A Rollout-based Joint Spectrum Sensing and Access Policy for Cognitive Radio Networks with Hardware Limitations

Lingcen Wu[†‡], Wei Wang[†‡], Zhaoyang Zhang[†], Lin Chen[§]
[†]Department of Information Science and Electronic Engineering,
Zhejiang Provincial Key Lab of Information Network Technology,
Zhejiang University, Hangzhou, P.R. China
[‡]State Key Laboratory of Integrated Services Networks,
Xidian University, Xi'an, P.R. China
[§]Laboratoire de Recherche en Informatique (LRI),
University of Paris-Sud, Orsay, France

***Abstract*—The practical hardware limitations bring technical challenges to cognitive radio, e.g. limited capability of spectrum sensing and certain frequency range of spectrum access. In this paper, we propose a rollout-based joint spectrum sensing and access policy incorporating the hardware limitations of both sensing capability and spectrum aggregation, in which the optimal policy is shown to be PSPACE-hard. Two heuristic policies are proposed to serve as base policies, based on which the developed rollout-based policy approximates the value function and determines the appropriate spectrum sensing and access actions. We establish mathematically that the rollout-based policy achieves better performance than the base policy. We also demonstrate that the low-complexity rollout-based policy leads to only slight performance loss compared with the optimal policy.**

I. Introduction

The proliferation of wireless mobile networks and the ever-increasing density of wireless devices underscore the necessity for efficient allocation and sharing of the radio spectrum resource. Cognitive radio (CR) [1], with its capability to flexibly configure its transmission parameters, has emerged in recent years as a promising paradigm to enable more efficient spectrum utilization. The objective of CR is to solve the imbalance between spectrum scarcity and under-utilization. With CR technique, secondary users are allowed to search for, identify, and exploit instantaneous spectrum opportunities while limiting the interference perceived by primary users (or licensees).

While conceptually simple, CR presents novel challenges, among which spectrum sensing and access are of primordial importance and thus have attracted considerable research attention in recent years. Among representative works, a decentralized MAC protocol is proposed in [2] where SUs search for spectrum opportunities without a centralized controller. The optimal sensing and channel selection schemes maximize the expected total number of bits delivered over a finite number of slots. The authors of [3] propose a Least Channel Switch (LCS) strategy for spectrum assignment considering the dynamic access of SUs with different bandwidth requirements. In [4], considering the fusion strategy of collaborative spectrum sensing, the authors design a multi-channel MAC protocol.

More recently, motivated the impact of hardware limitations and physical constraints on the performance of spectrum sensing and access, we have developed a joint spectrum sensing and access scheme by systematically incorporating the following practical constraints: (1) the continuous full-spectrum sensing being impossible, SUs can only sense and access a subset of spectrum channels; (2) only spectrum channels within a certain frequency range can be aggregated and accessed for data transmission [5]. A decision-theoretic approach has been proposed in [6] to model the joint spectrum sensing and access problem under these constraints as a Partially Observable Markov Decision Process (POMDP) [7]. By application of linear programming, the optimal policy is obtained which minimizes the times of channel switch, thus reducing the system overhead and maintaining its stability in dynamic environments. However, the formulated problem being PSPACE-hard, the practical application of the derived optimal policy is severely limited due to its exponential computation complexity. Therefore, a heuristic joint spectrum sensing and access policy is called for so as to strike a balanced between system performance and computation complexity.

In this paper, we develop a joint spectrum sensing and access policy based on the rollout algorithms, a class of suboptimal solution methods inspired by the policy iteration methodology of dynamic programming. Specifically, two heuristic policies are proposed to serve as base policies, based on which the developed rollout-based policy approximates the value function and determines the appropriate spectrum

This work was supported in part by National Key Basic Research Program of China (No. 2009CB320405), National Natural Science Foundation of China (Nos. 61001098, 60972057, 60972058), National Science and Technology Major Project of China (No. 2012ZX03002009), Xu Guangqi Project and the open research fund of State Key Laboratory of Integrated Services Networks, Xidian University.

sensing and access actions. We establish mathematically that the rollout-based policy achieves better performance than the base policies. We also demonstrate that the low-complexity rollout-based policy leads to only slight performance loss compared with the optimal policy.

The rest of this paper is organized as follows. Section II introduces the system model and the optimal scheme in the POMDP framework. The rollout-based suboptimal spectrum sensing and access scheme is proposed in Section III. Section IV provides the performance evaluation by simulation. Finally, this paper is concluded in Section V.

II. Joint Spectrum Sensing and Access: A POMDP Formulation

We consider a large-span licensed spectrum consisting of N independent channels, each of bandwidth BW. Let the vector $\mathbf{S}(t)$ denote the system state at time slot t,

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

where $S_n(t) \in \{0(\text{occupied}), 1(\text{idle})\}$ represents the state of channel $n \in \{1, ..., N\}$ at time slot t. The transition probability of system states $p_{ij} = Pr\{\mathbf{S}(t+\tau) = j | \mathbf{S}(t) = i\}$ can be calculated based on the state of each channel $P^n_{xy}(\tau)$,

$$P^n_{xy}(\tau) = Pr\{S_n(t+\tau) = y | S_n(t) = x\}, \forall x, y \in \{0,1\} \tag{2}$$

which can be estimated by the statistics of the primary network traffic and are assumed to be known by SUs [9].

At the beginning of each time slot, the SU chooses a set of channels A_1 to sense and a set of channels A_2 to access in order to satisfy the bandwidth requirement Υ. The size of A_1 is no more than L channels, and the channels in A_2 are within the frequency range Γ, which are characterized by the spectrum sensing and aggregation limitations, respectively. Before choosing A_1 and A_2, the SU checks whether its requirement Υ is still satisfied. If yes, only A_1 is selected and the spectrum access decision A_2 does not change; otherwise, the SU has to reselect appropriate A_1 and A_2 and trigger a channel switch. Define $\eta(t)$ as the expected number of channel switches from slot 0 to slot t, we focus on the SU's optimization problem of minimizing $\eta(t)$ by appropriately choosing A_1 and A_2. Such joint spectrum sensing and access problem can be formulated as follows:

$$\min_{A_1, A_2} \quad \lim_{t\to\infty} \frac{\eta(t)}{t} \tag{3}$$

$$s.t. \quad |A_1| \le L \tag{4}$$

$$D(i,j) \le \Gamma, \quad \forall i, j \in A_2 \tag{5}$$

$$\sum_{n\in A_2} S_n(t) \ge \frac{\Upsilon}{BW}, \quad \forall t \tag{6}$$

where $D(i,j)$ denotes the frequency distance between channel i and j. The first two constraints indicate the spectrum sensing and spectrum aggregation limitations respectively, and the last constraint guarantees that the bandwidth requirement is satisfied.

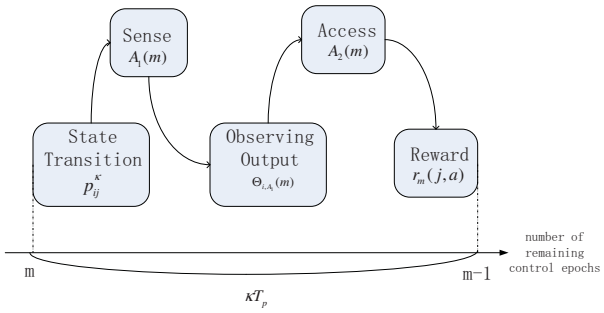

Fig. 1. The basic operations of POMDP

To better present our analysis, we divide time into *control epoches*, each composed of a number of consecutive time slots and delimited by channel switches. Formally, let $t_s(k)$ denote the time slot when the kth channel switch is triggered, the kth control epoch denotes the time from $t_s(k-1)$ to $t_s(k)$ with $t_s(0) = 0$. Clearly, the longer the current accessed channels can keep satisfying the bandwidth requirement of the SU, the longer is the corresponding control epoch.

Mathematically, the optimization problem faced by the SU can be cast into a class of POMDP frameworks [7] by incorporating the control epoch structure. The basic operations in each control epoch are shown in Fig. 1, in which T_p denotes the duration of one time slot.

Let T denote the number of control epoches within the time horizon t, and the index m denote the m-th last control epoch (i.e., the mth control epoch from slot t). The state transition probability expressed in control epoches is denoted by $p^{\kappa}_{ij} = Pr\{\mathbf{S}(m-1) = j | \mathbf{S}(m) = i\}$, where κ indicates the number of time slots in the control epoch. Taking both spectrum sensing and access as the ***action***, denoted by $a(m)$ for epoch m, and the sensing results as the ***observation***, denoted by $\Theta_{i,A_1}(m)$ for epoch m, we have

$$a(m) = \{A_1(m); A_2(m)\} = \{C_1, C_2, ..., C_L; C_{start}\} \tag{7}$$

$$\Theta_{i,A_1}(m) = \{S_{C_1}(m), S_{C_2}(m), ..., S_{C_L}(m)\} \tag{8}$$

where C_i is the index of the i-th sensed channel, C_{start} is the index of the first accessed channel in A_2, and $\Theta_{i,A_1}(m)$ indicates the observation output with the current system state i and the sensing action A_1.

A ***belief vector*** $\Delta(m)$ is introduced to represent the SU's estimation of the system state based on past decisions and observations, which is also a sufficient statistics for designing the optimal policy for future epoches. Formally,

$$\Delta(m) \quad = \quad (\delta_i(m))_{i\in\mathfrak{S}} \triangleq (\Pr\{\mathbf{S}(m) = i | H(m)\})_{i\in\mathfrak{S}} \tag{9}$$

where $H(m) = \{a(i), \Theta(i)\}_{i\ge m}$. A ***joint spectrum sensing and access policy*** (termed as policy for briefty) $\pi \triangleq (\mu_m, 1 \le m \le T)$ is defined as a mapping from the belief vector $\Delta(m)$ to the action $a(m)$ for each epoch: i.e.,

$$\mu_m : \Delta(m) \in [0,1]^{2^N} \to a(m) = \{A_1(m)\ A_2(m)\}. \tag{10}$$

To quantify the SU's objective, we define the ***reward*** of a control epoch as the number of time slots in the control epoch, i.e. the length of the control epoch. We now show that minimizing the number of channel switches equals to maximizing the total reward. To this end, let T denote the total number of control epoches over the whole time horizon (t slots) and $R(T)$ denote the total reward, we have

$$\eta(t) = \underset{T}{\operatorname{argmin}}\{R(T) \geq t\}. \tag{11}$$

It then follows that

$$\underset{\pi}{\operatorname{argmin}} \frac{\eta(t)}{t} = \underset{\pi}{\operatorname{argmax}} \frac{R(T)}{t}. \tag{12}$$

Moreover, it can be noted that given m, its reward for this control epoch is a Bernoulli random variable with probability density function (pdf) $p(\kappa)$ ($\kappa \in \mathbb{Z}^+$) derived as follows:

$$p(\kappa) = \zeta \cdot (1-\xi)^{\kappa-1} \cdot \xi, \tag{13}$$

where ζ is the probability that the channels in A_2 have available bandwidth more than Υ in current time slot, and ξ is the probability that the bandwidth requirement of the SU would not be satisfied by A_2 in the next time slot. Both the access probability ζ and the switching probability ξ can be calculated according to central limit theorem [12] and asymptotic analysis as in [6].

To find an optimal policy π^*, we express the cumulated reward in the recursive form by a function defined as the ***value function*** formalized as follows:

$$V^m(\Delta) = \max_{a \in \mathbf{A}} \left\{ \sum_i \delta_i \sum_\kappa p_\kappa \sum_j p_{ij}^\kappa \sum_\theta \Pr[\Theta_{j,A_1} = \theta][\kappa + V^{m-1}(\Omega(\Delta|a,\theta))] \right\} \tag{14}$$

with the initial condition $V^0(\Delta) = 0$, and the update rule operator of the belief vector Δ is denoted by $\Omega(\Delta|a,\theta)$.

It has been proved in [10] that $V^m(\Delta)$ is piecewise linear and convex. Specifically,

$$V^m(\Delta) = \max_\omega \left[\sum_i \delta_i \alpha_i^\omega(m) \right] \tag{15}$$

where the 2^N-dimensional vector $\vec{\alpha}^\omega(m)$ denotes the slopes associated with different convex regions divided from the space of belief vectors, which can be calculated as

$$\begin{aligned} \alpha_i(m) &= \sum_{j,\theta,\kappa} p_\kappa p_{ij}^\kappa \Pr[\Theta_{j,A_1} = \theta] \cdot \kappa \\ &\quad + \sum_{j,\theta,\kappa} p_\kappa p_{ij}^\kappa \Pr[\Theta_{j,A_1} = \theta] \alpha_j^\omega(m-1) \end{aligned} \tag{16}$$

Obviously, the calculation of a new α-vector yields an optimal action $a^*(m)$. By linear programming [11], the α-vectors and the corresponding optimal actions in all control epoches can be calculated by backward induction, and then stored in a table. For a given Δ, we can find the maximum α-vector through (15). By searching the table for the corresponding optimal action, the optimal sensing and access scheme is obtained, i.e. $\Delta \Rightarrow \vec{\alpha} \Rightarrow a^*$.

However, both the value function $V^m(\Delta)$ and the α-vectors are obtained by averaging over all possible state transitions and observations. Since the number of system states is exponential with respect to the number of channels, the implementation of the optimal scheme suffers from the curse of dimensionality and is computationally expensive or even prohibitive in some cases. Hence, a low-complexity policy is called for to achieve a desired balance between system performance and computation complexity, which is the subject of the sequent study.

III. Rollout-based Joint Spectrum Sensing and Access Policy

In this section, we exploit the structural properties of the problem and develop a joint spectrum sensing and access scheme with reduced complexity and limited performance loss.

The core part of the joint optimization of spectrum sensing and access is the calculation of the value function $V^m(\Delta)$, which is also the most computationally intensive component. To alleviate the complexity, we adopt the rollout algorithm [8], an approximation technique that can significantly reduce computation complexity. Rollout algorithm, as an approximate dynamic programming methodology based on policy iteration, has been widely used in various applications ranging from combinatorial optimization [13] to stochastic scheduling [14]. Its basic idea is one-step lookahead. To obtain the value function in an efficient way, the rollout algorithm tries to estimate the value function approximately rather than tracing the accurate value. The most widely used approximation approach is Monte Carlo method, which averages the results of a number of randomly generated samples. As the sample number is typically order-of-magnitude fewer compared to the total strategy space, the computational complexity can be significantly reduced.

We now develop a rollout framework to design the joint spectrum sensing and access policy. To this end, the problem-dependent heuristic method is proposed first as the base policy, whose reward will be used by the rollout algorithm to approximate the value function. Fig. 2 illustrates the procedure of the proposed rollout-based policy. For simplicity, we rewrite the value function (14) as

$$V^m(\Delta) = \max_{a \in \mathbf{A}} E\left\{\kappa^m(a) + V^{m-1}(\Omega(\Delta|a,\theta))\right\} \tag{17}$$

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

Base Policy To apply the rollout algorithm, a heuristic algorithm is needed to serve as the base policy:

$$\pi^{\mathcal{H}} = [\mu_1^{\mathcal{H}}, \mu_2^{\mathcal{H}}, ..., \mu_T^{\mathcal{H}}] \tag{18}$$

In our study, we develop two heuristic algorithms, namely Bandwidth-Oriented Heuristics (BOH) and Switch-Oriented Heuristics (SOH).

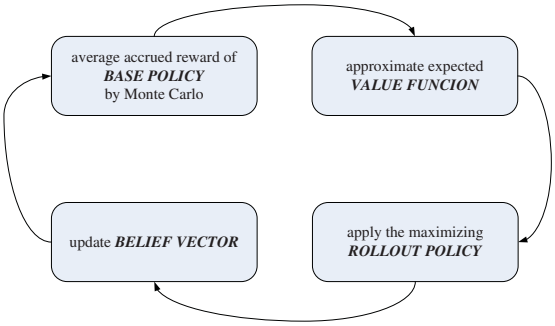

Fig. 2. Rollout-based joint spectrum sensing and access policy

In BOH, the sensing and access sets A_1 and A_2 are chosen to maximize the expected available bandwidth, i.e.,

$$\mu_m^{\mathcal{H}_1} : \Delta(m) \rightarrow a^{\mathcal{H}_1}(m) = \arg\max_{a\in\mathbf{A}} \sum_{i\in A_2} P_i(A_1)\cdot BW \quad (19)$$

where $P_i = \Pr\{S_i = 1\}$ can be updated based on the sensing action A_1. Intuitively, the wider the available bandwidth is, the better the requirement of SU would be satisfied, and the less likely a channel switch will be triggered in next time slot. However, in BOH, the statistics of the primary traffic is not taken into consideration to predict channel dynamics.

On the other hand, in SOH, the spectrum sensing and access actions are chosen to maximize the expected reward (i.e., the length of current control epoch),

$$\mu_m^{\mathcal{H}_2} : \Delta(m) \rightarrow a^{\mathcal{H}_2}(m) = \arg\max_{a\in\mathbf{A}} \sum_{\kappa^m} \kappa^m(a) p_{\kappa^m}(a) \quad (20)$$

where the calculation of p_{κ^m} includes the operation of predicting the access probability ζ and the switching probability ξ. Making full use of the dynamic statistics of the channels, the SOH algorithm is expected to perform better than BOH.

We would like to emphasis that both heuristic algorithms are greedy approaches with low computational complexity. Adopting either of them as the base policy, the expected reward from current control epoch to the end of the time horizon can be calculated in a recursion way with the initial condition $V_{\mathcal{H}}^0(\Delta) = 0$:

$$V_{\mathcal{H}}^m(\Delta) = E\left\{\kappa^m(a^{\mathcal{H}}) + V_{\mathcal{H}}^{m-1}(\Omega(\Delta|a^{\mathcal{H}},\theta))\right\} \quad (21)$$

Rollout Policy Based on the base policy $\pi^{\mathcal{H}}$, the rollout policy $\pi^{RL} = [\mu_1^{RL}, \mu_2^{RL}, ..., \mu_T^{RL}]$ is defined by the following operation.

$$\mu_m^{RL} : \Delta(m) \rightarrow a^{RL}(m) \quad (22)$$

$$a^{RL}(m) = \arg\max_{a\in\mathbf{A}} E\left\{\kappa^m(a) + V_{\mathcal{H}}^{m-1}(\Delta(m-1))\right\} \quad (23)$$

By rolling out the heuristic algorithm and observing the performance of a set of base policy solutions, useful information can be obtained to guide the search for the rollout policy solution. The rollout policy can approximate the value function according to the reward of the base policy, and consequently decide the action $a^{RL}(m)$.

In terms of efficiency, we establish in the following proposition that the rollout policy is guaranteed to improve substantially the performance of the base heuristics.

Proposition (Improving Property of Rollout Policy) The rollout policy is guaranteed to lead to better aggregated reward than the base policy. Mathematically, the following inequality holds:

$$\begin{aligned} V_{\mathcal{H}}^T(\Delta(T)) &\leq E\left\{\kappa^T(a^{RL}(T)) + V_{\mathcal{H}}^{T-1}(\Delta(T-1))\right\} \\ &\quad \cdots \\ &\leq E\{\kappa^T(a^{RL}(T)) + \kappa^{T-1}(a^{RL}(T-1)) \\ &\quad + \cdots + \kappa^m(a^{RL}(m)) + V_{\mathcal{H}}^{m-1}(\Delta(m-1))\} \\ &\quad \cdots \\ &\leq E\{\kappa^T(a^{RL}(T)) + \kappa^{T-1}(a^{RL}(T-1)) \\ &\quad + \cdots + \kappa^1(a^{RL}(1))\}. \end{aligned} \quad (24)$$

Proof: We prove the proposition by backward induction. For $m = T$, it follows from (23) that

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

Consequently,

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

The proposition holds for $m = T$.

Assume the proposition holds for $m < T$ i.e.:

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

It follows from (23) that

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

We then have

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

Consequently, it holds that

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

Therefore, the proposition holds for $m-1$. We thus complete the proof. ■

We now investigate the implementation of the proposed rollout policy. To that end, define the Q-factor $Q_m(a)$ as the expected reward that the SU can obtain from the current control epoch to the end of the time horizon, i.e.,

$$Q_m(a) \triangleq E\left\{\kappa^m(a) + V_{\mathcal{H}}^{m-1}(\Delta(m-1))\right\}. \quad (25)$$

TABLE I
SIMULATION CONFIGURATION

Parameter	Setting
Total number of channels N	20
Number of sensing channels L	5
Bandwidth per channel BW	10 MHz
Aggregation range Γ	80 MHz
Bandwidth requirement Υ	60 MHz
Duration of time slot T_p	2 ms

The rollout policy can be expressed as $a^{RL}(m) = \arg\max_{a\in\mathbf{A}} Q_m(a)$. Since the Q-factor may not be known in closed form, the rollout action $a^{RL}(m)$ cannot be calculated directly.

To overcome this difficulty, we adopt a widely applied approach to compute the rollout action, the Monte Carlo method [15]. Specifically, we define the *trajectory* as a sequence of the form

$$(\{\mathbf{S}(T), a(T)\}, \{\mathbf{S}(T-1), a(T-1)\}, \cdots, \{\mathbf{S}(1), a(1)\}). \tag{26}$$

To implement the Monte Carlo approach, we consider all possible actions $a \in \mathbf{A}$ and generate a number of trajectories of the system starting from the belief vector $\Delta(m)$, using a as the first action and the base policy $\pi^{\mathcal{H}}$ thereafter. Under this setting, a trajectory has the following form:

$$\left(\{\mathbf{S}(m), a\}, \{\mathbf{S}(m-1), a^{\mathcal{H}}(m-1)\}, \cdots, \{\mathbf{S}(1), a^{\mathcal{H}}(1)\}\right) \tag{27}$$

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 past actions and observations:

$$\Delta(i-1) = \begin{cases} \Omega(\Delta|a^{\mathcal{H}}(i), \theta) & i = m-1, m-2, \cdots, 1 \\ \Omega(\Delta|a, \theta) & i = m \end{cases} \tag{28}$$

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

$$\widetilde{a}^{RL}(m) = \arg\max_{a\in\mathbf{A}} \widetilde{Q}_m(a) \tag{29}$$

IV. PERFORMANCE EVALUATION

In this section, we evaluate the performance of the proposed rollout-based spectrum sensing and access scheme by simulation. The effects of both the number of Monte Carlo random trajectories and the proportion of sensing channels L/N are investigated. The primary network traffic statistics follows the model of Erlang-distribution [9]. The settings of parameters in the simulation are listed in Table I. For each policy, we run 100 simulations with random channel states to obtain the average performance, i.e. average times of channel switch per slot.

Fig. 3 traces the value of approximate Q-factor $\widetilde{Q}_m(a)$ with different number of Monte Carlo random trajectories.

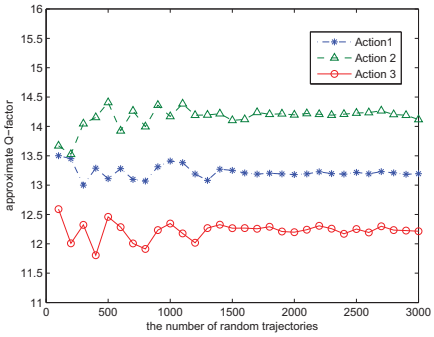

Fig. 3. Convergence with different number of random trajectories.

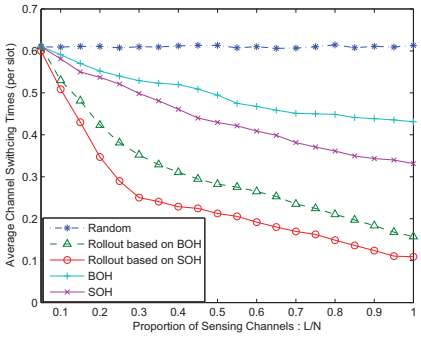

Fig. 4. Performance comparison

Three curves represent different rollout actions $a_1, a_2, a_3 \in \mathbf{A}$ chosen in the current control epoch. It is shown that, for all the three actions, the fluctuation range of $\widetilde{Q}_m(a)$ decreases with the increase of the number of random trajectories. When the number of trajectories exceeds 1500, the approximate value converges, which approaches the original value of Q-factor. In the rest of simulation results, we adopt 1500 random trajectories for approximation, which achieves the convergent performance.

Fig. 4 illustrates the effect of the proportion of sensing channels L/N on the performance of the rollout-based policy. The rollout policies based on both BOH and SOH are evaluated. The random scheme is adopted as a baseline for performance comparison, in which M channels are chosen randomly to access.

In Fig. 4, it is observed that the average times of channel switch using BOH, SOH, BOH-based and SOH-based rollout schemes reduces as the number of sensing channels L increases. This is because the more channels the SU senses, the more accurate information about the system state can be obtained. The access action determined on the basis of sensing results has better performance in minimizing the expected times of channel switches. On the contrary, 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 very limited spectrum can be sensed, the performances of all the five schemes are almost the same, for the reason that L

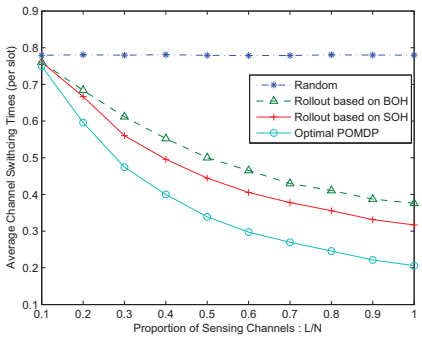

Fig. 5. Performance comparison with the optimal scheme.

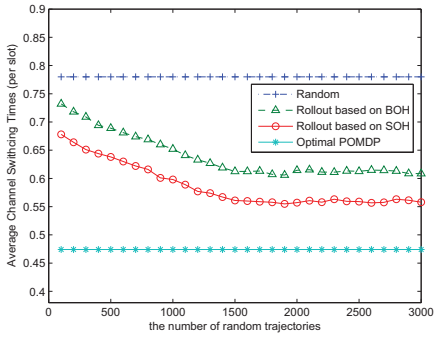

Fig. 6. Performance improvement with the increase of the number of random trajectories.

is the main limiting factor of the system performance for the moment. With larger L, the rollout-based spectrum sensing and access schemes achieve much better performance than the basis heuristics and the random scheme. Especially, the suboptimal scheme based on the SOH algorithm outperforms that based on BOH, which implies that the choice of the base policy has non-neglectable effects to the performance of the corresponding rollout policy. When the heuristic scheme performs good, the corresponding rollout policy based on it achieves relatively better performance.

For the performance comparison with the optimal scheme, due to the unacceptable computational complexity of the exact optimal policy, we adopt a new simulation setting in which $N = 10$ independent channels are considered, the maximum span of the aggregation region Γ is set to $40MHz$, and the bandwidth requirement $\Upsilon = 20MHz$.

Fig. 5 compares the performance of the proposed rollout-based policy with the optimal one. We can observe from the result that both the optimal and rollout-based policies achieve significant performance gain compared with the random selection policy with the optimal policy slightly outperforming the rollout-based policy.

Fig. 6 evaluates the performance of the rollout-based policy with different number of random trajectories when $L = 3$. The performance of the rollout-based policy becomes closer and closer to the optimal one until the number of random trajectories converge. When more than 1500 trajectories are considered, the performance gain is not significant. It can be also observed that the rollout-based policy with SOH as base heuristic performs better than that with BOH.

V. Conclusion

In this paper, we have studied the problem of joint spectrum sensing and access under the hardware limitations of both sensing capability and spectrum aggregation. Motivated by the analysis that the optimal policy is PSPACE-hard. We have developed a rollout-based policy in which two heuristic policies are proposed to serve as base policies, based on which the developed rollout-based policy approximates the value function and calculates the appropriate spectrum sensing and access actions. We have established mathematically that the rollout-based policy achieves better performance than the base policies. We have also demonstrated that the rollout-based policy leads to order of magnitude gain in terms of computation complexity compared with the optimal policy at the price of only slight performance loss.

References

[1] J. Mitola, G. Maguire, "Cognitive radio: making software radios more personal", *IEEE Personal Commun.*, vol. 6, no. 4, pp. 13–18, Aug 1999

[2] Q. Zhao, L. Tong, A. Swami, Y. Chen, "Decentralized Cognitive MAC for Opportunistic Spectrum Access in Ad Hoc Networks: A POMDP Framework", *IEEE J. Selected Areas in Commmun.*, vol. 25, no. 3, Apr 2007

[3] 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*, Apr 2010

[4] 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 Computing*, vol. 10, no. 7, pp. 1011–1027, Jul 2011

[5] W. Wang, Z. Zhang, A. Huang, "Spectrum Aggregation: Overview and Challenges", *Network Protocols and Applications*, vol. 2, no. 1, pp. 184–196, May 2010

[6] L. Wu, W. Wang, Z. Zhang, "A POMDP-based Optimal Spectrum Sensing and Access Scheme for Cognitive Radio Networks with Hardware Limitation", *Proc. of IEEE WCNC 2012*, Apr 2012

[7] G.E. Monahan, "A Survey of Partially Observable Markov Decision Processes: Theory, Models, and Algorithms", *Management Science*, vol. 28, no. 1,pp. 1–16, Jan 1982

[8] D.P. Bertsekas, J.N. Tsitsiklis, "Neuro-Dynamic Programming: an overview", *Proc. of 34th IEEE Conference on Decision and Control*, Dec 1995

[9] H. Kim and K.G. Shin, "Efficient Discovery of Spectrum Opportunities with MAC-Layer Sensing in Cognitive Radio Networks", *IEEE Trans. Mobile Computing*, vol. 7, pp. 533–545, May 2008

[10] R. Smallwood and E. Sondik, "The optimal control of partially observable Markov processes over a finite horizon", *Operation Research*, vol. 21, no. 5, pp. 1071–1088, 1973

[11] D. Braziunas, "*POMDP solution methods*", 2003

[12] B.V. Gendenko and A.N. Kolmogorov, "*Limit Distributions for Sums of Independent Random Variables*", MA: Addison-Wesley, 1954

[13] D.P. Bertsekas, J.N. Tsitsiklis, C. Wu, "Rollout algorithms for combinatorial optimization", *Journal of Heuristics*, vol. 3, no. 2, pp. 245–262, 1997

[14] D.P. Bertsekas, D.A. Castanon, "Rollout algorithms for stochastic scheduling problems", *Journal of Heuristics*, vol. 5, no. 1, pp. 89–108, 1998

[15] G. Tesauro, G.R. Galperin, "On-line policy improvement using Monte Carlo search", *Neural Information Processing Systems Conference*, 1996