IEEE Access
Multidisciplinary ⋮ Rapid Review ⋮ Open Access Journal

Received 3 April 2023, accepted 9 April 2023, date of publication 19 April 2023, date of current version 2 May 2023.

Digital Object Identifier 10.1109/ACCESS.2023.3268543

RESEARCH ARTICLE

Optimal Learning Paradigm and Clustering for Effective Radio Resource Management in 5G HetNets

MUHAMMAD USMAN IQBAL[1], EJAZ AHMAD ANSARI[1], SALEEM AKHTAR[1], MUHAMMAD FAROOQ-I-AZAM[1], SYED RAHEEL HASSAN[2], AND RAMEEZ ASIF[2]

[1]Department of Electrical and Computer Engineering, COMSATS University Islamabad (CUI), Lahore Campus, Lahore 54000, Pakistan
[2]School of Computing Sciences, University of East Anglia (UEA), NR4 7TJ Norwich, U.K.

Corresponding author: Muhammad Usman Iqbal (usmaniqbal@cuilahore.edu.pk)

The authors acknowledge the internal research start-up funding, reference: 1012606FA1, from University of East Anglia (UEA), Norwich, U.K.

ABSTRACT Ultra-dense heterogeneous networks (UDHN) based on small cells are a requisite part of the future cellular networks as they are proposed as one of the enabling technologies to handle coverage and capacity problems. But co-tier and cross-tier interferences in UDHN severely degrade the quality of service due to K-tiered architecture. Machine learning based radio resource management either through independent learning or cooperative learning is a proven efficient scheme for interference mitigation and quality of service provision in UDHN in a both distributive and cooperative manner. However, an optimal learning paradigm selection, i.e., either independent or cooperative learning and optimal cooperative cluster size in cooperative learning for efficient radio resource management in UDHN is still an open research problem. In this article, a Q-learning based radio resource management scheme is proposed and evaluated for both distributive and cooperative schemes using independent and cooperative learning. The proposed Q-learning solution follows the $\epsilon-$greedy policy for optimal convergence. The simulation results for the UDHN in an urban setup show that in comparison to the independent learning paradigm, cooperative learning has no significant impact on macro cell user capacity. However, there is a significant improvement in small cell user capacity and the sum capacity of the cooperating small cells in the cluster. A significant increase of 48.57% and 37.9% is observed in the small cell user capacity, and sum capacity of the cooperating small cells, respectively, using cooperative learning as compared to independent learning which sets cooperative learning as an optimal learning strategy in UDHN. The improvement in small cell user capacity is at cost of increased computational time which is directly proportional to the number of cooperating small cells. To solve the issue of computational time in cooperative learning, an optimal clustering algorithm is proposed. The proposed optimal clustering reduced the computational time by four times in cooperative Q-learning.

INDEX TERMS Heterogeneous networks, radio resource management, Q-learning, 5G.

I. INTRODUCTION

Evolution of wireless communication technologies in the last two decades results in an explosive increase in cellular networks users and quality of service (QoS) requirements like higher data rate, throughput, coverage, and capacity while reducing the latency to negligible value (nearly zero). The evolution of cellular networks from 1G to 5G results in improved QoS and Quality of Experience (QoE) key performance indicators (KPIs). Due to the massive increase in cellular network users, the concept of small cells based k-tiered UDHN was proposed for improved coverage and capacity [1], [2], [3], [4], [5]. Although the k-tiered UDHN successfully met the requirements of improved coverage and capacity, some related issues are effective radio resource management (RRM) for efficient interference mitigation as the UDHN deployment results in co-tier and cross-tier interferences which severely degrades the QoS for both macrocell

The associate editor coordinating the review of this manuscript and approving it for publication was Di Zhang.

users and small cell users. For effective utilization of k-tiered UDHN in 5G, co-tier and cross-tier interferences have to be mitigated through efficient RRM [5], [6], [7].

RRM is an essential aspect of wireless communication systems, especially in k-tiered UDHN which consists of multiple types of cells with different frequencies, technologies, and coverage areas. RRM includes but is not limited to load balancing, carrier aggregation, interference mitigation, and self-organizing networks (SON) implementation. Due to the large number of applications of RRM in UDHN, RRM is a widely researched topic in the context of 5G UDHN.

The RRM for effective interference mitigation in k-tiered UDHN is not an easy task due to the dynamic nature of UDHN. Many solutions, most of which were non-adaptive, were proposed in the literature but these non-adaptive solutions cannot handle the dynamic nature of UDHN where the density of small cells continuously changes and therefore interference conditions [6], [8]. In comparison to the non-adaptive RRM algorithms, recently some machine learning based RRM algorithms are proposed in literature which performed significantly better than the non-adaptive algorithms. Reinforcement learning which is a subdomain of machine learning is utilized in devising adaptive RRM through Q-learning in UDHN where the algorithm is utilized to optimize the allocation of network resources such as bandwidth, power, and spectrum to different nodes in the network by continuously learning and interacting with the environment [9].

Reinforcement learning based RRM in UDHN through Q-learning has shown remarkable performance in recently proposed solutions in literature [5], [6], [7]. Q-learning can be applied distributively through independent learning or cooperatively through cooperative learning. However, the literature is silent about the optimal learning scheme in real-time UDHN for 5G cellular networks. In this article, we explored the optimal learning strategy for efficient RRM in terms of various KPIs and the provision of QoS to both the macro cell and small cell users simultaneously.

A. RELATED WORK

Small cells are low-powered wireless access points that are used to provide coverage and capacity in densely populated areas. UDHN are networks that consist of a combination of small cells, macro cells, and other network elements to provide seamless and efficient coverage and capacity [5], [6], [7], [10]. Small cells in UDHN work by complementing macro cells and offloading traffic from them and creating a user balance among the tiers of the network. They are deployed in areas where there is high demand for data, such as shopping centers, airports, and sports stadiums. This helps to reduce congestion and improve the overall QoS and capacity for users through efficient user association [2], [4], [10], [11]. The deployment of small cells in HetNets can also help service providers to meet the increasing demand for high speed data services and support the growth of the Internet of Things (IoT). The combination of small cells and macro cells in a UDHN allows service providers to create a flexible and scalable network that can adapt to changing network conditions and user demand [2]. Recently, many solutions have also been proposed based on software-defined networking (SDN) architecture for efficient deployment of UDHN in the millimeter wave (mmW) spectrum [12], [13], [14].

Although the implementation of small cells has various advantages, the initial cost, overall system reliability, and interferences due to k-tiered architecture are unresolved problems [3]. Interference is one of the main challenges in small cell UDHN. In co-channel deployment mode, small cells operate in the same frequency bands as macro cells, and their proximity to each other can result in interference between small cells and between small cells and macro cells in the same network which is co-tier interference (I^{co}) and cross-tier interference (I^{cr}) respectively [15], [16]. In addition, a relatively small coverage area leads to multiple small cells being deployed close to each other resulting in a UDHN, further exacerbating the interference problem [15], [16], [17]. Therefore, this article focuses on interference mitigation in UDHN through optimal resource allocation.

Recently, researchers presented a number of strategies to improve the reliability, throughput, QoS, QoE, coverage, and capacity of small cells UDHN by mitigating I^{co} and I^{cr} through intelligent and adaptive schemes as the non-adaptive solution for RRM are not considered useful due to dynamic nature of k-tiered UDHN [15], [16], [17], [18]. Therefore, the concept of self-organizing networks (SON), outlined in LTE 3GPP TS 36.300 [19], is utilized in RRM techniques for adaptive solutions [20]. SON integration in UDHN has been also proven profitable and cost-effective for the network operators [21], [22]. However, the integration of SON in UDHN requires some source of cognition or intelligence which can be provided through reinforcement learning (RL).

RL is a type of machine learning algorithm that is used to optimize decision-making in dynamic environments. RL is applied in communication systems using Q-learning (QL) for optimization of network resource allocation, such as spectrum and power allocation, and traffic routing. QL algorithms learn from network conditions, such as traffic patterns and interference levels, and determine the best actions to take in real-time to improve network performance. This results in more efficient use of network resources, improved network coverage and capacity, and reduced latency [17], [18]. Overall, QL has the potential to be a key enabler for the successful deployment of 5G networks, providing the ability to optimize network operations and improve network performance in dynamic and complex environments [23].

QL can be implemented in many different ways. While the basic QL algorithm follows a similar structure, the differences between different QL schemes can be substantial. One QL scheme is different from another in terms of value function representation, exploration-exploitation tradeoff, reward function, learning rate, and discount factor [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34]. The novelty of different QL schemes lies in the development of the optimization problem and constraining it with network key performance indicators and the design of an efficient reward function to solve the optimization problem. The solutions proposed for adaptive power allocation to small cells define optimization problems in different ways and with different constraint and therefore a novel reward function to solve it. Each QL scheme may have a different reward function design to address the underlying optimization problem. Similarly, the discount factor and learning rate determine the convergence of the proposed QL scheme [23].

Despite several QL schemes based on distinct reward functions, discount factors, and other QL parameters [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34], are proposed in the literature to handle the I^{co} and I^{cr} through adaptive power control in small cells UDHN, a basic limitation is their inability to guarantee QoS to both macrocell users (M^{u}) and small cell users (S^{u}) simultaneously.

QL has been widely applied to the UDHN either through independent learning (${}^{i}L$) mode in a distributed manner or cooperative learning (${}^{c}L$) mode in a cooperative manner. Both of the learning paradigms have pros and cons. Independent Q-Learning is generally simpler to implement and more scalable, as it does not require communication between the cells. On the other hand, Cooperative Q-Learning can lead to more efficient decision-making and better network performance, as the cells can learn from each other's experiences. The reward function, an integral and critical part of QL is impacted by the learning paradigm. In ${}^{i}L$, each learning agent act according to individual reward optimization whereas in ${}^{c}L$, cooperating agents learn to form a joint RF [35]. QL solutions for RRM have been proposed in both ${}^{i}L$ and ${}^{c}L$. Although some solutions have been proposed in both ${}^{i}L$ and ${}^{c}L$ but optimal learning scheme for real-time implementation has not been proposed [24], [25], [26], [27], [28], [29], [30], [31], [32], [33], [34].

Literature review reveals multiple types of limitations of state of the art QL based adaptive power control schemes for UDHN like the value function representation and proposed reward functions in these QL schemes are either biased to M^{u} or S^{u} as discussed in [15] and [16]. Some of the proposed schemes could not set their superiority against the state of the art solutions. Many solutions proposed in ${}^{c}L$ performed better than ${}^{i}L$ but the performance was bottlenecked by the communication overhead and computational time [36], [37]. Therefore, there is still a need to investigate an ideal or optimal learning paradigm that can be deployed in real-time implementation to improve the performance of UDHN.

In this article, we proposed a QL scheme to address the above-mentioned limitations of the state of the art QL based adaptive power allocation schemes and also devised an optimal learning paradigm based on the performance of the proposed QL scheme in independent and cooperative learning paradigms.

B. CONTRIBUTIONS

In this article, we have investigated the performance of the QL-based RRM for small cell UDHN to provide QoS to both macrocell and small cell users by simultaneous mitigation of co-tier and cross-tier interferences. The proposed solution based is evaluated in both independent and cooperative learning paradigms to find an optimal learning paradigm for real-time deployment scenarios. The following are the major contributions of the paper:

- A QL-based adaptive power allocation scheme is proposed to handle the co-tier and cross-tier interferences simultaneously in the small cell UDHN which ensures QoS for both macrocell and small cell users.
- The proposed QL scheme models the small cell UDHN as a single or multi-agent Markov Decision Process (MDP) where small cells play the role of QL agents in the network and implement QL for RRM.
- The defined optimization problem which maximizes the capacity of macrocell users, small cell users, and the sum capacity of small cell users is constrained over the minimum required QoS thresholds to guarantee QoS for all users and is solved through the proposed reward function for the QL algorithm.
- The optimal learning paradigm is proposed by evaluating the proposed QL based adaptive power allocation scheme in both learning paradigms, i.e., independent and cooperative learning.
- Simulation results in various combinations of co-tier and cross tier interferences based on standard 3GPP simulation setup prove the optimality of the cooperative learning paradigm in terms of standard KPIs at the cost of increased computational time.
- For efficient deployment of small cell UDHN, an optimal clustering algorithm is also proposed and evaluated. The simulation results show that optimal learning, i.e. cooperative learning, is the efficient QL implementation scheme when deployed with optimal clustering technique which significantly reduces its computational time.

The paper is organized as follows: a system model for small cell UDHN is presented in section II for a comparison of QL implementation in ${}^{i}L$ and ${}^{c}L$ for RRM to QoS. In section III optimization problem in the underlying study is presented. QL algorithm in ${}^{i}L$ and ${}^{c}L$ paradigm to solve the optimization problem is presented in section IV whereas the simulation parameters and setup for evaluation of the proposed solution are discussed in section V. The results of Monte-Carlo simulations to compare the performance of ${}^{i}L$ and ${}^{c}L$ in 3GPP interference setups are presented in section VI whereas the conclusion of the paper is presented in section X.

II. SYSTEM MODEL

A system model composed of the k-tiered ${}^{s}C$ UDHN is presented in Fig.1 where small cells (${}^{s}C$) are deployed in the co-channel mode under the over laid macrocell (${}^{m}C$). The

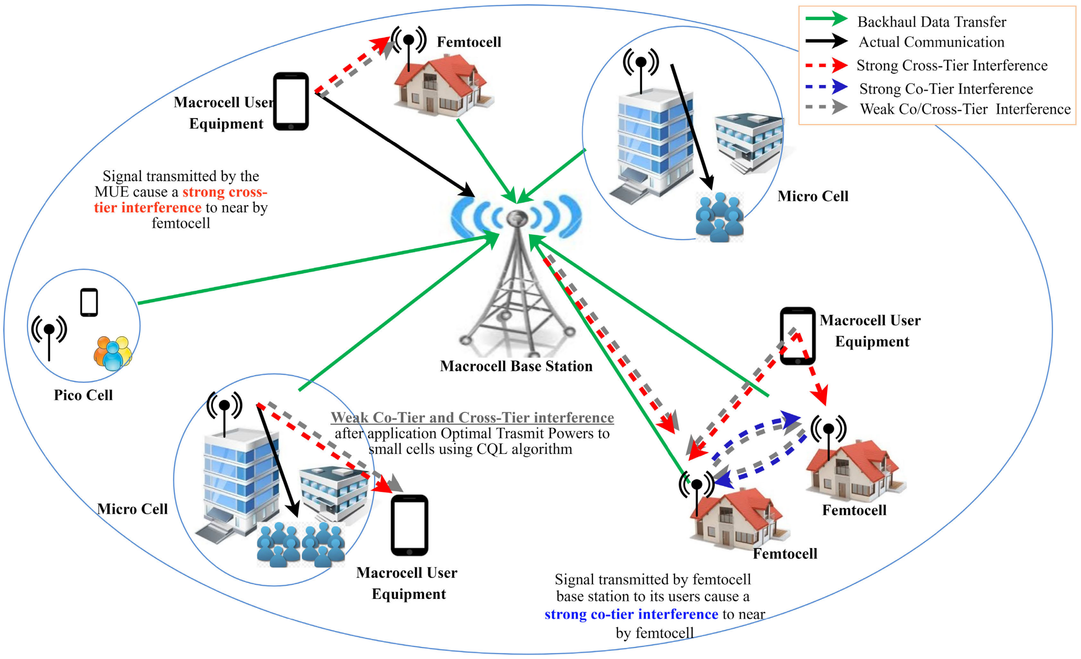

Backhaul Data Transfer
Actual Communication
Strong Cross-Tier Interference
Strong Co-Tier Interference
Weak Co/Cross-Tier Interference
Femtocell
Macrocell User Equipment
Signal transmitted by the MUE cause a strong cross-tier interference to near by femtocell
Micro Cell
Pico Cell
Macrocell Base Station
Macrocell User Equipment
Weak Co-Tier and Cross-Tier interference after application Optimal Trasmit Powers to small cells using CQL algorithm
Micro Cell
Macrocell User Equipment
Femtocell
Femtocell
Signal transmitted by femtocell base station to its users cause a strong co-tier interference to near by femtocell

FIGURE 1. ***k*-tiered small cell UDHN system model based on single macro cell users and multiple small cell users.**

TABLE 1. **Symbols/notations.**

Symbol	Description
$\mathcal{F}^{sub}$	F frequency subbands, $\mathcal{F}^{sub}=\{1,2,...,F\}$
$\mathcal{N}^{s}$	N ^{s}C, $\mathcal{N}^{s}=\{1,2,...,N\}$
$\daleth^{m}$	SINR threshold for QoS at M^{u}
$\daleth^{s}$	SINR threshold for QoS at S^{u}
ς_{i}^{m}	SINR at the i^{th} M^{u}
$\varsigma_{n,k}^{s}$	SINR at the k^{th} S^{u} of n^{th} ^{s}C
p_{i}^{m}	BS^{m} transmit power
$p_{n,k}^{s}$	BS^{s} transmit power
$h_{m,i}^{m}$	channel gain: BS^{m} to M$_{i}^{u}$
$h_{m,i}^{n}$	channel gain: n^{th} ^{s}C to M$_{i}^{u}$
$h_{n,k}^{n}$	channel gain: n^{th} ^{s}C to its S$_{n,k}^{u}$
$h_{n,k}^{j}$	channel gain: j^{th} ^{s}C to S$_{n,k}^{u}$
$h_{n,k}^{m}$	channel gain: BS^{m} to S$_{n,k}^{u}$
N_{o}	σ^{2} of AWGN
${}^{m}C_{i}$	Capacity of M$_{i}^{u}$
${}^{s}C_{n,k}$	Capacity of S$_{n,k}^{u}$ of n^{th} ^{s}C
${}^{s}C_{sum}$	Sum of capacities of all S$_{n,k}^{u}$
ξ^{m}	Capacity of M^{u} for QoS
ξ^{s}	Capacity of S^{u}for QoS
$\mathbb{P}$	BS^{s} Transmit powers $\mathbb{P}=\{p_{1},p_{2},...p_{max}\}$
π^{*}	Optimal policy
α	QL Rate
$\daleth$	Discount Factor
$\mathcal{R}_{k}^{t}$	Proposed RF

system model, Fig.1, is a standard HetNet simulation model based on Setup2b [38]. ^{m}C operates in the downlink over orthogonal subbands, $\mathcal{F}^{sub}$, where $\mathcal{F}^{sub}=\{1,2,3,\ldots.F\}$. A single macrocell base station, BS^{m}, is assumed to be placed in the center of ^{m}C whereas the M$_{i}^{u}$ are placed randomly in ^{m}C. from the set of M$_{i}^{u}$, $\mathcal{I}$, where and $\mathcal{I}=\{1,2,3,\ldots,\mathcal{I}\}$. According to the Setup2b [38], a cluster of ^{s}C, $\mathcal{N}^{sc}$, is considered in the coverage area of ^{m}C. $\mathcal{N}^{sc}$ is composed of N ^{s}C such that $\mathcal{N}^{sc}=\{1,2,3,\ldots.N\}$. Similar to the ^{m}C, each k^{th} user of n^{th} ^{s}C from $\mathcal{N}^{sc}$, S$_{n,k}^{u}$, where $k\in\mathcal{K}$ and $\mathcal{K}=\{1,2,3,\ldots\mathcal{K}\}$ are also deployed indoor randomly in the ^{s}C. The ^{s}C are operating in the co-channel deployment mode. BS^{m} and BS^{s} equally divide the transmission power to its their users [39]. It is assumed that QoS parameters ^{s}C are provided by the network operator in the SON procedures like self-configuration.

The presence of S$_{n,k}^{u}$ in ^{m}C due to the k-tiered ^{s}C UDHN results in I^{cr} to M$_{i}^{u}$ which affect its SINR. In the downlink, the SINR at any M$_{i}^{u}$, ς_{i}^{m}, can be calculated as follows in presence of I^{cr}

$$\varsigma_{i}^{m}=\frac{p_{i}^{m}|h_{m,i}^{m}|^{2}}{\underbrace{\sum_{n\in\mathcal{N}^{sc}}p_{n}^{s}|h_{m,i}^{n}|^{2}}_{\mathrm{I}^{cr}}+N_{o}}\tag{1}$$

where p_{i}^{m} and $h_{m,i}^{m}$ transmitted power and channel gain by BS^{m} to all M$_{i}^{u}$ operating in ^{m}C. p_{n}^{s} and $h_{m,i}^{n}$ is the transmitted power and the channel gain by n^{th} BS^{s} to all M$_{i}^{u}$. which results in I^{cr}. In addition to I^{cr}, AWGN also impacts the ς_{i}^{m} which is represented by the variance, σ^{2} in (1).

Unlike (1), the SINR at k^{th} S^{u} of n^{th} ^{s}C, S$_{n,k}^{u}$, $\varsigma_{n,k}^{s}$, in the downlink operating on the subband $f\in\mathcal{F}^{sub}$, is impacted by I^{cr} from BS^{m}, I^{co} from the neighboring BS^{s} and thermal noise. The $\varsigma_{n,k}^{s}$ is obtained as

$$\varsigma_{n,k}^{s}=\frac{p_{n,k}^{s}|h_{n,k}^{n}|^{2}}{\underbrace{\left[p^{m}|h_{n,k}^{m}|^{2}\right]}_{\mathrm{I}^{cr}}+\underbrace{\left[\sum_{j\in\mathcal{N},j\neq n}p_{j}^{s}|h_{n,k}^{j}|^{2}\right]}_{\mathrm{I}^{co}}+N_{o}}\tag{2}$$

where p^m, p_j^s and $p_{n,k}^s$ are the transmitted power by BS^m, BS^s of j^{th} and n^{th} ^{s}C to S$^u_{n,k}$ respectively. Similarly, $h^m_{n,k}$, $h^n_{n,k}$ and $h^j_{n,k}$ and are the channel gains from the BS^m and BS^s of n^{th} and j^{th} ^{s}C to the S$^u_{n,k}$ respectively.

From the ς_i^m and $\varsigma_{n,k}^s$ in (1) and (2), respectively, normalized capacities at the M$_i^u$ and S$^u_{n,k}$ are given below:

$$^mC_i = log_2(1 + \varsigma_i^m) \tag{3}$$

$$^sC_{n,k} = log_2(1 + \varsigma_{n,k}^s), \tag{4}$$

where ^mC_i and $^sC_{n,k}$ are the capacities of M$_i^u$ and S$^u_{n,k}$ respectively.

The accumulated value of capacities of all S$^u_{n,k}$ in the system, $^sC_{sum}$ is represented as follows:

$$^sC_{sum} = \sum^{s} C_{n,k} \quad \forall\, k \text{ } in\, \mathcal{K} \,\&\, n \in \mathcal{N}^{sc} \tag{5}$$

III. PROBLEM FORMULATION

The problem of RRM in 5G UDHN addressed in this research is one of the major research problems for many years in the domain of ultra-dense HetNets as 5G enabling technology. Recently, many solutions are proposed for QoS provision for all users in k-tiered ^{s}C UDHN architecture through optimal RRM and interference mitigation [15], [16], [27], [28], [29], [31], [34]. However, the fundamental difference among the recently proposed RRM techniques lies in an optimization problem in terms of optimization function and conditions.

The optimization problem (OP) defined in the underlying research strives to maximize the ^mC_i, $^sC_{n,k}$, and $^sC_{sum}$ while keeping ^mC_i and $^sC_{n,k}$ above the QoS capacity thresholds ξ^m and ξ^c through interference mitigation. The adaptive ^{s}C transmission power-based intelligent interference mitigation scheme handles the I^{co} and I^{cr} simultaneously and thus guarantees QoS to M$_i^u$ and S$^u_{n,k}$ by improving SINR. OP for adaptive power allocation is defined as follows by assuming that BS^s of n^{th} ^{s}C, operating over a subband, $f \in \mathcal{F}^{sub}$, can select a transmit power, p^s from the available set of powers, $\mathbb{P} = \{p_1, p_2, \ldots p_{max}\}$.

$$\max_{\mathbb{P}} \ {}^mC_i, {}^sC_{n,k}, {}^sC_{sum} \tag{6a}$$

$$\text{subject to } p_1 \leq p_n^s \leq p_{max}, \quad n \in \mathcal{N}^{sc} \tag{6b}$$

$$^mC_i \geq \xi^m, \quad i \in \mathcal{I}, \tag{6c}$$

$$^sC_{n,k} \geq \xi^s, \quad n \in \mathcal{N}^{sc} \,\&\, k \in \mathcal{K} \tag{6d}$$

where p_1 and p_{max} define the range of discrete values of transmit powers which any BS^s may select.

The objective function, (6a), maximize ^mC_i, $^sC_{n,k}$, and $^sC_{sum}$ whereas the constraints of the OP, (6b)-(6d), describe threshold/ values p_n^s, ^mC_i and $^sC_{n,k}$. The OP constraints in (6c) and (6d), ensure QoS provision to S^u and M^u simultaneously in the ^{s}C UDHN. OP in (6a) - (6d) can be solved through learning based adaptive solution by relating the p_n^s to the ^mC_i and $^sC_{n,k}$ while constraining over QoS capacity thresholds. The learning framework to solve (6a)–(6d) is discussed in the following sections.

IV. OPTIMAL RESOURCE ALLOCATION IN ^{s}C UDHN USING QL

The QL is an iterative algorithm to apply RL in a system where the environment is dynamic or unknown. QL agents interact with the environment and strive for a maximum reward through a learned optimal policy, π^*. However, learning π^* is a computationally extensive process that requires improving the π in each iteration. π^* can be found whether prior information on the environment is available or not. ^{s}C UDHN can be modeled as MDP to implement QL based RRM for interference mitigation and QoS provision. The detailed modeling of ^{s}C UDHN as MDP is provided in [15], [32], [33], [34], and [16].

A. PROPOSED QL ALGORITHM

Based on the rationale of QL and ^{s}C UDHN as MDP, we have proposed QL algorithm, 1 for optimal RRM in ^{s}C UDHN as MDP. The proposed QL algorithm is based on the definitions of ^{s}C UDHN as MDP where each BS^s acts as the QL agent and adaptively selects an action, a_n^t, which is transmission power based on learning of the QL agent. The actions of the agents, $a_n \in A$, are a discrete set of transmission powers, $\mathbb{P}$, of BS^s, as defined in section III. The step size between elements of $\mathbb{P}$ is calculated through the following equation [32], [33], [34].

$$step = \frac{P_{max} - P_{min}}{N_{Power}} \tag{7}$$

In the iterative process of learning and improvement, the QL agent updates Q-Table (QT) which is based on the actions, a_n^t, and states, x_n^t, of QL agent [15]. At the time, $t = 0$, QT is initialized with no entry in QT and a random state, x_n^t. During the iterative process, an action, a_n^t, can be selected based on the exploration-exploitation policy (EEP) [15].

$$a_t = \begin{cases} arg\, \max\limits_{a \in A} Q^t(x, b) & exploitation(1 - \epsilon) \\ \underset{a \in A}{rand}(a) & exploration(\epsilon) \end{cases} \tag{8}$$

After the selection of a_n^t at time t, reinforcement, R^{t+1} is applied according to the (9) resulting in the new state, x_c^{t+1} of the agent and QT is updated [16].

$$\begin{aligned} Q^{t+1}(x_t, a_t) = (1 - \alpha)Q^t(x_t, a_t) + \alpha\{R_{t+1} \\ + \daleth \underbrace{\max_{a'} Q^t(x_{t+1}, a')\}}_{\mathcal{R}_f^t} \end{aligned} \tag{9}$$

where the $\mathcal{R}_f^t$ is the reward function (RF) of the QL algorithm. For the proposed QL algorithm, $\mathcal{R}_f^t$ is defined as a function of $^mC_i, {}^sC_{n,k}$, Γ^m and Γ^s at any time t and is given below.

$$\mathcal{R}_f^t = w \underbrace{\{{}^mC_i^t\}^{zs}C_{n,k}^t}_{a} - w^{-2} \underbrace{\{B_m + B_k\}}_{b} \tag{10}$$

where

$$B_m = \{{}^{m}C_i - \Gamma^{m}\}^2$$
$$B_k = \{{}^{s}C_{n,k} - \Gamma^{s}\}^2$$
$$w = \frac{D_{BS^s - M_i^u}}{d_{th}}$$

The part a of the $\mathcal{R}_f^t$ in (10) encourages the system to maximize the capacities of the M_i^u and $S_{n,k}^u$ whereas the part b guarantee to meet the minimum QoS requirements. A value of $z > 1$ in (10) provides small preference to ${}^{m}C_i$ over the ${}^{s}C_{n,k}$ as M_i^u is the primary user in the network. z and d_{th} are user defined parameters which selected in line with the literature [15], [16], [32]. The systematic design and development of effective $\mathcal{R}_f^t$ is presented in [15].

The EEP in (9) involves the ^{c}L and ^{i}L learning paradigms. In the ^{i}L, each QL agent in the system learn and act independently regardless of learning of neighboring agents and impact of its actions on the other agents or environment. On the other hand, ^{c}L consider the learning of the other agents and consider the impact of their related actions while learning and therefore share the learning information with neighboring agents. As compared to the learning process in ^{i}L, learning agents cooperate with neighboring agents by sharing rows of their updated QT. The details of ^{i}L and ^{c}L are discussed in the following subsections.

B. INDEPENDENT LEARNING VS COOPERATIVE LEARNING

In the ^{s}C UDHN as MDP, BS^s are the learning agents of the QL which repeatedly interact with the surrounding environment to learn π^* from the π by improving in each iteration. According to the EEP, learning of the agents can be either in the ^{i}L and ^{c}L mode. Both learning modes are explained in the following subsections.

1) INDEPENDENT LEARNING

In ^{i}L, the learning of each agent in the UDHN is independent of the other agents in the environment. While learning in ^{i}L, agents assume anything around it as the environment even if there are other agent present and act selfishly somehow. No agent cooperates with other agents by not accepting or sharing any information. Therefore, no prior information is available to the agents in this learning paradigm. In the ^{i}QL, any agent does not cooperate to share any QL information with other neighboring agents.

2) COOPERATIVE LEARNING

In ^{i}L paradigm, each agent of ^{s}C UDHN learns π^* for RRM individually and without any prior information, therefore, learning the π^* more time and resources. Furthermore, in ^{i}L all agents learn an optimal policy, π^*, individually in a somehow greedy manner regardless of the negative impacts on the neighboring agents. Contrary to the ^{i}L, in ^{c}L ^{s}C cooperates with the neighboring agents to exchange the QL-related information. Unlike ^{i}L, prior information about agents and the environment is available to the new agents entering the system in ^{c}L. Therefore, learning π^* is more robust and effective in ^{c}QL as compared to ^{i}QL. The cooperation of the agents also helps the neighboring agents in learning and considering the environment in learning π^* in such a way

Algorithm 1 QL Algorithm for Optimal RRM in ^{s}C UDHN for 5G CN ^{i}L and ^{c}L Paradigm

```
Define size and coverage area of QL Agents
    $N^C$ where $\mathcal{N}^{sc} \leq N^C$
    $*R^C$ where $R^C \leq^m C$ coverage area
For each agent $n \in \mathcal{N}^{sc}$, Define
    Agent states $x_n^t \in X$
    Agent actions $a_c \in A$
    Initialize QT i.e. $Q_t(x_n^t, a_c^t)$
At $t = 0$,
if $n > 1$ and ${}^{c}L$ then
    Initialize QT as $Q^0(x_c^0, a_c^0)$
    Update QT with avaiable shared QT information
else
    Initialize QT as $Q^0(x_n^0, a_n^0)$
end
for All $n \in \mathcal{N}^{sc}$ (Parallel) do
    for $n_i \leq N_{itrations}$ do
        current $x_n^t = x_n^0$
        for $n_s \leq N_{step}$ Apply EEP do
            if rand $< \epsilon$ i.e. then
                    Random $a_n^t \in A$
            else
                if ${}^{c}L$ then
                    Share $Q_i^t(x_n^t, :)$ with cooperating
                    agents, $j$,
                    Collect $Q_j^t(x_j^t, :)$ from cooperating
                      agents, $j$,
                    $a_i^t \leftarrow arg\max_a \sum_{n \in \mathcal{N}^{sc}} Q_k^t(x_n^t, a^t)$
                else
                    % Independent
                    $a_i^t \leftarrow arg\max_a Q_c^t(x_n^t, a^t)$
                end
            end
            current action $a_n^t$
            Perform Reinforcement $R^{t+1}$
            Compute new state $x_n^{t+1}$
            Update QT
            Apply $x_n^t \leftarrow x_n^{t+1}$
            Apply $t \leftarrow t + 1$
            if convergence condition given in (11) met
              then
                visit next state-action pair
            else
                continue iterations
            end
        end
    end
    if $n > 1$ and ${}^{c}L$ then
        Share the updated QT with all agents in the system
    else
        Do not share the updated QT
    end
end
* All QL Agents in radius of $R^C$ from QL Agent nearest to
  ${}^{m}$C
```

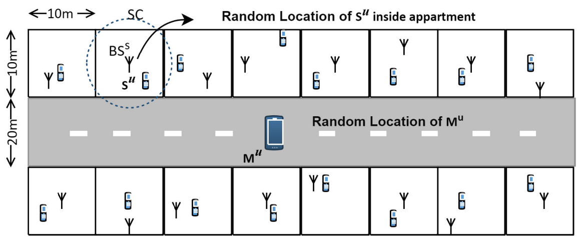


FIGURE 2. Apartment strips to create simulation setup based on single macrocell and small cell user per macrocell and small cell respectively [38].

that their impact on the performance of other agents is least. Therefore, ^{c}QL can further reduce the I^{co} in ^{s}C UDHN and convergence time for new agents entering the system.

Based on the above rationale, the ^{c}L is a step ahead of the ^{i}L where useful information is shared with the cooperating ^{s}C to optimally allocate RRM. In ^{c}L, ^{s}C cooperates by sharing one of the following three different types of information; *i)* episodic information, *ii)* instantaneous information, and *iii)* individually learned π^* [40].

Based on the above cooperating techniques, in the proposed algorithm QL agent shares QT with agents to cooperatively learn an π^* to adaptively allocate p_n^s to handle I^{co} and I^{cr} and improve the capacity of the ^{s}C while considering minimum required QoS parameters as proposed in [27], [28], and [32].

C. COOPERATING CLUSTERING

The limiting factor in the performance of ^{c}QL is the number of cooperative ^{s}C, n, in the cooperating cluster. As n increases, the size of QT also increases which results in increased computational time and overhead. To handle the issue of large cluster size, an optimal clustering algorithm is proposed presented in Fig. 10. The optimal clustering algorithm combines the neighboring ^{s}C into small overlapping clusters. Each ^{s}C decides its cooperating agents based on the distance threshold, d_{nt}, of the neighboring ^{s}C discovered earlier using automatic neighbor discovery (AND) in the self-configuration phase. Although there is no limit on the number of n in the cluster, the size of the cluster remains limited due to d_nt. According to the proposed algorithm, an ^{s}C can be part of multiple clusters due to the random deployment of ^{s}C in ^{s}C UDHN. All the clusters execute ^{c}QL in parallel which results in fast convergence of ^{c}QL in less computational time and with negligible computational overhead.

D. CONVERGENCE OF Q-LEARNING ALGORITHM

In Q-learning algorithm, an optimal policy (π^*) in MDP is learned through an iterative process. The algorithm works by updating estimates of the Q-values of the state-action pairs based on the received rewards and the Q-values of the next state-action pairs in each iteration. One of the key properties of Q-learning is its ability to converge to the optimal Q*-values, given certain conditions. Specifically, Q-learning is guaranteed to converge to the optimal Q*-values based on the following factors:

TABLE 2. Simulation parameters.

Parameter	Value
BS^m	1
Number of ^{s}C in UDHN	16
Number of M^u	1
Number of S^{u}in each ^{s}C	1
BS^m Coverage Area	350 m
^{s}C Coverage Area	10 m
Transmit power of BS^m , p^m	50 dBm
Transmit power of BS^s , p_n^s	-15 to 15 dBm
N_{Power}	32
^{m}C Operating Frequency	2.0 GHz
N_1	3
N_2	3
d^m	50, 150, 250
d^s	15, 25, 40
z	2
ξ^m	1 b/s/Hz
ξ^s	1 b/s/Hz
α	0.5
$\daleth$	0.9
Exploration Probability (ϵ)	10%, 1% , 0.1%
Maximum QL Iterations	75000
Channel Model	Dual Strip Model [43]
Traffic Model	Full Buffer [43]

- An appropriate selection of learning rate parameter (α), determines the degree to which new information overrides the old information. A high value of α may lead to no convergence whereas a low value may result in slow convergence.
- An appropriate selection of exploration rate in EEP (8). The exploration rate (ϵ) in EEP determines the degree to which the algorithm explores versus exploits the current best estimate of the Q-values. If the exploration rate is set too high, the algorithm may not converge, while if it is set too low, the algorithm may converge to a suboptimal policy.
- The state and action spaces must be finite.
- The MDP must satisfy the "Markov property," which means that the future state and reward depend only on the current state and action, and not on any previous states or actions.

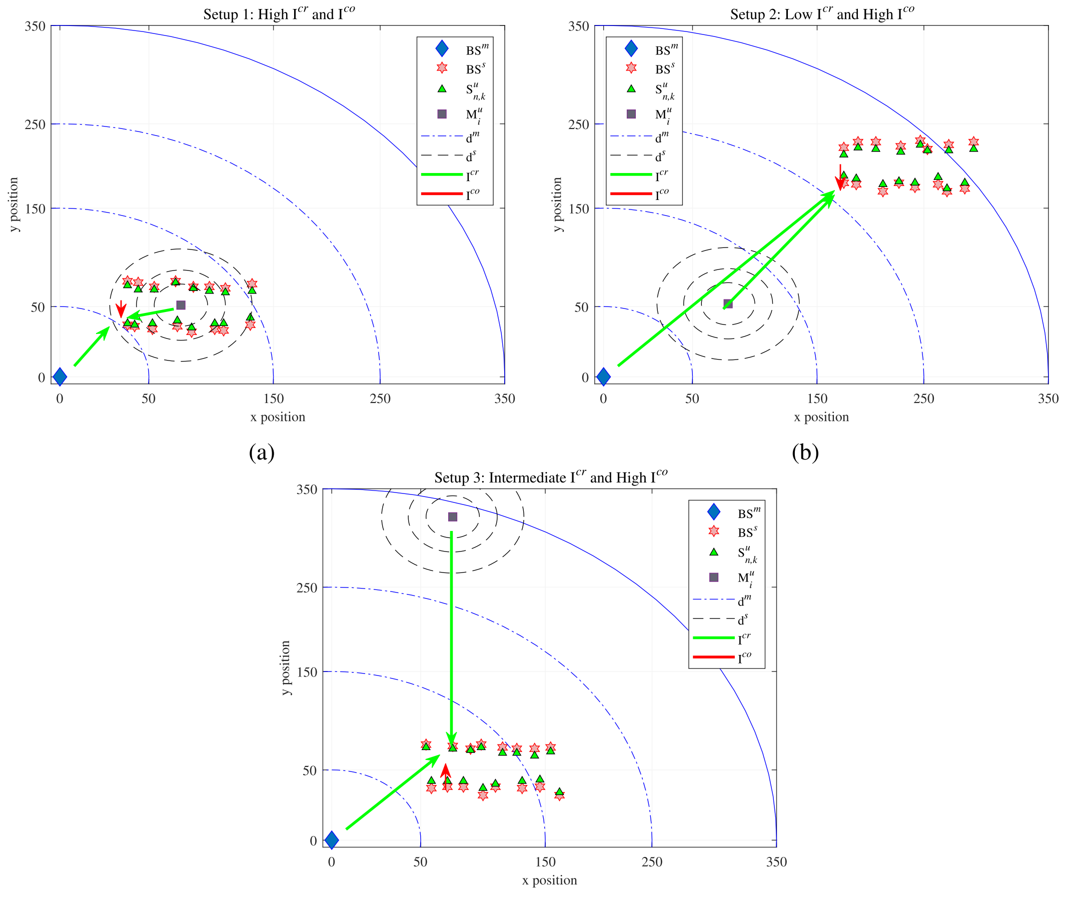
Based on the above factors, the convergence of the QL can be achieved through the $\epsilon-$greedy policy in EEP. π^* is an $\epsilon-$optimal policy for $\epsilon > 0$ and $\delta \in (0, 1]$, if the following condition is met [32], [41], [42].

$$Pr(||Q^* - Q_\pi|| < \epsilon) \geq 1 - \delta \tag{11}$$

where Q* represents the learned optimal Q*-value after QL iterations and Q$_\pi$ is the Q-value based on the current state-action pair. The proposed QL algorithm follows $\epsilon-$greedy policy for optimal convergence as in [15], [32], and [16].

V. SIMULATION SETUP AND PARAMETERS

The proposed QL algorithm is evaluated through Monte-Carlo simulations in a standard 3GPP setup [38] in MATLAB 2020a on a Corei7,16 GB memory machine. The UDHN



(a)

(b)

(c)

FIGURE 3. UDHN simulation setups based on Fig.2 (a) simulation setup 1: high I^{cr} and I^{co}, (b) simulation setup 2: low I^{cr} and high I^{co} setup, and (c) simulation setup 3: intermediate I^{cr} and high I^{co} setup.

simulation setups, Fig. 3 are created by variation in Setup 2b (sparse), Fig. 2a based on the urban dual strip model [38], [43]. The UDHN simulation setups, Fig. 3, are developed to cater to different combinations of I^{co} and I^{cr} based on the density of ^{s}C, and the number of M^u and S^u and the location of apartment strips in the ^{m}C. The simulation parameters of ^{m}C and ^{s}C are according to the 3GPP TR 36.872 [38]. The simulation parameters are also in line with the recently proposed RRM solutions through QL [27], [29], [31], [32], [33], [34].

A summary of the simulation parameters is provided in Table 2.

VI. RESULTS

The performance of the QL algorithm is evaluated in multiple interference scenarios presented in Fig. 3 by increasing the density of the ^{s}C in the system. The ^{s}C are added one by one where each ^{s}C performs QL either through ^{i}L or ^{c}L. According to the algorithm, all ^{s}C learn independently in parallel, however, share learned information in form of QT if operating in ^{c}L mode.

The simulation results of QL in ^{i}L or ^{c}L are measured through various KPIs to find an optimal learning policy for QL based RRM in ^{s}C UDHN for 5G CN. The performance of the QL algorithm is compared in terms of the capacities of macrocell and small cell users, the sum capacity of small cells, computational time, and the sum power transmitted by small cells in the network.

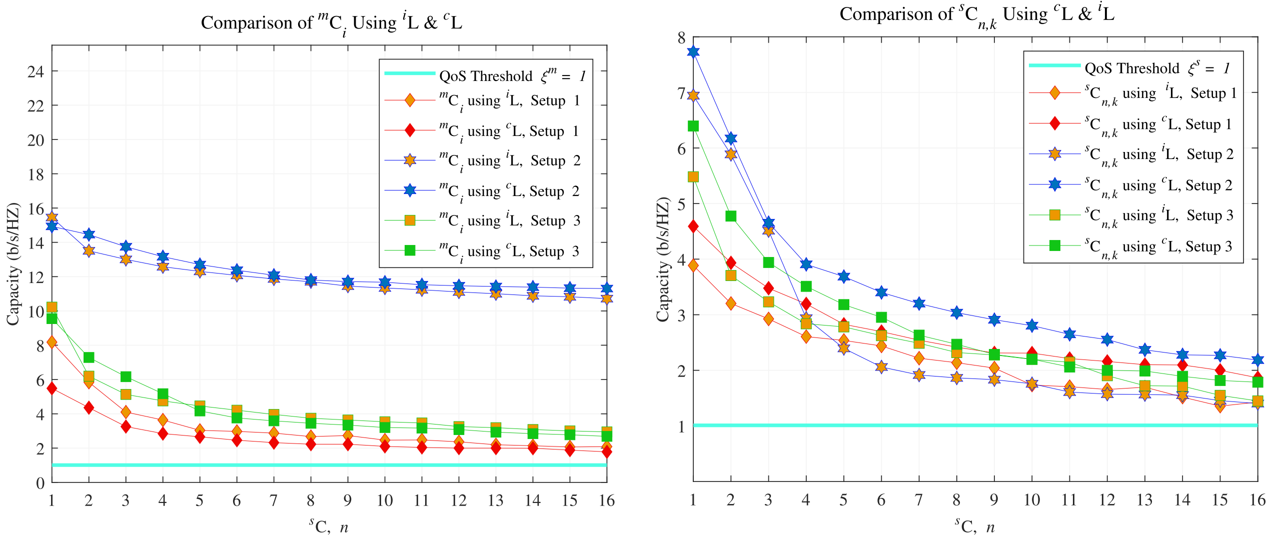
TABLE 3. Comparison of $^m C_i$ (b/s/Hz) using ^{i}L and ^{c}L.

Setup	^{i}QL	^{c}QL	^{c}QL based Analysis
1	2.08	2.02	↓2.97%
2	10.7	11.30	↑ 5.60%
3	2.94	3.08	↑ 4.60%

^{c}QL performed better than ^{i}QL

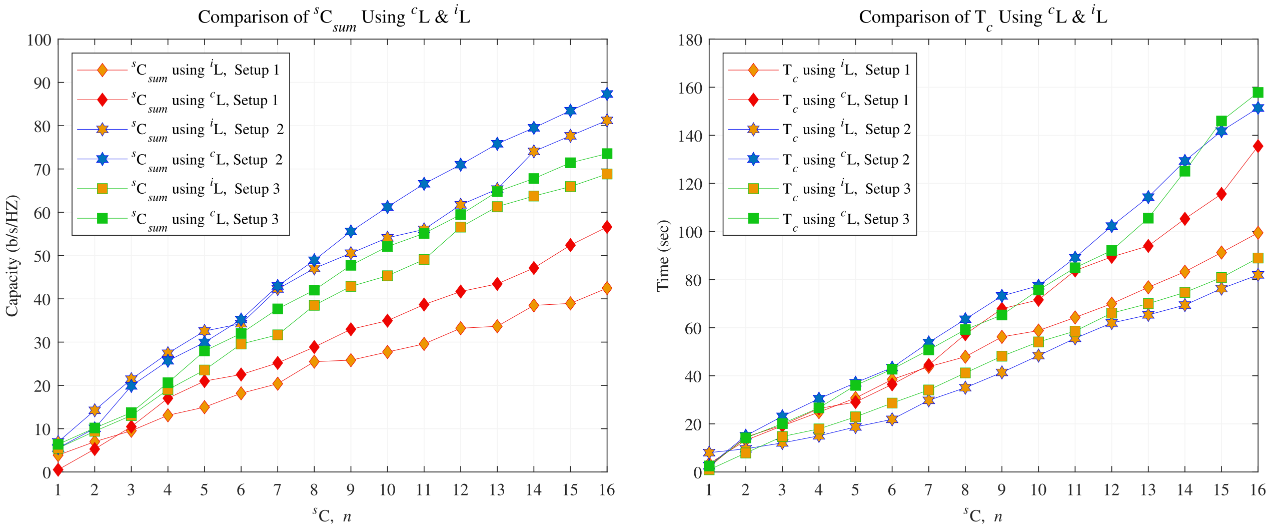
A. CAPACITY OF MACROCELL USERS

The capacity of macrocell users is computed using the QL in both learning paradigms, ^{i}L and ^{c}L, in all simulation setups of Fig. 3 and is presented in Fig. 4a. From Fig. 4a, it can be observed that the QL algorithm performed closely in both learning paradigms and provided the required capacity to the macrocell users for QoS. Hence, QoS requirements can be met through either of the learning schemes. The performance comparison is also summarized in Table 3. ^{i}L performed slightly better than the ^{c}L by providing 2.97% higher capacity



(a) (b)

FIGURE 4. Performance comparison of QL algorithm using ${}^{i}L$ and ${}^{c}L$ in all simulation setups in Fig. 3 (a) comparison of ${}^{m}C_{i}$, and (b) comparison of ${}^{s}C_{n,k}$.



(a) (b)

FIGURE 5. Performance comparison of QL algorithm using ${}^{i}L$ and ${}^{c}L$ in all simulation setups in Fig. 3 (a) comparison of ${}^{s}C_{sum}$, and (b) comparison of T_{c}.

to macrocell users in high I^{co} and I^{cr} setup, Fig. 3a. Whereas in the other two setups Fig. 3a and 3b, ${}^{c}L$ performed better than ${}^{i}L$ with 5.6% and 4.60% increase in macrocell user capacity respectively. Although the performance difference is negligible, ${}^{c}L$ provided higher macrocell user capacity. The similar performance of both learning paradigms, ${}^{i}L$ and ${}^{c}L$, for capacity of macrocell users is due to almost negligible impact of cooperation of ${}^{s}C$ in ${}^{c}L$ on I^{cr} mitigation.

B. MINIMUM CAPACITY OF SMALL CELL USERS

Both ${}^{i}QL$ and ${}^{c}QL$ provided small cell user capacity above the minimum required capacity threshold to ensure QoS for small cell users in simulation setups 1-3, Fig. 3a-3c. However, cooperation among the neighboring small cells in the clusters has shown a significant impact on the small cell capacity. Fig. 4b presents the performance comparison for the minimum capacity of small cells using ${}^{c}L$ and ${}^{i}L$ based QL in simulation setups 1-3, Fig. 3a-3c. In all three simulation setups, the ${}^{c}QL$ algorithm performed significantly better than the ${}^{i}QL$ algorithm in the same setup as shown in Fig.4b. Using ${}^{c}QL$, a minimum improvement of 23.61% is observed in setup

TABLE 4. Comparison of ${}^{s}C_{n,k}$ (b/s/Hz) using ${}^{i}L$ and ${}^{c}L$.

Setup	${}^{i}QL$	${}^{c}QL$	${}^{c}QL$ based Analysis
1	1.42	1.86	↑30.90%
2	1.40	2.08	↑ 48.57%
3	1.44	1.78	↑ 23.61%

${}^{c}QL$ performed better than ${}^{i}QL$

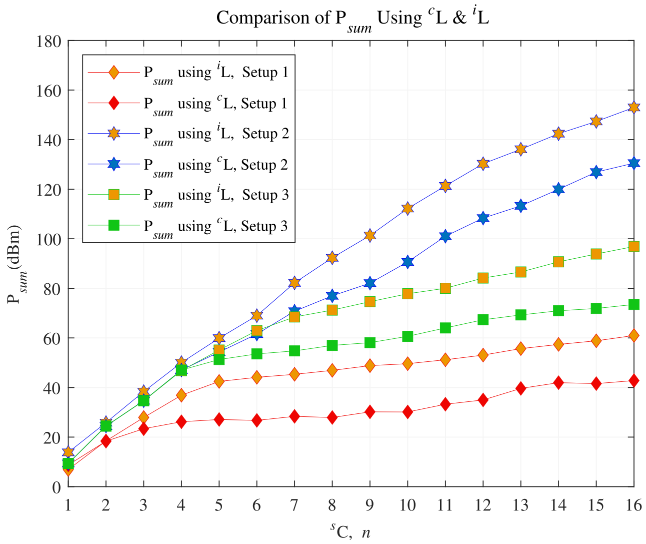


FIGURE 6. Comparison of P_{sum} using ^{i}QL and ^{c}QL in setup 1-3.

TABLE 5. Comparison of $^sC_{sum}$ (b/s/Hz) using ^{i}L and ^{c}L.

Setup	^{i}QL	^{c}QL	^{c}QL based Analysis
1	42.80	58.99	↑ 37.9%
2	81.22	87.32	↑ 7.40%
3	68.84	73.54	↑ 6.287%

^{c}QL performed better than ^{i}QL

TABLE 6. Comparison of T_c using ^{i}L and ^{c}L.

Setup	^{i}QL	^{c}QL	^{c}QL based Analysis
1	99.45	135	↑35.70%
2	81.95	151	↑ 84.25%
3	88.96	157	↑ 77.56%

^{i}QL performed better than ^{c}QL

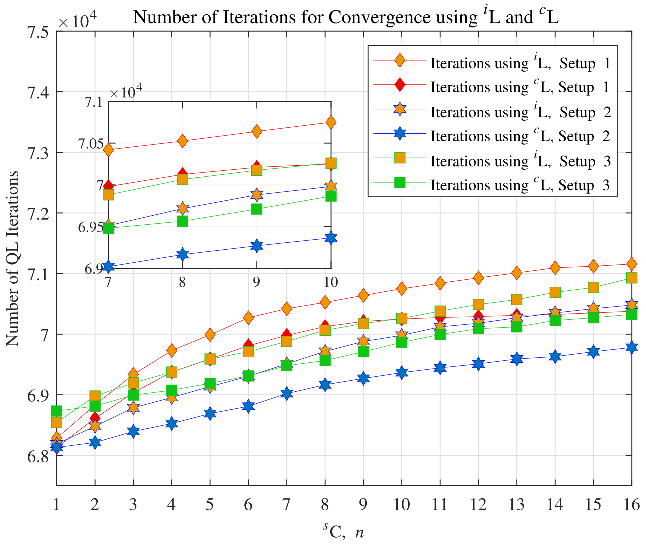


FIGURE 7. Comparison of the number of iterations for convergence of proposed algorithm using ^{i}QL and ^{c}QL in setup 1-3.

TABLE 7. Comparison of P_{sum} (dBm) using ^{i}L and ^{c}L.

Setup	^{i}QL	^{c}QL	^{c}QL based Analysis
1	60.98	42.79	↓ 29.71%
2	152.95	130.50	↓ 17.20%
3	96.85	70.50	↓ 27.20%

^{i}QL performed better than ^{c}QL

TABLE 8. Comparison of number of iterations using ^{i}L and ^{c}L for 16 ^sC.

Setup	^{i}QL	^{c}QL	^{c}QL based Analysis
1	71158	70378	↓ 1.10%
2	70480	69781	↓ 0.99%
3	70928	70328	↓ 0.85%

^{i}QL performed better than ^{c}QL

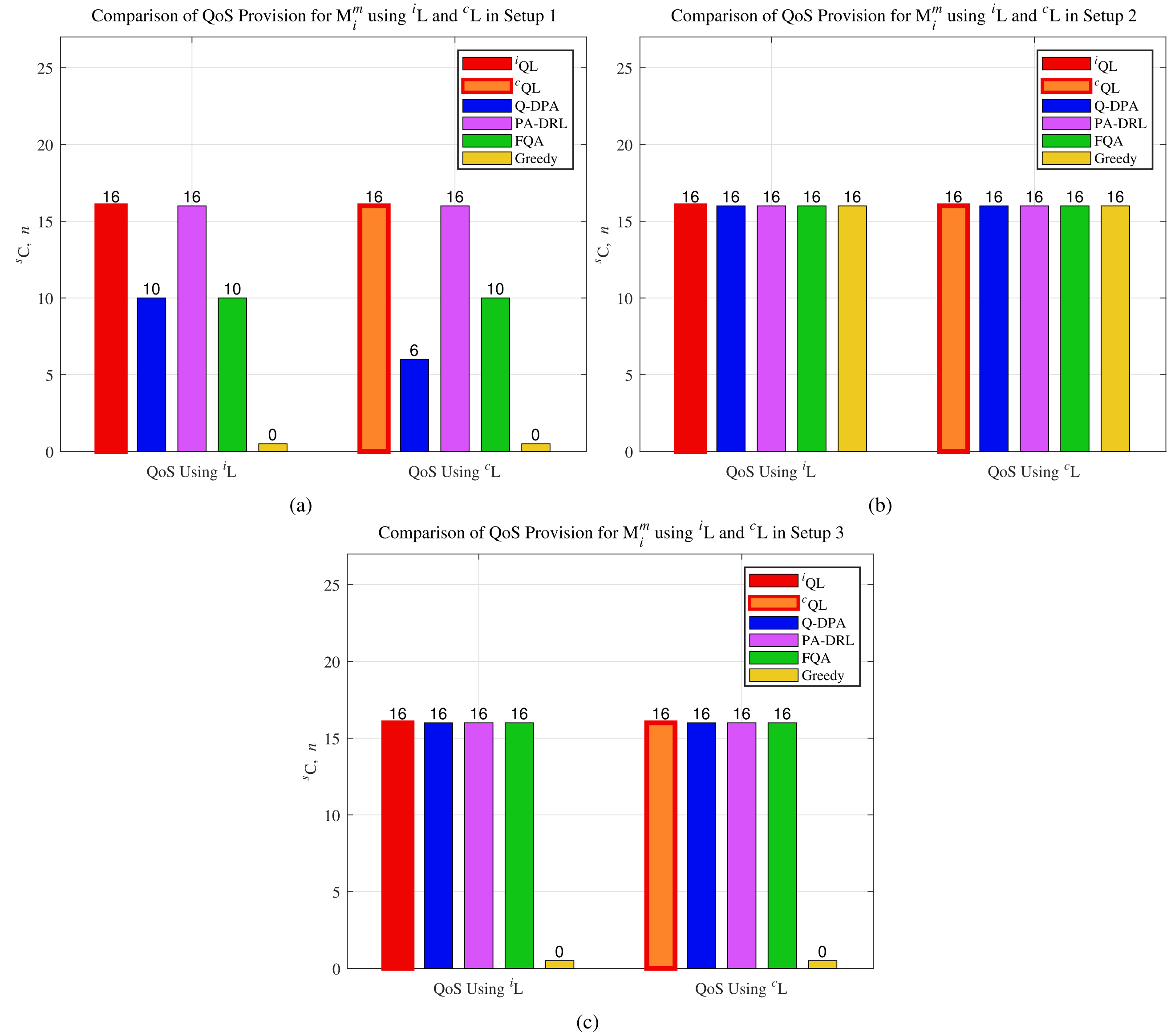
3 whereas the maximum improvement is in setup 2 which is 48.57%. The performance comparison of ^{i}QL and ^{c}QL is summarized in Table 4 which establishes that ^{c}QL provides a higher minimum capacity of small cells in UDHN as the density of the small cells increases. The improvement in the minimum capacity of small cell users using the ^{c}L paradigm is due to significant reduction in I^{co} by efficient cooperation among the ^{s}C.

C. SUM CAPACITY OF SMALL CELL USERS

Although the sum capacity of small cell users is not a QoS parameter in UDHN but a higher value of sum capacity represents that the utilized RRM technique can provide a higher minimum capacity to all small cells in the UDHN. As the sum capacity is the sum of capacities of all small cell users in the system, therefore an improvement in the minimum capacities of an individual small cell user is reflected in sum capacity. As the ^{c}L has improved the minimum capacity of small cell users in all interference setups in Fig. 3a-3c, therefore sum capacity provided by ^{c}L is also higher than the ^{i}L. The comparison of sum capacity using ^{i}L and ^{c}L is presented in Fig. 5a. The minimum improvement in sum capacity is for setup 3 which is 6.287% and the maximum improvement is in the highest interference setup 1 which is 37.90%. The improvement in sum capacity using ^{c}L for all interference setups is summarized in Table 5. From Table 5, it can be inferred that ^{c}L provided higher sum capacity as compared to ^{i}L and therefore can be opted as the optimal learning policy.

D. COMPUTATIONAL TIME

The significant improvements in minimum and sum capacities of small cells, and competitive performance for macrocell user capacity using the ^{c}L based QL algorithm as compared to the ^{i}L are at the cost of the increase in computational time, and overhead. In the ^{c}L, all the cooperating agents transmit and receive the entries of QT which result in communication overhead. The communication overhead is directly proportional to the density of small cells in the system. The increased communication overhead in the ^{c}L paradigm also increases the computational time as compared to the ^{i}L paradigm.



(a)

(b)

(c)

FIGURE 8. Comparison of QoS for M_i^u as function of n in ^{i}L and ^{c}L paradigm in (a) simulation setup 1, Fig. 3a, (b) simulation setup 2, Fig. 3b, and (c) simulation setup 3, Fig. 3c.

However, the increase in computational time is not significant as it is compensated by the decreased learning time due to the availability of prior information to small cells in form of QT rows shared by the neighboring small cells. The computational time of the proposed ^{c}L based QL algorithm is significantly less as compared to other CL algorithms in literature but ^{i}L has slightly less computational time as compared to ^{c}L which is evident in Fig. 5b. A similar trend is observed for the computational time in all three simulation setups 1-3, Fig. 3a-3c, using ^{c}L and ^{i}L. The analysis of the increase in computational time using ^{c}QL as compared to ^{i}QL is summarized in Table 6.

E. SUM POWER OF SMALL CELLS

Small cells transmitting at higher power levels result in strong I^{co} and I^{cr} and also reduce the EE of the system. ^{i}L reduced the sum power of the small cells operating in the cluster and also provided capacities above the minimum required thresholds to maintain minimum QoS for macrocell and small cell users. However, the cooperation among the small cells through ^{c}L further reduced the sum transmit power in all three simulation setups 1-3, Fig. 5a. The decrease in sum transmit power is indirectly related to improvements in the capacities of the macrocell, and small cells. A lower value of the sum transmit power in conjunction with improvements in the capacities of the macrocell, and small cell indicates that the proposed solution has the capability of mitigating I^{co} and I^{cr} simultaneously through adaptive power allocation to BS^s in UDHN. The decrease in sum transmit power using the ^{c}L is summarized in Table 7 which shows a minimum decrease in sum transmit power of 17.20% and a maximum decrease of 29.71% in setups 2 and 1 respectively.

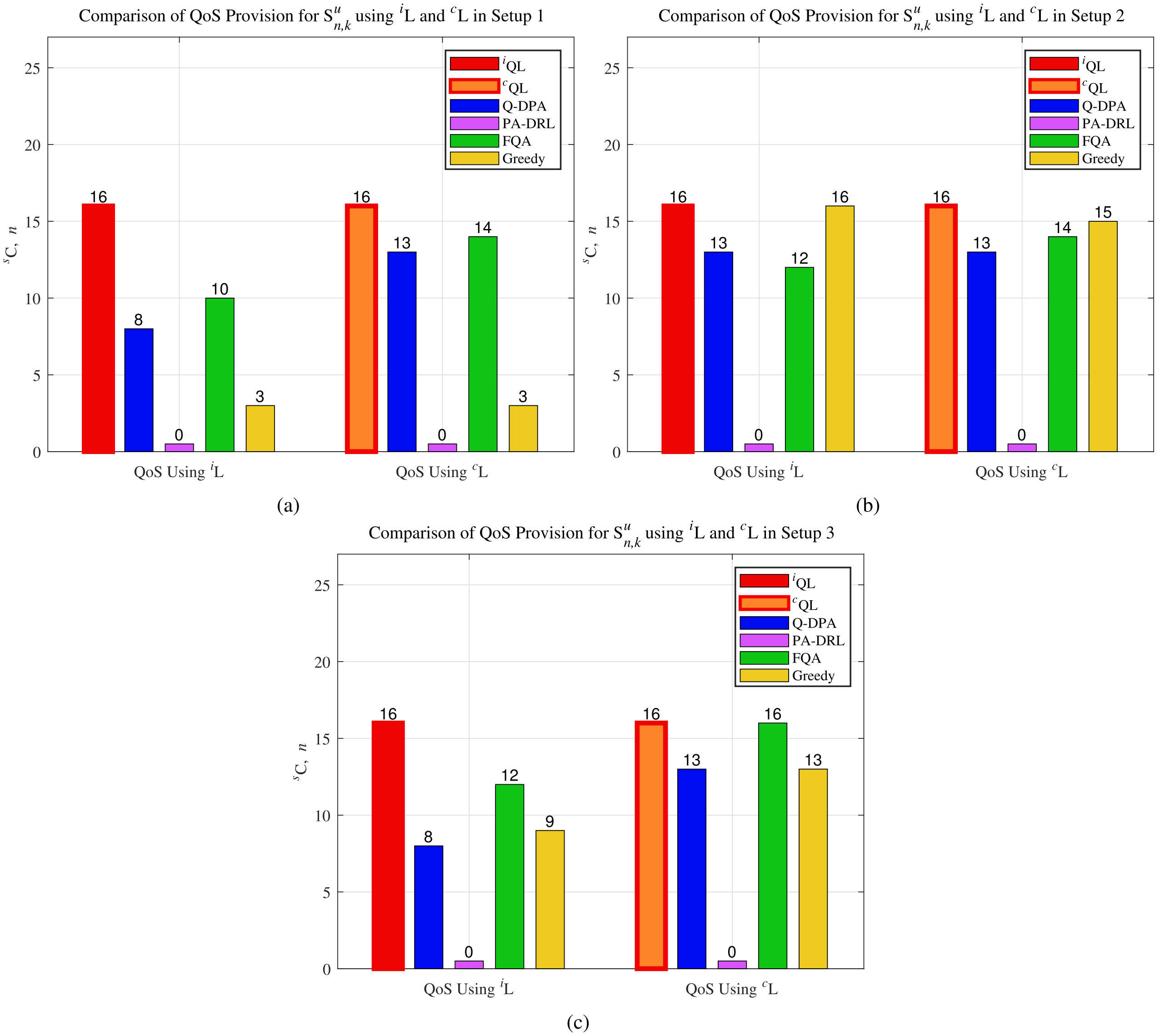


FIGURE 9. **Comparison of QoS for $S^u_{n,k}$ as function of n in ^{i}L and ^{c}L paradigm in (a) simulation setup 1, Fig. 3a, (b) simulation setup 2, Fig. 3b, and (c) simulation setup 3, Fig. 3c.**

F. CONVERGENCE ANALYSIS

For the convergence of the proposed algorithm in ^{i}L and ^{c}L paradigm, the ϵ-greedy policy is utilized based on the (8) and (11). To learn optimal policy (π^*) through convergence of the Q-value to optimal Q*-value, the QL parameters affecting the convergence, like learning rate (*alpha*), exploration probability (ϵ), finite state-action space, and MDP satisfying "Markov Property", are selected in line with the literature [32], [33], [34]. The simulation parameters are summarized in Table 2. The QL algorithm is considered to be converged if the regret magnitude in (11) is less than 0.001 for consecutive 1000 iterations where the maximum number of QL iterations is 75×10^3. The convergence of the regret to the defined threshold is presented in Appendix for both ^{i}L and ^{c}L in different simulation setups. It can be observed in Fig. 12a-12c that regret magnitude converges to the required threshold and then remains constant after a certain number of QL iterations. The regret minimization to the threshold levels is also shown in the magnified views in Fig. 12a-12c. The QL iterations for convergence of ^{i}L and ^{c}L are presented in Fig. 7 for all three simulation setups. The statistical comparison of the number of iterations for convergence is also presented in Table 8. It can be observed that both ^{i}L and ^{c}L converge in less number of iterations without reaching the maximum limit of QL iterations. However, the overall pattern of the number of iterations for convergence remains similar. A slight difference between number of iterations using ^{i}L and ^{c}L is due to the learning paradigm where ^{c}L converges slightly earlier than ^{i}L.

VII. QoS PROVISION USING IL AND CL

In UDHN, QoS provision to the macrocell and small cell users simultaneously through an effective interference mitigation scheme is a difficult task and one of the fundamental

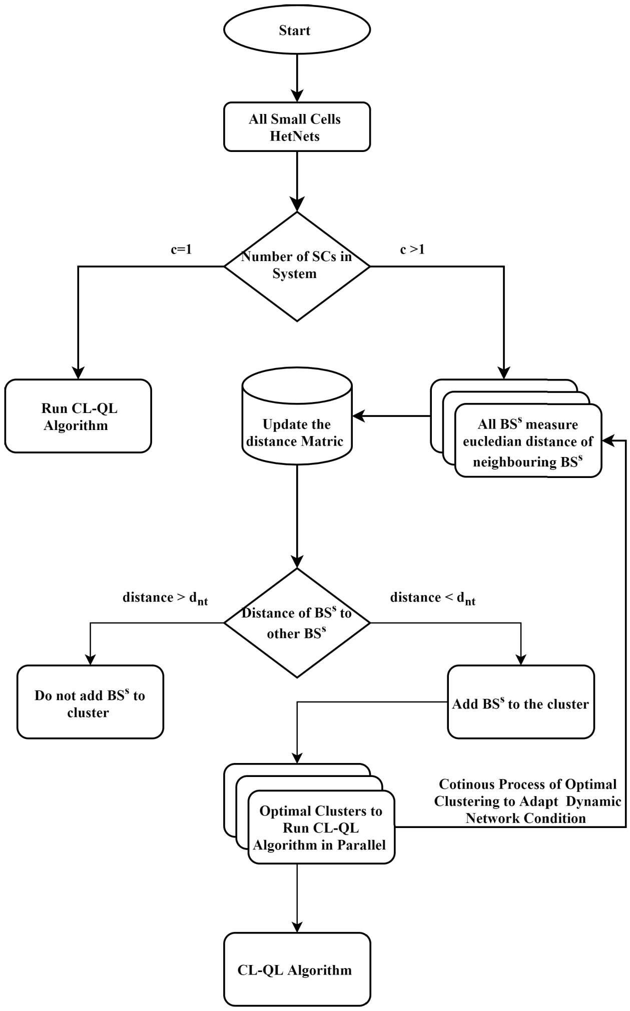

Start
All Small Cells HetNets
Number of SCs in System
c=1
c >1
Run CL-QL Algorithm
Update the distance Matric
All BS^{s} measure eucledian distance of neighbouring BS^{s}
distance > d$_{nt}$
distance < d$_{nt}$
Distance of BS^{s} to other BS^{s}
Do not add BS^{s} to cluster
Add BS^{s} to the cluster
Optimal Clusters to Run CL-QL Algorithm in Parallel
Cotinous Process of Optimal Clustering to Adapt Dynamic Network Condition
CL-QL Algorithm

FIGURE 10. Flow chart of clustering based ^{c}QL algorithm.

objectives of this research. In this section, we analyzed the performance of the QL using ^{i}L and ^{c}L for the provision of QoS to macrocell and small cell users simultaneously. The performance of ^{i}L and ^{c}L is also compared with other recently proposed solutions in literature Q-DPA [32], PA-DRL [33], FAQ [34] and Greedy in ^{i}L and ^{c}L paradigm.

A. QoS FOR MACROCELL USERS

QoS provision for macrocell users by the ^{i}QL, ^{c}QL and other recently proposed solutions in literature [32], [33], [34] is presented in Fig. 8 for all three simulation setups, Fig. 3a-3c. It can be observed from Fig. 8a-8c that proposed ^{i}QL and ^{c}QL algorithms provided QoS to macrocell users in all simulation setups in presence of a cluster of sixteen small cells whereas the other recently proposed solutions in the literature [32], [33], [34] failed to provide QoS in high interference setup, i.e. simulation setup 1. Therefore, the proposed QL based solution can effectively mitigate the I^{cr} and can provide QoS to macrocell users in highly dense urban UDHN. However, there is no significant difference in the provision of QoS to macrocell users in ^{i}L or ^{c}L paradigm which is also evident from Fig. 4a.

B. QoS FOR SMALL CELL USERS

In UDHN, small cell users are victims of high I^{cr} and I^{co}. The QL based proposed in ^{i}L paradigm not only provided QoS to all small cell users in the cluster of sixteen ^{s}C in all simulation setups but also outperformed the recently proposed solution [32], [33], [34] by providing QoS to the higher number of small cells in different interference setups. Similarly, ^{c}QL also not only provided QoS to all small cell users but also improved the minimum and sum capacities of small cell users where the other recently proposed solution in literature [32], [33], [34] could not provide QoS to all small cell users in ^{c}L paradigm as well.

VIII. OPTIMAL LEARNING PARADIGM FOR QL BASED RRM IN UDHN

From the performance comparison of various KPIs in section VI VII, it can be observed that ^{c}L has a negligible effect on the capacity of macrocell users as compared to ^{i}L in the case of UDHN. However, a significantly higher minimum capacity of small cell users and the sum capacity of the cooperating small cells in the cluster can be observed using the ^{c}L as compared to ^{i}L. This significant improvement is at the cost of increased T_c which is directly related to the number of small cells in the cluster. In this research, we simulated a cluster size of 16 small cells, which are 37.5% more small cells according to 3GPP TR36.872 by adding small cells in the cluster one by one. Simulation results show that the proposed ^{c}L based QL algorithm not only outperformed other recently proposed solutions in literature but it proved its significance over the ^{i}L paradigm. Therefore, ^{c}L is an optimal learning strategy for QL based RRM in UDHN.

IX. OPTIMAL CLUSTERING IN ^{c}L BASED QL

Although simulation results and their comparison have set the superiority of ^{c}L over the ^{i}L, the issues of increased computational time and overhead can be handled through optimal cooperative clustering in ^{s}C UDHN to further improve the performance of ^{c}L based QL. An optimal clustering algorithm is proposed in Fig. 10 to handle computational time and overhead. The optimal clustering technique finds an optimal size of the cooperative cluster which always guarantees an optimal size of the cluster to successfully handle the issue of increased computational time in cooperative learning. The simulation results of optimal clustering in terms of computation time and overhead and the optimal size of the cluster are discussed in the subsequent sections.

A. IMPACT OF CLUSTERING ON COMPUTATIONAL TIME

The simulation results for the proposed ^{c}L based QL algorithm and optimal clustering algorithm to evaluate the impact of reduced cluster size on computational time are presented in Fig. 11b. To evaluate the proposed solution, the simulation setup 1, Fig. 3a is utilized as it is the combination of the highest I^{co} and I^{cr}. The proposed optimal clustering algorithm divided the cluster of total small cells, i.e. 16, into multiple overlapping smaller clusters of 2, 3, and 4 small cells. The

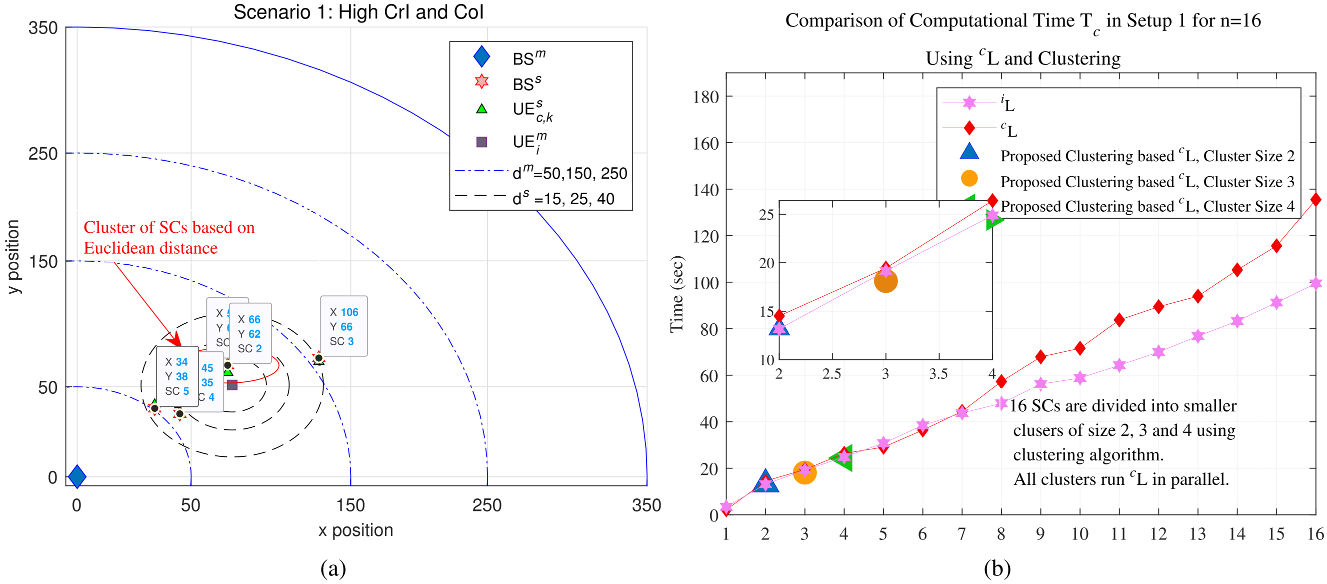


FIGURE 11. Comparison of QoS for $S^u_{n,k}$ as function of n in ^{i}L and ^{c}L paradigm in (a) simulation setup 1, Fig. 3a, (b) simulation setup 2, Fig. 3b, and (c) simulation setup 3, Fig. 3c.

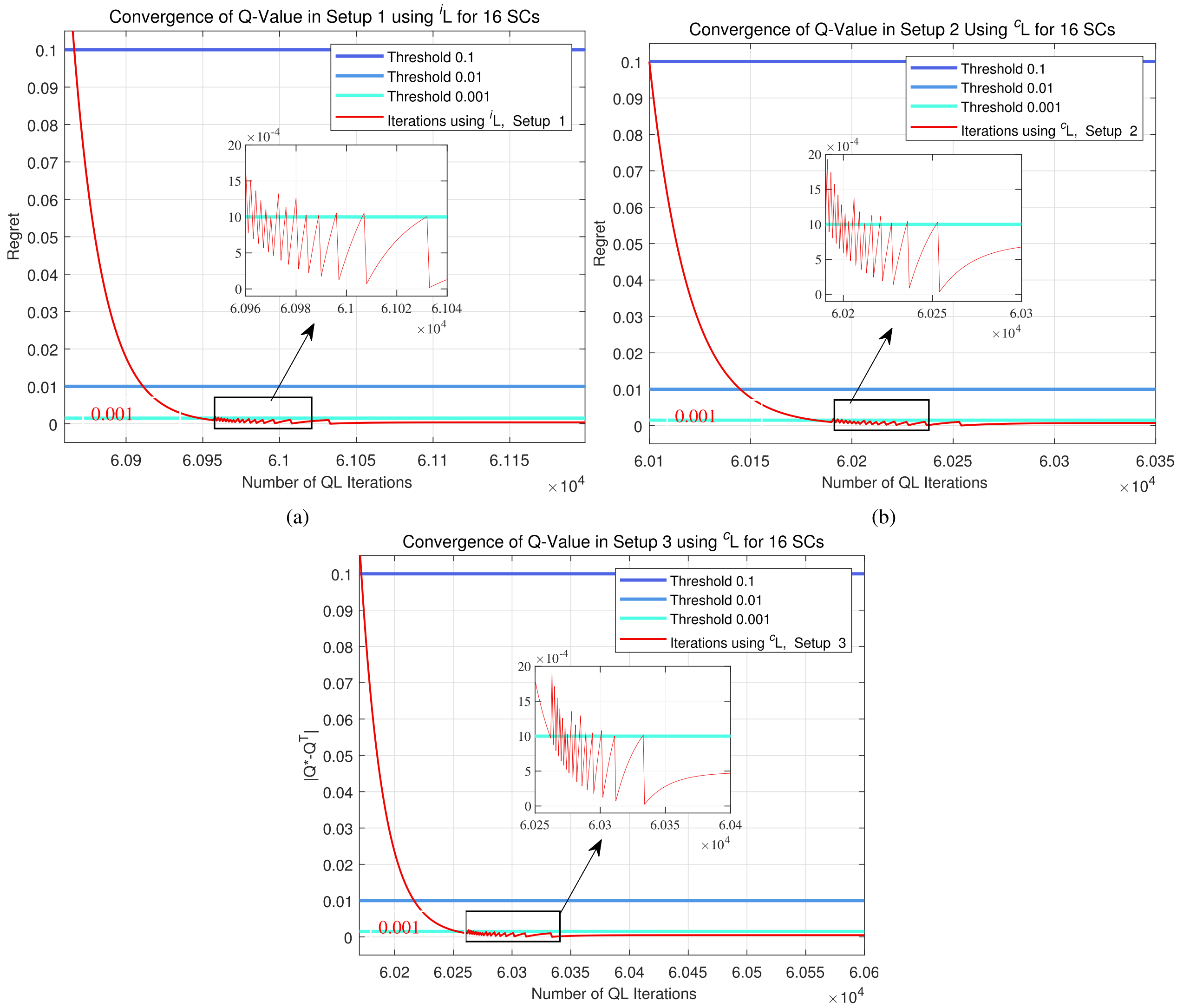
optimal clustering in simulation setup 1, Fig. 3a is presented in Fig. 11a. Therefore, the largest cluster size in terms of small cells is 4. However, there is no limit on the number of smaller overlapping clusters. All the clusters executed the ^{c}QL in parallel and therefore the largest cluster computational time is the maximum computational time. It is pertinent to mention that all small cells are always independent and work in parallel whether the learning scheme is independent or cooperative. Therefore, there is no such case where the small cells are not working in parallel. In the Monte-Carlo simulation, the maximum size remained limited to four small cells, and the computational time for this cluster is 24 sec as shown in Fig. 4.10. The computational time for 16 ^{s}C in the system is now reduced to approximately equal to 4 small cells in the system. Therefore, the ^{c}QL algorithm with optimal clustering is four times faster than a simple ^{c}L application in UDHN. The comparison of ^{c}L based QL and optimal clustering based ^{c}L QL is presented in Fig. 11b. The results for the optimal clustering in simulation setup 2-3, Fig. 3a-3c can be obtained similarly.

B. IMPACT OF CLUSTERING ON COMPUTATIONAL OVERHEAD

In ^{c}QL, the computational overhead is due to the large size of QT which is the combination of x^t_n and a^t_c. The size of the QT increases with the number of small cells in the system as they share it with cooperating agents. When all the small cells operate in a single cluster or unlimited size of the cluster, the size of the QT is large and therefore computational overhead while executing the QL algorithm. The optimal clustering size reduces the number of small cells in the system and thus the size of the QT which as result not only reduces computational overhead but also computational time. The optimal clustering algorithm, Fig. 10, divided the total 16 small cells into small clusters. The largest cluster size is composed of 04 small cells, therefore optimal clustering reduced computational overhead by four times as compared to using ^{c}L in a single cluster of 16 small cells in simulation setup Fig. 11a. However, the reduction in communication overhead can be varied due to the deployment scenario, the total number of 16 small cells in the system, and the dynamic behavior of the UDHN.

X. CONCLUSION

In this research article, the Q-Learning algorithm in the independent and cooperative learning paradigm is explored for RRM to handle the co-tier and cross-tier interferences simultaneously in UDHN for quality of service provision to both macro and small cell users. In the cooperative learning paradigm, learning agents share independently learned information with the other neighboring agents in the cluster and utilize their mutual experience to learn optimal policy by meeting the convergence conditions to improve system KPIs jointly. The Q-Learning algorithm successfully mitigated the co-tier and cross-tier interferences simultaneously in both independent and cooperative learning paradigms and provided QoS to all users in K-tiers in the cluster of 16 small cells in various setups of standard 3GPP interference scenarios where other recently proposed solutions in literature and greedy power allocation fail to meet the QoS requirements for both macro and small cell users at the same time. Cooperative learning provides higher macrocell and small cell users capacity and sum capacity of small cell users as compared to independent learning at the cost of increased computational time and sum power of small cells. Although the impact of cooperative learning is not very large on the capacity of macro cell users as compared to independent learning, significant improvement can be observed in small cell capacity in the



(a)

(b)

(c)

FIGURE 12. Converge of Q-Value using proposed algorithm in ^{i}L and ^{c}L paradigm (a) simulation setup 1, Fig. 3a, (b) simulation setup 2, Fig. 3b, and (c) simulation setup 3, Fig. 3c.

case of UDHN. A significant improvement of 48.57% and 37.9% in the small cell capacity set that ^{c}L is the optimal learning scheme of QL based RRM in UDHN for QoS provision. The performance improvement of ^{c}L results in increased computational time and overhead which are directly proportional to the size of the cluster. To handle the issue of computational time and overhead, an optimal clustering algorithm is proposed and evaluated. The optimal clustering in combination with a superior learning scheme, ^{c}L, improved the robustness and reduced computational complexity by a factor of 4 in a cluster size of 16 small cells, 37.5% more small cells according to 3GPP TR36.872. However, improvement in robustness and decrease in computational overhead is directly proportional to the number of small cells in UDHN.

APPENDIX

The convergence of regret to the required threshold using the proposed QL algorithm based on the ϵ-greedy policy in ^{i}L and ^{c}L paradigm for simulation scenarios, Fig. 3, is presented in Fig. 12 for the convenience of the reader.

REFERENCES

[1] R. N. Mitra and D. P. Agrawal, "5G mobile technology: A survey," *ICT Exp.*, vol. 1, no. 3, pp. 132–137, Dec. 2015.

[2] N. Panwar, S. Sharma, and A. K. Singh, "A survey on 5G: The next generation of mobile communication," *Phys. Commun.*, vol. 18, pp. 64–84, Mar. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1874490715000531

[3] I. F. Akyildiz, S. Nie, S.-C. Lin, and M. Chandrasekaran, "5G roadmap: 10 key enabling technologies," *Comput. Netw.*, vol. 106, pp. 17–48, Sep. 2016. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S1389128616301918

[4] A. Gupta and R. K. Jha, "A survey of 5G network: Architecture and emerging technologies," *IEEE Access*, vol. 3, pp. 1206–1232, 2015.

[5] S. Manap, K. Dimyati, M. N. Hindia, M. S. A. Talip, and R. Tafazolli, "Survey of radio resource management in 5G heterogeneous networks," *IEEE Access*, vol. 8, pp. 131202–131223, 2020.

[6] C. Niu, Y. Li, R. Q. Hu, and F. Ye, "Fast and efficient radio resource allocation in dynamic ultra-dense heterogeneous networks," *IEEE Access*, vol. 5, pp. 1911–1924, 2017.

[7] M. A. Adedoyin and O. E. Falowo, "Combination of ultra-dense networks and other 5G enabling technologies: A survey," *IEEE Access*, vol. 8, pp. 22893–22932, 2020.

[8] T. O. Olwal, K. Djouani, and A. M. Kurien, "A survey of resource management toward 5G radio access networks," *IEEE Commun. Surveys Tuts.*, vol. 18, no. 3, pp. 1656–1686, 3rd Quart., 2016.

[9] M. E. M. Cayamcela and W. Lim, "Artificial intelligence in 5G technology: A survey," in *Proc. Int. Conf. Inf. Commun. Technol. Converg. (ICTC)*, Oct. 2018, pp. 860–865.

[10] K. Jin, X. Cai, J. Du, H. Park, and Z. Tang, "Toward energy efficient and balanced user associations and power allocations in multiconnectivity-enabled mmWave networks," *IEEE Trans. Green Commun. Netw.*, vol. 6, no. 4, pp. 1917–1931, Dec. 2022.

[11] Z. Li, M. Chen, K. Wang, C. Pan, N. Huang, and Y. Hu, "Parallel deep reinforcement learning based online user association optimization in heterogeneous networks," in *Proc. IEEE Int. Conf. Commun. Workshops (ICC Workshops)*, Jun. 2020, pp. 1–6.

[12] G. Yu, R. Liu, Q. Chen, and Z. Tang, "A hierarchical SDN architecture for ultra-dense millimeter-wave cellular networks," *IEEE Commun. Mag.*, vol. 56, no. 6, pp. 79–85, Jun. 2018.

[13] M.-C. Chuang, J. H. Liu, C.-M. Lin, and C.-L. Chen, "STASH: SDN-based trajectory-aware seamless handover scheme with multiple antennas in ultra-dense smallcell networks," in *Proc. IEEE Region Conf. (TENCON)*, 2018, pp. 1295–1300.

[14] S. Dutta, M. Mezzavilla, R. Ford, M. Zhang, S. Rangan, and M. Zorzi, "Frame structure design and analysis for millimeter wave cellular systems," *IEEE Trans. Wireless Commun.*, vol. 16, no. 3, pp. 1508–1522, Mar. 2017.

[15] M. U. Iqbal, E. A. Ansari, and S. Akhtar, "Interference mitigation in HetNets to improve the QoS using Q-learning," *IEEE Access*, vol. 9, pp. 32405–32424, 2021.

[16] M. U. Iqbal, E. A. Ansari, S. Akhtar, and A. N. Khan, "Improving the QoS in 5G HetNets through cooperative Q-learning," *IEEE Access*, vol. 10, pp. 19654–19676, 2022.

[17] L. Xiao, H. Zhang, Y. Xiao, X. Wan, S. Liu, L.-C. Wang, and H. V. Poor, "Reinforcement learning-based downlink interference control for ultra-dense small cells," *IEEE Trans. Wireless Commun.*, vol. 19, no. 1, pp. 423–434, Jan. 2020.

[18] M. Zangooei, N. Saha, M. Golkarifard, and R. Boutaba, "Reinforcement learning for radio resource management in ran slicing: A survey," *IEEE Commun. Mag.*, vol. 61, no. 2, pp. 118–124, Feb. 2023.

[19] *Evolved Universal Terrestrial Radio Access (E-UTRA) and Evolved Universal Terrestrial Radio Access Network (E-UTRAN), Overall Description*, document TS 36.300, Version 16.3.0, Release 8, 3GPP, Oct. 2020. [Online]. Available: https://portal.3gpp.org/

[20] M. Nohrborg. (Oct. 2020). *Self-Organizing Networks*. [Online]. Available: https://www.3gpp.org/technologies/keywords-acronyms/105-son

[21] M. Dirani, Z. Altman, and M. Salaun, "Autonomics in radio access networks," in *Autonomic Network Management Principles*, N. Agoulmine, Ed. New York, NY, USA: Academic, 2011, ch. 7, pp. 141–166. [Online]. Available: https://www.sciencedirect.com/science/article/pii/B9780123821904000073, doi: 10.1016/B978-0-12-382190-4.00007-3.

[22] H. Nouira. (Mar. 2015). *SON in LTE: The What, the Where, and the Why*. [Online]. Available: https://www.nokia.com/blog/son-lte-what-where-and-why/

[23] M. Pesavento and F. Bahlke, "Machine learning for optimal resource allocation," in *Machine Learning for Future Wireless Communications*. Hoboken, NJ, USA: Wiley, 2020, ch. 5, pp. 85–103. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9781119562306.ch5, doi: 10.1002/9781119562306.ch5.

[24] A. Galindo-Serrano and L. Giupponi, "Distributed Q-learning for interference control in OFDMA-based femtocell networks," in *Proc. IEEE 71st Veh. Technol. Conf.*, May 2010, pp. 1–5.

[25] J. R. Tefft and N. J. Kirsch, "A proximity-based Q-learning reward function for femtocell networks," in *Proc. IEEE 78th Veh. Technol. Conf. (VTC Fall)*, Sep. 2013, pp. 1–5.

[26] B. Wen, Z. Gao, L. Huang, Y. Tang, and H. Cai, "A Q-learning-based downlink resource scheduling method for capacity optimization in LTE femtocells," in *Proc. 9th Int. Conf. Comput. Sci. Educ.*, Aug. 2014, pp. 625–628.

[27] H. Saad, A. Mohamed, and T. El Batt, "Distributed cooperative Q-learning for power allocation in cognitive femtocell networks," in *Proc. IEEE Veh. Technol. Conf. (VTC Fall)*, Sep. 2012, pp. 1–5.

[28] H. Saad, A. Mohamed, and T. El Batt, "A cooperative Q-learning approach for online power allocation in femtocell networks," in *Proc. IEEE 78th Veh. Technol. Conf. (VTC Fall)*, Sep. 2013, pp. 1–6.

[29] J. R. Tefft and N. J. Kirsch, "Accelerated learning in machine learning-based resource allocation methods for heterogenous networks," in *Proc. IEEE 7th Int. Conf. Intell. Data Acquisition Adv. Comput. Syst. (IDAACS)*, vol. 1, Sep. 2013, pp. 468–473.

[30] R. Amiri and H. Mehrpouyan, "Self-organizing mm wave networks: A power allocation scheme based on machine learning," in *Proc. 11th Global Symp. Millim. Waves (GSMM)*, May 2018, pp. 1–4.

[31] R. Amiri, H. Mehrpouyan, L. Fridman, R. K. Mallik, A. Nallanathan, and D. Matolak, "A machine learning approach for power allocation in HetNets considering QoS," in *Proc. IEEE Int. Conf. Commun. (ICC)*, May 2018, pp. 1–7.

[32] R. Amiri, M. A. Almasi, J. G. Andrews, and H. Mehrpouyan, "Reinforcement learning for self organization and power control of two-tier heterogeneous networks," *IEEE Trans. Wireless Commun.*, vol. 18, no. 8, pp. 3933–3947, Aug. 2019.

[33] Q. Su, B. Li, C. Wang, C. Qin, and W. Wang, "A power allocation scheme based on deep reinforcement learning in HetNets," in *Proc. Int. Conf. Comput., Netw. Commun. (ICNC)*, Feb. 2020, pp. 245–250.

[34] W. AlSobhi and A. H. Aghvami, "QoS-aware resource allocation of two-tier HetNet: A Q-learning approach," in *Proc. 26th Int. Conf. Telecommun. (ICT)*, Apr. 2019, pp. 330–334.

[35] R. S. Sutton and A. G. Barto, *Reinforcement Learning: An Introduction*, 2nd ed. Cambridge, MA, USA: MIT Press, 2018. [Online]. Available: http://incompleteideas.net/book/the-book-2nd.html

[36] M. Peng, C. Wang, J. Li, H. Xiang, and V. Lau, "Recent advances in underlay heterogeneous networks: Interference control, resource allocation, and self-organization," *IEEE Commun. Surveys Tuts.*, vol. 17, no. 2, pp. 700–729, 2nd Quart., 2015.

[37] 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. Surveys Tuts.*, vol. 19, no. 4, pp. 2392–2431, 4th Quart., 2017.

[38] *Small Cell Enhancements for E-UTRA and E-UTRAN—Physical Layer Aspects*, document TR 36.872, 3GPP, Version 12.1.0, Release 12, Dec. 2013. [Online]. Available: https://portal.3gpp.org/

[39] B. Abuhaija, "Performance analysis of LTE multiuser flat downlink power spectrum and radio resources scheduling," *J. High Speed Netw.*, vol. 18, no. 3, pp. 173–184, 2012.

[40] M. Tan, *Multi-Agent Reinforcement Learning: Independent vs. Cooperative Agents*. San Francisco, CA, USA: Morgan Kaufmann, 1997, pp. 487–494.

[41] M. Kearns and S. Singh, "Finite-sample convergence rates for Q-learning and indirect algorithms," in *Proc. Adv. Neural Inf. Process. Syst.*, vol. 11, 1998, pp. 1–7.

[42] M. T. Regehr and A. Ayoub, "An elementary proof that Q-learning converges almost surely," 2021, *arXiv:2108.02827*.

[43] *Evolved Universal Terrestrial Radio Access (E-UTRA); Further Advancements for E-UTRA Physical Layer Aspects*, document TR 36.814, 3GPP, Version 9.2.0, Release 9, Mar. 2017. [Online]. Available: https://portal.3gpp.org/

MUHAMMAD USMAN IQBAL received the B.Sc. degree in electrical engineering from the University College of Engineering and Technology (UCE&T), Bahauddin Zakariya University, Multan, Pakistan, in 2009, the M.S. degree in electrical engineering from the School of Electrical Engineering and Computer Science (SEECS), National University of Science and Technology (NUST), Islamabad, Pakistan, in 2013, and the Ph.D. degree in electrical engineering from the Department of Electrical and Computer Engineering (ECE), COMSATS University Islamabad (CUI), Lahore Campus, Pakistan, in 2023. He was a Lecturer in electrical engineering with the University College of Textile Engineering (UCTE), Bahauddin Zakariya University, from June 2009 to June 2014. He has been a Lecturer with the Department of ECE, CUI, Lahore Campus, since July 2014. His research interests include digital signal processing, digital image processing, wireless communications, and deep learning.

EJAZ AHMAD ANSARI received the B.Sc. degree (Hons.) in electrical engineering from the University of Engineering and Technology (UET), Lahore, Pakistan, in 1990, the M.Sc. degree in electrical engineering from the Georgia Institute of Technology (Georgia Tech), Atlanta, GA, USA, in 1995, the M.B.A. degree in marketing from the University of Punjab (PU), Lahore, in 2000, and the D.Eng. degree in telecommunications from the School of Engineering and Technology (SET), Asian Institute of Technology (AIT), Bangkok, Thailand, in 2009. Earlier, he was with the Water and Power Development Authority (WAPDA), Pakistan, from April 1991 to August 2000, in the capacity of Planning and System Study Engineer. Afterward, he joined academia and was a Lecturer with the School of Arts and Sciences (SoAS), Lahore University of Management and Sciences (LUMS), from September 2000 to December 2002. Later on, he joined the Department of Computer Engineering afterward changed to the Department of Electrical Engineering (DEE), COMSATS Institute of Information Technology (CIIT), Lahore, as an Incharge and Assistant Professor, in January 2003, and was a Pioneer and an important Faculty Member with DEE, till June 2014 in the same designation. In July 2014, he was promoted to the rank of Associate Professor with DEE, CIIT. He was an Associate Professor and the Associate Head of DEE, COMSATS University Islamabad (CUI), Lahore Campus, till May 2020. Since June 2020, he has been the Head of the Department of Electrical and Computer Engineering (ECE) and looking after its affairs (both academic and administrative) with CUI, Lahore Campus. He has been the Head of the Department of ECE, CUI, Lahore Campus, since June 2020. His research interests include multi-rate signal and image processing and their modeling, performance analysis of wireless networks, and communication systems theory. He is a Lifetime Member of the Pakistan Engineering Council (PEC), Islamabad, Pakistan, and an IEEE Reviewer of *Wireless Sensor Networks*.

SALEEM AKHTAR received the B.Sc. degree in electrical engineering from the University of Engineering and Technology, Lahore, Pakistan, in 1991, and the D.E.A. degree in digital telecommunication systems and the Ph.D. degree in mobile CN from École Nationale Supérieure des Télécommunications, Paris, France, in 1997 and 2001, respectively. From December 1997 to July 2001, he was a Research Associate with the Department of Network and Services, Institut National des Télécommunications, France, where he was a Research Fellow with the Department of Network and Services, from October 2001 to September 2002. He is currently a Principal Engineer with the Department of Electrical and Computer Engineering, COMSATS University Islamabad (CUI), Lahore. His primary research interests include quality-of-service (QoS) provisioning and radio resource management in heterogeneous wireless networks.

MUHAMMAD FAROOQ-I-AZAM received the B.Sc. degree (Hons.) in electrical engineering from the University of Engineering and Technology, Taxila Campus, Lahore, the M.Sc. degree in electrical engineering from the University of Engineering and Technology, Lahore Campus, the M.Sc. degree in computer science from the University of Punjab, and the Ph.D. degree from Lancaster University, U.K. He is currently a part of the faculty with the Department of Electrical and Computer Engineering, COMSATS University Islamabad (CUI), Lahore Campus. He has experience working in both industry and academia. He has also been interested in computers, programming, and operating systems throughout his professional career. He has experience to manage and administer enterprise computer networks running diverse operating systems. He has also managed and completed a number of industrial projects including projects in the power sector. His research interests include communication systems and networks, cognitive radio, wireless sensor networks, localization and navigation, information security, electric vehicles, smart grids, electric circuits, probability modeling, artificial intelligence, and machine learning.

SYED RAHEEL HASSAN received the Ph.D. degree from Universite de Franche-Comte, France. Before his Ph.D. study, he worked for four years in the industry as a Network Administrator. He has been in Academia for the last ten years and has published several articles. In the past, he has completed a Research Fellowship from Emory University, USA, and taught a few courses with Universite de Franche-Comte, from 2011 to 2012. During the last four years, he was affiliated with the Department of Computer Science, King Abdulaziz University, which is one of the largest universities in Saudi Arabia. He has recently joined the School of Computing Science, University of East Anglia. He has been involved in multiple projects related to network security, such as intrusion detection systems in distributed networks, security information and management systems, and smart authentication for future networks. Recently, he has started working on the management of security for IoT networks using blockchain.

RAMEEZ ASIF received the Ph.D. degree from Friedrich-Alexander-Universität Erlangen–Nürnberg, Germany, in collaboration with the Max Planck Institute for the Science of Light, in 2012. Later, he joined leading research institutes, including Denmark Technical University (DTU) and the University of Cambridge, U.K. He was an Assistant Professor with the Department of Electronics and Electrical Engineering, University of Strathclyde, Glasgow, U.K., where he extensively researched energy systems, electrical vehicles, smart grids, and teleprotection. He has been an Associate Professor with the School of Computing Sciences, University of East Anglia (UEA), Norwich, U.K., since 2021. He has more than 120 publications in different international journals and conferences. His research interests include wireless communications, embedded systems, cloud computing, and data processing.