Energy Efficient Power Allocation in Multi-tier 5G Networks Using Enhanced Online Learning

Ismail AlQerm, *Student Member, IEEE*, Basem Shihada, *Senior Member, IEEE*

***Abstract*—The multi-tier heterogeneous structure of 5G with dense small cells deployment, relays, and device-to-device (D2D) communications operating in an underlay fashion is envisioned as a potential solution to satisfy the future demand for cellular services. However, efficient power allocation among dense secondary transmitters that maintains quality of service (QoS) for macro (primary) cell users and secondary cell users is a critical challenge for operating such radio. In this paper, we focus on the power allocation problem in the multi-tier 5G network structure using a non-cooperative methodology with energy efficiency consideration. Therefore, we propose a distributive intuition-based online learning scheme for power allocation in the downlink of the 5G systems, where each transmitter surmises other transmitters power allocation strategies without information exchange. The proposed learning model exploits a brief state representation to account for the problem of dimensionality in online learning and expedite the convergence. The convergence of the proposed scheme is proved and numerical results demonstrate its capability to achieve fast convergence with QoS guarantee and significant improvement in system energy efficiency.**

***Index Terms*—5G networks, online learning, energy efficiency, power allocation**

I. Introduction

Digital mobile communications have penetrated mass markets with data services, high number of mobile users, and service providers. The existing wireless systems will not be able to cope with the upcoming increase in mobile data usage by new applications. 5G is a promising technology to satisfy the future demand for data services as it is expected to provide high data rates up to 10 Gbps 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) [2]. One of the envisioned 5G structures is the multi-tier network with different sizes, transmission powers, and unprecedented numbers of smart and heterogeneous wireless devices [3]. The multi-tier structure, which we consider in this paper, consists of two tiers: primary tier and secondary tier. The primary tier includes high power macrocells that serve macro users (MUEs), while the secondary tier comprises picocells, femtocells and Device to Device (D2D) communications. The secondary base stations (BSs) including pico and femto BSs, and D2D transmitters are denoted as secondary transmitters (STs). They utilize the available resources (e.g., bandwidth and power) in an underlay mode as long as the interference caused to the macro tier remains below certain threshold. However, they tend to increase their transmission power to maximize their performance, which causes severe interference to the primary tier and increases the power consumption [4] [5]. Due to the conflict of interests among the STs, it is more suitable to address the power allocation problem in a non-cooperative fashion. This also reduces the overhead of either appointing central entity for information broadcasting and information exchange among the STs. However, the non-cooperative approach may cause severe interference, increase in the power consumption, and degradation in QoS of the macro users and the secondary users due to the lack of the environment awareness [6].

In this paper, we consider the power allocation problem from the energy efficiency perspective in the multi-tier 5G heterogeneous network. Therefore, we proposed a distributive intuitive online learning power allocation scheme for the STs, which allows each ST to conjecture other STs power allocation strategies with only local information from direct interactions with the environment and making use of the past experience. This accounts for the lack of information exchange among STs, which is essential to reach optimal power allocation. In addition, the proposed scheme maintains QoS represented by signal to interference plus noise ratio (SINR) for all users in the network. The proposed scheme provides the following key highlights: First, the intuition feature allows each ST to update its learning information using its private past experience and this eliminates the overhead of cooperation. Second, the paper contributes to the general literature of online learning [7] as the traditional online learning relies on full information from all agents in the environment, which is difficult to achieve in such dynamic environment. Third, the proposed online scheme exploits a brief representation feature that significantly reduces the scheme computation. The brief feature approximates the Q-value of the online learning as a function of much smaller set of variables, which reduces the state space and expedites the algorithm convergence.

The rest of the paper is organized as follows. A brief motivation to emphasize the energy efficiency problem in 5G networks and the related work are presented in section II. Section III describes the system model and the problem formulation. The power allocation learning model along with the challenging issues is described in section IV. Section V describes the approximated intuitive scheme for power allocation. The proposed scheme numerical results are presented in section VI and the paper concludes in section VII.

II. Related Work

Although the multi-tier approach in 5G is promising to improve the performance of the network in order of magnitudes compared to the legacy cellular networks, it may lead to a significant interference between the primary tier and the secondary tier and also between the secondary tier

devices [8]. This interference impacts energy efficiency and degrades the QoS experienced by all users. STs tend to increase their transmission power unnecessarily to overcome interference and this leads to significant sacrifice in energy efficiency. The interference is a crucial problem in 5G due to the following reasons: heterogeneity and dense deployment of wireless devices, various transmission powers of different transmitters, which may cause imbalance in the traffic load and coverage, public or private access restrictions in different tiers that lead to diverse interference levels and priorities in accessing different portions of spectrum plus the impact of carrier aggregation and D2D communications.

Therefore, sophisticated power allocation mechanisms are necessary to account for the interference problem and enhance the system performance, which consequently reduces the power consumption and maintains QoS for different tier users [9] [10] [11] [12] [13].

Power allocation problem in the multi-tier heterogeneous environment has become an interesting topic in the current research of wireless communication. Authors in [14] proposed a utility based power adaptation algorithm to mitigate the cross tier interference at the macrocell from the femtocells. In [15], authors proposed game-theoretic framework in heterogeneous network, which enables both the small cells and the macrocells to strategically decide on their downlink power control policies. They formulate the power allocation problem as a stackelberg game to maximize the data rate of each cell. The work in [16] proposed a hierarchical game theoretical framework for optimal resource allocation on the uplink of a heterogeneous network with femtocells overlaid on the edge of a macrocell. Authors in [17] summarized the challenges and opportunities to improve energy efficiency while increasing the network capacity in both multi-radio access technology (RAT) and single-RAT heterogeneous networks. The downlink resource allocation problem in Orthogonal Frequency Division Multiple Access (OFDMA) heterogeneous networks consisting of macrocells and small cells sharing the same frequency band was investigated in [18]. The authors aim to devise an energy efficient scheme that allows shared spectrum access to small cells, while ensuring a minimum level of QoS for the macro users. Heterogeneous networks based on large-scale user behavior was proposed in [19], where the heterogeneity of large-scale user behavior is quantitatively characterized and exploited to study the energy efficiency performance. The authors in [20] proposed a new resource allocation scheme presenting a low computational overhead and a low sub-band handoff rate in a dynamic ultra-dense heterogeneous network. The work in [21] addressed the energy efficiency optimization problem for downlink two-tier heterogeneous networks, where the power allocation problem is decomposed into multiple optimization problems with single inequality constraint. Those optimization problems are solved using a suboptimal solution based on the zero forcing precoding approach. An energy-efficient radio resource allocation algorithm (EERRAA) was proposed in [22] for interference management, maximization of throughput and energy efficiency to enhance the performance of a heterogeneous deployment of ultra-dense femtocells overlaying the macrocells. The proposed scheme exploits cognitive radio technology and stochastic process to perform resource allocation. The authors in [23] proposed a multiuser MIMO precoding scheme that is capable to reduce the negative impact of interference in heterogeneous networks. The formulated optimization problem is solved using heuristics based techniques. Both works in [21] and [23] are cooperative and rely on information exchange among STs. The work in [24] aimed at network utility maximization via jointly optimizing user association, resource allocation and power control in a load-coupled heterogeneous network. The authors in [25] addressed a non-convex energy efficient optimization problem with resource assignment and power allocation for the OFDM heterogeneous cloud radio access network. They found closed-form expressions for the energy-efficient resource allocation solution to jointly allocate the resource block and transmit power. Reinforcement Learning (RL) [26] was considered as suitable solution for interference management and power allocation in such autonomous environment. An RL based algorithm was proposed to optimize the network performance by managing power allocation in femtocells in [27]. However, non of the above work consider 5G structure where D2D is involved and the network is dense with various types of small cells deployed i.e.(all of the previous work considers only one type of ST). In addition, our work explores an improved online learning theory for a stochastic non-cooperative power allocation in heterogeneous 5G network, which achieve better results with less computations and without pre-defined network model.

III. System Description

This section describes the considered system including the system model and the power allocation problem formulation with energy efficiency consideration.

A. System Model

Our system considers the downlink transmission in a spectrum sharing heterogeneous 5G with two tiers: primary tier and secondary tier, where the primary tier consists of the macrocell with its associated users (MUEs) and the secondary tier consists of two types of cells including picocells and femtocells noted as secondary tier cells and D2D communications as in Figure 1. All the secondary tier BSs and D2D users

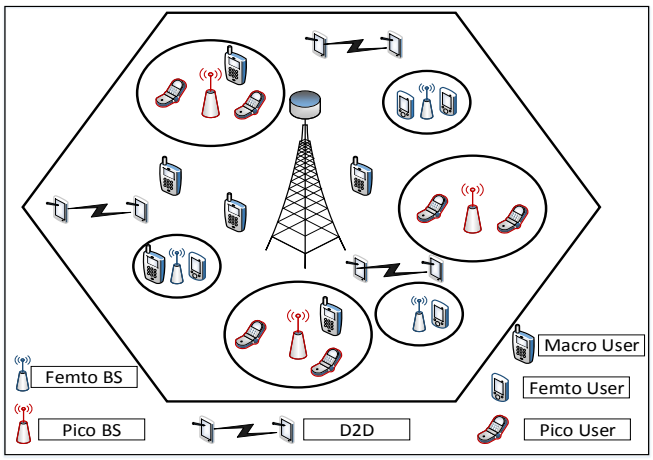

Figure 1. System Model

are assumed to be uniformly distributed under the coverage of the macrocell. The set of secondary BSs is donated by $\mathbf{N} = \{1, 2...., N\}$, the set of secondary BSs associated users (SUEs) are denoted by $\mathbf{X} = \{1, 2, ..., X\}$, the set of MUEs is represented by $\mathbf{K} = \{1, 2, ..., K\}$, and the set of the active D2D pairs as $\mathbf{D} = \{1, 2....., D\}$. The dth D2D pair ($d \in \mathbf{D}$) consists of the D2D transmitter $d_T \in \mathbf{D}_T$ and D2D receiver $d_R \in \mathbf{D}_R$, where $\mathbf{D}_T = \{1, 2....., D_T\}$ and $\mathbf{D}_R = \{1, 2....., D_R\}$. The set of SUEs associated with the nth secondary BS is referred as $\mathbf{X}_n$ with assumption that each SUE can associate with at most one secondary BS. Their resource allocation problem includes SUE association and power allocation. For SUE association, each SUE can be associated with a single BS. To specify the SUE association, we define υ_n^x as the association indicator for SUE x with BS n which is a binary variable. If $\upsilon_n^x = 1$, it indicates that xth SUE ($x \in \mathbf{X}$) is associated with nth cell and it is zero otherwise. For transmission power allocation, the nth secondary BS can select a random power level P_L from a set of discrete power levels as follows,

$$P_L = \begin{cases} \in [1, P_L^*], & \text{if the } n\text{th BS serves one SUE} \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where P_L^* is the maximum integral level of the secondary BS transmission power. To illustrate, the power allocated to nth secondary BS P_n belongs to the set $[0, \frac{1}{P_L^*}P_n^{max}, \frac{2}{P_L^*}P_n^{max}, ..., \frac{P_L}{P_L^*}P_n^{max}, ..., P_n^{max}]$, where P_n^{max} is the maximum allowed transmission power for the nth secondary BS. On the other hand, the dth D2D pair can select a power level $P_d \in \{0, 1, 2, ..., P_L^*\}$ that satisfies minimum transmission power requirements P_d^{min} and ensures that the D2D receiver is located within the proximity of the D2D transmitter $R_d \leq R_d^{max}$. The channel inversion power control is exploited to compensate the large scale fading and make the average received power at the D2D receiver larger than the minimum sensitivity η_{min} [28]. Thus, the D2D proximity is calculated as,

$$R_d = \left(\frac{P_d P_d^{max}}{P_L^* \eta_{min}}\right)^{\alpha} \tag{2}$$

As a result, the minimum power level allocated to the dth D2D pair P_d is found as

$$P_d^{min} = \frac{P_L^* \eta_{min}}{P_d^{max}} R_d^{\alpha} \tag{3}$$

where P_d^{max} is the maximum allowed transmission power for the D2D transmitter.

The SINR of the xth SUE associated with the nth secondary BS is calculated as follows,

$$\gamma_{n,x} = \frac{P_n G_{n,x}}{I_{\mathbf{D},x} + I_{q,x} + I_{\mathbf{M},x} + \sigma} \tag{4}$$

where $G_{n,x}$ is the power gain between the nth secondary BS and the xth SUE, and σ is the noise power. The aggregate interference at the xth SUE from all D2D transmitters is defined as,

$$I_{\mathbf{D},x} = \sum_{d_T \in D_T} P_d G_{d_T,x} \tag{5}$$

Moreover, the interference at the xth SUE from all other secondary BSs q is defined as,

$$I_{q,x} = \sum_{q \in \mathbf{N}/n} P_q G_{q,x} \tag{6}$$

where $G_{d_T,x}$ and $G_{q,x}$ are the power gains between the xth SUE and both D2D transmitters and the other secondary BSs q respectively. P_q is the transmission power of the secondary BS q. The interference from the macro BS to the xth SUE is defined as follows,

$$I_{\mathbf{M},x} = P_m G_{m,x} \tag{7}$$

where P_m is the transmission power of the macro BS and $G_{m,x}$ is the power gain between the macro BS and the xth SUE. Note that the macro BS interference comes from the macro BS that the SUE operate under its coverage.

The SINR of the dth D2D pair is calculated as follows,

$$\gamma_d = \frac{P_d G_{d_T,d_R}}{I_{y,d} + I_{\mathbf{N},d} + I_{\mathbf{M},d} + \sigma} \tag{8}$$

where G_{d_T,d_R} is the power gain between the dth D2D transmitter and the dth D2D receiver and σ is the noise power. The aggregate interference at the d_Rth D2D from all other D2D transmitters is defined as,

$$I_{y,d} = \sum_{y \in \mathbf{D}_T/d_T} P_y G_{y,d_R} \tag{9}$$

Moreover, the interference at the d_Rth D2D from all the secondary BSs that belongs to $\mathbf{N}$ is defined as,

$$I_{\mathbf{N},d} = \sum_{n \in \mathbf{N}} P_n G_{n,d_R} \tag{10}$$

where G_{y,d_R} and G_{n,d_R} are the power gains between the d_Rth D2D and both D2D transmitters y and other secondary BSs n respectively. P_y is the transmission power of the D2D transmitter y. The interference from the macro BS to the dth receiver is found as follows,

$$I_{\mathbf{M},d} = P_m G_{m,d_R} \tag{11}$$

where G_{m,d_R} is the power gain between the macro BS and the dth D2D receiver.

B. Problem Formulation

To realize the non-cooperative energy efficient power allocation, we define the energy efficiency in the power allocation process as the ratio of the data rate to the power consumed by both secondary BSs and D2D transmitters as follows,

$$EE_i = \frac{B \log_2(1 + \gamma)}{P_i + P_{cc}} \tag{12}$$

where B is the bandwidth and P_{cc} is the power consumed by the ST circuit. Note that the index i refers to the secondary transmitter (ST) including both secondary BSs and D2D transmitters. γ is the SINR achieved at the secondary receiver whether it is $\gamma_{n,x}$ or γ_d. Formally, The non-cooperative power allocation optimization problem in the 5G heterogeneous structure is defined as follows,

$$\max_{P_i} EE_i \tag{13}$$

subjected to

$$\text{C.1: } \gamma_{n,x} \geq \gamma^*_{n,x} \ \ \forall x \in \mathbf{X} \ \ and \ \ \gamma_d \geq \gamma^*_d \ \ \forall d_R \in \mathbf{D}_R$$

$$\text{C.2: } \gamma_k \geq \gamma^*_k \quad \forall k \in \mathbf{K}$$

$$\text{C.3: } \sum_{n \in \mathbf{N}} \upsilon^x_n = 1 \ \ \forall x \in \mathbf{X}$$

$$\text{C.4: } \sum_{x \in \mathbf{X}} \upsilon^x_n \leq \kappa \ \ \forall n \in \mathbf{N}$$

The constraint C.1 is to guarantee that the SINR of the xth SUE and dth D2D do not fall below the thresholds $\gamma^*_{n,x}$ and γ^*_d respectively. C.2 is the constraint to maintain the SINR of the MUEs above a designated threshold γ^*_k provided that the SINR of the macro tier MUE is defined as follows,

$$\gamma_k = \frac{P_m G_{m,k}}{\sum_{n \in \mathbf{N}} P_n G_{n,k} + \sum_{d_T \in \mathbf{D}_T} P_d G_{d_T,k} + \sigma} \tag{14}$$

where $G_{m,k}$, $G_{d_T,k}$ and $G_{n,k}$ are the power gains between the kth MUE and macro BS, D2D transmitters d_T and secondary BSs n respectively. Note that C.2 is supported by the assumption that macro BS can exchange its associated MUEs SINR information with both secondary BSs and D2D transmitters as they operate under its coverage. The constraint C.3 indicates that each SUE can be associated with only one secondary BS and C.4 emphasizes that each secondary BS can serve at most κ SUEs.

IV. Power Allocation Learning Model

In this section, we establish the ST power allocation model using online learning defined as $\zeta = (\mathbf{N}, \mathbf{D}_T, P_i, EE_i)$. The action space available for all STs is defined as $P = \prod_{i \in \mathbf{N} \cup \mathbf{D}_T} P_i$. We consider a slotted time structure for spectrum access for macro and secondary tier during the long time learning process. The continuous action profile $P_i = [P_i^{min}, P_i^{max}]$ is discretized to be compatible with the online learning framework according to (1) in the system model. We designate $a_i \in A_i = \{0, \frac{1}{P^*_L} P_i^{max}, \frac{2}{P^*_L} P_i^{max}, ..., \frac{P_L}{P^*_L} P_i^{max}, ..., P_i^{max}\}$ as the STs action and $A = \prod_{i \in \mathbf{N} \cup \mathbf{D}_T} A_i$ as the action space for all STs. Therefore, ζ is converted to the discrete form $\zeta' = (\mathbf{N}, \mathbf{D}_T, \{A_i\}, \{EE_i\})$. In the following subsection, we define the main components of the online learning mechanism.

A. Online Learning Structure

Each ST has the role of a learning agent, which aims to reach optimal power allocation strategy for different network state. The online learning parameters are defined as follows:

- **State** since there is no cooperation among the competing STs, they only rely on the local observation to define their environment state at certain time slot t. The state observed by ST i is defined as,

$$s^t_i = (i, P_i) \tag{15}$$

- **Action:** the action is defined as the ST transmission power $(a_i) = (P_i)$.
- **Reward:** the reward function is defined for the state/action pair as $R_i(s_i, a_i)$ and is evaluated as follows,

$$R_i(s_i, a_i) = \begin{cases} R_i(a_i) = EE_i, & \text{if C.1 to C.4 are satisfied} \\ 0, & \text{otherwise} \end{cases} \tag{16}$$

Specifically, it is a return of selecting power level action (a_i) in state s_i to guarantee the transmission QoS requirement as well as to achieve the desired energy efficiency. This indicates that the reward is achieved if the conditions in C.1 to C.4 are satisfied.

- **Transition Function:** the transition from state s^t_i to s^{t+1}_i is determined by the ST stochastic behavior. Each ST selects the strategy $\pi_i(s_i)$ independently to maximize its total expected reward. The strategy $\pi_i(s_i)$ is defined to be a probability vector $\pi_i(s_i) = [\pi_i(s_i, 0), ..., \pi_i(s_i, P_i^{max})]$ where $\pi_i(s_i, a_i)$ represents the probability at which the ST i selects action a_i at the state s_i.

For the case of having complete information about all other STs strategies $\pi_{-i} = (\pi_1, ..., \pi_{i-1}, \pi_{i+1}, ..., \pi_N)$, the total expected discounted reward of ST i over an infinite time slots is defined as follows,

$$\begin{aligned} V_i(s_i, \pi_i, \pi_{-i}) &= E\big[\sum_{t=0}^{\infty} \beta^t R_i(s^t_i, \pi_i(s^t_i), \pi_{-i}(s^t_i)), s^0_i = s_i\big] \\ &= E[R_i(s_i, \pi_i(s_i), \pi_{-i}(s_i))] + \\ &\beta \sum_{s'_i \in S_i} T_{s_i, s'_i}(\pi_i(s_i), \pi_{-i}(s_i)) V_i(s'_i, \pi_i, \pi_{-i}) \end{aligned} \tag{17}$$

where $T_{s_i, s'_i}(.)$ is the state transition probability, and

$$\begin{aligned} &E[R_i(s_i, \pi_i(s_i), \pi_{-i}(s_i))] = \\ &\sum_{(a_i, a_{-i}) \in A} [R_i(s_i, a_i, a_{-i}) \prod_{j \in N} \pi_j(s_j, a_j)] \end{aligned} \tag{18}$$

where a_{-i} represents the action selected by other STs for state s_i. In the stochastic learning, each ST has the task to learn the optimal power allocation strategy π^*_i for each environment state s_i. The following condition must be satisfied in order to reach the optimal strategy π^*_i for each ST $i \in \mathbf{N} \cup \mathbf{D}_T$,

$$V_i(s_i, \pi^*_i, \pi^*_{-i}) \geq V_i(s_i, \pi_i, \pi^*_{-i}), \forall \pi_i \in \Pi_i \tag{19}$$

The optimal strategy satisfies the Bellman's optimality equation [26], that is, for ST i

$$\begin{aligned} V_i(s_i, \pi^*_i, \pi^*_{-i}) &= \max_{a_i \in A_i} \big\{E[R_i(s_i, a_i, \pi^*_{-i}(s_i))] \\ &+ \beta \sum_{s'_i \in S_i} T_{s_i, s'_i}(a_i, \pi^*_{-i}(s_i)) V_i(s'_i, \pi^*_i, \pi^*_{-i})\big\} \end{aligned} \tag{20}$$

where

$$\begin{aligned} &E[R_i(s_i, a_i, \pi^*_{-i}(s_i))] = \\ &\sum_{a_{-i} \in A} [R_i(s_i, a_i, a_{-i}) \prod_{j \in N/\{i\}} \pi^*_j(s_j, a_j)] \end{aligned}$$

Thus, we can evaluate the optimal Q-value of ST i as the current expected reward plus a future discounted reward when all other STs follow the optimal strategy as follows,

$$Q^*_i(s_i, a_i) = E[R_i(s_i, a_i, \pi^*_{-i}(s_i))]$$

$$+\beta \sum_{s_i' \in S_i} T_{s_i,s_i'}(a_i, \pi^*_{-i}(s_i)) V_i(s_i', \pi^*_i, \pi^*_{-i}) \quad (21)$$

By combining (20) and (21), we get,

$$Q^*_i(s_i, a_i) = E[R_i(s_i, a_i, \pi^*_{-i}(s_i))]$$
$$+\beta \sum_{s_i' \in S_i} T_{s_i,s_i'}(a_i, \pi^*_{-i}(s_i)) \max_{b_i \in A_i} Q^*_i(s_i', b_i) \quad (22)$$

The employed online learning scheme aims to reach the optimal Q-value defined in (22) in a recursive way using the information $(a_i, s_i, s_i', \pi_i^t)$ with the two states $s_i = s_i^t$ and $s_i' = s_i^{t+1}$ observed at the time slot t and $t+1$ respectively, a_i and π_i^t are the ST action taken at the end of time slot t and the power allocation strategy during time slot t respectively. The update rule for the online learning employed to reach the optimal Q-value is given by,

$$Q_i^{t+1}(s_i, a_i) = (1-\alpha^t) Q_i^t(s_i, a_i) + \alpha^t \Big\{ \sum_{a_{-i} \in A_{-i}} [R_i(s_i, a_i, a_{-i}) \times \prod_{j \in N \ i} \pi_j^t(s_j, a_j)] + \beta \max_{b_i \in A_i} Q_i^t(s_i', b_i) \Big\} \quad (23)$$

where $\alpha \in [0, 1)$ is the learning rate. Once the transmission power level action (a_i) is selected and the ST i achieved the expected reward, the corresponding Q-value is updated by combining the old value and the new expected reward.

B. System Design Challenging Issues

As the goal of this paper is to create a stochastic energy efficient power allocation scheme that is non-cooperative and can guarantees QoS for different tier users, we can notice in the online learning structure presented in section IV.A that it requires information about other STs strategies and the reward of each ST is a function of joint actions of other STs. This creates a challenging problem due to the following reasons:

- Every ST may not be aware of the number of other ST existing in the system.
- Each ST can only obtain its local information such as environment state, its transmission strategy and received historical reward.
- The 5G heterogeneous system has a large space. Therefore, the curse of dimensionality increases the required computations and makes it unfeasible to use the typical online learning methodology to maintain the Q-value for each state/action pair, which slows the system convergence.

According to the update rule derived in (23), we can deduce that the stochastic power allocation problem cannot be solved directly because STs cannot observe other STs strategies in the non-cooperative power allocation fashion.

V. Power Allocation with Approximated-Intuition Based Online Learning

In this section, we account for the problem of power allocation in 5G heterogeneous network without being aware of other STs power allocation strategies. In addition, the slow convergence problem due to the large space of state/action Q-values in such environment is considered. Thus, we propose a novel approximated-intuition based online learning scheme, which allows each ST to surmise other STs power allocation strategies without explicit information exchange. In addition, it uses a brief representation for the Q-values in which they are approximated as a function of much smaller set of variables. This expedites the convergence and reduces the algorithm related computations.

A. Intuition Based Power Allocation

The intuition idea is derived from the concept that different STs follow similar power allocation strategies at the same network state. To be able to estimate other STs power allocation strategies $\pi^t_{-i}(s_i) = (\pi^t_1(s_1), ..., \pi^t_{i-1}(s_{i-1}), \pi^t_{i+1}(s_{i+1}), ..., \pi^t_{N+D}(s_{N+D}))$ using non-cooperative learning scheme, we define an intuition factor as,

$$\mu_i^t(s_i, a_{-i}) = \prod_{j \in \mathbf{N} \cup \mathbf{D}/\{i\}} \pi_j^t(s_j, a_j) \quad (24)$$

for ST i at time slot t. This function conjectures the change in the Q-value $Q_i^{t+1}(s_i, a_i)$ in the next time slot $t+1$ as a result of certain strategies employed by the other STs. $\mu_i^t(s_i, a_{-i})$ is assumed to be the only information that the learning agent knows about other STs and it is found based on local information. The probability that ST i experience environment state s^{t+1} which is the same as the probability that ST i achieves a reward $R_i(s_i, a_i, a_{-i})$ is defined as follows,

$$\Gamma_i = \pi_i^t(s_i, a_i) \mu_i^t(s_i, a_{-i}) \quad (25)$$

The probability calculated in (25) is also the probability that ST i achieves the reward function defined in (16). Let us assume that δ is the number of consecutive time slots that ST i achieved the same reward. Consequently, δ has an independent and identical distribution with probability $\Gamma_i = \frac{1}{1+\delta'}$, where δ' is the mean of δ and can be found by ST i through observing its reward history. Thus, the intuition factor can be estimated as $\mu_i^t(s_i, a_{-i}) = \frac{1}{(1+\delta')\pi_i^t(s_i, a_i)}$ as ST i is aware of its own power allocation strategy $\pi_i^t(s_i, a_i)$. As the action available to ST i is to choose the transmission power according to its strategy $\pi_i^t(s_i)$, the simplest method to express the intuition factor as a function of ST i power allocation strategy is the following expression,

$$\mu_i^t(s_i, a_{-i}) = \mu_i^{'}(s_i, a_{-i}) + w_i[\pi_i^t(s_i, a_i) - \pi_i'(s_i, a_i)] \quad (26)$$

where $\mu_i^{'}(s_i, a_{-i})$ and $\pi_i'(s_i, a_i)$ are the reference points for specific intuition and probability of certain action selection, and w is a positive scalar for linearization. The reference points are considered to be given of common knowledge. They are determined with assumption that other STs can observe ST i deviation from its reference points $\pi_i'(s_i, a_i)$ and $\mu_i^{'}(s_i, a_{-i})$ by a quantity proportional to $[\pi_i^t(s_i, a_i) - \pi_i'(s_i, a_i)]$. If the reference points are $\mu_i^{'}(s_i, a_{-i}) = \prod_{j \in \mathbf{N} \cup \mathbf{D}/\{i\}} \pi_j^*(s_j, a_j)$ and $\pi_i'(s_i, a_i) = \pi_i^*(s_i, a_i)$, then the optimal intuition factor is $\mu_i^*(s_i, a_{-i}) = \prod_{j \in \mathbf{N} \cup \mathbf{D}/\{i\}} \pi_j^*(s_j, a_j)$ and this will lead to an optimal transmission. The STs revise their reference points

based on their historical local information about transmissions that achieved maximum Q-value. We define the following rule for the intuition factor of STs to update their reference points in time slot t,

$$\mu_i^t(s_i, a_{-i}) = \mu_i^{t-1}(s_i, a_{-i}) + w_i[\pi_i^t(s_i, a_i) - \pi_i^{t-1}(s_i, a_i)] \tag{27}$$

where $\mu_i^{'}(s_i, a_{-i})$ and $\pi_i^{'}(s_i, a_i)$ are set to $\mu_i^{t-1}(s_i, a_{-i})$ and $\pi_i^{t-1}(s_i, a_i)$ respectively. This means that each ST believes that any modifications to its current strategy, will induce other STs to perform changes in the next time slot. We consider the deviation of ST i from its reference points as the model to capture the strategies variation of other STs as follows,

$$\mu_i^t(s_i, a_{-i}) - \mu_i^{t-1}(s_i, a_{-i}) = w_i[\pi_i^t(s_i, a_i) - \pi_i^{t-1}(s_i, a_i)] \tag{28}$$

Now, we can adjust the updating rule in (23) by placing the intuition of ST i in place of the allocation strategies of other STs. The new rule becomes,

$$Q_i^{t+1}(s_i, a_i) = (1-\alpha^t)Q_i^t(s_i, a_i) + \alpha^t \Big\{ \sum_{a_{-i} \in A_{-i}} R_i(s_i, a_i, a_{-i}) [\mu_i^t(s_i, a_{-i}) - \mu_i^{t-1}(s_i, a_{-i})] + \beta \max_{b_i \in A_i} Q_i^t(s_i', b_i) \Big\} \tag{29}$$

The updating rule in (29) emphasizes the point that ST i exploits its intuition factor variation to estimate how other STs strategies change in the stochastic learning process. Balancing exploration and exploitation is an essential issue in the stochastic power allocation process. Exploration aims to try new allocation strategies so it does not only apply the strategies it already knows to be good but also explores new ones. Exploitation is the process of using well-established strategies. The most common technique to achieve this balance is to use the ϵ-greedy selection [29]. However, this approach selects equally among the available actions i.e. (the worst action is likely to be chosen as the best one). In order to overcome the drawback of the ϵ-greedy approach, the action selection probabilities are varied as a graded function of the Q-value. The best power level is given the highest selection probability while all other levels are ranked according to their Q-values. The learning algorithm exploits Boltzmann probability distribution to determine the probability of the power allocation action that fulfill the energy efficiency maximization constraints in C.1 to C.4. Thus, ST i selects the action a_i in state s_i at time slot t with the following probability,

$$\pi_i^t(s_i, a_i) = \frac{e^{Q_i^t(s_i, a_i)/\tau}}{\sum_{b \in A_i} e^{Q_i^t(s_i, b_i)/\tau}} \tag{30}$$

where τ is a positive integer that controls the selection probability. With high value of τ, the action probabilities become nearly equal. However, low value of τ causes big difference in selection probabilities for actions with different Q-values.

B. Approximated-Intuition Based Power Allocation

The computational complexity of the system increases along with the size of the states and action spaces. The simple look-up table where separate Q-value is maintained for each state/action pair is not feasible in large space with massive number of states like our system. Therefore, we propose a brief representation for the Q-values in which they are approximated as a function of much smaller set of variables. The compact representation of Q using function approximator $Q' : S' \times A$ is achieved by employing a vector $\xi = \{\xi_z\}_{z=1}^Z$ to minimize the metric of difference between the optimal Q-value $Q^*(s_i, a_i)$ and the approximated one $Q_i^{'t}(s_i, a_i, \xi)$. The approximated Q-value is expressed as follows,

$$Q_i^{'t}(s_i, a_i, \xi) = \sum_{z=1}^{Z} \xi_z \psi_z(s_i, a_i) = \xi \psi^T(s_i, a_i) \tag{31}$$

where T denotes the transpose operator, each scalar $\psi_z(s_i, a_i)$ is defined as the basis function (BF) over $S' \times A$, and ξ_z are the associated weights. The right hand side of (31) presents the vectors of the corresponding variables. We use the gradient function $\psi(s_i, a_i)$ to combine the online learning model with the brief representation. As a result, the update rule for the Q-value stated in (29) takes the following form,

$$\xi_i^{t+1}\psi^T(s_i, a_i) = \Big\{(1-\alpha^t)\xi_i^t\psi^T(s_i, a_i) + \alpha^t \Big[\sum_{a_{-i} \in A_{-i}} R_i(s_i, a_i, a_{-i})[\mu_i^t(s_i, a_i) - \mu_i^{t-1}(s_i, a_i)] + \beta \max_{b_i \in A_i} \xi_i^t \psi^T(s_i', b_i) \Big] \Big\} \psi(s_i, a_i) \tag{32}$$

The gradient function $\psi(s_i, a_i)$ is a partial derivative with respect to the elements of ξ^t. Moreover, the probability of selecting certain action presented in (30) is updated with the Q-value approximation as follows,

$$\pi_i^t(s_i, a_i) = \frac{e^{\xi_i^t \psi^T(s_i, a_i)/\tau}}{\sum_{b \in A_i} e^{\xi_i^t \psi^T(s_i, b_i)/\tau}} \tag{33}$$

The intuition based online learning process with approximated Q-value is illustrated in Algorithm 1. The algorithm starts by initializing the power allocation strategy, intuition factor and the approximated Q-value for each state belong to the reduced state space. Once the state is initialized, certain transmission power is selected for the corresponding ST according to the probability in (33). If the conditions C.1 to C.4 are satisfied, then, the reward is achieved and the Q-value, intuition factor and power allocation strategy are updated and the new state is observed.

C. Power Allocation Algorithm Convergence

In this section, we prove the convergence of the proposed approximated intuition-based online learning algorithm for power allocation. Our proof relies on exploiting ordinary differential equations (ODE) to acquire the necessary conditions for convergence. The following assumptions are required to proceed with the proof:

Assumption 1. *The basis functions $\psi_z(s_i, a_i)$ are linearly independent for all (s_i, a_i) and all the properties of $Q_i^t(s_i, a_i)$ in previous discussion are applicable to the dot product for the vectors $\xi_i^t \psi^T(s_i, a_i)$.*

Assumption 2. *For every $z = (1, 2....Z)$, $\psi_z(s_i, a_i)$ is bounded, which means $E\{\psi_z^2(s_i, a_i)\} < \infty$ and the reward function satisfies $E\{R_i^2(s_i, a_i, a_{-i})\} < \infty$.*

Algorithm 1 Approximated intuition based online learning algorithm for power allocation

Require: $\pi_i^t(s_i, a_i)$, t, $w_i > 0$, $\gamma_{n,x}^*$,γ_d^* and γ_k^*
Ensure: Transmission power allocation for STs
initialization
Let $t = 0$
for each(s_i, $a_i \in A_i$) **do**
  initialize power allocation strategy $\pi_i^t(s_i, a_i)$;
  initialize approximated Q-value $\xi_i^t \psi^T(s_i, a_i)$;
  initialize intuition factor $\mu_i^t(s_i, a_{-i})$;
end for
evaluate the state $s_i = s_i^t$
while (true) **do**
  Select action a_i according to $\pi_i^t(s_i, a_i)$;
  Measure the received $\gamma_{n,x}$, γ_d and γ_k with feedback from the receiver and observe the state s_i' by identifying P_i and comparing SINR;
  if ($\gamma_{n,x} \geq \gamma_{n,x}^*$, $\gamma_d \geq \gamma_d^*$ and $\gamma_k \geq \gamma_k^*$) **then**
    $R_i(s_i, a_i, a_{-i})$ is achieved ;
  else
    $R_i(s_i, a_i, a_{-i}) = 0$ as the receiver could not receive the data correctly
  end if
  Update $\xi_i^{t+1}\psi^T(s_i, a_i)$ based on $\mu_i^t(s_i, a_{-i})$ according to (32)
  Update $\pi_i^{t+1}(s_i, a_i)$ according to (33)
  Update $\mu_i^{t+1}(s_i, a_{-i})$ according to (27)
  $s_i = s_i^{t+1}$
  $t = t + 1$
end while

Assumption 3. *The learning rate satisfies* $\sum_{t=1}^{\infty} \alpha^t = \infty$ *and* $\sum_{t=1}^{\infty} (\alpha^t)^2 < \infty$.

Definition 1. *Let* $\Psi = E[\psi^T(s_i, a_i)\psi(s_i, a_i)]$. *For the parameter vector* ξ *and a particular network state* $s_i \in S'$, *we define a vector* $\psi(s_i, \xi) = [\psi_z(s_i, a_i)]$ *for* $z = 1 \rightarrow Z$ *where* $a_i \in \{a_i = arg\ \max_{b_i \in A_i} \xi_i \psi^T(s_i, b_i)\}$ *is the set of optimal power allocation actions for* s_i. *We define the following* ξ*-dependent matrix:*

$$\Psi' = E[\psi^T(s_i, \xi)\psi(s_i, \xi)] \tag{34}$$

Proposition 1. *With the assumptions 1-3 and Definition 1, the intuition based online learning with approximation converges with probability (w.p) 1, if*

$$\Psi' < \Psi,\ \forall \xi \tag{35}$$

Proof: The proof of convergence is linked to finding stable fixed points of the ODE defined based on the expectation of the derivative of the update rule in (32) with respect to t as follows,

$$\xi_i^{\cdot t} = E[(\sum_{a_{-i} \in A_{-i}} [\mu_i^t(s_i, a_i) - \mu_i^{t-1}(s_i, a_i)] R_i(s_i, a_i, a_{-i})$$

$$+\beta \xi^t \psi^T(s_i', \xi^t) - \xi^t \psi^T(s_i, a_i))\psi(s_i, a_i)] \tag{36}$$

where $\xi_i^{\cdot t} = \frac{\partial \xi}{\partial t}$ as $\alpha \rightarrow 0$. From the definition of the intuition factor in (27), we state,

$$\sum_{a_{-i} \in A_{-i}} \big[[\mu_i^t(s_i, a_i) - \mu_i^{t-1}(s_i, a_i)] R_i(s_i, a_i, a_{-i})]\big]$$

$$= \sum_{a_{-i} \in A_{-i}} w_i [\pi_i(s_i, a_i) - \pi_i'(s_i, a_i)] R_i(s_i, a_i, a_{-i}) \tag{37}$$

By substituting the value of $\pi_i(s_i, a_i)$ from (33), and when τ is large, we obtain,

$$e^{\xi_i^t \psi^T(s_i, a_i)/\tau} = 1 + \frac{\xi_i \psi^T(s_i, a_i)}{\tau} + \rho(\frac{\xi_i \psi^T(s_i, a_i)}{\tau})$$

where $\rho(\frac{\xi_i \psi^T(s_i, a_i)}{\tau})$ is a polynomial of order $O((\frac{\xi_i \psi^T(s_i, a_i)}{\tau})^2)$, we can easily find,

$$\sum_{b \in A_i} e^{\xi_i^t \psi^T(s_i, b_i)/\tau} = m_i + 1 + \big[\frac{\xi_i \psi^T(s_i, b_i)}{\tau} + \rho \frac{\xi_i \psi^T(s_i, b_i)}{\tau}\big]$$

where m_i is the number of power levels considered in certain range. Consequently, we get the following,

$$\pi_i(s_i, a_i) = \frac{1}{m_i + 1} + \frac{1}{m_i + 1} . \frac{\xi_i \psi^T(s_i, a_i)}{\tau} + \varrho(\frac{\xi_i \psi^T(s_i, b_i)}{\tau}) \tag{38}$$

where $\varrho(\frac{\xi_i \psi^T(s_i, b_i)}{\tau})$ is a polynomial of order smaller than $O(\frac{\xi_i \psi^T(s_i, a_i)}{\tau})$, Note that the coefficient of the polynomial is independent of the vector Q-value. The reference strategy is evaluated according to the historical optimal actions as follows,

$$\pi_i'(s_i, a_i) = \frac{1}{m_i + 1} + \frac{1}{m_i + 1} . \frac{\xi_i \psi^T(s_i, \xi)}{\tau} + \varrho(\frac{\xi_i \psi^T(s_i, b_i)}{\tau}) \tag{39}$$

By substituting (38) and (39) in (37), we get,

$$\sum_{a_{-i} \in A_{-i}} \big[[\mu_i^t(s_i, a_i) - \mu_i^{t-1}(s_i, a_i)] R_i(s_i, a_i, a_{-i})]\big]$$

$$= \frac{\sum_{a_{-i} \in A_{-i}} w_i R_i(s_i, a_i, a_{-i})}{\tau} . \frac{1}{m_i + 1} [\xi_i \psi^T(s_i, a_i) -$$

$$\xi_i \psi^T(s_i, \xi)] + \varrho(\frac{\xi_i \psi^T(s_i, b_i)}{\tau}) - \varrho(\frac{\xi_i \psi^T(s_i, b_i)}{\tau})$$

If we assume a large value of τ, then

$$\sum_{a_{-i} \in A_{-i}} \big[[\mu_i^t(s_i, a_i) - \mu_i^{t-1}(s_i, a_i)] R_i(s_i, a_i, a_{-i})]\big]$$

$$\leq \frac{1 - \beta}{m_i + 1} \big[\xi_i \psi^T(s_i, a_i) - \xi_i \psi^T(s_i, \xi)\big]$$

Let us assume that $V = \frac{1-\beta}{m_i+1}$ to simplify the notation. Now, (36) can be expressed as follows,

$$\xi^{\cdot t} = E\big[\big(V[\xi^t \psi^T(s_i, a_i) - \xi^t \psi^T(s_i, \xi^t)] + \beta \xi^t \psi^T(s_i', \xi^t)$$

$$-\xi^t \psi^T(s_i, a_i)\big)\psi(s_i, a_i)\big] \tag{40}$$

We define two trajectories of the ODE ξ_1^t and ξ_2^t that have different initial conditions and satisfies $\xi_0^t = \xi_1^t - \xi_2^t$. Then, we have

$$\frac{\partial \|\xi_0^t\|^2}{\partial t} = 2(\xi_1^{\cdot t} - \xi_2^{\cdot t})(\xi_0^t)^T$$

$$= E\big[(-2V\xi_1^t\psi^T(s_i,\xi_1^t)+2\beta\xi_1^t\psi^T(s_i',\xi_1^t))\psi(s_i,a_i)(\xi_0^t)^T- (-2V\xi_2^t\psi^T(s_i,\xi_2^t)+2\beta\xi_2^t\psi^T(s_i',\xi_2^t))\psi(s_i,a_i)(\xi_0^t)^T\big] +(2V-2)\xi_0^t\Psi(\xi_0^t)^T \quad (41)$$

From Definition 1, we can deduce the following two inequalities,

$$\xi_1^t\psi^T(s_i',\xi_1^t) \le \xi_1^t\psi^T(s_i',\xi_2^t) \quad (42)$$

$$\xi_2^t\psi^T(s_i',\xi_2^t) \le \xi_2^t\psi^T(s_i',\xi_1^t) \quad (43)$$

As the expectation E in (41) is taken over different states and different actions, we can define two sets $\Lambda_+ = \{(s_i,a_i) \in S_i \times A_i | \xi_0^t\psi^T(s_i,a_i) > 0\}$ and $\Lambda_- \in S_i \times A_i - \Lambda_+$. If we combine (42) and (43) in (41), we get,

$$\frac{\partial\|\xi_0^t\|^2}{\partial t} \le E\big[(-2V\xi_0^t\psi^T(s_i,\xi_2^t)+2\beta\xi_0^t\psi^T(s_i',\xi_2^t))\psi(s_i,a_i)(\xi_0^t)^T\big|\Lambda_+\big] +E\big[(-2V\xi_0^t\psi^T(s_i,\xi_1^t)+2\beta\xi_0^t\psi^T(s_i',\xi_1^t))\psi(s_i,a_i)(\xi_0^t)^T\big|\Lambda_-\big] +(2V-2)\xi_0^t\Psi(\xi_0^t)^T \quad (44)$$

After the application of Holder's inequality [30] to the expectation in (44), we get

$$\frac{\partial\|\xi_0^t\|^2}{\partial t} \le \bigg(-2V\sqrt{E\big[\big(\xi_0^t\psi^T(s_i,\xi_2^t)\big)^2\big|\Lambda_+\big]} +2\beta\sqrt{E\big[\big(\xi_0^t\psi^T(s_i',\xi_2^t)\big)^2\big|\Lambda_+\big]}\bigg)\times\sqrt{E\big[\big(\psi(s_i,a_i)(\xi_0^t)^T\big)^2\big|\Lambda_+\big]} + \bigg(-2V\sqrt{E\big[\big(\xi_0^t\psi^T(s_i,\xi_1^t)\big)^2\big|\Lambda_-\big]}+ 2\beta\sqrt{E\big[\big(\xi_0^t\psi^T(s_i',\xi_1^t)\big)^2\big|\Lambda_-\big]}\bigg)\times\sqrt{E\big[\big(\psi(s_i,a_i)(\xi_0^t)^T\big)^2\big|\Lambda_-\big]} +(2V-2)\xi_0^t\Psi(\xi_0^t)^T$$

$$\le \bigg(-2V\sqrt{E\big[\big(\xi_0^t\psi^T(s_i,\xi_2^t)\big)^2\big]}+2\beta\sqrt{E\big[\big(\xi_0^t\psi^T(s_i',\xi_2^t)\big)^2\big]}\bigg) \times\sqrt{E\big[\big(\psi(s_i,a_i)(\xi_0^t)^T\big)^2\big|\Lambda_+\big]} + \bigg(-2V\sqrt{E\big[\big(\xi_0^t\psi^T(s_i,\xi_1^t)\big)^2\big]}+2\beta\sqrt{E\big[\big(\xi_0^t\psi^T(s_i',\xi_1^t)\big)^2\big]}\bigg) \times\sqrt{E\big[\big(\psi(s_i,a_i)(\xi_0^t)^T\big)^2\big|\Lambda_-\big]}+(2V-2)\xi_0^t\Psi(\xi_0^t)^T$$

If we apply the definition of Ψ' in Definition 1, we get,

$$\le (-2V+2\beta)\sqrt{max\big[\xi_0^t\Psi_1'(\xi_0^t)^T,\xi_0^t\Psi_2'(\xi_0^t)^T\big]} \times\bigg(\sqrt{E\big[\big(\psi(s_i,a_i)(\xi_0^t)^T\big)^2\big|\Lambda_+\big]}+ \sqrt{E\big[\big(\psi(s_i,a_i)(\xi_0^t)^T\big)^2\big|\Lambda_-\big]}\bigg)+(2V-2)\xi_0^t\Psi(\xi_0^t)^T$$

$$\le (-2V+2\beta)\sqrt{max\big[\xi_0^t\Psi_1'(\xi_0^t)^T,\xi_0^t\Psi_2'(\xi_0^t)^T\big]} \times\sqrt{E\big[\big(\psi(s_i,a_i)(\xi_0^t)^T\big)^2\big]}+(2V-2)\xi_0^t\Psi(\xi_0^t)^T \quad (45)$$

According to the condition in (35), we can state that,

$$\frac{\partial\|\xi_0^t\|^2}{\partial t} < (-2V+2\beta)\xi_0^t\Psi(\xi_0^t)^T+(2V-2)\xi_0^t\Psi(\xi_0^t)^T = (2\beta-2)\xi_0^t\Psi(\xi_0^t)^T < 0 \quad (46)$$

which means that ξ_0^t converges to the origin and this confirms that there exists a stable point of the ODE in (36). Thus, the proposed intuition based online learning with Q approximation converges w.p 1. ■

Consequently, the stable point $\xi^*$of the ODE in (36) indicates that,

$$0 = E[(\sum_{a_{-i}\in A_{-i}}[\mu_i^t(s_i,a_i)-\mu_i^{t-1}(s_i,a_i)]R_i(s_i,a_i,a_{-i}) +\beta\xi^*\psi^T(s_i',\xi^*)-\xi^*\psi^T(s_i,a_i))\psi(s_i,a_i)] \quad (47)$$

and ξ^* can be found as follows,

$$\xi^* = E[(\sum_{a_{-i}\in A_{-i}}[\mu_i^t(s_i,a_i)-\mu_i^{t-1}(s_i,a_i)]R_i(s_i,a_i,a_{-i}) +\beta\xi^*\psi^T(s_i',\xi^*))\psi(s_i,a_i)\big]\Psi^{-1} \quad (48)$$

As a result, the optimal intuition based approximated online Q-function is defined as follows

$$Q'(s_i,a_i,\xi^*) = \xi^*\psi(s_i,a_i) \quad (49)$$

VI. Numerical Results

In this section, we present the simulation results obtained to demonstrate the capability of our proposed scheme for energy efficient power allocation. The simulation environment comprises multi-tier heterogeneous 5G network constructed according to the network model in section III, where multiple picocells, femtocells and D2D share the spectrum resources with one macrocell in an underlay fashion. The multi-tier network is composed of one macrocell, 4 picocell, 8 femtocells and variable number of D2D connections. The pico, femto BSs and D2D users are uniformly distributed within an area centered by the macro BS with radiuses of 400m for the macrocell, 75 m for the picocell, and 20 m for the femtocell. All the users are assumed to have identical and independent Rayleigh fading channels. The channel gain is given as $G = F(d)^{-c}$ where F is shadowing factor, which is a random number generated with log normal distribution with mean 0 and variance 6 dB, and d is the physical distance between the user and its BS. The learning rate α is chosen to be dynamic according to the Win or learn fast principle, which states that the learning agent should learn faster when it is losing and more slowly upon winning [31]. The learning rates that we used are $\alpha = 0.2$ for rewarded solution and $\alpha = 0.6$ for the punished one. The rest of the simulation parameters are presented in Table I.

The service request arrival in each time t follows Poisson distribution with an arrival rate of 6λ where $\lambda = 0.3$. The thresholds $\gamma_{n,x}^*$, γ_d^* and γ_k^* are specified as follows, $\gamma_{n,x}^*$ is 20 dB and 15 dB for pico and femto receivers respectively. The γ_d^* is -8 dB, while γ_k^* is 30 dB. We compare the performance of our proposed scheme with non-cooperative online learning

Simulation Parameter	Value
Bandwidth	20 MHz
Discount factor (β)	0.9
Number of MUEs	10
Pico BS Tx power	20 to 28 dBm
Femto BS Tx power	12 to 17 dBm
Macro BS Tx power	45 dBm
DUE Tx power	-8 to -2 dBm
p_{cc} for macro. pico, femto, D2D transmitters respectively	130 W, 15 W, 5 W, 0.1 W
Thermal noise density (σ)	-174 dBm/Hz

Table I
5G ENVIRONMENT SIMULATION PARAMETERS

(NCOL) where STs act selfishly to improve their own rewards. In addition, we compare it to the stackelberg game (SG) theory for power allocation proposed in [15], heuristic based scheme in [23] and sub-optimal algorithm (SOA) proposed in [21]. Note that schemes in [23] and [21] require awareness of the network model and rely on information exchange between the secondary BSs. The proposed scheme performance is demonstrated in four sets of simulations.

In the first simulation set, the average achieved energy efficiency (EE) for the whole system is evaluated. The convergence behavior is estimated for each scheme by showing how the system behaves as a function of the number of epoch as in Figure 2. The impact of the number of STs on the system

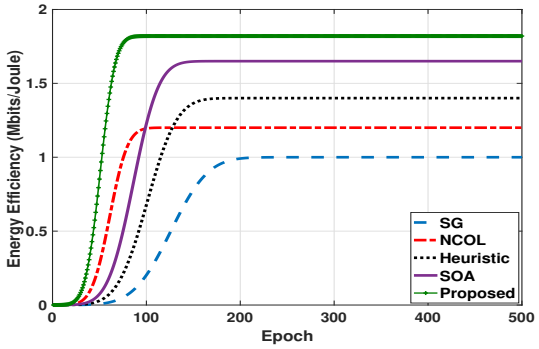

Figure 2. Average system energy efficiency of different schemes

energy efficiency is evaluated in Figure 3, where there are 4 pico BSs, 8 femto BSs and 6 D2D connections. Figure 4 presents the achieved system energy efficiency across the channel gain to noise ratio (CNR) from 0 to 30 dB. Unlike the other schemes, we notice that the proposed scheme converges faster than other schemes due to the approximation of Q-value which reduces the state space and information exchange overhead. Consequently, this validates **Proposition 1**. Moreover, our scheme records high energy efficiency compared to others. This is due to the fact that intuition about other agents' strategies through online learning approach improves the quality of the action selected in each state. One observation to note is that the energy efficiency first increases then decreases as the number of STs increases. This is because when there are only few STs in the heterogeneous 5G network, the inter-cell interference caused by the macro BS and other STs is comparably low and there exists enough signal spaces to compensate for its impact. However, when there are more STs, the inter-cell interference becomes significant and more power is required to satisfy the minimum rate target, resulting in a lower energy efficiency.

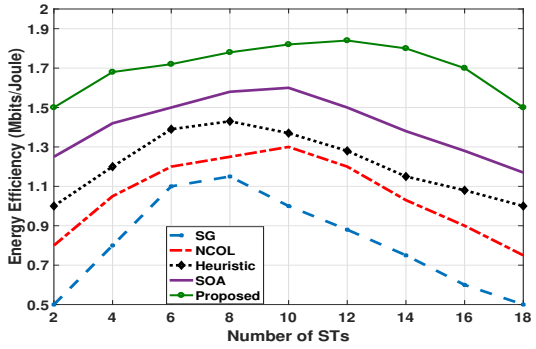

Figure 3. Impact of the number of STs on system energy efficiency

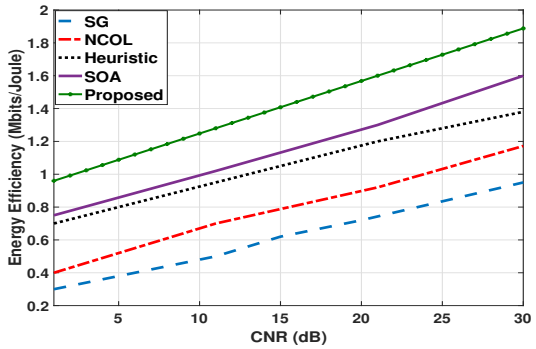

Figure 4. Average system energy efficiency under different CNR ratio

The second simulation set demonstrate the scheme performance in terms of spectral efficiency (SE). This simulation part measures the speed of convergence as in Figure 5, which presents the average spectral efficiency of the whole 5G system. In addition, we study the impact of the number of

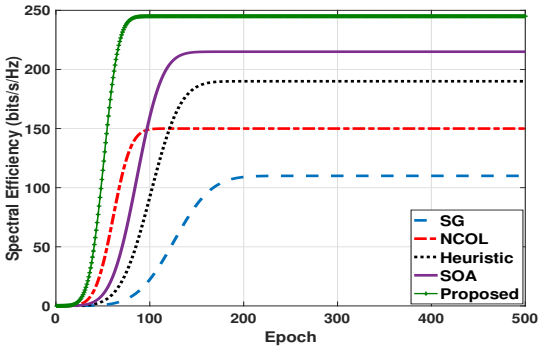

Figure 5. Average system spectral efficiency of different schemes

STs from certain type on the achieved spectral efficiency per each ST of that type. For example, the average spectral efficiency per each femtocell is plotted in Figure 6 against the number of femtocells in the system. Another evaluation metric considers the impact of the choice of the maximum transmission power of STs on the system spectral efficiency. We select picocells as an instance to measure the impact of the maximum transmission power choice. Figure 7 shows the achieved system spectral efficiency with variable maximum transmission power of pico BS. It is clear that our scheme achieved the highest spectral efficiency with the fastest convergence among other schemes. Another notice is that the spectral efficiency decreases with the increase in the number of STs

(femto BSs) in Figure 6. We also observe that the interference threshold imposed by the macrocell transmission prevents any performance improvement even when increasing the pico BS transmission power as in Figure 7.

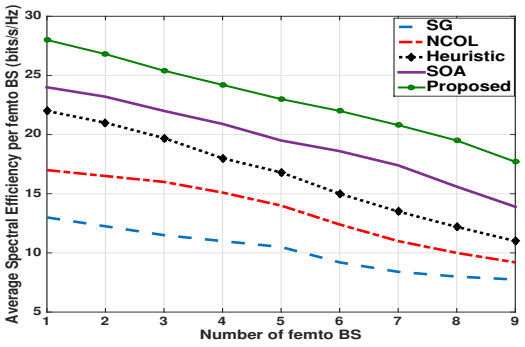

Figure 6. Average spectral efficiency per each femtocell with variable number of femtocells

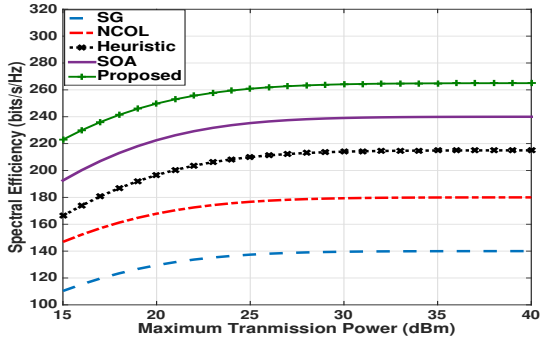

Figure 7. Average system spectral efficiency as a function of maximum transmission power of pico BS

The third simulation set focus on the evaluation of users QoS satisfaction. Figure 8 and Figure 9 present the average SINR measured at the macro receivers and pico receivers respectively for all the schemes. We notice that our scheme

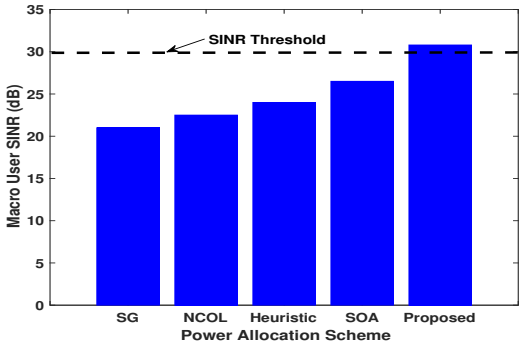

Figure 8. Average SINR for macro users of different schemes

achieved higher SINR than others and maintains SINR above threshold for both receivers. This confirms the satisfaction of constraints C.1 and C.2 as the proposed scheme managed to maintain QoS for both macro and secondary tier users.

In the fourth simulation set, we aim at demonstrating the capability of our learning scheme to produce acceptable results in terms of QoS with variable number of users (UEs) in the network. Thus, we plot as in Figure 10, the achieved SINR

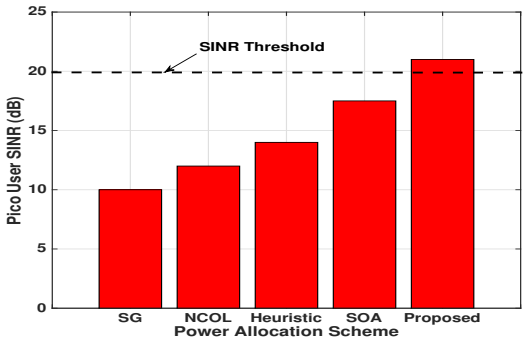

Figure 9. Average SINR for pico users of different schemes

for MUEs against the total number of UEs. In addition, the SINR for each pico user, which is selected as a representation for the SUEs is plotted in Figure 11. Figure 12 presents the achieved SINR for D2D receiver with variable number of UEs. The contribution of each tier UEs in the total number of network UEs is determined with the following percentages: 25 % MUEs, 25 % pico users, 25 % femto users, and 25 % D2D users. We notice that our scheme achieved the highest SINR for both tier users compared to other schemes. In addition, it managed to maintain the minimum SINR required by each tier even with high number of UEs. These results are anticipated as the machine learning implemented by our power allocation scheme enforces the QoS constraints during the learning process.

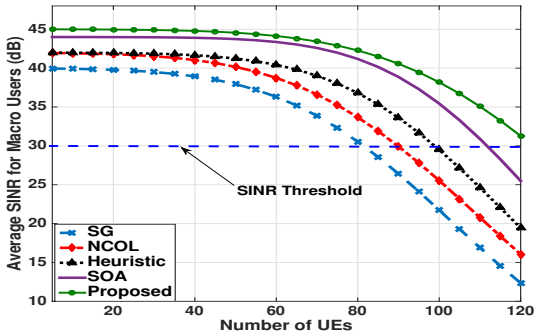

Figure 10. Average SINR for each macro UE with variable number of UEs in the network

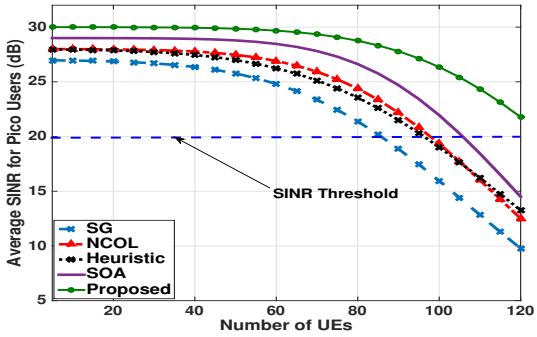

Figure 11. Average SINR for each pico UE with variable number of UEs in the network

VII. Conclusion

In this paper, power allocation problem is tackled for the downlink transmission in a spectrum sharing multi-tier 5G environment, where small cells and D2D access the spectrum in

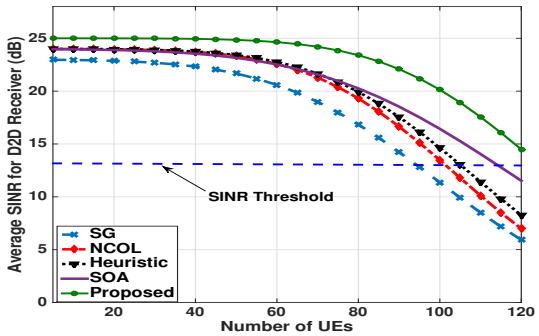

Figure 12. Average SINR for each D2D receiver with variable number of UEs in the network

an underlay fashion. We proposed an enhanced online learning based scheme to allocate transmission power to reduce the overall power consumption while maintaining QoS for both primary tier and secondary tier. The online learning exploits an intuition based approach to account for the impact of other STs actions on the selected transmission power strategy. In addition, the scheme employs an approximation mechanism for the Q-value, which reduces the state/action space and expedites the speed of convergence. The performance of the proposed scheme was demonstrated through simulation where it achieved faster convergence and higher energy efficiency compared to other schemes.

REFERENCES

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

[2] M. Agiwal, A. Roy, and N. Saxena, "Next generation 5g wireless networks: A comprehensive survey", *IEEE Communications Surveys Tutorials*, vol. 18, no. 3, pp. 1617–1655, thirdquarter 2016.

[3] 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.

[4] T. O. Olwal, K. Djouani, and A. M. Kurien, "A survey of resource management toward 5g radio access networks", *IEEE Communications Surveys Tutorials*, vol. 18, no. 3, pp. 1656–1686, thirdquarter 2016.

[5] R. Q. Hu and Y. Qian, "An energy efficient and spectrum efficient wireless heterogeneous network framework for 5g systems", *IEEE Communications Magazine*, vol. 52, no. 5, pp. 94–101, May 2014.

[6] S. Buzzi, C. L. I, T. E. Klein, H. V. Poor, C. Yang, and A. Zappone, "A survey of energy-efficient techniques for 5g networks and challenges ahead", *IEEE Journal on Selected Areas in Communications*, vol. 34, no. 4, pp. 697–709, April 2016.

[7] Abhijit Gosavi, "Reinforcement learning: A tutorial survey and recent advances", *INFORMS J. on Computing*, vol. 21, no. 2, pp. 178–192, Apr. 2009.

[8] C. Yang, J. Li, M. Guizani, A. Anpalagan, and M. Elkashlan, "Advanced spectrum sharing in 5g cognitive heterogeneous networks", *IEEE Wireless Communications*, vol. 23, no. 2, pp. 94–101, April 2016.

[9] S. Guruacharya, D. Niyato, Dong In Kim, and E. Hossain, "Hierarchical competition for downlink power allocation in ofdma femtocell networks", *IEEE Transactions on Wireless Communications,*, vol. 12, no. 4, pp. 1543–1553, April 2013.

[10] I. AlQerm and B. Shihada, "A cooperative online learning scheme for resource allocation in 5g systems", in *2016 IEEE International Conference on Communications (ICC)*, May 2016, pp. 1–7.

[11] Ismail AlQerm and Basem Shihada, "Cognitive aware interference mitigation scheme for LTE femtocells", in *Cognitive Radio Oriented Wireless Networks - 10th International Conference, CROWNCOM 2015, Doha, Qatar, April 21-23, 2015, Revised Selected Papers*, 2015, pp. 607–619.

[12] M. J. Abdel-Rahman, M. AbdelRaheem, and A. B. MacKenzie, "Stochastic resource allocation in opportunistic lte-a networks with heterogeneous self-interference cancellation capabilities", in *IEEE International Symposium on Dynamic Spectrum Access Networks (DySPAN)*, Sept 2015, pp. 200–208.

[13] C. Bouras and G. Diles, "Resource management in 5g femtocell networks", in *2015 10th International Conference on Broadband and Wireless Computing, Communication and Applications (BWCCA)*, Nov 2015, pp. 353–358.

[14] V. Chandrasekhar, J.G. Andrews, Tarik Muharemovict, Zukang Shen, and Alan Gatherer, "Power control in two-tier femtocell networks", *IEEE Transactions on Wireless Communications,*, vol. 8, no. 8, pp. 4316–4328, August 2009.

[15] M. Haddad, P. Wiecek, O. Habachi, and Y. Hayel, "A game theoretic analysis for energy efficient heterogeneous networks", in *12th International Symposium on Modeling and Optimization in Mobile, Ad Hoc, and Wireless Networks (WiOpt),*, May 2014, pp. 241–246.

[16] H. Munir, S. A. Hassan, H. Pervaiz, and Q. Ni, "A game theoretical network-assisted user-centric design for resource allocation in 5g heterogeneous networks", in *IEEE 83rd Vehicular Technology Conference (VTC Spring)*, May 2016, pp. 1–5.

[17] S. Navaratnarajah, A. Saeed, M. Dianati, and M. A. Imran, "Energy efficiency in heterogeneous wireless access networks", *IEEE Wireless Communications*, vol. 20, no. 5, pp. 37–43, October 2013.

[18] A. Saeed, E. Katranaras, A. Zoha, A. Imran, M. A. Imran, and M. Dianati, "Energy efficient resource allocation for 5g heterogeneous networks", in *IEEE 20th International Workshop on Computer Aided Modelling and Design of Communication Links and Networks (CAMAD)*, Sept 2015, pp. 119–123.

[19] Y. Huang, X. Zhang, J. Zhang, J. Tang, Z. Su, and W. Wang, "Energy-efficient design in heterogeneous cellular networks based on large-scale user behavior constraints", *IEEE Transactions on Wireless Communications*, vol. 13, no. 9, pp. 4746–4757, Sept 2014.

[20] 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.

[21] J. Tang, D. K. C. So, E. Alsusa, K. A. Hamdi, and A. Shojaeifard, "Resource allocation for energy efficiency optimization in heterogeneous networks", *IEEE Journal on Selected Areas in Communications*, vol. 33, no. 10, pp. 2104–2117, Oct 2015.

[22] M. Adedoyin and O. Falowo, "An energy-efficient radio resource allocation algorithm for heterogeneous wireless networks", in *IEEE 27th Annual International Symposium on Personal, Indoor, and Mobile Radio Communications (PIMRC)*, Sept 2016, pp. 1–6.

[23] Elmahdi Driouch, Wessam Ajib, and Chadi Assi, "Power control and clustering in heterogeneous cellular networks", *Wireless Networks*, pp. 1–12, 2016.

[24] Z. Yang, W. Xu, H. Xu, J. Shi, and M. Chen, "User association, resource allocation and power control in load-coupled heterogeneous networks", in *IEEE Globecom Workshops (GC Wkshps)*, Dec 2016, pp. 1–7.

[25] M. Peng, K. Zhang, J. Jiang, J. Wang, and W. Wang, "Energy-efficient resource assignment and power allocation in heterogeneous cloud radio access networks", *IEEE Transactions on Vehicular Technology*, vol. 64, no. 11, pp. 5275–5287, Nov 2015.

[26] Richard S. Sutton and Andrew G. Barto, *Introduction to Reinforcement Learning*, MIT Press, Cambridge, MA, USA, 1st edition, 1998.

[27] M. Bennis, S. Guruacharya, and D. Niyato, "Distributed learning strategies for interference mitigation in femtocell networks", in *IEEE Global Telecommunications Conference (GLOBECOM)*, Dec 2011, pp. 1–5.

[28] J. Papandriopoulos and J. S. Evans, "Scale: A low-complexity distributed protocol for spectrum balancing in multiuser dsl networks", *IEEE Transactions on Information Theory*, vol. 55, no. 8, pp. 3711–3724, Aug 2009.

[29] Eduardo Rodrigues Gomes and Ryszard Kowalczyk, "Dynamic analysis of multiagent q-learning with -greedy exploration", in *Proceedings of the 26th Annual International Conference on Machine Learning*, New York, NY, USA, 2009, ICML '09, pp. 369–376, ACM.

[30] J.M. Aldaz, "A stability version of hlder's inequality", *Journal of Mathematical Analysis and Applications*, vol. 343, no. 2, pp. 842 – 852, 2008.

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