Distributed Q-Learning Based Dynamic Spectrum Management in Cognitive Cellular Systems: Choosing the Right Learning Rate

Nils Morozs, Tim Clarke, David Grace and Qiyang Zhao
Department of Electronics, University of York
Heslington, York YO10 5DD, United Kingdom
E-mail: {nm553, tim.clarke, david.grace, qiyang.zhao}@york.ac.uk

Abstract—**This paper presents the concept of the Win-or-Learn-Fast (WoLF) variable learning rate for distributed Q-learning based dynamic spectrum management algorithms. It demonstrates the importance of choosing the learning rate correctly by simulating a large scale stadium temporary event network. The results show that using the WoLF variable learning rate provides a significant improvement in quality of service, in terms of the probabilities of file blocking and interruption, over typical values of fixed learning rates. The results have also demonstrated that it is possible to provide a better and more robust quality of service using distributed Q-learning with a WoLF variable learning rate, than a spectrum sensing based opportunistic spectrum access scheme, but with no spectrum sensing involved.**

Keywords—Self-Organisation, Distributed Q-learning, Dynamic Spectrum Management

I. Introduction

One of the fundamental tasks of a cellular system is spectrum management, concerned with dividing the available spectrum into a set of resource blocks and assigning them to voice calls and data transmissions in a way which would provide a good quality of service (QoS) to the users. Flexible dynamic spectrum management (DSM) techniques play a key role in utilising the given spectrum efficiently. This gave rise to the novel wireless communication systems such as cognitive radio networks [1] and cognitive cellular systems [2]. Such networks employ intelligent opportunistic spectrum access techniques instead of the inefficient static spectrum allocation methods used in most current cellular systems.

An emerging state-of-the-art technique for intelligent DSM is reinforcement learning (RL), which is a machine learning technique aimed at building up solutions to decision problems only through trial-and-error [3]. It has been successfully applied to a range of DSM problems and scenarios, such as cognitive radio networks [4], LTE pico-cells [5], femto-cell networks [6] and multi-hop backhaul networks [7]. The most widely used RL algorithm in both artificial intelligence and wireless communications domains is Q-learning. Therefore, most of the literature on RL based DSM focuses on Q-learning and its variations. Furthermore, this paper is concerned with *distributed* Q-learning based DSM, where no information exchange is assumed among the individually learning base stations. The distributed Q-learning approach has advantages over centralised methods in that no communication overhead is required to achieve the learning objective, and the network operation does not rely on a single computing unit. It also allows for easier insertion and removal of base stations from the network, if necessary. For example, such distributed opportunistic types of protocols are well suited to temporary event networks and disaster relief scenarios, where rapidly deployable network architectures with unplanned or variable topologies may be required to supplement any existing wireless infrastructure [8].

The learning rate is a crucial parameter in Q-learning algorithms that can significantly influence the dynamics of the learning process. So far, there is nothing in the DSM literature on the best selection of learning rate values. One of the rare examples where the value of the learning rate is at least specified is [6], where the authors have arbitrarily chosen a value of 0.5, which is simply in the middle of its allowed range of $[0, 1]$. In [4] the authors have swept all possible values of the fixed learning rate to compare different exploration strategies, but do not comment on difference in performance due to the difference in learning rate values. The majority of other examples in DSM literature do not even specify the learning rate they have chosen, making it impossible to replicate their results.

The purpose of this paper is to present the concept of the Win-or-Learn-Fast (WoLF) variable learning rate [9] from the artificial intelligence literature, and show how it can be applied in the DSM context and what performance improvements can be achieved using it, in terms of the QoS provided by the network. The simulation results shown in this paper also aim to demonstrate the importance of choosing the right learning rate, and serve as a reference guide for researchers who design Q-learning based DSM algorithms. Although the WoLF principle has been mentioned in our previous work [10][11], we have only tried using one pair of values there. This paper provides a full two-dimensional sweep across all feasible learning rate values and discovers those learning rates that outperform the ones we or anyone else has used before in the wireless communications domain.

The rest of the paper is organised as follows: in Section II a cognitive cellular system model designed for stadium temporary events is introduced. In Section III the distributed Q-learning algorithm used for DSM is described. Section IV presents the concept of the Win-or-Learn-Fast variable learning rate, and discusses the simulation results obtained by using it together with the distributed Q-learning algorithm. Finally, the conclusions are given in Section V.

II. Stadium Temporary Event Network

A. Network Model

The cognitive cellular system investigated in this paper is modelled for a stadium event scenario, where a temporary network architecture is installed in a large stadium to provide an increase in mobile data capacity to the users attending the event. The network architecture is depicted in Figure 1, where the users are located in a circular spectator area 53.7 - 113.7 m from the centre of the stadium. The spectator area is covered by 78 eNodeBs (eNBs) arranged in three rings at 1 m height, e.g. with antennas attached to the backs of the seats or to the railings between the different row levels. Seat width is assumed to be 0.5 m, and the space between rows - 1.5 m, which yields the total capacity of 43,103 seats.

The other assumptions used in the network model are listed below:

- The 18 MHz transmission bandwidth of an available 20 MHz LTE channel is split into 25 subchannels (each subchannel consists of 4 physical resource blocks, i.e. the bandwidth of each subchannel is 4×180 kHz) [12]. Only downlink communications are considered.
- The 2.6 GHz frequency band is used by all user equipment (UE).
- The UE receiver noise floor is -94 dBm, obtained by assuming 290 K temperature, 20 MHz total bandwidth and a 7 dB noise figure.

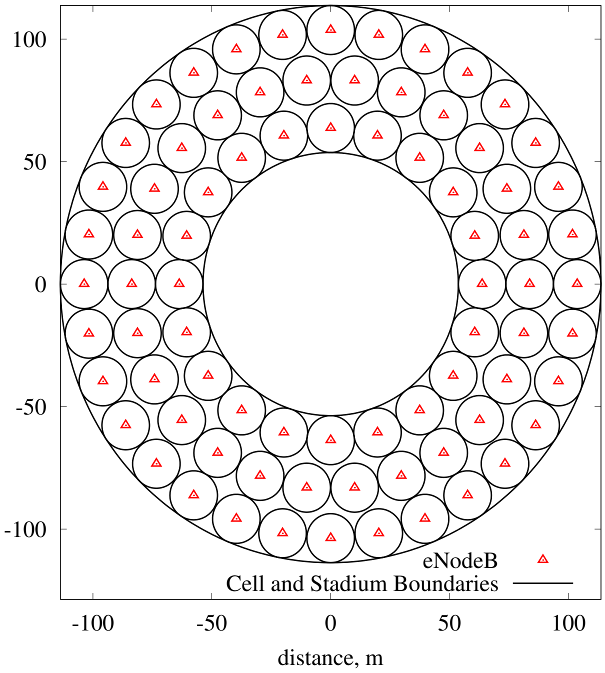


Figure 1: Stadium network architecture

- Each UE is associated with an eNB with a minimum estimated downlink pathloss to it, based on the Reference Signal Received Power (RSRP).
- Open loop control of the eNB transmit power is assumed, such that its signal power received at the UE is constant at -74 dBm (20 dB Signal to Noise Ratio).
- The minimum Signal to Interference plus Noise Ratio (SINR) that can support data transmission is 1.8 dB. The minimum SINR at which a new file transmission can be admitted is 5 dB.
- The radio propagation model used in the simulations is WINNER II A1 [13] designed for indoor communications, due to its support of small antenna heights and short transmission distances.

B. Traffic Model

The file inter-arrival times and file sizes are assumed to follow Pareto distributions. Such heavy-tailed distributions are well suited for describing typical internet traffic [14]. The mean file size is 1Mb and the mean file arrival rate is varied to obtain different values of offered traffic. The file transmission time length is determined by the data rate which is calculated using the Truncated Shannon Bound [15] as follows:

$$DR = \begin{cases} 0 & \gamma < \gamma_{min} \\ \alpha W log_2(1+\gamma) & \gamma_{min} \leq \gamma \leq \gamma_{max} \\ \alpha W log_2(1+\gamma_{max}) & \gamma > \gamma_{max} \end{cases} \quad (1)$$

where DR is the data rate in Mb/s, W is the subchannel bandwidth in MHz, $\alpha = 0.65$ - the implementation loss of the Shannon Bound, γ - the SINR on the link, γ_{min} - the minimum allowed SINR of 1.8 dB, γ_{max} - the SINR level which contributes to the maximum achievable data rate (21 dB).

C. Performance Metrics

The metrics used to assess the performance of the network are the probability of a file transmission being blocked and rescheduled ($P(blocking)$), if a suitable subchannel cannot be assigned for it, and the probability of interruption ($P(interruption)$), if the SINR on the given subchannel drops below 1.8 dB during the file transmission. The network is assumed to be serviceable only if $P(blocking)$ does not exceed 5% and $P(interruption)$ does not exceed 0.5%. These strict limits are especially relevant to the real-time traffic, e.g. video streaming and Voice Over IP, where transmission delays cannot be tolerated. They are chosen, since $P(blocking)$ and $P(interruption)$ in file-based traffic are equivalent to the probabilities of blocking and dropping in call-based traffic, and the 5% and 0.5% limits have been used for these performance metrics before, e.g. [10][16]. File interruptions or call dropping are in general significantly less tolerable than blocked files or calls. Therefore, it is justifiable to set the limit for $P(interruption)$ 10 times lower than that for $P(blocking)$.

III. Distributed Q-Learning Based Dynamic Spectrum Management

In distributed reinforcement learning based DSM the task of every eNB is to learn to prioritise among the available subchannels only by trial-and-error, with no spectrum sensing or frequency planning involved, and with no information exchange with the other eNBs.

A. Reinforcement Learning

Reinforcement learning is a model-free type of machine learning which is aimed at learning the desirability of taking any available action in any state of the environment only by trial-and error [3]. This desirability of an action is represented by a numerical value known as the Q-value - an expected cumulative reward for taking a particular action in a particular state. The job of an RL algorithm is to estimate the Q-values for every action in every state, which are all stored in an array known as the Q-table. In some cases where an environment is not represented by states, only the action space and a 1-dimensional Q-table are considered [17]. This is also the case investigated in this paper.

B. Stateless Q-Learning

One of the most successful and widely used RL algorithms is Q-learning, introduced in [18]. Since the learning problem described in the previous section does not require a state representation, a simple stateless variation of this algorithm, as formulated in [17], is used in this paper.

Each eNB maintains a Q-table such that every subchannel has an expected reward or Q-value associated with it. The Q-value represents the desirability of assigning a particular channel to a file transmission. Upon each file arrival, the eNB either assigns an available subchannel to its transmission or blocks it if no subchannels are available.

The Q-table is updated by the corresponding eNB each time it attempts to assign a subchannel to a file transmission. The update equation for stateless Q-learning, as defined in [17], is given below:

$$Q'(c) = Q(c) + \alpha(r - Q(c)) \tag{2}$$

where $Q(c)$ and $Q'(c)$ represent the Q-value of the selected subchannel c, before and after the update respectively, r is the reward associated with the most recent trial and determined by the reward function, and $\alpha \in [0, 1]$ is the learning rate parameter which weights recent experience with respect to previous estimates of the Q-values.

C. Q-table Initialisation and Reward Function

The values in the Q-tables are initialised to zero, so all eNBs start learning with equal choice among all available subchannels.

The reward function returns two discrete values:

- $r = -1$, if the file is blocked due to SINR being lower than the minimum admission threshold (5 dB) on the selected subchannel, or if it is later interrupted due to SINR being lower than the minimum transmission threshold of 1.8 dB.
- $r = 1$, if the file is successfully transmitted using the subchannel chosen by the eNB, i.e. if SINR is higher than 5 dB at the start and higher than 1.8 dB until the end of transmission.

D. Action Selection Strategy

The main role of an action selection strategy is to provide a balance between exploration and exploitation in an RL problem [3]. However, the problem discussed in this paper is simpler than most classical RL problems in one fundamental aspect - it is stateless. It is also a multi-agent (i.e. distributed) RL problem, which means that the decisions made by each eNB will affect the learning process of the other independent eNBs.

Therefore, a greedy action selection policy is used in this algorithm, i.e. each eNB always selects an available subchannel with the highest Q-value, if any. In this way, if an eNB discovers a good set of subchannels, it will continue using it to maximise performance and to make it easier for neighbouring base stations to learn to avoid the same subchannels.

E. Learning Rate

Every eNB in the network learns independently, and the learning environment, as perceived by each individual learning agent, depends on the choices made by other learning agents. Therefore the environment is locally dynamic from the viewpoint of every individual eNB.

Time-invariant values of the learning rate (α) are well-suited to such dynamic learning problems, since they essentially introduce the effect of a moving window, where the impact of older rewards on the current estimate gradually fades away, as seen from Equation (2).

Subsection IV-A explores the effect of varying the learning rate values across an appropriate range. It also presents and analyses a novel concept of the Win-or-Learn-Fast variable learning rate for distributed Q-learning based DSM.

IV. Choosing the Learning Rate

A. Win-or-Learn-Fast Learning Rate

This section of the paper explores the benefits of using the Win-or-Learn-Fast (WoLF) variable learning rate in distributed Q-learning based DSM. The WoLF principle states that the learning agent should learn faster when it is losing and more slowly when winning [9]. The adaptation of this variable learning rate principle in stateless Q-learning applied to DSM has been used by us in previous preliminary work [10][11]. However, that work focused on other topics, and the WoLF method of choosing the learning rate was not optimised, analysed or described in any depth.

The idea of a WoLF learning rate, proposed by us in [10] in the same multi-agent stateless Q-learning setting, is to split the value of the learning rate α into two cases - α_{win} and α_{lose} - when the subchannel chosen by the eNB successfully supported the file transmission and when it failed (blocking or interruption) respectively. If $\alpha_{win} < \alpha_{lose}$, the WoLF principle holds, since the agent is learning slower on successful trials (α_{win}) and faster on the failed ones (α_{lose}).

One of the advantages of using a WoLF learning rate is that it encourages thorough exploration in the early stages of learning. Since all values in the Q-tables are initially set to zero and the greedy action selection strategy is followed, if an eNB has several successful trials on a particular subchannel, its Q-value will increase and it will continue to be used. If, later on, the interference from other eNBs on this subchannel significantly increases, it will take fewer failed trials for its Q-value to fall below zero than it would if a fixed value of α was used. This can be proven by splitting the learning rate value into two cases - α_{win} and α_{lose}, substituting the reward values into Equation (2) and rearranging the terms to yield the expression for the change in Q-value $\Delta Q(c) = Q'(c) - Q(c)$ shown below:

$$\Delta Q(c) = \begin{cases} -\alpha_{win} Q(c) + \alpha_{win} & r = 1 \\ -\alpha_{lose} Q(c) - \alpha_{lose} & r = -1 \end{cases} \tag{3}$$

Comparing $|\Delta Q(c)|$ in both cases proves that, if $Q(c) > \frac{\alpha_{win} - \alpha_{lose}}{\alpha_{win} + \alpha_{lose}}$ and $\alpha_{win} < \alpha_{lose}$, $|\Delta Q(c)|$ is greater when $r = -1$, i.e. the changes in the Q-values are bigger when the negative rewards are received.

It would also have a similar effect if there has been a change in the learning environment, e.g. a change in network topology or traffic distribution. In such cases an eNB would start exploring other subchannels sooner. Another advantage of the WoLF learning rate is that at any stage of the operation of the network the ratio of successful to failed trials would need to be higher for a subchannel to maintain a high Q-value and keep being assigned, which is consistent with the goal of achieving low probabilities of blocking and interruption in a cellular system.

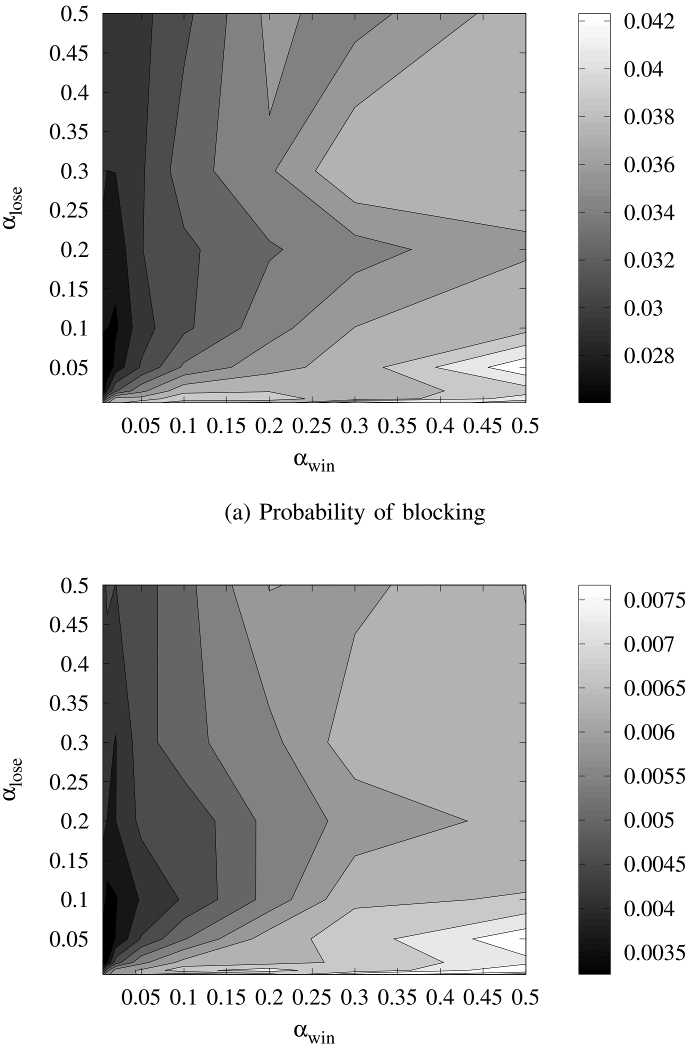
The simulation model of a stadium network presented in Section II is used to test the QoS provided to the UEs, using different combinations of the values of α_{win} and α_{lose}. 25% of the overall stadium capacity is randomly filled with wireless subscribers, i.e. $\approx$ 10,776 randomly distributed UEs.

B. Simulation Results

The contour plots in Figure 2 show the probabilities of file blocking and interruption after performing the simulations of the distributed Q-learning based DSM algorithm described in the previous section, using different combinations of α_{win} and α_{lose}. The simulations lasted 800,000 transmissions, which constituted 800,000 reinforcement learning trials for all eNBs in total, and took approximately 50 minutes at 256 Mb/s offered traffic. The $P(blocking)$ and $P(interruption)$ values shown on the contour plots are calculated over the last 240,000 transmissions, when the eNBs have had sufficient time to learn mature DSM policies. The values of α_{win} and α_{lose} varied in the range $[0.005, 0.5]$.

Both plots demonstrate the performance improvement when the WoLF principle of varying the learning rate is used. The best performance is achieved in the small darkest region of the plots around the point $(0.01, 0.05)$. The fixed learning rate values lie on the 45^o diagonal (where $\alpha_{win} = \alpha_{lose}$), and perform significantly worse than those in the "WoLF region" above and to the left of the diagonal.

Figure 3 shows the difference in the QoS time response (i.e. how QoS improves over time) of the distributed Q-learning based DSM algorithm with a typical choice of the fixed learning rate value of 0.1, and the WoLF variable learning rate of $\{0.01, 0.05\}$. The first $P(blocking)$ and $P(interruption)$ points on the graphs at zero time are obtained by simulating a random dynamic spectrum access scheme, where all eNBs randomly choose among all available subchannels. This effectively represents the starting point of a Q-learning algorithm with a Q-table initialised to zero.

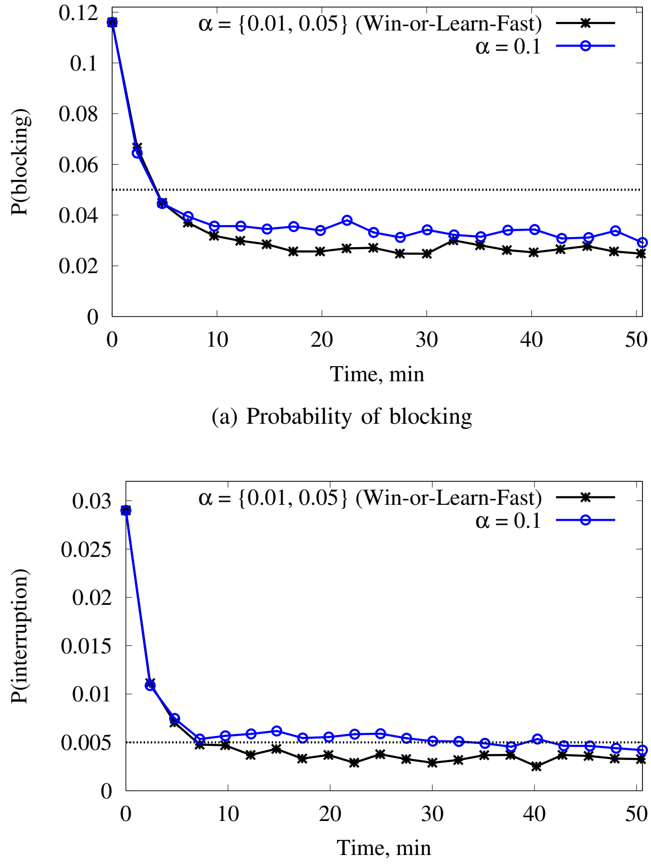


(a) Probability of blocking

(b) Probability of interruption

Figure 2: Network-wide quality of service at 256 Mb/s offered traffic, with different values of the learning rates for the positive (α_{win}) and the negative outcome (α_{lose})

In the early stages of learning, in $\approx$10 minutes, the WoLF learning rate achieves better QoS due to its increased adaptivity to changes in the policies of all eNBs, which are affecting the learning process of every individual eNB. However, after $\approx$40 minutes, long after the Q-learning algorithm has reached its steady state, the QoS achieved using the WoLF learning rate is still significantly better, which suggests that fixed learning rates cause the Q-learning algorithm to converge on poorer solutions, compared to the WoLF variable learning rates. The results presented in the next subsection show that this improvement is consistent across a range of different traffic loads. Therefore, it is not necessary to optimise the WoLF learning rate values for each of them individually.



(a) Probability of blocking

(b) Probability of interruption

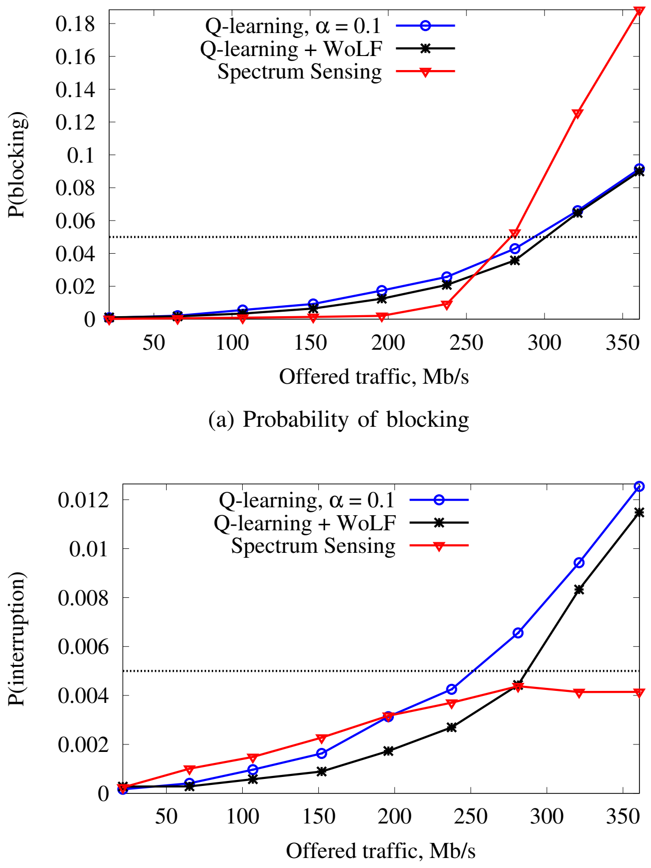
Figure 3: The difference between the quality of service time responses at 256 Mb/s offered traffic, using the fixed learning rate of 0.1, and the Win-or-Learn-Fast variable learning rate of $\{0.01, 0.05\}$

C. *Comparison with Spectrum Sensing Based DSM*

In Figure 4 the difference in steady state probabilities of file blocking and interruption, using the fixed learning rate of 0.1 and the WoLF learning rate $\{0.01, 0.05\}$, is demonstrated across the full range of serviceable traffic loads. It also compares the results with the performance of a simple opportunistic spectrum sensing based DSM scheme described by the flow diagram in Figure 5. The interference threshold used for it was -84 dBm, which is 10 dB above the noise floor and 10 dB below the received power at the UE receivers.

Firstly, the WoLF learning rate introduced a significant and consistent improvement in the QoS, especially the probability of interruption, achieved by the distributed Q-learning based DSM algorithm across all traffic loads above ≈80 Mb/s. At lower traffic loads $P(blocking)$ and $P(interruption)$ are very low in both cases. Secondly, the spectrum sensing based scheme performed better in terms of $P(blocking)$ at low and medium traffic loads due to its *listen-before-talk* capability. However, as the traffic load increased, its performance deteriorated significantly faster than that achieved by both Q-learning schemes, exceeding the 5% limit at the offered traffic of ≈280 Mb/s. The most significant result from these plots is the difference in $P(interruption)$ achieved by the spectrum sensing based DSM scheme and the Q-learning algorithm with the WoLF variable learning rate. The "Q-learning+WoLF" scheme has achieved a ≈44% lower $P(interruption)$ on average, compared to the spectrum sensing based DSM scheme at the traffic loads between 65 and 280 Mb/s. At higher traffic loads $P(interruption)$ achieved by the spectrum sensing based scheme is lower due to the rapid increase in $P(blocking)$. The "Q-learning+WoLF" scheme has also managed to provide a slightly wider range of traffic loads with acceptable QoS, i.e. where $P(blocking) < 5\%$ and $P(interruption) < 0.5\%$, than that provided by the spectrum sensing based scheme (by $\approx 3\%$), whereas the regular Q-learning algorithm with a fixed learning rate was significantly outperformed by both schemes in this regard.

Although no spectrum sensing was involved in the "Q-learning + WoLF" approach, it has outperformed the spectrum sensing based DSM algorithm in terms of the probability of interruption, the range of usable traffic loads, and the overall robustness to changes in the network traffic load. The only



(a) Probability of blocking

(b) Probability of interruption

Figure 4: Steady state quality of service at different traffic loads, using Q-learning with $\alpha = 0.1$, Q-learning with the Win-or-Learn-Fast variable learning rate, and spectrum sensing based dynamic spectrum management

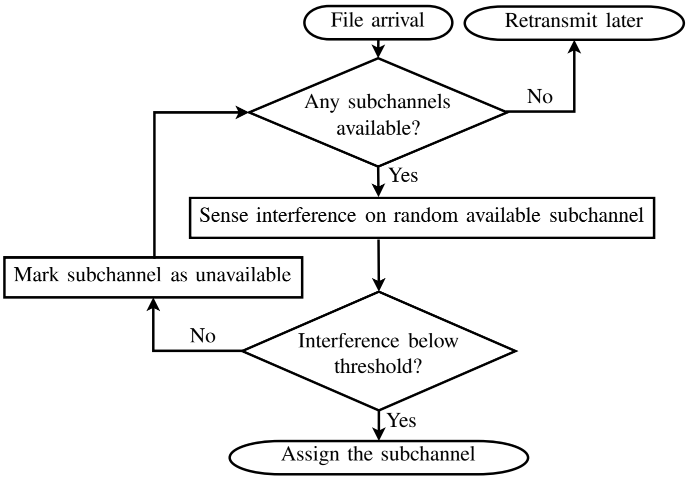

File arrival
Retransmit later
Any subchannels available?
No
Yes
Sense interference on random available subchannel
Mark subchannel as unavailable
No
Interference below threshold?
Yes
Assign the subchannel

Figure 5: Flow diagram of the spectrum sensing based opportunistic spectrum access scheme used for baseline comparison

disadvantage of the Q-learning approach is the initial learning period, where the QoS starts at a very poor level due to the lack of information in the Q-tables and it takes the eNBs time to learn mature DSM policies (e.g. time responses in Figure 3). However, this poor initial performance can be significantly mitigated by extending the simple Q-learning algorithm to more advanced schemes, such as transfer learning [19], case-based reinforcement learning [11], heuristically accelerated reinforcement learning [20], e.g. using spectrum sensing and/or a radio environment map [21] as heuristic acceleration, etc.

V. Conclusion

In this paper we demonstrate the importance of the learning rate parameter in distributed Q-learning based dynamic spectrum management (DSM) in cognitive cellular systems, and its effect on the quality of service (QoS) provided to the users of the network. A concept of the Win-or-Learn-Fast (WoLF) variable learning rate is presented in the context of DSM. We empirically demonstrate that it is possible to achieve significant QoS performance improvements, in terms of the probabilities of file blocking and interruption, simply by choosing an appropriate WoLF learning rate for a distributed Q-learning based DSM algorithm. Large scale simulations of a stadium temporary event network show that a distributed Q-learning algorithm with a WoLF variable learning rate can outperform a spectrum sensing based opportunistic spectrum access scheme in terms of the probability of interruption and overall range of serviceable traffic loads, with no spectrum sensing involved. This ensures that the QoS provided by a simple distributed Q-learning based DSM algorithm is maximised or near-maximised, before it is extended to more sophisticated schemes, such as transfer learning, case-based reinforcement learning, reinforcement learning heuristically accelerated by spectrum sensing and radio environment maps.

Acknowledgment

This work has been funded by the ABSOLUTE Project (FP7-ICT-2011-8-318632), which receives funding from the 7th Framework Programme of the European Commission.

References

[1] Y. Saleem, A. Bashir, E. Ahmed, J. Qadir, and A. Baig, “Spectrum-aware dynamic channel assignment in cognitive radio networks,” in *International Conference on Emerging Technologies (ICET)*, 2012.

[2] J. Sachs, I. Maric, and A. Goldsmith, “Cognitive cellular systems within the TV spectrum,” in *IEEE Symposium on New Frontiers in Dynamic Spectrum*, 2010.

[3] R. Sutton and A. Barto, *Reinforcement Learning: An Introduction.* MIT Press, Cambridge, MA, 1998.

[4] A. Ko, R. Sabourin, and F. Gagnon, “Performance of distributed multi-agent multi-state reinforcement spectrum management using different exploration schemes,” *Expert Systems with Applications*, vol. 40, no. 10, pp. 4115 – 4126, 2013.

[5] A. Feki, V. Capdevielle, and E. Sorsy, “Self-organized resource allocation for LTE pico cells: A reinforcement learning approach,” in *IEEE Vehicular Technology Conference (VTC Spring)*, 2012.

[6] M. Bennis and D. Niyato, “A Q-learning based approach to interference avoidance in self-organized femtocell networks,” in *2010 IEEE GLOBECOM Workshops (GC Wkshps)*, 2010.

[7] Q. Zhao and D. Grace, “Application of cognition based resource allocation strategies on a multi-hop backhaul network,” in *IEEE International Conference on Communication Systems (ICCS)*, 2012, pp. 423–427.

[8] R. Valcarce, et al., “Airborne base stations for emergency and temporary events,” in *International Conference on Personal Satellite Services*, 2013.

[9] M. Bowling and M. Veloso, “Multiagent learning using a variable learning rate,” *Artificial Intelligence*, vol. 136, pp. 215–250, 2002.

[10] N. Morozs, T. Clarke, and D. Grace, “A novel adaptive call admission control scheme for distributed reinforcement learning based dynamic spectrum access in cellular networks,” in *International Symposium on Wireless Communication Systems (ISWCS)*, 2013.

[11] N. Morozs, D. Grace, and T. Clarke, “Case-based reinforcement learning for cognitive spectrum assignment in cellular networks with dynamic topologies,” in *Military Communications and Information Systems Conference (MCC)*, 2013.

[12] 3GPP, “LTE; Evolved Universal Terrestrial Radio Access (E-UTRA); Physical layer procedures (3GPP TS 36.213 version 11.5.0 Release 11) ,” Dec. 2013.

[13] P. Kyösti, et al., “IST-4-027756 WINNER II D1.1.2 v1.2 WINNER II channel models,” Feb. 2008.

[14] M. Crovella, M. Taqqu, and A. Bestavros, “A practical guide to heavy tails,” R. Adler, R. Feldman, and M. Taqqu, Eds. Cambridge, MA, USA: Birkhauser Boston Inc., 1998, ch. Heavy-tailed Probability Distributions in the World Wide Web, pp. 3–25.

[15] A. Burr, A. Papadogiannis, and T. Jiang, “MIMO truncated shannon bound for system level capacity evaluation of wireless networks,” in *IEEE Wireless Communications and Networking Conference Workshops (WCNCW)*, 2012.

[16] D. Akerberg and F. Brouwer, “On channel definitions and rules for continuous dynamic channel selection in coexistence etiquettes for radio systems,” in *IEEE Vehicular Technology Conference (VTC)*, 1994.

[17] C. Claus and C. Boutilier, “The dynamics of reinforcement learning in cooperative multiagent systems,” in *Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence*, 1998.

[18] C. Watkins, “Learning from Delayed Rewards,” Ph.D. dissertation, University of Cambridge, England, 1989.

[19] Q. Zhao, T. Jiang, N. Morozs, D. Grace, and T. Clarke, “Transfer learning: A paradigm for dynamic spectrum and topology management in flexible architectures,” in *IEEE Vehicular Technology Conference (VTC Fall)*, 2013.

[20] R. Bianchi, M. Martins, C. Ribeiro, and A. Costa, “Heuristically-accelerated multiagent reinforcement learning,” *Cybernetics, IEEE Transactions on*, vol. 44, pp. 252–265, 2014.

[21] M. Pesko, T. Javornik, M. Štular, and M. Mohorčič, “The comparison of methods for constructing the radio frequency layer of radio environment map using participatory measurements,” in *Workshop of COST Action IC0902 on Cognitive Radio and Networking for Cooperative Coexistence of Heterogeneous Wireless Networks*, 2013.