Model-Free Adaptive Rate Selection in Cognitive Radio Links*

Álvaro Gonzalo-Ayuso and Jesús Pérez

University of Cantabria,
Department of Communication Engineering,
Santander-39140, Spain
{alvaro,jperez}@gtas.dicom.unican.es

Abstract. In this work we address the rate adaptation problem of a cognitive radio (CR) link in time-variant fading channels. Every time the primary users (PU) liberate the channel, the secondary user (SU) selects a transmission rate (from a finite number of available rates) and begins the transmission of fixed sized packets until a licensed user reclaims the channel back. After each transmission episode the number of successfully transmitted packets is used by the SU to update its optimal rate selection ahead of the next episode. The problem is formulated as an n-armed bandit problem and it is solved by means of a Monte Carlo control algorithm.

Keywords: Cognitive radio (CR), rate control, n-armed bandit problem, reinforcement learning (RL).

1 Introduction

In this work we focus on opportunistic spectrum access (OSA) in hierarchical cognitive radio (CR) networks where the secondary users (SU's) only use the licensed spectrum when primary users (PU) are not transmitting (in the following we use "primary user" or PU to refer to the aggregate of primary users). Every time the PU liberates the channel, the SU begins transmitting until, without prior notice, the PU reclaims the channel again at any given time.

We consider noncooperative spectrum sharing where each SU makes its own decision on the spectrum access strategy, based on its error free channel sensing and the number of data packets successfully transmitted over time. In this work we focus on a single SU link and we do not take into consideration the competition between the different SU's. Nonetheless, the proposed scheme would still work in an scenario with multiple SU's competing to access the channel.

In this work we assume that the SU's support automatic repeat request (ARQ) protocol, so when a frame is decoded with error, its data is retransmitted in a further frame. Rate adaptation of SU links in CR has been widely addressed in the

* This work was supported by the Spanish Government, Ministerio de Ciencia e Innovación (MICINN), under projects COSIMA (TEC2010-19545-C04-03) and COMONSENS (CSD2008-00010, CONSOLIDER-INGENIO 2010).

D. Pesch et al. (Eds.): MONAMI 2013, LNICST 125, pp. 188–201, 2013.

technical literature, [1], [2], [3]. However, none of the above works consider frames retransmission. In [4] frames retransmission was taken into account, but assuming a time independent channel occupancy model. In [5] the authors present a similar problem, however they do not consider time variables scenarios or time dependent occupancy models. In [6] and [7] we considered frames retransmission and we made use of the acknowledgments (ACK's) for rate control, however, in these cases we assumed perfect knowledge of the channel fading and occupancy statistical models. To the best of our knowledge, optimal rate adaptation while considering retransmissions of failed frames, time-dependent channel occupancy and fading models have not been addressed so far in the context of OSA.

In this work we aim to go one step further, we introduce a rate adaptation algorithm in which we do not require any additional information about the channel state and occupancy. The ACK's sent back by the SU receiver are the only information exchanged with the SU transmitter. We propose an energy efficient scheme with a reduce computational cost and hardware complexity with relaxed requirements in terms of delay and transmission rate. This is the main novelty of this work.

We formulate the adaptive rate selection as an n-armed bandit problem [8] where the actions are the different available transmission rates, and the rewards are based on the number of successfully transmitted packets over time and its duration. We propose a Monte-Carlo based algorithm [8] capable of tracking changes in the received signal to noise ratio (SNR), the channel occupancy process and others variables over time to select the optimal rate.

The remaining of this paper is organized as follows; system model is presented in section 2. In section 3 we formulate the problem and we introduce the solving algorithm. In section 4 we present numerical results to evaluate the tracking capacity, the robustness, the speed of convergence and the performance of the algorithm under different scenarios. Finally, section 5 presents the conclusions of this work.

2 System Model

We consider an SU that periodically senses the channel ideally (with zero probability of miss detection and false alarm). Once it detects that the channel is idle, it begins the transmission of a sequence of fixed size data packets until the PU reclaims the channel. Each one of these packets is encoded into a single frame. The SU has the capability of adapting its transmission rate, i.e. the duration of each frame.

The aim of the SU is to maximize its own throughput during the sojourn time of the PU. To achieve this goal, the SU selects a transmission rate from a set $\mathcal{A}$ of $K = |\mathcal{A}|$ different types of available frames, each one with duration T_a and frame error rate (FER) denoted by $p(a, SNR)$, where $a \in \mathcal{A}$ and SNR is the signal to noise ratio at the receiver during the frame transmission. We assume a block fading channel model, namely, the SNR does not change during the transmission of a frame, but it can change from frame to frame.

We consider a conventional and ideal ARQ mechanism to detect frame transmission errors. When the receiver receives a frame, it sends back an ACK packet to the transmitter through an instantaneous error-free feedback channel to inform whether the frame has been correctly decoded or not. Whenever a frame is decoded with error, the corresponding packet must be retransmitted in a further frame.

The SU transmitter does not have access to any information regarding the channel state, the receiver SNR or the PU channel occupancy patterns. The information related to the channel state available to the SU transmitter is:

1. The ACK's sent back by the receiver.
2. The availability of the channel by means of perfect sensing.

We assume that the transmit power constrains on the SU avoid significant interference on the normal operation of the PU. Consistently, whenever the PU occupies the channel, the SU frames are lost.

Primary User Channel Access Model

Figure 1 depicts the channel occupation process by the PU. The channel state changes alternatively between idle and busy periods over time. The duration of the idle/busy periods is given by two random variables denoted by d_i and d_b respectively. Let $F_i(d_i)$ and $F_b(d_b)$ be the corresponding cumulative distribution functions (CDF) that model the occupancy process.

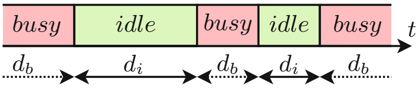

Fig. 1. Channel occupancy process

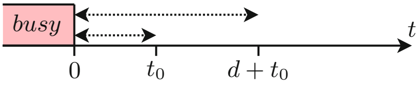

Fig. 2. Idle period

Regarding figure 2, let $\beta(d|t_0)$ denote the conditional probability that the channel remains idle at time $t_0 + d$ given that it was idle at time t_0. Using Bayes' theorem:

$$\beta(d|t_0) = \begin{cases} \dfrac{1 - F_i\left(t_0 + d\right)}{1 - F_i\left(t_0\right)}, & t_0 > 0 \\ 1 - F_i\left(d\right), & t_0 = 0 \end{cases} \tag{1}$$

Generally, $\beta(d|t_0)$ depends on t_0.

Sensing and Transmission Strategy

Figure 3 illustrates the adopted sensing strategy. The SU periodically senses the channel every T_i seconds and every sensing instance takes T_s seconds. Once it detects that the channel is idle, it begins its own transmission. After transmitting each frame, the SU senses the channel again. As long as the presence of the PU is not detected the SU continues transmitting. Whenever the channel is sensed as busy the SU stops transmitting.

Notice that, when the channel is idle, the time interval between two consecutive sensings depends on the frame duration and it may differ from the interval when the channel is busy. If the PU reclaims the channel during the transmission of a frame, the SU transmitter has no way to detect the collision during the transmission of the current frame. This last frame is usually lost. Figure 3 illustrate this situation, transmission of frame 3 is completed even thought the channel is busy during the last half of the transmission and the frame is therefore lost.

In the following, we will refer to time windows in which the SU transmits, as episodes and we will use T_{eps} to denote their duration. Notice that the actual time interval in which the channel is idle, will in general differ from the observed episode.

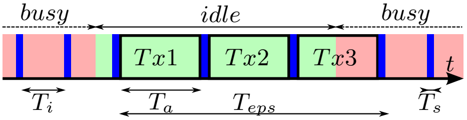

Fig. 3. Sensing and transmission strategy

3 Problem Formulation

We formulate this rate adaptation problem as a n-armed bandit problem [8] with ε-greedy policy, where the set of actions, $\mathcal{A}$, is formed by the set of available rates.

The SU maintains and updates an action-value function, $Q(a)$, which can be understood as an estimated measure of performance for each action (rate). In particular the entries of $Q(a)$ are an estimation of the expected throughput when transmitting each type frame during an episode. After a rate is selected at the beginning of the episode, it will be maintained throughout the entire episode. This approach guarantees that the transmission rate only needs to be changed once per episode (at most). In [6] we showed that for non fading channels, this is optimal or close to optimal. However, it is easy to see that under fading channels and, unless the coherence time of the channel is higher than the average episode duration, this scheme is not optimal, or close to optimal, anymore [7].

Two possibilities arise, exploration and exploitation. Usually, we are interested in exploitation, selecting the action that we expect to yield the best performance, namely the one with the highest Q. However, it is also important to try, or explore, the other actions occasionally in order to keep their estimated values updated. This is of critical importance in a dynamic environment since the expected performance of each action will vary over time. To handle the trade-off between exploration and exploitation we propose to use an ε-greedy policy [8]. With probability $1-\varepsilon$ the SU selects exploitation, the action with highest current value, $a^* = \arg\max_a Q(a)$, is selected. With probability ε the SU explores and the action is randomly picked from $\mathcal{A}$. An ε-greedy policy is usually expressed as

$$\pi(a) = \begin{cases} 1-\varepsilon+\varepsilon/K, & a = a^* \\ \varepsilon/K, & a \neq a^*, \end{cases} \tag{2}$$

for $a \in \mathcal{A}$ and where $\pi(a)$ is the probability of selecting action a. The greedier a policy is, the higher the probability of choosing the optimal action.

After each episode concludes, a reward is granted to the decision maker in order to update its value-function. These rewards are a measure of *how good* performed the selected action (rate). In particular, they are an estimation of the throughput during the episode

$$r = \frac{P_{Tx}}{T_{eps}}, \tag{3}$$

where P_{Tx} is the number of data packets that have been successfully transmitted during the episode (number of received ACK's).

Given the reward, the value of the selected action, a, is updated with the following Monte-Carlo [8] rule:

$$Q(a) = Q(a) + \mu(a)\left[r - Q(a)\right], \tag{4}$$

where $0 < \mu(a) \leq 1$ is the so called adaptation step. Notice that when $\mu(a) = 1$, the updated value of $Q(a)$ is simply the reward r, the algorithm has no memory of the past episodes.

Large adaptation steps provide a faster convergence in the action-value estimations at the price of smaller precision. On the other hand, smaller values provide slower convergence but can achieve higher precision. Noisy or inaccurate action-value estimations can lead to a noisy ε-greedy policy. If the noise level is high, it can affect the optimal rate selection causing the ε-greedy policy to be unstable continually changing its choice of a^* even in stationary environments. On the other hand, actions that are not currently optimal, are selected less frequently, meaning that their values are most likely outdated. For these actions it makes sense to use a large adaptation step to be sure that a few, or even a single adaptation step, is enough to get a reasonable update of the value.

We propose to use two different step sizes. Optimal action-value estimation is refined with a smaller adaptation step, μ_{opt}, while the other values are kept updated only with less precise estimations by using a larger step size μ_{exp}. Therefore we define $\mu(a)$as

$$\mu(a) = \begin{cases} \mu_{opt}, & a = a^* \\ \mu_{exp}, & a \neq a^*, \end{cases} \tag{5}$$

with $\mu_{opt} \leq \mu_{exp}$. Table 1 illustrates the complete adaptive rate selection algorithm.

4 Numerical Results

In this section we first present the general simulation framework and then, the specific simulations to evaluate the performance of the rate selection algorithm under different realistic scenarios.

Table 1. Adaptive Rate Selection Algorithm

```
$Q \leftarrow$ Initialize arbitrarily
$\pi \leftarrow$ Initialize $\varepsilon$-greedy policy from $Q$
Repeat forever
    $ChannelState \leftarrow$ Channel Sensing Output
    If $ChannelState$ = $busy$
        wait $T_i$
    Else
        Select rate $a$ using policy $\pi$
        $P_{Tx} = 0$, $T_{eps} = 0$
        Repeat until $ChannelState$ = $busy$
            Transmit a type $a$ frame
            If an $ACK$ is received $\rightarrow P_{Tx} = P_{Tx} + 1$
            $T_{eps} = T_{eps} + T_a$
            $ChannelState \leftarrow$ Channel Sensing Output
            $T_{eps} = T_{eps} + T_s$
        End
        $r = P_{Tx}/T_{eps}$
        $Q(a) = Q(a) + [r - Q(a)] \cdot \begin{cases} \mu_{opt}, & a = a^* \\ \mu_{exp}, & a \neq a^* \end{cases}$
        $\pi \leftarrow \varepsilon$-greedy policy from $Q$
    End
End
```

4.1 Simulation Framework

Frame Types. Throughout section 4 we assume that there are four different rates available to the SU transmitter, all carrying a payload of 1024 bits. Table 2 shows the frame duration, T_a, for each rate. We also associate each rate with one of the FER curves shown in figure 4. The FER curves of many practical system can be approximated by the exponential function, for this reason we choose to work with the following generic exponential expression to resemble realistic systems

$$FER(a) = A \cdot e^{-B \cdot [SNR + C(a)]},$$

where $A = 10^3$ and $B = 0.6$ are constants, SNR is given in dBs and $C(a)$ represents a SNR shift in dBs and it is also given in table 2 for each rate. All frames encode a single packet of 1024 bits.

In the following we will use *operational SNR range* to refer to the range of SNR values in which a particular rate is optimal. For example, operational SNR range of rate 2 is approximately between 7 and 12 dBs.

Sensing Strategy. We assume that the sensing period $T_i = 0.1$ ms, and negligible sensing time, $T_s = 0$ ms.

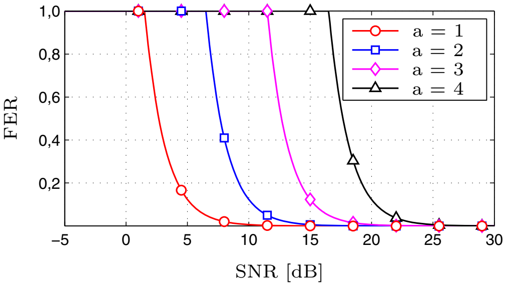

Fig. 4. Frame Error Rate for the four types of frames available

Table 2. Parameters for the four types of frames

rate	T_a [ms]	$C(a)$ [dBs]
1	1.6	10
2	0.8	5
3	0.4	0
4	0.2	-5

PU Channel Access Model. Without loss of generality, and only for simulation purposes, in the following we model d_i and d_b as generalized exponential (GE) random variables with CDF

$$GE_x(x) = \left[1 - e^{-\lambda(x-\mu)}\right]^{\alpha}, \qquad x \geq \mu \tag{6}$$

where $x \geq \mu > 0$, $\lambda > 0$ and $\alpha > 0$. Notice that for $\alpha = 1$ and $\mu = 0$ the GE distribution becomes the exponential distribution with parameter λ

$$\beta(d|t_0) = \frac{1 - \left(1 - e^{-\lambda(t_0+d)}\right)}{1 - \left(1 - e^{-\lambda(t_0)}\right)} = e^{-\lambda d}, \tag{7}$$

which does not depend on t_0. This memoryless exponential model has been extensively used in the literature, however, in many practical cases it is not a realistic one [9].

Specifically, throughout the simulations section we make use of the four different GE distributions described in table 3.

We propose three different distributions, GE_{i1}, GE_{i2} and GE_{i3}, for d_i. In these cases, given $\mu = 0$, d_i can take any value greater than zero. Parameter λ is chosen so that the three distributions share the same mean value. Since $\alpha = 1$,

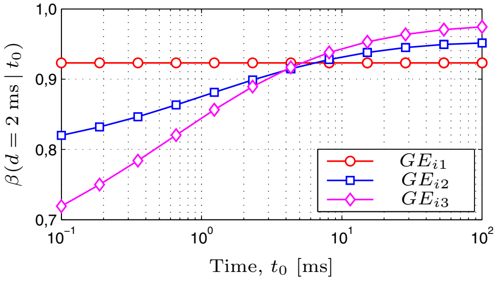

Fig. 5. $\beta\,(d = 1.6\text{ ms}|t_0)$ for the GE_i distributions described in table 3

Table 3. Parameters of the selected CDFs, μ and the mean value in ms

	λ	μ	α	Mean
GE_b	200	2	1	7
GE_{i1}	50	0	1	20
GE_{i2}	30.69	0	0.5	20
GE_{i3}	14.41	0	0.2	20

both GE_{i1} and GE_b are memoryless distributions. Figure 5 illustrates $\beta(d|t_0)$ for $d = 1.6$ ms (the duration of the longest frame), as a function of t_0, for the three GE_i distributions.

The distribution of d_b has an important influence on the performance of the adaptation scheme. In time varying environments, the probability that the learned value-function resembles the real channel state decreases with the duration of the busy intervals.

Channel Fading. We consider both, simple additive white Gaussian noise channels (AWGN) and fading channels. For the fading channels we use a Rayleigh block fading model so the channel gain remains constant at least during the transmission of the longest frame. For our experiments we consider three values for the Doppler frequency for these fading channels: $f_D = \{0.1, 0.5, 1\}$ Hz.

The AWGN channel has a constant unitary channel gain (0 dB), in turn we assume an average channel gain of 0 dB for the fading channels. In all cases the average received SNR is assumed to be 15 dB.

Adaptive Algorithm. Unless otherwise indicated, we also assume that:

- The exploration adaptation step is always $\mu_{exp} = 1$.
- The exploration parameter is fixed as $\varepsilon = 0.1$.
- The action-value function, Q, is initialized to zero.

Performance Metrics and Upper Bonds. As a measure of performance, we use the averaged throughput per episode,

$$Th_{eps} = \mathrm{E}\left[r\right] \cdot payload,$$

assuming that the four types of frames carry the same payload of 1024 bits. Unless it is otherwise indicated, all the numerical results have been obtained averaging ten thousand independent simulations.

Upper bounds are computed as the throughput achieved by selecting the optimal rate on each episode but taking the exploration parameter into account. The optimal actions are selected *a priori* based on the channel occupancy statistical models, the variable FER of each type of frame and its duration. The upper bound on AWGN channels and fading channels might not be the same, even if the average SNR is the same, simply because the channel gain is not a random variable but a constant in the AWGN channels. In fading channels the upper bound is expected to be above the throughput yield by any of the stationary policies (selecting the same type of frame on every episode). In turn, for the AWGN channel the upper bound will match the throughput achieved by one of the stationary policies.

4.2 Adaptation Step and Tracking Capability

In this section we aim to show how the choice of the adaptation step can affect the tracking capability of the algorithm and hence its performance. We consider a fading channel, with $f_D = 0.1$ Hz, a channel occupancy model described by GE_{i2} and GE_b, three adaptation steps, $\mu_{opt} = \{0.1, 0.5, 1\}$, and we average ten thousand independent runs of the algorithm. Figure 6 show the throughput and the corresponding averaged policy. We can see how there is little difference between $\mu_{opt} = 0.1$ and 0.5, however, for $\mu_{opt} = 1$ the throughput sinks. In the latter case the algorithm relays solely on the last reward obtain with each rate, therefore, and on the face of incomplete information, the algorithm tends to select the safest rate, the one which can work with a lower SNR. Our experiments show than in general a fading channel will lead to a less greedy policy simply because different rates become optimal at different times. The fact that the adaptation step depends on the selected rate is what gives the algorithm its robustness against the selection of μ_{opt}, in general the performance of the algorithm is not strongly dependent on μ_{opt} as long as we choose a small value.

A key feature of our algorithm is its tracking capability at a reduced computational cost and complexity. To illustrate how the algorithm is capable of tracking the changes in the SNR, we generate an independent sampled sequence of the channel fading process with $f_D = 0.5$ Hz, then we run ten thousand independent simulations over the same fading sequence and, considering a channel occupancy process with the memoryless distributions GE_{i1} and GE_b.

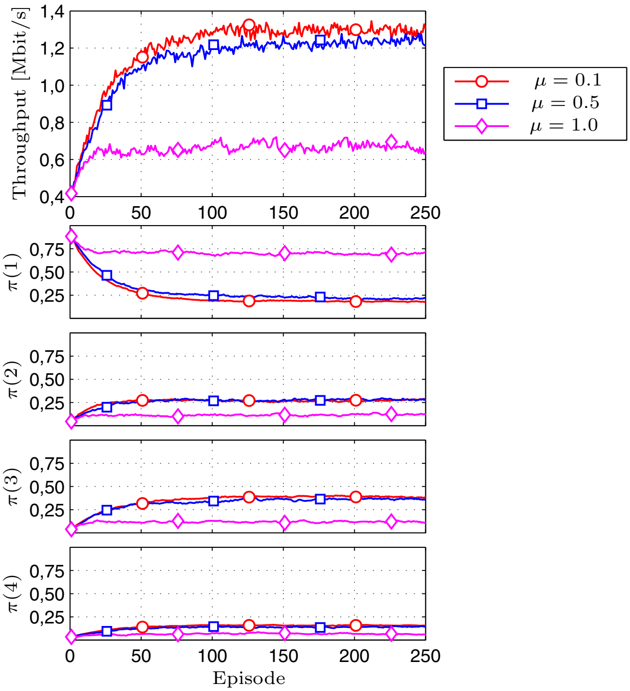

Fig. 6. Throughput per episode (top) and its corresponding policy (bottom) for several values of μ. Fading channel with $f_D = 0.1$ Hz and occupancy model given by GE_{i2} and GE_b.

Figure 7 illustrates the SNR evolution, the achieved throughput and the corresponding policy and action value function. We can see how the throughput varies overtime following the changes in the SNR. By looking at the policy figure we can see how rates 2 and 3 share most of the probability throughout the whole simulation, this is because the SNR remains approximately within the operational SNR ranges of rates 2 and 3. A closer look to the action value figure reveals that rate 3 is optimal most of the time, however when the SNR falls below approximately 10 dB rate 2 becomes optimal. On the other hand rates 1 and 4 are deactivated most of the time because the SNR is above and below their operational SNR ranges respectively. We can see how only rate 4 gains some value and probability of selection precisely when the SNR is maximum.

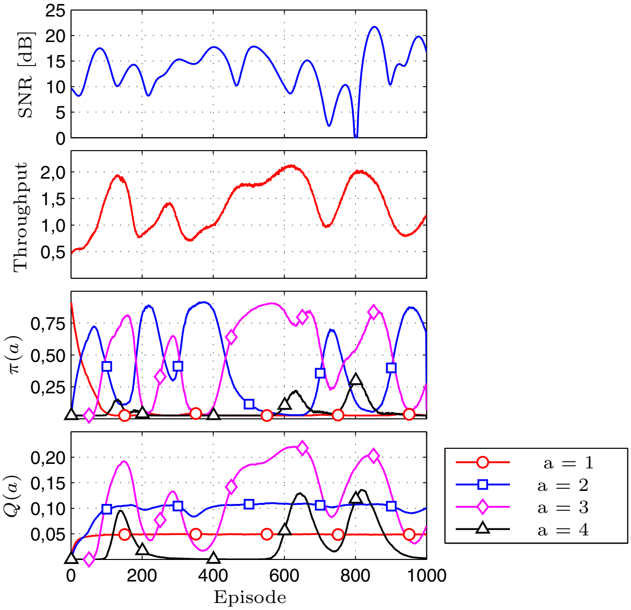

Fig. 7. From top to bottom, Signal to noise ratio evolution, Throughput per episode in Mbit/s, policy and Q value

4.3 Performance over Fading Channels

In this subsection we measure the performance degradation due to the limited tracking capability and we show that the algorithm is still capable of exploiting the transmission opportunities. To do so we consider distributions GE_{i1} and GE_b for the channel occupancy model and we run the simulations for the AWGN channel and for three different values of Doppler frequency, $f_D = \{0.1, 0.5, 1\}$ Hz.

Figure 8 depicts the throughput evolution along with the corresponding achievable rates and the averaged policy. As we increase f_D, the SNR variation within an episode increases reducing the probability that the selected rate remains optimal. Even more important is the fact that, as we increase f_D the correlation between two consecutive episodes decreases, for this reason it is harder for the algorithm to forecast which rate is optimal for the next episode. The gap between the achievable throughput when transmitting over the AWGN channel and when transmitting over a fading channel is due to the random nature of the channel gain in the latter case.

As for the averaged policy evolution we notice that for the AWGN we obtain a greedy policy after convergence. For the other three cases, the policy is roughly the same and it is less greedy because the probability is spread more evenly among the rates. In this particular case $\pi(2)$ and $\pi(3)$ stand out meaning that the SNR oscillates within the SNR operational range of rates 2 and 3 most of the time.

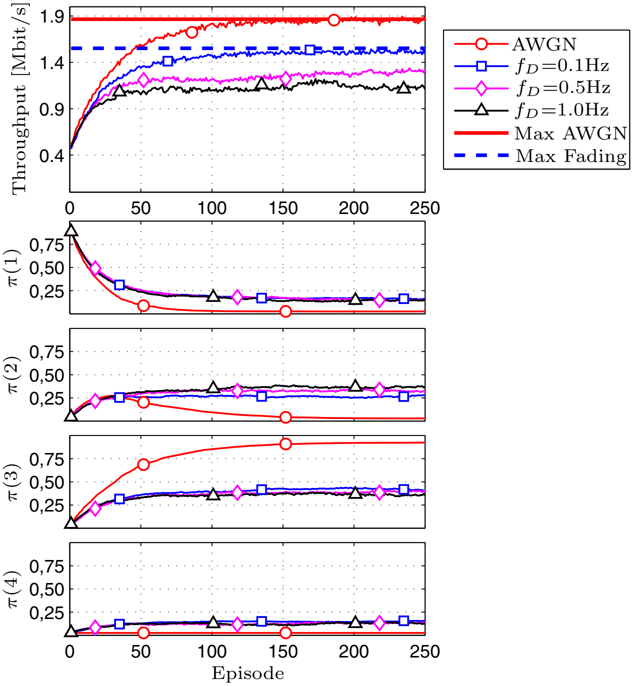

Fig. 8. Throughput per episode and achievable rate (top) and the corresponding policy (bottom) for several values of f_D

4.4 Channel Occupancy Model

Next, we explore the effect of different degrees of memory on the channel occupancy model. To do so, we repeat the same simulation but this time only for the AWGN channel and for the three distributions GE_{i1}, GE_{i2} and GE_{i3}. Figure 9 depicts the throughput evolution along with the corresponding achievable rates and the averaged policy. In this case the achievable rate is strongly influenced by the occupancy model, however, the algorithm is still capable of converging towards the upper bounds.

As the memory effect is more noticeable, episodes much shorter and much longer than the average are more likely to occur. Hence, suboptimal rates might gain value and be selected more often eventually leading to a less greedy policy and a degraded performance. We can see how for the distribution with the largest memory (GE_{i3}), the gap between the achieved rate and the upper bond is larger than for the other distributions. Correspondingly, by looking at the policy evolution we can also see how $\pi(3)$ decreases and $\pi(2)$ increases resulting in a less greedy policy. Notice that this effect is different from the one observed for the fading channels, there is no possibility of tracking here, only the fact that as the episode duration is more variable, some rates may randomly gain value introducing noise in the $Q(a)$ estimation and therefore interfering with the rate selection.

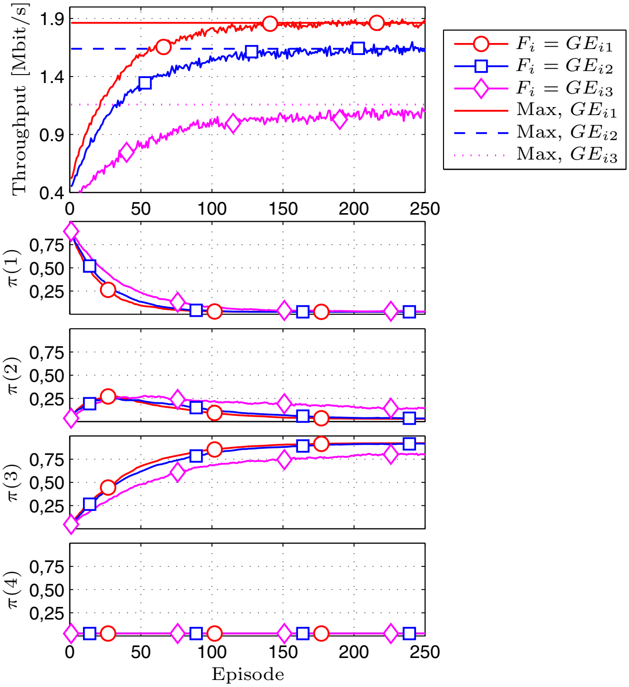

Fig. 9. Throughput per episode and achievable rates (top) and the corresponding policy (bottom) for the three distributions in table 3

4.5 Exploration Parameter, ε

Theory suggests that as we increase the value of ε we explore more and therefore we exploit less, in other words, it takes less time to learn the ε-policy but it learns a less greedy policy which might lead to a lower achievable throughput. This is exactly what we found when we run simulations on static environments, with no fading channels or other variables varying over time. However, when the environment varies over time, for example due to a fading channel, a larger ε does not necessarily lead to a lower throughput. Experiments reveled that in this cases it is critical to explore sufficiently often in order to maintain an estimation of Q that resembles the true value of each rate over time. On the other hand, exploring too often means that we spend less time exploiting what we learned. The optimal ε depends on how fast the environment changes, slow changes require less exploration and fast changes require higher values of ε.

5 Conclusions

In this work we have presented a novel adaptive rate selection scheme for cognitive radio links. We introduce a model free adaptive rate selection with a reduce computational cost and hardware complexity; these are the main novelties of this work. We consider a SU that opportunistically accesses the channel with the goal of transmitting an infinite number of data packets. Every time the SU begins transmitting it has to select the transmission rate following an ε-greedy policy. After the PU reclaims the channel the decision maker in the SU receives a reward and updates its value function. These updates are done with different adaptation steps depending on the nature of the episode, exploration or exploitation. This trick allows the algorithm to maintain an updated and yet accurate estimation of the action-value function improving the tracking capability.

Experiments illustrate the tracking capabilities of the algorithm under variable SNR channels. We also study the performance and converging properties for different degrees of fading and different channel occupancy models.

References

1. Chai, C.C.: On power and rate adaptation for cognitive radios in an interference channel. In: Proc. 71th IEEE Vehicular Technology Conference, PIMRC, Taipei, Taiwan, vol. 2, pp. 1–5 (May 2010)
2. Gao, L., Cui, S.: Power and rate control for delay-constrained cognitive radios via dynamic programming. IEEE Transactions on Vehicular Technology 58, 4819–4827 (2009)
3. Wang, H.H.J., Zhu, J., Li, S.: Optimal policy of cross-layer design for channel access and transmission rate adaptation in cognitive radio networks. EURASIP Journal on Advances in Signal Processing, vol. 2010 (2010)
4. Pérez, J., Khodaian, M.: Optimal rate and delay performance in non-cooperative opportunistic spectrum access. In: 9th International Symposium on Wireless Communication Systems (ISWCS 2012), Paris, France (August 2012)
5. Jouini, W., Ernst, D., Moy, C., Palicot, J.: Upper confidence bound based decision making strategies and dynamic spectrum access. In: 2010 IEEE International Conference on Communications (ICC), pp. 1–5 (2010)
6. Gonzalo-Ayuso, A., Pérez, J.: Dynamic rate adaptation in cognitive radio considering time-dependent channel access models. In: 8th International Conference on Cognitive Radio Oriented Wireless Networks, CROWNCOM (2013)

7. Gonzalo-Ayuso, Á., Pérez, J.: Rate adaptation in cognitive radio links with time-varying channels. In: 21st European Signal Processing Conference 2013, EUSIPCO 2013 (2013)
8. Sutton, R.S., Barto, A.G.: Reinforcement Learning: An Introduction. M.I.T. Press (1998)
9. López-Benítez, M.: Spectrum usage models for the analysis, design and simulation of cognitive radio networks. Ph.D. dissertation, Universitat Politècnica de Catalunya (UPC), Barcelona (May 2011)