Cell Selection Using Distributed Q-Learning in Heterogeneous Networks

Toshihito Kudo and Tomoaki Ohtsuki
Keio University
3-14-1, Hiyoshi, Kohokuku, Yokohama, 223-8522, Japan
Email: kudo@ohtsuki.ics.keio.ac.jp, ohtsuki@ics.keio.ac.jp

***Abstract*—Cell selection with cell range expansion (CRE) that is a technique to expand a pico cell range virtually by adding a bias value to the pico received power, instead of increasing transmit power of the pico base station (PBS), can make coverage, cell-edge throughput, and overall network throughput improved. Many studies about CRE have used a common bias value among all user equipments (UEs), while the optimal bias values that minimize the number of UE outages vary from one UE to another. The optimal bias value that minimizes the number of UE outages depends on several factors such as the dividing ratio of radio resources between macro base stations (MBSs) and PBSs, it is given only by the trial and error method. In this paper, we propose a scheme to select a cell by using Q-learning algorithm where each UE learns which cell to select to minimize the number of UE outages from its past experience independently. Simulation results show that, compared to the practical common bias value setting, the proposed scheme reduces the number of UE outages and improves network throughput in the most cases. Moreover, instead of the degradation of the performances, it also solves the storage problem of our previous work.**

I. Introduction

Heterogeneous networks (HetNets) whereby low power base stations (BSs) are deployed within the macro cell, has recently received significant attention because of the rapid increase of the traffic amount [1]. HetNets are discussed as one of the proposed solutions as part of the Long Term Evolution-Advanced (LTE-Advanced) by the third generation partnership project (3GPP) [2]. Among the low power BSs, for instance, pico BS (PBS), femto BS, and relay BS, PBSs are mostly considered, because they usually have the same backhaul as MBS and are placed near the hotspot where the traffic amount is high [3]. If pico cells cover the hotspot areas, PBSs can serve UEs within those areas and improve the throughput of the downlink (DL) channel. However, because the hotspot's location and amount of traffic change dynamically, PBSs cannot always cover that area and UEs may have to access the MBSs even if the PBS may be closer to them.

In [1], the authors discuss cell range expansion (CRE), which is a technique that adds a bias value to pico received power during the handover as if pico cell range is expanded, and many works focus on this topic [1], [3]–[5]. CRE can make more UEs to access PBSs even if the received power of MBS's signals is larger than the that of PBS's signals. However, those UEs that access PBSs whose the received power from PBS is smaller than that from MBSs, referred to as expanded region (ER) UEs [1], are affected by a large amount of interference from MBSs. Therefore, to eliminate the interference, inter-cell interference coordination (ICIC) may be needed, and many papers have worked on optimal configuration of ICIC.

Many papers about CRE apply ICIC realized by dividing the radio resource: between two categories of MBS and PBS, ICIC is usually realized by stopping MBS's transmission on some radio resources [5]. In 3rd-generation partnership project long term evolution (3GPP-LTE) system, Resource blocks (RBs) introduced as blocks of subcarriers [6] can also realize ICIC by dividing them. However, in terms of the scalability, to decide the bias value or the connected cell should be focused on rather than to setting the dividing ratio of radio resources because using different spectrum between MBS and PBS is recently discussed in [7]. There are no differences between to decide the bias value and the connected cell under applying CRE because both allow UEs to send their access requests to PBS even if the powers of the MBS signals are larger than those of the PBS ones.

In general, UEs are set to use the same, fixed, bias value [1], [3]–[5]. The appropriate bias values of each UE depend on the ratio of RBs of each BS and the location of UEs and BSs that is hard to get [4]. From the aforementioned reason, the optimal bias values are obtained only by the trial and error method.

Instead of the trial and error method, we propose to use Q-learning [8], a reinforcement learning (RL) technique, to determine the connected cell. Using RL in a radio communication system is becoming popular [9]–[11], because the recent complicated situations that have different radio systems in the same area make it harder to adjust parameters. Q-learning has been applied to many other areas such as: cognitive radio [9] and self-optimization of capacity and coverage scheme in HetNets [10]. Moreover, it has also been applied to set transmit powers, radio resources, and a bias value of CRE [11]. However, this work optimizes each PBS bias value although the bias value should have been defined for each UE [4]. Though our previous work [12] also applied it to set a bias value of CRE independently-learned by each UE, there was a storage problem of the Q-table because all UEs have to store the Q-values of all bias values. Because of this, it is largely affected by the curse of dimensionality, and has little scalability to add other types of BSs.

In this paper, each UE learns which cell to connect to min-

imize the number of UE outages individually by Q-learning. Simulation results show that the proposed scheme can make the size of a Q-table smaller than our previous work. Moreover, compared to the practical common bias value setting, the proposed scheme reduces the number of UE outages and improves network throughput in the most cases.

II. Heterogeneous Network

Though HetNets encompass many types of BSs, out of concern for simplicity, this work shall be limited to the case where only two types of BSs, namely MBS and PBS, as this is also the case in the majority of the related works. PBSs are typically deployed within macro cells for capacity enhancement and coverage extension. Moreover, they usually have the same back-haul and access features as MBSs [1].

Although PBSs are deployed within macro cells to avoid hotspot UEs from accessing MBSs, the limited coverages of PBSs still cause many UEs to be outage with the reference signal received power (RSRP) based cell selection that has always allowed UEs to connect the BSs that serve the strongest received power of the reference signal (pilot signal). Path loss based cell selection schemes have also been discussed to balance loads [1]. Since it allows UEs to connect to the BSs that have the smallest path loss, more UEs tend to connect to the PBSs. However, in terms of load balancing, UEs should adapt the connected cell for the surrounding environment, which is realized by CRE [3] explained in the subsequent paragraph.

A. Cell Range Expansion

CRE is usually applied with RSRP based cell selection [1]. A bias value is added to the pico received signal, and more UEs can connect to PBSs, which is as if pico cell range is expanded, that is, UEs connect to:

$$\text{MBS, } \textbf{when } (p_{\text{M}})_{\text{dB}} > (p_{\text{P}})_{\text{dB}} + (\Delta bias)_{\text{dB}} \quad (1)$$

$$\text{PBS, } \textbf{when } (p_{\text{M}})_{\text{dB}} < (p_{\text{P}})_{\text{dB}} + (\Delta bias)_{\text{dB}} \quad (2)$$

where $(p_{\text{M}})_{\text{dB}}$, $(p_{\text{P}})_{\text{dB}}$, and $(\Delta bias)_{\text{dB}}$ represent the decibel value of pilot signal power from MBS and PBS, and bias value, respectively [1].

In this way, the pico cell range can be artificially expanded. However, since ER UEs connect to BSs that do not provide the strongest received power, they suffer from interference from MBS [1].

Thus, we need ICIC that can eliminate the interference from MBS to PBS. We apply ICIC by dividing the radio resource between MBSs and PBSs to avoid the interference between them [1]. Although each PBS's signal can interfere with other PBSs' ones, it is not a big problem because they have almost the same transmit power.

Although the bias values in eqs. (1) and (2) are defined to be the common one by BSs, they are possible to be decided differently by each UE. Each UE should define bias values or should select the cell because the optimal bias values of each UE that minimize the number of UE outages are affected by the ratio of radio resource and UEs' distribution [4]. However, because of the difficulty to find the optimal one suitable for those factors, most papers use the common bias value [1], [3].

III. Reinforcement Learning

Although supervised learning is effective, it may be hard to get training data on this field. Thus, RL represents a suitable alternative as it only uses the experiences of agents that learn automatically from the environment. In the RL, instead of the training data, agents get scalar values referred to as costs, and only these costs provide knowledge to agents [8].

A. Q-Learning

Q-learning is one of the typical methods of RL that is proved to converge [8]. Agent i at time t has the following parameters:

- State $s_t^i \in \mathcal{S}$; $\mathcal{S}$ is a set of states.
- Action $a_t^i \in \mathcal{A}$; $\mathcal{A}$ is a set of states.
- Cost $c_t^i \in \mathcal{S} \times \mathcal{A} \to \mathbb{R}$;

The goal of the agents is to minimize costs after selecting actions. RL will consider not only instant costs but also cumulative costs in the future that are represented as scalar value referred to as Q-value. It is defined as follows:

$$Q(s,a) = E\left\{\sum_{t=0}^{\infty} \gamma^t c(s_t, a_t) | s_0 = s, a_0 = a\right\} \quad (3)$$

where γ, $c(s_t, a_t)$, s_0, and a_0 represent discount factor ($0 \leq \gamma \leq 1$), the cost of the set of state s_t and action a_t, initial state, and initial action, respectively [9]. If the terminal state can be defined, costs are calculated up to the final one with eq. (3). However, since it can be rarely defined, the final time becomes infinity and future costs make Q-values diverse.

To make it converge, Q-learning provides the agents Q-tables storing the sets of states, actions, and Q-values that represent the effectiveness of the sets. The Q-values of all the state and action pairs are stored and updated repetitively, which realizes eq. (3) directly. Because of this, it can be said that the Q-table may inherently have a memory problem. Since this concept is simple, it makes the analysis of algorithm easier.

We describe the flow of Q-learning, illustrated in Fig. 1, as follows.

1. Agent i observe their states s_t^i from the environment and find the sets that have the state s_t^i in the Q-table. They also get costs c_t^i from the environment as the evaluation of the selected actions.
2. Using the state s_t^i and cost c_t^i that are known at step 2, the Q-value selected at the previous state and action is updated.
3. Following an action selection policy, for instance ε-greedy policy mentioned later, an action a_t^i is selected making use of the Q-values of observed states at step 1.

Through above steps, eq. (3) has been realized in Q-learning.

Q-value is updated as follows:

$$Q(s_t, a_t) \leftarrow (1-\alpha) Q(s_t, a_t) + \alpha \left[c_{t+1} + \gamma \min_a Q(s_{t+1}, a)\right] \quad (4)$$

where α represents the learning rate ($0 < \alpha \leq 1$) that controls the amount of the change of Q-value and "$\leftarrow$" means update.

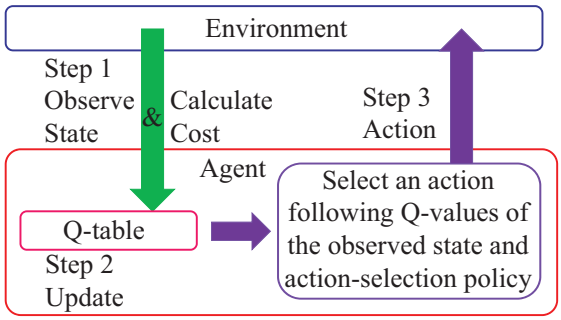


Fig. 1. The flow of Q-learning.

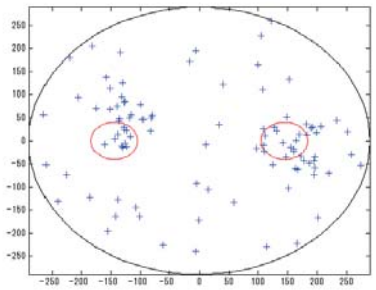


Fig. 2. UE's distribution (+ expresses UEs, red line means pico cell and MBS is the center of this circle.)

IV. Learning based Cell Selection

The bias value defined by BSs is usually set to be the common one among all UEs although the optimal bias values that minimize the number of UE outages vary from one UE to another [4]. To solve it, we propose learning based cell selection scheme that allows UEs to learn which cell to select to minimize the number of UE outages from its past experience independently with Q-learning. While the usual CRE technique provides the same bias value among all UEs, our proposed scheme allows UEs to select cells, which makes a better effective use of the frequency vacancies of the BSs.

Because all UEs learn by themselves and never share their Q-tables, this system is a multi-agent system, referred to as distributed Q-learning in [9]. Fig. 2 describes the example of UE distribution where some UEs are allocated in the hotspot areas around PBSs.

We use RBs as radio resources, blocks of subcarriers in this paper, that are the basic resource allocation units for scheduling in 3GPP-LTE system [6]. Although one or more RBs are considered to be allocated to UEs in 3GPP-LTE system [6], UEs can be allocated only one RB in this paper. To eliminate the interference from MBSs to ER UEs, RBs should be divided into MBSs and PBSs [1]. If UEs use the same RBs simultaneously, there will be interference among the UEs. UEs that do not get allocated any RB cannot access radio services.

A. Definition of State, Action, and Cost

The definition of state, action, and cost is as follows.

- **State:** The state of agent i at time t is defined as:

$$s_t^i = \{p_{\mathrm{M}}^i, p_{\mathrm{P}}^i\} \tag{5}$$

  where p_{M}^i and p_{P}^i denote the received powers of the pilot signals from MBS and PBS, respectively. Although UEs can hear many signals from various BSs, they use the largest macro and pico ones.
- **Action:** The action of agent i at time t is defined as:

$$a_t^i = j \tag{6}$$

  where j denotes the category of cells, that is, macrocell or picocell. UE will send an access request to the BS that serves the largest received power in the selected cell type.
- **cost:** The cost of agent i at time t is defined as:

$$c_t^i = n \tag{7}$$

  where n denotes the number of UEs that cannot get the radio service because of no spectrum vacancy or weak received power, referred to as UE outages. Using the backhaul between BSs, we can calculate this number and broadcast it to UEs.

On this definition, UEs decide the cell that they send an access request to minimize the number of UE outages depending on the received power from each BS. Furthermore, considering the amount of radio resources, when there are many macro RBs (MRBs), access to the MBS may be better even if the difference is small, and vice versa. Each UE can cope with aforementioned situations and decide the appropriate cell by using Q-learning.

In our system, when the agents find a new state, if they always add them to the Q-table, the size of Q-table increases, which is not allowed by the memory constraint. Moreover, this makes the learning time longer. We quantize received powers used as the state and set upper and lower limits to check and remove outlier values. After outlier checking and quantization, the state is added. By introducing this, the required memory size becomes smaller and the convergence becomes faster.

UEs keep having the data of the Q-table when they move to another PBS coverage area because even if the situation changes and if situations may have some similarities, the data got in one situations helps to learn in another situation [13]. UEs use the data as the initial values of next learning, because we expect that it helps a learning algorithm to converge faster. Even in different situations, UEs learn environment so that the table is updated.

V. Simulation Model and Results

Each PBS has one hotspot, and hotspots are placed randomly around PBSs. A hotspot area has 25 UEs inside it and they are uniformly distributed. The rest 50 UEs are also uniformly distributed inside the macro cell. The learning parameters are set as $\alpha = 0.5$, $\gamma = 0.5$, and $\varepsilon = 0.1$. We

Algorithm 1 Q-learning algorithm for UE i.

Initialize:
  let $t = 0$
  for each $s \in \mathcal{S}$, $a \in \mathcal{A}$ **do**
    initialize the Q-value, $Q(s_t^i, a_t^i)$.
  end for
Learning:
  loop
    receive pilot signals from all BSs.
    choose the largest p_{M}^i and p_{P}^i.
    if Q-table of UE i does not have $s_t^i = \{p_{\mathrm{M}}^i, p_{\mathrm{P}}^i\}$ **then**
      add s_t^i to the Q-table.
    end if
    generate a random number r $(0 \leq r \leq 1)$.
    if $r < \varepsilon$ **then** $\{\varepsilon$-greedy policy$\}$
      select a cell type a_t^i randomly.
    else
      select the cell type a_t^i that has minimum Q-value.
    end if
    send an access request based on eqs. (1) and (2).
    each UE is allocated to each RB by BSs randomly.
    get the number of UE outages as a cost from BSs.
    update the Q-value $Q(s_t^i, a_t^i)$ based on eq. (4).
    $s_t^i = s_{t+1}^i$
  end loop

TABLE I
SIMULATION PARAMETERS [1], [3]

Macro cell radius	289 m
Pico cell radius	40 m
Carrier Frequency	2.0 GHz
Bandwidth	10 MHz
RBs	50
Thermal noise density	-174 dBm/Hz
Macro BSs	1
Pico BSs	2
hotspots	2
Macro BS transmit power	46 dBm
Pico BS transmit power	30 dBm
Macro path loss model	128.1 + 37.6$\log_{10}(R)$ dB (R [km])
Pico path loss model	140.1 + 36.7$\log_{10}(R)$ dB (R [km])
Velocity of UEs	3 km/h
Channel	Rayleigh fading

show the simulation parameters in Table I. Furthermore, in this simulation, when macro signals are stronger than pico ones, as far as the difference of them is smaller than 32 dB, UEs can send access requests to PBSs.

At first, we would like to mention the storage problem that Q-learning inherently has because Q-learning has to store the Q-values. As for states, agents in our scheme add new one to Q-table if they find it. Because of this characteristic, the number of states is not fixed. During the simulation, about 1600 states are observed. In Table II, the number of observed Q-values compares with our previous work [12]. Our proposed

TABLE II
THE APPROXIMATE NUMBER OF THE OBSERVED Q-VALUES.

UE-bias [12]	UE-CS (proposed)
38000	3300

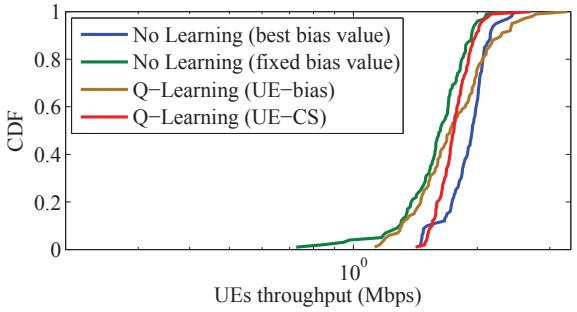
scheme has about 3300 Q-values that are two seventeenth of our previous work, which is equal to the difference of the number of the actions.

From now on, we compare four schemes after 5×10^5 trials: the proposed Q-learning scheme (UE-CS), our previous Q-learning scheme (UE-bias), no learning scheme (best bias value), and no learning scheme (fixed bias value). Both no learning schemes use common bias values among all UEs and the trial and error method, and search the bias value that minimizes the number of UE outages. No learning scheme (best bias value) searches the bias value that minimizes the number of UE outages with the trial and error method every time. Although it can get the minimum number of UE outages by a common bias value, this is not practical because the best bias value can be found after checking the number of UE outages of all bias values. Since the channel condition changes dynamically, they check these values every trial, in other words, this approach has the best performance in the case using common bias value. However, since it takes a bit long time to do that, it is not suitable in the real environment. Because of this, no learning scheme (fixed bias value) uses the trial and error method only at the first trial as a practical scheme. The difference between our previous work and proposed one is the action: the action of our previous one is a bias value of each UE, while that of our proposed one is a cell that try to connect.

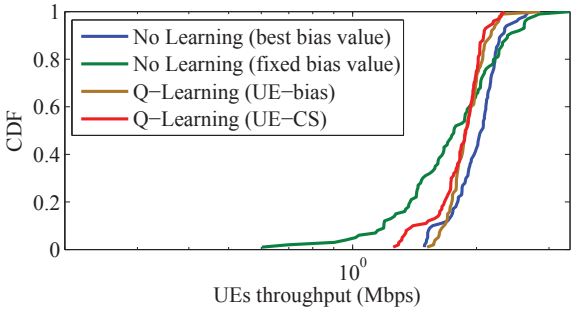
As shown in both Figs. 3, 4, the number of UE outages and the UE's average throughput change depending on the ratio of pico RBs (PRBs). This is because bias values that minimize the number of UE outages also differ according to the ratio of RBs between MBS and PBS.

No learning scheme (best bias value) has less number of UE outages than any other schemes in Fig. 3, though it is impractical scheme. Comparing with No learning scheme (fixed bias value), our proposed scheme can decrease the number of UE outages more than no learning scheme (fixed bias value) except when the ratio of PRBs is 80 %. When the ratio of PRBs is 80 %, because many UEs distributed uniformly in the macro cell try to connect to MBSs, many UEs become outage in the proposed scheme. The previous Q-learning scheme always has less UE outages than the proposed one. It is because its large numbers of the actions can express the surrounding situations of UEs well.

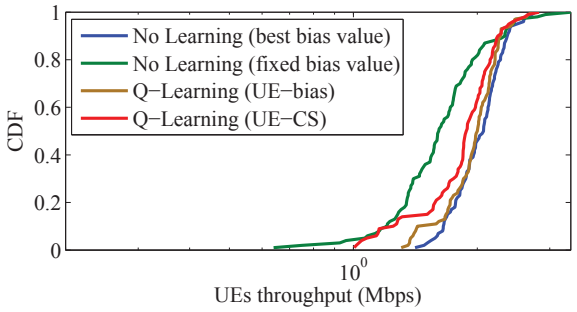
The average throughput of no learning scheme (best bias value) is also the highest one in the schemes of Fig. 4, though it is impractical scheme. Although our proposed scheme has lower average throughputs than our previous Q-learning scheme, it still keeps higher values than No learning scheme (fixed bias value).



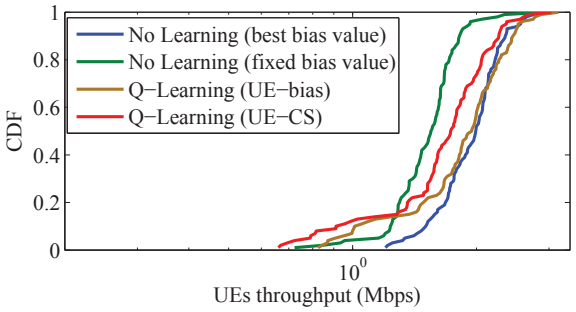
(a) CDF when the ratio of PRB is 20 %



(b) CDF when the ratio of PRB is 40 %



(c) CDF when the ratio of PRB is 60 %



(d) CDF when the ratio of PRB is 80 %

Fig. 5. CDF of UE throughput

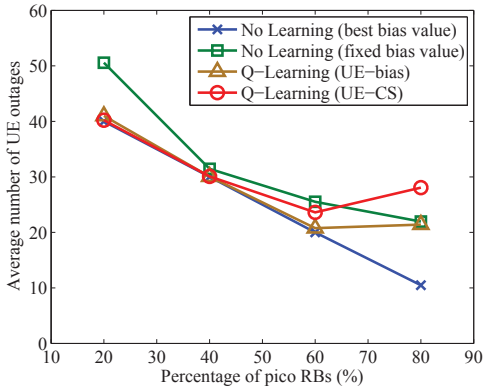


Fig. 3. Average number of UE outages at each ratio of RBs.

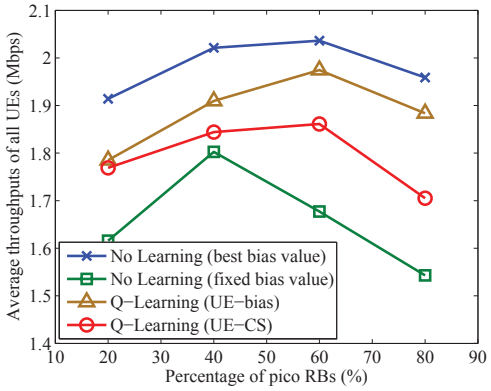


Fig. 4. Average throughput of all UEs at each ratio of RBs.

The CDF of average throughputs of all UEs through all trials is shown in Fig. 5. In the two cases that PRB ratio is more than 50 %, macro UEs are desirable to connect to PBSs because the number of allocated PRBs is larger than the number of hotspot UEs. The throughputs of cell-edge UE become worse with the increase of the PRB ratio in our proposed scheme. This is because the cell-edge UE in the macro cell do not connect to PBSs. However, comparing with No learning scheme (fixed bias value), it can be seen that almost all UEs have higher throughputs in all PRB ratios.

VI. Conclusion and Future Work

We propose a cell selection scheme without the RSRP based cell selection or usual CRE schemes. In CRE, the bias value of each UE depends on several factors such as the dividing ratio of radio resource between MBSs and PBSs, and it is determined only by the trial and error method. Thus, in this

paper, we proposed a scheme using Q-learning that UEs learn which cell to connect to minimize the number of UE outages from past experience.

We got the results of the average throughput which show that after thousands of trials, the proposed approach can perform better than the practical common bias value setting. Moreover, instead of the degradation of the throughput and the number of UE outages, our proposed scheme has less Q-values than our previous one. For these results, our proposed scheme is most preferred in some severe storage situations.

In the real environment, more types of BSs are placed in the same cell than our system. As one of the future work, we expect to add other types of BSs, for instance, femto BS, to this system to apply our work in the real situation.

REFERENCES

[1] D. Pérez-López and X. Chu, "Inter-Cell Interference Coordination for Expanded Region Picocells in Heterogeneous Networks," *IEEE ICCCN*, June/Aug. 2011, pp. 1–6.

[2] A. Damnjanovic *et al.*, "A survey on 3GPP heterogeneous networks," *IEEE Wireless Communications,* vol. 18, no. 3, pp. 10–21, June 2011.

[3] J. Sangiamwong *et al.*, "Investigation on Cell Selection Methods Associated with Inter-cell Interference Coordination in Heterogeneous Networks for LTE-Advanced Downlink," *European Wireless*, Apr. 2011, pp. 1–6.

[4] M. Shirakabe *et al.*, "Performance Evaluation of Inter-cell Interference Coordination and Cell Range Expansion in Heterogeneous Networks for LTE-Advanced Downlink." *ISWCS*, Nov. 2011, pp. 844–848.

[5] İ. Güvenç *et al.*, "Range Expansion and Inter-Cell Interference Coordination (ICIC) for Picocell Networks," in *IEEE VTC Fall*, Sep. 2011, pp. 1–6.

[6] M. Lee and S. K. Oh, "On resource block sharing in 3GPP-LTE system," *APCC*, Oct. 2011, pp. 38–42.

[7] H. Ishii *et al.*, "A novel architecture for LTE-B: C-plane/U-plane split and Phantom Cell concept," *GC Wkshps*, Dec. 2012, pp. 624–630.

[8] R. S. Sutton and A. G. Barto, *Reinforcement Learning.* Cambridge, MA: MIT Press 1998.

[9] A. Galindo-Serrano and L. Giupponi, "Distributed Q-Learning for Aggregated Interference Control in Cognitive Radio Networks," *IEEE Trans. Veh. Technol.,* vol. 59, no. 4, pp. 1823–1834, May 2010.

[10] R. Razavi *et al.*, "Self-optimization of capacity and coverage in LTE networks using a fuzzy reinforcement learning approach," *IEEE PIMRC*, Sep. 2010, pp. 1865–1870.

[11] M. Simsek *et al.*, "Dynamic Inter-Cell Interference Coordination in HetNets: A Reinforcement Learning Approach," *IEEE GLOBECOM*, Dec. 2012, pp. 5668–5672.

[12] T. Kudo and T. Ohtsuki, "Cell range expansion using distributed Q-learning in heterogeneous networks," *EURASIP Journal on Wireless Communications and Networking,* vol. 2013, no. 61, Mar. 2013.

[13] A. Galindo-Serrano *et al.*, "Learning from Experts in Cognitive Radio Networks: The Docitive Paradigm," *CROWNCOM*, June 2010, pp. 1–6.