A Novel Adaptive Call Admission Control Scheme for Distributed Reinforcement Learning Based Dynamic Spectrum Access in Cellular Networks

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

***Abstract*—This paper introduces a novel Q-value based adaptive call admission control scheme (Q-CAC) for distributed reinforcement learning (RL) based dynamic spectrum access (DSA) in mobile cellular networks, which provides a good quality of service (QoS) without the need for spectrum sensing. A DSA algorithm has been developed in this paper using the stateless Q-learning algorithm with Win-or-Learn-Fast (WoLF) learning rates. Its performance was analysed using the spatial distribution of the probabilities of call blocking (BP) and dropping (DP) across the network and compared to that of a 100% accurate spectrum sensing based DSA scheme. The Q-CAC scheme demonstrated good controllability of the blocking probability using a Q-value based call admission threshold parameter. It significantly reduced spatial fluctuations in BP and DP, thus providing more cells with acceptable quality of service (QoS).**

***Keywords*—Dynamic Spectrum Access, Adaptive Call Admission Control, Distributed Reinforcement Learning**

I. Introduction

One of the fundamental tasks of a mobile cellular network is to divide the available spectrum into a set of channels and set up a protocol for assigning them to incoming calls in a way which provides a good quality of service (QoS) to the users. Modern communication systems, such as cognitive radio and LTE networks, require more sophisticated and intelligent schemes for channel assignment than static spectrum allocation. Such schemes belong to the area of dynamic spectrum access (DSA).

Reinforcement learning (RL) is a machine learning technique for learning solutions to various decision problems only by trial-and-error [1]. In terms of DSA, RL is a state-of-the-art technique widely investigated within the area of wireless communications. It has been successfully applied to a range of problems such as LTE pico cells [2], cognitive radio [3] [4] and multi-hop backhaul networks [5].

However, there is little evidence of work on applying fully distributed RL to DSA at the base station level in mobile cellular networks. Notable examples can be found in [6] and [7]. The distributed DSA approach has a significant advantage over centralised methods in that no information exchange is required among independently learning base stations and the network operation does not rely on a single computing unit. Also, RL techniques eliminate the requirement for spectrum sensing during the channel allocation process. The channel assignment policies are obtained purely by trial-and error. This RL-based trial-and-error approach has both advantages and disadvantages compared with spectrum sensing based DSA methods such as [8]. Its significant disadvantage is that the fundamental source of information about the channel availability is removed. The challenge is then to learn a desired set of channels only from experience as opposed to making instantaneous sensing measurements. However, if this challenge can be overcome, the RL approach introduces some advantages over spectrum sensing based DSA. The design of the radio equipment is greatly simplified by eliminating the need for spectrum sensing functionality. It also makes the decisions made by the base stations independent of the reliability of the spectrum sensing data.

The purpose of this paper is to present a simple distributed Q-learning based DSA algorithm together with a novel adaptive Q-value based call admission control (CAC) scheme (Q-CAC) which provides a feasible alternative to spectrum sensing based DSA methods. The performance of the algorithm is evaluated using probabilities of call blocking (BP) and dropping (DP) and is compared to that of a perfect spectrum sensing based DSA scheme. The results are analysed in terms of their spatial distribution per base station, as opposed to their average network-wide values. This type of analysis is especially important for large scale networks, since it is better to provide acceptable QoS across the whole network rather than having a mixture of high QoS cells and coverage holes.

The rest of the paper is organised as follows: in Section II the DSA problem and the network model are defined. In Section III the development of the distributed RL algorithm for DSA is described. In Section IV the Q-CAC scheme for RL based DSA is introduced, followed by a large scale simulation of the developed algorithm in Section V. Finally, conclusions are given in Section VI.

II. Problem Description

A. Mobile Cellular Network

The network model used in this paper consists of a square rural service area covered by a grid of base stations spaced 2 km apart. The initial experiments in Sections III and IV use a small network of 4 base stations covering a 4x4 km area. After developing the DSA and Q-CAC algorithms using this model, it is tested on a larger 14x14 km service area covered by 49 base stations in Section V. The general network architecture is depicted in Fig. 1. We stress that the DSA and CAC schemes developed in this paper are fully distributed and do not employ any backhaul communications among the base stations or a centralised control unit.

The assumptions used in the reported simulations are listed

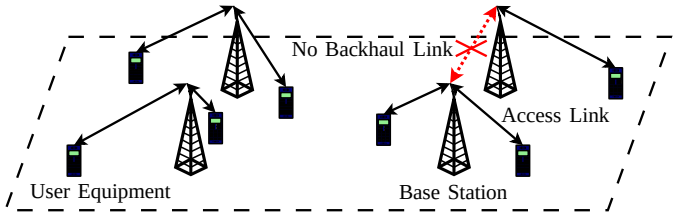

Fig. 1. Network architecture

below:

- The available resources are divided into 36 logical channels. Adjacent-Channel Interference is assumed to be negligible and only uplink communications are considered.
- Fixed transmission power of 23 dBm and 2.6 GHz frequency band are used by all user equipment (UE).
- The minimum Signal to Interference plus Noise Ratio (SINR) for accepting a new call is 5 dB, and the calls are dropped if the SINR falls below 1.8 dB.
- The receiver noise floor is -100 dBm, obtained by assuming 290 K temperature, 10 MHz bandwidth and 4dB noise figure.
- Each UE chooses which base station to connect to such that the overall attenuation of its signal is minimised.
- The transmission is continuous until a call is completed.

B. Radio Propagation Model

The propagation model used to calculate the path loss between the UE transmitters and the base station receivers is the WINNER 2 model, described in detail in [9]. In particular, the variation designed for a Line-of-Sight (LOS) rural macrocell scenarios is used (WINNER 2 D1), since it is most relevant to the network architecture discussed in this paper. The formula for calculating path loss using this model is given below:

$$PL = 21.5log_{10}(d) + 44.2 + 20log_{10}(0.2f_c) + SFL \quad (1)$$

where PL is the path loss in dB, d is the distance between the receiver and the transmitter in metres, f_c is the carrier frequency in GHz and SFL is the log-normal shadow fading loss with the standard deviation of 4dB and 0dB mean.

C. Traffic Model

The call arrival rate is modelled as a Poisson process with a constant mean arrival rate of λ_{UE} (calls per minute per UE) for all UEs in the network. The call duration is also an exponentially distributed random variable with the mean holding time of 1 minute.

D. Learning Objective

The objective of the learning problem investigated in this paper is for all base stations to prioritise among the available channels in a fully distributed fashion, only by trial-and-error. No communication between the base stations is assumed in order to achieve this objective. Therefore, it is a problem of distributed DSA.

The metrics used to evaluate the performance of the algorithm are the probabilities of call blocking (BP) and dropping (DP). The network is assumed to be serviceable only if the BP does not exceed 5% and the DP does not exceed 0.5%. In general, call dropping is considered significantly less tolerable than blocking. Therefore, it is justifiable to set the DP threshold 10 times lower than that for BP [10] [11]. The primary aim of the DSA and CAC schemes discussed in this paper is to provide acceptable QoS to as many parts of the network as possible, as opposed to optimising the average network-wide BP and DP.

III. Reinforcement Learning Algorithm

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 [1]. 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 a 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 [12]. 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 [13]. Since the learning problem described in the previous section does not require a state representation, a simple stateless variation of this algorithm, formulated in [12], is used in this paper.

Each base station maintains a Q-table such that every channel has an expected reward or Q-value associated with it. The Q-value represents the desirability of assigning a particular channel to an arriving call. Upon each call arrival, the base station has a choice of either assigning an available channel to the call or blocking it if no channel can be assigned.

The Q-table is updated by the corresponding base station each time it attempts to assign a channel to an arriving call. The update formula for stateless Q-learning, as defined in [12], is given below:

$$Q'(c) = Q(c) + \alpha(r - Q(c)) \quad (2)$$

where $Q(c)$ and $Q'(c)$ represent the Q-value of the selected channel before and after the update respectively, r is the reward associated with the most recent trial and determined by the reward function, and α 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-table are initialised to zero, so all base stations start learning with equal choice among all available channels.

The reward function returns two discrete values:

- -1, if the call is blocked due to SINR being lower than 5 dB on the selected channel.
- +1, if the connection is successfully established using the channel chosen by the base station, i.e. if SINR is higher than 5 dB.

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 [1]. 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 learning agent will affect the learning process of the other independent agents.

Therefore, a greedy action selection policy is used in this paper, i.e. each base station always selects an available channel with the highest Q-value, if any. In this way, if a base station discovers a good set of channels, it will continue using it to maximise performance and to make it easier for neighbouring base stations to learn to avoid the same channels. Investigating the effect of different action selection strategies on the algorithm performance is beyond the scope of this paper.

E. Learning Rate

Each base station 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, even though the environment is globally static, it is essentially dynamic from the viewpoint of each individual base station.

Fixed values of the learning rate 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 [1], as seen from Equation (2).

The DSA algorithm developed in this paper also adopts the principle of the Win-or-Learn-Fast (WoLF) algorithm for variable learning rate, as introduced in [14]. The WoLF principle states that the learning agent should learn faster when it is losing and more slowly when winning. Since there are only two possible outcomes associated with learning - blocking ("lose") and successful call arrival ("win") - it is sufficient to assign a fixed learning rate to each.

Fig. 2 shows the effect of varying the learning rate for the positive outcome (α_{pos}) from 0 to 0.4, whilst keeping the learning rate for the negative outcome (α_{neg}) fixed at 0.2. The graph was obtained by simulating the 4 base station model described in the previous section with a 0.17 Erlang traffic intensity per channel. The vertical axis displays the steady-state blocking and dropping probabilities, i.e. those to which the RL algorithm has converged.

There is a significant degradation of performance for very low values of α_{pos}. However, the best point on this graph occurs around 0.05, which demonstrates the benefit of the WoLF principle. Therefore, the learning rate used in the experiments in this paper is 0.2 for call blocking and 0.05 for successful call arrivals.

IV. CAC FOR Q-LEARNING BASED DISTRIBUTED DSA

So far, the RL algorithm introduced in the previous section does not use any form of CAC. Each base station always assigns a channel to an arriving call, unless the whole channel set is occupied.

In this section a novel Q-value based adaptive CAC scheme (Q-CAC) for improving the spatial distribution of BP and DP in a cellular network is introduced which can be used in conjunction with RL algorithms for DSA, such as developed in the previous section. A new parameter is incorporated into the algorithm - the Q-value based call admission threshold (CAT).

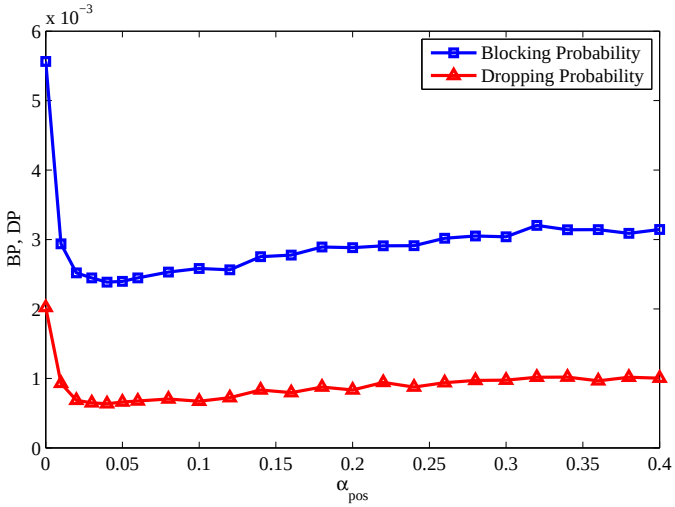

Fig. 2. Steady state probability of blocking (BP) and dropping (DP) of the stateless Q-learning algorithm with different learning rates for positive outcome (α_{pos}), while α_{neg} is constant at 0.2

Each base station maintains its own Q-table which ranks the channels from best to worst based on their Q-values, $Q(c) \in [-1, 1]$ for all channels. The CAT is defined as the minimum Q-value with which a channel can be assigned to an arriving call. Any channels with Q-values less than the CAT are considered unavailable for channel assignment, thus reducing the size of the channel set available to the given base station.

A classical negative feedback control structure [15] was employed to facilitate dynamic tuning of CAT values for controlling the BP at each base station. This control loop is shown in Fig. 3.

It exploits the relationship between the CAT and BP measured at a given base station. When CAT is at -1, it does not cut off any channels, therefore it does not have any effect on the BP performance. However, as it increases, fewer channels are available and, as a result, more calls are blocked.

The reference input of this control system (BP_{ref}) is the interval of desired BP values, and the output is the actual BP (BP) measured at a given base station. The experiments in this paper use an interval of $[0.04, 0.045]$ which leaves a small safety margin between its upper bound and a maximum acceptable BP of 0.05. The control scheme for tuning the CAT is described in Algorithm 1.

Note that the CAT is set to -1, when the measured BP exceeds the upper limit of the desired BP interval. This eliminates overshoots in the BP response, as it is crucial not to exceed the 5% BP limit to continuously provide acceptable QoS. A unity negative feedback control law is used only when the measured BP is below the lower limit of the reference interval (lines 9-11 in Algorithm 1). K is the gain which converts the BP error (BP_{err}) into the CAT correction term,

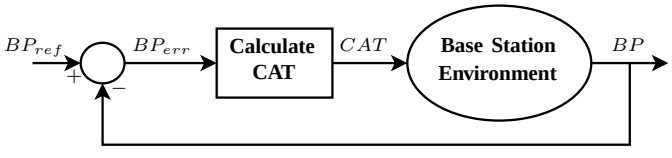

Fig. 3. Negative feedback control loop for CAT tuning

Algorithm 1 CAT tuning algorithm

$CAT = -1$, $BP_{ref} = [0.04, 0.045]$
while base station is on **do**
  Wait for a call arrival
  Try to assign a channel and measure BP
  if $BP > max(BP_{ref})$ **then**
    $CAT_{new} = -1$
  else if $BP \in BP_{ref}$ **then**
    $CAT_{new} = CAT$
  else if $BP < min(BP_{ref})$ **then**
    $BP_{err} = mean(BP_{ref}) - BP$
    $CAT_{new} = CAT + K * BP_{err}$
  end if
  $CAT = CAT_{new}$
end while

since the error in CAT is linearly proportional to the error in BP but not necessarily of the same magnitude. K directly affects the rate at which the CAT responds to the errors in BP, as well as stability of this response.

The BP is measured using a moving window which stores the outcome of the last N call arrivals in a binary vector. Each element is either 1 for blocked call or 0 for successful call. Due to this binary nature of BP measurements, the estimated BP value over last N calls often experiences small increments and decrements with every update of the vector. An input interval $[0.04, 0.045]$ was used in favour of a single reference value (e.g. 0.0425), to prevent the CAT value from persistent corrections when the measured BP value oscillates within a close neighbourhood of the desired value.

Fig. 4 shows an example of a BP time response at an arbitrarily selected base station, using the small 4x4 km network model described in Section II with 0.17 Erlang traffic intensity per channel. The BP is successfully controlled using the Q-CAC scheme with the feedback gain $K = 2$, which was activated after the base stations had enough time to learn mature channel assignment policies (in this case, after 3000 call arrivals at the given base station). Therefore, the Q-CAC scheme with feedback gain of $K = 2$ is used in the large scale simulation discussed in the next section.

V. Simulation Results

The developed algorithm was simulated on a 14x14 km network of 49 base stations with 2000 UEs randomly distributed across the service area. The performance of the Q-learning based DSA algorithm with and without Q-CAC was compared to that of a spectrum sensing based DSA scheme [16]. This scheme was assumed to be able to measure the interference-plus-noise power on each channel prior to making a decision and predict the achievable SINR with 100% accuracy. Upon each call arrival it assigns an available channel with the highest achievable SINR, unless all of them are below the acceptance threshold of 5dB, in which case a call gets blocked.

Fig. 5 shows the cumulative distribution function (CDF) of the BP measured at each individual base station for the 3 schemes described above. The network-wide traffic load is 100 Erlangs obtained by setting the arrival rate to $\lambda_{UE} = 0.05$ calls per minute per user. All 3 schemes were simulated using identical call arrival and holding times to ensure a fair comparison of their performance. Firstly, the BP does not exceed the 5% limit in any single cell for both Q-learning based methods. As expected, the spectrum sensing based approach yields superior BP performance, keeping it at zero for all base stations. However, the important point here is that the network is equally 100% serviceable in terms of BP using either a spectrum sensing based approach or the Q-learning based DSA algorithm with no spectrum sensing. Secondly, all base stations have successfully kept their BP in the desired interval of $[0.04, 0.045]$ or slightly above it using the Q-CAC scheme, thus significantly reducing the spatial fluctuations in BP across the network.

The benefits of using the Q-CAC scheme are demonstrated in Fig. 6. It shows the CDF of the DP at each individual base station using the 3 schemes. Firstly, the DP exceeds the 0.5% limit at some base stations for all algorithms. Therefore, it is the main QoS constraint. Secondly, the Q-CAC scheme significantly improves on the spatial distribution of the DP for the Q-learning based DSA algorithm. The Q-learning scheme without Q-CAC provided only 71% of the network with acceptable QoS, whereas the Q-CAC scheme raised it to 86% which is significantly closer to the result of the spectrum

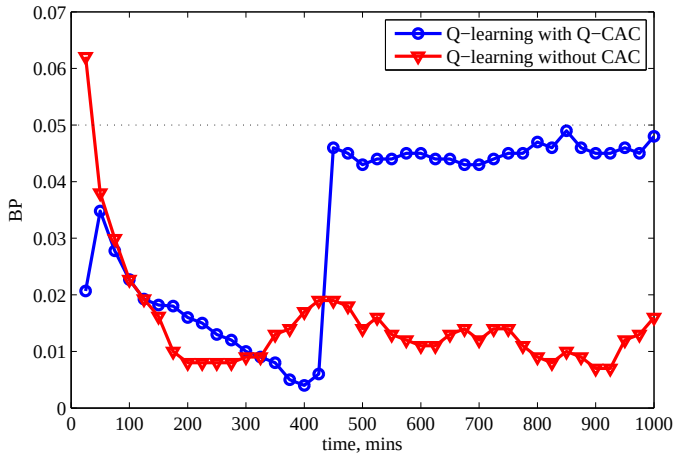

Fig. 4. The effect of the Q-CAC scheme on the blocking probability (BP) time response at an individual base station running the Q-learning based DSA algorithm

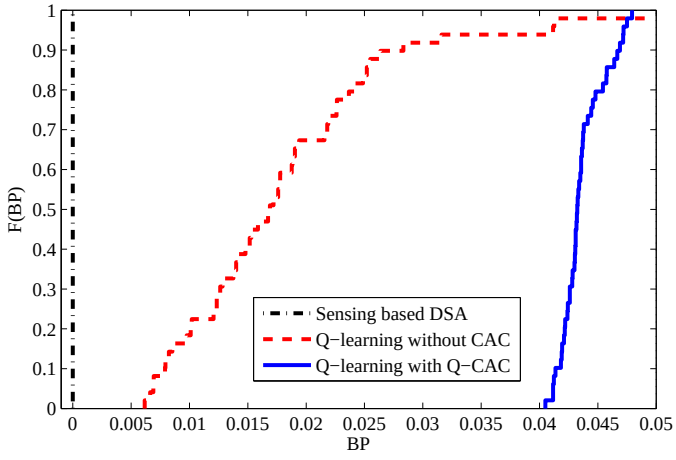

Fig. 5. CDF of the blocking probability (BP) at individual base stations for sensing based DSA and Q-learning based DSA with and without Q-CAC at a high traffic load

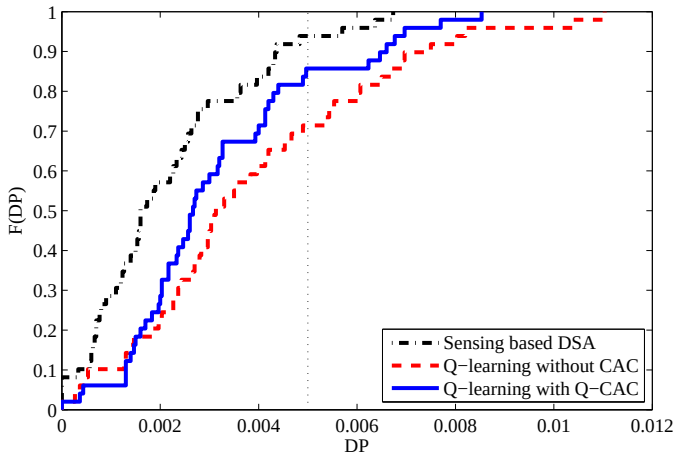

Fig. 6. CDF of the dropping probability (DP) at individual base stations for sensing based DSA and Q-learning based DSA with and without Q-CAC at a high traffic load

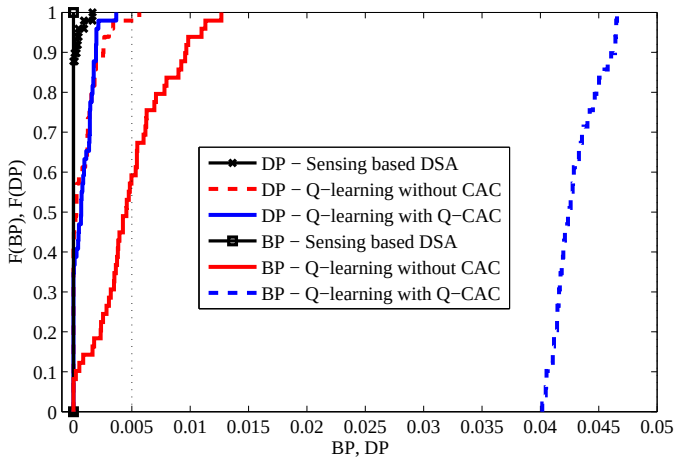

Fig. 7. CDF of the probabilities of blocking (BP) and dropping (DP) at individual base stations for sensing based DSA and Q-learning based DSA with and without Q-CAC at a medium traffic load

sensing based algorithm (94%), but with no sensing involved.

Fig. 7 shows an example of the network performance at a lighter traffic load - 50 Erlangs. It includes the CDFs of BP and DP of all 3 schemes. Here, the Q-CAC scheme has once again improved the DP performance of the Q-learning based DSA algorithm. The network is 100% serviceable in terms of BP and DP for both the spectrum sensing based algorithm and the Q-learning + Q-CAC scheme, which is the primary concern of the DSA and CAC methods investigated in this paper. It also shows that the Q-CAC scheme behaves adaptively, since it provides similar BP performance regardless of the traffic load.

VI. Conclusion

We have developed a fully distributed stateless Q-learning based DSA algorithm with a novel Q-value based adaptive CAC scheme (Q-CAC) for RL based DSA. It significantly reduces the spatial fluctuations in BP and DP across a large scale network and provides more cells with acceptable QoS without the need for spectrum sensing. The combination of the Q-learning based scheme with Q-CAC has been shown to have similar performance to a 100% accurate spectrum sensing based DSA scheme. Therefore, it is a viable yet simpler alternative to the spectrum sensing based DSA methods.

The main advantage of the DSA and Q-CAC schemes developed in this paper is that each base station only uses local information about its own trials, yet delivering comparable performance to spectrum sensing based methods. These algorithms are simple, flexible and easy to implement in a real network. They are also well suited to dynamic environments due to their deliberately designed fixed learning rates.

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

[2] A. Feki, V. Capdevielle, and E. Sorsy, "Self-organized resource allocation for lte pico cells: A reinforcement learning approach," in *Vehicular Technology Conference (VTC Spring), 2012 IEEE 75th*, 2012, pp. 1–5.

[3] Y. Teng, Y. Zhang, F. Niu, C. Dai, and M. Song, "Reinforcement learning based auction algorithm for dynamic spectrum access in cognitive radio networks," in *Vehicular Technology Conference Fall (VTC 2010-Fall), 2010 IEEE 72nd*, 2010, pp. 1–5.

[4] T. Jiang, D. Grace, and P. D. Mitchell, "Efficient exploration in reinforcement learning-based cognitive radio spectrum sharing," *Communications, IET*, vol. 5, no. 10, pp. 1309–1317, Jul. 2011.

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

[6] N. Lilith and K. Dogancay, "Distributed dynamic call admission control and channel allocation using sarsa," in *Communications, 2005 Asia-Pacific Conference on*, Oct., pp. 376–380.

[7] ——, "Distributed reduced-state sarsa algorithm for dynamic channel allocation in cellular networks featuring traffic mobility," in *Communications, 2005. ICC 2005. 2005 IEEE International Conference on*, vol. 2, 2005, pp. 860–865 Vol. 2.

[8] A. Hoang, Y. Liang, Y. Zeng, and D. Wong, "Distributed opportunistic spectrum access with imperfect spectrum sensing," in *Communication Systems (ICCS), 2010 IEEE International Conference on*, 2010, pp. 87–91.

[9] P. Kyösti, J. Meinilä, L. Hentilä, X. Zhao, T. Jämsä, C. Schneider, M. Narandzić, M. Milojević, A. Hong, J. Ylitalo, V. Holappa, M. Alatossava, R. Bultitude, Y. de Jong, and T. Rautiainen, "IST-4-027756 WINNER II D1.1.2 v1.2 WINNER II channel models," Feb. 2008.

[10] D. Akerberg and F. Brouwer, "On channel definitions and rules for continuous dynamic channel selection in coexistence etiquettes for radio systems," in *Vehicular Technology Conference, 1994 IEEE 44th*, 1994, pp. 809–813 vol.2.

[11] D. Grace, "Distributed Dynamic Channel Assignment for the Wireless Environment," Ph.D. dissertation, University of York, UK, 1998.

[12] 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*. American Association for Artificial Intelligence, 1998, pp. 746–752.

[13] C. Watkins, "Learning from Delayed Rewards," Ph.D. dissertation, University of Cambridge, England, 1989.

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

[15] N. Nise, *Control Systems Engineering*, 4th ed. Wiley, 2004.

[16] D. Goodman, S. Grandhi, and R. Vijayan, "Distributed dynamic channel assignment schemes," in *Vehicular Technology Conference, 1993., 43rd IEEE*, 1993, pp. 532–535.