Dynamic Cooperative Secondary Access in Hierarchical Spectrum Sharing Networks

Liping Wang and Viktoria Fodor
School of Electrical Engineering and the ACCESS Linnaeus Center
KTH Royal Institute of Technology
Email: {lipingw,vfodor}@kth.se

***Abstract*—We consider a hierarchical spectrum sharing network consisting of a primary and a cognitive secondary transmitter-receiver pair, with non-backlogged traffic. The secondary transmitter may utilize cooperative transmission techniques to relay primary traffic while superimposing its own information, or transmit opportunistically when the primary user is idle. The secondary user meets a dilemma in this scenario. Choosing cooperation it can transmit a packet immediately even if the primary queue is not empty, but it has to bear the additional cost of relaying, since the primary performance needs to be guaranteed. To solve this dilemma we propose dynamic cooperative secondary access control that takes the state of the spectrum sharing network into account. We formulate the problem as a Markov Decision Process (MDP) and prove the existence of a stationary policy that is average cost optimal. Then we consider the scenario when the traffic and link statistics are not known at the secondary user, and propose to find the optimal transmission strategy using reinforcement learning. With extensive numerical evaluation, we demonstrate that dynamic cooperation with state aware sequential decision is very efficient in spectrum sharing systems with stochastic traffic, and show that dynamic cooperation is necessary for the secondary system to be able to adapt to changing load conditions or to changing available energy resource. Our results show, that learning based access control, with or without known primary buffer state, has close to optimal performance.**

***Index Terms*—Hierarchical spectrum sharing, cooperative transmission, queuing systems, Markov decision process, reinforcement learning.**

I. Introduction

Hierarchical spectrum sharing among users of different networks is a promising solution to improve the spectrum efficiency, and thus to alleviate the spectrum shortage problem caused by the rapidly growing demand for wireless applications and services. Under hierarchical spectrum sharing the higher priority, primary users (PUs) have performance guarantees, whereas the secondary, low priority users (SUs) need to be cognitive, and adjust their access strategies so that the primary performance does not degrade. One traditional paradigm to facilitate hierarchical spectrum sharing is opportunistic spectrum sharing [1], where SUs identify the time-frequency resources unused by the PUs [2][3][4], and exploit them for their own transmissions [5][6][7].

Thanks to the development of advanced signal processing and interference management techniques, cooperative spectrum sharing is considered as an alternative way to share spectrum more efficiently between primary and secondary users. Instead of transmitting in idle time or on idle frequency, the SUs relay primary packets, and transmit their own packets with superimposed signal [8][9][10][11], or with a time-[12][13][14], frequency-[15][16], or space-[17] division based cooperative relaying scheme. With appropriate cooperation between the two networks, the throughput or the power efficiency of the PUs can be guaranteed or improved, whereas the SUs gain more transmission opportunities. The literature on cooperative spectrum sharing networks addresses the optimal relay selection and resource allocation [12][13][14], and presents scaling laws [18][19][20], assuming that users are always willing to cooperate and always have packets to transmit. Non-backlogged traffic is considered for various cooperative spectrum sharing schemes in [21][22][23][24]. These works characterize the stable-throughput region under stochastic packet arrival and transmission processes, however, still assuming that the SUs always cooperate.

In this paper we formulate the dilemma of the secondary user that needs to transmit a stream of randomly arriving packets. As the primary performance needs to be guaranteed, the SU may have significant cost of cooperation, in terms of increased transmission power or very low probability of successful transmission. Therefore, under dynamic load, the SU may instead wait for idle time and transmit opportunistically. Note that the PU does not have this dilemma, since its transmission performance is guaranteed in the hierarchical spectrum sharing scenario. We define the cost of secondary access as the combination of the cost of cooperation and the cost of additional packet delay, and define dynamic cooperative secondary access schemes that aim at minimizing the long-run average cost. We provide a Markov Decision Process (MDP) formulation to derive optimal access schemes when the load and the network parameters are known for the secondary system, and evaluate the performance of reinforcement learning for the cases when this knowledge is not available.

The MDP framework and its variations such as constrained MDP and Partially Observable MDP (POMDP) [25] have been used extensively for optimizing control strategies in discrete time stochastic systems in general. In the area of spectrum sharing networks with opportunistic secondary access MDP has been used to design sensing and access strategies for the SUs, when the primary traffic can be modeled with some known stochastic processes [2][3][5][6]. Reinforcement learning techniques, such as Q-learning [26] and R-learning [27] provide online optimization tools that can solve MDPs it-

eratively without a priori knowledge of the state transition probabilities, and therefore, has been used for secondary access or interference control design, when the stochastic behavior of the primary system, or of the other interfering secondary systems are not known a priori [4][6][7].

In this paper we consider a primary and a secondary node pair, both with dynamic traffic and unreliable transmission channel. The SU aims at trading off the cost of cooperative relaying and the increased delay of opportunistic transmission. The main contributions of the paper are summarized as follows:

- We define four secondary access strategies in the hierarchical spectrum sharing system with opportunistic, cooperative, sequential decision and random cooperation based access, and derive the stable-throughput region, that is, the maximum arrival rates when both the primary and the secondary transmission queues are strongly stable.
- To find optimal sequential decision policy, we formulate the dilemma of the secondary user as an MDP with the objective of minimizing the long-run average cost, where the cost of an action includes the cost of cooperative transmission and the cost of secondary queuing delay. We show that the long-run average cost is upper bounded within the stable throughput region. We prove the existence of an optimal sequential decision that is stationary and propose an approximation method based on linear programming to find the optimal policy and the minimum long-run average cost for infinite buffer systems.
- We consider the case of unknown primary and secondary traffic and link statistics, and propose R-learning to find optimal sequential decision policy based on the experienced cost, including the scenario when the secondary user receives request for cooperation from the primary user but is not aware of the primary queue length.
- We investigate the performance of optimal sequential decision by extensive simulations, and show that it can significantly reduce the average cost compared to the cost of pure opportunistic and pure cooperative spectrum sharing schemes, especially when the traffic intensity of the secondary user is moderate and that of the primary is not very low. Random cooperation, in turn, has lower gain, but can achieve a performance close the the better one of the static opportunistic or cooperative schemes. We show that the performance of sequential decision under learning is close to optimal, even with limited knowledge on the primary queue size.

The rest of the paper is organized as follows. We introduce the system model and the four different spectrum sharing schemes in Section II. The stable-throughput regions of the considered schemes are evaluated in Section III. In Section IV, we define and discuss the MDP to achieve the optimal sequential decision of the secondary system, and in Section V give the R-learning formulation. The performance of optimal sequential decision based spectrum sharing is evaluated in Section VI. Finally, we conclude the paper in Section VII.

II. Spectrum sharing schemes and System model

We consider a spectrum sharing network, where a primary user and a secondary user, PT and ST, intend to transmit packets to their respective destinations, PR and SR, via a shared wireless channel. Time is slotted and the transmission of each packet takes one time slot. The PT and the ST can transmit in separate time slots directly to the destinations, or they can use a cooperative transmission scheme, where the ST relays the primary packet toward the PR, at the same time superimposing its own packet to the SR [8].

We compare two static and two dynamic secondary access schemes:

- Opportunistic spectrum sharing: the PT transmits directly to the PR. The ST senses the channel at the beginning of each time slot. If the channel is idle, the ST transmits a packet directly to PR, and it keeps silent otherwise.
- Cooperative spectrum sharing: the ST always relays the primary packet, and superimposes its own packet, if the the PT transmission queue is non-empty. If the PT is idle, the ST transmits directly to the SR.
- Sequential decision: in each time slot the ST decides about cooperating or not, where the decision can depend on the current state of the system. The ST needs to be aware of the state of the two queues, that is, the number of packets waiting, Q_p and Q_s, and needs to inform the PT about its decision. If the PT does not transmit, the ST uses direct transmission.
- Random cooperation: in each time slot the ST makes a random decision, to cooperate or to transmit opportunistically, with fixed probabilities, and informs the PT about its decision. Again, the ST transmits directly to SR if the PT is silent. This scheme can be considered as a special case of the sequential decision scheme where the decision does not depend on the system state.

To compare the performance of the above schemes we consider the stable-throughput region and the long-run average cost of the secondary system. To reflect the secondary dilemma of to wait or to cooperate, the cost reflects the increased delay of opportunistic access and the increased power consumption of cooperative transmission. Specifically, in each time slot the accumulated secondary cost is increased with $C_h Q_s$ that reflects the secondary queuing delay and in addition with C_c, the cost of cooperation, if the secondary node performs cooperative transmission.

We model the primary and the secondary system as follows. Since packets may not be received successfully due to the impairments of the radio channel, we model the packet transmission on each end-to-end communication link via independent Bernoulli processes, as an abstraction of various cooperative transmission techniques and channel models. Let q_{pd} and q_{sd} denote the probabilities of successful packet transmission in a time slot under direct transmission at the primary and secondary systems, respectively, whereas q_{pc} and q_{sc} represent the respective transmission success probabilities in the cooperative transmission case. The expressions for calculating these probabilities can be found for instance in [8] for a spectrum sharing networks using a two-phase cooperative

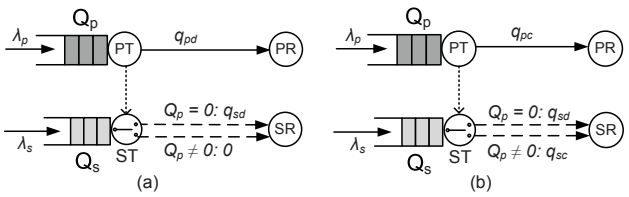

Fig. 1. Queuing network modeling the spectrum sharing system with (a) opportunistic and (b) cooperative spectrum sharing.

decode-and-forward relaying protocol under Rayleigh fading channels. Since the secondary cooperative transmission needs to improve or at least guarantee the probability of successful packet transmission at the PT, but may decrease the transmission success probability at the ST, we consider $q_{pc} \geq q_{pd}$ and $q_{sc} \leq q_{sd}$. We model the packet arrival at the primary and secondary users by independent Bernoulli processes with per slot packet arrival probabilities λ_p and λ_s, respectively. Our work can be extended to consider Markov modulated arrival and loss processes. Both PT and ST have a buffer of infinite capacity for storing the incoming packets and they follow FIFO transmission policy. Packets that are not received successfully are retransmitted. We assume that ACK/NACK messages from the PR and the SR do not get lost.

The resulting queuing networks for opportunistic and cooperative spectrum sharing are shown in Fig. 1(a) and (b), respectively. We can see that in both cases the primary and the secondary queues are coupled, more precisely, the service rate of the ST depends on the status of the queue at the PT.

III. The Stable-Throughput Region

Let us first evaluate the stable-throughput regions of the considered spectrum sharing methods. We follow the notion of *strong stability* given in [28][29], for slotted systems with time varying service process. To define strong stability, let us denote the queue length at the beginning of a time slot n as $Q(n)$, the number of arrivals in time slot n as $A(n)$, and the service rate as $B(n)$. We assume that arrivals happen at the end of the time slot, and can be served only in the following slot. Then, the queue evolves, as:

$$Q(n+1) = \max[Q(n) - B(n), 0] + A(n).$$

Definition 1. *The queue is strongly stable, if [28]:*

$$\limsup_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} E[Q(n)] < \infty. \quad (1)$$

That is, the queue is strongly stable, if it has a bounded average queue length. A network of queues is strongly stable, if all queues are strongly stable. The stable-throughput region of a queuing network is given by the arrival rate vectors for which the network of queues is strongly stable.

Proposition 1. *The following conditions are sufficient for strong stability.*

1) *The arrival and service processes are rate convergent, that is:*

$$\lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} E[A(n)] = \lambda, \ \lim_{N\to\infty} \frac{1}{N} \sum_{n=1}^{N} E[B(n)] = \mu,$$

and for each positive δ there exists an N, such that, regardless of past history:

$$E\left\{\frac{1}{N}\sum_{n=1}^{N} A(n_0+n)\right\} \leq \lambda+\delta, \ E\left\{\frac{1}{N}\sum_{n=1}^{N} B(n_0+n)\right\} \geq \mu-\delta;$$

2) *in each time-slot the arrival process is bounded in second moment, and the service process is bounded, regardless of past history; and finally*
3) *the arrival rate is less than the average service rate, that is $\lambda < \mu$.*

Proof: The proof of the proposition is given in [29]. ■

Theorem 1. 1) *The stable-throughput regions $\mathcal{S}_D$ and $\mathcal{S}_C$, for opportunistic and cooperative spectrum sharing respectively, are:*

$$\mathcal{S}_D = \left\{(\lambda_p, \lambda_s) : \lambda_p < q_{pd} \text{ and } \lambda_s < q_{sd} - \frac{q_{sd}}{q_{pd}}\lambda_p\right\},$$

and

$$\mathcal{S}_C = \left\{(\lambda_p, \lambda_s) : \lambda_p < q_{pc} \text{ and } \lambda_s < q_{sd} + \frac{q_{sc}-q_{sd}}{q_{pc}}\lambda_p\right\},$$

while $\mathcal{S}_R$, the stable-throughput region of random cooperation with cooperation probability p_c is:

$$\mathcal{S}_R = \left\{(\lambda_p, \lambda_s) : \begin{array}{l} \lambda_p < (1-p_c)q_{pd} + p_c q_{pc} \\ \lambda_s < q_{sd} + \frac{p_c q_{sc} - q_{sd}}{(1-p_c)q_{pd} + p_c q_{pc}}\lambda_p \end{array}\right\}.$$

2) *If the ST performs sequential decisions Π, then the stable-throughput region $\mathcal{S}_\Pi$ is respectively lower- and upper-bounded by $\mathcal{S}_D$ and $\mathcal{S}_C$:*

$$\mathcal{S}_D \subseteq \mathcal{S}_\Pi \subseteq \mathcal{S}_C.$$

Proof: To derive the stable-throughput region of the spectrum sharing system, we have to prove that conditions 1 and 2 in Proposition 1 hold for the secondary and primary arrival and service processes, and find the arrival rates λ_p and λ_s, where condition 3 holds as well.

Let $A_p(n)$ and $A_s(n)$ denote the number of arrivals in time slot n at the PT and ST, respectively, whereas $B_p(n)$ and $B_s(n)$ are the respective service rates. Since the considered primary and secondary arrival processes are iid Bernoulli processes with $E[A_p(n)] = \lambda_p$ and $E[A_s(n)] = \lambda_s$, they are rate convergent with bounded second moment, that is, conditions 1 and 2 on the arrival processes hold.

The service processes depend on the spectrum sharing scheme. For all schemes however $B_p(n)$ and $B_s(n)$ can take values 0 or 1, that is, condition 2 for the service processes holds. For all schemes, the service processes are either iid Bernoulli processes, or are controlled by the state of the queues, that in turn, when stable, can be described with ergodic discrete time birth-deaths processes, that converge monotonically to steady state [30]. (For example, under opportunistic spectrum sharing, the primary service process is an iid Bernoulli process with $E[B_p(n)] = q_{pd}$, while the secondary service process is controlled by the primary queue length Q_p, $E[B_s(n)|Q_p(n) = 0] = q_{pd}$ and $E[B_s(n)|Q_p(n) > 0] = 0$.) Consequently, the primary and secondary service processes are rate convergent, and thus, condition 1 for the service

processes is fulfilled as well, for all considered spectrum sharing schemes.

Let us now find the λ_p, λ_s pairs when condition 3 holds. These arrival rate pairs give the stable-throughput region of the system.

Consider first opportunistic spectrum sharing. The PT transmits a packet with success probability q_{pd} whenever its queue is non-empty, independently from the state of the ST. However, the ST can transmit a packet with success probability q_{sd} if and only if the primary queue is empty. Therefore μ_{pd} and μ_{sd}, the primary and secondary service rates under direct transmission can be derived as:

$$Pr[B_p(n)=1]=q_{pd},\quad Pr[B_p(n)=0]=1-q_{pd},\\ \Rightarrow\ \mu_{pd}=q_{pd}, \tag{2}$$

$$\left.\begin{aligned} &Pr[B_s(n)=1]=q_{sd}\mathbf{P}[Q_p(n)=0],\\ &Pr[B_s(n)=0]=1-q_{sd}\mathbf{P}[Q_p(n)=0],\end{aligned}\right\}\\ \Rightarrow\ \mu_{sd}=q_{sd}\mathbf{P}[Q_p=0]. \tag{3}$$

Since $\mathbf{P}[Q_p=0]=1-\lambda_p/\mu_{pd}$, condition 3, that is $\lambda_p<\mu_{pd}$ and $\lambda_s<\mu_{sd}$, gives the following stable-throughput region for opportunistic spectrum sharing:

$$\mathcal{S}_D=\left\{(\lambda_p,\lambda_s):\lambda_p<q_{pd}\text{ and }\lambda_s<q_{sd}-\frac{q_{sd}}{q_{pd}}\lambda_p\right\}.$$

Similarly, we can derive the stable-throughput region under cooperative spectrum sharing. Then, the PT transmits a packet with success probability q_{pc} whenever its queue is non-empty, using cooperative transmission. The ST transmits a packet with success probability q_{sd} if the primary queue is empty, and with success probability q_{sc} otherwise. Consequently, the average service rates under cooperation, μ_{pc} and μ_{sc}, respectively seen by the primary and secondary sources are:

$$\mu_{pc}=q_{pc}, \tag{4}$$

$$\mu_{sc}=q_{sd}\mathbf{P}[Q_p=0]+q_{sc}\mathbf{P}[Q_p\neq 0]. \tag{5}$$

Under $\lambda_p<\mu_{pc}$ and $\lambda_s<\mu_{sc}$ we get $\mathcal{S}_C$, the stable-throughput region of cooperative spectrum sharing:

$$\mathcal{S}_C=\left\{(\lambda_p,\lambda_s):\lambda_p<q_{pc}\text{ and }\lambda_s<q_{sd}+\frac{q_{sc}-q_{sd}}{q_{pc}}\lambda_p\right\}.$$

Clearly, $\mathcal{S}_D\subseteq\mathcal{S}_C$ for all $q_{pc}\geq q_{pd}$ and $q_{sc}\leq q_{sd}$.

To consider random the cooperation scheme, let p_c denote the probability that the secondary user chooses to cooperate, given that $Q_p\neq 0$. The average service rates under random cooperation, μ_{pr} and μ_{sr}, become:

$$\mu_{pr}=(1-p_c)q_{pd}+p_cq_{pc},$$

$$\mu_{sr}=q_{sd}\mathbf{P}[Q_p=0]+p_cq_{sc}\mathbf{P}[Q_p\neq 0],$$

which gives $\mathcal{S}_R$, the stable-throughput region of the random cooperation:

$$\mathcal{S}_R=\left\{(\lambda_p,\lambda_s):\ \begin{array}{l}\lambda_p<(1-p_c)q_{pd}+p_cq_{pc}\\ \lambda_s<q_{sd}+\frac{p_cq_{sc}-q_{sd}}{(1-p_c)q_{pd}+p_cq_{pc}}\lambda_p\end{array}\right\}.$$

Opportunistic and cooperative spectrum sharing are special cases of random cooperation with $p_c=0$ and $p_c=1$, respectively. As expected, $\mathcal{S}_R$ is equivalent to $\mathcal{S}_D$ when $p_c=0$, and to $\mathcal{S}_C$ when $p_c=1$.

Let us now evaluate the stable-throughput region of the sequential decision scheme, following the *dominant system* approach. We consider a system X to be a dominant system of Y, if the queue sizes in X are, at all times, at least as large as those in Y. The stable-throughput region of the dominant system X inner bounds that of Y [23][31].

By comparing the average service rate of PT in (2) with that in (4), and the average service rate of ST in (3) with that in (5), we get $\mu_{pd}\leq\mu_{pc}$ and $\mu_{sd}\leq\mu_{sc}$. So for any sequential decision scheme Π, the primary and secondary service rates are bounded as $\mu_p^\Pi\in[\mu_{pd},\mu_{pc}]$ and $\mu_s^\Pi\in[\mu_{sd},\mu_{sc}]$. Consequently, any sequential decision scheme stochastically dominates the opportunistic one, and is dominated by the cooperative one, that is, $\mathcal{S}_D\subseteq\mathcal{S}_\Pi\subseteq\mathcal{S}_C$. ■

Fig. 2 gives an example of the stable-throughput region. The shaded area shows the improvement achieved by cooperation, which is significant if q_{pc} is larger than q_{pd}, and q_{sc} is close to q_{sd}. According to Theorem 1, the boundary of the stable-throughput region of any $\mathcal{S}_\Pi$ is located in the shaded area.

IV. Optimal Sequential Decision Policy for the Secondary System

A. MDP formulation of sequential decision

We use a Markov Decision Process (MDP) to model the sequential decision of the secondary user, and to find the optimal decision policy when the system parameters, that is, the packet arrival probabilities (λ_p,λ_s) and the successful transmission probabilities $(q_{pd},q_{pc},q_{sd},q_{sc})$, as well as the system state, that is, the primary and secondary queue lengths (Q_p,Q_s) are known for the ST.

In general, an MDP describes a stochastic control system, whose state can be observed in discrete time. At each time slot, the decision maker chooses an action depending on the present state or the history of the process. An immediate cost (or reward) incurs after taking the action, and the system moves to a state with some transition probability that is determined by the present state and the selected action.

The MDP we formulated is defined as MDP$\langle\mathcal{S},\mathcal{A},\mathbb{A},p,c\rangle$, where

- $\mathcal{S}=\{(Q_p,Q_s),Q_p\in\mathbb{N}^0\text{ and }Q_s\in\mathbb{N}^0\}$: the countable set of discrete states, each of which is defined as the queue length pair.
- $\mathcal{A}=\{0,1\}$: the set of control actions taken by the secondary system, where 0 denotes the case that the ST chooses to access the channel opportunistically, and 1 refers to cooperative transmission.
- $\mathbb{A}:\mathcal{S}\to\mathcal{P}(\mathcal{A})$: the action constraint function. $A(s)$ denotes the set of allowed actions in state s. We have:
$$A(s)=\begin{cases}\{0\} & \text{if } s\in\{(Q_p,Q_s),Q_p=0\text{ and }Q_s\in\mathbb{N}^0\}\\ \{0,1\} & \text{if } s\in\{(Q_p,Q_s),Q_p\neq 0\text{ and }Q_s\in\mathbb{N}^0\}\end{cases}.$$
- $p:\mathcal{S}\times\mathcal{A}\to\Delta(\mathcal{S})$: the transition function, where $\Delta(\mathcal{S})$ denotes the set of all probability distributions on $\mathcal{S}$. The probability that the process moves to state s' after taking action a in state s is given by $p_a(s,s')=$

$\mathbf{P}(s_{t+1} = s' | s_t = s, a_t = a)$, which depends on the arrival rates, and also on the state and action dependent service rates. The derivation of state transition probabilities is straightforward, examples are given in [32].

- $c : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$: the cost function $c(s, a)$ denoting the immediate cost that depends on the present state and the selected action. A general cost function for queue-length controlled system is $c(s, a) = B(a) + H(s)$, where $B(a) \geq 0$ represents the cost of selecting action a, and $H(s) \geq 0$ is a cost that depends on the system state.
  We consider:

$$c(s, a) = B(a) + H(s) = C_t a + C_h Q_s, \tag{6}$$

  that is, we set $B(a) = C_t a$ to represent the additional power consumption of relaying the primary packet, and $H(s) = C_h Q_s$, where C_h denotes the cost of the ST for holding one packet in its queue for one time slot and thus $H(s) = C_h Q_s$ represents the cost of secondary queuing delay.

In each state, actions are chosen by following a policy Π, which defines a rule for decisions that may depend on the current state, on the past history of the process, and on the time. A policy is Markovian, if the choice does not depend on the history, and is stationary if it does not depend on the time either. The policy is random, if several actions can be selected in a state with some probabilities, and is deterministic otherwise.

Note, that both the opportunistic spectrum sharing policy Π_D and the cooperative spectrum sharing policy Π_C are deterministic stationary policies. They can be expressed respectively as:

$$\Pi_D = \{\pi_n = 0, n \in \mathbb{N}^+\}; \tag{7}$$

$$\Pi_C = \{\pi_n = i, n \in \mathbb{N}^+\} \text{ with } i = \begin{cases} 0 \text{ if } s_n = (0, j), j \in \mathbb{N}^0 \\ 1 \text{ otherwise.} \end{cases}$$

The sequential decision degrades to opportunistic spectrum sharing if $a = 0$, and to cooperative spectrum sharing if $a = 1$ for all $Q_p \neq 0$ and Q_s. It degrades to random cooperation if $\mathbf{P}(a = 0)$ is constant and independent from Q_p and Q_s for all $Q_p \neq 0$.

Our objective is to find the optimal policy Π^* that minimizes the long-run average cost $C(\Pi)$, defined as:

$$C(\Pi) = \lim_{N\to\infty} \frac{1}{N} \mathbb{E}_\Pi \left[\sum_{n=1}^{N} c(s_n, \pi_n) | s_1 = (0, 0) \right], \tag{8}$$

where s_1 denotes the initial state of the network, π_n denotes the action taken in the nth time slot according to policy Π, and $\mathbb{E}_\Pi$ is the expectation taken under policy Π.

B. Long-run average cost upper bound

Theorem 2. 1) *The long-run average cost defined in (8) achieved by any policy Π is upper bounded when the arrival rates lie within the stable-throughput region $\mathcal{S}_\Pi$. Moreover, upper bounded average cost implies the strong stability of Q_s.*

2) *If the ST makes sequential decisions according to the optimal policy Π^*, then the long-run average cost is upper bounded when the arrival rates lie within the stable-throughput region $\mathcal{S}_C$.*

Proof: For $c(s, a) = C_t a + C_h Q_s$, the long-run average cost in (8) becomes:

$$C(\Pi) = C_c \lim_{N\to\infty} \frac{1}{N} \mathbb{E}_\Pi \left[\sum_{n=1}^{N} \pi_n \right] + C_h \lim_{N\to\infty} \frac{1}{N} \mathbb{E}_\Pi \left[\sum_{n=1}^{N} Q_s(n) \right]$$
$$\leq C_c + C_h \lim_{N\to\infty} \frac{1}{N} \mathbb{E}_\Pi \left[\sum_{n=1}^{N} Q_s(n) \right] < \infty.$$

The first inequality holds because the first limit is upper bounded by always taking action 1 in every time slot. The second inequality holds as a consequence of the definition of the stable-throughput region introduced in Definition 1, considering strong stability according to (1).

Similarly, by selecting action 0, we get a lower bound on the first limit, and consequently,

$$C(\Pi) \geq C_h \lim_{N\to\infty} \frac{1}{N} \mathbb{E}_\Pi \left[\sum_{n=1}^{N} Q_s(n) \right],$$

which implies the strong stability of Q_s for $C(\Pi) < \infty$. However, the stability of Q_p is not guaranteed. This proves the first part of Theorem 2.

As the optimal policy Π^* achieves the minimum average cost under given λ_p and λ_s, this cost has to be upper bounded by the average cost of the cooperation based spectrum sharing, that is, $C(\Pi^*) \leq C(\Pi_C)$. As $C(\Pi_C) < \infty$ within $\mathcal{S}_C$, $C(\Pi^*) < \infty$ within $\mathcal{S}_C$ as well. ■

C. The existence of optimal stationary policy

We modeled the sequential decision with an infinite state MDP with unbounded costs and finite action set. We prove the existence of a stationary policy that is average cost optimal, building on the results of [33].

Following [33], let us introduce a discount factor $0 < \beta < 1$, and give the total expected discounted cost incurred by policy Π as:

$$V_{\Pi,\beta}(i, j) = \mathbb{E}_\Pi \left[\sum_{n=1}^{\infty} \beta^n c(s_n, \pi_n) | s_1 = (i, j) \right].$$

Let $V_\beta(i, j) = \inf_\Pi V_{\Pi,\beta}(i, j)$ and $h_\beta(i, j) = V_\beta(i, j) - V_\beta(0, 0)$.

Proposition 2 specifies the conditions that must be satisfied for the average cost optimal stationary policy to exist.

Proposition 2. *There exists a stationary policy that is average cost optimal for the MDP$\langle \mathcal{S}, \mathcal{A}, \mathbb{A}, p, c \rangle$ if the following conditions are satisfied:*

1) *$V_\beta(i, j)$ is finite for all (i, j) and β;*
2) *There exists a nonnegative N such that $h_\beta(i, j) \geq -N$ for all (i, j) and β;*
3) *There exists nonnegative $M_{i,j}$ such that $h_\beta(i, j) \leq M_{i,j}$ for every (i, j) and β. For every (i, j), there exists an action a such that $\sum_{k,l} p_a((i, j), (k, l)) M_{k,l} < \infty$.*

Proof: See [33]. ■

Corollary 1. *Under condition 1, the quantity $V_\beta(i,j)$ satisfies the optimality equation:*

$$V_\beta(i,j) = \min_a \left\{ c((i,j),a) + \beta \sum_{k,l} p_a((i,j),(k,l)) V_\beta(k,l) \right\}.$$

Proof: See [33]. ■

Now we can prove the existence of the optimal stationary policy for sequential decision based spectrum sharing. We give the proof for the general case when $H(s)$ in (6) is a polynomial of degree m in Q_s, and is nonnegative and nondecreasing with Q_s. The cost function considered in the paper is a specific case with $m = 1$.

Theorem 3. *For the MDP$\langle \mathcal{S}, \mathcal{A}, \mathbb{A}, p, c \rangle$, there exists a stationary policy Π^* that minimizes the long-run average cost.*

Proof: We prove the theorem by proving that conditions 1-3 in Proposition 2 hold. For the stationary policy of always accessing the spectrum opportunistically, with Π_D given in (7), the following holds:

$$V_\beta(i,j) \le V_{\Pi_D,\beta}(i,j) \le \sum_{n=1}^{\infty} \beta^n c(s_n,0)|_{s_n=(i+n,j+n)}. \quad (9)$$

If $H(s)$ in (6) is a polynomial of degree m in Q_s, the right side of (9) is a polynomial of degree m with the leading item $\sum_{n=0}^{\infty} \beta^n n^m$.

Since $\sum_{n=0}^{\infty} \beta^n n^m = (1-\beta)^{-m-1} \sum_{j=1}^{m} a_j^{(m)} \beta^j$, where $a_1^{(m)} = a_m^{(m)} = 1$ and $a_j^{(m)} = j a_j^{(m-1)} + (m-j+1) a_{j-1}^{(m-1)}$ for $j = 2, ..., m-1$, that is, $\sum_{n=1}^{\infty} \beta^n n^m$ is a summation of finite terms, $\sum_{n=0}^{\infty} \beta^n n^m < \infty$. Therefore, $V_\beta(i,j) < \infty$, and condition 1 is satisfied. Specifically, for $m = 1$, $V_\beta(i,j) \le C_h j/(1-\beta)^2$.

As the cost function is a polynomial and non-decreasing on i and j, from Corollary 1 we can observe that $V_\beta(i,j)$ is also non-decreasing in i and j. Therefore, condition 2 on finding a nonnegative N is straightforward.

Finally, let us consider condition 3. Since $V_\beta(i,j)$ is positive and non-decreasing in i and j, we have:

$$h_\beta(i,j) \le V_\beta(i,j) - 0 \le \sum_{n=0}^{\infty} \beta^n c(s_n,0)|_{s_n=(i+n,j+n)}.$$

Let $M_{i,j} = \sum_{n=0}^{\infty} \beta^n c(s_n,0)|_{s_n=(i+n,j+n)}$. Then the first part of condition 3 is fulfilled. As in the considered system there is a finite number of possible transitions from each state, the second part of the condition holds as well. ■

D. LP approximation

From Theorem 3 we know that there exists an optimal policy that is stationary. However, obtaining the optimal stationary policy is computationally prohibitive since it involves solving an MDP with a countably infinite state-space. To make the problem tractable, we aim at approximating the original MDP by a finite-state MDP with tunable number of states. Specifically, we consider the system where the PT and the ST have finite buffers for storing at most N_p and N_s packets, respectively, and arriving packets are dropped if there is no space in the corresponding queue. In this case, the state space becomes $\mathcal{S} = \{(Q_p, Q_s), Q_p \in \{0,1,...,N_p\} \text{ and } Q_s \in \{0,1,...,N_p\}\}$, whereas the action space and the cost function remains the same.

Proposition 3. *If the arrival rates are inside the stable-throughput region $\mathcal{S}_C$, the long-run average cost from the LP approximation converges as $N_p \to \infty$ and $N_s \to \infty$.*

Proof: From Theorem 2 the long-run average cost under policy Π^* is bounded in the infinite buffer system. On the other hand, simple evaluation of the underlaying discrete-time birth-death processes shows that the average secondary queue length, and with that, the average cost, are nondecreasing function of N_p and N_s. Therefore, the long-run average cost converges when $N_p \to \infty$ and $N_s \to \infty$. ■

We find the optimal stationary policy for the finite-state system by solving the following linear program (LP) [25]:

$$\begin{aligned} \{z^*_{s,a}\}_{s\in\mathcal{S}, a\in A(s)} = \arg\min & \sum_{s\in\mathcal{S}} \sum_{a\in A(s)} z_{s,a} c(s,a), \\ \text{s.t.} & \sum_{s\in\mathcal{S}} \sum_{a\in A(s)} z_{s,a} = 1, \\ & \sum_{a\in A(s')} z_{s',a} = \sum_{s\in\mathcal{S}} \sum_{a\in A(s)} z_{s,a} p_a(s,s'), \ \forall s' \in \mathcal{S}, \\ & z_{s,a} \ge 0, \ \forall a \in A(s), s \in \mathcal{S}, \end{aligned} \quad (10)$$

where $z_{s,a}$ denotes the probability that the system is in state s and chooses action a. With the optimal solution to the LP, the optimal randomized stationary policy $\Pi^* = \{\pi^*_{s,a}\}$ that minimizes the long-run average cost per unit time is computed as $\pi^*_{s,a} = z^*_{s,a} / \sum_{b\in A(s)} z^*_{s,b}$, and the objective function gives the minimum long-run average cost. As it is shown in [25], in each state there is only one action that has positive $z^*_{s,a}$, and therefore the optimal policy achieved by (10) is the optimal deterministic stationary policy. We estimate the optimal stationary policy for the original infinite buffer MDP by letting $N_p \to \infty$ and $N_s \to \infty$.

V. Sequential Decision with Online Reinforcement Learning

In the case when the system parameters or the system state are not known for the ST, MDP based optimization of sequential decision can not be applied. Therefore, we propose to adapt online reinforcement learning, specifically R-Learning to find the average cost optimal policy.

We consider two scenarios, full-state and reduced-state. In both of the cases the ST does not have full knowledge on the arrival and successful transmission probabilities.

- Full-state scenario: when it requests cooperation, the PT provides its state, Q_p to the ST, that is, the ST has full knowledge on the primary and secondary queue lengths.
- Reduced-state scenario: the PT sends only a request for cooperation, but no information on Q_p. Therefore, the ST only knows whether the primary queue is empty or not. Therefore, in this case, the original state space $\mathcal{S} = \{(Q_p, Q_s), Q_p \in \mathbb{N}^0, Q_s \in \mathbb{N}^0\}$ is reduced to $\mathcal{S}' = \{(Q_p, Q_s), Q_p \in \{0,1\}, Q_s \in \mathbb{N}^0\}$, where 0 denotes that the primary queue is empty, and 1, otherwise.

R-learning [6][27] is an average-reward reinforcement learning method, and thus can be used in both scenarios to try to find the average cost optimal policy. R-learning is based on the iterative updating of the state dependent action-value functions

called the R-factors, and the experienced average cost ρ, via a sample path. The R-factor $R_t(s,a)$ represents the expected cost of taking action a in state s given that an optimal policy is applied for all future steps. The R-learning algorithm works as follows:

1) At time $t = 1$, all the R-factors are initialized to a finite value (for instance 0), and the average cost to $\rho_1 = 0$. Let s denote the current state.
2) Action $a = \arg\min_{b\in A(s)} R_t(s,b)$ is selected with probability $1-\alpha_t$, whereas with probability α_t, an exploratory action a is chosen uniformly from $A(s)$, to ensure that the action space is explored and the learning does not converge to a local optimum.
3) Let $c(s,a)$ and s' denote the incurred cost and the next state, respectively. The R-factor is updated as:

$$R_{t+1}(s,a) \leftarrow (1-\beta_r)R_t(s,a)+\beta_r\left[c(s,a)-\rho_t+\min_{b\in A(s')} R_t(s',b)\right]. \tag{11}$$

In the case $a = \arg\min_{b\in A(s)} R_t(s,b)$ was selected, the average cost is updated as well:

$$\rho_{t+1} \leftarrow (1-\beta_\rho)\rho_t+\beta_\rho\left[c(s,a)+\min_{b\in A(s')} R_t(s',b)-\min_{b\in A(s)} R_t(s,b)\right]. \tag{12}$$

4) Let $t = t+1$ and $s = s'$, and go to step 2.

In (11) and (12), β_r and β_ρ denote the update rate of the R-factors and ρ, with $0 \le \beta_r \le 1$ and $0 \le \beta_\rho \le 1$. After convergence, the decision in state s is set to $\arg\min_{b\in A(s)} R(s,b)$.

The above R-learning algorithm runs similarly in the full-state and the reduced-state scenarios. The main difference is that $(N_p+1)\times(N_s+1)$ R-factors need to be learned in the former case, whereas only $2(N_s+1)$ ones in the latter one. For R-learning in the considered infinite buffer system the state space is extended dynamically according to the maximum experienced queue lengths.

We apply R-learning algorithm with semi-uniform exploration, with constant α_t. Other exploratory methods, like Boltzmann exploration, uncertainty estimation (UE) exploration are presented and compared in [27][34]. While the convergence of R-learning to the optimal value is not proved, the detailed evaluations show that R-learning finds near optimal solutions in most scenarios [27].

VI. Case Study

In this section, we compare the optimal sequential decision (OSD) with the opportunistic (OPP), and the cooperative (COOP) schemes, as well as with optimal random cooperation (ORC), that is, random cooperation with optimal state-independent cooperation probability, that minimizes the long-run average cost. For OSD we obtain the optimal sequential decision policy by solving the LP in (10), and perform simulations to validate the analytic solution. We find the optimal cooperation probability for ORC with exhaustive search. Moreover, we evaluate whether R-learning (RL) can provide OSD average-cost close to the optimal one.

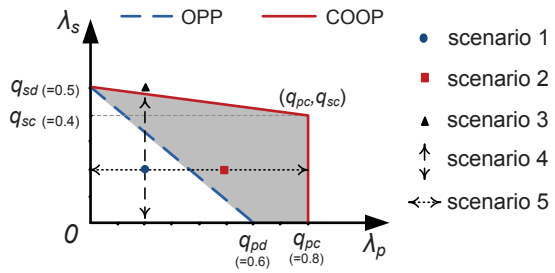

Fig. 2. Stable-throughput region for the spectrum sharing schemes.

The length of each time slot is assumed to be one time unit. All simulation results shown are average values from 10 runs, each of which lasting for 50,000 time slots.

A. Stable-throughput region and simulation scenarios

Fig. 2 shows the stable-throughput region when the PT and the ST have infinite buffers and the probabilities of successful packet transmission are set as $q_{pd} = 0.6$, $q_{pc} = 0.8$, $q_{sd} = 0.5$, and $q_{sc} = 0.4$. We keep these parameters fixed, and consider five scenarios with different sets of (λ_p, λ_s). Under scenarios 1-3, the arrival rates (λ_p, λ_s) are $(0.2, 0.2)$, $(0.5, 0.2)$ and $(0.2, 0.5)$, respectively. As shown in Fig. 2, under these three scenarios the (λ_p, λ_s) pairs are in $\mathcal{S}_\mathrm{D}$, in $\mathcal{S}_\mathrm{C}$ and outside $\mathcal{S}_\mathrm{C}$ respectively. Under scenario 4 we fix $\lambda_p = 0.2$, giving a moderate primary load, and increase λ_s until it reaches the upper bound of $\mathcal{S}_C$, whereas under scenario 5, we fix $\lambda_s = 0.2$ and increase λ_p.

B. Average cost under increasing buffer size

Let us first evaluate whether the LP approximation with limited buffer size can provide good estimate of the average cost in infinite buffer OSD systems. We set the unit costs as $C_h = C_c = 1$, and increase the buffer size of PT and ST in scenarios 1-3. Fig. 3 shows that if the (λ_p, λ_s) pair is inside the stable-throughput region of a scheme, the average cost converges as the buffer sizes increase and the packet loss ratio reduces to zero; if the (λ_p, λ_s) pair is outside the stable-throughput region of a scheme, the average cost goes to infinity, as it happens for OPP in scenario 2 and for all schemes in scenario 3 (shown in Fig. 3(c) and (d), respectively). As we see, the results in Fig. 3 verify Proposition 3, that is, the LP approximation introduced in Section IV-D gives good estimation of the OSD average cost for stable systems with infinite buffer capacity, though the required buffer size increases with the load. Fig. 3 also indicates that for OSD the analytic and simulation results are consistent. For the rest of the evaluation, unless specified, we use fixed buffer size $N_p = N_s = 30$, and use the analytic LP results to show the average cost under OSD.

C. On the increase of the cost for cooperation

The cost of cooperation depends on the preferences of the secondary system, and may change, for example increase as energy resources become scarce. Therefore, we evaluate, how

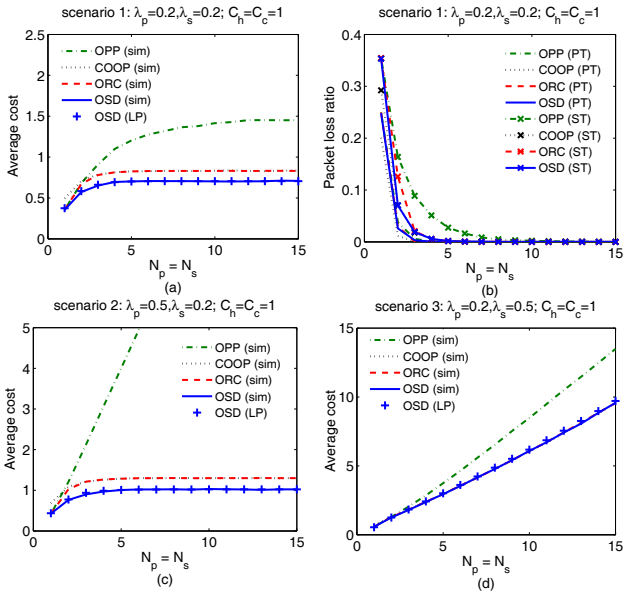

Fig. 3. (a) The average cost and (b) the packet loss ratio vs. the buffer sizes under scenario 1. The average cost vs. the buffer sizes under (c) scenario 2 and (d) scenario 3.

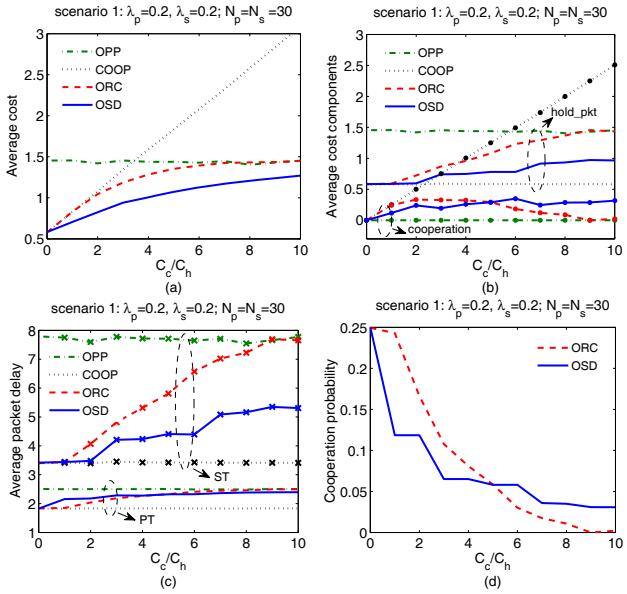

Fig. 4. (a) The average cost, (b) the components of the average cost, (c) the average PT and ST packet delay, and (d) the optimal cooperation probability vs. C_c/C_h under scenario 1.

the different schemes adapt to increasing cooperation cost C_c. We consider scenario 1, where (λ_p, λ_s) is located inside the stable-throughput regions of all the schemes. We keep the cost of holding a packet constant $C_h = 1$, and increase the cost of cooperative transmission $C_c = 0, 1, 2, \ldots 10$.

As it is shown in Fig. 4(a), the cost of OPP does not depend on C_c/C_h, while the cost of COOP increases linearly, as expected. OSD achieves the lowest average cost, which converges to the cost of COOP at low C_c, and to the cost of OPP at high C_c values, as OSD trades off the cost of cooperation and the cost of holding packets. ORC can as well trade off these costs, however with lower efficiency. For the considered parameters OSD can decrease the average cost with ca. 15% at high C_c values, compared to OPP, while it can achieve up to 40% gain for medium C_c values, where none of OPP or COOP is efficient.

Fig. 4(b) shows the components of the average cost, that is, the cost of cooperation and the cost of holding packets. The average cost shown in Fig. 4(a) is the sum of these two components. OPP does not allow cooperation so it has zero cost for cooperation, whereas under COOP the secondary user always cooperates when the primary queue is not empty, so the average cost for cooperation in this case increases linearly with C_c. Both of these schemes have constant average cost for holding packets since increasing C_c does not change their access strategies. For OSD and ORC the cost of cooperation is kept more or less constant, or is even decreased, while the cost of holding packets increases sublinearly with C_c. Note, that when OSD and ORC have the same cooperation cost (at ca. $C_c/C_h = 5$), OSD achieves significantly lower cost of holding packets, which shows the efficiency of the state dependent decision policy.

Fig. 4(c) shows the average packet delay at the PT and the ST. Compared to OPP, the delay experienced by PT is decreased or at least guaranteed, motivating that the primary system will allow cooperation. Comparing the secondary packet delay in Fig. 4(c) with the packet holding cost in Fig. 4(b), we can see that they are proportional, which indicates that the introduced state depended cost reflects the experienced delay, and delay sensitive secondary systems can tune C_h to achieve the preferred performance.

To see the reason of the delay increase, in Figure 4 (d) we compare the probability that cooperation is performed in an arbitrary time-slot for OSD and ORC. Both schemes reduce the cooperation probability when the unit cost for cooperation increases, though with different rate. As C_c increases, these schemes move from always cooperating when $Q_p \neq 0$ towards transmitting opportunistically. At $C_c/C_h = 9$, ORC starts to transmit purely opportunistically, since the random cooperation strategy is not efficient enough to achieve the delay gain needed to compensate for the increased cost of cooperation.

We can conclude that a dynamic secondary access scheme, OSD or ORC, is necessary to adapt to changing cooperation cost. The introduced state depend cost allows delay control in these systems. OSD can achieve significant (up to 40% in the considered case) cost decrease, while ORC has lower efficiency due to the sate independent control.

D. On the increase of the arrival rates

Let us now evaluate, how the proposed solutions adapt to changing load conditions, first considering an increasing load at ST. We keep the unit costs $C_h = 1$, $C_c = 2$ and $\lambda_p = 0.2$ constant, and increase λ_s according to scenario 4. Fig. 5(a) shows that OPP performs well at low, while COOP at high λ_s, and OSD and ORC balances well between the two deterministic solutions. Note, however, that for low and high loads ORC performs just as well as the more effective deterministic scheme. To better understand the gain of OSD,

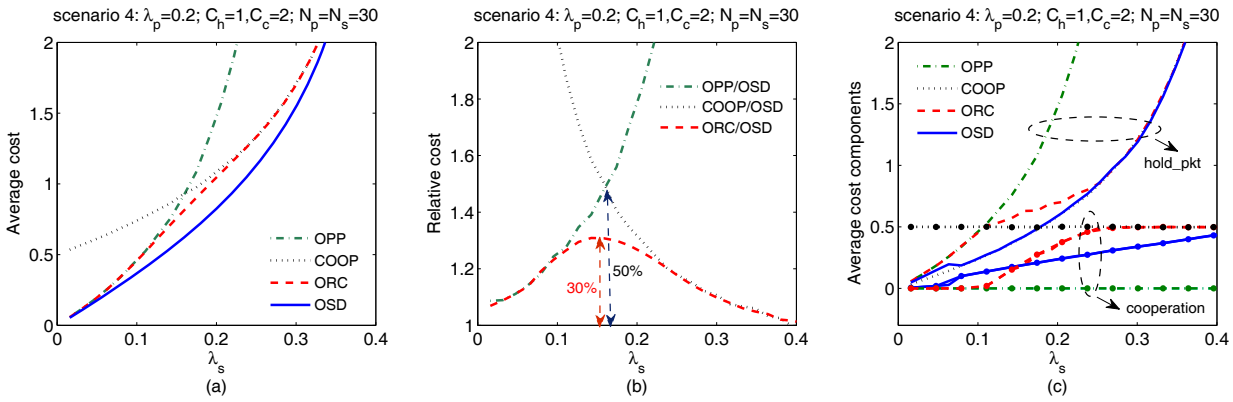

Fig. 5. (a) The average cost, (b) the relative cost, and (c) the components of the average cost vs. the arrival rate at the ST under scenario 4.

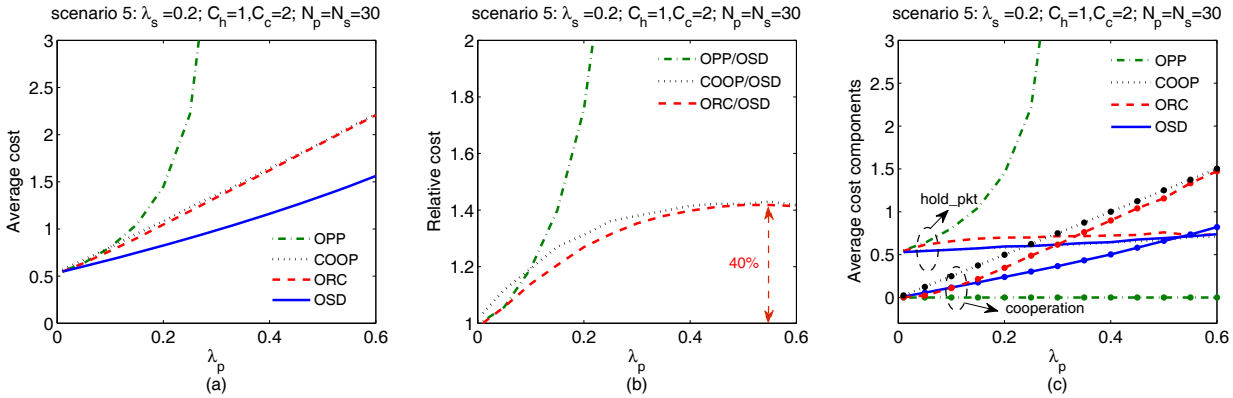

Fig. 6. (a) The average cost, (b) the relative cost, and (c) the components of the average cost vs. the arrival rate at the PT, for scenario 5.

in Fig. 5(b) we shows the costs relative to the OSD one. We see that the OSD gain is significant at moderate λ_s, where the cost of OPP and COOP is higher with 50% and the cost of ORC with 30%.

The components of the average cost are shown in Fig. 5(c), where the cooperation cost curves indicate that the two dynamic secondary access schemes (ORC and OSD) increase the cooperation probability when the load at the secondary user increases, and cooperation is necessary to keep the ST queuing delay low. Again we see that the state independent decision forces ORC to transmit opportunistically at low λ_s and to start to always cooperate rather early, while OSD increases the cooperation probability nearly linearly with λ_s. The packet holding cost increases exponentially for OPP and COOP, as expected. Interestingly, for this scenario OSD can manage nearly the same packet holding cost as COOP, but with lower cooperation cost since it can make state-dependent decision.

Let us now keep $\lambda_s = 0.2$ constant, and increase λ_p according to scenario 5. Fig. 6(a) and (b) show that OSD always achieves the lowest average cost, and its gain is significant when λ_p is not very low. At λ_p close to zero the four schemes achieve the same average cost since in all the schemes ST transmit directly to SR if the primary queue is empty. As it is visible in Fig. 6(b) the average cost of OPP is slightly lower than that of COOP when λ_p is low, but increases very fast and the system becomes unstable. ORC balances between OPP and COOP again, but it is not efficient, for the considered scenario its cost is up to 40% higher than that of OSD.

Fig. 6(c) shows the components of the average cost. As we can see, cooperation can decrease the cost of holding packets significantly, in the considered scenario this cost is nearly independent from λ_p under COOP, OSD and ORC. Of course the price is the increased cost of cooperation, which increases nearly linearly with λ_p not only for COOP, but also for OSD and ORC.

Our results considering the changing primary or secondary load show that although ORC can trade off the cost of cooperation and delay, its performance is not significantly better than the better one of the deterministic (OPP or COOP) schemes. The average cost can be significantly reduced with OSD. By utilizing cooperation when it effectively decreases the packet holding cost, OSD can achieve secondary queuing delays close to the one of COOP, however cooperating less frequently.

E. On the performance of R-learning

Finally we consider the case when the traffic and link statistics and possibly even the primary system state are not known for the SU, and the optimal sequential decision policy needs to be found by R-learning (RL), as defined in Section V. The parametrization of R-learning is non-trivial [27]. We consider fixed exploration probability $\alpha_t = 0.5$ for $t \in \mathbb{N}^+$ and

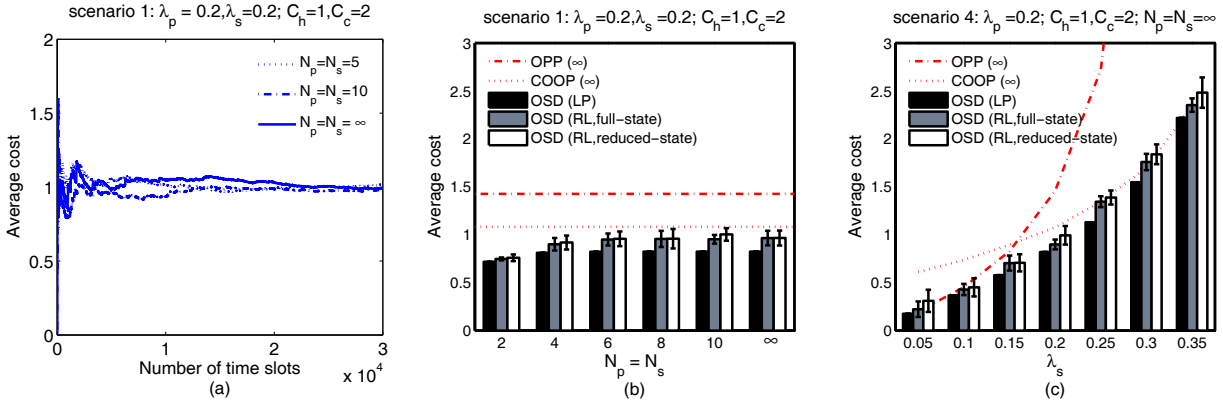

Fig. 7. (a) The average cost from a single RL simulation run under the full-state scenario. The average cost vs. (b) the buffer sizes under scenario 1, and (b) the secondary arrival rate under scenario 4.

update rates $\beta_r = 0.5$ and $\beta_\rho = 0.05$. The optimization of these parameters is beyond the focus of this paper.

We evaluate the convergence speed of RL in Fig. 7(a), showing the average cost from time zero as a function of the number of iteration steps. We consider scenario 1 with $C_h = 1$ and $C_c = 2$, and single RL simulation runs for the full-state scenario with buffer sizes $N_p = N_s = 5, 10, +\infty$, respectively. Similar convergence can be shown under the reduced-state scenario. The figure shows that the average cost converges fast and changes little after step $10,000$. The convergence speed is almost unaffected by the buffer size, which shows that RL can efficiently discover the action space in the states where the system resides with high probability.

Fig. 7(b) and (c) show the average cost with 95% confidence interval, based on 10 simulation runs. Each simulation run starts with a learning phase of $50,000$ iteration steps. Then the policy is fixed and the average cost is calculated for the next $50,000$ time slots.

In Fig. 7(b) we compare the average cost of OSD under scenario 1 with different buffer sizes, considering the LP solution with full knowledge on the traffic and link statistics, and full-, and reduced-state RL. We show the OPP and COOP performance for comparison. For all the three OSD solutions the cost first increases with the buffer size, as packet losses are avoided, and remains constant once the packet loss ratio is reduced almost to zero, showing that RL reaches a close to optimal solution even at infinite buffer size. RL achieves a little higher average cost than LP, however, still performs better than the two deterministic schemes, OPP and COOP. Comparing RL with full- and reduced-state in Fig. 7(b) we see that the secondary system can perform close to optimal dynamic access control even without knowing the exact state of the primary queue.

Fig. 7(c) shows the average cost under the changing secondary load of scenario 4, where $\lambda_p = 0.2$ kept constant and λ_s increased. Similarly to Fig. 5(a), Fig. 7(c) indicates that OPP and COOP performs well respectively at low and high λ_s, and OSD with LP based optimization can achieve a tradeoff between the two deterministic schemes. Most importantly we see, that RL can achieve a similar trade-off, with a very slight increase of the average cost. The efficiency of the RL algorithms shown in Fig. 7 motivates well its use to find dynamic secondary access control policies, when the secondary user does not know a priori the traffic and link statistics, and even the present queue size of the primary user.

VII. Conclusion

In this paper we considered cognitive spectrum sharing networks with non-backlogged traffic, where the secondary user has the possibility to trade off cooperation cost and channel access delay. We defined two dynamic secondary access schemes, sequential decision and random cooperation, and compared them to static solutions with pure opportunistic or pure cooperative access. We formulated the problem of optimal sequential decision as an MDP with a cost function that combines the cost of cooperation, e.g., increased secondary energy consumption, and the cost of queuing delay, and proved the existence of a stationary policy that is average cost optimal. We showed that dynamic secondary access with optimal sequential decision can achieve significant gain compared to the static schemes. It can adapt the transmission policy to changing cooperation cost or changing traffic intensities, and can achieve secondary delays as low as the one of pure cooperative transmission, with significantly lower cooperation cost. Optimal random cooperation can trade off cooperation cost and delay as well, though with lower efficiency. In general, our results show that dynamic secondary cooperation control is necessary, if the cost of cooperation is not negligible for the secondary system.

We addressed the case when the primary user does not share traffic and link statistics and even buffer state with the secondary one. In this case the secondary user can adapt online reinforcement learning to find optimal stationary access policy. Specifically, we considered R-learning, originally proposed to find average-cost optimal solution in MDPs with unknown transition probabilities. Our results demonstrate that reinforcement learning performs close to optimal, and therefore can provide a very powerful solution in spectrum sharing networks where traffic and link statistics are not known.

References

[1] Q. Zhao and B. Sadler, "A survey of dynamic spectrum access," *IEEE Signal Proc. Mag.*, vol. 24, no. 3, pp. 79–89, 2007.

[2] Y. Chen, Q. Zhao, and A. Swami, "Joint design and separation principle for opportunistic spectrum access in the presence of sensing errors," *IEEE Trans. Inf. Theory*, vol. 54, no. 5, pp. 2053–2071, 2008.

[3] A. T. Hoang, Y.-C. Liang, and Y. Zeng, "Adaptive joint scheduling of spectrum sensing and data transmission in cognitive radio networks," *IEEE Trans. Commun.*, vol. 58, no. 1, pp. 235–246, 2010.

[4] U. Berthold, F. Fu, M. V. der Schaar, and F. Jondral, "Detection of spectral resources in cognitive radios using reinforcement learning," in *Proc. of IEEE DySPAN*, 2008.

[5] X. Li, Q. Zhao, X. Guan, and L. Tong, "Optimal cognitive access of Markovian channels under tight collision constraints," *IEEE J. Sel. Areas Commun.*, vol. 29, no. 4, pp. 746–756, Apr. 2011.

[6] M. Levorato, S. Firouzabadi, and A. Goldsmith, "A learning framework for cognitive interference networks with partial and noisy observations," *IEEE Trans. Wireless Commun.*, vol. 11, no. 9, pp. 3101–3111, Sept. 2012.

[7] A. Galindo-Serrano and L. Giupponi, "Distributed Q-learning for aggregated interference control in cognitive radio networks," *IEEE Trans. Veh. Technol.*, vol. 59, no. 4, pp. 1823–1834, 2010.

[8] Y. Han, A. Pandharipande, and S. H. Ting, "Cooperative decode-and-forward relaying for secondary spectrum access," *IEEE Trans. Wireless Commun.*, vol. 8, no. 10, pp. 4945–4950, Oct. 2009.

[9] Y. Han, S. H. Tingand, and A. Pandharipande, "Cooperative spectrum sharing protocol with secondary user selection," *IEEE Trans. Wireless Commun.*, vol. 9, no. 9, pp. 2914–2923, Sept. 2010.

[10] I. Krikidis, "Multilevel modulation for cognitive multiaccess relay channel," *IEEE Trans. Veh. Technol.*, vol. 59, no. 6, pp. 3121–3125, Jul. 2010.

[11] B. Cao, L. X. Cai, and et. al., "Cooperative cognitive radio networking using quadrature signaling," in *Proc. of IEEE INFOCOM*, 2012.

[12] T. Elkourdi and O. Simeone, "Spectrum leasing via cooperation with multiple primary users," *IEEE Trans. Veh. Technol.*, vol. 61, no. 2, pp. 820–825, Feb. 2012.

[13] Y. Yi, J. Zhang, Q. Zhang, T. Jiang, and J. Zhang, "Spectrum leasing to multiple cooperating secondary cellular networks," in *Proc. of IEEE ICC*, 2011.

[14] L. Duan, L. Gao, and J. Huang, "Contract-based cooperative spectrum sharing," in *Proc. of IEEE DySPAN*, May 2011.

[15] W. Su, J. Matyjas, and S. Batalama, "Active cooperation between primary users and cognitive radio users in heterogeneous ad hoc networks," *IEEE Trans. Inf. Theory*, vol. 60, no. 4, pp. 1796–1805, Apr. 2012.

[16] W. Lu, Y. Gong, S. H. Ting, X. Wu, and N. Zhang, "Cooperative OFDM relaying for opportunistic spectrum sharing: protocol design and resource allocation," *IEEE Trans. Wireless Commun.*, vol. 11, no. 6, pp. 2126–2135, Jun. 2012.

[17] S. Hua, H. Liu, M. Wu, and S. Panwar, "Exploiting MIMO antennas in cooperative cognitive radio networks," in *Proc. of IEEE INFOCOM*, 2011.

[18] L. Gao, R. Zhang, C. Yin, and S. Cui, "Throughput and delay scaling in supportive two-tier networks," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 2, pp. 415–424, Feb. 2012.

[19] Y. Han, S. H. Ting, M. Motani, and A. Pandharipande, "On throughput and delay scaling with cooperative spectrum sharing," in *Proc. of IEEE ISIT*, Aug. 2011.

[20] L. Wang and V. Fodor, "On the gain of primary exclusion region and vertical cooperation in spectrum sharing wireless networks," *IEEE Trans. Veh. Technol.*, vol. 61, no. 8, pp. 3746–3758, Oct. 2012.

[21] O. Simeone, Y. Bar-Ness, and U. Spagnolini, "Stable throughput of cognitive radios with and without relaying capacity," *IEEE Trans. Wireless Commun.*, vol. 55, no. 12, pp. 2351–2360, Dec. 2007.

[22] I. Krikidis, J. N. Laneman, J. S. Thompson, and S. McLaughlin, "Protocol design and throughput analysis for multi-user cognitive cooperative systems," *IEEE Trans. Wireless Commun.*, vol. 8, no. 9, pp. 4740–4751, Sep. 2009.

[23] S. Kompella, G. Nguyen, J. Wieselthier, and A. Ephremides, "Stable throughput tradeoffs in cognitive shared channels with cooperative relaying," in *Proc. of IEEE INFOCOM*, 2011.

[24] B. Rong and A. Ephremides, "Cooperative access in wireless networks: stable throughput and delay," *IEEE Trans. Inf. Theory*, vol. 58, no. 9, pp. 5890–5907, sept. 2012.

[25] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. New York, NY: John. Wiley & Sons, Inc., 1994.

[26] D. P. Bersekas, *Dynamic programming and optimal control*, 2nd ed. Belmont, MA: Athena Scientific, 2001.

[27] S. Mahadevan, "Average reward reinforcement learning: foundations, algorithms, and empirical results," *Machine Learning*, vol. 22, no. 1, pp. 159–195, 1996.

[28] L. Georgiadis, M. J. Neely, and L. Tassiulas, "Interference in large wireless networks," *Found. Trends Netw.*, vol. 1, no. 1, 2006.

[29] M. J. Neely, "Dynamic power allocation and routing for satellite and wireless networks with time varying channels," Ph.D. dissertation, Massachusetts Institute of Technology, 2003.

[30] P. Coolen-Schrijner and E. A. V. Doorn, "On the convergence to stationarity of birth-death processes," *Journal of Applied Probability*, vol. 38, no. 3, pp. pp. 696–706, 2001.

[31] R. Rao and A. Ephremindes, "On the stability of interacting queues in a multiple-access systems," *IEEE Trans. Inf. Theory*, vol. 34, no. 5, pp. 918–930, Sep. 1988.

[32] L. Wang and V. Fodor, "Cooperate or not: the secondary user's dilemma in hierarchical spectrum sharing network," in *Proc. of IEEE ICC 2013*, to appear.

[33] L. Sennott, "Average cost optimal stationary policies in infinite state markov decision processes with unbounded cost," *Operation Research*, vol. 37, pp. 626–633, 1989.

[34] S. B. Thrun, "The role of exploration in learning control," in *Handbook of Intelligent Control: Neural, Fuzzy, and Adaptive Approaches*, 1992.