Using Bayesian Networks for Convergence Analysis of Intelligent Dynamic Spectrum Access Algorithms

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*—In this paper we propose a novel Bayesian network based model for analysing convergence properties of reinforcement learning (RL) based dynamic spectrum access (DSA) algorithms. It uses a minimum complexity DSA problem for probabilistic analysis of the joint policy transitions of RL algorithms. A Monte Carlo simulation of a distributed Q-learning DSA algorithm shows that the proposed approach exhibits remarkable accuracy of predicting convergence behaviour of such algorithms. Furthermore, their behaviour can also be expressed in the form of an absorbing Markov chain, derived from the novel Bayesian network model. This representation enables further theoretical analysis of convergence properties of RL based DSA algorithms. The main benefit of the analysis tool presented in this paper is that it enables the design and theoretical evaluation of novel DSA schemes by extending the proposed Bayesian network model.**

***Keywords*—Distributed Reinforcement Learning; Bayesian Networks; Dynamic Spectrum Access**

I. Introduction

One of the fundamental tasks of a cellular system is spectrum management, concerned with dividing the available spectrum into a set of resources 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 access (DSA) techniques play a key role in utilising the given spectrum efficiently. This has given rise to 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 DSA is reinforcement learning (RL); 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 DSA problems and scenarios, such as cognitive radio networks [4], femto-cell networks [5], cognitive wireless mesh networks [6], as well as generic cellular networks [7]. This paper investigates distributed RL based DSA. The distributed RL 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 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]. An example of such a scenario investigated in the EU FP7 ABSOLUTE project is a temporary cognitive cellular infrastructure that is deployed in and around a stadium to provide extra capacity and coverage to the mobile subscribers and event organizers involved in a temporary event, e.g. a football match or a concert [9].

An important step in designing RL algorithms not only for DSA applications, but also for any other type of learning problems, is to perform theoretical analysis of their convergence. There is a large amount of previous work on probabilistic analysis of RL algorithms applied to wireless communications problems, where the researchers have stochastically modelled the RL problems to derive their optimal solutions and compare them with the solutions obtained through learning. For example, Pandana and Liu [10] model the problem of average throughput maximization per total consumed energy in a wireless sensor network as a Markov decision proccess (MDP), derive an optimal solution analytically, and compare it with one achieved by an RL algorithm. In another example Song and Jamalipour [11] model a vertical handoff decision problem as a semi-MDP and use Q-learning to solve this model directly. However, none of the stochastic models proposed in the wireless communications domain help to understand the dynamics of the RL algorithms themselves, as opposed to the learning problems they are applied to.

The purpose of this paper is to propose a simple Bayesian network model for analysing convergence properties of RL based DSA algorithms. This model is based on a minimum complexity inter-cell interference problem and provides a platform for theoretical evaluation of RL algorithms before they are applied to complex real-world DSA problems. It is briefly introduced by us in [15] for aiding the design of a novel RL based DSA scheme. However, this paper provides a significantly more detailed description and an empirical validation of this model. In previous work on combining Bayesian networks and RL, the purpose of Bayesian networks was to enhance the performance of RL algorithms by being used as a framework for reasoning under uncertainty, e.g. [12][13]. There is no evidence in the literature of using Bayesian networks as an analysis tool for RL algorithms.

The rest of the paper is organized as follows: Section II introduces the problem of DSA in cellular networks. In Section III we explain distributed RL and give details of the distributed Q-learning algorithm used in this paper. Section IV presents our novel Bayesian network model for joint policy transition analysis, and assesses its accuracy using a Monte Carlo simulation. The conclusions are given in Section V.

II. Dynamic Spectrum Access

In DSA networks all base stations are allowed opportunistic access to the whole spectrum pool available to the network. This approach has been shown to be significantly more flexible and efficient than fixed spectrum allocation methods [14]. The main limiting factor for network throughput and QoS performance in DSA networks is inter-cell interference, since all cells are allowed to use the same spectrum. This section presents a simple network model used for the analysis of inter-cell interference in this paper.

A. Simple Inter-Cell Interference Model

Figure 1 shows a small and analytically tractable DSA network model which can be related to most inter-cell interference problems in general. The aim of this model is to provide a small yet sufficiently complex DSA problem for theoretical analysis of RL algorithms which can then be extrapolated to larger and more realistic scenarios.

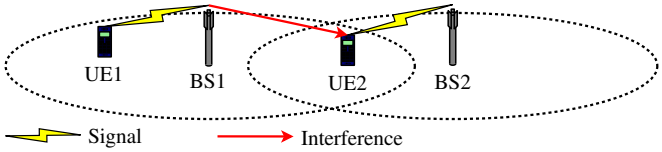


Figure 1. 2 BS 2 UE network model [15]

The network consists of 2 base stations (BSs) and 2 user equipments (UEs), each connected to its own BS. If one of the UEs is located within the interference range of the other BS, it suffers from harmful co-channel interference from it. The network is assumed to be allocated 2 subchannels, and the task of both BSs is to learn to use their own subchannel through distributed RL.

III. Distributed Reinforcement Learning

In distributed RL based DSA the task of every BS is to learn to prioritise among the available subchannels only through trial-and-error, with no frequency planning involved, and with no information exchange with other BSs, e.g. [7]. In this way, frequency reuse patterns emerge autonomously using distributed artificial intelligence with no requirement for any prior knowledge of a given environment.

A. Reinforcement Learning

RL 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 through trial-and error [3]. This desirability of an action is represented by a numerical value known as the Q-value - the expected cumulative reward for taking a particular action in a particular state, as shown in the equation below:

$$Q(s,a) = E\left[\sum_{t=0}^{T} \gamma^t r_t\right] \tag{1}$$

where $Q(s,a)$ is the Q-value of action a in state s, r_t is the numerical reward received t time steps after action a is taken in state s, T is the total number of time steps until the end of the learning process or episode, and $\gamma \in [0,1]$ is a discount factor.

The task of an RL algorithm is to estimate $Q(s,a)$ for every action in every state, which are then stored in an array known as the Q-table. In some cases where an environment does not have to be represented by states, only the action space and a 1-dimensional Q-table $Q(a)$ can be considered [16]. The job of an RL algorithm then becomes simpler; it aims to estimate an expected value of a single reward for each action available to the learning agent:

$$Q(a) = E[r_t] \tag{2}$$

This is also applicable to distributed Q-learning based DSA in cellular systems, e.g. [7][17].

B. Distributed Stateless Q-Learning

One of the most widely used RL algorithms is Q-learning [18]. In particular, a simple stateless variant of this algorithm, as formulated in [16], has been shown to be effective for several distributed DSA learning problems, e.g. [7][17].

Each BS maintains a Q-table $Q(a)$ such that every subchannel a has a Q-value associated with it. Upon each file arrival, the BS either assigns a subchannel to its transmission or blocks it if all subchannels are occupied. It decides which subchannel to assign based on the current Q-table and the greedy action selection strategy described by the following equation:

$$\hat{a} = \underset{a}{\operatorname{argmax}}(Q(a)) \tag{3}$$

where $\hat{a}$ is the subchannel chosen for assignment, and $Q(a)$ is the Q-value of subchannel a.

The values in the Q-tables are initialised to zero, so all BSs start learning with equal choice among all available subchannels. A Q-table is updated by a BS each time it attempts to assign a subchannel to a file transmission in the form of a positive or a negative reinforcement. The recursive update equation for stateless Q-learning, as defined in [16], is given below:

$$Q(a) \leftarrow (1-\alpha)Q(a) + \alpha r \tag{4}$$

where $Q(a)$ represents the Q-value of the subchannel a, r is the reward associated with the most recent trial and is determined by a 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.

The reward function, which is generally applicable to a wide range of RL problems and which has been successfully applied to DSA problems in the past [4][17], returns two values:

- $r = -1$ (negative reinforcement), if the file transmission failed due to excessive interference on the selected subchannel.
- $r = 1$ (positive reinforcement), if the file transmission is completed using the selected subchannel.

The choice of the learning rate values for this type of distributed Q-learning based DSA problems is thoroughly investigated in [17]. The best performance is achieved by using

the Win-or-Learn-Fast (WoLF) principle [19] where a lower value of α is used for successful trials, and a higher value of α is used for failed trials. In this way, the learning agents learn faster when "losing" and more slowly when "winning". For example, the WoLF learning rates used in this paper are the following:

$$\alpha = \begin{cases} 0.01 & r = 1 \\ 0.1 & r = -1 \end{cases} \quad (5)$$

where the value of α when $r = -1$ is significantly higher than when $r = 1$.

IV. Joint Policy Transition Analysis

Bayesian networks are a powerful tool for modelling conditional dependencies among stochastic variables [20]. This section explains our proposed Bayesian network based approach for analysing convergence of distributed RL algorithms by modelling joint policy transitions of the learning agents.

A. Bayesian Network Model

Figure 2 presents the Bayesian network which describes the behaviour of distributed RL introduced in Section III when applied to the simple DSA network model from Figure 1.

The variables used to denote the Bayesian network nodes are the following:
$\Pi_n \in \{Same, Diff\}$ - the joint policy of the BSs after n learning iterations. The policy of a BS is defined as its preferred subchannel $\pi_x \in \{1, 2\}$ and is derived from the Q-table based on (3). Π_n takes two values of interest - whether the policies of 2 BSs are the same or different ($\Pi_n = Diff$ is the learning objective).
$I_{UEx} \in \{Yes, No\}$ - whether or not $UE1$ or $UE2$ is located within the interference range of the adjacent BS during the current file arrival.
$TxOL \in \{Yes, No\}$ - whether file transmissions to $UE1$ and $UE2$ overlap in time during the current iteration.
$R_{UEx} \in \{S, F\}$ - whether a file transmission to $UE1$ or $UE2$ was successful (S), or whether it failed (F) due to interference. It is conditionally dependent on Π_n, I_{UEx} and $TxOL$.
$\Pi_{n+1} \in \{Same, Diff\}$ - the joint policy after the Q-learning updates described in (4), as a result of the outcome at the current iteration. It is conditionally dependent on Π_n, R_{UE1} and R_{UE2}.

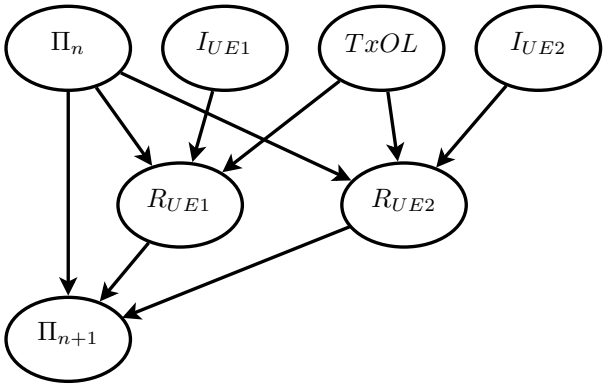


Figure 2. Bayesian network describing the behaviour of distributed Q-learning

Based on the conditional dependencies described above and depicted in Figure 2, the equation for calculating the joint probability distribution over all variables $P_{joint} = P(\Pi_{n+1}, \Pi_n, R_{UE1}, R_{UE2}, I_{UE1}, I_{UE2}, TxOL)$ is the following:

$$\begin{aligned} P_{joint} &= P(\Pi_{n+1}|\Pi_n, R_{UE1}, R_{UE2}) \\ &\times P(R_{UE1}|\Pi_n, I_{UE1}, TxOL)\ P(R_{UE2}|\Pi_n, I_{UE2}, TxOL) \\ &\times P(\Pi_n)\ P(I_{UE1})\ P(I_{UE2})\ P(TxOL) \end{aligned} \quad (6)$$

which consists of a number of prior probabilities of the form $P(X)$, and conditional probabilities of the form $P(X|Y_1...Y_n)$.

The prior probability distributions that appropriately describe the scenario depicted in Figure 1 are defined in Table I. Before any file arrivals at either BS, the Q-tables of both BSs are initialised to zero for both subchannels. Therefore, there is a 50% chance of the BSs choosing the same subchannel, since both of them choose a subchannel at random, i.e. $P(\Pi_0 = Same) = 0.5$. Furthermore, it is assumed that the interference range overlap of the BSs is such that there is a 40% chance of a UE being located in it, i.e. $P(I_{UEx} = Yes) = 0.4$. Finally, the offered traffic level is assumed to produce a 60% chance of transmissions to both UEs overlapping in time at any given learning iteration, thus potentially resulting in inter-cell interference: $P(TxOL = Yes) = 0.6$. The values chosen for $P(I_{UEx})$ and $P(TxOL)$ only affect the relative difficulty of the DSA problem. They can be changed without the loss of generality of the proposed probabilistic model.

The conditional probability distributions are defined in Table II. The values used for the $P(R_{UEx}|\Pi_n, I_{UEx}, TxOL)$ distribution state that a transmission to $UE1$ or $UE2$ will fail with a probability of 1 ($R_{UEx} = F$), if the given UE is within the interference range of the other BS ($I_{UEx} = Yes$), transmissions to both UEs overlap in time ($TxOL = Yes$) and both BSs have chosen the same subchannel ($\Pi_n = Same$). Whereas, if $\Pi_n = Diff$, $I_{UEx} = No$ or $TxOL = No$, then the transmission will be successful: $R_{UEx} = S$.

The $P(\Pi_{n+1}|\Pi_n, R_{UE1}, R_{UE2})$ table defines how the Q-learning policies of both BSs (Π_{n+1}) are likely to change,

TABLE I. Prior probability distributions

$P(\Pi_0)$		$P(I_{UEx})$		$P(TxOL)$	
Same	Diff	Yes	No	Yes	No
0.5	0.5	0.4	0.6	0.6	0.4

TABLE II. Conditional probability distributions

$P(R_{UEx}\|\Pi_n, I_{UEx}, TxOL)$								
S	0	1	1	1	1	1	1	1
F	1	0	0	0	0	0	0	0
	Same	Same	Same	Same	Diff	Diff	Diff	Diff
	Yes	Yes	No	No	Yes	Yes	No	No
	Yes	No	Yes	No	Yes	No	Yes	No
	$\Pi_n, I_{UEx}, TxOL$							

$P(\Pi_{n+1}\|\Pi_n, R_{UE1}, R_{UE2})$					
Same	1	Low	Low	$High$	0
Diff	0	$High$	$High$	Low	1
	Same	Same	Same	Same	Diff
	S, S	S, F	F, S	F, F	S, S
	Π_n, R_{UE1}, R_{UE2}				

given their current joint policy Π_n, and the result of transmissions to both UEs (R_{UE1} and R_{UE2}). Both BSs are running a stateless Q-learning algorithm introduced in Subsection III-B. Firstly, if the transmissions to both UEs are successful ($R_{UE1} = R_{UE2} = S$), then both BSs will reward their respective subchannels and maintain the same policies regardless whether they are the same or different ($\Pi_{n+1} = \Pi_n$). Secondly, if Π_n is $Same$ and only a transmission to one of the UEs failed ($\{S, F\}$ or $\{F, S\}$), this UE is more likely to change its policy due to the WoLF learning rate used in its Q-learning algorithm, described by (5). Therefore, there is a relatively high probability of the policies being different at the next iteration: $P(\Pi_{n+1} = Diff) = High$. If transmissions to both UEs fail ($\{F, F\}$), both BSs are likely to change their policies, thus making $\Pi_{n+1} = Same$ a more likely outcome: $P(\Pi_{n+1} = Same) = High$. The remaining 3 combinations of Π_n, R_{UE1} and R_{UE2} values are not considered, since they can never occur according to the $P(R_{UEx}|\Pi_n, I_{UEx}, TxOL)$ conditional probability distribution. Regardless of the values used for these combinations in the $P(\Pi_{n+1}|\Pi_n, R_{UE1}, R_{UE2})$ table, they will be multiplied by zero during the calculation of the joint probability distribution defined in (6).

The aim of the Bayesian network model described above is to establish the marginal likelihood of the joint Q-learning policy at the next iteration $P(\Pi_{n+1})$ by taking a sum over all other variables in P_{joint} as follows:

$$P(\Pi_{n+1}) = \sum_{\Pi_n} \sum_{R_{UE1}} \sum_{R_{UE2}} \sum_{I_{UE1}} \sum_{I_{UE2}} \sum_{TxOL} P_{joint} \quad (7)$$

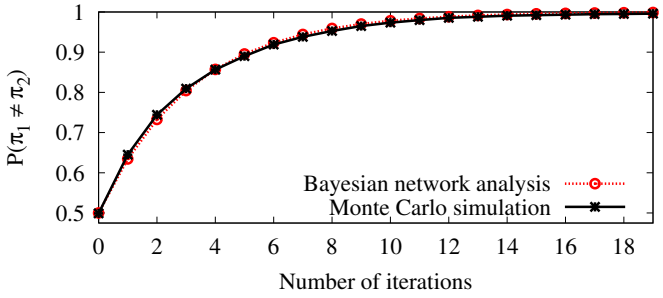
The resulting distribution can then be substituted as the prior for the next learning iteration: $P(\Pi_n) \leftarrow P(\Pi_{n+1})$. This enables iterative evaluation of the Bayesian network model which shows how the probability of transmission failure $P(R_{UEx})$ and the probability of BSs using different subchannels $P(\Pi_n)$ change over time, as the learning process progresses. The $P(R_{UEx})$ distribution can be obtained using the principle of marginalization as follows:

$$P(R_{UE1/2}) = \sum_{\Pi_{n+1}} \sum_{\Pi_n} \sum_{R_{UE2/1}} \sum_{I_{UE1}} \sum_{I_{UE2}} \sum_{TxOL} P_{joint} \quad (8)$$

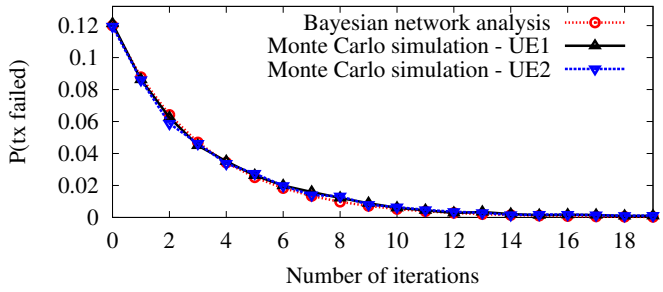
This probabilistic analysis is only valid for the 2 BS 2 UE network model depicted in Figure 1, and is not designed to be scalable to larger and more realistic networks. The purpose of this model is to enable theoretical analysis of the relative behaviour of RL algorithms using a simple and tractable problem. An additional, useful approach to evaluating such algorithms, that is outside of the scope of this paper, is performing realistic large scale simulations and assessing similarities between the simulation results and the theoretical predictions.

B. Probabilistic Analysis vs Monte Carlo Simulation

Figure 3 shows the expected convergence behaviour of distributed Q-learning analytically derived through iterative evaluation of the Bayesian network model presented in this section. The values for $High$ and Low in the conditional probability distributions in Table II are assumed to be $\{0.9, 0.1\}$. However, changing these values to other interpretations of "high" and "low" probabilities had negligible effect on the convergence



(a) Probability of BSs having different policies



(b) Probability of a UE being blocked or interrupted

Figure 3. Convergence of distributed Q-learning using Bayesian network analysis and a Monte Carlo simulation

patterns shown in Figure 3. The analytical results are compared with a Monte Carlo simulation, where the Q-learning algorithm from Subsection III-B is applied to the scenario described in Subsection II-A. At every transmission arrival the simulation experiment drew the inter-cell interference parameters from the prior probabilities defined in Table I. The probabilities plotted for every learning iteration were obtained by averaging over 10,000 independent runs.

The comparison of the convergence behaviour predicted by the Bayesian network model and that achieved by the Monte Carlo simulation demonstrates remarkable accuracy of the joint policy transition analysis tool proposed in this paper. Therefore, it is a valid and effective approach for stochastic modelling of RL based DSA algorithms. It can be used for designing and analysing the convergence of more sophisticated RL algorithms by adding nodes and edges to the Bayesian network from Figure 2. The added nodes and edges would represent additional functionality and conditional dependencies introduced by the new schemes. This approach would clearly demonstrate in what ways other schemes designed in future using this method extend the basic distributed RL approach depicted in Figure 2. For example, a novel heuristically accelerated RL scheme was successfully analysed using an extension to the Bayesian network model proposed in this paper, and was then also evaluated using large scale network simulations in [15].

C. Absorbing Markov Chain Formulation

Figure 4 shows an alternative formulation of the convergence properties of distributed Q-learning derived from the Bayesian network model introduced in Subsection IV-A. It is a Markov chain describing the probabilities of transitions be-

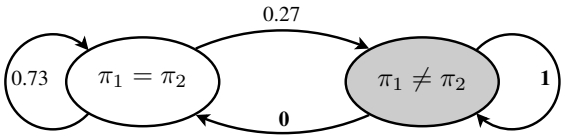


Figure 4. An absorbing Markov chain describing the transitions between two states of the joint policy derived from Bayesian network model of the 2 BS 2 UE cellular network

tween two different states of the joint policy - $Same(\pi_1 = \pi_2)$ and $Diff(\pi_1 \neq \pi_2)$. The transition probabilities are taken from the $P(\Pi_{n+1}|\Pi_n)$ distribution, which in turn is calculated using the following definition of conditional probability:

$$P(\Pi_{n+1}|\Pi_n) = \frac{P(\Pi_{n+1}, \Pi_n)}{P(\Pi_n)} \tag{9}$$

where $P(\Pi_{n+1}, \Pi_n)$ is obtained by marginalizing all other variables from the overall joint distribution as follows:

$$P(\Pi_{n+1}, \Pi_n) = \sum_{R_{UE1}} \sum_{R_{UE2}} \sum_{I_{UE1}} \sum_{I_{UE2}} \sum_{TxOL} P_{joint} \tag{10}$$

Firstly, the Markov chain in Figure 4 shows that "$\pi_1 \neq \pi_2$" is an absorbing state, i.e. a state that cannot be left, since the probability of going from "$\pi_1 \neq \pi_2$" to "$\pi_1 = \pi_2$" is zero. Therefore, this is an absorbing Markov chain which formally demonstrates that the RL algorithm is guaranteed to converge on the desired absorbing state "$\pi_1 \neq \pi_2$". The speed of convergence is controlled by the probability of going from "$\pi_1 = \pi_2$" to "$\pi_1 \neq \pi_2$", which in this case is 0.27. The objective of future more advanced RL algorithms designed using the method proposed in this paper is to increase this transition probability to speed up the convergence, whilst preserving the absorbing state where "$\pi_1 \neq \pi_2$".

V. Conclusion

The Bayesian network based joint policy transition analysis methodology proposed in this paper is able to provide a simple and accurate probabilistic model of distributed RL algorithms applied to a minimum complexity DSA problem. A Monte Carlo simulation of a distributed Q-learning based DSA algorithm shows that the proposed approach demonstrates remarkably accurate prediction of the convergence behaviour of such algorithms. Furthermore, their behaviour can also be expressed in the form of an absorbing Markov chain, derived from the novel Bayesian network model. This representation enables further theoretical analysis of convergence properties of RL based DSA algorithms. Finally, the main benefit of the analysis tool presented in this paper is that it enables the design and theoretical evaluation of novel RL based DSA algorithms by extending the proposed Bayesian network model, that describes a standard distributed Q-learning scheme.

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] H. Sun, A. Nallanathan, C.-X. Wang, and Y. Chen, "Wideband spectrum sensing for cognitive radio networks: a survey," *Wireless Communications, IEEE*, vol. 20, pp. 74–81, 2013.

[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] T. Jiang, D. Grace, and P. D. Mitchell, "Efficient exploration in reinforcement learning-based cognitive radio spectrum sharing," *Communications, IET*, vol. 5, pp. 1309–1317, 2011.

[5] M. Bennis, S. Perlaza, P. Blasco, Z. Han, and H. Poor, "Self-organization in small cell networks: A reinforcement learning approach," *Wireless Communications, IEEE Transactions on*, vol. 12, pp. 3202–3212, 2013.

[6] X. Chen, Z. Zhao, and H. Zhang, "Stochastic power adaptation with multiagent reinforcement learning for cognitive wireless mesh networks," *Mobile Computing, IEEE Transactions on*, vol. 12, pp. 2155–2166, 2013.

[7] 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.

[8] R. Valcarce, T. Rasheed, K. Gomez, S. Kandeepan, L. Reynaud, R. Hermenier, A. Munari, M. Mohorcic, M. Smolnikar, and I. Bucaille, "Airborne base stations for emergency and temporary events," in *International Conference on Personal Satellite Services*, 2013.

[9] L. Reynaud, S. Allsopp, P. Charpentier, H. Cao, D. Grace, R. Hermenier, A. Hrovat, G. Hughes, C. Ioan, T. Javornik, A. Munari, M. Vidal, J. Strother, R. Valcarce, and S. Zaharia, "FP7-ICT-2011-8-318632-ABSOLUTE/D2.1 Use cases definition and scenarios description," 2014.

[10] C. Pandana and K. Liu, "Near-optimal reinforcement learning framework for energy-aware sensor communications," *Selected Areas in Communications, IEEE Journal on*, vol. 23, pp. 788–797, 2005.

[11] Q. Song and A. Jamalipour, "A quality of service negotiation-based vertical handoff decision scheme in heterogeneous wireless systems," *European Journal of Operational Research*, vol. 191, pp. 1059 – 1074, 2008.

[12] P. Poupart, N. Vlassis, J. Hoey, and K. Regan, "An analytic solution to discrete bayesian reinforcement learning," in *International Conference on Machine Learning (ICML)*, 2006.

[13] M. Kearns and D. Koller, "Efficient reinforcement learning in factored MDPs," in *International Joint Conference on Artificial Intelligence (IJCAI) - Volume 2*, 1999.

[14] Y. Xing, R. Chandramouli, S. Mangold, and S. N, "Dynamic spectrum access in open spectrum wireless networks," *Selected Areas in Communications, IEEE Journal on*, vol. 24, pp. 626–637, 2006.

[15] N. Morozs, T. Clarke, and D. Grace, "Distributed heuristically accelerated Q-learning for robust cognitive spectrum management in LTE cellular systems," submitted to *Mobile Computing, IEEE Transactions on*, 2015.

[16] 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.

[17] N. Morozs, T. Clarke, D. Grace, and Q. Zhao, "Distributed Q-learning based dynamic spectrum management in cognitive cellular systems: Choosing the right learning rate," in *IEEE International Symposium on Computers and Communications (ISCC)*, 2014.

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

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

[20] T. Nielsen and F. Jensen, *Bayesian networks and decision graphs*. Springer, 2009.