A Cooperative Online Learning Scheme for Resource Allocation in 5G Systems

Ismail AlQerm and Basem Shihada
CEMSE Division, King Abdullah University of Science and Technology, Saudi Arabia,
{ismail.qerm, basem.shihada}@kaust.edu.sa

***Abstract*—The demand on mobile Internet related services has increased the need for higher bandwidth in cellular networks. The 5G technology is envisioned as a solution to satisfy this demand as it provides high data rates and scalable bandwidth. The multi-tier heterogeneous structure of 5G with dense base station deployment, relays, and device-to-device (D2D) communications intends to serve users with different QoS requirements. However, the multi-tier structure causes severe interference among the multi-tier users which further complicates the resource allocation problem. In this paper, we propose a cooperative scheme to tackle the interference problem, including both cross-tier interference that affects macro users from other tiers and co-tier interference, which is among users belong to the same tier. The scheme employs an online learning algorithm for efficient spectrum allocation with power and modulation adaptation capability. Our evaluation results show that our online scheme outperforms others and achieves significant improvements in throughput, spectral efficiency, fairness, and outage ratio.**

***Index Terms*—5G networks, online learning, interference mitigation, resource allocation**

I. Introduction

Digital mobile communications have penetrated mass markets with data services, increasing the number of mobile users, and service providers. The existing wireless systems are not capable to handle the increase in mobile data consumed by new applications and services such as pervasive 3D multimedia, HDTV, VoIP, and games. 5G is a promising technology to satisfy the future demand for data services as they are expected to provide high data rates up to 300 Mbps for the downlink and 60 Mbps for the uplink with end to end latency of 2 to 5 milliseconds [1]. The vision of 5G networks is to have a global unified platform that provides seamless connectivity among existing standards (e.g., HSPA, LTE-A, and WiFi). 5G networks are composed of multiple network tiers of different sizes, transmission powers, and an unprecedented numbers of smart and heterogeneous wireless devices [2]. These tiers are expected to utilize advanced physical communications technology such as high-order spatial multiplexing multiple-input multiple-output (MIMO) communications which can provide higher aggregate capacity for more simultaneous users. However, all these technologies rely on spectrum sharing between multiple tiers (i.e., primary and secondary tier). Cross-tier and co-tier interference obstruct the path for seamless communication among users belong to these tiers. Therefore, interference mitigation and power control becomes a critical challenge in resource allocation in the 5G context.

Different approaches have been proposed to reduce the interference in cellular networks, including frequency reuse schemes such as fractional reuse and soft reuse schemes. Authors in [3] proposed a context-aware resource allocation scheme for cellular networks where the base station's scheduler observes context information from the user's environment and utilizes this knowledge for an efficient throughput-delay trade-off. Numerous power control schemes have been proposed in the literature for single-tier cellular networks such as Target-SIR-tracking power control (TPC) [4], TPC with gradual removal (TPC-GR) [5], opportunistic power control (OPC) [6], and dynamic-SIR tracking power control (DTPC) [7]. The aforementioned distributed power control schemes for satisfying various objectives in single-tier wireless cellular networks are unable to address the interference management problem in 5G multi-tier networks. This is due to the fact that they can not guarantee that the total interference caused by the secondary tier users to the primary tier users remains within tolerable limits. In additions, the schemes proposed for interference control exclude the QoS of the other tiers (e.g., fairness and data rate).

In this paper, we propose a cooperative online learning scheme which aims at solving the resource allocation and interference problems in the 5G systems. The cooperation considered involves information exchange between secondary tier users of the same type which is either femtocell or D2D connection. Online learning exploits environment awareness to allocate frequency resource blocks (RBs) and control interference. It creates optimal allocation policies without any prior model of the environment (in our settings, a prior model can not be achieved due to the unplanned placement of the secondary tier users and the dynamics of the wireless environment). Moreover, online learning allows the secondary tier to take actions while they are learning (i.e., without a centralized controller) which reduces the complexity of the system implementation. These features make online learning suitable to be applied at the distributed secondary tier network setting in the form of the so called multi-agent online learning. Our proposed scheme has the following contributions:

- Efficient RBs allocation using cooperative online learning in a priority based distributed fashion, where user location determines its assigned resources. This allocation proceeds in conjunction with system parameters adaptation including transmission power and modulation.

- Cross-tier interference mitigation between primary tier (macro) and secondary tier (femto and D2D links) and co-tier interference between secondary tier devices in the downlink.
- Maintain quality of service (QoS) of both macro users and secondary tier users. This includes zero outage and SINR above threshold for macro users, maximum data rate and minimum outage for the secondary tier users, and high level of fairness.

The rest of the paper is organized as follows. Section II briefly highlights the main characteristics of the 5G systems and describes the basics of online learning. Section III presents the system model, the proposed learning system parameters and the scheme mechanism for RBs allocation, interference mitigation, and system adaptation. Section IV presents the evaluation to demonstrate our scheme capability. Finally, Section V concludes the paper and discusses future directions.

II. Background

5G belongs to the set of wireless cellular networks that are based on Orthogonal Division Multiple Access (OFDMA). Some of the emerging trends in 5G include: multi-tier heterogeneous networks, D2D, densification of the heterogeneous base stations (e.g., extensive use of relays and small cells), massive and 3D MIMO, millimeter wave, and full duplex communications [8]. The multi-tier network architecture consists of macrocells, femtocells, and D2D networks to serve users with different QoS requirements. Macrocells are the cells that provide radio coverage served by a high power cellular base station while femtocells are small, low-power cellular base stations, typically designed for use in a home or small business. D2D communication comprises any human-centric wireless devices with Internet connectivity such as smart phones, superphones, and tablets. The deployment of heterogeneous multi-tier devices in 5G systems will significantly have much higher density than todays conventional single-tier (e.g., macrocell) networks [2]. There is a common set of radio resources available to the network tiers (e.g., macro, femto, and D2D). The femto users and D2D users utilize the available resources (e.g., bandwidth and power level) in an underlay mode as long as the interference caused to the macro tier remains below a given threshold. Frequency scheduled in the unit of RBs [9] where bandwidth is divided into certain number of RBs that network link uses. Each RB consists of successive subcarriers over a certain number of OFDMA symbols. Power and modulation are grouped in sets within certain ranges.

Online learning is a learning algorithm that uses reinforcement Q-learning [10] principles to determine an optimal policy π_s^* for decision-making without detailed modeling of the radio environment. This implies that Q-learning represents the performance metrics of interest and improves it as a whole. For instance, instead of tackling factors that affect network performance such as the wireless channel condition and mobility, Q-learning monitors the feedback of its actions such as throughput monitoring. Online learning includes four parameters which are state s, action a, probabilistic transition function from one state to the other $P_{s,s'}$ and reward function $r_{s,a}$. The state may describe internal phenomena, which are within the agent, such as instantaneous queue size, or external to the agent, such as the usage of the wireless medium. The reward function reflects the feedback for the quality of the action taken and consequently the system gains the experience. The interaction between the agent and the environment at time t occurs as follows, the agent observes the environment state s_t. The action a_t is selected based on the state s_t. According to a_t and $P_{s,s'}$, the environment makes transition to the new state and the reward function $r_t = R(s_t, a_t)$ achieved as a result of this transition is recorded and fed back to the agent. The optimal Q-value is the metric defined for each state-action pair in the process to find the optimal policy π_s^*, and it is evaluated as follows,

$$Q^*(s,a) = E\{r(s,a)\} + \gamma \sum_{s' \in S} P_{s,s'}(a) \max_{b \in A} Q^*(s',b) \quad (1)$$

where S, A are the sets of the available states and actions respectively and γ is the discount factor. The optimal policy can be determined by $\pi_s^* = \arg\max_{a \in A} Q^*(s,a)$. The online learning algorithm finds $Q^*(s,a)$ in a recursive manner using the following rule:

$$Q(s,a) = (1-\alpha)Q(s,a) + \alpha(r(s,a) + \gamma \max_{b \in A} Q^*(s',b)) \quad (2)$$

Where α is the learning rate. An appropriate action is rewarded and its Q-value is increased. In contrast, an inappropriate action is punished and the Q-value is decreased. The Q-value is maintained in a two-dimensional lookup Q-table with size (state x action). It was proved in [11] that this update rule converges to the optimal Q-value when each state-action pair is visited infinitely often. Therefore, the learning comprises two processes that control its actions selections: exploitation and exploration. Exploitation is the process of selecting the best action based on the found optimal policy π_s^* while exploration is the process of selecting non-optimal actions randomly and discovering new ones. The exploration rate ϵ is the parameters that control the level of exploration against exploitation.

III. Online Learning Scheme for Resource Allocation and Interference Mitigation

This section describes our scheme for resource allocation and interference mitigation which includes the system model and the online learning based algorithm proposed for RBs allocation and interference mitigation.

A. System Model

Our system model consists of a heterogeneous multi-tier network with secondary tier underlaid within the primary (macro) tier coverage with constraint that the interference caused to the macro users (MUE) remains below certain threshold as in Fig. 1. All the underlay network devices including femtocells base stations (FBS), and D2D transmitters (DUE) share the same radio resources with the macrocell. The network in Fig. 1 is a multi-tier heterogeneous since each of the network tiers (i.e., macro tier and underlay tier

which comprises femtocells and D2D UEs) has different transmission power range, coverage region and specific set of users with different application requirements. It is assumed that the user association with the base stations (either MBS or FBS) is completed prior to resource allocation. In addition, the potential DUEs are discovered during the D2D session setup by transmitting known synchronization or reference signal (i.e., beacons) [12]. Only one MUE is assumed to be served on each RB to avoid co-tier interference within the macro tier. However, multiple underlay devices compete to reuse the same RB to improve the spectrum utilization. This reuse causes severe cross-tier interference to the MUEs, and also co-tier interference within the underlay tier. The main objective is to allocate resources to the underlay transmitters (FBS or DUE sender) with the goal of maximizing their data rate and spectral efficiency while maintaining the SINR of the affected MUE.

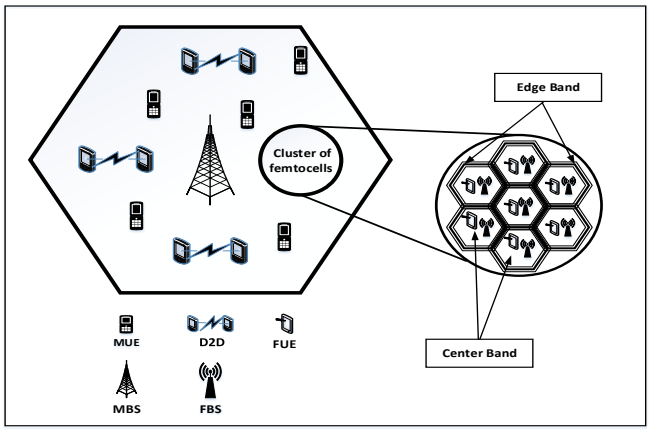


Fig. 1. System Model

Each transmitter in the underlay tier selects one RB from a set of N RBs for transmission with certain transmission power and modulation order. The underlay transmitters are capable of selecting the transmission power and modulation level from a finite set of power levels, $p = \{1, 2, 3....P\}$ and finite set of modulation indexes $m = \{1, 2, 3....M\}$, respectively. As we have two types of underlay transmissions (i.e., femtocell and D2D), each type select transmission power and modulation from its corresponding finite set. Each transmitter selects a suitable $\{n, p, m\}$ level combination. We call this combination transmission package (Tp). The transmitting FBS or D2D transmitter is denoted by y. Its transmission power over the RB is determined by the vector $p_y = \{p_y^1, p_y^2, p_y^3....p_y^N\}$ where $p_y^n > 0$ denotes the transmission power level of the transmitter y over RB n. The same thing is applicable to the modulation levels. Note that if the RB n is not allocated to the transmitter y, the corresponding power variable will be $p_y^{(n)} = 0$. Since we assume that each underlay transmitter selects only one RB, only one element in the power vector p_y is non-zero. The frequency band b is assumed to be divided into two disjoint sub-bands: edge sub-band e, where users experienced the weakest signal and the center band c with stronger signal as highlighted in Fig. 1. These two bands differ in the assigned transmission power for the underlay transmissions located within their coverage: users located in c communicate with less power than the ones located in e where transmissions occur at the higher power to compensate for the poor signal experienced. Therefore, we adopt a priority based resource allocation in which UEs with the weakest signal quality located at e benefit from the maximum transmission power of their transmitters. The signal quality is quantified based on the measured SINR and thus, all the underlay users are sorted according to their received SINR.

B. Problem Formulation

The resource allocation problem in the 5G context with multiple tiers has the objective of maximizing the data rate of the underlay users with minimal interference to the macro tier. This objective relies on certain factors including SINR of underlay UEs and SINR of MUE. In order to calculate the SINR at an underlay receiver u whether it is an (femto UE) or D2D receiver, we need to define the sources of interference that impact its signal assuming that the interfering underlay transmitters y' share the same RB used by y. The interference perceived by u is found as follows,

$$I_u = p_k G(k, u) + \sum_{y' \in Y, y' \neq y} \lambda_{y',b} p_{y',b} G(y', u) \tag{3}$$

where p_k is the transmission power of the MBS k, $G(k, u)$ is the interference gain of the MBS in the direction of the receiver u, $p_{y',b}$ is the transmission power of the interfering underlay transmitters , $\lambda_{y',b}$ represents the load of sub-band b and is defined as the ratio of the number of RBs allocated in sub-band b, and the total number of RBs available in that sub-band, and $G(y', u)$ is the gain of underlay transmitter y in the direction of u. The gain G of certain transmitter k or y' in the direction of u over RB n is defined as follows,

$$G(y', u) = \beta_{y'u} d_{y'u}^{-\alpha^*} \tag{4}$$

where $\beta_{y'u}$ denotes the channel fading component between y' and u over RB n, $d_{y'u}^{-\alpha^*}$ is the distance between y' and u, and α^* is the path-loss exponent. The SINR for u over RB n in sub-band b is given by

$$SINR_u = \frac{p_{y,b} G(y, u)}{I_u + \sigma^2} \tag{5}$$

where $p_{y,b}$ is the transmission power of the transmitter y and the $G(y, u)$ is the gain between the underlay transmitter y and the receiver u. The variable $\sigma^2 = N_0 B_{RB}$ where B_{RB} is the bandwidth corresponding to an RB and N_0 is the thermal noise. Similarly, the SINR for the MUE z over RB n can be written as follows:

$$SINR_z = \frac{p_k G(k, z)}{\sum\limits_{y \in Y} p_{y,b} G(y, z) + \sigma^2} \tag{6}$$

Given the SINR, the data rate of the UE u over RB n in sub-band b can be calculated according to the Shannon's formula, i.e., $R_u = B_{RB}(1 + SINR_u)$. The assignment of transmission package i.e. $Tp = \{n, p, m\}$ for for any underlay transmitter

y to maximize the data rate is denoted by a binary decision variable $x_y^{(n,p,m)}$ where

$$x_y^{(n,p,m)} = \begin{cases} 1, \text{ if s is transmitting over n with p and m} \\ 0, \text{otherwise} \end{cases} \tag{7}$$

Thus, the achievable data rate of an underlay UE u with the corresponding transmitter y is written as,

$$R_{uy} = \sum_{n=1}^{N} \sum_{p=1}^{P} \sum_{m=1}^{M} x_y^{(n,p,m)} B_{RB}(1 + SINR_u) \tag{8}$$

In order to maintain the SINR of the MUEs, the interference caused by the underlay transmissions I_z over RB n must be maintained below the threshold of the maximum tolerable interference by the MUE I_{TH} as follows,

$$I_z = \sum_{y \in Y} \sum_{p=1}^{P} \sum_{m=1}^{M} x_y^{(n,p,m)} G(y,z) p_{y,b} \leq I_{TH} \tag{9}$$

The resource allocation problem can be expressed by using the following optimization formulation,

$$\max_{x_y^{(n,p,m)}, p_{y,b}, m_{y,b}} \sum_{y \in Y} \sum_{n=1}^{N} \sum_{p=1}^{P} \sum_{m=1}^{M} x_y^{(n,p,m)} B_{RB}(1 + SINR_{uy}) \tag{10}$$

subjected to,

$$C.1 \quad I_z \leq I_{TH} \; \forall n \in N$$

$$C.2 \; \sum_{n=1}^{N} \sum_{p=1}^{P} \sum_{m=1}^{M} x_y^{(n,p,m)} \leq 1 \; \forall y \in FBS \sqcup DUE$$

$$C.3 \; x_y^{(n,p,m)} \in \{0,1\} \; \forall n \in N, \forall y, \forall p \in P, \forall m \in M$$

where,

$$SINR_u = \frac{p_{y,b} G(y,u)}{p_k G(k,u) + \sum\limits_{y' \in Y_{y' \neq y}} \sum\limits_{p'=1}^{P} \sum\limits_{m'=1}^{M} x_y^{(n,p,m)} \lambda_{y',b} p_{y',b} G(y',u) + \sigma^2} \tag{11}$$

The allocation problem presented in (10) aims at maximizing the data rate of the femto UE or DUE while fulfilling the constrains C.1, C.2, and C.3. In the constraint C.1, the interference caused to the MUEs by the underlay transmitters on certain RB is limited by a predefined threshold. The constraint in C.2 indicates that the number of RBs selected by each underlay transmitter should be at most one and each transmitter can only select one power level and one modulation level at each RB. The binary assignment variable for Tp selection is represented by the constraint in C.3. The resource allocation problem is a combination non-convex non-linear optimization problems. Considering the computational overhead, it not feasible to solve the resource allocation problem by a centralized approach.

C. Online Learning System Parameters

We consider a cooperative multi-agent online learning system where each agent is either an FBS or DUE transmitter. Only agents of the same type (i.e., FBS or D2D transmitter) coordinate with each other. The agent collects the environment information or performance indicators from its own and its neighboring agents to define the system state s_t at time t, and perform a local action a_t. Agents enforce the cooperative learning using a global reward, which comprises the sum of rewards achieved by all the neighboring agents. The state action table (Q-table) is common and shared by the underlay transmitters. Thus, the agents learn together a common strategy by feeding a single Q-table. In addition to fast convergence time and fairness, the system benefits from a diversified experience learned by the cooperating agents. The fundamental parameters of the online learning system are defined as follows:

- **State:** at time instant t, the environment state is defined as $s_t = \{y, n, p, SINR_c u, SINR_e u, SINR_z\}$ where y is the underlay tier transmitter, n is the available RB, p is the transmission power of the underlay tier transmitter, $SINR_c u$ is the SINR measured at the underlay receiver u in the center band, $SINR_e u$ is the SINR measured at the underlay receiver in the edge band, and $SINR_z$ is the SINR measured at the macro receiver. The aggregated state information of the neighboring agents ($s(y'_{all})$), is defined as a weighted sum over the performance indicators (of the same type) ($s(y')$) and it is given by:

$$s(y'_{all}) = \sum_{y' \in Y} w_{y'} s(y') \tag{12}$$

  where $w_{y'}$ is the weighing coefficient that reflects the degree of neighborhood of agent y to y. The weight represents the normalized traffic flux between agents y and y with respect to the total traffic flux between y and all its neighbors.
- **Action:** the action at time instant t is defined as $a_t = \{\text{allocation of } n, p_c, p_e, m\}$ where p_c and p_e are the selected transmission power of the underlay transmitters y located at the center band and edge band respectively. m is the transmission modulation level.
- **Reward:** the reward achieved is represented as $r_t = \{R_{uy}, R_{zk}, SE_{uy}\}$ where $R_{zk} = Blog_2(1 + SINR_z)$ is the data rate achieved by the MUE and SE_{uy} is the spectral efficiency achieved at the underlay receiver u, which is found by mapping it to $SINR_u$ using quality tables (obtained using a link-level simulator) incorporated within the network simulator. The reward function is evaluated with rationale of maximizing the underlay tier users' data rate while maintaining the macrocell users' data rate above certain threshold R'_{zk} as follows,

$$r_t^y = \begin{cases} e^{-(R_{zk} - R'_{zk})^2} - e^{-R_{uy}}, \text{ subjected to } C.1, C.2, C.3 \\ -2, \; otherwise \end{cases} \tag{13}$$

As the considered learning scheme is cooperative, the instantaneous global reward is a sum of all the individual rewards of the cooperating agents of the same type and is given by

$$r_y = \sum_{y \in Y} r_t^y \tag{14}$$

D. Online Resource Allocation

The online learning mechanism exploits the collected state information and uses it to allocate resources and control interference. As the exploited learning approach is cooperative, it necessitates sending state-action information from the corresponding agent Q-table to all the neighbors and receive their state-action information. This information supports the exploitation phase in which the action is selected according to the highest Q-value, which is recursively updated as follows,

$$Q_y(s_y, a_y) = (1-\alpha) Q_y(s_y, a_y) + \alpha (r_y(s_y, a_y) + \gamma \max_{l \in A_y} Q^(s_{y*}, l)) \tag{15}$$

where s_y is the current state of the agent y and s_{y*} is the previous state of the agent y. However, the learning mechanism relies on the performance metric (i.e., SINR, data rate) measured to take actions in the exploration phase besides their usage to evaluate the reward function. There are three main characteristics of the proposed learning mechanism: First, the environment state information are assumed to be available upon request during learning which is the task of the employed environment awareness methodology. Second, the mechanism starts with high exploration rate ϵ and this rate is decreased gradually to ensure that there are enough actions with high Q-value to exploit. Finally, the learning rate α in (15) is assumed to be dynamic and follow the Win-or-Learn-Fast principle which states that the learning agent should learn faster when it is losing and more slowly when winning [13]. The learning rates that we used are $\alpha = 0.05$ for rewarded solution and $\alpha = 0.2$ for the punished one. The learning mechanism based on the proposed model is illustrated in Fig. 2.

The process starts by collecting the network state information. Then, the state information exchange process engages where each agent y shares the row of its Q-table that corresponds to its current state and the optimal action where $Q_y(s_y, a_y) \geq Q_y(s_y, a'_y)$ for all $a'_y \in A$ with all other cooperating agents y' (i.e., underlay tier in the same range). At the same time, it receives the current state and all optimal actions $Q_{y'}(s_{y'}, :)$ from other agents y'. This helps the agent y to determine its joint action with the highest cumulative Q-value in the exploitation phase as follows,

$$a_y = argmax(\sum_{y \in Y} Q(s_y, a_y)) \tag{16}$$

The Q-value is found and updated according to the rule in (15). The global Q-value found in (16) is decomposed into a linear combination of local agent-dependent Q-values: $Q(s,a) = \sum_{y=1}^{Y} Q_y(s,a)$. Thus, if each agent y maximized its own Q-value, the global Q-value will be maximized. Based on this observation, choosing the action based on (16) will maximize the global Q-value. After the information exchange, the mechanism decides to employ exploration or exploitation process according to the value of ϵ. In the exploration process, the action to assign RB and the selection of power and modulation is taken randomly at the first trial. Then, the measurement of SINR and the location of the underlay receivers are exploited to improve resource allocation. For instance, if the selected transmission power to communicate with user at c is p_c, the allocated power to communicate with a receiver at e will be $p_e = \Gamma p_c$ where Γ is decided based on the the maximum allowed transmission power. Modulation is selected to be high for high SINR users and vise verse. In addition, the conditions C.1, C.2, and C.3 are checked for satisfaction. On the other hand, the transmission package with highest cumulative Q-value is selected according to (16) in the exploitation process. For example, if there are two agents i and j, each agent has one state s and three actions a_1, a_2 and a_3, the reward for each agent is its capacity and the Q-values for both agents are as follows: $Q_i(s, a_1) = 1$, $Q_i(s, a_2) = 2$, $Q_i(s, a_3) = 3$, $Q_j(s, a_1) = 2.5$, $Q_j(s, a_2) = 7$ and $Q_j(s, a_3) = 4.5$. Both agents will choose action a_2 (the maximum of the summation of the Q-values is 2 + 7), thus maximizing the aggregate capacity. After that, the reward function is evaluated for the selected Tp and the new state is observed. Finally, the Q-table is updated with the new Q-value as in (15) according to the state action paired selected. The overhead of information exchange in this cooperative scheme is calculated based on the size of Tp and the number of cooperative transmitters. So if the number of transmitters is Y^*, then the overhead is $Y^*(Y^* - 1)$ messages. Each message has the size of Tp.

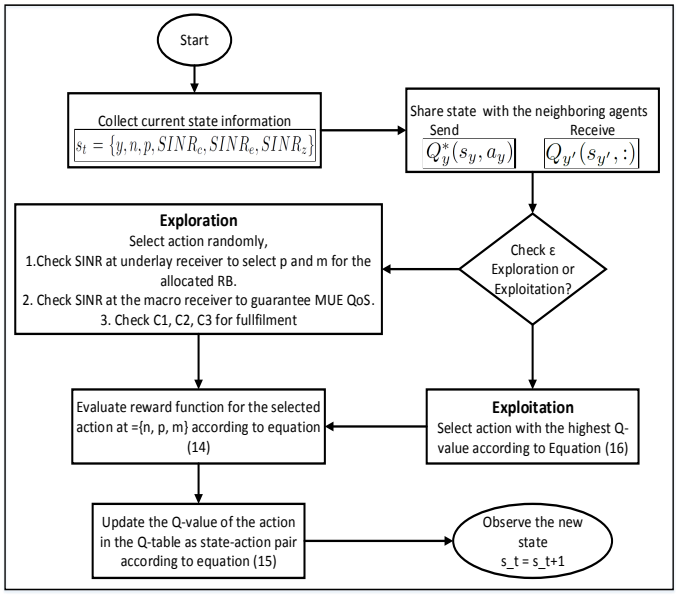
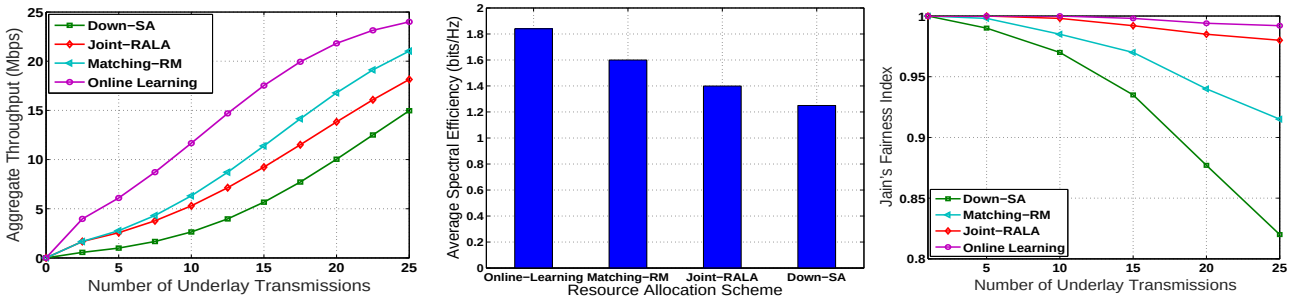


Fig. 2. Online resource allocation mechanism

IV. Performance Evaluation

We conducted extensive simulations to demonstrate the capability of using online learning scheme for 5G resource allocation. The considered simulation environment consists of a multi-tier network where multiple femtocells and D2D



(a) Aggregate system throughput achieved against number of active transmissions

(b) Spectral efficiency achieved by different resource allocation schemes

(c) Jain's fairness index with different number of active transmissions

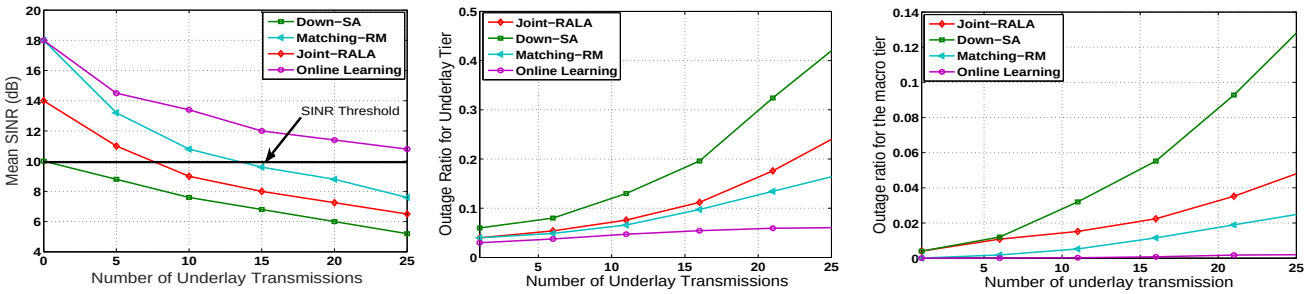
Fig. 3. Performance comparison of different resource allocation schemes in terms of throughput, spectral efficiency and fairness

communications share the spectrum resources with one macro-cell in an underlay fashion. Note that the number of D2D transmitters contribute to 20% of the number of the underlay transmitters while the rest are FBS in all the simulations. The MUEs are uniformly distributed in their respective cell coverage area. Each FBS is assumed to serve one UE which is randomly located in the coverage area. The simulation parameters and the online mechanism related parameters are listed in Table I.

Simulation Parameter	**Value**
System Bandwidth	10 MHz
Number of RBs	10
number of MUEs	10
FBS Tx power	12 to 22 dBm
MBS Tx power	48 dBm
DUE TX power	-10 to -2 dBm
Modulation QPSK	16-QAM: 1/2, 2/3, 3/4
and coding rate	64-QAM: 1/2, 2/3, 3/4, 4/5
Path loss	$PL = 127 + 37log(d)$, d = distance between underlay transmitter and UE
Shadowing	log-normal distribution (mean = $0dB$, standard deviation = $8dB$)
Thermal noise density	-174 dBm/Hz
Learning Parameters	Value
Learning Rate (α)	dynamic
Exploration Rate (ϵ)	dynamic
Discount Factor (γ)	0.9

TABLE I
5G ENVIRONMENT AND ONLINE SCHEME SIMULATION PARAMETERS

We compare the performance of our online learning scheme to validate its performance with reference schemes including: Down-SA [14], Joint-RALA [15], and Matching-RM schemes [16]. Down-SA proposes an architecture that consists of a Decision Support System (DSS) and a data collection system to dynamically manage and control the spectrum allocation process in the heterogeneous 5G networks. Joint-RALA proposes a joint algorithm for resource allocation and link adaptation to support carrier aggregation functionality for downlink in 5G LTE-A network. Matching-RM employs matching theory to find a stable match for the resource allocation problem to maximize the throughput of small cells under the cross-tier interference constraint. We evaluate the performance of our scheme compared to others in terms of the underlay average system throughput, spectral efficiency in bits/Hz and fairness of resource allocation among the underlay transmitters competes for the available RBs as in Fig. 3. In addition, the average SINR of the MUE receivers to evaluate the scheme capability to maintain QoS for the macro-tier and the outage ratio for both macro and the underlay tier are plotted in Fig. 4. Fig. 3 (a) and (b) show the aggregate system throughput with respect to the number of underlay transmitters involved and the spectral efficiency achieved by our allocation scheme in comparison to other schemes respectively. It is clear that our allocation scheme achieved the highest throughput and spectral efficiency, compared to other schemes. The reason is that our scheme formalizes the allocation problem with throughput maximization objective and it exploits online learning as a tool to solve it. Online learning is convenient to make decisions to perform allocation in such heterogeneous environment that needs plenty of interactions. The fairness performance in terms of Jains fairness index $f(x_1, x_2....) = \frac{(\sum_{i=1}^{J} x_i)^2}{J\sum_{i=1}^{J} x_i^2}$ as in [17] versus the number of underlay UEs for each scheme is plotted in Fig. 3 (c). The value of the fairness index lies in the range between [0, 1], and the value of 1 represents that all users have the same average data rate. We notice that the online allocation scheme achieved the maximum fairness as it employs cooperative learning which not only maintain fairness but also enhance the system performance. Only Joint-RALA shows comparable fairness to our scheme. The reason is that Joint-RALA adopts proportional fairness based scheduling. The average SINR achieved by the MUEs is plotted in Fig. 4 (a). This measurement aims to distinguish schemes that accounts for the macro tier QoS. We can notice that our scheme achieved the highest SINR as this is a fundamental constraint in the allocation problem. Matching-RM is the only other scheme account for this constraints. Therefore, it achieved a comparable SINR. Fig. 4 (b) and (c) present the outage ratio versus the number of underlay transmitters for the underlay users and the MUEs respectively. The outage ratio of a particular tier can be expressed as the ratio of the number of UEs supported by a tier with their minimum target SINRs and the total number of UEs in that tier. Our online scheme recorded almost zero outage ratio for the underlay tier and minimum outage for the macro tier in comparison with other schemes. Nevertheless, that Down-SA, Joint-RALA, and Matching-RM are proposed for resource allocation in the downlink of 5G, their performance is limited by several drawbacks. For instance, Down-RA does



(a) Mean SINR at MUE receiver (b) Average Outage ratio at the underlay tier (c) Average outage ratio at the macro tier

Fig. 4. Performance comparison of different resource allocation schemes in terms of SINR and outage ratio

not consider QoS in RB allocation and interference problem is out of its scope although it is the main problem in resource allocation. Joint-RALA proposed an algorithm to allocate RBs while maximizing the achieved data rate. However, power control and macro tier QoS maintenance issue are not taken in consideration. Matching-RM came with the goal to handle both allocation with QoS and interference problems but it follows centralized approach for resource management which is not feasible with all the computational overhead in such network. In addition, non of the proposed schemes consider the impact of the D2D communications in 5G.

V. Conclusion

We proposed a resource allocation scheme with embedded online learning algorithm to allocate RB and control interference in 5G networks. It is designed with high abstraction control functions for resource allocation. Our overall system goal was to maximize data rate and spectral efficiency with QoS guarantees for the macro tier. Radio environment awareness and learning are used to improve network efficiency and respond to changes in network conditions in fast and convenient manner. The performance of our online resource allocation mechanism was compared with other schemes designed for resource allocation in 5G networks. Our scheme shows a better achievement in terms of throughput, spectral efficiency, fairness, and outage ratio for different tier transmissions.

References

[1] Ekram Hossain and Monowar Hasan, "5g cellular: Key enabling technologies and research challenges", *CoRR*, vol. abs/1503.00674, 2015.

[2] Woon Hau Chin, Zhong Fan, and R. Haines, "Emerging technologies and research challenges for 5g wireless networks", *IEEE Wireless Communications,*, vol. 21, no. 2, pp. 106–112, April 2014.

[3] Magnus Proebster, Matthias Kaschub, Thomas Werthmann, and Stefan Valentin, "Context-aware resource allocation for cellular wireless networks", *EURASIP Journal on Wireless Communications and Networking*, vol. 2012, no. 1, 2012.

[4] G.J. Foschini and Z. Miljanic, "A simple distributed autonomous power control algorithm and its convergence", *IEEE Transactions on Vehicular Technology,*, vol. 42, no. 4, pp. 641–646, Nov 1993.

[5] M. Rasti, A.R. Sharafat, and J. Zander, "Pareto and energy-efficient distributed power control with feasibility check in wireless networks", *IEEE Transactions on Information Theory*, vol. 57, no. 1, pp. 245–255, Jan 2011.

[6] Kin-Kwong Leung and Chi Wan Sung, "An opportunistic power control algorithm for cellular network", *IEEE/ACM Transactions on Networking,*, vol. 14, no. 3, pp. 470–478, June 2006.

[7] M. Rasti, A.R. Sharafat, and J. Zander, "A distributed dynamic target-sir-tracking power control algorithm for wireless cellular networks", *IEEE Transactions on Vehicular Technology,*, vol. 59, no. 2, pp. 906–916, Feb 2010.

[8] Cheng-Xiang Wang, F. Haider, Xiqi Gao, Xiao-Hu You, Yang Yang, Dongfeng Yuan, H. Aggoune, H. Haas, S. Fletcher, and E. Hepsaydir, "Cellular architecture and key technologies for 5g wireless communication networks", *IEEE Communications Magazine,*, vol. 52, no. 2, pp. 122–130, February 2014.

[9] S-E Elayoubi and B. Fourestie, "On frequency allocation in 3g lte systems", in *IEEE 17th International Symposium on Personal, Indoor and Mobile Radio Communications*, 2006, pp. 1–5.

[10] C. J. Watkins and P. Dayan, "Technical note: Q-learning", in *Mach. Learn*, 1992, vol. 8, pp. 279–292.

[11] ChristopherJ.C.H. Watkins and Peter Dayan, "Technical note: Q-learning", *Machine Learning*, vol. 8, no. 3-4, pp. 279–292, 1992.

[12] G. Fodor, E. Dahlman, G. Mildh, S. Parkvall, N. Reider, G. Miklos, and Z. Turanyi, "Design aspects of network assisted device-to-device communications", *IEEE Communications Magazine*, vol. 50, no. 3, pp. 170–177, March 2012.

[13] Michael Bowling and Manuela Veloso, "Multiagent learning using a variable learning rate", *Artificial Intelligence*, vol. 136, no. 2, pp. 215 – 250, 2002.

[14] T.R. Omar, A.E. Kamal, and J.M. Chang, "Downlink spectrum allocation in 5g hetnets", in *International Wireless Communications and Mobile Computing Conference (IWCMC)*, Aug 2014, pp. 12–17.

[15] S. Rostami, K. Arshad, and P. Rapajic, "A joint resource allocation and link adaptation algorithm with carrier aggregation for 5g lte-advanced network", in *22nd International Conference on Telecommunications (ICT)*, April 2015, pp. 102–106.

[16] S. M. Ahsan Kazmi, Nguyen H. Tran, Tai Manh Ho, Thant Zin Oo, Tuan LeAnh, Seung Il Moon, and Choong Seon Hong, "Resource management in dense heterogeneous networks", in *17th Asia-Pacific Network Operations and Management Symposium, APNOMS 2015, Busan, South Korea, August 19-21, 2015*, 2015, pp. 440–443.

[17] Dah-Ming Chiu and Raj Jain, "Analysis of the increase and decrease algorithms for congestion avoidance in computer networks", *Comput. Netw. ISDN Syst.*, vol. 17, no. 1, pp. 1–14, June 1989.