Received ; | Revised ; | Accepted

DOI: xxx/xxxx

ARTICLE TYPE

Reinforcement learning for high performance computing in heterogeneous networks †

Amalorpava Mary Rajee S[1] | Salman A.AlQahtani[2] | Ahila A[3] | Deema Mohammed alsekait[4]

[1]Associate Professor, Mallareddy Engineering College for Women, Secundrabad,Telengana, India
[2]Research Chair of New Emerging Technologies and 5G Networks and Beyond,Computer Engineering Department , College of computer and Information Sciences,King Saud University, State name, Country name
[3]Post doc researcher, Indian Institute of Technology,Madras, Tamilnadu, India
[4]Department of Computer Science and Information Technology, Apllied College,Princess Nourah bint Abdulrahman University, ,

Correspondence
*Ahila A, Email: ahilaamarnath27@gmail.com

Present Address
This is sample for present address text this is sample for present address text

Summary

Next-generation networks powered by millimetre wave (mmW) technology offer improved data rates in the range of a few Gigabits per second. Heterogeneous networks (HetNets) with mmWave capability must be reliable, adaptable, and energy-efficient. It is suggested to maximise the transmission power of small cells to get better performance. This leads to the use of intelligent techniques namely, Q-learning approach that will allow seamless connectivity. A higher degree of automation, cooperation, and intelligence in distributed HetNets are applied using Q-learning with Markov decision process (MDP) based technique. The first objective is to perform cooperative online learning scheme to allocate power based on MDP in the Two-tier HetNet. Maximizing energy efficiency through effective power distribution is the second objective. The suggested technique improves the network's overall energy efficiency and capacity by optimizing judicious power usage based on the probability of an MDP state transition. Cooperative learning techniques and the right Markov state models can enhance both macrocell and femtocell service, enhancing user experience.

KEYWORDS:
Heterogeneous Network, Markov decision process, Q-Learning

1 | INTRODUCTION

It is a modern trend to use a multi-tier network structure with differences in size, transmitted powers, and exceptional smart wireless devices to maximize the network capacity. In the context of the next-generation network, femtocells are the most promising indoor base stations that have gained considerable attention in recent research. The studies consider macrocell overlaid with femtocells and control of an autonomous power distribution. This heterogeneity structure is expected to improve key network features such as capacity and energy efficiency. For decades, numerical optimization has been playing an important role in addressing power control problems.

In this paper, It is considered mmWave enabled HetNet operating at 28GHZ to address the problem of dynamic power control based on MDP. The major determinants enabling the performance of a power allocation strategy employing the Value Iteration Algorithm in MDP are dynamic channel load and connection quality [1]. In order to improve the Quality of service, Base Stations (BS) transmit power is optimised using the machine learning approach known as Q-Learning (QoS). Users' transmission rates are improved via cooperative Q-learning, and a predetermined threshold is set [2]. The closeness of multiple base stations (BSs) to the associated user allows for the creation of a fair power distribution method. The reward mechanism is based on the proximity of BSs in the network [3]. One common approach to manage the interference is to learn from the dynamic environment formed by the coexistence of multi-agents such as femtocells and macrocells. Their power transmission level can be adjusted during learning from the environment; Reinforcement learning technique called Q-learning is a common tool to achieve this learning, and has been widely used in [4]. An advantage of Q-learning is that no prior knowledge of the environmental state is needed. This can be utilized in HetNet

† Reinforcement learning for high performance computing in heterogeneous networks
[0]**Abbreviations:** ANA, anti-nuclear antibodies; APC, antigen-presenting cells; IRF, interferon regulatory factor

regardless of the number or spatial locations of femtocells. Moreover, the learning can be independent or cooperative [5] where the femtocells share information with each other and the decision is made on the basis of MDPs. This is a classical formalization of decision making in sequential order, actions are not just immediate rewards, but also subsequent situations, or states, and thus future reward which can be utilized for mobile users to infer decision making under unknown network conditions such as channel availability and resource allocation.[6]. The authors in [7,8] proposed cooperative learning among macro cells and femto cells to improve user QoS while keeping the transmission rate above the threshold. This is a traditional formalisation of sequential decision-making, where actions include not just immediate rewards but also following situations or states. As a self-organizing mechanism to enable power adaptation utilising MDP, Roohollah Amiri et al. offer a Q-learning based distributed power allocation algorithm (Q-DPA) [9]. Singular value decomposition and Q-Learning optimization were used by H. Sun et al. to create their difficult real-time implementation [10]. In the future generation mmWave enabled networks, blockages have heavy impact on designing learning outcomes [11]. Reward based functions are used in HetNets to limit the macrocell capacity and increase the power consumption [12-14].

1.1 | Power Saving Techniques - Overview

A more general view of the existing strategies that researchers and vendors adopt to save power is as follows,

Intelligent management of network

The current cellular network adopts sleep mode techniques which is from “always switched on" to “always make available" design topology. This approach monitors the traffic load and decides when to switch on and when to switch off the network elements. Through this an improved energy performance can be achieved in access nodes. The overall network power consumption has reduced globally by adopting energy saving modes and sleep mode strategy. The existing energy saving modes and sleep mode strategies utilize queuing theory or information theory to measure steady-state parameters such as blocking or dropping rate, throughput and traffic load. The functionalities of these two theories are dynamically activated and the total energy budget focuses on the hardware parts to shut down or to enable active networks. (Jingjin Wu et al. 2015)

Energy-saving resource allocation

The improved energy efficiency is facilitated through radio resource allocation strategy. The goal is to reduce the overall energy consumption of UE products while maintaining key performance metrics like BER, QoS and Maximum Received power. This could be achieved through complex problems such as Virtualization techniques, fractional programming and so on. The performance of network components such as server and host workload are continuously monitored and through that energy utilization is finalized.(Abdul Hameed et al. 2014)

User-centric Approach

Traditional energy efficient network is based on a User oriented approach where the system monitors the on-going user data transactions and accordingly broadcasts necessary control information to the host to provide access within the cell area coverage. The objective is to design on the dynamic optimization of transmitted power in uplink and downlink using beam forming techniques, Cell Zooming techniques and so on. An optimization problem based on linear programming and fractional programming are adopted to perform resource-aware energy-saving techniques that minimize the power consumption.

In this study, a new approach based on Q-learning with MDP decision-making algorithm [15] is proposed to provide better capacity and energy efficiency performance throughout the HetNet. In the studies cited above, the design of reward function affects the overall sum capacity, The key novelty in this proposed work is Energy efficient reward function. The aim is to let the FUEs achieve as high energy efficiency as possible without subjecting the MUE to lower energy efficiency to overcome this challenge.
The main contributions of this work:

1. This paper proposes a cooperative Q learning approach that adapts MDP decision-making algorithm to make the power optimization solution.
2. A new energy efficient reward function is developed to guarantee the minimum threshold rate per user and energy efficiency for each femto user and the network.
3. A power optimization analysis is carried out to train the agent to explore the environment to solve power allocation problem in Two Tier HetNet.Figure 1 depicts a two-tier heterogeneous network.

The remainder of the work is structured as follows. The key ideas of the suggested system model are introduced in Section II. The issue of power optimization is covered in Section III. Performance findings and their analysis are presented in Section IV. The paper is concluded in Section V.

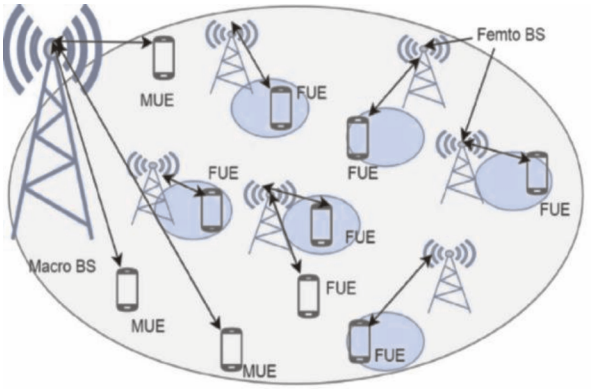

FIGURE 1 An Illustration of Two-Tier Heterogeneous Network

2 | SYSTEM MODEL

2.1 | An Illustration of Hetnet

A HetNet is a grouping of k-tier BSs that can be identified by their transmit strengths and densities. The primary and secondary tiers of the Hetnet system are two tiers. High power macro cells that service macro user equipment (MUEs) are found in the primary layer, and femtocells that serve femto user equipment are found in the second tier (FUEs).

A two-tier heterogeneous downlink cellular network is modeled on the Euclidean plane. It comprises k femtocells and one macrocell. For each of the cells, there is an associated user. This HetNet simplified model impact the accuracy of the learning algorithm.

mm-wave spectrum is the operated frequency of all users. However, due to the static/dynamic blockages, femtocells are allocated to the indoor user with line of sight (LOS) link and macrocells are allocated to the outdoor user. Cell selection is based on the Markovian state transition probability.

This work presents a simple two-state Markovian wireless channel model for capturing LOS link. A femto user's (FUE) channel is modelled into LOS state and Non-LOS (NLOS) state in this case due to the external circumstances. A simplified version of the markovian channel model is offered for these cases, with two alternate states for each markovian channel connection. When a direct link is available, then the state is said to be in LOS; when a direct link is not available, then the state is said to be in NLOS;

Based on the properties of Markovian processes, a channel state transition probability matrix is given by

$$q^n = \begin{bmatrix} q_{00}^n & q_{01}^n \\ q_{10}^n & q_{11}^n \end{bmatrix} = \begin{bmatrix} q_{00} & q_{01} \\ q_{10} & q_{11} \end{bmatrix}^n \tag{1}$$

Where q_{01}^n is a one-step transition probability from the state 0 = LOS into the state 1 = NLOS after n steps transition. The distance dependent path loss model is given as

$$L_{mm}(r) = \begin{cases} Cr^{-\alpha_L} \text{if LOS } \alpha_L = 2 \\ Cr^{-\alpha_N} \text{if NLOS } \alpha_N = 4 \end{cases} \tag{2}$$

The mmWave link's distance-dependent path loss is shown as $L_{mm}(r)$. Which is the function of the radial distance r. $C = \left(\frac{\lambda}{4\pi}\right)^2$ and λ is the wavelength. α_L and α_N are the exponents of the path loss in LOS and NLOS mmWave links.

The signal to interference noise ratio (SINR) is calculated using a mathematical relationship for each of the users and also compute the capacity of the macrocell base stations

(MBSs) and the femtocell base stations (FBSs) in the system[16]. The instantaneous SINR of FUE i is associated with its designated FBS and is defined as:

$$SINR_{FUE_i} = \gamma_i = \frac{P_i g_{FBS_i,FUE_i} L_{mm}(r)}{P_{MBS} g_{MBS,FUE_i} + \sum_{j=1}^{k} P_j g_{FBSj,FUE_i} + \sigma^2} \tag{3}$$

Where P_i indicates the i^{th} transmission power of FUE, g_{FBS_i,FUE_i} indicates the channel gain between i^{th} FUE and designated FBS. $P_{MBS} g_{MBS,FUE_i} + \sum_{j=1}^{k} P_j g_{FBSj,FUE_i}$ denotes the set of MUEs and FUEs that are interference from the same sub channel. σ^2 denotes the Additive white Gaussian noise(AWGN). The instantaneous SINR of MUE associated with its designated MBS is defined as:

$$SINR_{MUE} = \gamma_m = \frac{P_{MBS} g_{MBS,MUE} L_{mm}(r)}{\sum_{i=1}^{k} P_i g_{FBSi,MUE} + \sigma^2} \tag{4}$$

Where P_{MBS} indicates the transmission power of MBS, $g_{MBS,MUE}$ indicates the channel gain between MUE and designated MBS, The denominator notations are the same as defined in (3).

The formulas used to determine the femtocell, macrocell, and overall system capacities are:

$$C_{FUE_i} = \log_2(1 + SINR_{FUE_i}), i = 1, 2, 3, ...k \tag{5}$$

$$C_{MUE} = \log_2(1 + SINR_{MUE}) \tag{6}$$

The maximum achievable capacity for the HetNet is given by

$$C_{system} = BW(C_{MUE} + C_{FUE_i}), i = 1, 2, 3, ...k \tag{7}$$

The bandwidth of the system BW is multiplied with the individual macro and femto users. The two-tier HetNet energy Efficiency model is created by using the maximum capacity that can be reached.

$$\begin{aligned} EE_{eff} &= \frac{\text{Maximum achievable capacity}}{\text{Average Network Power Consumption}} \\ &= \frac{C_{system}}{P_{MBS} + \sum_{i=1}^{k} P_i} \end{aligned} \tag{8}$$

In which P_{MBS} and P_i is the transmission power at MBS and FBS. The main objective of these parameters is to guarantee, that MUE has a certain Quality-of-Service (QoS) requirement that is above a defined threshold at every time instant. Another goal is to ensure that FUEs equally have a reasonable QoS. This criterion will be discussed further in section 3 where we discuss the concept of “rewards" in Q-learning.

3 | POWER OPTIMIZATION ANALYSIS

3.1 | Q-Learning

One or more agents and an environment make up the Q-learning issue. It is used to engage in trial-and-error interactions with the dynamic environment. Decisions are formed through interactions, and this process is modelled as an MDP. The agent can choose from a variety of actions to transition between different states using state transition probability. The agent's trail of actions determines the policy. The dynamic environment will reward the agent with each transition; as a result of activity, a cumulative reward is stored. The agent will carry on acting to earn more rewards.Q learning is illustared in Figure 2. FUE and MUE, which are taken to be agents, can interact with the dynamic environment to determine the optimum decision policy through trial and error.

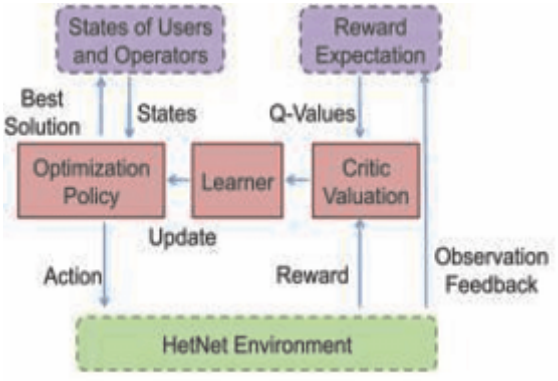

FIGURE 2 Q-Learning Illustration

Q learning can be used as a tuple $(S, A, P_{s,s*}, R(s, a))$where $S = (s_1, s_2, ...s_n)$ and $A = (a_1, a_2, ...a_i)$ are the finite set of environment states and actions, The state transition probability function is denoted by $P_{s,s*}(a)$ the agents are oscillated from state s to the new state $s*$ after taking action a. The reward function is denoted as $R(s, a)$ and given to the agents based on action a in state s. The optimal policy $\pi * (s, a)$ discovered to find an expected discounted reward in order to maximize the total reward. Hence the agents are required to find an optimal policy [17]. The

expected discounted reward over infinite time can be written as

$$Q(s,a) = E\left\{\sum_{t=0}^{\infty}\gamma^t R(s_t, \pi(s_t), s_{t+1}) | s_0 = s\right\} \tag{9}$$

The significance of future rewards is determined using the discount factor $\gamma\varepsilon[0,1]$. The agent will ignore learning or put emphasis on future rewards based on the value of $\gamma = 0$ or $\gamma = 0.9$.

$Q(s, a)$ is a Q value or the current state. The Q value (expected discounted reward) is maximized if an agent selects an optimal policy $\pi * (s, a)$. An optimal Q value $Q^*(s, a)$ is defined as:

$$Q(s,a) = E\{R(s,a)\} + \gamma\sum_{t=0}^{\infty} P_{s,s'}(a)\max_{b\in A} Q^*(s',b) \tag{10}$$

Power optimization problem through co-operative learning in the HetNet structure is defined as follows,

$$\begin{aligned} &\max\ \mathrm{EE_{eff}} \\ &\text{Subject to} \\ &\mathrm{C.1}: \mathrm{C_{FUE_i}}\gamma_{\mathrm{T}}^{*}\ \mathrm{i} = 1, 2, \ldots \mathrm{k} \\ &\mathrm{C.2}: \mathrm{C_{MUE}}\gamma_{\mathrm{T}}^{*} \end{aligned} \tag{11}$$

If the SINR of the femto and macro tier defined in equations 5 and 6 is met, then C.1 and C.2 are provided to ensure that the SINR of the i^{th} FUE and MUE does not fall below the predefined levels.

Figure 3 illustrates one FUE decision process modeled as MDP with three states and one possible action associated with each state. It shows a simplified example of how the interaction process of the FUE can be modeled as MDP. However, in practice, there are multiple agents and they have multiple states and actions.

Agents: The learning agents are the FUEs and MUE associated with the only MBS in the considered channel.

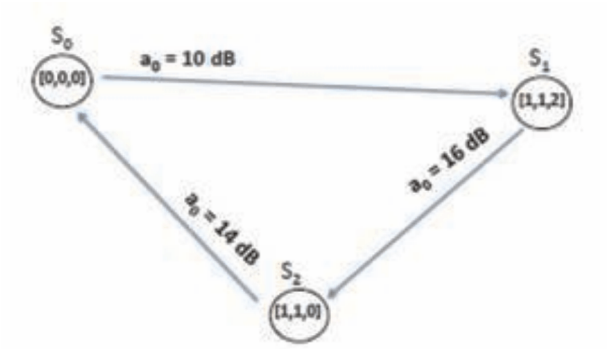

FIGURE 3 The MDP for one FUE with three states and one possible action

Actions: The actions of the learning agents are predefined transmission power levels. For illustration purposes, The actions with uniform step size are considered, 2dB,4dB,.....18dB with step size of 2dB.

States: Basically, there are 12 possible states for each FUE and 6 possible states for each MUE. The different state variables of $S_0(0\ 0\ 0)$ through $S_{11}(1\ 1\ 2)$ and $S_0(0\ 0)$ to $S_5(1\ 2)$ are given in the table 1.

The state variable for FUE and MUE at a particular time step is represented as a tuple of indicators: $S_i^{FUE} = \{I_{\gamma_m}, I_{\gamma_i}, I_r\}$ and $S_i^{MUE} = \{I_{\gamma_m}, I_r\}$. Here I_{γ_m} represents the instantaneous SINR condition of the MUE and I_{γ_i} indicates the SINR condition of FUE i. $I_r = \gamma_i/p_i$ is the SINR and energy ratio of FUE i, used to measure the energy efficiency of the FUE i. These three indicators are defined in equation (15). $S_1(1\ 1\ 2) - S_1$ values show whether or not the FUE is supported by its minimum SINR. Here $I_{\gamma_m} = 1, I_{\gamma_i} = 1, I_r = \gamma_i/p_i = 2$

$S_4(11)$ The state variables of S_4 values show whether or not the MUE is supported by its minimum SINR. Here

$I_{\gamma_m} = 1, I_r = \gamma_i/p_i = 1$

FUE i states consider not only the SINR of the FUE i and the energy efficiency of the FUE i but also the SINR of the MUE.

TABLE 1 Possible States for the FUE and MUE

FUE STATES	MUE STATES
$S_0(000)$	$S_0(00)$
$S_1(001)$	$S_1(01)$
$S_2(002)$	$S_2(02)$
$S_3(100)$	$S_3(10)$
$S_4(101)$	$S_4(11)$
$S_5(102)$	$S_5(12)$
$S_6(010)$	$S_7(011)$
$S_8(012)$	$S_9(110)$
$S_{10}(111)$	$S_{11}(112)$

This is because our aim is to let the FUEs achieve as high efficiency as possible and protect the MUE at the same time, However, the MUE only needs to care about its own SINR and try to achieve high energy efficiency.

4 | MDP ALGORITHM

1. Initialize $S_i(0)$ arbitrarily and $Q_i(s, a) = 0 \forall\, s, a$
2. **for all** episodes $t = 1, 2, \ldots$ **do**
3. Perform action a^t at according to policy $\pi * (s, a)$
4. Observe Reward r_i^m, r_i^f
5. If C.1 and C.2 of equation (13) are satisfied

$$r_i^m = \left\{ \begin{array}{c} 100 \text{ if } \gamma_m \gamma_T \\ -1 \text{ otherwise} \end{array} \right\}$$

$$r_i^f = \left\{ \begin{array}{c} 100 \text{ if } \frac{\gamma_m}{\gamma_T} 1, \frac{\gamma_i}{\gamma_T} 1 \\ -1 \text{ if } \frac{\gamma_m}{\gamma_T} 1, \frac{\gamma_i}{\gamma_T} < 1 \\ -1 \text{ if } \frac{\gamma_m}{\gamma_T} < 1, \frac{\gamma_i}{\gamma_T} < 1 \end{array} \right\}$$

   else

   $r_i^m = 0$ and $r_i^f = 0$

6. Update $Qi(s, a)$ by using Q-learning for all episodes for each agent i.
7. Update t until the total number of episodes are completed, otherwise, return to Step 2.

   end for

$$I_{\gamma_m} = \left\{ \begin{array}{ll} 1 & \text{if } \gamma_m \gamma_T \\ 0 & \text{otherwise} \end{array} \right\} I_{\gamma_i} = \left\{ \begin{array}{ll} 1 & \text{if } \gamma_{\gamma_i} \gamma_T \\ 0 & \text{otherwise} \end{array} \right\}$$

$$I_r = \left\{ \begin{array}{ll} 0 & \text{if } \gamma_i / p_i t_a \\ 1 & \text{if } t_a < \gamma_i / p_i t_b \\ 2 & f \gamma_i / p_i t_b \end{array} \right\}$$

here I_{γ_m} is the threshold SINR for reliable communication, t_a and t_b are the two predefined threshold values. As a result of network feedback, a reward is received, which is an action selected at state x_t.

- All the agents choose a random action simultaneously.
- Each agent learns the actions of other agents and uses this information to estimate its instantaneous SINR based on Equation 5 for the FUE and based on Equation 6 for the MUE.

- The SINR and the predefined threshold values determine the next state of the agent.

Let's assume the agent FUEi is at initial random state $S_0[0,0,0]$ and chooses a random action a^t which corresponds to a predefined transmission power level of 10dB (see Figure 3). Since the other agents simultaneously choose their respective random actions, the agent FUE_i can now estimate its SINR based on the other agents' actions. This is the only information shared with the agents via the designated base station. This random process gives the FUE_i an instantaneous SINR(i) of 8dB which is calculated using Equation 5. We also assume the following values: $t_b = 10dB, t_a = 2dB$ and $pi = 1dB$. We note that these constants are deterministic.

Reward

The reward function must be carefully chosen because it establishes the system's goal and learning potential. The choice of the reward function for the MUE and FUE are calculated as

$$r_i^m = \left\{ \begin{array}{ll} 100 & \text{if } \gamma_m \gamma_T \\ -1 & \text{otherwise} \end{array} \right\}$$

$$r_i^f = \left\{ \begin{array}{ll} 100 & \text{if } \frac{\gamma_m}{\gamma_T} 1, \frac{\gamma_i}{\gamma_T} 1 \\ -1 & \text{if } \frac{\gamma_m}{\gamma_T} 1, \frac{\gamma_i}{\gamma_T} < 1 \\ -1 & \text{if } \frac{\gamma_m}{\gamma_T} < 1, \frac{\gamma_i}{\gamma_T} < 1 \end{array} \right\} \tag{12}$$

where γ_m and γ_i are the instantaneous SINR of the MUE and FUE respectively, and p_i is the transmission power of the FUE. t_a and t_b are two predetermined thresholds.

The purpose of these reward functions is to allow FUEs to use as much of the same spectrum as they can without interfering with the MUE while maintaining high capacity and protection for the MUE. If the MUE's SINR is exceeds the threshold, it will receive a 100 reward and punished -1 if its SINR is below the threshold because the aim is to give the MUE higher SINR and at the same time to obtain a reasonable SINR for the FUE. Having a negative reward such as -1 would result in a low Q-value. For the FUE, there are three different cases. The FUE will only be rewarded 100 if both its own SINR and the MUE's SINR are above the thresholds. The FUE will be punished -1 when either its own the MUE's SINR is below the threshold. Since the reward function determines the Q-value which is constantly updated as iteration progresses, this choice of values for the rewards might further bridge the gap between the rewards for the FUE and MUE. Through this design, FUEs will have to consider the MUE when selecting transmission powers. The agents will choose the actions that have the highest Q-value at every state, after a specified number of iterations.

5 | RESULTS AND DISCUSSION

A dense urban scenario of multi-tier heterogeneous network is considered to have one macrocell with radius $R_m = 500m$, whereas for every femtocell $R_f = 50m$. There are two to forty femtocells that support multiple MUEs and FUEs. There is only one active FUE and MUE at every given time slot.

As a result, online learning is appropriate for decision-making to carry out power allocation in a heterogeneous environment that requires numerous interactions. Apparently, the FBS and the MBS share the same channel bandwidth at $f = 28$ GHz. q_{00} & q_{01} chosen as 0.8 and 0.2.

The common parameters such as signal to interference plus noise ratio (SINR), capacity (C) and energy efficiency (EE) are the main source of performance evaluation in this system. Simulation parameter values used to perform Q-Learning are listed in Table 2. In this section, the evaluated performance is based on energy efficiency, capacity, and QoS of the system. This evaluation shows how effectively the online learning system can allocate power and sustain users' quality of service. These parameters are as defined in the related literature. The main objective of these parameters is to guarantee, that the MUE has a certain QoS requirement that is above a defined threshold at every time instant.

Another goal is to ensure that FUEs have a reasonable QoS when the condition for the MUE is met. Initially, the distance of all agents is assumed to be in LOS, this implies that constant path loss is used in the Q -learning algorithm. Other parameters are same as illustrated in section 2 unless otherwise stated. Due to the impact on the same distance at each base station, the SINR will be subjected to change by choosing transmission.

The threshold parameters are set as follows, using a Q-learning iteration count of 2000 $(tb = 10; ta = 2; \gamma T = 10)$, at each iteration, agents choose random actions in the first half of the iterations to one of the corresponding transmission power levels in the range of 2dB; 4dB;. . . ; . . . 18dB.

The resulted capacities in oscillating mode where the agents continue to interact with the environment and learn the actions of other agents. Finally, Q-table information of each agent is updated and the results of the FUEs and the MUE are converged during the second half of the iteration.

TABLE 2 Hetnet Simulation Parameters

Symbol	Quantity	Value
BW	Bandwidth	20MHz
	Learning Rate	0.5
	Discount factor	0.9
	FBS Transmit Power	17 dBm
	MBS Transmit Power	45 dBm
	FUE Minimum Power	2 dBm
	FUE Maximum Power	18 dBm
	Predefined threshold	1:1:10
	Thermal noise density	-174 dBm/Hz
	Carrier frequency	28 GHz
	Simulation Time	3200 seconds
	Pathloss Exponent	2,4

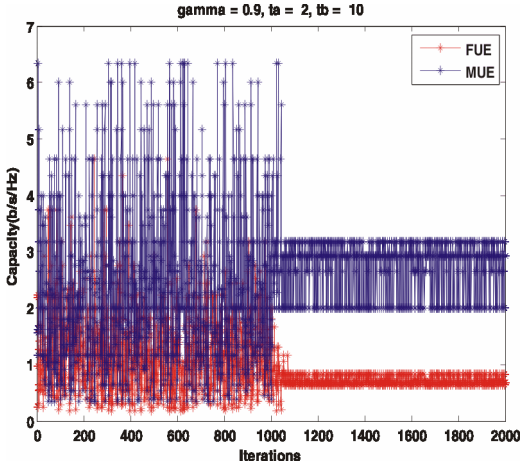

FIGURE 4 Capacity of macrocell and femtocells as a function of iterations

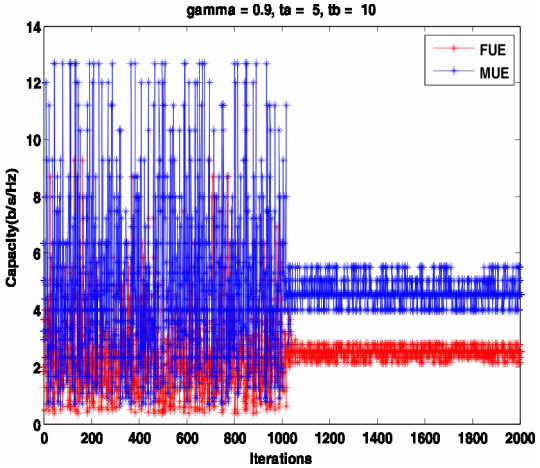

FIGURE 5 Capacity of macrocell and femtocells as a function of iterations.

The effect of cooperative Q learning is described with the help of different state sets (12 states for one FUEs) and (6 states for one MUE). The capacity of the MUE and FUE simulation results as a function of iterations is displayed.

Figures 4 and 5 provide a comparison between the MUE's and FUEs' capacities. The results are displayed to show that there was oscillation between performances prior to 1000 iterations and that the learning process attained a global optimum after 1000 iterations. The speed at which a solution is reached is a second type of optimality. It is observed that the capacity of MUEs is higher than the respective capacity of FUEs, which is due to the condition to prefer for high QoS especially at the macrocell.

The comparison of proposed MDP strategy with conventional sleeping strategy is presented in figure 6. For 20 number of small cells the power consumption is 60W whereas random sleeping strategy consumes 80W and No sleeping strategy consumes 118W almost double when compared to MDP strategy. Small cells confirm power optimization without affecting the QoS constraints C.1 and C.2, guaranteed the SINR of FUEs.

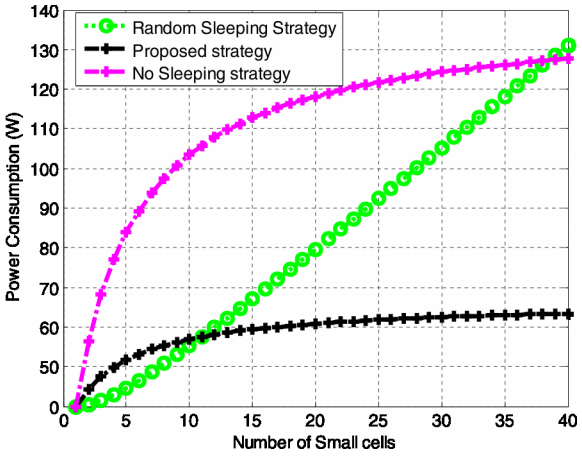

FIGURE 6 Comparison of power consumption

Figure 6 shows how the system's power usage changes when the number of tiny BSs rises. It is compared to what is known as the sleeping approach. We can see that the co-tier interference in dense HetNets causes the power consumption to rise for the random sleep strategy and the no sleep approach. Particularly, the suggested approach lowers co-tier interference through optimization and incorporates the MDP decision process. As a result, the states and action spaces grow in size and the technique successfully lowers the system's power usage.

The system's capacity, QoS, and energy efficiency are taken into account while evaluating performance. This study confirms the online learning system's capacity for effective power allocation and user QoS maintenance. As stated in the relevant literature, these parameters. These parameters' primary goal is to ensure that the MUE has a specific QoS requirement that is greater than a predetermined threshold at all times. Table 3 is the performance outcome of mmWave-enabled Hetnet. Due to the massive demand for traffic, communication networks are now a constant requirement.

TABLE 3 Hetnet Simulation Parameters

No. of FBS	Capacity of Femto users (b/sec/Hz)	Capacity of Macro users (b/sec/Hz)	Power Consumption (Watts)
4	10	6.5	45
10	12	5	50
14	14	4.8	60

In order to address these needs, this work focused attention towards capacity and power consumption on 5G networks. Millimeter Wave (mmWave) band immigration is the solution for new generation Heterogeneous Networks.

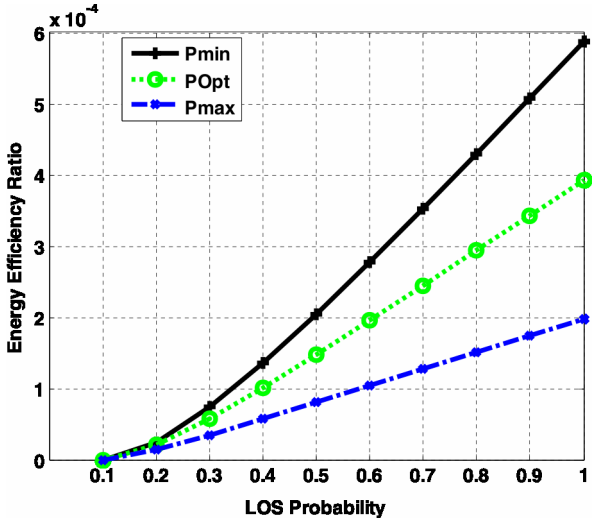

FIGURE 7 Effect of Power control on LOS

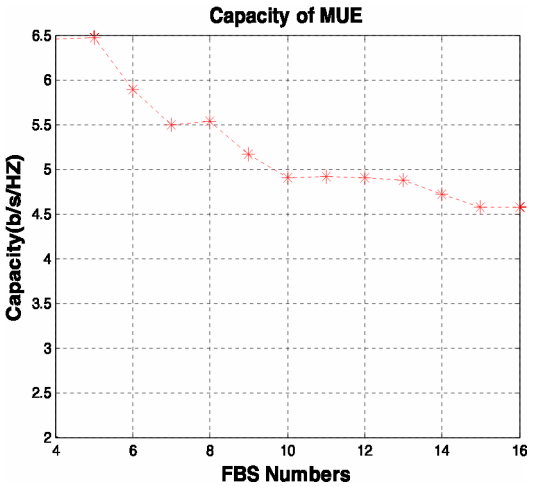

FIGURE 8 Sum Capacity of Macrocells

The proposed reward function results in achieving the maximum energy efficiency against the number of iterations. The system computational complexity increases along with the increase in FBS number, consequently, the size of the states and action spaces increase.

In a big space like our system, the Q-value look-up table is not practical. Q-values are represented as a function of a much smaller set of variables since dimensionality is annoying. The convergence of the proposed scheme reaches within a short time as seen. The disparity between the approximated Q-value and the ideal one is reduced by the online learning Q-value approximation.

In figure 7, It is observed that the energy efficiency Ratio increases with LOS probability. This is because the network throughput decreases at a faster rate than the savings in power consumption when LOS probability is decreased. The femtocell power control can be represented in three ways P_{min}, P_{max} or P_{opt}. In this part, the achieved energy efficiency as a function of the number of small cells is evaluated.

The cumulative capacity can be raised while the MUE's transmission rate is reduced to the level of an exhaustive solution without dropping below the minimum needed rate. The framework enables additional femtocell to adapt its transmission power and it is observed in Figure 8 and 9.

The results of simulation show that the proposed distributed coordinated learning algorithm outperforms previous learning algorithms in terms of learning efficacy, network data rate, and likelihood of QoS satisfaction shown in fig.10.To efficiently learn the optimum intelligent resource management policy, a multi-agent RL network-based approach is suggested and tested through episodes. By estimating the state-value and action advantage functions, the MDP algorithm used networks to learn the action-value distribution. The learning algorithm can quickly converge to the optimum policy under the distributed coordinated learning method and centralized learning methods.

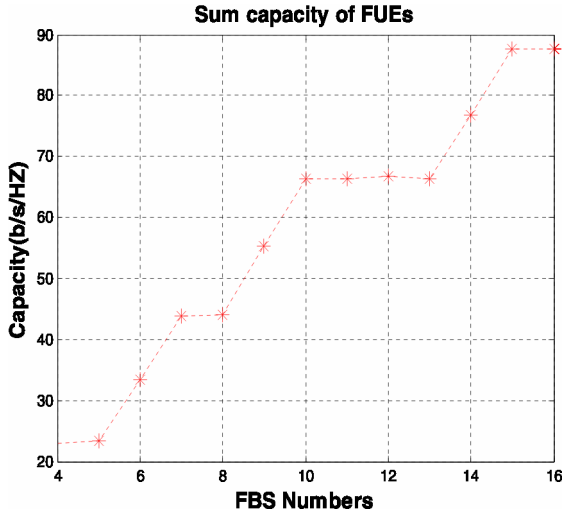

FIGURE 9 Sum Capacity of Femtocells

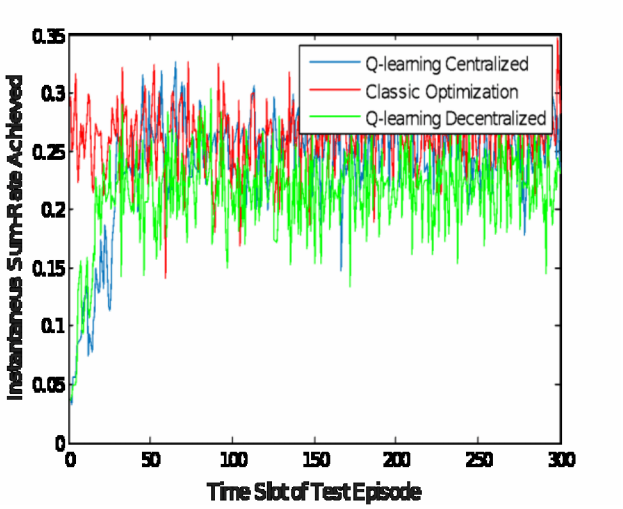

FIGURE 10 Sum Capacity of agents

6 | CONCLUSION

The simulation results are shown is helpful in analyzing the performance of Cooperative learning through MDP algorithm to allocate power and make judicious power optimization. This proposed framework enables femto cells to learn to adapt transmit power to support users. The results have shown how to control and update Q-values during the co-operative learning using the Q-table after a fixed number of iterations, Subsequently, the MDP algorithm has been implemented to take optimal decision among multiple agents, the optimal decision policy for subsequent actions are taken in the network. As a consequence, the energy efficiency of the two-tier heterogeneous network is improved. The proposed MDP strategy is compared with the conventional power control techniques such as sleeping strategy. In terms of system capacity, power usage, and energy efficiency ratio, MDP produces better results. Comparing the observed performance to the traditional power control technique, it is better. It would be interesting to take into account the mobility of UEs and blockage effects in the learning outcome of mmWave-enabled HetNets in future studies.

7 | ACKNOWLEDGEMENT

The authors are grateful to the Deanship of Scientific Research at King Saud University for funding this work through the Vice Deanship of Scientific Research Chairs: Research Chair of New Emerging Technologies and 5G Networks and Beyond.

REFERENCES

1. Marco Mezzavilla et al “An MDP Model for Optimal Handover decisions in mmWave Cellular Networks" arXiv:1507.00387v4 [cs.NI] 6 Jun 2016
2. H. Saad, A. Mohamed, and T. ElBatt, “Distributed cooperative Q-learning for power allocation in cognitive femtocell networks," in Proc. IEEE Veh. Technol. Conf., pp. 1-5, Sep 2012.
3. J. R. Tefft and N. J. Kirsch, “A proximity-based Q-learning reward function for femtocell networks," in Proc. IEEE Veh.Technol. Conf., pp. 1-5, Sep 2013.
4. Amalorpava Mary Rajee,S., Merline.A.: Machine Intelligence Technique for Blockage Effects in Next-Generation Heterogeneous Networks.
5. M. Peng, D. Liang, Y. Wei, J. Li, and H. Chen, “Self-configuration and self-optimization in LTE-advanced heterogeneous networks," IEEE Commun. Mag., vol. 51, no. 5, pp.36-45, May 2013.
6. Roohollah Amiri, Hani Mehrpouyan, Lex Fridman, Ranjan K.Mallik,Arumugam Nallanathan, David Matolak. “A Machine Learning Approach for Power Allocation in HetNets Considering QoS", IEEE International Conference on Communications (ICC), 2018 available online at. arXiv:1803.06760v1[cs.IT] 18 Mar2018.
7. Z. Gao, B. Wen, L. Huang, C. Chen, and Z. Su, “Q-learning-based power control for LTE enterprise femtocell networks," IEEE Syst. J., vol. 11, no. 4, pp. 2699-2707, Dec 2017.
8. Roohollah Amiri, Mojtaba Ahmadi Almasi, Jeffrey G. Andrews,Hani Mehrpouyan, “Reinforcement Learning for self organization and power control of Two-Tier Heterogeneous Networks,IEEE Transactions on Wireless Communications, 2019, available online at. arXiv:1812.09778v2 [cs.IT] 17 Mar 2019
9. H. Sun, X. Chen, Q. Shi, M. Hong, X. Fu, and N. D. Sidiropoulos, “Learning to optimize: Training deep neural networks for interference management," IEEE Trans. Signal Processing, vol. 66, no. 20, pp. 5438-5453, Oct 2018.
10. S. Niknam, R. Barazideh, and B. Natarajan, “Cross-layer Interference Modeling for 5G MmWave Networks in the Presence of Blockage," ArXiv e-prints, Jul. 2018.
11. P. V. Klaine, M. A. Imran, O. Onireti, and R. D. Souza,“A survey of machine learning techniques applied to self-organizing cellular networks," IEEE Commun. Surv. Tutor.,vol. 19, no. 4, pp. 2392-2431, Fourthquarter 2017
12. R. Li, Z. Zhao, X. Zhou, G. Ding, Y. Chen, Z. Wang, and H.Zhang, “Intelligent 5G: When cellular networks meet artificial intelligence," IEEE Wirel. Commun., vol. 24, no. 5, pp. 175-183,Oct 2017.
13. M. Weichold, M. Hamdi, M. Z. Shakir, M. Abdallah, G. K.Karagiannidis,and M. Ismail, Eds., Cognitive Aware Interference Mitigation Scheme for LTE Femtocells. Cham: Springer International Publishing, 2015, pp. 607 Available:http://dx.doi.org/10.1007/978-3-319-24540-9 50.
14. van Otterlo M., Wiering M. (2012) Reinforcement Learning and Markov Decision Processes. In: Wiering M., van Otterlo M.(eds) Reinforcement Learning. Adaptation, Learning, and Optimization, vol 12. Springer, Berlin, Heidelberg.