Reinforcement Learning for Dynamic Spectrum Management in WCDMA

Nemanja Vučević, Jordi Pérez-Romero, Oriol Sallent, and Ramon Agustí

Abstract—**Low use of licensed spectrum imposes a need for the advanced spectrum management for wise spectrum usage with the release of unneeded frequency bands for the secondary markets and opportunistic access. In this paper we present the possibilities to apply reinforcement learning in WCDMA to enable the autonomous decision on spectrum repartition among cells and release of frequency bands for possible secondary usage. The proposed solution increases spectrum efficiency while ensuring maximum outage probability constraints in WCDMA uplink. We give two possible approaches to implement reinforcement learning in this problem area and compare their behavior. Simulations demonstrate the capability of two methods to successfully achieve desired goals.**

Keywords — **dynamic spectrum management, reinforcement learning, WCDMA**

I. Introduction

THE constantly increasing demand for throughput and quality in telecommunication systems leads to continuous search for wise resource management, especially present in radio communications that rely on a scarce number of frequency bands with obvious limitations in capacity. Research in actual frequency usage of the existing communications systems demonstrates however that the main limiting factor reducing spectrum efficiency is the low use of the licensed frequency bands, studied by Federal Communications Commission (FCC) [1]. Triggered by this fact, the envisaged evolution of the existing telecommunications systems moves towards spectrum sharing concepts, in which secondary access can be allowed to those licensed spectrum bands that are temporarily or spatially unused by the licensee (i.e. the primary user), and provided that no harmful interference is generated to primary users. This concept will be supported by current developments in the field of cognitive radio networks [2].

The future radio access is to be flexible, where a wireless user will fulfill its requirements through the most appropriate frequency band, technology or operator at that moment in a certain location. With that aim, the frequency bands are expected to be dynamically distributed among pretenders on its usage, supposing collaboration and opportunistic access. The potential of the primary systems, the users of the licensed bands, should be used efficiently, thus to enable the delivery of the unused frequency bands to the secondary markets [2].

One of the disciplines that emerge in the above conditions is the Dynamic Spectrum Management (DSM). DSM gives the answer to previously described problems and decides on intelligent repartition of frequency bands. More bandwidth liberation gives more potential to cognitive secondary users [3], so that better spectrum utilisation can be achieved.

In practice, current 3G providers often dispose with more than one frequency band. WCDMA (Wideband Code Division Multiple Access) systems generally apply a frequency reuse of one. However, the time space migrations of users cause traffic variations that make traditional fixed spectrum allocation patterns inadequate. Thus, dynamic assignment/releasing of frequency bands in such systems may increase spectral efficiency.

In this paper, we propose the use of reinforcement learning (RL) [4] to evaluate and adjust spectrum assignment to cells in WCDMA with the aim to increase the spectral efficiency of such a system. We build an autonomous solution, where one primary system decides on releasing of some spectrum based on self-evaluation – liberating bandwidth while preserving its own users' quality satisfaction.

The introduction of cognitive processes by means of reinforcement learning is envisaged as a prosperous tool to achieve reconfigurable networks that extend the scope from radio access to the entire networks [5]. The learning adaptable systems should increase the desired benefit of communication resources. In this work we apply RL algorithms, give a detailed description and demonstrate how learning can be used to archive predefined spectral efficiency goals. This rather new problem area has been studied in recent works that apply genetic algorithms [6] and simulated annealing with coupling matrices [7].

The rest of this paper is organized as follows. In section II we expose the main goals of the proposed solution. We give a detailed description of the RL based DSM proposal in section III. This section is followed by section IV containing simulation results. Section V concludes this paper.

II. Performance Objectives

The final goal is to satisfy the primary system requirements with a minimum number of carriers, so that those carriers that are not used can be released to e.g. a secondary market. Satisfaction of the primary system means preservation of the quality level for its users.

In this paper we will assume that user's quality conditions are considered to be satisfied in WCDMA when a user is not in outage, meaning that the bit energy over noise spectral density (*Eb/No*) is higher than or equal to a certain tar-

This work has been supported by the Spanish Research Council and FEDER funds under COGNOS grant (ref. TEC 2007-60985); and MEC-FPU Scholarship (AP2005-3739).

N. Vučević, J. Pérez-Romero, O. Sallent, R. Agustí are with the Signal Theory and Communications Department, Universitat Politècnica de Catalunya, Barcelona, Spain
e-mails: {vucevic, jorperez, sallent, ramon}@tsc.upc.edu

get that will depend on a specific service. For the WCDMA uplink with ideal power control this can be formulated as:

$$\frac{\frac{w}{R_{bi}}\frac{P_{Ti}}{L_{i,j}}}{P_{TOTj}-\frac{P_{Ti}}{L_{i,j}}} \ge \left(\frac{E_b}{N_0}\right)_{i,TARGET} \tag{1}$$

where w is the chip rate, R_{bi} is user's i bit-rate, P_{Ti} is the transmitted power of the mobile i and $L_{i,j}$ is the path loss for mobile i to cell j. The maximum transmit power of a mobile (P_{Ti}^{MAX}) is the mobile's limiting factor that determines the outage. $P_{TOT,j}$ is the total received power by cell j, from all the users (k) of that cell, intercell interference (χ_j) and background noise power in uplink (P_N):

$$P_{TOT,j}=\sum_k \frac{P_{Tk}}{L_{k,j}}+\chi_j+P_N \tag{2}$$

In the sequel the Key Performance Indicators (KPI) used in the paper are explained: Spectral Efficiency and Useful Released Surface (URS) factor. Note here that the URS factor is not the direct objective of our algorithm, but is used further in the paper to compare the proposed solutions.

A. *Spectral efficiency*

The primary objective of the proposed dynamic spectrum assignment algorithm is to improve the spectral efficiency in WCDMA uplink. When M different services (of data rate R_b^m) are provided in a cell j, spectral efficiency (θ_j) at cell j is easily calculated as:

$$\theta_j=\frac{1}{g_j\cdot BW}\sum_{m=1}^{M} n_j^m\cdot R_b^m\cdot\left(1-\Theta_{\mathrm{r},j}^m\right) \tag{3}$$

Here n_j^k is the number of users using service m in cell j, and $\Theta_{r,j}^{\;m}$ is the outage probability of the users using service m in cell j. The number of nominal bandwidths BW in use in cell j is denoted as g_j.

B. *URS factor*

Useful Released Surface (URS), denoted as U, is a KPI defined in [8] to reflect the profitability of the released spectrum for the potential secondary usage, in accordance with the size of geographical areas where the spectrum is being released.

URS is defined as:

$$U=\sum_{f=1}^{F} BW\cdot\sum_{c=1}^{C(f)} s_c^f\cdot\omega_c^f=\sum_{f=1}^{F} BW\cdot\sum_{c=1}^{C(f)}\frac{\left(s_c^f\right)^2}{S} \tag{4}$$

where F is the number of frequency carriers, S is the overall surface under study, and BW is the nominal bandwidth of a WCDMA carrier. In turn, *C(f)* is the set of non-contiguous areas where the carrier f could be used by a secondary network, s_c^f is the surface of the area c in relation with carrier f and ω_c^f is the weight given to this area depending on the expected number of secondary users in this area. We calculate weights in expression (4) as if the secondary users are equally probable in the entire area.

III. Reinforcement learning for DSM

This paper gives a model that uses reinforcement learning to find the appropriate frequency-to-cell assignment in order to ensure the primary user requirements with maximum spectral efficiency, trying to release spectrum for secondary markets.

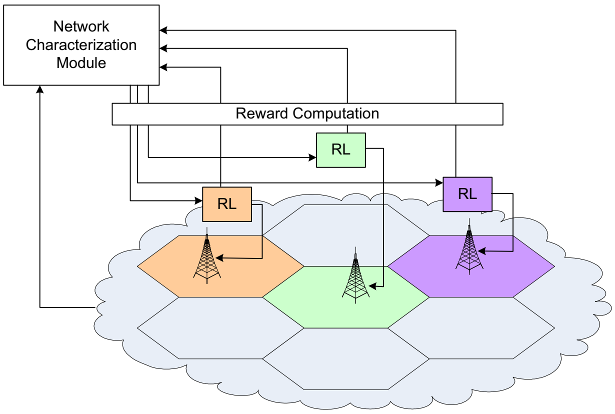

Fig. 1: DSM with RL.

The proposed mechanism is presented in Fig. 1. It consists of reinforcement learning (RL) units in each cell and a network characterization module. The learning mechanism collects the system state from the environment and passes it to the network characterization module, which mimics the network behavior based on the measurements of different mobiles. The learning process then executes on this module and for each cell decides whether a certain frequency band will or will not be used.

A. *Reinforcement Learning*

Reinforcement learning is an area of machine learning that through interaction with an environment considers determination of actions an agent should take with the objective of maximizing some long term goals. In this study we focus on actor-critic RL algorithms [4], as these methods require minimal computation in order to select actions and can learn optimal probabilities of selecting various actions.

We use softmax method with the Boltzmann distribution function to generate actions. In our solution each cell is assigned a RL agent built of one or more RL units. Each RL unit will select an action to make a transition from one state to another. The mapping of the states into frequency allocation will be explained in subsection III-*C*. In our case, in each time instant t the probability to select any action a (a_t=a) that makes a transition to a state s_a (s_t=s_a) at RL unit x is:

$$\pi_{j,x,t}(a,s_a)=\frac{e^{p_{j,x}(a,s_a)/\tau}}{\sum_{b=1}^{A} e^{p_{j,x}(b,s_b)/\tau}} \tag{5}$$

Here τ is a positive, so called temperature parameter and A is the set of possible actions. The learning process of the algorithm is achieved through the update of the parameter $p(a_t,s_t)$. In this study we use REINFORCE [9] learning techniques to make this update:

$$p_{j,x}(a_t,s_t)\leftarrow p_{j,x}(a_t,s_t)+\beta\cdot(r_{j,t}-\hat{r}_{j,t})\cdot(1-\pi_{j,x,t}(a_t,s_t)) \tag{6}$$

where β is a positive step-size parameter, $r_{j,t}$ is the reward estimate at cell j in time instant t.

After this update is provided and new action selection probabilities (5) have been calculated, the reward accumulate is updated for the following iterations:

$$\hat{r}_{j,t}\leftarrow\hat{r}_{j,t}+\gamma\cdot(r_{j,t}-\hat{r}_{j,t}) \tag{7}$$

where γ is another positive step-size parameter.

Note that the RL algorithm tends to achieve maximization of the reward function r through time. Thus the further study supposes an appropriate definition of desired goals in the reward.

B. *Reward function*

The reward function is the evaluation of the actions taken by RL agents through the algorithm's evolution. In order to have the frequency band assignment according to the desired system behavior, the predefined goals have to be projected in the reward function. For each cell j, we define the reward in instance t as:

$$r_j = \begin{cases} \theta_j & \Theta_{r,j} < \Theta_r^{MAX} \\ 0 & \Theta_{r,j} > \Theta_r^{MAX} \end{cases} \qquad (8)$$

The primary objective of the algorithm – spectral efficiency θ_j at each cell is directly introduced as a reward function. However, as we set the outage higher than Θ_r^{MAX} to determine the limit of the system capacity, the reward equals zero when such a case occurs. This reward function ensures that spectral efficiency is maximized while primary user requirements are kept.

C. *Frequency allocation*

The application of the RL to a system supposes offline learning of the algorithm over a network characterization module as previously explained. Each action will lead the system to one state. During a learning process, the actions are selected following output probabilities from (5) so the resulting states determine frequency assignment. After the RL agent has converged, the system output may be applied to a real system.

In this study we compare two possible approaches to decide on frequency allocation:

- Method 1: Each cell has an agent consisting of one RL unit whose number of states is equal to the maximum number of frequencies. Each state defines the number of consecutive frequencies that are assigned to that cell (e.g. if state is N this means that frequencies 1, 2, ..., N will be assigned to that cell). The frequency assignment is always following the same order (Fig. 2-a).
- Method 2: Each cell has an agent built of several RL units. Each RL unit is assigned to one frequency carrier and has only two states to decide on activation/deactivation of that specific frequency carrier in the cell. This decision is independent for each frequency carrier in a cell, thus, the frequency assignment does not have to be consecutive (Fig. 2-b).

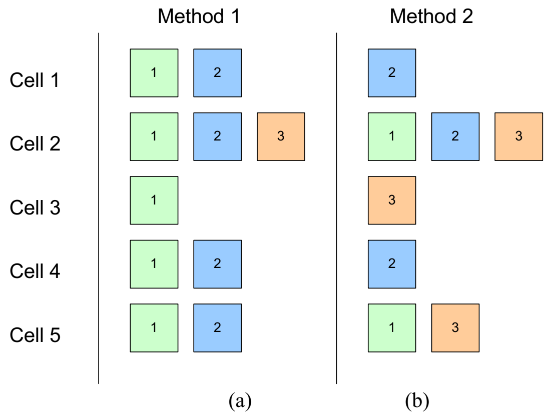

Fig. 2: Example of allocation of 3 frequency bands in 5 cells with (a) Method 1 and (b) Method 2.

IV. Simulations

For the evaluation purposes of the RL possibilities to learn and configure the frequency allocation, we build a model of 19 cells with four of them presenting a traffic hotspot. The model is illustrated in Fig. 3. Users are homogeneously distributed in the scenario, except in cells CT1 which contain four times more users than the rest of cells, denoted as CT2 cells. We consider a case where a WCDMA system may use up to 3 nominal UMTS frequency bands. The system limits are defined with Θ_r^{MAX}=0.05 set as the maximum allowed outage. Additional simulation parameters may be seen in Table 1.

We compare the proposed solution to the commonly applied static frequency assignment in practice, where we measure system performances for the case where all the cells have one (ST-I), two (ST-II) or three (ST-III) frequency bands in use.

In both the static and the RL case we suppose that in those cells having more than one carrier, the load in all these carriers is the same (e.g. in a cell with 2 carriers half of the users will be connected to the first one and half to the second one).

The number of users in the system ranges from 400-1800 users. All the results have been averaged over 100 snapshots. The average results are presented separately for the more loaded cells (CT 1) and the less loaded cells (CT 2).

The values of reinforcement learning parameters have been assigned based on experience gained on different trials not shown here for the sake of brevity. These values are γ=0.2, β=0.1, whereas τ=0.005 for method 1 and τ=0.01 for method 2.

A. *Results*

In Fig. 4-a and Fig. 4-b the spectral efficiency obtained with the two RL methods (denoted as M1 and M2) is presented for cell types CT1 and CT2, respectively. For comparison purposes, the spectral efficiency in the static methods ST-I, ST-II and ST-III is also presented. In addition, Fig. 4-c shows the average number of assigned frequency

TABLE 1
SIMULATION PARAMETERS

Parameter	Value
Cell Radius (outer hexagon circle)	0.5 km
Data Rate (*Rb*)	12.2 kbps
Chip rate (*w*)	3.84 Mchips/s
Nominal bandwidth (*BW*)	5 MHz
Eb/No target	6 dB
Maximal Mobile Transmission Power (P_{Ti}^{MAX})	21 dBm
Background noise in uplink (P_N)	-104 dBm

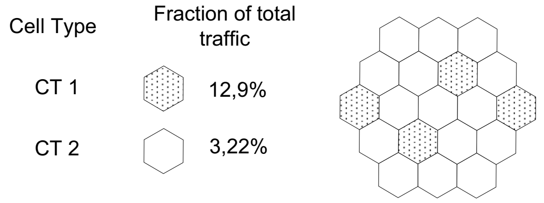

Fig. 3: Simulation model.

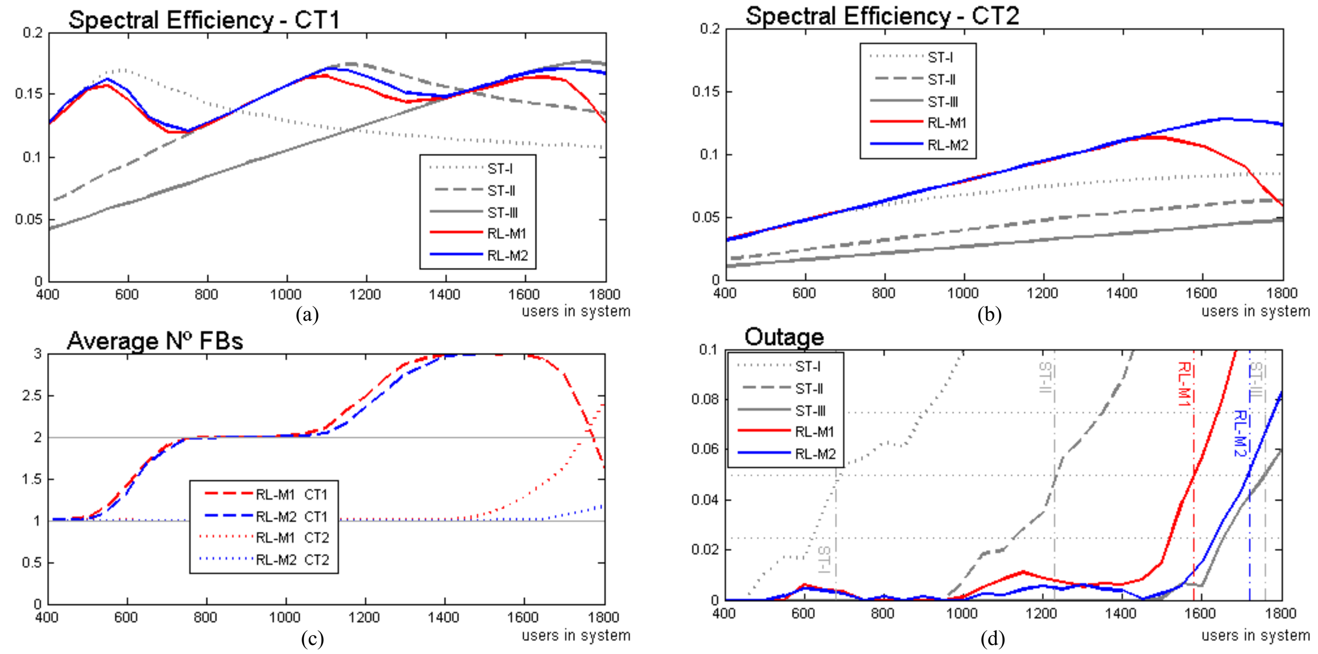

Fig. 4: (a) Spectral efficiency in CT1, (b) Spectral efficiency in CT2, (c) average number of assigned frequency bands and (d) maximum system outage, for the reinforcement learning Method 1 and Method 2 and static cases (ST-I, ST-II, ST-III).

bands with all the approaches. Results show that both RL algorithms manage to increase spectral efficiency by assigning fewer frequency bands than the static cases and still preserve a satisfactory system quality. As expected, the main contribution in spectral efficiency is in the cells that have a lower load (CT2). Nevertheless, higher loaded cells (CT1) also manage to achieve high performances always around the value of the best static case (within its capacity limits).

The maximum system outage values are presented in Fig. 4-d, which also includes the capacity limits (marked with vertical dash-dotted lines), defined as the maximum number of users that ensures the outage probability to be below the limit of 5%. Notice that ST-III case exhibits the highest system capacity limit (around 1760 users in the scenario) but RL-method 2 is able to reach up to 97.8% of this limit while exhibiting a better spectral efficiency. Lastly, method 1 reaches only 89.8% of the capacity limit.

Finally, Fig. 5 depicts the normalized value of URS factor. It can be observed that method 1 provides better results than method 2 in terms of URS factor, particularly for low and medium loads. However, for very high loads, close to the capacity limit, better performance in terms of capacity and outage provided by method 2 are also translated into a better URS.

V. Conclusion

This paper presents a dynamic spectrum assignment strategy based on reinforcement learning to decide the carrier to cell assignment in WCDMA. The objective is to satisfy users' requirements with the minimum number of frequencies (correspondingly maximizing spectral efficiency), so that the unused carriers can be released for e.g. secondary usage. The results show that the evaluation of the system performances may lead to a successful configuration of the spectrum. The applied algorithms have given satisfactory results with advantages under different load conditions.

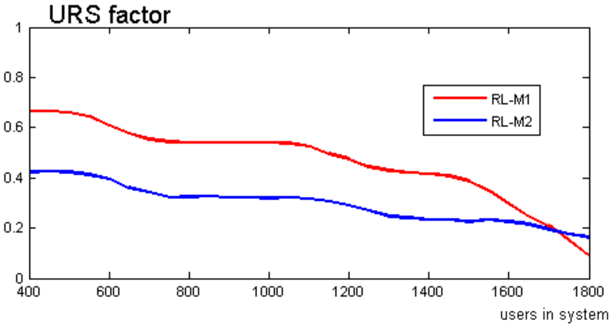

Fig. 5: URS factor values.

References

[1] FCC: "Report of the Spectrum Efficiency Working Group", Nov. 2002

[2] I. F. Akyildiz, W.-Y. Lee, M. C. Vuran, S. Mohanty, "Next Generation/Dynamic Spectrum Access/Cognitive Radio Wireless Networks: A Survey", *Computer Networks*, vol. 50, issue 13, pp. 2127-2159, Sept. 2006

[3] P.N. Anggraeni, N.H. Mahmood, J. Berthod, N. Chaussonniere, L. My, H. Yomo, "Dynamic Channel Selection for Cognitive Radios with Heterogeneous Primary Bands", *Wireless Personal Communications*, Vol. 45, num. 3, pp. 369-384, May 2008

[4] R. S. Sutton, A.G. Barto, *Reinforcement Learning: An Introduction*, A Bradford Book, MIT Press, Cambridge, MA 1998

[5] R.W. Thomas, D.H. Friend, L.A. DaSilva, A.B. MacKenzie, "Cognitive networks: adaptation and learning to achieve end-to-end performance objectives", *IEEE Communications Magazine*, vol. 44, issue 12, pp. 51-57, Dec. 2006

[6] D. Thilakawardana, K. Moessner, "Enhancing Spectrum Productivity through Cognitive Radios facilitating Cell-by-Cell Dynamic Spectrum Allocation", in proceedings of Software Defined Radio Technical Conference (SDR'07), Nov. 2007

[7] J. Nasreddine, J. Pérez-Romero, O. Sallent, R. Agusti, "Simulated Annealing-Based Advanced Spectrum Management Methodology for WCDMA Systems," in proceedings of IEEE International Conference on Communications (ICC), May 2008

[8] J. Nasreddine, J. Pérez-Romero, O. Sallent, R. Agusti, "Primary Spectrum Management Solution Facilitating Secondary Usage Exploitation", in proceedings of ICT Mobile Summit, June 2008

[9] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning", *Machine Learning*, vol. 8, pp 229-256, 1992