Traffic Offloading in Two-Tier Multi-Mode Small Cell Networks over Unlicensed Bands: A Hierarchical Learning Framework

Youming Sun[1,2], Hongxiang Shao[2,3], Xin Liu[4], Jian Zhang[1], Junfei Qiu[2] and Yuhua Xu[2]
[1] National Digital Switching System Engineering & Technological Research Center
Zhengzhou 450000 - China
[e-mail: {sunyouming10, zhangjianndsc}@163.com]
[2] College of Communications Engineering, PLA University of Science and Technology
Nanjing 210007 - China
[e-mail: shaohongxiang2003@163.com, leo_nanjing@126.com, junfeiqiu@163.com, yuhuaenator@gmail.com]
[3]Luoyang Institute of Science and Technology
Luoyang 471000 - China
[4]Comprehensive Trainning Base of Guangzhou Military Area
Guilin 541002 - China
*Corresponding author: Hongxiang Shao

Received July 12, 2015; revised September 8, 2015; accepted September 20, 2015; published November 19, 2015

Abstract

This paper investigates the traffic offloading over unlicensed bands for two-tier multi-mode small cell networks. We formulate this problem as a Stackelberg game and apply a hierarchical learning framework to jointly maximize the utilities of both macro base station (MBS) and small base stations (SBSs). During the learning process, the MBS behaves as a leader and the SBSs are followers. A pricing mechanism is adopt by MBS and the price information is broadcasted to all SBSs by MBS firstly, then each SBS competes with other SBSs and takes its best response strategies to appropriately allocate the traffic load in licensed and unlicensed band in the sequel, taking the traffic flow payment charged by MBS into consideration. Then, we present a hierarchical Q-learning algorithm (HQL) to discover the Stackelberg equilibrium. Additionally, if some extra information can be obtained via feedback, we propose an improved hierarchical Q-learning algorithm (IHQL) to speed up the SBSs' learning process. Last but not the least, the convergence performance of the proposed two algorithms is analyzed. Numerical experiments are presented to validate the proposed schemes and show the effectiveness.

Keywords: Small cells, Stackelberg game, Q-learning, Traffic offloading,

Part of this paper had been submitted to the 2015 IEEE International Conference on Wireless Communications and Signal Processing (WCSP).
This work was supported by the National Science Foundation of China under Grant 61401508 and Grant 61172062.

1. Introduction

With the blasting increase of mobile data from the introduction of smart phones, tablet computers and other new mobile devices, it brings a number of challenges to cellular network operators (CNOs) such as boost system capacity and improve the coverage simultaneously. The next generation cellular networks, known as 5G, require 1000-fold capacity improvement comparing to the current Long Term Evolution (LTE) 4G networks [1]-[4]. The effective solution to CNO is to make use of traffic offloading to transfer part of the traffic load elsewhere off the main networks. The main object of traffic offloading is to support more capacity-hungry applications simultaneously maintain the satisfactory quality of experience (QoE) of end users[5][6]. From the perspective of industry, small cells and Wi-Fi networks are the candidates undertaking the traffic offloading. Recently, the concept of multi-mode small cell base station [7][8], owing the ability of simultaneously accessing licensed and unlicensed band, is emerging and attracts significant interest of many industrial companies such as Qualcomm and Huawei Corp. [9][10].

Due to the limited available licensed bands for CNOs, the European telecommunications standards institute reconfigurable radio systems (ETSI RRS) technical standardization committee has released a detailed outline of a new concept of licensed shared access (LSA) to promote the spectrum sharing in 5G networks [11]. LSA is a spectrum sharing approach designed to serve short-term to mid-term industry needs through a quasi-static allocation of shared spectrum to CNOs. Specifically, the incumbent spectrum holders negotiate their spectrum with demanders in underutilized location with quality of service (QoS) guarantee. Thus, LSA approach can supply extra spectrum complement to alleviate the scarcity of operators licensed band. On the other hand, there are a number of available unlicensed bands, i.e. 2.4GHz ISM (Industrial, Scientific and Medical) and 5GHz U-NII (Unlicensed National Information Infrastructure) bands. In addition, United States and Europe recently published rules to access the TV white spaces (TVWS) [12]. In the current LTE system, the advanced carrier aggregation (CA) technology, defined in LTE Rel-10 and Rel-12, makes it possible to utilize the unlicensed band for traffic offloading [13], where the unlicensed carriers are operated as secondary carriers associated to and controlled by the licensed LTE primary carriers. Extensive attention has been paid to the research on the LTE in unlicensed band, also known as LTE-U [14][15].

In this paper, we address the traffic offloading issue in two-tier multi-mode small cell networks over unlicensed bands. Each small cell base station (SBS) transfers part of traffic load to the sharing band may affect the capacity performance of the neighbors who access the unlicensed band at the same time. This implies the traffic-offloading optimization of each SBS is coupled with its neighbors. Convex optimization technologies have advantages on the centralized optimization and have been widely applied to cooperative resource allocation in mobile communication, such as [16][17]. In the context of randomness of SBS's activity and lack of mutual coordination, it is desirable to design a distributed scheme to handle traffic offloading, so that there is no need of timely cross-tier and co-tier information exchange which may bring heavily overhead burden to communication system especially in large scale scenario.

Game theory is a powerful tool to analyze and predict the outcomes of interactive decision makers [18][19][20]. It has been shown that the Stackelberg game model is suitable for analyzing the hierarchical competition in the two-tier networks consisting macro base

station (MBS) and underlying SBSs 0-[24].

Therefore, we formulate this problem of traffic offloading in two-tier small cell networks as a Stackelberg game. To be specific, the MBS is modeled as leader and SBSs are followers. A pricing mechanism is adopted by MBS to balance the traffic load between the primary and complementary network. To begin with, MBS broadcasts current price information to all SBSs firstly, then each SBS allocates the traffic load in licensed and unlicensed band via appropriate power allocation in the sequel, considering the traffic flow payment charged by MBS in licensed band. The goal of all players in the hierarchical game is to maximize their revenue. Furthermore, we propose a hierarchical learning framework to discover the Stackelberg equilibrium (SE). A hierarchical Q-learning algorithm (HQL) is proposed firstly and then an improved hierarchical Q-learning algorithm (IHQL) is presented to speed up the convergence of SBSs' learning when some feedback information can be obtained. During the process of learning, MBS and SBSs learn their optimal policies through interaction with the network environment. Our preliminary work has been reported in [25]. In this paper, we further provide rigorous theoretical proofs and performance analysis. In addition, extensive simulations are conducted to verify the effectiveness of the proposed schemes.

The rest of this paper is organized as follows. Section 2 introduces the related works. In Section 3, the system model and problem formulation are presented. Specifically, we apply Stackelberg game to model the agents' behaviors in two-tier small cell networks. In Section 4, we propose a hierarchical learning framework and then present two hierarchical Q-learning algorithms to discover the SE. In Section 5, simulation results are given. Finally, the conclusion is drawn in Section 6.

2. Related Work

In this section, some related studies are presented. In the following, we shall use interchangeably capacity offloading and traffic offloading. There exist some efforts on the traffic offloading in heterogeneous cellular networks such as [26]-[29]. Specifically, in [26], Chiang *et al.* proposed a scheme based on reinforcement learning to intelligently offload traffic in a stochastic macro cell. In [27], Chen *et al.* provided a brief survey on existing traffic offloading techniques in wireless networks. Moreover, small cells are low-power and short-range access points that can operate in a flexible and economic way. It is regarded as a promising approach to implement the traffic offloading and improve the system capacity [1].

Since the coupling of co-tier interference across the SBSs and the ad-hoc topology of small cell networks introduced by randomness of SBS's activity, many existing studies such as 0-[24] resort to the hierarchical game, also known as Stackelberg game or leader-follower game, to model the players' behaviors in two-tier networks in the limited coordination scenario.

Recently, the multi-mode SBS is emerging as an advanced technology to expand the LTE capacity to meet the traffic demands by integrating numerous unlicensed bands into the current LTE system especially at the standpoint of the industrial field [9][10]. Whereas, there are limited studies on traffic offloading over unlicensed bands in multi-mode cell networks [7][30][31]. To be specific, in [30], Bennis *et al.* investigated cross system learning by means of which SBSs self-organize and autonomously steer their traffic flows across different radio access technologies. In [31], Zhang *et al.* studied the wireless service providers (WSP) utilize unlicensed band to transfer partial traffic load and formulated this problem as a non-cooperative capacity offloading game. In [7], Liu *et al.* focused on maximizing the total

user satisfaction/utility of the small cell user over traffic balancing over unlicensed band.

However, the significant difference between our proposed scheme and the existing schemes mentioned above lies in the followings:

(i) Our framework considers the hierarchical feature in the two-tier small cell networks and models the hierarchical interaction between MBS and SBSs by Stackelberg game rather than the normal non-cooperative game, only considering the relationship among SBSs.

(ii) The existing works mentioned above don't consider the access cost in licensed band from the perspective of economy; that is to say, SBS needs to pay the traffic flow in the licensed band to the corresponding wireless operator, which is a common commercial mode[32].

(iii) In our schemes, MBS can make a tradeoff between the revenue and the traffic load of primary networks via adjusting the traffic flow price.

3. System Model and Problem Formulation

3.1 System Model

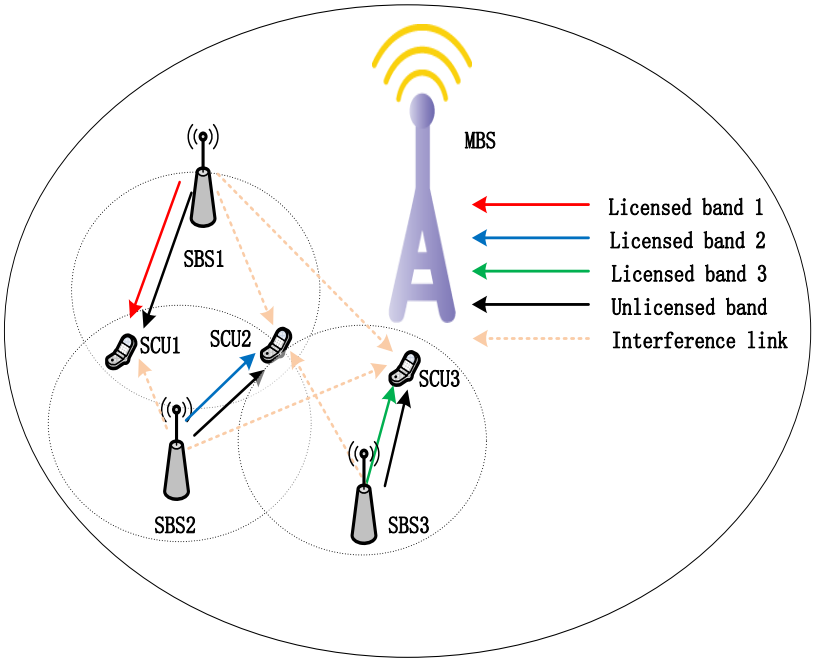


Fig. 1. System model.

As shown in **Fig. 1,** we consider the downlink transmission of a two-tier OFDMA small cell network consisting a central MBS and N multi-mode SBSs. Each SBS services a certain number of wireless devices and operate in closed access manner, i.e. only licensed wireless devices can communicate with the corresponding SBS. For analytical tractability, we assume that at most one user of SBS can be scheduled at a specific time slot.

Let 0 denote the index of the MBS and the SBSs' set is denoted by $\mathbf{B}=[1,2\cdots N]$. Moreover, let available scheduled SBSs' user set is denoted by $\mathbf{SCU}=[1,2,\cdots N]$. For SBSs, although they can simultaneously access licensed and unlicensed band, they need taking the

differences of bands into consideration, i.e. accessing licensed bands means the guarantee of end user's QoS while relatively high accessing cost charged by wireless operator comparing to free unlicensed bands with no QoS provisioning. In this paper, we address the traffic offloading issue of SBSs over unlicensed bands. Suppose the split-spectrum scheme is adopted in two-tier small cell networks and each SBS *i* occupies orthogonal licensed band, whose bandwidth is B_i^{L} Hertz. In addition, all SBS can access the sharing unlicensed band, which bandwidth is B^{U} Hertz. Therefore, the total available bandwidth B_{total}, consisting of all SBS's licensed bands and the unlicensed band; that is, $B_{\text{total}} = B^{\mathrm{U}} + \sum_i B_i^{\mathrm{L}} = (1+\sum_i \alpha_i) B^{\mathrm{U}}$ Hertz, where $\alpha_i = B_i^{\mathrm{L}} / B^{\mathrm{U}}$.

We assume that the SBS *i* has a maximum power constraint p_i, and that it can arbitrarily allocate its power on its own licensed band and the unlicensed band. Denote by $\beta_i \in [0,1]$ the fraction of SBS *i*'s power allocated on its licensed band; that is to say the SBS *i* allocates $\beta_i p_i$ in its private band, and the residual power $(1-\beta_i) p_i$ in the unlicensed band. It is reasonable for each SBS use up its all power. If there is power left unallcoation, SBS can allocate this residual power on the unlicensed band to increase its achievable rate without additional cost.

Firstly, we give the utility function of SBS *i* as follows (normalized by B^{U})

$$U_i(\lambda_0, \boldsymbol{\beta}) = \log_2\left(1 + \frac{p_i h_{i,i}^{\mathrm{U}} (1-\beta_i)}{n_0 + \sum_{j \neq i} p_j h_{j,i}^{\mathrm{U}} (1-\beta_j)}\right) + \alpha_i \log_2\left(1 + \frac{p_i h_{i,i}^{\mathrm{L}} \beta_i}{n_0 \alpha_i}\right)(1-\lambda_0), \tag{1}$$

where $h_{j,i}^{\mathrm{U}}$ and $h_{j,i}^{\mathrm{L}}$ denote the normalized channel gain between the BS *j* and the user equipment (UE) *i* in unlicensed and licensed band, respectively. $\boldsymbol{\beta} = [\beta_1, \beta_1, \ldots, \beta_N]$ is the vector of all SBSs' power allocation strategies. λ_0 is the unit traffic flow price charged by operator and n_0 denotes the Gaussian noise power in unlicensed band. In addition, the term $n_0 + \sum_{j \neq i} p_j h_{j,i}^{\mathrm{U}} (1-\alpha_j)$ denotes the received interference in the unlicensed band.

We can see the SBS's utility consists of two parts, the first part is the revenue from the unlicensed band and the second part represents the profit form licensed band. The object of SBS *i* is to optimize its power allocation fraction β_i to maximum its utility:

$$(\mathbf{P}1): \ \beta_i = \max_{\beta_i} U_i(\lambda_0, \boldsymbol{\beta}). \tag{2}$$

From the MBS's side, MBS's revenue comes from pricing the SBS's traffic flow in the licensed band and its utility function is as follows:

$$U_0(\lambda_0, \boldsymbol{\beta}) = \lambda_0 \sum_i^N \alpha_i \log_2\left(1 + \frac{p_i h_{i,i}^{\mathrm{L}} \beta_i}{n_0 \alpha_i}\right) \tag{3}$$

The object of MBS is to maximize its revenue and find the optimal price, the optimization problem of MBS is:

$$(\mathbf{P}2): \ \lambda_0 = \max_{\lambda_0} U_0(\lambda_0, \boldsymbol{\beta}) \tag{4}$$

3.2 Stackelberg Game solution

In two-tier small cell networks, it is suitable and natural to apply Stackelberg game to model

the hierarchical interaction between MBS and SBS0-[24]. Specifically, The MBS is modeled as leader and move first. In the sequel, SBSs are followers and take their best response dynamic based on the observation of leaders' actions. Mathematically, the Stackelberg game is defined as

$$G=\left\{\mathbf{B}\bigcup 0,\{\lambda_0\},\{\beta_i\}_{i\in\mathbf{B}},\{U_0\},\{U_i\}_{i\in\mathbf{B}}\right\}, \tag{5}$$

where $\{\lambda_0\}$ and $\{\beta_i\}_{i\in\mathbf{B}}$ denote the strategy space of MBS (leader) and FBS i (follower), respectively.

The solution of Stackelberg game is to find the Stackelberg equilibrium (SE).

Definition 1 (Stackelberg Equilibrium, SE): A strategy profile $(\lambda_0^*,\boldsymbol{\beta}^*)$ is called Stackelberg equilibrium if λ_0^* maximizes the utility of the MBS (leader) and $\boldsymbol{\beta}^*$ is the best response of SBS. Mathematically, for any strategy profile $(\lambda_0,\boldsymbol{\beta})$, the following conditions are satisfied:

$$U_0(\lambda_0^*,\boldsymbol{\beta}^*)\geq U_0(\lambda_0,\boldsymbol{\beta}^*) \tag{6}$$

$$U_i(\beta_i^*,\lambda_0^*,\boldsymbol{\beta}_{-i}^*)\geq U_i(\beta_i,\lambda_0^*,\boldsymbol{\beta}_{-i}^*),\forall i\in\mathbf{B} \tag{7}$$

where $\boldsymbol{\beta}_{-i}^*$ denotes that all SBSs take the best response strategies except SBS i.

Note that the Stackelberg game is the extension of normal non-cooperative game. SE is a stable operation point and that means no player can improve its utility by deviating unilaterally in the hierarchical competition structure. Given the traffic flow price of the MBS, then SBSs, which are selfish and rational, play a strictly non-cooperative capacity offloading subgame.

Theorem 1: The SE always exists in our defined Stackelberg game.

Proof: For a given $\forall\lambda_0$, the Stackelberg game reduced to the non-cooperative game $G_f=\{\lambda_0,\mathbf{B}_f,\{\beta_i\},\{U_i\}\}$. To prove the existence of NE in the lower subgame, we introduce the following Lemma 1.

Lemma 1[33]: A NE exists in game $G_f=\left\{\mathbf{B}_f,\{\beta_i\},\{U_i\}\right\}$ if for all $i=1,2,\cdots,N$

(1) $\{\beta_i\}$ is a nonempty, convex and compact subset of some Euclidean space R^N.

(2) U_i is continuous in $\boldsymbol{\beta}$ and quasi-concave in β_i.

SBS i's strategy space is defined to be $\{\beta_i:0\leq\beta_i\leq 1\}$, and it is nonempty, convex and compact subset of some Euclidean space R^N.

From (1), we can see the $U_i(\lambda_0,\boldsymbol{\beta})$ is continuous in $\boldsymbol{\beta}$. Next, we take the second-order derivative with respect to β_i to prove its concavity, i.e.

$$\frac{\partial U_i}{\partial\beta_i}=\frac{1}{\ln 2}\left[\frac{\alpha_i(1-\lambda_i)p_ih_{i,i}^{\mathrm{L}}}{n_0\alpha_i+p_ih_{i,i}^{\mathrm{L}}\beta_i}-\frac{p_ih_{i,i}^{\mathrm{U}}}{n_0+\sum_j p_jh_{j,i}^{\mathrm{U}}(1-\beta_j)}\right], \tag{8}$$

$$\frac{\partial^2 U_i}{(\partial\beta_i)^2}=-\frac{1}{\ln 2}\frac{\alpha_i(1-\lambda_0)\left(p_ih_{i,i}^{\mathrm{L}}\right)^2}{\left(n_0\alpha_i+p_ih_{i,i}^{\mathrm{L}}\beta_i\right)^2}-\frac{1}{\ln 2}\frac{\left(p_ih_{i,i}^{\mathrm{U}}\right)^2}{\left(n_0+\sum_j p_jh_{j,i}^{\mathrm{U}}(1-\beta_j)\right)^2}<0. \tag{9}$$

Thus, the $U_i(\lambda_0, \boldsymbol{\beta})$ is concave in β_i. Following the **Lemma 1,** there exists NE in the lower subgame $G_f = \{\lambda_0, \mathbf{B}, \{\beta_i\}, \{U_i\}\}$. Let $\mathbf{NE}(\lambda_0)$ denotes the best response dynamics of the followers with given MBS's strategy λ_0, the SE can be equivalently defined as

$$U_0(\lambda_0^*, \mathbf{NE}(\lambda_0^*)) \geq U_0(\lambda_0, \mathbf{NE}(\lambda_0))$$

Therefore, there exists λ_0^* satisfying the following:

$$U_0(\lambda_0^*, \mathbf{NE}(\lambda_0^*)) = \sup_{\lambda_0} U_0(\lambda_0, \mathbf{NE}(\lambda_0)).$$

It implies that SE always exists in our defined Stackelberg game. To better understand the process of this proof, we give the corresponding flow chart in **Fig. 2**.
This completes the proof.

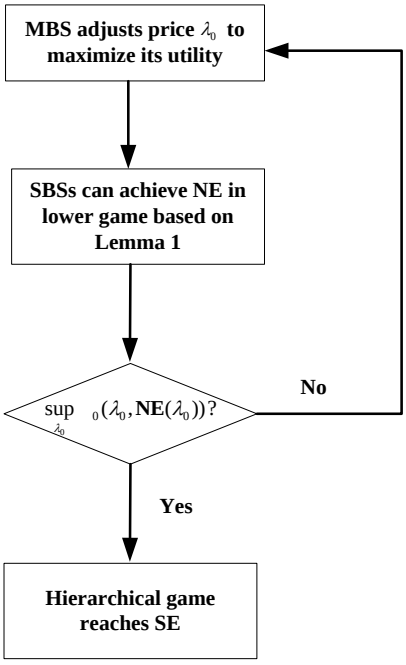


Fig. 2. The process of obtaining SE

4. Hierarchical Learning Framework

In the developed hierarchical learning framework, the SBS and MBS are assumed to behave as intelligent agents and have self-learning ability to automatically optimize their configuration. In this section, we apply hierarchical stochastic learning based on the Stackelberg game framework to implement the traffic offloading.

To be compatible with reinforcement learning mechanism and we assume that each player has a finite and discretized action set. In stochastic learning game, each player aims at maximizing its own expected utility function and commits a best policy. Specifically, a policy of agent i at time slot t is defined to be a probability vector $\boldsymbol{\pi}_i^t = \left(\pi_i^t(a_{i,1}), \pi_i^t(a_{i,2}), \cdots, \pi_i^t(a_{i,|\mathbf{A}_i|})\right)$, where $\pi_i^t(a_{i,c_i})$ means the probability with which agent i

chooses the action a_{i,c_i}, that is, an price of MBS and the power allocation fraction β_{i,a_i} of SBS from the available action set $\mathbf{A}_i = \left(a_{i,1}, a_{i,2}, \cdots, a_{i,|\mathbf{A}_i|}\right), i \in \mathbf{B} \bigcup \{0\}$, which satisfies $\sum_{a_{i,c_i}} \pi_i^t(a_{i,c_i}) = 1$. To be specific, the SBS i's and MBS's available action set are defined as $\mathbf{A}_i = \left(\beta_{i,1}, \beta_{i,2}, \cdots \beta_{i,|\mathbf{A}_i|}\right)$ and $\mathbf{A}_0 = \left(\lambda_{0,1}, \lambda_{0,2}, \cdots, \lambda_{0,|\mathbf{A}_0|}\right)$, respectively, where the $|\mathbf{A}_i|$ denotes the cardinal number of $\mathbf{A}_i$. The action space for all agents is denoted by $\mathbf{A} = \otimes_{i \in \mathbf{B} \cup \{0\}} \mathbf{A}_i$, where $\otimes$ is the Cartesian product. Note that in some practical scenario such as the 3GPP LTE only support discrete power control in the downlink.

Then, the expected utility of agent i can be expressed as

$$u_i(\boldsymbol{\pi}_i^t, \boldsymbol{\pi}_{-i}^t) = \mathrm{E}\left[U_i \,|\, \boldsymbol{\pi}_i^t, \boldsymbol{\pi}_{-i}^t \right] = \sum_{\mathbf{a}^t \in \mathbf{A}} U_i(\mathbf{a}^t) \prod_{i \in \mathbf{B} \cup \{0\}} \pi_{i,c_i}^t, \tag{10}$$

where $\mathbf{a}^t = \left[a_{0,c_0}^t, a_{1,c_1}^t, \cdots, a_{N,c_N}^t \right]$ denotes the chosen action by agent i based on current policy $\boldsymbol{\pi}_i^t$.

Based on above analysis, we have the following definition of an SE for the hierarchical learning framework. The MBS's objective is to maximize its revenue as

$$(\mathbf{P}3): \max_{\boldsymbol{\pi}_0} u_0(\boldsymbol{\pi}_0, \boldsymbol{\pi}_{-0}). \tag{11}$$

Similarly, each SBS i's objective is

$$(\mathbf{P}4): \max_{\boldsymbol{\pi}_i} u_i(\boldsymbol{\pi}_i, \boldsymbol{\pi}_{-i}). \tag{12}$$

In the following, we will give the similar definition of the SE in hierarchical learning framework.

Definition 2. A stationary policy profile $(\boldsymbol{\pi}_0^*, \boldsymbol{\pi}_{-0}^*)$ is the SE for hierarchical learning framework if for any policy profile $(\boldsymbol{\pi}_0, \boldsymbol{\pi}_{-0})$ the followings hold

$$\begin{cases} u_0(\boldsymbol{\pi}_0^*, \boldsymbol{\pi}_{-i}^*) \geq u_0(\boldsymbol{\pi}_0, \boldsymbol{\pi}_{-0}^*); \\ u_i(\boldsymbol{\pi}_i^*, \boldsymbol{\pi}_{-i}^*) \geq u_i(\boldsymbol{\pi}_i, \boldsymbol{\pi}_{-i}^*). \end{cases} \tag{13}$$

Then, we prove the **Theorem 2** as follows and then show the SE always exists in the hierarchical learning framework.

Theorem 2: Given MBS's policy $\boldsymbol{\pi}_0$, there always exists a mixed strategy $(\boldsymbol{\pi}_i, \boldsymbol{\pi}_{-\{i,0\}}, \boldsymbol{\pi}_0)$ satisfies $u_i(\boldsymbol{\pi}_i^*, \boldsymbol{\pi}_{-\{i,0\}}^*, \boldsymbol{\pi}_0) \geq u_i(\boldsymbol{\pi}_i, \boldsymbol{\pi}_{-\{i,0\}}^*, \boldsymbol{\pi}_0)$, which is a Nash equilibrium (NE) point.

Proof: Given MBS's policy $\boldsymbol{\pi}_0$, the follower game is a non-cooperation game. Due to every finite strategic game has at least one mixed strategy equilibrium as shown in [34], i.e., there exists NE point. This completes the proof.

In the following, similar as the proof in **Theorem 1,** the MBS's optimal strategy can be expressed as

$$\boldsymbol{\pi}_0^* = \arg\max_{\boldsymbol{\pi}_0} u_0(\boldsymbol{\pi}_0, \mathbf{NE}(\boldsymbol{\pi}_0)) \tag{14}$$

Therefore, we can conclude that $(\boldsymbol{\pi}_0^*, \mathbf{NE}(\boldsymbol{\pi}_0^*))$ constitutes a SE in the sense of stationary mixed strategy, which implies there exists SE in the considered hierarchical learning framework.

4.1. Hierarchical Q-learning algorithm (HQL)

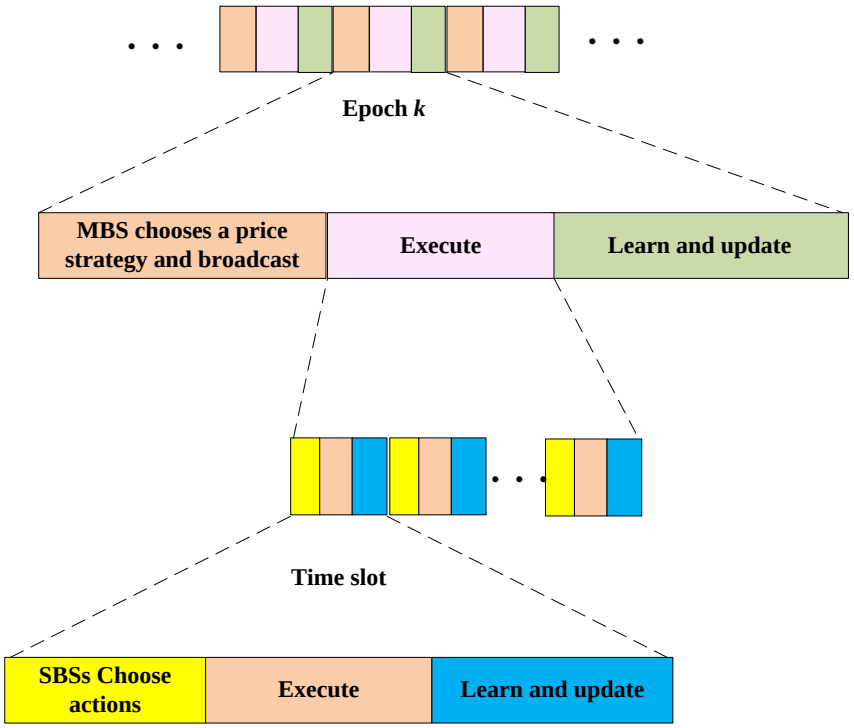


Fig. 3. Time structure of the proposed hierarchical learning framework.

Reinforcement learning (RL) algorithms are closely related to dynamic programming, which is always applied in optimal control. In standard RL framework, a learning agent observes the state of its outer environment and then selects and performs an action. Performing the action changes the state of the world and an immediate payoff can be obtained by agent. Positive payoffs are regarded as rewards, whereas, the negative payoffs are viewed as punishments. In the face of reward and punishment, each agent must choose actions to maximize its long term sum or average payoffs in future [37]. Therefore, in the hierarchical learning framework, each smart agent's overall goal is to learn to optimize its individual long-term cumulative reward via repeatedly interacting with network environment. In this paper, we adopt the Q-learning algorithm [35], which is a common reinforcement learning method and widely used in self-organized small cell networks [38].

In Q-learning process, agents' strategies are parameterized through so called Q-functions, which characterize the relative utility of a particular action and are updated during the course of the agent's interaction with the network environments. With the update of Q-functions, the policies that yield high rewards are reinforced and the optimal Q-function is found in a recursive way. To be specific, let $Q_i^t(a_{i,c_i}^t)$ denote the corresponding Q-function of agent i's action a_{i,c_i}^t based on current policy $\boldsymbol{\pi}_i^t$ at time slot t. Then, after selecting action a_{i,c_i}^t, the corresponding Q-value is updated as follows

$$Q_i^{t+1}(a_{i,c_i}^{t+1}) = (1-\kappa_i^t)Q_i^t(a_{i,c_i}^t) + \kappa_i^t u_i(a_{i,c_i}^t, \boldsymbol{\pi}_{-i}^t), \tag{15}$$

where $\kappa_i^t \in [0,1)$ is the learning rate, satisfying $\sum_{t=0}^{\infty}\kappa_i^t = \infty, \sum_{t=0}^{\infty}(\kappa_i^t)^2 < \infty$. $u_i(a_{i,c_i}^t, \boldsymbol{\pi}_{-i}^t)$ is the observed reward for action *i* at time slot *t*,

$$u_i(a_{i,c_i}^t, \boldsymbol{\pi}_{-i}^t) = \sum_{\mathbf{a}_{-i}^t \in \mathbf{A}_{-i}} U_i(a_{i,c_i}^t, \mathbf{a}_{-i}^t) \prod_{j \in \mathbf{B}\cup\{0\}/i} \pi_{j,c_j}^t, \tag{16}$$

where $\mathbf{a}_{-i}^t = \left[a_{0,c_0}^t, \ldots, a_{j-1,c_{j-1}}^t, a_{j+1,c_{j+1}}^t, \cdots a_{N,c_N}^t\right]$ and $\mathbf{A}_{-i} = \otimes_{j\in\mathbf{B}\cup\{0\}/i}\mathbf{A}_j$.

Each BS update its policy based on Boltzmann distribution

$$\pi_i^t(a_{i,k}) = \frac{\exp\left[Q_i^t(a_{i,k}^t)/\tau_i\right]}{\sum_{c_i \in \mathbf{A}_I}\exp\left[Q_i^t(a_{i,c_i}^t)/\tau_i\right]}, \tag{17}$$

where the temperature $\tau_i > 0$ controls the tradeoff between exploration and exploitation. To be specific, for $\tau_i \to 0$, agent greedily and chooses the policy corresponding to the maximum Q-value which means pure exploitation, whereas for $\tau_i \to \infty$, agent's policy is completely random which means pure exploration [39].

In the learning process, the MBS can learn the optimal price policy through recursively updating the Q-functions based on (15) and (17). However, each SBS can know neither other competing SBSs' policies nor the utility $U_i(a_{i,c_i}^t, \mathbf{a}_{-i}^t)$ before taking action a_{i,c_i}^t in the time slot *t*. The only information the SBSs can obtain is the MBS's price information a_{0,c_0}^t, which is regarded as the common knowledge for followers in the hierarchical learning framework.

Suppose that MBS update the policy every *T* time slot, which is defined as one epoch. After the action chosen by MBS and then broadcast it to all SBSs, then each SBS learn to the optimal best response and feedback corresponding policy to MBS at the end of each epoch. Therefore, the SBS *i*'s Q-function updates as

$$Q_i^{t+1}(a_{i,c_i}) = (1-\kappa_i^t)Q_i^t(a_{i,c_i}) + \kappa_i^t u_i(a_{i,c_i}, a_{0,c_0}), \tag{18}$$

where the estimated expected utility $u_i(a_{i,c_i}, a_{0,c_0})$ can be expressed as:

$$u_i^t(a_{i,c_i}) = \begin{cases} \dfrac{U_i(a_{i,c_i}^t, a_{0,c_0}, \mathbf{a}_{-(i,0)}^t) - u_i^{t-1}(a_{i,c_i}^t, a_{0,c_0})}{n_i^t(a_{i,c_i}, a_{0,c_0}) + 1} + u_i^{t-1}(a_{i,c_i}^t, a_{0,c_0}); \ a_{i,c_i} = a_{i,c_i}^t \\ u_i^{t-1}(a_{i,c_i}^t); \text{otherwise}. \end{cases} \tag{19}$$

where the $n_i^t(a_{i,c_i}, a_{0,c_0})$ is the record number of the combination (a_{i,c_i}, a_{0,c_0}) in each epoch. We can see that the policy update for MBS and SBS are based on different time scales. The time structure of the proposed hierarchical learning framework is shown in the **Fig. 3.** To be specific, SBS updates its Q-functions in each time slot while the MBS executes at the end of every one epoch. At the end of the *k*-th epoch, the Q-function update of MBS is shown as follows:

$$Q_0^{k+1}(a_{0,c_0}) = (1-\kappa_0)Q_0^k(a_{0,c_0}) + \kappa_i u_0^k(a_{0,c_0}, \boldsymbol{\pi}_{-i}^{kT}) \tag{20}$$

Next, we present the details of the proposed learning algorithm in Algorithm 1.

Algorithm 1: Hierarchical Q-learning Algorithm (HQL)

Step1: Initialize the Q-functions $Q_i(a_{i,c_i})$ at time slot t=1 for $a_{i,c_i} \in \mathbf{A}_i$

Step2: Learning process of SBSs

(1) At the beginning of epoch k, the MBS chooses a price a_{0,c_0} according to its policy $\boldsymbol{\pi}_0$ and broadcasts it to all SBSs in the network.

(2) Each SBS i selects a power allocation fraction a_{i,c_i} according to its policy $\boldsymbol{\pi}_i^t$.

(3) Each SBS i computes its achieved utility $U_i(a_{i,c_i}^t, a_{0,c_0}, \mathbf{a}_{-(i,0)}^t)$ via the feedback information, and updates the estimated expected $u_i^t(a_{i,c_i})$ in (19).

(4) Each SBS updates the Q-values according to (18) and the corresponding policy $\pi_i^t(a_{i,c_i})$ according to (17).

(5) All SBSs send the current policy to MBS at the end of each epoch.

Step3: The MBS computes the accrued utility $u_0^k(a_{0,c_0}, \boldsymbol{\pi}_{-i}^{kT})$. .

Step4: The MBS updates the Q-values and corresponding policy according to (15) and (17), respectively.

Step5: The MBS selects an action based on the updated policy.

Step6:Update k=k+1 and jump to $k = k_{\max}$.

4.2. Improved Hierarchical Q-learning algorithm (IHQL)

In practical implementations, the convergence speed may be one limitation when the action set of each agent is relatively large. In order to achieve fair performance, each action's Q function must be sampled sufficiently. In the proposed HQL algorithm, each SBS only can update one action's Q-value. Nevertheless, we found that there exist rooms to improve the convergence of HQL. Assume that each SBS have the cognitive ability and can measure the interference level in the unlicensed band, we can speed up HQL by updating all actions' Q-function simultaneously.

For instance, at the end of time slot t, SBS i get the feedback information, containing the channel estimation $(h_{i,i}^{\mathrm{U}}, h_{i,i}^{\mathrm{L}})$ and the interference level $I_i^{\mathrm{U}} = n_0 + \sum_{j \neq i} p_j h_{j,i}^{\mathrm{U}}(1-\beta_j)$ in the unlicensed band, from served user. Thus we can compute all $|\mathbf{A}_i| - 1$ virtual received utilities $U_i(a_{i,k}^t, a_{0,c_0}, \mathbf{a}_{-(i,0)}^t)$, $a_{i,k} \in \mathbf{A}_i / a^t{}_{i,c_i}$ as follows:

$$U_i(a_{i,k}^t) = \log_2(1 + \frac{p_i h_{i,i}^{\mathrm{U}}(1-a_{i,k}^t)}{n_0 + \sum_{j \neq i} p_j h_{j,i}^{\mathrm{U}}(1-a_{i,k}^t)}) + \alpha_i \log_2(1 + \frac{p_i h_{i,i}^{\mathrm{L}} a_{i,k}^t}{n_0 \alpha_i})(1-\lambda_0), a_{i,k} \in \mathbf{A}_i / a^t{}_{i,c_i}. \quad (21)$$

Therefore, combine with the one real sample $U_i(a_{i,c_i}^t, a_{0,c_0}, \mathbf{a}_{-(i,0)}^t)$ and we can simultaneously update all Q-values via (15). For notation convenient, we named the improved algorithm as IHQL for short. The details of IHQL are given in Algorithms 2.

Algorithm 2: Improved Hierarchical Q-learning Algorithm (IHQL)

Step1:Initialize the Q-functions $Q_i(a_{i,c_i})$ at time slot t=1 for $a_{i,c_i} \in \mathbf{A}_i$

Step2: **Learning process in the lower tier**

(1) In epoch k, the MBS chooses a price a_{0,c_0} according to its policy $\boldsymbol{\pi}_0$ and broadcasts it to all SBSs in the network.

(2) Each SBS i selects a power allocation fraction a_{i,c_i} according to its policy $\boldsymbol{\pi}_i^t$, and then SBS sends the relevant policy information to the MBS.

(3) Each SBS i computes its achieved utility $U_i(a_{i,c_i}^t, a_{0,c_0}, \mathbf{a}_{-(i,0)}^t)$ via the feedback information, and updates the estimated expected $u_i^t(a_{i,c_i})$ in (19).

(4) Each SBS i computes other $|\mathbf{A}_i|-1$ virtual samples $U_i(a_{i,k}^t, a_{0,c_0}, \mathbf{a}_{-(i,0)}^t)$, $a_{i,k} \in \mathbf{A}_i / a^t{}_{i,c_i}$ according to (21).

(5) Each SBS updates all action's Q-values according to (18) and the corresponding policy $\pi_i^t(a_{i,k})$ according to (17).

(6) All SBSs send the current policy to MBS at the end of each epoch.

Step3: The MBS computes the accrued utility $u_0^k(a_{0,c_0}, \boldsymbol{\pi}_{-0}^{kT})$. .

Step4: The MBS updates the Q-values and corresponding policy according to (15) and (17), respectively.

Step5: The MBS selects an action based on the updated policy.

Step6:Update k=k+1 and jump to $k = k_{\max}$.

4.3. Performance analysis

Along with the discussion in [39], we obtain the following differential equation describing the evolution of the Q-values:

$$\frac{dQ_i^t(a_{i,c_i}^t)}{dt} = \begin{cases} \kappa_0^t \left(u_0(a_{0,c_0}^t, \pi_{-0}^t) - Q_0^t(a_{0,c_0}^t) \right), \text{if } i = 0; \\ \kappa_i^t \left(u_i(a_{i,c_i}^t, \pi_{-i}^t) - Q_i^t(a_{i,c_i}^t) \right), \text{otherwise} \end{cases} \tag{22}$$

In the following, we would like to express the dynamics in terms of strategies rather than the Q values. Toward this end, we differentiate (17) with respect to time and use (22). We can obtain the equations as follows:

$$\frac{d\pi_i^t(a_{i,k})}{dt} = \begin{cases} \dfrac{\kappa_0^t}{\tau_0} \pi_0^t(a_{0,k}) \left\{ \left[u_0^{t-1}(a_{0,k}) - \sum_{c_0 \in \mathbf{A}_0} \pi_0^t(a_{0,c_0}) u_0^{t-1}(a_{0,c_0}) \right] - \tau_0 \sum_{c_0 \in \mathbf{A}_i} \pi_0^t \ln \dfrac{\pi_0^t(a_{0,k})}{\pi_0^t(a_{0,c_0})} \right\}; \text{if } i = 0 \\ \dfrac{\kappa_i^t}{\tau_i} \pi_i^t(a_{i,k}) \left\{ \left[u_i^{t-1}(a_{i,k}) - \sum_{c_i \in \mathbf{A}_i} \pi_i^t(a_{i,c_i}) u_i^{t-1}(a_{i,c_i}) \right] - \tau_i \sum_{c_i \in \mathbf{A}_i} \pi_i^t \ln \dfrac{\pi_i^t(a_{i,k})}{\pi_i^t(a_{i,c_i})} \right\}; \text{otherwise.} \end{cases} \tag{23}$$

The first term in braces of (23) asserts that the probability of selecting action a_{i,c_i} increases with a rate proportional to the overall efficiency of that strategy, while the second term describes the BS's tendency to randomize over possible actions. The steady state strategy profile $\pi_i^s(a_{i,k})$ can be obtained [39]

$$\pi_i^s(a_{i,k}) = \begin{cases} \dfrac{\exp\left[u_0(a_{0,k})/\tau_0\right]}{\sum_{c_0 \in \mathbf{A}_0} \exp\left[u_0(a_{0,c_0})/\tau_0\right]}; \text{if } i = 0 \\ \dfrac{\exp\left[u_i(a_{i,k})/\tau_i\right]}{\sum_{c_i \in \mathbf{A}_i} \exp\left[u_i(a_{i,c_i})/\tau_i\right]}; \text{ otherwise.} \end{cases} \tag{24}$$

Let $\mathbf{\Pi}^t = \left(\boldsymbol{\pi}_0^t, \cdots, \boldsymbol{\pi}_N^t\right)$ the strategy profile of all players at time *t*. To capture the convergence of $\mathbf{\Pi}^t$ approximately, we resort to an ordinary differential equation (ODE) [15]. The right-hand side of (22) can be represented by a function $f(\mathbf{\Pi}^t)$ as $\kappa_i^t \to 0$. $\mathbf{\Pi}^t$ will converges weakly to $\mathbf{\Pi}^* = \left(\boldsymbol{\pi}_0^*, \mathrm{NE}(\boldsymbol{\pi}_0^*)\right)$, which is the solution of $\dfrac{d\mathbf{\Pi}}{dt} = f(\mathbf{\Pi}), \mathbf{\Pi}^0 = \mathbf{\Pi}_0$, with any initial $\mathbf{\Pi}^0 = \mathbf{\Pi}_0$.

Theorem 3: The proposed algorithms can discover a mixed strategy SE.

Proof: For brevity, the convergence of the proposed HQL algorithm can be found in [40]. For the IHQL algorithm, changing more than one Q value per iteration, along with the discussion in [35], as long as the stochastic convergence condition $\sum_{t=0}^{\infty} \kappa_i^t = \infty, \sum_{t=0}^{\infty} (\kappa_i^t)^2 < \infty$ are still satisfied, the modified action-reply process (ARP) still tends to the real process in the original manner, so the convergence of IHQL can be guaranteed. Therefore, the proposed algorithms can converge to the optimal policy and discover a mixed strategy SE. This completes the proof.

5. Performance Evaluation

In this section, several numerical examples are presented to evaluate the performance of the proposed learning schemes. For simplicity, we adopt a two-tiered small cell networks with one MBS and two SBSs. Without loss of generality, the channel gains in the unlicensed bands are chosen as follows: $h_{1,1}^{\mathrm{U}} = 1$, $h_{2,2}^{\mathrm{U}} = 1$, $h_{1,2}^{\mathrm{U}} = 0.3$, $h_{2,1}^{\mathrm{U}} = 0.1$. The channel gains in licensed band are normalized as $h_{1,1}^{\mathrm{L}} = 1$, $h_{2,2}^{\mathrm{L}} = 1$. The maximum transmission power of each SBS is set as 20dBm, and noise power is normalized as 0 dBm. For each SBS, the action set of power allocation fraction is $\{0.2, 0.4, 0.6, 0.8\}$. Moreover, the MBS's price set is $\{0.2, 0.4, 0.6, 0.8\}$. Each epoch contains *T*=100 time slots. Other parameters used in simulations are set as follows: $\alpha_i = 1$, $k_{\max} = 100$.

Firstly, we show simulation results of the proposed algorithms' convergence in **Fig. 4**. **Fig. 4** shows the learning trajectories of the MBS's and SBSs' expected utilities under HQL and IHQL algorithms. We can see the curves are smoother obtained by IHQL algorithm than HQL and the reason lies in the learning process of SBSs (followers) can quickly converge to stable state (NE in lower subgame) in IHQL algorithm.

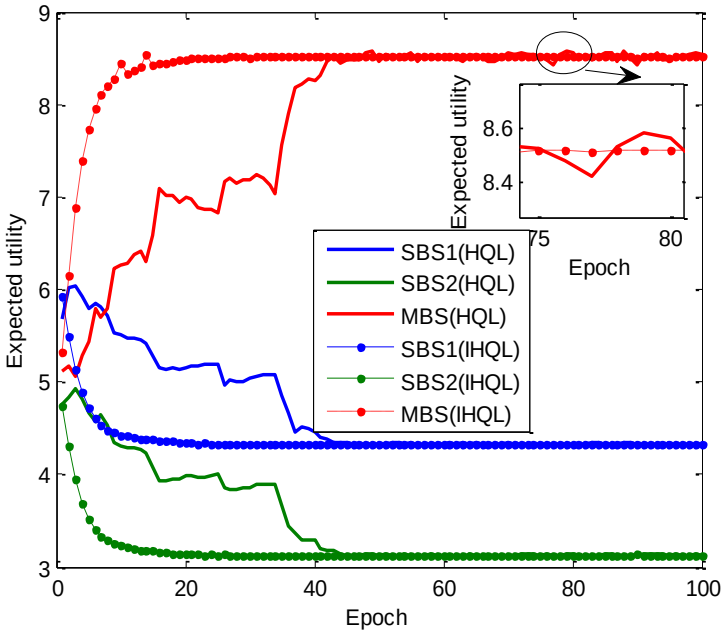


Fig. 4. One-shot learning process of SBSs and MBS.

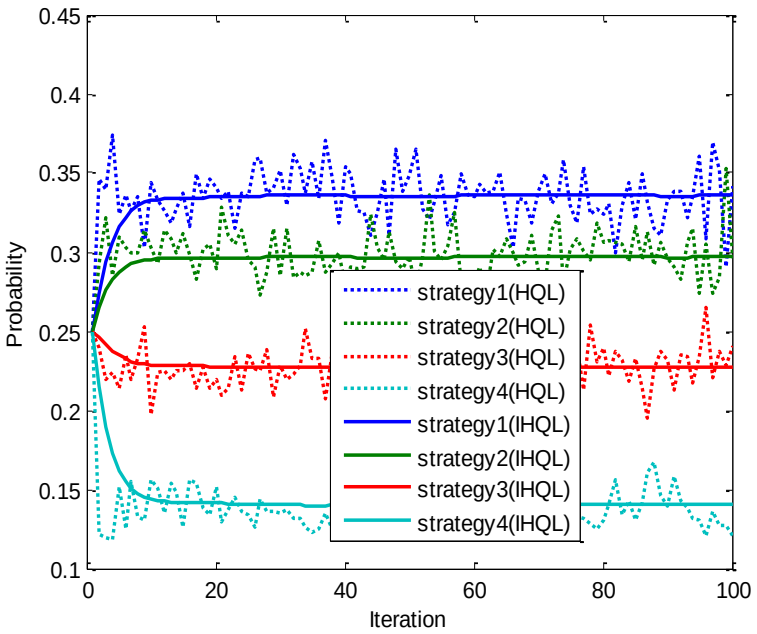


Fig. 5. Learning process of SBS1 in the first epoch.

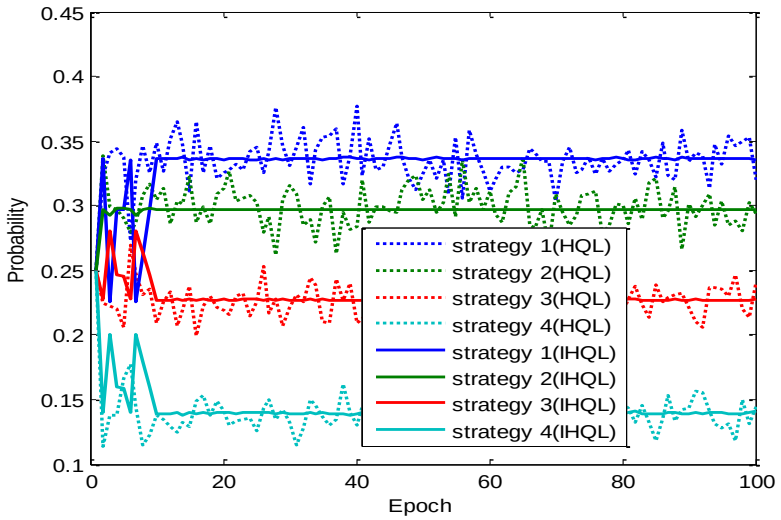


Fig. 6. Learning process of SBS1 over epochs.

Then, in **Fig. 5**, we present the SBS1's action selection probability update in the first epoch. Moreover, in **Fig. 6**, we give the SBS1's action selection probability update over the 100 epochs, which validates that IHQL can speed up the SBSs' learning process. **Fig. 6** shows the SBS 1's learning process converges to a mixed strategy which maximizes its expected utility.

In the following, let $h_{1,1}^{\mathrm{L}} = h_{2,2}^{\mathrm{L}} = h$ and we show the learning process of MBS with various h under IHQL scheme in **Fig. 7.** The value of h reflects the relative channel gain compared to the unlicensed band. We can see that MBS's expected utility in stable state increase with the growth of the h in **Fig. 7.** The reason lies in that the better channel condition in licensed band can attract SBS to allocate more traffic load to primary networks, thus it brings more profit to MBS. On the contrary, when then channel gain in licensed band is unsatisfactory to SBSs (such as h=0.1 in **Fig. 7**), then they prefer to allocate more traffic load to shared spectrum bands which are free of charge.

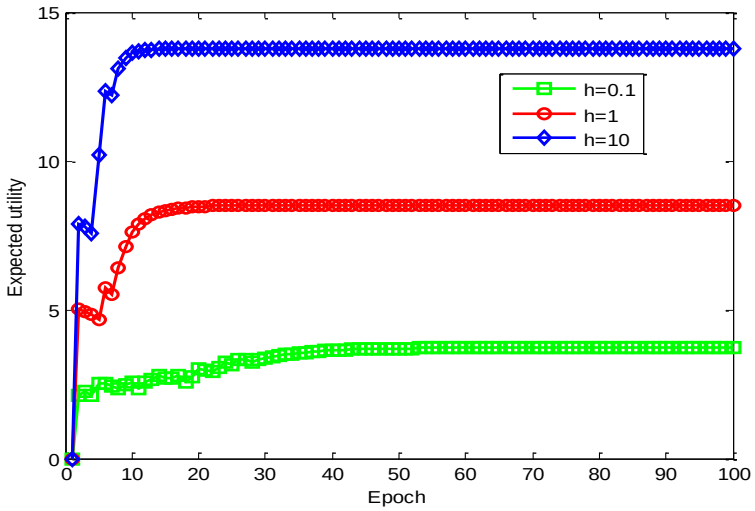


Fig. 7. Learning process of MBS over epochs with various h

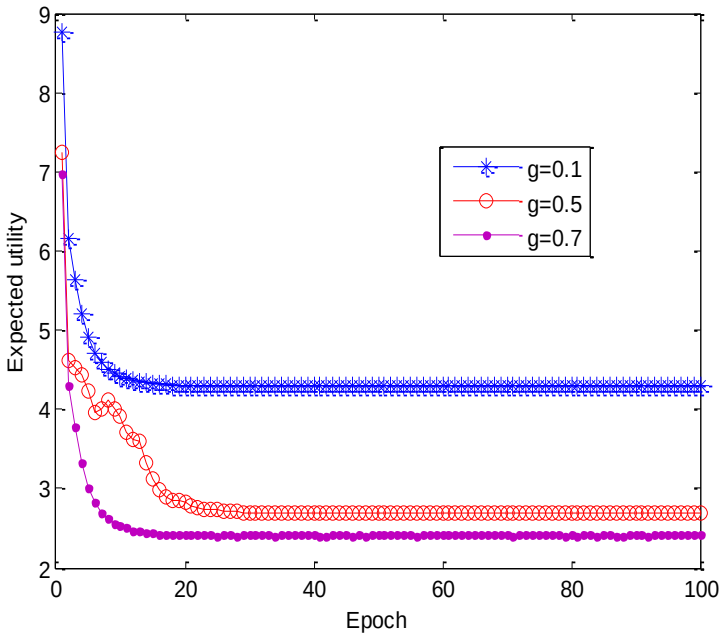


Fig. 8. Learning process of SBSs over epochs with various g

In **Fig. 8**, we show the impacts of cross-channel gain on the SBSs' learning process under IHQL. Firstly, let $h_{1,2}^{\mathrm{U}} = h_{2,1}^{\mathrm{U}} = g$ for simplicity (note that the SBS1's and SBS2's learning process are the same, so we just plot one curve corresponding to a given g). The larger g means the severer mutual interference among SBSs. It is observed that the SBSs' expected utilities in stable state increase with the decrease of g. The result is intuitive and it implies the traffic offloading of each SBS is coupled by its nearby neighbors because of the co-tier interference.

6. Conclusion

In this paper, we investigate the traffic offloading in two-tier small cell networks over unlicensed bands. The unit traffic flow pricing mechanism is adopted by macro base station and the small base station can autonomously make its traffic offloading decision. To capture the self-optimizing abilities of SBSs and MBS, we propose a hierarchical learning framework based on Q-learning mechanism. In the proposed framework, the MBS is a leader and moves first, then SBSs are modeled as non-cooperative followers and move subsequently, who only can communicate with MBS. At last, a hierarchical Q-learning algorithm is proposed to discover the SE and an improved hierarchical Q-learning algorithm is presented to speed up the followers' learning process. Simulation results validate the effectiveness of the proposed algorithms.

References

[1] J. G. Andrews, H. Claussen, M. Dohler, et al. "Femtocells: past, present, and future," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 497-508, Apr. 2012. Article (CrossRef Link)

[2] J. G. Andrews, S. Buzzi, W. Choi, et al. "What will 5G be?," *IEEE J. Sel. Areas Commun.*, vol. 32, no. 6, pp. 1065-1082, Jun. 2014. Article (CrossRef Link)

[3] Y. Xu, J. Wang, Q. Wu, et al. "A game theoretic perspective on self-organizing optimization for cognitive small cells," *IEEE Commun. Mag.*, vol. 53, no. 7, pp. 100-108, Jul. 2015. Article (CrossRef Link)

[4] M. Jo, T. Maksymyuk, B. Strykhalyuk, et al. "Device-to-Device Based Heterogeneous Radio Access Network Architecture for Mobile Cloud Computing," *IEEE Wireless Commun.*, vol. 22, no.3 , pp. 50-58, Jun. 2015. Article (CrossRef Link)

[5] Z. Du, Q. Wu, P. Yang, et al. "Exploiting user demand diversity in heterogeneous wireless networks," *IEEE Trans. Wireless Commun.*, vol. 14, no. 8, pp. 4142-4155, Aug. 2015. Article (CrossRef Link)

[6] Z. Du, Q. Wu, P. Yang, et al. "Dynamic user demand driven online network selection," *IEEE Trans. Veh. Tech.* vol. 18, no. 3, pp. 1089-7798. Mar. 2014. Article (CrossRef Link)

[7] F. Liu, E. Bala, E. Erkip, et al. "Small cell traffic balancing over licensed and unlicensed Bands,"*IEEE Trans. Veh. Tech.*, to appear. Article (CrossRef Link)

[8] M. Bennis, M. Simsek, A. Czylwik, et al. "When cellular meets WiFi in wireless small cell networks," *IEEE Commun. Mag.*, vol. 51, no. 6, pp. 44-50. Jun. 2013. Article (CrossRef Link)

[9] Qualcomm. "LTE in unlicensed spectrum: harmonious coexistence with Wi-Fi," *white paper*, Jun. 2014. Article (CrossRef Link)

[10] Huawei, "U-LTE: unlicensed spectrum utilization of LTE," *white paper*, Feb. 2014. Article (CrossRef Link)

[11] M. Mustonen, T. Chen, H. Saarnisaari, et al, "Cellular architecture enhancement for supporting the european licensed shared access concept," *IEEE Wireless Mag.*, vol. 21, no. 3, pp. 37-43. Jun. 2014. Article (CrossRef Link)

[12] Federal Communications Commission, Unlicensed operation in the TV broadcast bands, 47 CFR Part 15, Federal Register, vol. 74, no. 30, Feb. 2009.Article (CrossRef Link)

[13] M. Iwamura, K. Etemad, M.-H. Fong, R. Nory, and R. Love, "Carrier aggregation framework in 3GPP LTE-advanced [WiMAX/LTE update]," *IEEE Commun. Mag.*, vol. 48, no. 8, pp. 60-67, Aug. 2010. Article (CrossRef Link)

[14] A. al-Dulaimi, S. al-rubaye, Q. Ni, et al. "Pursuit of more capacity triggers LTE in unlicensed band," *IEEE vehicular Technology Mag.*, vol. 10, no. 1, pp. 43-51. 2015. Article (CrossRef Link)

[15] 3GPP-TSG-RAN-Meetin65, "Study on licensed assisted access using LTE," Tech. Rep. RP141664, Sep. 2014. Article (CrossRef Link)

[16] L. Zhou, M. Chen, Y. Qian, et al. "Fairness Resource Allocation in Blind Wireless Multimedia Communications," *IEEE Transactions on Multimedia*, vol. 15, no. 4, pp. 946-956, Jun. 2013. Article (CrossRef Link)

[17] L. Zhou, Z. Yang, Y. Wen, et al. "Resource Allocation with Incomplete Information for QoE-driven Multimedia Communications," *IEEE Transactions on Wireless Communications*, vol. 12, no. 8, pp. 3733-3745, Aug. 2013. Article (CrossRef Link)

[18] Y. Xu, A. Anpalagan, Q. Wu, et al. "Decision-theoretic distributed channel selection for opportunistic spectrum access: Strategies, challenges and solutions," *IEEE Commun. Surveys and Tutorials*, vol. 15, no. 4, pp. 1689-1713, Fourth Quarter, 2013. Article (CrossRef Link)

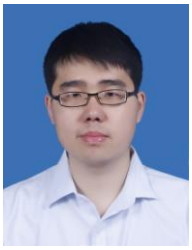
[19] Y. Xu**,** J. Wang, Q. Wu, et al. "Opportunistic spectrum access in cognitive radio networks: Global optimization using local interaction games," *IEEE Journal of Selected Topics in Signal Processing*, vol. 6, no. 2, pp. 180-194, Apr. 2012 .Article (CrossRef Link)

[20] Y. Xu, J. Wang, Q. Wu, et al. "Opportunistic spectrum access in unknown dynamic environment: A game-theoretic stochastic learning solution," *IEEE Transactions on Wireless Communications*, vol. 11, no. 4, pp. 1380-1391, Apr. 2012. Article (CrossRef Link)

[21] X. Kang, R. Zhang, and M. Motani, "Price-based resource allocation for spectrum-sharing femtocell networks: a Stackelberg game approach," *IEEE J. Sel. Areas Commun.*, vol. 30, no. 3, pp. 538-549, Apr. 2012. Article (CrossRef Link)
[22] S. Guruacharya, D. Niyato, D. I. Kim, and E. Hossain,"Hierarchical competition for downlink power allocation in OFDMA femtocell networks," *IEEE Trans. Wireless Commun.*, vol. 12, no. 4, pp. 1543-1553, Apr. 2013. Article (CrossRef Link)
[23] Q. Han, B. Yang, X. Wang, et al. "Hierarchical-game-based uplink power control in femtocell networks," *IEEE Trans. Veh. Technol.*, vol. 63, no. 6, pp. 2819-2835, Jul. 2014. Article (CrossRef Link)
[24] S. Bu, F. R. Yu, Y. Cai and H. Yanikomeroglu. "Interference-aware energy-effcient resource allocation for OFDMA-based heterogeneous networks with incomplete channel state information." *IEEE Trans. Veh. Technol.*, vol. 64, no. 3, pp. 1036-1050, Jun. 2014. Article (CrossRef Link)
[25] Y. Sun, H. Shao, J. Qiu, et al. "Capacity offloading in two-tier small cell networks over unlicensed band: a hierarchical learning framework," in *Proc. of the IEEE International Conference on Wireless Communications and Signal Processing(WCSP 15)*, Nanjing, Oct. 2015, to appear. Article (CrossRef Link)
[26] Y. Chiang and W. Liao. "Genie: An optimal green policy for energy saving and traffic offloading in heterogeneous cellular networks," in *Proc. of the IEEE International Conference on Communications (ICC)*, Budapest , Jun. 2013, pp. 6230–6234.Article (CrossRef Link)
[27] X. Chen, J. Wu, Y. Cai et al. "Energy-efficiency oriented traffic offloading in wireless networks: a brief survey and a learning approach for heterogeneous cellular networks," *IEEE J. Sel. Areas Commun.*, vol. 33, no. 4, pp. 627-640, Apr. 2015.Article (CrossRef Link)
[28] X. Kang, Y. Chia, S. Sun et al. "Mobile data offloading through a third-party WiFi access point: An operator's perspective," *IEEE Trans. Wireless Commun.*, vol. 13, no. 10, pp. 5340–5351, Oct. 2014.Article (CrossRef Link)
[29] L. Gao, G. Iosifidis, J. Huang, and L. Tassiulas, "Economics of mobile data offloading," in *Proc. of IEEE INFOCOM SDP Workshop*, Turin, Italy, Apr. 2013, pp. 3303-3308. Article (CrossRef Link)
[30] M. Bennis, S. M. Perlaza, P. Blasco, et al. "Self-organization in small cell networks: a reinforcement learning approach," *IEEE Trans. Wireless Commun.*, vol. 12, no. 7, pp. 3202-3212, Jul. 2013.Article (CrossRef Link)
[31] F. Zhang, W. Zhang and Q. Ling. "Non-cooperative game for capacity offload." *IEEE Trans. Wireless Commun.*, vol. 11, no. 4, pp. 1565-1575, Apr. 2012. Article (CrossRef Link)
[32] R. Mahindra, H. Viswanathan, K. Sundaresan, et al. "A practical traffic management system for integrated LTE-WiFi networks," in *Proc. of the 20th annual international conference on Mobile computing and networking(Mobicom 14)*, Hawaii,USA , Sep. 2014, pp. 189-200. Article (CrossRef Link)
[33] J.B. Rosen. "Existence and uniqueness of equilibrium points for concave *N*-person games," *Ecnometrica*, vol. 33, no. 3, pp. 520-534, Jul. 1965. Article (CrossRef Link)
[34] D. Fudenburg and J. Tirole. "Game theory," *The MIT press*, 1991. Article (CrossRef Link)
[35] C. J. C. H. Watkins and P. Dayan. "Q-learning," *Mach. Learn.*, vol. 8, pp. 279-292, 1992. Article (CrossRef Link)
[36] Z. Du, Q. Wu and P. Yang. "Dynamic user demand driven online network selection," *IEEE Commun. Letters*, vol. 18, no. 3, pp. 419-422, Mar. 2014.Article (CrossRef Link)
[37] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*. Cambridge, MA, USA: MIT Press, 1998.Article (CrossRef Link)
[38] X. Chen, H. Zhang, T. Chen, et al. "Improving energy efficiency in femtocell networks: a hierarchical reinforcement learning framework," in *Proc. of the IEEE International Conference on Communications (ICC)*, Budapest, Jun. 2013, pp. 2241-2245.Article (CrossRef Link)
[39] A. Kianercy and A. Galstyan. "Dynamics of Boltzmann Q-learning in two-player two-action games," *Physical Review E*, vol. 85, no. 4, pp.1-10, Apr. 2012.Article (CrossRef Link)
[40] P. S. Sastry, V. V. Phansalkar, and M. Thathachar, "Decentralized learning of nash equilibria in

multi-person stochastic games with incomplete information," *IEEE Trans. on Systems, Man and Cybernetics.*, vol. 24, no. 5, pp. 769-777, May. 1994. Article (CrossRef Link)

Youming Sun received his B.S. degree in electronic and information engineering from Yanshan University, Qinhuangdao, China, in 2010 and M.S. degree from National Digital Switching System Engineering & Technological Research Center (NDSC), Zhengzhou, China, in 2013, respectively. He is currently a Ph.D. candidate in NDSC. His research interests include resource allocation in small cell networks, game theory and statistical learning.



Hongxiang Shao received the B.S. from Southwest University Of Science and Technology, Sichuan, China in 2007,and M.S. from Henan University Of Science and Technology, Henan, china in 2014. He is currently working toward the PH.D. degree in communications and information system in Institute of Communications Engineering, PLA University of Science and Technology. His research interests focus on opportunistic spectrum access, learning theory, game theory, and optimization techniques.

Xin Liu received his B.S. degree in communications engineering, M.S. degree and Ph.D. degree in communications and information system from institute of Communications Engineering, PLA University of Science and Technology, Nanjing, China, in 2004, 2007, 2011, respectively. His current research interests focus on cognitive radio, resource allocation, distributed optimization and game theory.

Jian Zhang received his Ph.D. degree in Zhengzhou Information Science and Technology Institute, Zhengzhou, China in 2007. He is currently a Full Professor in National Digital Switching System Engineering & Technological Research Center (NDSC), Zhengzhou, China. His current research interests focus on visible light communication, statistical signal processing and distributed optimization.

Junfei Qiu received his B.S. degree in electronic and information engineering from Wuhan University of Science and Technology, Wuhan, China, in 2013. He is currently pursuing the M.S. degree in communications and information system in College of Communications Engineering, PLA University of Science and Technology, Nanjing, China. His research interests include machine learning, statistical signal processing, big data analytics over wireless networks, game theory and wireless communication

Yuhua Xu received his B.S. degree in Communications Engineering, and Ph.D. degree in Communications and Information Systems from College of Communications Engineering, PLA University of Science and Technology, in 2006 and 2014 respectively. He has been with College of Communications Engineering, PLA University of Science and Technology since 2012, and currently as an Assistant Professor. His research interests focus on opportunistic spectrum access, learning theory, game theory, and distributed optimization techniques for wireless communications. He has published several papers in international conferences and reputed journals in his research area. He is an Editor for the KSII Transactions on Internet and Information Systems. In 2011 and 2012, he was awarded Certificate of Appreciation as Exemplary Reviewer for the IEEE Communications Letters.