A Reinforcement Learning based Game Theoretic Approach for Distributed Power Control in Downlink NOMA

Ashish Rauniyar[†*], Anis Yazidi[*§], Paal Engelstad[†*], Olav N. Østerbø[‡]
[†]Department of Technology Systems, University of Oslo (UiO), Norway
[*]Department of Computer Science, OsloMet - Oslo Metropolitan University, Norway
[§]Department of Computer Science, Norwegian University of Science and Technology (NTNU), Norway
[‡]Telenor Research, Norway
Email: (ashish.rauniyar, paal.engelstad, anisy)@oslomet.no, olav.osterbo@getmail.no

***Abstract*—Optimal power allocation problem in wireless networks is known to be usually a complex optimization problem. In this paper, we present a simple and energy-efficient distributed power control in downlink Non-Orthogonal Multiple Access (NOMA) using a Reinforcement Learning (RL) based game theoretical approach. A scenario consisting of multiple Base Stations (BSs) serving their respective Near User(s) (NU) and Far User(s) (FU) is considered. The aim of the game is to optimize the achievable rate fairness of the BSs in a distributed manner by appropriately choosing the power levels of the BSs using trials and errors. By resorting to a subtle utility choice based on the concept of marginal price costing where a BS needs to pay a virtual tax offsetting the result of the interference its presence causes for the other BS, we design a potential game that meets the latter objective. As RL scheme, we adopt Learning Automata (LA) due to its simplicity and computational efficiency and derive analytical results showing the optimality and convergence of the game to a Nash Equilibrium (NE). Numerical results not only demonstrate the convergence of the proposed algorithm to a desirable equilibrium maximizing the fairness, but they also demonstrate the correctness of the proposal followed by thorough comparison with random and heuristic approaches.**

***Index Terms*—Game Theory, NOMA, IoT, Power Allocation, Reinforcement Learning, Nash Equilibrium**

I. Introduction

With the exponential growth in the Internet of Things (IoT) applications, it is estimated that global data traffic will cross up to 75 trillion gigabytes [1]. Orthogonal Multiple Access (OMA) is currently used for the underlying wireless network such as the Fourth Generation (4G) and Long Term Evolution (LTE) by assigning orthogonal resources to the multiple users and thereby avoiding inter-cell interference [2]. However, it has been predicted that more than 125 billion IoT devices will be connected to the internet by 2030 [3]. Hence, the next generation of networks is expected to serve the massive connectivity of the IoT devices and maximize resource efficiency. In this regard, OMA is considered as spectrally inefficient for the design and optimization of the next-generation wireless systems [4].

To provide a higher data rate, higher capacity, and massive connectivity of the IoT devices, Non-Orthogonal Multiple Access (NOMA) has been contemplated as one of the key technologies to support these requirements of the next generation networks [5], [6]. In NOMA, multiple users can be multiplexed in the power domain at the same time, frequency, and code [7]. Specifically, multiple users signal are superimposed in the power domain at the transmitter side, and the user signals are separated using the Signal-to-Interference Cancellation (SIC) technique at the receiver side [8].

Since resources are not used in an orthogonal manner in NOMA, it is important to efficiently manage interference among multiple users to maximize the system throughput or capacity. Moreover, NOMA is based on the principle of SIC which is known to be very fragile to interference as the decoding failure propagates in the SIC chain to weaker users [9]. Therefore, the power must be properly allocated such that the interfering signals can be correctly decoded and subtracted from the certain users' received signal to recover the desired signal [10]. This is particularly more important in large scale networks where Base Stations (BSs) might be densely placed to serve their associated multiple NOMA users [11]. If power control optimization is not used, the BS serving at a higher power levels to satisfy the individual achievable data rate of its associated NOMA users will create interference on the NOMA users associated with other BSs and thus jeopardize their achievable data rates.

Most of the power control schemes in the literature are based on scheduling, optimization techniques, heuristics, subcarrier allocation, and power allocation techniques. These techniques may not be globally efficient and thus may result into sub-optimal solutions. Furthermore, latter techniques are usually centralized and require large message exchanges between BSs and users. As BSs aspire to maximize the individual data rate of its associated NOMA users, they might act selfishly by raising their power level at the detriment of other users from other BSs which might get affected by the interference and thus fail in the SIC phase. There has recently been a growing interest in examining distributed power control in wireless networks from a game-theoretical perspective. Game theory

has its root in economics and strategic decision making and deals with the analysis of several decision-makers' competitive relationships [12], [13]. Game Theory is a powerful modeling tool in many systems where the outcome of a player does not only depend on its decision or action, but also on the decisions taking by other players. Rational users make calculated decisions to maximise their pay-off functions. Game theory methods are also one of the most viable candidates for distributed power control and management in downlink NOMA, whereby BS needs to choose their power levels for overall better network performance. It is worth mentioning that game theory has found a plethora of applications within the field of wireless networks, for a comprehensive survey we refer the reader to a book by Han et al. [12]. On the other hand, some Reinforcement Learning (RL) strategies, such as Learning Automata (LA) [14], would ultimately yield the optimum strategy as the learning parameter gets sufficiently small [15]–[17]. LA is one of the simplest and yet efficient RL schemes that are shown to reach Nash Equilibrium (NE) in a large set of games [15]. LA has been used extensively in the literature for different game wireless related problems such as power control [18] cooperative spectrum sensing [19], opportunistic spectrum sharing in cognitive networks [20], [21], sensor fusion [22], anti-jamming in wireless communication [23] to mention few recent applications. Therefore, in this paper, we propose and study an LA based game-theoretic approach for distributed power control in downlink NOMA systems.

In particular, the major contributions of this paper are as follows:

- We first formulate distributed power control in downlink NOMA as a strategic game and derive the Nash Equilibrium of the game.
- We prove that the distributed power control game we designed is an exact potential game. We then propose a LA based game-theoretic approach for distributed power control that provides a full characterization of the best achievable performance for the potential function of the game.
- We show that our proposed distributed power control algorithm that is designed as a game is guaranteed to converge to an NE.
- We conduct a thorough theoretical analysis that demonstrates the convergence of the proposed algorithm that is maximizing the achievable rate fairness for the respective NUs and FUs and thus achieving the higher energy efficiency in downlink NOMA systems.

The rest of the paper is organized as follows. In Section II, we survey the related works. The system model is presented in Section III. Our proposed RL based game-theoretic approach for distributed power control algorithm is explained in Section IV. Theoretical proofs of the derived game are also provided in this section. Experimental results and analysis are carried out in Section V. Conclusions, and future works are drawn in Section VI.

II. Related Works

Fu et al. studied distributed downlink power control for the NOMA system with two interfering cells [24]. The authors formulated the distributed downlink power control mathematically as an optimization problem that aimed to minimize the total transmit power of the two BSs. Similarly, Sung et al. investigated game theoretic analysis of uplink power control with two interfering cells for the uplink NOMA systems [25]. Furthermore, a game-theoretic approach is studied in [26] where NOMA is applied to ALOHA for deciding the transmission probability. Based on the Glicksberg game, Aldebes et al. proposed a power allocation algorithm for cellular downlink NOMA networks [27]. In particular, for the power allocation algorithm, the authors proposed a price-based user's utility function, which is shown to be restrictive if the allocated power beyond a threshold value causes a decrease in the utility value. In [28], a joint utility-based power control via S-modular theory in multi-service wireless networks is addressed. A RL-based power control scheme for downlink NOMA in the presence of smart jamming is studied in [29], where the authors formulated a Stackelberg equilibrium of the antijamming NOMA transmission game. A power control based on evolutionary game theory for uplink NOMA systems is examined in [30]. A power allocation based on optimization and deep reinforcement learning approach for cache-aided NOMA systems is proposed in [31]. Moreover for a hybrid NOMA systems, a joint channel selection and power control based on game theory is proposed in [32]. Although a lot of work for power allocation for NOMA and wireless networks based on game theory has been carried out in the literature, some interesting questions still remain to be answered. How to optimize distributed power control, especially for multicell NOMA networks where multiple BSs compete with each other based on the fairness of achievable data rate among its users so as to achieve overall system fairness for the downlink NOMA systems. Therefore, in this paper, we propose distributed power control in a multicell downlink NOMA system based on the joint application of RL and game theory.

III. System Model

We consider a downlink NOMA scenario with multiple base stations (BSs) located geographically in close vicinity so that they might cause interference on each other, namely BS_1, BS_2 $\cdots$ BS_N where each BS is serving two User Equipments (UEs) - its Near User (NU) and Far User (FU). Please note that our work generalises for more than two UEs in a straightforward manner but for the sake of simplicity we content ourselves to two UEs per BS[1].

Due to data transmission to their respective UEs by their serving BSs, each of the BSs induces an external interference factor to the UEs. The nodes are assumed to be operating in

[1]Two UEs per frequency band has been already adopted as a standard by the Third Generation Partnership Project (3GPP) Long Term Evolution (LTE) [33] under the name of Multi-User Superposition Transmission (MUST). Please note that a BS might have a certain number of frequency bands, and thus, more pairs of users can be served in different frequency bands.

the half-duplex mode. We have assumed that Channel State Information (CSI) is perfectly known to the receiver. Each of the nodes is equipped with a single antenna. The channel between any two nodes is subjected to the independent Rayleigh block fading plus additive white Gaussian noise in which the channel remains constant during the transmission of a block and varies independently from one block to another. $h_{(2j-1)} \sim CN(0, \lambda_{h(2j-1)})$ is the complex channel co-efficient between near UEs and BS j with zero mean and variance $\lambda_{h(2j-1)}$, where $j = 1, 2, \cdots N$.

Furthermore, $h_{2j} \sim CN(0, \lambda_{h_{2j}})$ is the complex channel co-efficient between BS_j and far UEs node with zero mean and variance $\lambda_{h_{2j}}$, where $j = 1, 2, \cdots N$. $\hat{h}_{(2\hat{j}-1)} \sim CN(0, \lambda_{\hat{h}(2\hat{j}-1)})$ is the complex channel co-efficient between BS j and near UEs associated with BSs other than j BS with zero mean and variance $\lambda_{\hat{h}(2\hat{j}-1)}$, where $\hat{j} = 1, 2, \cdots N$ and $\hat{j} \neq j$. Similarly, $\hat{h}_{(2\hat{j})} \sim CN(0, \lambda_{\hat{h}(2\hat{j})})$ is the complex channel co-efficient between BS j and far UEs associated with BSs other than j BS with zero mean and variance $\lambda_{\hat{h}(2\hat{j})}$.

A. Signal-to-Interference Noise Ratios (SINR)

Let y_j denote the power level of BS_j. BSs are using NOMA to transmit the data to their respective UEs, i.e., NU and FU. Since, UE_{2j-1} is a near user and UE_{2j} is a far user for BS_j, BS_j allocates $\alpha_{2j-1}y_j$ power for transmitting the information to UE_{2j-1} and $(1-\alpha_{2j})y_j$ power for transmitting the information to UE_{2j}.

Following the downlink NOMA protocol, the BS_j transmits a superimposed composite signal Z_j which consists of UE_{2j-1} information x_{2j-1} and UE_{2j} information x_{2j}. The superimposed composite signal Z_j from the BS_j, following the downlink NOMA protocol can be given as:

$$Z_j = \sqrt{\alpha_{2j-1}y_j}x_{2j-1} + \sqrt{(1-\alpha_{2j})y_j}x_{2j} \tag{1}$$

The received SINR at $BS_j \rightarrow UE_{2j-1}$ link is given by:

$$\gamma_{UE_j \rightarrow x_{2j}} = \frac{(1-\alpha_{2j})y_j|h_{2j-1}|^2}{\Big(\alpha_{2j}y_j|h_{2j-1}|^2 + \sum_{n=1,n\neq j}^{N}(\alpha_{2n-1}y_n |\hat{h}_{2n-1}|^2 + \alpha_{2n}y_n|\hat{h}_{2n}|^2) + N_0\Big)} \tag{2}$$

$$\gamma_{UE_j \rightarrow x_{2j-1}} = \frac{\alpha_{2j-1}y_j|h_{2j-1}|^2}{\Big(\sum_{n=1,n\neq j}^{N}(\alpha_{2n-1}y_n|\hat{h}_{2n-1}|^2 + \alpha_{2n}y_n|\hat{h}_{2n}|^2) + N_0\Big)} \tag{3}$$

where $\gamma_{UE_j \rightarrow x_{2j}}$ is the SINR required at UE_i to decode and cancel x_{2j}, i.e. to perform SIC at UE_i.

The received SINR at $BS_j \rightarrow UE_{2j}$ link is given by:

$$\gamma_{UE_{2j} \rightarrow x_{2j}} = \frac{(1-\alpha_{2j})y_j|h_{2j}|^2}{\Big(\alpha_{2j-1}y_j|h_{2j}|^2 + \sum_{n=1,n\neq j}^{N}(\alpha_{2n-1}y_n |\hat{h}_{2n-1}|^2 + \alpha_{2n}y_n|\hat{h}_{2n}|^2) + N_0\Big)} \tag{4}$$

B. Achievable Data Rate

According to our system model, the achievable data rate associated with the far user UE_{2j-1} is given by:

$$\begin{aligned} R_{UE_{2j-1}} &= B\log_2(1 + \gamma_{UE_{2j-1} \rightarrow x_{2j-1}}) \\ &= B\log_2\left(1 + \frac{\alpha_{2j-1}y_j|h_{2j-1}|^2}{\Big(\sum_{n=1,n\neq j}^{N}(\alpha_{2n-1}y_n|\hat{h}_{2n-1}|^2 + \alpha_{2n}y_n|\hat{h}_{2n}|^2) + N_0\Big)}\right) \end{aligned} \tag{5}$$

Simiarly, the achievable data rate associated with the near user UE_{2j} is given by:

$$\begin{aligned} R_{UE_{2j}} &= B\log_2(1 + \gamma_{UE_{2j} \rightarrow x_{2j}}) \\ &= B\log_2\left(1 + \frac{(1-\alpha_{2j})y_j|h_{2j}|^2}{\Big(\alpha_{2j-1}y_j|h_{2j}|^2 + \sum_{n=1,n\neq j}^{N}(\alpha_{2n-1}y_n |\hat{h}_{2n-1}|^2 + \alpha_{2n}y_n|\hat{h}_{2n}|^2) + N_0\Big)}\right) \end{aligned} \tag{6}$$

Let y_{max} be the maximum transmission power and y_{min} be the minimum transmission power for a BS. Since the actual transmission power may usually only be set to a finite number of levels, the power is discretized into a finite number of levels, i.e. between y_{min} and y_{max}. Since, y_j is the power of BS_j, we have $y_{min} \leq y_j \leq y_{max}$.

Our aim is to optimize the fairness of the whole system [34]. The players here correspond to BSs. Thus, the fairness criterion for each BS depends on its total achievable data rate by its users in the presence of interference from other BSs.

Assuming the total transmission power of the BS is limited to y_{max}, the maximization problem could be formulated as:

$$\begin{aligned} &\textbf{\textit{Maximize:}} \\ &R_{Sum} = \sum_{j=1}^{N}\log_2(R_{UE_{2j-1}} + R_{UE_{2j}}) \\ &\textit{Subject to} \\ &y_j \leq y_{max} \\ &\forall y_j \geq y_{min}. \end{aligned} \tag{7}$$

In the above Equation 7, the $\log_2(R_{UE_{2j-1}} + R_{UE_{2j}})$ captures the achievable data rate of NU and FU associated with BS j according to the criterion of achievable data rate fairness [34].

IV. Reinforcement Learning based Game Theoretic Approach

We now formulate the distributed power control as game G in which each player (here BS) optimizes its power allocation so as to maximize its individual fairness. Let N denote the number of players or base stations in our case. The payoff of user j obtained from using power level y_j is denoted by

$$u_j(y_j, y_{-j}) \tag{8}$$

Note that u_j is a function of the strategy chosen by player j, and of y_{-j}, its opponents' current strategic profile. Players

will selfishly pick actions that improve their utility functions, taking into account the other players' current strategies. The main problem, therefore, is the choice of u_j, so that the players' individual actions lead to a theoretically optimal result.

Furthermore, the players are assumed to be selfish in the sense that they will try to maximise their own utility. They will compete to maximize their own utility functions, given the most recent action of the other players, then the process will converge to a Nash Equilibrium (NE) regardless of the order of play. To compensate for the selfishness of players, the utility function can be defined as:

$$u_j(y_j, y_{-j}) = C(y_j, y_{-j}) - \sum_{i=1, i\neq j}^{N} \big(C_{-j}(y_i) - C(y_i, y_{-i}, y_j)\big) \tag{9}$$

The above utility reflects the idea of marginal price costing where the player should pay a tax offsetting the nuisance his presence causes for the other players [35].

In the above Equation 9, $C_j(y_j, y_{-j}) = \log_2(R_{UE_{2j-1}} + R_{UE_{2j}})$ is the fairness of the achievable rate associated with player j, i.e. sum of the achievable rate fairness of NU and FU associated with BS j. $C{-}j(y_i)$ denotes the achievable rate fairness of the opponent player i other than j when it is not affected by the presence of player j. The second term in the above expression signifies the increase in the achievable rate fairness of the neighboring BSs i, $i \neq j$ if BS j was not present and causing interference. Hence, it is calculated as the difference between these achievable rate fairnesses of the different BSs i without and with the presence of BS i.

We now formulate a decentralized stochastic RL algorithm to evolve to the Nash Equilibrium (NE) of game G. NE defines the stability of the game that occurs when players behave according to correspondence from their best response (BR) in the game [36], [37]. The best response of the player j can be defined as:

Definition 1: action $a_j^* \in BR(a_{-j})$ if:

$$u_j(a_j^*, a_{-j}) \geq u_j(a_j, a_{-j}); \forall a_j \tag{10}$$

The informed reader observes that an action for player j in our power model is rather denoted y_j but here for the sake of presenting a definition, we denoted it by a_j since it is a common nomenclature in game theory when presenting definitions. When it comes to potential games, it is known in the literature that if the players adopt sequentially their BR strategies that a NE will be reached which corresponds to the maximizer of the potential function [38]. In our game, the choices of the players are distributed and without central control, thus, we rather adopt RL as a choice mechanism. The above definition implies that the players will not change their actions, which coincides with their BR if no other players in the game have an incentive to deviate from their action. That is, the game has reached a stable state, i.e. to the NE.

We now expand the game G to a mixed strategy for characterizing the learning algorithm. Let $P_j(t) = [p_{j,1}(t), p_{j,2}(t), \cdots, p_{j,k}(t)]$ be the mixed strategy of the player j, where $p_{j,k}$ denotes the probability with which the jth player chooses the kth pure strategy at instant t. We suppose that each BS has K power levels to choose among. Let $L = \{l_1, l_2, \cdots, l_K\}$ denote those power levels. Therefore $y_j L$. Each player is represented by an LA, and the actions of the automaton are the pure strategies of the player. The normalized payoff of the jth player will be a feedback $u_j(t)$ to the jth automaton based on his action and the action adopted by the rest of the players. At each iteration, each player samples individually an action according to its current action probability vector. Repeatedly the game is played for learning the NE until convergence of their corresponding action probability vectors, The LA-based distributed power control game in downlink NOMA is given in Algorithm 1.

Algorithm 1 LA based Distributed Power Control Game in Downlink NOMA

1: Start
2: Set the components of initial probability vector for player j, $P_j(0)$ as $p_{j,k}(0) = \frac{1}{K}$, $j = 1, 2, \cdots, N$, and $k = 1, 2, \cdots, K$.
3: For every time instant t, each of the players chooses j an action $a_j(t)$ according to its action probability vector $P_j(t)$.
4: The selected player j obtains $u_j(t)$ based on the set of all actions of the players in the game G. Let $r_j(t)$ be the normalized version of $u_j(t)$ using Equation 11.
5: Each player j updates it action probability through the following rule:
$p_{j,k}(t+1) = p_{j,k}(t) + \lambda r_j(t)(1 - p_{,jk}(t)), \quad k = a_j(t)$
$p_{j,k}(t+1) = p_{j,k}(t) - \lambda r_j(t) p_{j,k}(t), \quad k \neq a_j(t)$
where λ is a learning parameter.
6: If there is a component of $P_j(t)$ that is larger than 0.99, then set it to one and set the rest of components to zero and stop. If not, go to Step 3.
7: Stop

Since the utility $u_j(t)$ might take values outside the interval $[0, 1]$, it is common in the field of LA to use a normalized version of the feedback according to the following normalization procedure that is described by [39] and which is given by:

$$r_j = \frac{u_j - \min_j u_j}{\max_j u_j - \min_j u_j} \tag{11}$$

A. Theoretical Results

In game theory, a game is considered to be a potential game if a single global function called the potential function can be used to represent the incentives of all players to adjust their strategy. We will now prove that our proposed game is an exact potential game.

Theorem 1. *The game defined by the utility function given in Equation 9 is an exact potential game.*

Proof:

Let us define the following potential function:

$\Phi(y_j, y_{-j}) = \sum_{i=1}^{N} C(y_i, y_{-i}, y_j)$

We suppose that the j player changes its strategy from y_j to $\tilde{y}_j$.

$$u_j(y_j, y_{-j}) = C(y_j, y_{-j}) - \sum_{i=1, i\neq j}^{N} \big(C_{-j}(y_i) - C(y_i, y_{-i}, y_j)\big) \tag{12}$$

$$u_j(\tilde{y}_j, y_{-j}) = C(\tilde{y}_j, y_{-j}) - \sum_{i=1, i\neq j}^{N} \big(C_{-j}(y_i) - C(y_i, y_{-i}, \tilde{y}_j)\big) \tag{13}$$

We will show that:

$\Phi(y_j, y_{-j}) - \Phi(\tilde{y}_j, y_{-j}) = u_j(y_j, y_{-j}) - u_j(\tilde{y}_j, y_{-j})$.

Indeed,

$$\begin{aligned}
&u_j(y_j, y_{-j}) - u_j(\tilde{y}_j, y_{-j}) : \\
&= C_(y_j, y_{-j}) - \sum_{i=1, i\neq j}^{N} \big(C_{-j}(y_i) - C(y_i, y_{-i}, y_j)\big) - \\
&C_(\tilde{y}_j, y_{-j}) + \sum_{i=1, i\neq j}^{N} \big(C_{-j}(y_i) - C(y_i, y_{-i}, \tilde{y}_j)\big) \\
&= C(y_j, y_{-j}) + \sum_{i=1, i\neq j}^{N} C(y_i, y_{-i}, y_j) \\
&- C_(\tilde{y}_j, y_{-j}) - \sum_{i=1, i\neq j}^{N} C(y_i, y_{-i}, \tilde{y}_j) \\
&= \sum_{i=1}^{N} C(y_i, y_{-i}, y_j) - \sum_{i=1}^{N} C(y_i, y_{-i}, \tilde{y}_j) \\
&= \Phi(y_j, y_{-j}) - \Phi(y_j, \tilde{y}_j)
\end{aligned}$$

This completes the proof of Theorem 1.

B. Theoretical Treatment of the LA Scheme

We denote

$$\mathbb{E}[r_j | \boldsymbol{P}, y_j = l_k] \tag{14}$$

as the expected normalized reward of the base j if it employs its kth pure strategy while the rest of the base stations employ the mixed strategy $\boldsymbol{P}$ where $\boldsymbol{P}$ is the selection probability vector [2].

Theorem 2. *For sufficiently small λ, the selection probability matrix of the different power levels by the different base stations can be approximately characterized by the following Ordinary Differential Equation (ODE).*

$$\begin{aligned}
\frac{dp_{j,k}}{dt} =& p_{j,k} \Bigg(\sum_{k'} p_{j,k'} (\mathbb{E}[r_j | \boldsymbol{P}, y_j = l_k] - \mathbb{E}[r_j | \boldsymbol{P}, \\
& y_j = l_{k'}]) \Bigg)
\end{aligned} \tag{15}$$

[2]Please note here we use an abuse of notation as we could have used $\boldsymbol{P}_{-j}$ to denote the mixed strategy of the rest of base stations except j in the same line as the nomenclature used for action profile.

Proof:

$$\begin{aligned}
\frac{dp_{j,k}}{dt} =& p_{j,k}(1 - p_{j,k}) \mathbb{E}[r_n | \boldsymbol{P}_{-n}, y_j = l_k] + \sum_{k' \neq k} p_{j,k'} \\
& (-p_{j,k}) \mathbb{E}[r_j | \boldsymbol{P}, y_j = l_{k'}] \\
=& p_{j,k} \sum_{k' \neq k} p_{j,k'} \mathbb{E}[r_j | \boldsymbol{P}, y_j = l_k] - p_{j,k} \\
& \sum_{k' \neq k} p_{j,k'} \mathbb{E}[r_j | \boldsymbol{P}, y_j = l_{k'}] \\
=& p_{j,k} \Bigg(\sum_{k' \neq k} p_{j,k'} . (\mathbb{E}[r_j | \boldsymbol{P}, y_j = l_k] - \mathbb{E}[r_j \\
& | \boldsymbol{P}, y_j = l_{k'}]) \Bigg) \\
=& p_{j,k} \Bigg(\sum_{k'} p_{j,k'} (\mathbb{E}[r_j | \boldsymbol{P}, l_k] - \mathbb{E}[r_j | \boldsymbol{P}, \\
& y_j = l_{k'}]) \Bigg) \quad \forall j, k
\end{aligned}$$

This completes the proof of Theorem 2.

Theorem 3. *The following statements are true about the Ordinary Differential Equation (ODE) obtained from the learning algorithm*

- *All the stable stationary points of the ODE are Nash equilibria.*
- *All Nash equilibria are the stationary points of the ODE.*

Proof:

The proof of above Theorem 3 can be found in [15]. Hence the proof is omitted for the sake of brevity.

Theorem 4. *With a sufficiently small step-size λ, our adaptive power algorithm converges to a stable stationary point of the ODE given in Equation 15.*

Proof:

The proof is given in Appendix A.

V. Experimental Results

The simulation parameters are given in Table I. We run the Monte-Carlo simulation by averaging over 10^5 random realizations of Rayleigh block fading channels between BSs and UEs. For simplicity, we have assumed that there are two BSs (players) in the system and each of the BS is serving its associated NUs and FUs as shown in Fig. 1. Also, we have taken bandwidth $B = 1$ and $N_0 = 1$ for all our experiments.

Each BS possesses K discrete power level. Let l_1 be the lowest power level called y_{min} and l_K be the highest power level called y_{max}. We suppose that the power levels are

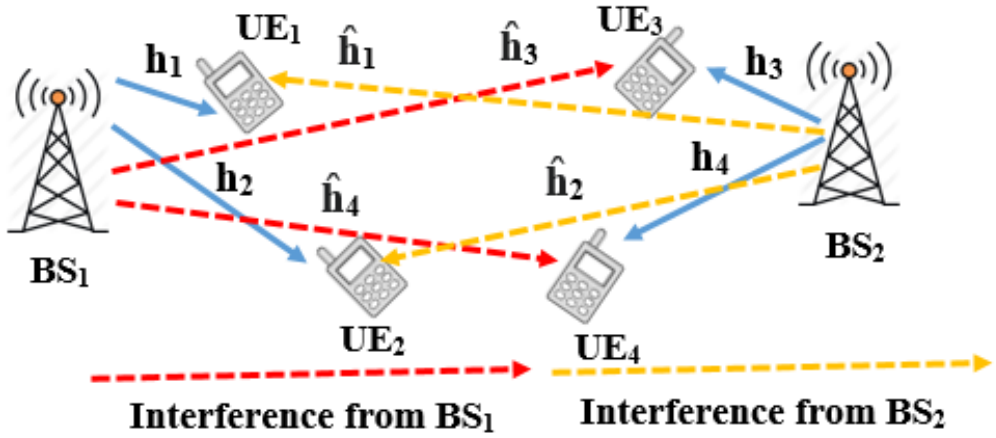

Fig. 1. Considered System Model Scenario for Experiments

TABLE I
SIMULATION PARAMETERS

Parameter	Symbol	Values
Mean of variance between BS_1 and UE_1	v_1	3.0
Mean of variance between BS_1 and UE_2	v_2	2.0
Mean of variance between BS_2 and UE_3	v_3	4.0
Mean of variance between BS_2 and UE_4	v_4	3.0
Mean of variance between BS_1 and UE_3	v_5	0.3
Mean of variance between BS_1 and UE_4	v_6	0.5
Mean of variance between BS_2 and UE_1	v_7	0.6
Mean of variance between BS_2 and UE_2	v_8	0.8
Path Loss Factor	v	4
Power of BSs - BS_1 and BS_2	y	50-500
Power Allocation Factor for NOMA	α	0.2

increasing. Let β determine the non-linearity of the discretization [40]. The kth discrete power level is given by:

$$l_k = y_{min} + \frac{k^\beta}{(K-1)^\beta}(y_{max} - y_{min}) \tag{16}$$

where $k = 1, 2, \cdots K$. Also, $\beta = 1$ for linear discretization and $\beta > 1$ for non-linear discretization. For all our experiments, we use linear discretization.

In Fig. 2, we plot the fairness of the system where two players, i.e. $P1$ and $P2$, has 25 discrete power levels each. We observe that for most of the cases, when both $P1$ and $P2$ power levels are high, the achievable sum rate fairness of the system is low. It means that when both BSs use maximum power for the data transmission for their associated NOMA UEs, they affect adversely each other achievable rate performance, and as a result, the fairness of the system decreases. Also, one can observe that when one of the BS transmits at a power level $P1 = 312$ and the other BS transmits at a low power level such as $P2 = 200$, then the fairness achieves the optimum value of the system. This indicates that both BSs

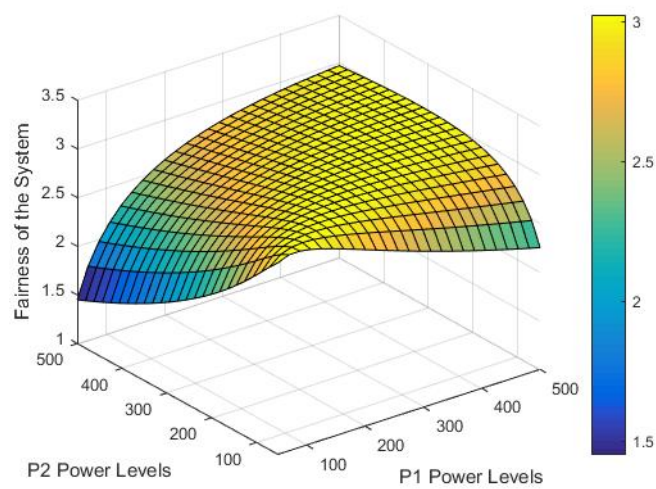

Fig. 2. Fairness of the System

cannot transmit at higher power levels at the same time as both of their achievable rate is severely affected by each other presence.

In Table II, we present our findings for the overall fairness of the system with different power levels at learning parameter $\lambda = 0.1$. Also, to compensate for the randomness of the probabilities in our experiments, we run all our experiments for 100 number of times and report the average performance of the system together with 95% confidence interval in Table III. From Table II and Table III, we can observe that, for learning parameter $\lambda = 0.1$, as we increase the power levels from 3 to 9 to 27, the average fairness of the system increases and the average iteration for convergence also increases. The best

TABLE II
FAIRNESS OF THE SYSTEM WHEN LEARNING PARAMETER $\lambda = 0.1$

Power levels K	Learning Parameter λ	Average Iteration	Average Fairness	95% CI Lower Range	95% CI Upper Range
3	0.1	244	2.9632	2.9467	2.9798
9	0.1	259	2.9770	2.9647	2.9893
27	0.1	276	2.9846	2.9740	2.9952

TABLE III
ITERATIONS CORRESPONDING TO LEARNING PARAMETER $\lambda = 0.1$

Power levels K	Learning Parameter λ	Average Iteration	95% CI Lower Range	95% CI Upper Range
3	0.1	244	216	272
9	0.1	259	234	283
27	0.1	276	252	300

average fairness of the system 2.9846 is achieved when there are 27 power levels which converges at an average iteration of 276. It should be noted that even with 27 different power levels for each of the player, the average iteration for convergence is 276, which is significant compared to having just three power levels where the average iteration is 244.

Similarly, in Table IV, we present our findings for the overall fairness of the system with different power levels at learning parameter $\lambda = 0.01$. In Table V, we report the average performance of the system together with 95% confidence interval at $\lambda = 0.01$. From Table IV, we observe that as we reduce the learning parameter λ from 0.1 to 0.01, the average number of iterations increases significantly. Although we can see higher fairness of the system, this comes at the cost of convergence time, which is a trade-off factor. The best average fairness 3.0186 is achieved when there are 27 power levels, which converges at an average iteration of 24634. Unlike Table II and Table III, one can observe a significantly higher number

TABLE IV
FAIRNESS OF THE SYSTEM WHEN LEARNING PARAMETER $\lambda = 0.01$

Power levels K	Learning Parameter λ	Average Iteration	Average Fairness	95% CI Lower Range	95% CI Upper Range
3	0.01	10384	3.0152	3.0127	3.0177
9	0.01	19894	3.0185	3.0172	3.0198
27	0.01	24634	3.0186	3.0173	3.0199

TABLE V
ITERATIONS CORRESPONDING TO LEARNING PARAMETER $\lambda = 0.01$

Power levels K	Learning Parameter λ	Average Iteration	95% CI Lower Range	95% CI Upper Range
3	0.01	10384	9062	11705
9	0.01	19894	18212	21576
27	0.01	24634	21767	27501

TABLE VI
COMPARISON OF FAIRNESS OF THE SYSTEM

Learning Parameter λ	Power Levels K	Random Method	Exhaustive Method	LA-GT Method
0.1	3	2.5648	3.0195	2.9632
0.1	9	2.7537	3.0248	2.9770
0.1	27	2.8009	3.0249	2.9846

of iterations in Table IV and V for the convergence as we increase the power levels of both of the players from 3 to 27. This is because the algorithm is learning at a lower rate, i.e. $\lambda = 0.01$. Also, it has to choose from the combination of power levels which is 27×27 when both $P1$ and $P2$ has 27 power levels each.

In Table VI, we present the comparison of the fairness of the system through our LA based Game-Theoretic (LA-GT) approach with the random and exhaustive search method. Unlike our proposed method, in a random method, the players choose the actions randomly with equal probability and no active learning parameters in each iteration. We can see that, at different power levels, the fairness of the LA-GT approach is higher compared to the random method. This signifies the importance of having LA in combination with game theory to improve the fairness of the system by distributed power control over a range of players with different power levels in the system. Also, we can observe that fairness of the LA-GT approach is quite competitive compared to exhaustive search method. It should be noted that exhaustive search method is the heuristics method which finds the best solution by including all power levels. Although exhaustive method gives the best fairness of the system, it is not desirable as it is usually centralized and require large message exchanges between BSs and users. Hence, energy-efficiency of the system cannot be achieved through exhaustive search method. Our LA-GT converges much faster at just an average iteration of 244 to achieve 2.9846 average fairness of the system which is quite competitive compared to exhaustive search method and better than random method.

In Fig. 3 and Fig. 4, we plot the evolution of the action probabilities of player 1 and player 2 with three power levels, respectively, with learning parameter $\lambda = 0.1$. It should be noted that the power selection probability vector evolves from $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ to $(0, 0, 1)$ for player 1 and $(0, 1, 0)$ for player 2. We observe that for both players, only one action probability of the power level converges to 1, which corresponds to the NE of the game G in this case. With the increase in number of iterations over time, the other two action probabilities for both player 1 and player 2 reduces to zero. Also, it should be noted that with less number of power levels, i.e. 3 for both players, our LA based game-theoretic approach converges much faster around 250 number of iterations.

Similarly, in Fig. 5 and Fig. 6, we plot the evolution of the action probabilities of player 1 and player 2, respectively, with more number of power levels, i.e. 6 with learning parameter $\lambda = 0.1$. It should be noted that the power

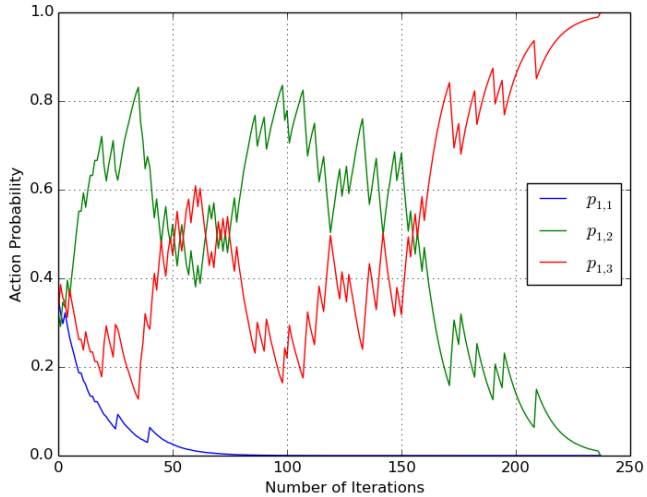

Fig. 3. Evolution of Action Probabilities of Player 1 with 3 Power Levels

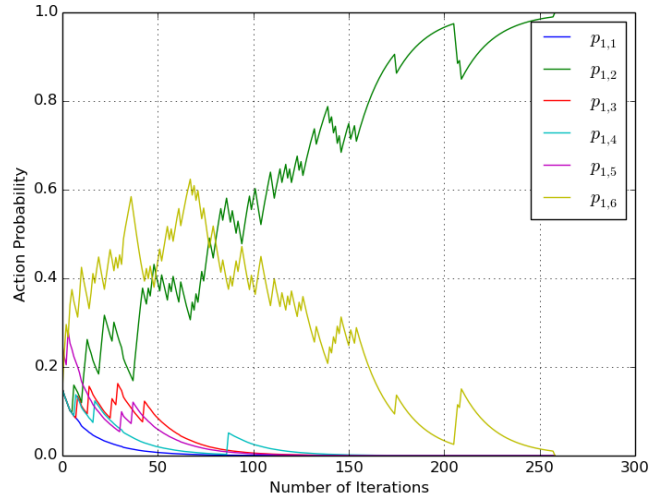

Fig. 5. Evolution of Action Probabilities of Player 1 with 6 Power Levels

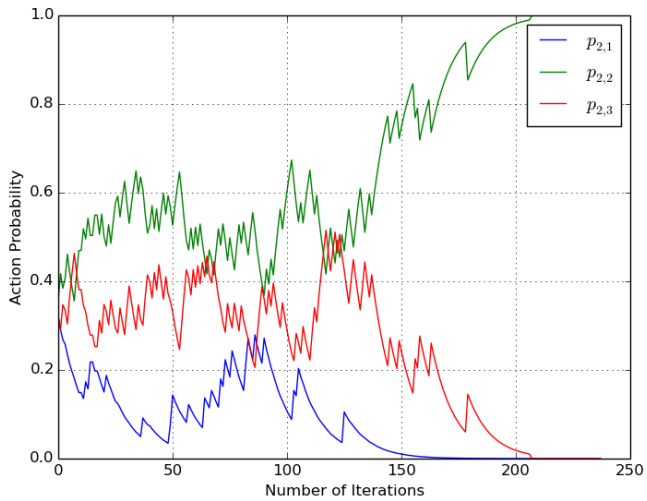

Fig. 4. Evolution of Action Probabilities of Player 2 with 3 Power Levels

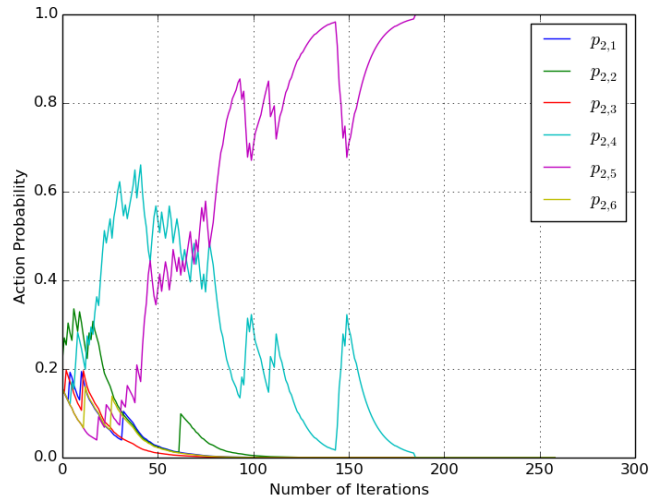

Fig. 6. Evolution of Action Probabilities of Player 2 with 6 Power Levels

selection probability vector evolves from $(\frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6}, \frac{1}{6})$ to $(0, 1, 0, 0, 0, 0)$ for player 1 and $(0, 0, 0, 0, 1, 0)$ for player 2 in this case. We observe that as the number of power levels increases, the convergence time increases for the action probabilities for both player 1 and player 2, which is expected as the players have to choose the power level from the combination of different power levels that will eventually maximize their utility function. Although having 6 different power levels, the convergence time of the evolution of one of the action probabilities for both player 1 and player 2 for our LA based game-theoretic approach is considerably faster, which is around 300 iterations.

VI. Conclusion and Future Work

In this paper, we formulated a RL-based game theoretic approach for distributed power control in multicell downlink NOMA systems. First, we designed the distributed power control in multicell downlink NOMA systems as a game. Next, for the considered game scenario, a utility function was also designed where the players used RL that provided full characterization of the best achievable rate fairness performance of the system. We showed that our proposed distributed power control algorithm is guaranteed to converge to an NE. We also proved that the distributed power control game we designed is an exact potential game. Numerical results demonstrated that our proposed RL based game-theoretic approach converges much faster at an average iteration of few hundreds for the

power levels of the BSs. We also demonstrated the correctness of the proposal followed by a thorough comparison with random and heuristic approaches.

In this work, we only considered discrete power level that used linear discretization for downlink NOMA systems where the power levels are equi-spaced. Nevertheless, for future work, we would like to extend our model and study the scenario with non-linear discretization power levels for uplink and downlink NOMA systems. Studying and designing a game for distributed power control in multicell NOMA systems in the presence of smart jammers and eavesdroppers is also an interesting research direction for future work.

Appendix A
Proof of Theorem 4

Proof. As defined earlier in the paper $\mathbb{E}[r_j|\boldsymbol{P}, y_j = l_k]$ denotes the expected normalized reward of the BS j given that it employs its kth pure strategy while the rest of the base stations employ the mixed strategy $\boldsymbol{P}$.

$$\mathbb{E}[r_j|\boldsymbol{P}, y_j = l_k] = \sum_{(y_1,\cdots,y_{j-1},y_{j+1},\cdots,y_N)} r_j(y_j = l_k, \boldsymbol{y}_{-j}) \prod_{j',j'\neq j} p_{j'y_{j'}}$$

In addition, we define the probabilistic potential function

$$F(\mathbf{P}) = \Phi(\mathbf{P}) = \sum_{\boldsymbol{y}} \Phi(\boldsymbol{y}) \prod_j p_{j,y_j}, \tag{17}$$

We compute term by term:

$$\frac{dF(\boldsymbol{P})}{dt} = \sum_{j,k} \frac{\partial F(\boldsymbol{P})}{\partial p_{j,k}} \frac{dp_{j,k}}{dt} \tag{18}$$

Using Equation 17 we obtain:

$$\frac{\partial F(\boldsymbol{P})}{\partial p_{j,k}} = \sum_{y_1,\cdots,y_{j-1},y_{j+1},\cdots,y_N} \Phi(y_j = l_k, \boldsymbol{y}_{-j}) \prod_{j',j'\neq j} p_{j'y_{j'}} \quad \forall j, \forall k$$

The above formula can be written as

$$\frac{\partial F(\boldsymbol{P})}{\partial p_{j,k}} = \mathbb{E}[\Phi(k, \boldsymbol{y}_{-j})|\boldsymbol{P}, y_j = l_k]$$

Now,

$$\frac{dF(\boldsymbol{P})}{dt} = \sum_{j,k} \mathbb{E}[\Phi(k, \boldsymbol{y}_{-j})|\boldsymbol{P}, y_j = l_k] p_{j,k} \left(\sum_{k'} p_{j,k'} (\mathbb{E}[r_j|\boldsymbol{P}, y_j = l_k] - \mathbb{E}[r_j|\boldsymbol{P},\ y_j = l_{k'}]) \right)$$

Where $\Phi(k, \boldsymbol{y}_{-j})$ denotes $\Phi(y_j = l_k, \boldsymbol{y}_{-j})$ for the sake of simplifying the notation.

Here,

$$\frac{dF(\boldsymbol{P})}{dt} = \sum_{j,k,k'} \mathbb{E}[\Phi(k, \boldsymbol{y}_{-j})|\boldsymbol{P}, y_j = l_k] p_{j,k} p_{j,k'} (\mathbb{E}[r_j|\boldsymbol{P}, y_j = l_k] - \mathbb{E}[r_j|\boldsymbol{P}, y_j = k'])$$

However we note that,

$$\sum_{j,k,k'} \mathbb{E}[\Phi(m, \boldsymbol{y}_{-j})|\boldsymbol{P}, y_j = l_k] p_{j,k} p_{j,k'} (\mathbb{E}[r_j|\boldsymbol{P}, y_j = l_k] - \mathbb{E}[r_j|\boldsymbol{P}, .y_j = l_{k'}])$$
$$= \sum_{j,k,k'} \mathbb{E}[\Phi(k', \boldsymbol{y}_{-j})|\boldsymbol{P}, y_j = l_k] p_{j,k'} p_{j,k} (\mathbb{E}[r_j|\boldsymbol{P}, y_j = l_{k'}] - \mathbb{E}[r_j|\boldsymbol{P}, .y_j = l_k])$$

Therefore,

$$\frac{dF(\boldsymbol{P})}{dt} = \frac{1}{2} \sum_{j,k,k'} p_{j,k} p_{j,k'} (\mathbb{E}[\Phi(k, \boldsymbol{y}_{-j})|\boldsymbol{P}, y_j = k] - \mathbb{E}[\Phi(k', \boldsymbol{y}_{-j})|\boldsymbol{P}, y_j = k'])(\mathbb{E}[r_j|\boldsymbol{P}, y_j = k] - \mathbb{E}[r_j|\boldsymbol{P}, y_j = k'])$$

By exploiting the fact that the game is an exact potential game, we can obtain:

$$\mathbb{E}[\Phi(k, \boldsymbol{y}_{-j})|\boldsymbol{P}, y_j = k] - \mathbb{E}[\Phi(k', \boldsymbol{y}_{-j})|\boldsymbol{P}, y_j = k'] = \mathbb{E}[r_j|\boldsymbol{P}, .y_j = k'] - \mathbb{E}[r_j|\boldsymbol{P}, y_j = k]$$

Therefore,

$$\frac{dF(\boldsymbol{P})}{dt} = \frac{1}{2} \sum_{j,k,k'} p_{j,k} p_{j,k'} (\mathbb{E}[\Phi(k, \boldsymbol{y}_{-j})|\boldsymbol{P}, y_j = k] - \mathbb{E}[\Phi(k', \boldsymbol{y}_{-j})|\boldsymbol{P}, y_j = k'])^2$$

We have $\frac{dF(\boldsymbol{P})}{dt} \geq 0$ which implies that $F(\boldsymbol{P})$ is increasing. Since, according to Equation 17, $F(\boldsymbol{P})$ is upper-bounded, therefore $F(\boldsymbol{P})$ will converge to a maximum point characterized by:

$$\frac{dF(\boldsymbol{P})}{dt} = 0$$

$$\Rightarrow \mathbb{E}[r_j|\boldsymbol{P}, y_j = k] - \mathbb{E}[r_j|\boldsymbol{P}, y_j = k'] = 0, \forall j, k, k'$$

$$\Rightarrow \frac{dp_{j,k}}{dt} = 0, \forall j, k$$

$$\Rightarrow \frac{d\boldsymbol{P}}{dt} = 0$$

The last equation shows that $\boldsymbol{P}$ eventually converges to the stationary point of ODE.

This completes the proof of Theorem 4. □

REFERENCES

[1] D. Reinsel, J. Gantz, and J. Rydning, "Data age 2025: The digitization of the world from edge to core," *Seagate https://www.seagate.com/files/www-content/our-story/trends/files/idc-seagate-dataage-whitepaper.pdf*, 2018.

[2] L. Lei, D. Yuan, C. K. Ho, and S. Sun, "Power and channel allocation for non-orthogonal multiple access in 5g systems: Tractability and computation," *IEEE Transactions on Wireless Communications*, vol. 15, no. 12, pp. 8580–8594, 2016.

[3] J. Howell, "Number of connected iot devices will surge to 125 billion by 2030, ihs markit says," *URL https://technology. ihs. com/596542/number-of-connected-iot-devices-will-surge-to-125-billion-by-2030-ihs-markit-says*, 2017.

[4] Z. Wu, K. Lu, C. Jiang, and X. Shao, "Comprehensive study and comparison on 5g noma schemes," *IEEE Access*, vol. 6, pp. 18 511–18 519, 2018.

[5] A. Rauniyar, P. Engelstad, and O. N. Østerbø, "On the performance of bidirectional noma-swipt enabled iot relay networks," *IEEE Sensors Journal*, 2020.

[6] M. B. Shahab, R. Abbas, M. Shirvanimoghaddam, and S. J. Johnson, "Grant-free non-orthogonal multiple access for iot: A survey," *IEEE Communications Surveys & Tutorials*, 2020.

[7] C.-H. Liu and D.-C. Liang, "Heterogeneous networks with power-domain noma: Coverage, throughput, and power allocation analysis," *IEEE Transactions on Wireless Communications*, vol. 17, no. 5, pp. 3524–3539, 2018.

[8] X. Su, A. Castiglione, C. Esposito, and C. Choi, "Power domain noma to support group communication in public safety networks," *Future Generation Computer Systems*, vol. 84, pp. 228–238, 2018.

[9] M. L. Merani, "Recovery failure probability of power-based noma on the uplink of a 5g cell for an arbitrary number of superimposed signals," in *2018 IEEE International Conference on Communications (ICC)*. IEEE, 2018, pp. 1–6.

[10] Z. Ding, Z. Yang, P. Fan, and H. V. Poor, "On the performance of non-orthogonal multiple access in 5g systems with randomly deployed users," *IEEE signal processing letters*, vol. 21, no. 12, pp. 1501–1505, 2014.

[11] Y. Liu, Z. Qin, M. Elkashlan, A. Nallanathan, and J. A. McCann, "Non-orthogonal multiple access in large-scale heterogeneous networks," *IEEE Journal on Selected Areas in Communications*, vol. 35, no. 12, pp. 2667–2680, 2017.

[12] Z. Han, D. Niyato, W. Saad, and T. Başar, *Game Theory for Next Generation Wireless and Communication Networks: Modeling, Analysis, and Design*. Cambridge University Press, 2019.

[13] H.-Y. Shi, W.-L. Wang, N.-M. Kwok, and S.-Y. Chen, "Game theory for wireless sensor networks: a survey," *Sensors*, vol. 12, no. 7, pp. 9055–9097, 2012.

[14] A. Yazidi, X. Zhang, L. Jiao, and B. J. Oommen, "The hierarchical continuous pursuit learning automation: a novel scheme for environments with large numbers of actions," *IEEE Transactions on Neural Networks and Learning Systems*, vol. 31, no. 2, pp. 512–526, 2019.

[15] P. Sastry, V. Phansalkar, and M. Thathachar, "Decentralized learning of nash equilibria in multi-person stochastic games with incomplete information," *IEEE Transactions on systems, man, and cybernetics*, vol. 24, no. 5, pp. 769–777, 1994.

[16] R. Vafashoar and M. R. Meybodi, "Reinforcement learning in learning automata and cellular learning automata via multiple reinforcement signals," *Knowledge-Based Systems*, vol. 169, pp. 1–27, 2019.

[17] A. Yazidi, H. L. Hammer, K. Samouylov, and E. Herrera-Viedma, "Game-theoretic learning for sensor reliability evaluation without knowledge of the ground truth," *IEEE Transactions on Cybernetics*, pp. 1–11, 2020.

[18] P. Zhou, Y. Chang, and J. A. Copeland, "Reinforcement learning for repeated power control game in cognitive radio networks," *IEEE Journal on Selected Areas in Communications*, vol. 30, no. 1, pp. 54–69, 2011.

[19] W. Yuan, H. Leung, W. Cheng, and S. Chen, "Optimizing voting rule for cooperative spectrum sensing through learning automata," *IEEE transactions on vehicular technology*, vol. 60, no. 7, pp. 3253–3264, 2011.

[20] Y. Xu, J. Wang, Q. Wu, A. Anpalagan, and Y.-D. Yao, "Opportunistic spectrum access in unknown dynamic environment: A game-theoretic stochastic learning solution," *IEEE transactions on wireless communications*, vol. 11, no. 4, pp. 1380–1391, 2012.

[21] Y. Xu, Q. Wu, J. Wang, L. Shen, and A. Anpalagan, "Opportunistic spectrum access using partially overlapping channels: Graphical game and uncoupled learning," *IEEE Transactions on Communications*, vol. 61, no. 9, pp. 3906–3918, 2013.

[22] A. Yazidi, M. A. Pinto-Orellana, H. Hammer, P. Mirtaheri, and E. Herrera-Viedma, "Solving sensor identification problem without knowledge of the ground truth using replicator dynamics," *IEEE Transactions on Cybernetics*, 2020.

[23] L. Jia, Y. Xu, Y. Sun, S. Feng, L. Yu, and A. Anpalagan, "A game-theoretic learning approach for anti-jamming dynamic spectrum access in dense wireless networks," *IEEE Transactions on Vehicular Technology*, vol. 68, no. 2, pp. 1646–1656, 2018.

[24] Y. Fu, Y. Chen, and C. W. Sung, "Distributed downlink power control for the non-orthogonal multiple access system with two interfering cells," in *2016 IEEE International Conference on Communications (ICC)*. IEEE, 2016, pp. 1–6.

[25] C. W. Sung and Y. Fu, "A game-theoretic analysis of uplink power control for a non-orthogonal multiple access system with two interfering cells," in *2016 IEEE 83rd Vehicular Technology Conference (VTC Spring)*. IEEE, 2016, pp. 1–5.

[26] J. Choi, "A game-theoretic approach for noma-aloha," in *2018 European Conference on Networks and Communications (EuCNC)*. IEEE, 2018, pp. 54–9.

[27] R. Aldebes, K. Dimyati, and E. Hanafi, "Game-theoretic power allocation algorithm for downlink noma cellular system," *Electronics Letters*, vol. 55, no. 25, pp. 1361–1364, 2019.

[28] E. E. Tsiropoulou, P. Vamvakas, and S. Papavassiliou, "Joint customized price and power control for energy-efficient multi-service wireless networks via s-modular theory," *IEEE Transactions on Green Communications and Networking*, vol. 1, no. 1, pp. 17–28, 2017.

[29] L. Xiao, Y. Li, C. Dai, H. Dai, and H. V. Poor, "Reinforcement learning-based noma power allocation in the presence of smart jamming," *IEEE Transactions on Vehicular Technology*, vol. 67, no. 4, pp. 3377–3389, 2017.

[30] S. Riaz, J. Kim, and U. Park, "Evolutionary game theory-based power control for uplink noma." *KSII Transactions on Internet & Information Systems*, vol. 12, no. 6, 2018.

[31] K. N. Doan, M. Vaezi, W. Shin, H. V. Poor, H. Shin, and T. Q. Quek, "Power allocation in cache-aided noma systems: Optimization and deep reinforcement learning approaches," *IEEE Transactions on Communications*, vol. 68, no. 1, pp. 630–644, 2019.

[32] E. E. Tsiropoulou, P. Vamvakas, and S. Papavassiliou, "Joint customized price and power control for energy-efficient multi-service wireless networks via s-modular theory," *IEEE Transactions on Green Communications and Networking*, vol. 1, no. 1, pp. 17–28, 2017.

[33] 3rd Generation Partnership Project (3GPP), "Study on downlink multi-user superposition transmission for LTE, TSG RAN Meeting 67," *Tech. Rep. RP-150496*, Mar 2015.

[34] L. Daniel and K. Narayanan, "Congestion control 2: Utility, fairness, and optimization in resource allocation," *Mathematical Modelling for Computer Networks-Part I*, pp. 2–1, 2013.

[35] R. Cole, Y. Dodis, and T. Roughgarden, "Pricing network edges for heterogeneous selfish users," in *Proceedings of the thirty-fifth annual ACM symposium on Theory of computing*, 2003, pp. 521–530.

[36] A. E. Abdulla, Z. M. Fadlullah, H. Nishiyama, N. Kato, F. Ono, and R. Miura, "An optimal data collection technique for improved utility in uas-aided networks," in *IEEE INFOCOM 2014-IEEE Conference on Computer Communications*. IEEE, 2014, pp. 736–744.

[37] F. Salehisadaghiani and L. Pavel, "Distributed nash equilibrium seeking: A gossip-based algorithm," *Automatica*, vol. 72, pp. 209–216, 2016.

[38] J. R. Marden and J. S. Shamma, "Game theory and distributed control," in *Handbook of game theory with economic applications*. Elsevier, 2015, vol. 4, pp. 861–899.

[39] Y. Xing and R. Chandramouli, "Stochastic learning solution for distributed discrete power control game in wireless data networks," *IEEE/ACM Transactions on networking*, vol. 16, no. 4, pp. 932–944, 2008.

[40] O.-C. Granmo, B. J. Oommen, S. A. Myrer, and M. G. Olsen, "Learning automata-based solutions to the nonlinear fractional knapsack problem with applications to optimal resource allocation," *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, vol. 37, no. 1, pp. 166–175, 2007.