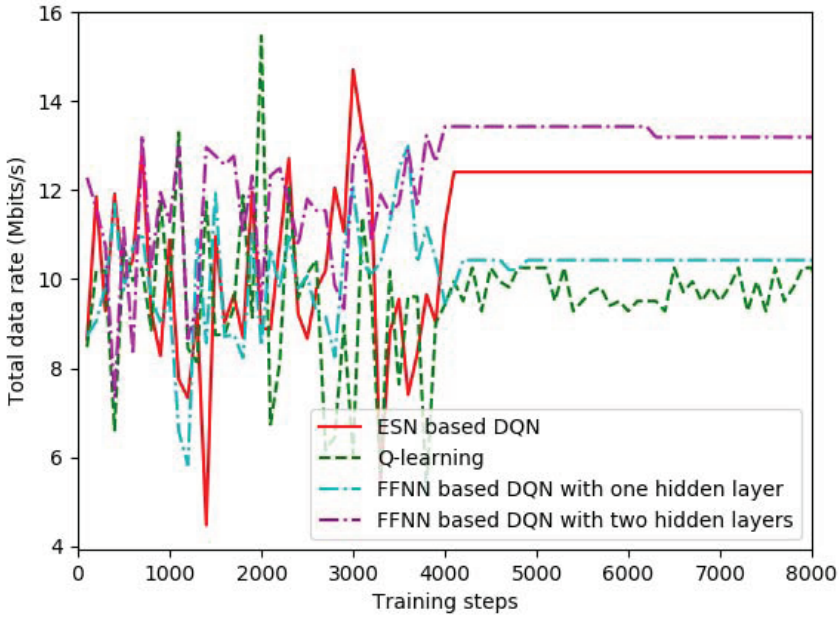

Figure 2.13: Total data rate with the specific interference feedback.

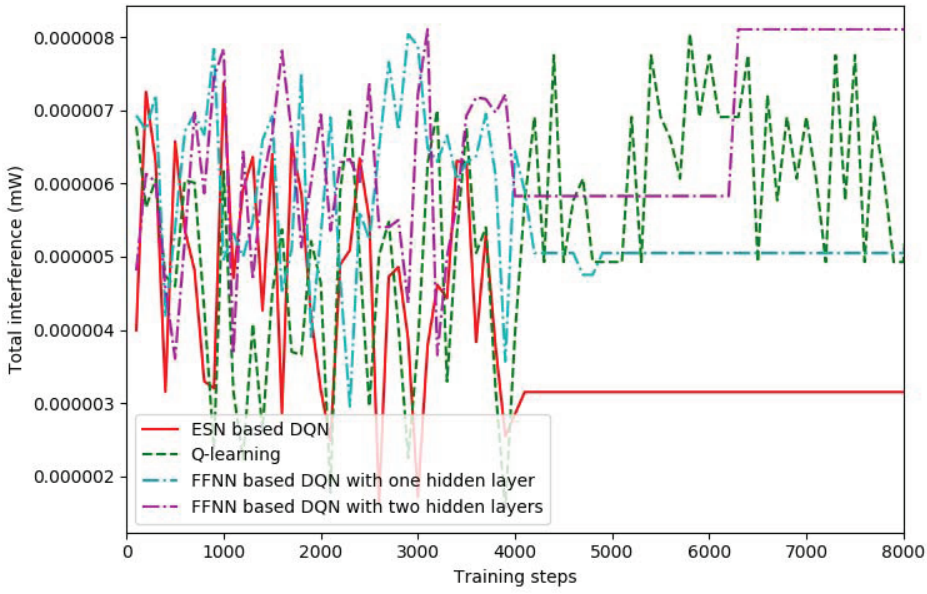

Figure 2.14: Total interference with the specific interference feedback.

Chapter 3

Summary

In a distributed DSA network, the spectrum management is very challenging, as lack of centralized control makes DSA users have to carry out spectrum managements, including spectrum access and power allocations, independently. In this thesis, a spectrum management approach leveraging Q-learning is proposed, which could enable DSA users to manage their spectrum resources just through interacting with environments without depending on channel estimations and training date. However, a new challenge is aroused that Q-learning is not able to handle a large Q-table size. In other words, the amount of wireless channels or DSA users becomes large, it is very difficult for Q-learning to converge. Thus, deep Q-network is employed to carry out my proposed spectrum management scheme, in which powerful neural networks are utilized for better performance, efficient training and fast convergence. Through extensive simulation studies, it has been proven that the proposed spectrum management scheme with ESN based DQN can achieve the higher reward with both the achievable data rate and PU protections considered. In addition, using ESN in the proposed scheme has better convergence behaviors.

Bibliography

[1] CISCO Whitepaper, "Cisco Visual Networks Index: Global Mobile Data Traffic Forecast Update, 2017-2022 White Paper," Feb 2019.

[2] Shared Spectrum Company (SSC), Available: http://www.sharedspectrum.com.

[3] Federal Communications Commission, "Spectrum Policy Task Force," Rep. ET Docket 02-135, Nov 2002.

[4] H. Song, X. Fang, L. Yan, and Y. Fang, "Control/User Plane Decoupled Architecture Utilizing Unlicensed Bands in LTE Systems," *IEEE Wireless Communications*, vol. 24, no. 5, pp. 132-142, Oct 2017.

[5] H. Song, X. Fang, and C. Wang, "Cost-Reliability Tradeoff in Licensed and Unlicensed Spectra Interoperable Networks With Guaranteed User Data Rate Requirements," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 1, pp. 200-214, Jan 2017.

[6] Hao Song and Xuming Fang, "A Spectrum Etiquette Protocol and Interference Coordination for LTE in Unlicensed Bands (LTE-U)," *IEEE International Conference on Communication Workshop (ICCW)*, London, UK, pp.2338-2343, June 2015.

[7] [Online]. Available: http://transition.fcc.gov/Daily_Releases/Daily_Business/2016/db0714/DOC-340310A1.pdf

[8] S. Bhattarai, J. J. Park, B. Gao, K. Bian, and W. Lehr, "An Overview of Dynamic Spectrum Sharing: Ongoing Initiatives, Challenges, and a Roadmap for Future Research," *IEEE Transactions on Cognitive Communications and Networking*, vol. 2, no. 2, pp. 110-128, June 2016.

[9] B. Jabbari, R. Pickholtz, and M. Norton, "Dynamic Spectrum Access and Management [Dynamic Spectrum Management]," *IEEE Wireless Communications*, vol. 17, no. 4, pp. 6-15, Aug 2010.

[10] H. Song, X. Fang, and Y. Fang, "Unlicensed Spectra Fusion and Interference Coordination for LTE Systems," *IEEE Transactions on Mobile Computing*, vol. 15, no. 12, pp. 3171-3184, Dec 2016.

[11] M. Tang, M. Vehkapera, X. Chu, and R. Wichman, "LI Cancellation and Power Allocation for Multipair FD Relay Systems With Massive Antenna Arrays," *IEEE Wireless Communications Letters*, vol. 8, no. 4, pp. 1077-1081, Aug 2019.

[12] Y. Wang, A. Klautau, M. Ribero, A. C. K. Soong and R. W. Heath, "MmWave Vehicular Beam Selection With Situational Awareness Using Machine Learning," *IEEE Access*, vol. 7, pp. 87479-87493, 2019.

[13] Y. Wang, M. Narasimha, and R. W. Heath, "MmWave Beam Prediction with Situational Awareness: A Machine Learning Approach," *2018 IEEE 19th International Workshop on Signal Processing Advances in Wireless Communications (SPAWC)*, Kalamata, pp. 1-5, 2018.

[14] S. Xu, P. Liu, R. Wang, and S. S. Panwar, "Realtime Scheduling and Power Allocation Using Deep Neural Networks," [Online]. Available: https://arxiv.org/abs/1811.07416

[15] R. Sutton and A. Barto, "Reinforcement Learning: An Introduction," *The MIT press*, Nov 2017.

[16] R. Pascanu, T. Mikolov, and Y. Bengio, "On the Difficulty of Training Recurrent Neural Networks," *ICML*, pp. 1310-1318, Feb 2013.

[17] M. Sewak, "Deep Q-Network (DQN), Double DQN, and Dueling DQN," *A Step Towards General Artificial Intelligence*, Springer, June 2019.

[18] 3GPP TS 36.331, "A Radio Resource Control (RRC) Protocol Specification," v10.0.0, Mar 2013.

[19] 3GPP TS 36.214, "Physical Layer Measurements," v11.1.0, Dec 2012.

[20] C. J. C. H. Watkins and P. Dayan, "Q-learning," *Machine Learning*, vol. 8, pp. 279–292, May 1992.

[21] E. R. Gomes and R. Kowalczyk, "Dynamic Analysis of Multiagent Q-learning with ε-greedy Exploration," *ICML '09 Proceedings of the 26th Annual International Conference on Machine Learning*, pp. 369–376, Jun 2009.

[22] H. V. Hasselt, A. Guez, and D. Silver, "Deep Reinforcement Learning with Double Q-learning," *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence (AAAI-16)*, pp. 2094-2100, 2016.

[23] I. Goodfellow, Y. Bengio, A. Courville, and Y. Bengio, "Deep Learning," *MIT press*, 2016

[24] D. P. Mandic and J. A. Chambers, "Recurrent Neural Networks for Prediction: Learning Algorithms Architecture and Stability," *Wiley*, 2001.

[25] T. Mikolov, M. Karafiát, L. Burget, J. Cernocky, and S. Khudanpur, "Recurrent Neural Network Based Language Model," *Inter-Speech*, pp. 1045-1048, Sept 2010.

[26] M. Lukosevicius, "A Practical Guide To Applying Echo State Networks," *Neural Networks: Tricks of the Trade*, Springer, pp. 659–686, 2012.

[27] P. Kyosti, "WINNER II channel models," D1.1.2, V1.2, Sep. 2007.

[28] H. Song, L. Liu, H. Chang, J. Ashdown, and Y. Yi, "Deep Q-Network Based Power Allocation Meets Reservoir Computing in Distributed Dynamic Spectrum Access Networks," *2019 IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS)*, Paris, France, pp. 774-779, 2019.

[29] H. Song, L. Liu, J. Ashdown, and Y. Yi, "Distributed Dynamic Spectrum Access Using Deep Reinforcement Learning," *Submitted to IEEE Transactions on Wireless Communications*, 2019.