Wang *et al. EURASIP Journal on Wireless Communications and Networking* (2015) 2015:130
DOI 10.1186/s13638-015-0365-7

EURASIP Journal on Wireless Communications and Networking
a SpringerOpen Journal

RESEARCH Open Access

Joint spectrum sensing and access for stable dynamic spectrum aggregation

Wei Wang[1,2*], Lingcen Wu[1], Zhaoyang Zhang[1] and Lin Chen[3]

Abstract

Spectrum aggregation is an emerging technology to satisfy the data rate requirement of broadband services for next-generation wireless communication systems. In dynamic spectrum environment, in which the spectrum availability is time-varying, it is quite challenging to maintain the stability of spectrum aggregation. In this paper, we investigate the spectrum sensing and access schemes to minimize the times of channel switching for achieving stable dynamic spectrum aggregation, taking into consideration the hardware limitations of spectrum sensing and aggregation capability. We develop an analytical framework for the joint spectrum sensing and access problem based on partially observable Markov decision process (POMDP). Especially, we derive the reward function by estimation of the stability of different spectrum sensing and access strategies. Based on the POMDP framework, we propose a rollout-based suboptimal spectrum sensing and access scheme which approximates the value function of POMDP, and propose a differential training method to improve its robustness. It is proved that the rollout policy achieves performance improvement over the basis heuristics. The simulation results show that the proposed POMDP-based spectrum sensing and access scheme improves the system stability significantly and achieves near-optimal performance with a much lower complexity.

Keywords: Cognitive radio; Spectrum aggregation; Spectrum sensing; POMDP

1 Introduction

Spectrum aggregation [1,2] enables the utilization of discrete spectrum bands or fragments to support broadband services. By spectrum aggregation, the discrete spectrum bands can provide the same transmission service as continuous spectrum bands. Recently, spectrum aggregation becomes one of the key features during LTE-advanced standardization. The performance on the system efficiency and fairness of spectrum aggregation is investigated in [3] and [4]. The energy efficiency of spectrum aggregation is also considered in [5].

The introduction of cognitive radio (CR) [6,7] increases spectrum efficiency by utilizing the spectrum dynamically, and further facilitates the application of spectrum aggregation. To exploit the instantaneous spectrum opportunities in dynamic spectrum environment, the secondary users (SUs) identify available spectrum resources by spectrum sensing and then access the available channels without interrupting primary users (PUs). *Dynamic spectrum aggregation (DSA)* provides a feasible way to support the broadband services in dynamic spectrum environment. With DSA, multiple available spectrum bands discovered via CR can be aggregated dynamically to fulfill the service requirement.

There have been a few existing publications on spectrum sensing and access schemes in dynamic spectrum environment. In [8], a decentralized MAC protocol is proposed for the SUs to sense the spectrum opportunities. The optimal sensing and channel selection are investigated to maximize the expected total number of bits delivered over a finite number of slots. In [9] and [10], the authors investigate the impacts of sensing errors on the system performance and try to alleviate their negative effects. In [11], by adopting the fusion strategy of collaborative spectrum sensing, the authors design a multi-channel MAC protocol. However, these existing works have all focused on the cases that each user uses

*Correspondence: wangw@zju.edu.cn
[1]Department of Information Science and Electronic Engineering, Zhejiang Key Lab of Information Network Technology, Zhejiang University, Hangzhou 310027, P.R. China
[2]State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, P.R. China
Full list of author information is available at the end of the article

only a single channel without considering the cases with spectrum aggregation. In [12], we propose a Maximum Satisfaction Algorithm (MSA) for admission control and a Least Channel Switch (LCS) strategy for DSA, but the spectrum sensing and access schemes are not considered jointly. In [13], we provides some preliminary results on a general POMDP framework for cognitive radio networks.

In this paper, we investigate the joint spectrum sensing and access for DSA considering several practical limitations:

- *Spectrum Sensing Limitation*: Due to the limitation of spectrum sensing capability, it is not always possible to sense all the spectrum bands for a large-span spectrum. Each SU chooses only a subset of channels (i.e., a part of the spectrum) to sense. As a result, the system is lack of the perfect information on channel availability, which brings new technical challenges for DSA.
- *Spectrum Aggregation Limitation*: Due to the hardware capability, only the spectrum bands within a certain range can be aggregated together for a single user. The spectrum aggregation range leads to an additional constraint when the SUs access the spectrum.
- *Channel Switch Overhead*: When an SU adjusts his access strategy and switches to other channels, it is unavoidable for the channel switch to result in extra system overhead, such as rendezvous, synchronization, etc. When designing the spectrum sensing and access scheme with the overhead consideration, it is necessary to reduce as many times of channel switch as possible.

Taking the above practical issues into consideration, we propose a decision-theoretic approach by casting the design of joint spectrum sensing and access for stable DSA in the *partially observable Markov decision process (POMDP)* [14] framework. In order to provide the reward function for the POMDP framework, the probability of channel switch is estimated based on Markov chain. Since the optimal solution of POMDP is very intensive computationally due to the *curse of dimensionality*, i.e., the computational time and storage requirements grow exponentially with the number of channels. We further introduce an approximation technique called rollout [15] to design the suboptimal joint spectrum sensing and access scheme. A heuristics is proposed first as the base policy, which can greedily choose the spectrum sensing and access actions to reduce the channel switch times. By rolling out the base policy, the proposed rollout algorithm can approximately calculate the value function defined in the POMDP framework and reduces the times of channel switch. A theoretical analysis is provided to prove the performance improvement of rollout policy over the heuristics. Furthermore, we propose a differential training method which reduces the sensitivity to approximation errors. The performance of the proposed scheme is evaluated by simulation which demonstrates that the proposed policies in the POMDP framework reduce the times of channel switch significantly, and the rollout-based scheme achieves a near-optimal performance compared to the optimal scheme.

The rest of this paper is organized as follows. Section 2 describes the system model and formulates the problem. Section 3 introduces the POMDP framework and the approach to estimate the access and switching probabilities. In Section 4, the rollout-based suboptimal spectrum sensing and access schemes are proposed. Section 5 provides the performance evaluation by simulation. Finally, Section 6 summarizes this paper.

2 System model and problem formulation

2.1 Dynamic spectrum aggregation model

We consider a large-span licensed spectrum consisting of N channels, which have the same bandwidth BW. Time is slotted and the duration of each time slot is T_p. The availabilities of channels, which depends on the PU activities, are modeled as the following assumption:

Assumption 1 (Channel Availability)**.** *The availabilities of N channels compose a system which can be modeled as a discrete-time Markov process with* 2^N *states,*

$$\mathbf{S}(t) = [S_1(t), \ldots, S_N(t)] \in \mathfrak{S} = \{0,1\}^N, \tag{1}$$

where $S_n(t) \in \{0(occupied), 1(idle)\}$ *denotes the occupancy state of channel* $n \in \{1, \ldots, N\}$ *at time slot t, which is independent over channels.* □

Denote p_{ij} as the transition probability from state i to state j, i.e.,

$$p_{ij} = \Pr\{\mathbf{S}(t+T_p) = j | \mathbf{S}(t) = i\}, \forall i, j \in \mathfrak{S}, \tag{2}$$

which can be obtained by multiplying the transition probability of each channel[a]

$$p_{ij} = \prod_{n=1}^{N} \Pr\{S_n(t+T_p) = j_n | S_n(t) = i_n\}, \tag{3}$$

where $i_n \in \{0,1\}$ and $j_n \in \{0,1\}$ are the nth element of the system states i and j, respectively. For simplicity of expression, we denote $P_n(t) = \Pr\{S_n(t) = 1\}$.

The SUs sense the presence of PUs and access the spectrum opportunistically in a decentralized manner. Here, we consider the spectrum sensing and access of a single SU[b]. At the beginning of each time slot t, the SU chooses a set of channels $A_1(t)$ to sense.

Assumption 2 (Spectrum Sensing)**.** *Due to the spectrum sensing capability, the SU can only sense at most L channels, which means the size of $A_1(t)$ is no more than L, i.e., $|A_1(t)| \leq L$. When $L < N$, the SU only obtains the availability information of a subset of channels.* □

Note that although L channels can be sensed by the SU, the availability states of these L channels are not always accurate because of the existence of sensing errors. The SU performs a binary hypotheses test:

- $\mathcal{H}_0$: Null hypothesis indicating that the sensed channel is available.
- $\mathcal{H}_1$: Alternative hypothesis indicating that the sensed channel is occupied.

If the SU obtains an incorrect sensing result $\mathcal{H}_1$ when the channel state is $\mathcal{H}_0$, i.e., false alarm, the SU will refrain from transmitting and a spectrum opportunity is wasted. On the other hand, if the SU obtains an incorrect sensing result $\mathcal{H}_0$ when the channel state is $\mathcal{H}_1$, i.e., miss detection, the SU will collide with a PU. Let P_f and P_m denote the probabilities of false alarm and miss detection, respectively.

Based on the spectrum sensing results, the SU aggregates a set of channels A_2 for the data transmission with spectrum aggregation.

Assumption 3 (Spectrum Aggregation)**.** *Due to the spectrum aggregation limitation, the SU can only aggregate the channels within Γ, which means that the channels in $A_2(t)$ are within the frequency range Γ, i.e., $D(i,j) \leq \Gamma, \ \forall i,j \in A_2(t)$, where $D(i,j)$ indicates the frequency distance between channel i and channel j. The total bandwidth of the available channels in $A_2(t)$ should satisfy the SU's bandwidth requirement, denoted as Υ.*

We illustrate $A_1(t)$ and $A_2(t)$ in a large-span spectrum in Figure 1.

2.2 Problem formulation

The SU can detect the return of PUs and utilize the channels unoccupied by PUs to avoid the interference to PUs. To satisfy the bandwidth requirement Υ in the dynamic spectrum environment, the SU adopts different spectrum sensing and access strategies according to the number of the current available channels $R(t)$ within $A_2(t)$ for time slot t, which is defined as $R(t) = \sum_{n \in A_2(t)} S_n(t)$.

- If $R(t) \geq \Upsilon/BW$, the SU reselects only $A_1(t)$. The spectrum aggregation decision does not change, i.e., $A_2(t) = A_2(t - T_p)$.
- If $R(t) < \Upsilon/BW$, the SU has to reselect both $A_1(t)$ and $A_2(t)$ and trigger a *channel switch*.

With the consideration of reducing the system overhead and maintaining the stability of dynamic spectrum aggregation, our aim is to minimize the expected times of channel switches[c] by adjusting the spectrum sensing and access strategies, $A_1(t)$ and $A_2(t)$. Denote $\eta(t)$ as the expected times of channel switches from time slot 0 to t. The joint spectrum sensing and access optimization problem for stable DSA can be formulated formally as follows:

$$\begin{aligned} \min_{A_1,A_2} \lim_{t\to\infty} \frac{\eta(t)}{t} \\ s.t. \ \ |A_1(t)| &\leq L \ \ \forall t \\ D(i,j) &\leq \Gamma, \ \ \forall i,j \in A_2(t), \forall t \\ R(t) &= \sum_{n \in A_2(t)} S_n(t) \geq \frac{\Upsilon}{BW}, \ \ \forall t \end{aligned} \tag{4}$$

The first two constraints indicate the spectrum sensing and spectrum aggregation limitations respectively, and the last constraint guarantees the satisfactory of the SU's bandwidth requirement.

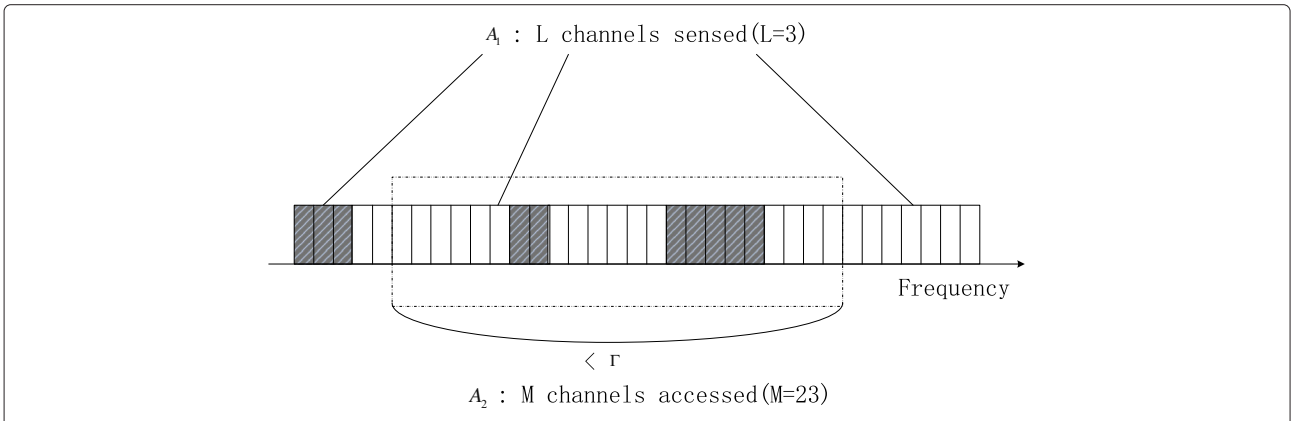


Figure 1 Illustration of spectrum sensing and access.

3 A POMDP framework for dynamic spectrum aggregation

In this section, we propose a decision-theoretic framework for DSA based on POMDP [14]. Especially, we convert minimizing the times of channels switches into a new objective, i.e., maximizing the time interval of channel switches, and provide an approach to estimate this interval as the reward of the POMDP model, which is challenging in dynamic spectrum environment.

If the SU is able to sense the whole spectrum accurately in the network, all the elements of $\mathbf{S}(t)$ can be obtained and the optimization problem is a standard Markov decision process (MDP) since the channel availability states $\mathbf{S}(t)$ is a discrete-time Markov process. However, in the practical situation with the limitation of spectrum sensing ability and the existence of sensing errors, the SU can only obtain the imperfect occupancy states of a part of channels, which means $L < N$ and $\mathbf{S}(t)$ is partially and inaccurately observable. As a result, we need to cast the optimization problem into the POMDP framework, which is a particular case of MDP in which the state of the system is partially observed by the decision maker.

3.1 POMDP framework

Before the discussion of POMDP framework, we first introduce a new concept called control interval, each of which is composed of a number of consecutive time slots and delimited by channel switches. It is obvious that the length of a control interval is uncertain depending on how long time the current aggregated channels keep satisfying the SU's bandwidth requirement. Incorporating the control interval structure, the joint spectrum sensing and access scheme are designed based on the POMDP framework, and the framework in [8] is no longer suitable.

Let T denotes the number of control intervals within the whole time horizon (finite number of time slots), and the index m indicates the mth last control interval. Denote $t_s(m)$ as the time slot when the mth channel switch is triggerred. If the current control interval includes κ time slots, the state transition probability at the next control interval is

$$p_{ij}^{\kappa} = \Pr\{\mathbf{S}(m-1) = j | \mathbf{S}(m) = i\}. \tag{5}$$

Now, we define the key components of the POMDP framework for DSA. For simplicity of expression, we adopt $A_1(m)$ and $A_2(m)$ instead of $A_1(t_s(m))$ and $A_2(t_s(m))$, respectively.

Action The actions of the SU have two stages: determining $A_1(m)$ to sense and $A_2(m)$ to access. Define $a(m)$ as the SU action for the mth last control interval,

$$a(m) = \{A_1(m); A_2(m)\} = \{C_1, C_2, \ldots, C_L; C_{start}\}, \tag{6}$$

where C_i is the index of the ith sensed channel, $\forall i \in \{1, \ldots, L\}$ and C_{start} is the starting index of the accessed channel set $A_2(m)$. Define $\mathfrak{A}$ as the set of all possible actions, i.e., $a(m) \in \mathfrak{A}$.

Observation By sensing the channels in $A_1(m)$, the SU can obtain their occupancy states inaccurately. Let $\Theta_{i,A_1}(m)$ denotes the sensing results based on the current system state i and the sensed channel set $A_1(m)$. The observing output in the mth last control interval is expressed as

$$\Theta_{i,A_1}(m) = \{\boldsymbol{\theta}_{C_1}(m), \boldsymbol{\theta}_{C_2}(m), \ldots, \boldsymbol{\theta}_{C_L}(m)\} \tag{7}$$

where $\boldsymbol{\theta}_{C_j}(m) \in \{0, 1\}$, $\forall C_j \in A_1(m)$. Although $\boldsymbol{\theta}_{C_j}(m)$ may be different to the actual channel availability state $S_{C_j}(m)$ due to the sensing errors, they are correlated and $\boldsymbol{\theta}_{C_j}(m) \in \{0, 1\}$ provides useful information for estimating $S_{C_j}(m)$. Specifically, the conditional probabilities of the channel states can be calculated by the Bayes rule [16]. If $\boldsymbol{\theta}_{C_j}(m) = 0$, we have

$$\begin{aligned} &\Pr\{S_{C_j}(m) = 1 | \boldsymbol{\theta}_{C_j}(m) = 0\} \\ &= \frac{\Pr\{S_{C_j}(m) = 1, \boldsymbol{\theta}_{C_j}(m) = 0\}}{\Pr\{\boldsymbol{\theta}_{C_j}(m) = 0\}} \\ &= \frac{\Pr\{\boldsymbol{\theta}_{C_j}(m) = 0 | S_{C_j}(m) = 1\} \cdot \Pr\{S_{C_j}(m) = 1\}}{\Pr\{\boldsymbol{\theta}_{C_j}(m) = 0\}} \end{aligned} \tag{8}$$

where $\Pr\{\boldsymbol{\theta}_{C_j}(m) = 0 | S_{C_j}(m) = 1\} = P_f$, $\Pr\{S_{C_j}(m) = 1\} = P_{C_j}(m)$, and

$$\begin{aligned} &\Pr\{\boldsymbol{\theta}_{C_j}(m) = 0\} \\ &= \Pr\{S_{C_j}(m) = 1\} \cdot \Pr\{\boldsymbol{\theta}_{C_j}(m) = 0 | S_{C_j}(m) = 1\} \\ &\quad + \Pr\{S_{C_j}(m) = 0\} \cdot \Pr\{\boldsymbol{\theta}_{C_j}(m) = 0 | S_{C_j}(m) = 0\} \\ &= P_{C_j}(m) P_f + (1 - P_{C_j}(m))(1 - P_m) \end{aligned} \tag{9}$$

Thus, we have

$$\begin{aligned} &\Pr\{S_{C_j}(m) = 1 | \boldsymbol{\theta}_{C_j}(m) = 0\} \\ &= \frac{P_f P_{C_j}(m)}{P_{C_j}(m) P_f + (1 - P_{C_j}(m))(1 - P_m)} \end{aligned} \tag{10}$$

Similarly, if $\boldsymbol{\theta}_{C_j}(m) = 1$, we can obtain that

$$\begin{aligned} &\Pr\{S_{C_j}(m) = 1 | \boldsymbol{\theta}_{C_j}(m) = 1\} \\ &= \frac{(1 - P_f) P_{C_j}(m)}{P_{C_j}(m)(1 - P_f) + (1 - P_{C_j}(m)) P_m} \end{aligned} \tag{11}$$

Belief Vector In the optimization problem (4), the channel availability state $\mathbf{S}(m)$ is partially and inaccurately observed, which means the internal states of the system cannot be obtained specifically. Consequently, we introduce an important metric $\Delta(m)$ called belief vector to represent the SU's estimation of the system states based on the past decisions and observations as

$$\Delta(m) = (\delta_i(m))_{i\in\mathfrak{S}} \tag{12}$$

where $\delta_i(m) = \Pr\{\mathbf{S}(m) = i|H(m)\}$ and $H(m) = \{a(i), \Theta(i)\}_{i\geq m}$.

The 2^N-dimensional belief vector $\Delta(m)$ is updated according to the action and the observation in the last control interval:

$$\Delta(m-1) = \Omega(\Delta(m)|a(m), \Theta(m)) = (\delta_j(m-1))_{j\in\mathfrak{S}} \tag{13}$$

where $\Omega(\cdot|a(m), \Theta(m))$ indicates the update operator. The updated belief vector can be calculated by the Bayes rule, which also depends on the length of the control interval under consideration,

$$\delta_j^{\kappa}(m-1) = \frac{\sum\limits_{i} \delta_i(m) p_{ij}^{\kappa} \Pr\{\Theta_{i,A_1}(m) = \theta\}}{\sum\limits_{i,j} \delta_i(m) p_{ij}^{\kappa} \Pr\{\Theta_{i,A_1}(m) = \theta\}} \tag{14}$$

in which $\Pr\{\Theta_{i,A_1}(m) = \theta\}$ can be obtained through the information provided by Equations (10) and (11).

It has been proved in [14] that the belief vector $\Delta(m)$ is a sufficient statistics for determining the optimal actions for future control intervals.

Policy Denote the policy vector as $\pi = [\mu_1, \mu_2, \ldots, \mu_T]$, where μ_m is a mapping from a belief vector $\Delta(m)$ to an action $a(m)$ for each control interval. The set of all possible policies is denoted as Π, i.e., $\pi \in \Pi$.

$$\mu_m : \Delta(m) \in [0,1]^{2^N} \rightarrow a(m) = \{A_1(m)\ A_2(m)\} \tag{15}$$

A policy is said to be stationary if the mapping μ_m only depends on the belief vector $\Delta(m)$ and is independent to the number of remaining control intervals m. Denote the set of stationary policies as Π_s, and it is usual to restrict the set of policies to Π_s in POMDP. In our framework, a spectrum sensing and access scheme are essentially a policy of this POMDP.

Reward To quantify the SU's objective, we define the reward of a control interval as the length of this control interval, i.e., the number of time slots it includes, denoted as $U(m)$. We now demonstrate that minimizing the number of channel switches equals to maximizing the average reward. For a given total number of time slots t, we have

$$\eta(t) = \min\left\{ T \left| \sum_{m=1}^{T} U(m) \geq t \right. \right\}. \tag{16}$$

It then follows that

$$\arg\min_{\pi} \lim_{t\to\infty} \frac{\eta(t)}{t} = \arg\max_{\pi} \lim_{T\to\infty} \frac{\sum_{m=1}^{T} U(m)}{T} \tag{17}$$

which means that over the finite time horizon, the longer the control intervals are, the less expected total times of channel switches will occur, and our objective can be converted into maximizing the average reward.

For control interval m, a set of accessed channels $A_2(m)$ is determined according to the belief vector $\Delta(m)$. To evaluate the reward of $A_2(m)$, we first define the access probability and the switching probability as follows.

Definition 1 (Access Probability)**.** *The access probability indicates the probability that the bandwidth of the available channels in $A_2(m)$ is more than the number of required channels Υ, i.e.,*

$$\zeta(a(m)) = Pr\{R(t_s(m)) \geq \Upsilon | A_2(m)\}. \tag{18}$$

Definition 2 (Switching Probability)**.** *The switching probability indicates the probability that the bandwidth of the available channels in $A_2(m)$ at the next time slot is not more than the number of required channels Υ, i.e.,*

$$\xi(a(m)) = Pr\{R(t_s(m) + T_p) < \Upsilon | A_2(m)\}. \tag{19}$$

Both the access probability ζ and the switching probability ξ can be calculated based on the sensing and access action a, which will be discussed with details in the next subsection. We omit a in the notations of both probabilities for simplicity of expression.

The reward $U(m)$ is a function of the sensing and access action a when the system state is j in the mth last control interval, which is a Bernoulli random variable with probability density function (p.d.f.) $p(\kappa)$ ($\kappa \in \mathbb{Z}^+$) derived as follows:

$$p(\kappa) = \zeta \cdot (1-\xi)^{\kappa-1} \cdot \xi \tag{20}$$

Using Equations (14) and (20), we can update the belief vector as

$$\begin{aligned} \delta_j(m-1) &= \sum_{\kappa} p(\kappa) \delta_j^{\kappa}(m-1) \\ &= \sum_{\kappa} p(\kappa) \frac{\sum\limits_{i} \delta_i(m) p_{ij}^{\kappa} \Pr\{\Theta_{i,A_1} = \theta\}}{\sum\limits_{i,j} \delta_i(m) p_{ij}^{\kappa} \Pr\{\Theta_{i,A_1} = \theta\}} \end{aligned} \tag{21}$$

In summary, our goal in the POMDP framework is to find the optimal policy π^* to maximize the average reward as follows:

$$\pi^* = \arg\max_{\pi\in\Pi} \lim_{T\to\infty} \frac{1}{T} \sum_{m=1}^{T} \mathbb{E}_{\pi} [U(m)|\, \Delta(m)] \tag{22}$$

and the POMDP framework for joint spectrum sensing and access for DSA is illustrated in Figure 2, which also indicates the Markovian dynamics of the system.

3.2 Estimation of access probability and switching probability

In order to obtain the reward function of POMDP, the access probability ζ and the switching probability ξ need to be estimated.

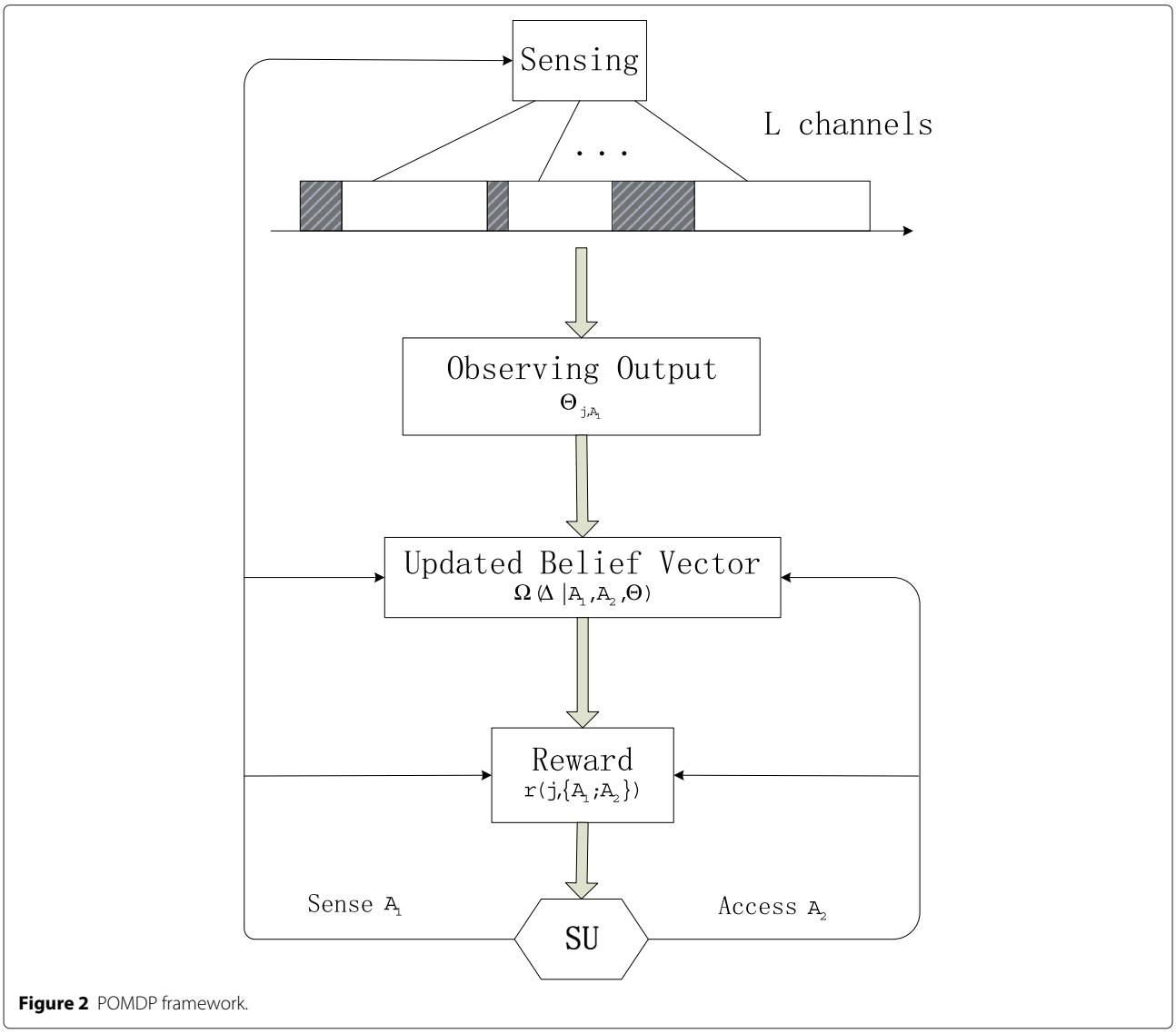


Figure 2 POMDP framework.

For the channels in the access set $A_2(m)$, we have the state vector $[S_{C_{start}}, S_{C_{start}+1}, \ldots, S_{C_{start}+M-1}]$. All the states in the vector can be taken as independent random variables, with $\mu_i = P_i$ and $\sigma_i^2 = P_i - P_i^2$ for $i \in \{C_{start}, C_{start}+1, \ldots, C_{start}+M-1\}$. According to the central limit theorem [17], we have

$$R(t) = \sum_{i=C_{start}}^{C_{start}+M-1} S_i(t) \sim N(\mu, \sigma^S) \tag{23}$$

where $\mu = \sum_{i=C_{start}}^{C_{start}+M-1} P_i$ and $\sigma^2 = \sum_{i=C_{start}}^{C_{start}+M-1} (P_i - P_i^2)$, as illustrated in Figure 3.

Based on the distribution of $R(t)$, we calculate the access probability ζ and the switching probability ξ in the following two propositions.

Proposition 1 (Calculation of Access Probability)**.** *If the spectrum sensing obtains the accurate channel availability information of all channels, the access probability ζ is*

$$\zeta = \begin{cases} 1 & if\ R \geq \frac{\Upsilon}{BW} \\ 0 & if\ R < \frac{\Upsilon}{BW} \end{cases} \tag{24}$$

Otherwise, the access probability ζ is calculated as

$$\zeta = \int_{\frac{\Upsilon}{BW}}^{M} \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(R-\mu)^2}{2\sigma^2}} dR \tag{25}$$

Proof. In the case that the spectrum sensing obtains the accurate channel availability information of all channels, R is a deterministic variable and Equation (24) can be obtained easily. On the other hand, if the spectrum sensing is incomplete or inaccurate, R is a random variable, whose p.d.f. is

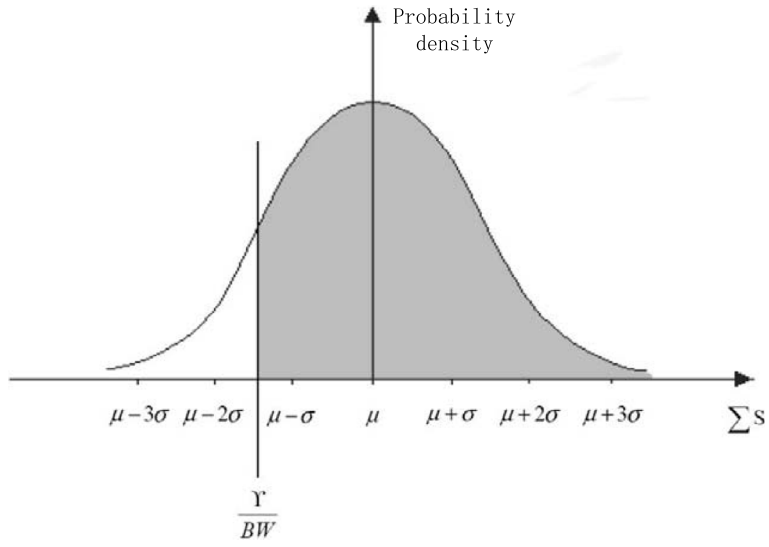


Figure 3 The distribution of $R(t)$.

$$f(R) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(R-\mu)^2}{2\sigma^2}} \tag{26}$$

The access probability is the probability of $R \geq \frac{\Upsilon}{BW}$, which is also shown by the shadow region in Figure 3. Therefore, Equation (25) is obtained and the proposition holds. □

Now we estimate the switching probability ξ by *asymptotic analysis*, in which the sensing period T_p is equally divided into k slim time spans. The situation within one slim time span $\frac{T_p}{k}$ is analyzed firstly, and then the period T_p is investigated by considering multiple slim time spans.

1) The case with complete and accurate sensing.

There are R available channels and $M-R$ channels occupied by the PUs in set $A_2(m)$ at the beginning of period T_p. The sensing period T_p is equally divided into k parts, in which the parameter k is large enough, so that we can assume that only one single channel's state is altered during one slim time span.

During the slim time span $\frac{T_p}{k}$, the number of available channels R in set $A_2(m)$ has three possible situations: increased by one, decreased by one, or unchanged. The probabilities of these three situations are denoted by P_{up}, P_{down}, and P_{hold}, respectively.

Here, we have two assumptions for approximation which are proposed based on the actual facts. First, the number of occupied channels $M-R$ stays unchanged at the beginning of each slim time span $\frac{T_p}{k}$, which is the most likely case. Second, we take the geometric average of all the channels in set A_2 to approximately calculate the probability of each channel to keep occupied, since the application of geometric average reflects the influence of small probabilities. The probability that the state of one channel keeps occupied P_{00} is calculated as follows:

$$P_{00} = \sqrt[M]{\prod_{n=C_{start}}^{n=C_{start}+M-1} P_{00}^n} \tag{27}$$

where P_{ij}^n be the transition probability of channel n from state i to state j during time $\frac{T_p}{k}$.

According to these assumptions, we can obtain

$$P_{up} = 1-(P_{00})^{M-R} = 1-\left(\sqrt[M]{\prod_{n=C_{start}}^{n=C_{start}+M-1} P_{00}^n}\right)^{M-R} \tag{28}$$

Similarly, we have

$$P_{down} = 1-(P_{11})^{R} = 1-\left(\sqrt[M]{\prod_{n=C_{start}}^{n=C_{start}+M-1} P_{11}^n}\right)^{R} \tag{29}$$

Based on P_{up} and P_{down}, we have $P_{hold} = 1-P_{up}-P_{down}$.

According to the above analysis, we obtain the expression of switching probability ξ in the following proposition.

Proposition 2 (Calculation of Switching Probability)**.** *For a given R, the switching probability $\xi(R)$ is*

$$\xi(R) = \sum_{l=\left\lceil \frac{R-\frac{\Upsilon}{BW}+H}{2} \right\rceil+1}^{H} \Pr\{H_d = l\} \tag{30}$$

where

$$Pr\{H_d = l\} = C_H^l \left(\frac{P_{down}}{P_{up}+P_{down}}\right)^l \left(\frac{P_{up}}{P_{up}+P_{down}}\right)^{H-l} \tag{31}$$

$$H = \lceil k(1-P_{hold})\rceil = \left\lceil k(P_{up}+P_{down})\right\rceil \tag{32}$$

Proof. During the sensing period T_p, there are $H = \lceil k(1-P_{hold})\rceil = \left\lceil k(P_{up}+P_{down})\right\rceil$ alterations of channel state in total, in which we assume that there are l times of decrease and $H-l$ times of increase of the number of available channels

If a channel switch occurs, which means that the bandwidth requirement of the SU is no longer satisfied, then the number of available channels R is less than $\frac{\Upsilon}{BW}$. In other words, $l-(H-l) > R-\frac{\Upsilon}{BW}$, and consequently $l > \frac{R-\frac{\Upsilon}{BW}+H}{2}$. We can calculate the switching probability as

$$\xi(R) = \Pr\left\{l > \frac{R-\frac{\Upsilon}{BW}+H}{2}\right\} \tag{33}$$

Based on Equation (33), Equation (30) can be obtained. □

2) The case with partial or inaccurate sensing.

Similar to the case for estimating ζ, R is a random variable instead of a deterministic variable, which has the p.d.f. of $f(R) = \frac{1}{\sqrt{2\pi}\sigma}e^{-\frac{(R-\mu)^2}{2\sigma^2}}$. Based on the result in Proposition 2, we can calculate the average switching probability by

$$\xi = \int_{R=0}^{M} f(R)\xi(R)dR \tag{34}$$

4 Joint spectrum sensing and access: Rollout policy

On basis of the proposed POMDP framework in Section 3, we can derive the optimal spectrum sensing and access scheme. For optimality, the value function $V^m(\Delta)$ is computed by averaging over all possible state transitions and observations. Since the number of system states grows exponentially with the number of channels, the realization of the optimal scheme suffers from the curse of dimensionality and is computationally overwhelming. In this section, we exploit the specific structure of the problem and develop a rollout-based suboptimal spectrum sensing and access scheme with a much lower complexity.

4.1 Rollout policy

The most essential issue of designing the spectrum sensing and access scheme is the calculation of the value function $V^m(\Delta)$, which is also the most computationally intensive part. To alleviate the complexity, we adopt an approximation technique that can offer an effective and computation-saved solution. Rollout algorithm [15], as an approximate dynamic programming methodology based on policy iteration ideas, has been successfully applied to various domains such as combinatorial optimization [18] and stochastic scheduling [19]. Instead of tracing the accurate value, the rollout algorithm can estimate the value function approximately. By use of Monte Carlo method, the results of a number of randomly generated samples are averaged, and the number of samples is typically smaller than the dimensionality of the total strategy space. When the sample number is large enough, we can obtain a joint spectrum sensing and access scheme with reduced complexity and limited performance loss.

To obtain a suboptimal solution, a problem-dependent heuristics is proposed first as the base policy, and then the reward of the base policy can be used by the rollout algorithm in a one-step lookahead method to approximate the value function. The procedure of the rollout-based scheme is illustrated in Figure 4.

The value function of POMDP can be written as

$$V^m(\Delta) = \max_{a\in\mathbf{A}} \mathbb{E}\left\{\kappa^m(a) + V^{m-1}(\Omega(\Delta|a,\theta))\right\} \tag{35}$$

where $\kappa^m(a)$ denotes the amount of time slots included in the m-th last control interval, namely the reward function which depends on the action choice a.

Base Policy In the rollout algorithm, a heuristic algorithm is needed to serve as the base policy, which is also designed on the basis of control interval structure.

$$\pi^{\mathcal{H}} = \left[\mu_1^{\mathcal{H}}, \mu_2^{\mathcal{H}}, \ldots, \mu_T^{\mathcal{H}}\right] \tag{36}$$

Here, we propose two different heuristics based on our designing objective, namely *Bandwidth-Oriented Heuristics (BOH)* and *Switch-Oriented Heuristics (SOH)*.

In BOH, we simply choose the sensing and access sets A_1 and A_2 which can obtain the widest expected available bandwidth currently,

$$\begin{aligned}\mu_m^{\mathcal{H}_1} : \Delta(m) &\rightarrow a^{\mathcal{H}_1}(m) \\ &= \arg\max_{a\in\mathfrak{A}} \sum_{i\in A_2(m)} P_i(A_1(m))\cdot BW\end{aligned} \tag{37}$$

where $P_i = \Pr\{S_i = 1\}$, which can be updated according to $A_1(m)$. Intuitively, the wider the available bandwidth is, the better the requirement of SU will be satisfied, and it is less possible to trigger the channel switch in the next time slot. But in this heuristics, the statistics of the PU traffic is not taken into consideration to predict the future dynamic behaviors of the channels.

In SOH, we choose the sensing and access actions that can maximize the expected current reward

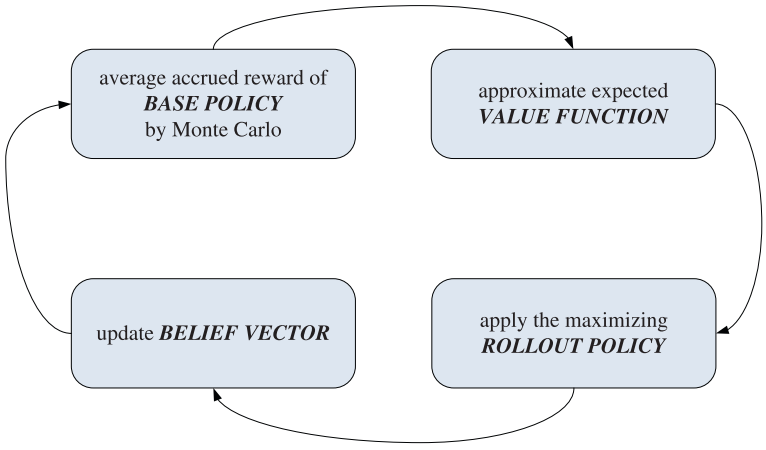


Figure 4 Rollout-based joint spectrum sensing and access scheme.

$$\mu_m^{\mathcal{H}_2} : \Delta(m) \to a^{\mathcal{H}_2}(m) = \arg\max_{a\in\mathfrak{A}} \sum_{\kappa^m} \kappa^m(a) p_{\kappa^m}(a) \tag{38}$$

where the calculation of p_{κ^m} includes the operation of prediction on the access probability ζ and the switching probability ξ. Making full use of the dynamic statistics of the channels, SOH is more sophisticated and achieves better performance than BOH.

Both heuristics are greedy and require low computational complexity. With either of the two heuristics as the base policy, the relevant expected reward from current control interval to termination can be calculated by recursion,

$$V_{\mathcal{H}}^m(\Delta) = \mathbb{E}\left\{\kappa^m(a^{\mathcal{H}}) + V_{\mathcal{H}}^{m-1}(\Omega(\Delta|a^{\mathcal{H}},\theta))\right\} \tag{39}$$

with the initial condition $V_{\mathcal{H}}^0(\Delta) = 0$.

Rollout Policy The rollout policy based on the base policy $\pi^{\mathcal{H}}$ is denoted by

$$\pi^{RL} = \left[\mu_1^{RL}, \mu_2^{RL}, \ldots, \mu_T^{RL}\right] \tag{40}$$

and is defined through the following operation.

$$\mu_m^{RL} : \Delta(m) \to a^{RL}(m) \tag{41}$$

where

$$a^{RL}(m) = \arg\max_{a\in\mathfrak{A}} \mathbb{E}\left\{\kappa^m(a) + V_{\mathcal{H}}^{m-1}(\Delta(m-1))\right\} \tag{42}$$

The rollout policy can approximate the value function by the use of the reward of the base policy, and consequently decide the near-optimal action $a^{RL}(m)$. We prove by theoretical deduction that the rollout policy is guaranteed to substantially improve the performance of the heuristics as the base policy.

Proposition 3 (Rollout Improving Property)**.** *The rollout policy is guaranteed to obtain better aggregated reward than the base policy.*

$$\begin{aligned} V_{\mathcal{H}}^T(\Delta(T)) &\le \mathbb{E}\left\{\kappa^T(a^{RL}(T)) + V_{\mathcal{H}}^{T-1}(\Delta(T-1))\right\} \\ &\quad \cdots \\ &\le \mathbb{E}\{\kappa^T(a^{RL}(T)) + \kappa^{T-1}(a^{RL}(T-1)) \\ &\quad + \cdots + \kappa^1(a^{RL}(1))\} \end{aligned} \tag{43}$$

Proof. The proposition is proved by backward mathematical induction.

For $m = T$, according to the essence of the rollout policy (42), we obtain

$$a^{RL}(T) = \arg\max_{a\in\mathfrak{A}} \mathbb{E}\left\{\kappa^T(a) + V_{\mathcal{H}}^{T-1}(\Delta(T-1))\right\} \tag{44}$$

Consequently, we have

$$\begin{aligned} V_{\mathcal{H}}^T(\Delta(T)) &= \mathbb{E}\left\{\kappa^T(a^{\mathcal{H}}) + V_{\mathcal{H}}^{T-1}(\Delta(T-1))\right\} \\ &\le \mathbb{E}\left\{\kappa^T(a^{RL}) + V_{\mathcal{H}}^{T-1}(\Delta(T-1))\right\} \end{aligned} \tag{45}$$

Hence, the proposition holds for $m = T$.

Assume it holds for $m < T$, i.e.,

$$\begin{aligned} V_{\mathcal{H}}^T(\Delta(T)) &\le \mathbb{E}\left\{\kappa^T(a^{RL}(T)) + V_{\mathcal{H}}^{T-1}(\Delta(T-1))\right\} \\ &\quad \cdots \\ &\le \mathbb{E}\{\kappa^T(a^{RL}(T)) + \kappa^{T-1}(a^{RL}(T-1)) \\ &\quad + \cdots + \kappa^m(a^{RL}(m)) + V_{\mathcal{H}}^{m-1}(\Delta(m-1))\} \end{aligned}$$

Using the essence of the rollout policy (42) again,

$$a^{RL}(m-1) = \arg\max_{a\in\mathfrak{A}} \mathbb{E}\left\{\kappa^{m-1}(a) + V_{\mathcal{H}}^{m-2}(\Delta(m-2))\right\} \tag{46}$$

we can obtain that

$$V_{\mathcal{H}}^{m-1}(\Delta(m-1)) = \mathbb{E}\left\{\kappa^{m-1}(a^{\mathcal{H}}) + V_{\mathcal{H}}^{m-2}(\Delta(m-2))\right\}$$

$$\leq \mathbb{E}\left\{\kappa^{m-1}(a^{RL}(m-1)) + V_{\mathcal{H}}^{m-2}(\Delta(m-2))\right\} \quad (47)$$

Consequently, we have

$$\begin{aligned} & V_{\mathcal{H}}^{T}(\Delta(T)) \\ \leq\ & \mathbb{E}\left\{\kappa^{T}(a^{RL}(T)) + V_{\mathcal{H}}^{T-1}(\Delta(T-1))\right\} \\ & \cdots \\ \leq\ & \mathbb{E}\{\kappa^{T}(a^{RL}(T)) + \kappa^{T-1}(a^{RL}(T-1)) + \cdots \\ & +\kappa^{m}(a^{RL}(m)) + V_{\mathcal{H}}^{m-1}(\Delta(m-1))\} \\ \leq\ & \mathbb{E}\{\kappa^{T}(a^{RL}(T)) + \kappa^{T-1}(a^{RL}(T-1)) + \cdots \\ & +\kappa^{m}(a^{RL}(m)) + \kappa^{m-1}(a^{RL}(m-1)) \\ & +V_{\mathcal{H}}^{m-2}(\Delta(m-2))\} \end{aligned} \quad (48)$$

Therefore, the property holds for $m-1$. According to the mathematical induction, the proposition is proved. □

4.2 Suboptimal spectrum sensing and access

Focusing on the implementation of the proposed rollout policy (42), we define Q-factor as

$$Q_m(a) = \mathbb{E}\left\{\kappa^{m}(a) + V_{\mathcal{H}}^{m-1}(\Delta(m-1))\right\} \quad (49)$$

which indicates expected reward that the SU can accrue during the lifetime of the process from current control interval, and then the rollout action can be expressed as $a^{RL}(m) = \arg\max_{a\in\mathfrak{A}} Q_m(a)$.

However, as the key point of the rollout policy, the Q-factor may not be known in a closed form, which makes the computation of $a^{RL}(m)$ an nontrivial issue [20]. To overcome this difficulty, we adopt a widely used Monte Carlo method [21].

Here, we define the *trajectory* as a sequence of the form

$$(\mathbf{S}(T), a(T), \mathbf{S}(T-1), a(T-1), \cdots, \mathbf{S}(1), a(1)) \quad (50)$$

Using the Monte Carlo method, we consider any possible action $a \in \mathfrak{A}$ and generate a large number of trajectories of the system starting from belief vector $\Delta(m)$, using a as the first action and the base policy $\pi^{\mathcal{H}}$ thereafter. Thus, a trajectory has the form as

$$\left(\mathbf{S}(m), a, \mathbf{S}(m-1), a^{\mathcal{H}}(m-1), \cdots, \mathbf{S}(1), a^{\mathcal{H}}(1)\right) \quad (51)$$

where the system states $\mathbf{S}(m), \mathbf{S}(m-1), \cdots, \mathbf{S}(1)$ are randomly sampled according to the belief vectors which are updated based on the action and observation history:

$$\Delta(i-1) = \begin{cases} \Omega(\Delta|a^{\mathcal{H}}(i), \theta) & i = m-1, m-2, \cdots, 1 \\ \Omega(\Delta|a, \theta) & i = m \end{cases} \quad (52)$$

The rewards corresponding to these trajectories are averaged to compute $\widetilde{Q}_m(a)$ as an approximation to the Q-factor $Q_m(a)$. The approximation value becomes increasingly accurate as the number of trajectories increases. Once the approximate Q-factor $\widetilde{Q}_m(a)$ corresponding to each action $a \in \mathfrak{A}$ is computed, we can obtain the approximate rollout action $\widetilde{a}^{RL}(m)$ by the maximization

$$\widetilde{a}^{RL}(m) = \arg\max_{a\in\mathfrak{A}} \widetilde{Q}_m(a) \quad (53)$$

This rollout-based suboptimal spectrum sensing and access scheme can reduce the computational complexity a lot by estimating the value function approximately rather than tracing the accurate value.

4.3 Robustness via differential training

It is obvious that, in a stochastic environment, the Monte Carlo method of computing the rollout policy is particularly sensitive to the approximation error, which is closely related to the number of trajectories. In this subsection, we adopt *differential training* [22] in the proposed rollout-based suboptimal scheme to improve the robustness. In the differential training method, we estimate the relative Q-factor difference rather than absolute Q-factor value, which is a suitable improvement of the recursively generating rollout policy in the context of Monto Carlo-based policy iteration methods.

In order to compute the rollout action $a^{RL}(m) = \arg\max_{a\in\mathfrak{A}} Q_m(a)$, the Q-factor differences $Q_m(a_1) - Q_m(a_2)$, $\forall a_1, a_2 \in \mathfrak{A}$ should be computed accurately. By comparing the Q-factor differences with 0, these possible actions can be accurately compared. Unfortunately, in a stochastic environment, the approximation $\widetilde{Q}_m(a)$ fluctuated around the accurate Q-factor value, bigger or smaller than $Q_m(a)$ randomly, as a result of which, the preceding differences computing operation enlarges the approximation error. For example, in the case that a_1 performs better than a_2 and thus, $Q_m(a_1)$ is definitely bigger than $Q_m(a_2)$, which results in $Q_m(a_1) - Q_m(a_2) > 0$. However, when using stochastic Monte Carlo method, the approximate $\widetilde{Q}_m(a_1)$ may be smaller than the accurate value $Q_m(a_1)$, and meanwhile $\widetilde{Q}_m(a_2)$ may be bigger than the accurate value $Q_m(a_2)$, which makes it quite possible that $\widetilde{Q}_m(a_1) - \widetilde{Q}_m(a_2) < 0$, and this computation result will lead to a fatal error when determining which action is chosen for spectrum sensing and access.

To reduce the negative effects of the approximation error discussed above, we adopt the differential training method. Specifically, we take the Q-factor value of the base policy $\pi^{\mathcal{H}}$ as a reference to enhance the robustness, which can be viewed as a variance reduction technique. Instead of approximating the independent Q-factor, the approximate rollout action $\widetilde{a}^{RL}(m)$ is obtained

by maximizing the approximation of the Q-factor difference $Q_m(a) - Q_m(a^{\mathcal{H}})$,

$$\widetilde{a}^{RL}(m) = \arg\max_{a\in\mathfrak{A}} \left\{\widetilde{Q}_m(a) - \widetilde{Q}_m(a^{\mathcal{H}})\right\} \tag{54}$$

The reference $\widetilde{Q}_m(a^{\mathcal{H}})$ has the same fluctuation monotonicity as $\widetilde{Q}_m(a)$, which is caused by the approximation error due to the limited number of trajectories. We take the same example that a_1 actually performs better than a_2 and $Q_m(a_1) > Q_m(a_2)$. If the approximate $\widetilde{Q}_m(a_1)$ is smaller than the accurate value, so is $\widetilde{Q}_{m1}(a^{\mathcal{H}})$. Similarly, if $\widetilde{Q}_m(a_2)$ is larger than the accurate value, so is $\widetilde{Q}_{m2}(a^{\mathcal{H}})$. Using the differential training operation $\widetilde{Q}_m(a) - \widetilde{Q}_m(a^{\mathcal{H}})$, the effect of approximation error can be eliminated. Thus, it probably holds that $\widetilde{Q}_m(a_1) - \widetilde{Q}_{m1}(a^{\mathcal{H}}) > \widetilde{Q}_m(a_2) - \widetilde{Q}_{m2}(a^{\mathcal{H}})$, consequently the SU will choose the better action a_1.

From the above discussion, the approximate Q-factor difference $\widetilde{Q}_m(a) - \widetilde{Q}_m(a^{\mathcal{H}})$ is more robust than the approximate independent Q-factor $\widetilde{Q}_m(a)$. By the differential training of the rollout policy, the approximation error caused by Monte Carlo method can be reduced a lot and the proposed suboptimal spectrum sensing and access scheme performs more robustly.

5 Simulation results

In this section, we evaluate the performance of the proposed joint spectrum sensing and access scheme by simulation. We investigate the effect of the number of Monte Carlo random trajectories, the proportion of sensing channels L/N, and the ratio of aggregation range to bandwidth requirement Γ/γ. The PU traffic statistics follows the model of Erlang-distribution [23], and the simulation configuration is listed in Table 1. The average simulation results are obtained by 100 runs with random channel states.

In Figure 5, for different number of Monte Carlo random trajectories, we compute the value of the approximate independent Q-factor $\widetilde{Q}_m(a)$ and the value of the approximate Q-factor difference $\widetilde{Q}_m(a) - \widetilde{Q}_m(a^{\mathcal{H}})$, respectively. Two pairs of curves represent two different rollout actions $a_1, a_2 \in \mathfrak{A}$ chosen in the current control interval. It is shown that, no matter which action to take, the fluctuation range of $\widetilde{Q}_m(a)$ and $\widetilde{Q}_m(a) - \widetilde{Q}_m(a^{\mathcal{H}})$ converges with the increase of the number of random trajectories. With a small number of random trajectories, $\widetilde{Q}_m(a) - \widetilde{Q}_m(a^{\mathcal{H}})$ has less fluctuation than that of $\widetilde{Q}_m(a)$. If the trajectory number exceeds about 1,500, both approximate values converge. It is indicated that the approximate Q-factor difference is more robust than the approximate independent Q-factor, and the differential training of the rollout policy can reduce the negative effect of the approximation error a lot. In the case that the trajectory number is large enough, both approximate values are nearly accurate compared with the original values.

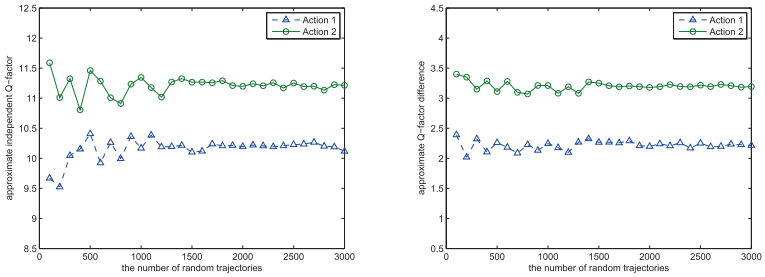
Table 1 Simulation configuration

Total number of channels N	10
Number of sensing channels L	3
Bandwidth per channel BW	10 MHz
Aggregation range Γ	40 MHz
Bandwidth requirement Υ	20 MHz
Duration of time slot T_p	2 ms
Probability of false alarm P_f	0.03
Probability of miss detection P_m	0.08

Figure 6 illustrates the effect of the proportion of sensing channels L/N on the performance of both optimal and suboptimal schemes. The random scheme is adopted as a baseline for performance comparison, in which M channels are chosen randomly to access. Besides, we also evaluate the performance of the base policies, the suboptimal rollout schemes based on BOH and SOH, and the POMDP-based optimal scheme. Here, we adopt the performance with 1,500 random trajectories for approximation in rollout policies, which can achieve the performance in convergence.

In Figure 6, as the number of sensing channels L increases, the numbers of channel switches decrease for the BOH, SOH, BOH- and SOH-based rollout, and optimal POMDP-based schemes. When the whole spectrum can be sensed ($L/N = 1$), these schemes achieve their corresponding best performance because the more channels the SU senses, the more information about the system state can be obtained. The spectrum aggregation action determined on the basis of sensing results has better performance in minimizing the expected times of channel switches. For the random access scheme, which determines the access channels without considering the sensing results, the performance does not change with the increase of L.

When L is small, which means that a small number of channels can be sensed, the performances of all schemes are almost the same because the system performance is limited by L in this case. With the increasing of L, the POMDP-based optimal scheme performs the best, and the rollout-based suboptimal schemes achieve much better performance than the basis heuristics and the random scheme. Especially, the SOH-based rollout scheme achieves a performance gain over the BOH-based rollout scheme, which verifies that the choice of the base policy affects the performance of the corresponding rollout policy. When the heuristic is good, the rollout scheme based on it can achieve relatively better performance. Compared with the optimal POMDP-based scheme, the rollout-based suboptimal scheme only has a slight performance



(a) (b)

Figure 5 Algorithm convergence. **(a)** Convergence of approximate independent Q-factor and **(b)** Convergence of approximate Q-factor difference.

loss, but makes significant improvement in reducing the computational complexity.

Figure 7 illustrates the effect of the ratio of aggregation range to bandwidth requirement Γ/γ on the performance of both optimal and suboptimal schemes. The performance comparison of these schemes is similar to those in Figure 6. The performance gaps of different schemes are large when the aggregation range is small, because sophisticated schemes can select the channels whose availability state is stable. On the contrary, all the schemes can achieve a good performance when the aggregation range is large enough.

In Figure 8, we evaluate the performance of the rollout-based schemes with different number of random trajectories when $L = 3$. The performances of the random and optimal schemes stay constant with the increase of the number of random trajectories. The performance of the rollout-based scheme approaches to the optimal scheme until the number of random trajectories reach the converging boundary which is 1,500 in this simulation. The differential training method improves the performance of the rollout-based schemes significantly when the number of random trajectories is small. After convergence, the advantage of differential training is small since the performance is not so sensitive to the approximation error with a large enough number of random trajectories. It is proved again in Figure 8 that the SOH-based rollout outperforms the BOH-based one.

6 Conclusion

In this paper, we investigate the spectrum sensing and access schemes to minimize the channel switching times for achieving stable DSA, taking into consideration the practical limitations of both spectrum sensing and

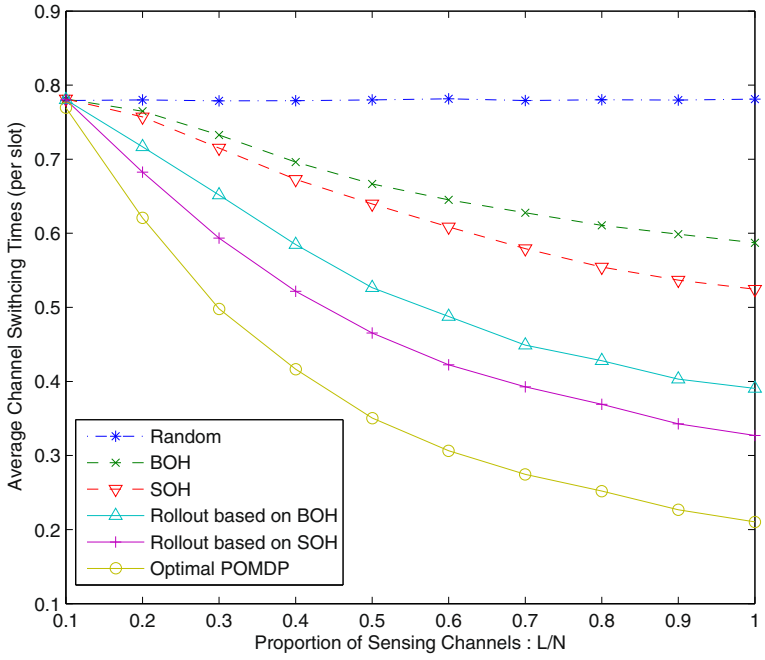


Figure 6 Performance comparison with different spectrum sensing capability.

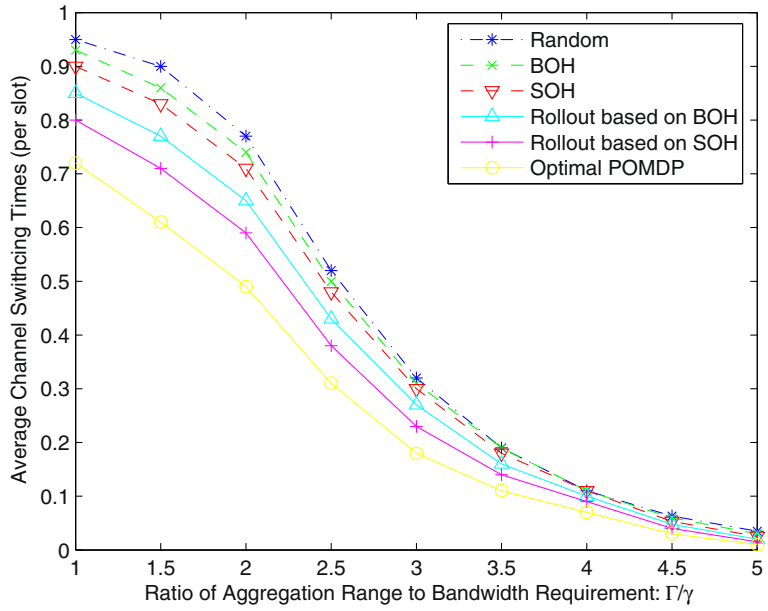


Figure 7 Performance comparison with different spectrum aggregation capability.

aggregation capability. We develop an POMDP framework for joint spectrum sensing and access. Especially, we derive the reward function by estimation of the stability of different spectrum sensing and access strategies. Based on the POMDP framework, we propose a rollout-based suboptimal spectrum sensing and access scheme which approximates the value function of POMDP. It is proved that the rollout policy achieves performance improvement over the basis heuristics. By numerical evaluation, we find that with the increase of number of random trajectories, the performance of the proposed rollout-based scheme gets close to the optimal performance. When the number of random trajectories is large enough, the proposed scheme performs near-optimally with a lower complexity,

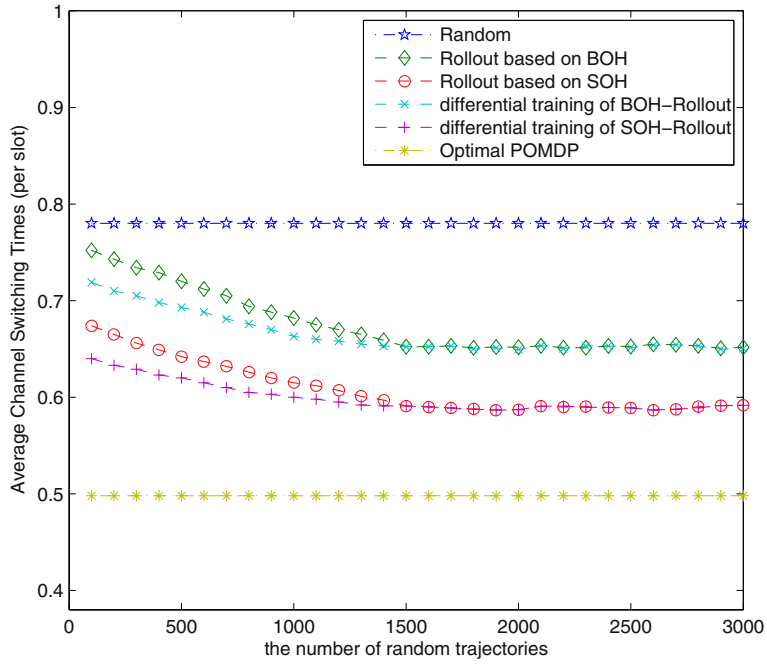


Figure 8 Performance comparison with different number of random trajectories.

which also achieves a significant improvement over the base policy. In the rollout-based schemes, the basis heuristics affects the performance of its corresponding rollout policy, and the differential training method improves the robustness to the approximation error.

Endnotes

[a]The transition probabilities can be estimated by the statistics of the channel availabilities of two adjacent slots and is assumed to be known by the SUs [23].

[b]The schemes proposed in this paper can be easily extended to multiple SU cases by adopting the RTS/CTS scheme [24] for the access coordination between SUs.

[c]Minimizing the expected times of channel switches can also be treated equally as maximizing the throughput with the consideration of the system overhead of channel switches.

Competing interests
The authors declare that they have no competing interests.

Acknowledgements
This work was supported in part by National Natural Science Foundation of China (No. 61261130585), National Key Basic Research Program (No. 2012CB316006), National Hi-Tech R&D Program (No. 2014AA01A702), Fundamental Research Funds for the Central Universities, and Open Research Fund of State Key Laboratory of Integrated Services Networks (No. ISN13-08).

Author details
[1]Department of Information Science and Electronic Engineering, Zhejiang Key Lab of Information Network Technology, Zhejiang University, Hangzhou 310027, P.R. China. [2]State Key Laboratory of Integrated Services Networks, Xidian University, Xi'an 710071, P.R. China. [3]Laboratoire de Recherche en Informatique (LRI), Department of Computer Science, University of Paris-Sud 11, Orsay 91405, France.

Received: 1 October 2014 Accepted: 17 April 2015
Published online: 08 May 2015

References

1. W Wang, Z Zhang, A Huang, Spectrum aggregation: Overview and challenges. Net. Protoc. Appl. **2**(1), 184–196 (2010)
2. QinetiQ Ltd, A study of the provision of aggregation of frequency to provide wider bandwidth services. Final report for Office of Communications (Ofcom) (2006). http://www.indepen.uk.com/docs/aggregation[1].pdf
3. F Wu, Y Mao, S Leng, *et al*, A carrier aggregation based resource allocation scheme for pervasive wireless networks. Proc. of IEEE DASC 2011, 196–201 (2011)
4. H Shajaiah, A Abdel-Hadi, C Clancy, Utility proportional fairness resource allocation with carrier aggregation in 4G-LTE. Proc. of IEEE Milcom 2013, 412–417 (2013)
5. F Liu, K Zheng, W Xiang, *et al*, Design and performance analysis of an energy-efficient uplink carrier aggregation scheme. IEEE J. Sel. Areas Commun. **32**(2), 197–207 (2014)
6. J Mitola, G Maguire, Cognitive radio: making software radios more personal. IEEE Pers. Commun. **6**(4), 13–18 (1999)
7. W Wang, KG Shin, W Wang, Joint spectrum allocation and power control for multi-hop cognitive radio networks. IEEE Trans. Mobile Comput. **10**(7), 1042–1055 (2011)
8. Q Zhao, L Tong, A Swami, Y Chen, Decentralized cognitive MAC for opportunistic spectrum access in ad hoc networks: a POMDP framework. IEEE J. Sel. Areas Commmun. **25**(3), 589–600 (2007)
9. AA El-Sherif, KJR Liu, Joint design of spectrum sensing and channel access in cognitive radio networks. IEEE Trans. Wireless Commum. **10**(6), 1743–1753 (2011)
10. W Wang, K Wu, H Luo, G Yu, Z Zhang, Sensing error aware delay-optimal channel allocation scheme for cognitive Radio Networks. Telecommun. Sys. **52**(4), 1895–1904 (2013)
11. J Park, P Pawelczak, D Cabric, Performance of joint spectrum sensing and MAC algorithms for multichannel opportunistic spectrum access Ad Hoc networks. IEEE Trans. Mobile Comput. **10**(7), 1011–1027 (2011)
12. F Huang, W Wang, H Luo, G Yu, Z Zhang, Prediction-based spectrum aggregation with hardware limitation in cognitive Radio Networks. Proc. of IEEE VTC. 2010 (2010)
13. L Wu, W Wang, Z Zhang, L Chen, A POMDP-based optimal spectrum sensing and access scheme for cognitive radio networks with hardware limitation. Proc. of IEEE WCNC 2012 (2012)
14. GE Monahan, A survey of partially observable Markov decision processes: Theory, models, and algorithms. Manage. Sci. **28**(1), 1–16 (1982)
15. DP Bertsekas, JN Tsitsiklis, Neuro-dynamic programming: an overview. Proc. of IEEE CDC 1995 (1995)
16. R Smallwood, E Sondik, The optimal control of partially observable Markov processes over a finite horizon. Oper. Res, 1071–1088 (1971)
17. BV Gendenko, AN Kolmogorov, *Limit Distributions for Sums of Independent Random Variables*, (Addison-Wesley, 1954)
18. DP Bertsekas, JN Tsitsiklis, C Wu, Rollout algorithms for combinatorial optimization. J. Heuristics. **3**(2), 245–262 (1997)
19. DP Bertsekas, DA Castanon, Rollout algorithms for stochastic scheduling problems. J. Heuristics. **5**(1), 89–108 (1998)
20. L Wu, W Wang, Z Zhang, L Chen, A Rollout-based joint spectrum sensing and access policy for cognitive radio networks with hardware limitation. Proc. of IEEE Globecom 2012 (2012)
21. G Tesauro, GR Galperin, On-line policy improvement using Monte Carlo search. Proc. of Neural Inf Process. Syst. Conf (1996)
22. DP Bertsekas, Differential training of rollout policies. Proc. of Allerton Conference on Communication, Control, and Computing (1997)
23. H Kim, KG Shin, Efficient discovery of spectrum opportunities with MAC-layer sensing in cognitive radio networks. IEEE Trans. Mobile Comput. **7**, 533–545 (2008)
24. G Bianchi, Performance analysis of the IEEE 802.11 distributed coordination function. IEEE J. Sel. Areas Commun. **18**(3), 535–547 (2000)